NUMERICAL ANALYSIS – THEORY AND APPLICATION

Edited by **Jan Awrejcewicz**

INTECHWEB.ORG

Numerical Analysis – Theory and Application
Edited by Jan Awrejcewicz

Published by InTech
Janeza Trdine 9, 51000 Rijeka, Croatia

Publishing Process Manager Ana Nikolic
Technical Editor Teodora Smiljanic
Cover Designer Jan Hyrat
Image Copyright pashabo, 2011. Used under license from Shutterstock.com

First published August, 2011
Printed in Croatia

A free online edition of this book is available at www.intechopen.com
Additional hard copies can be obtained from orders@intechweb.org

Numerical Analysis – Theory and Application, Edited by Jan Awrejcewicz
p. cm.
ISBN 978-953-307-389-7

Contents

Preface

This book focuses on introducing theoretical approaches of numerical analysis as well as applications of various numerical methods to either study or solving numerous physical and engineering problems.

Since a large number of pure theoretical research is proposed and a large amount of applications oriented numerical simulation results is given, the book can be useful for both theoretical and applied research aimed at numerical simulations.

In addition, in many cases the presented approaches can be applied directly either by theoreticians or engineers.

The book consists of two parts devoted to theory and application. Part 1 (Theory) consists of eleven chapters. In **chapter 1.1** Aklilu T. G. Giorges illustrates numerical solutions of elliptic and parabolic equations using both finite element and finite difference methods. Author showed how finite element method used discrete elements to obtain the approximate solution of the governing differential equation. Furthermore, author explained how the final system equation was constructed from the discrete element equations and also how finite difference method used points over intervals to define the equation and the combination of all the points to produce the system equation.

Chapter 1.2 authored by Eduardo B. De Amôres, summarized the utilization of large data sets in galactic astronomy where most of them covered almost entire area of the sky in several wavelengths. For both the diffuse data were provided by IRAS, DIRBE/COBE, molecular and hydrogen surveys and point sources catalogues were provided by stellar large-scale surveys such as DENIS, 2MASS, SDSS, among others. A brief specification of these surveys and how to access them in the context of Virtual Observatory was introduced. Concerning HI model to describe spiral arms positions from HI data, the results presented allowed to obtain the spiral arm positions based on HI distribution obtaining the spiral arm parameters (r_0, θ_0, i, $\Delta\theta$), which reproduced the main observed features in the -v diagrams for HI. Using the Besançon Galaxy Model and the 2MASS data, Dr. Amôres performed a detailed analysis of the tangential directions from near infrared star counts.

The aims of **chapter 1.3** coauthored by Sergey Abramov et al., were to consider different approaches to robust regression, to compare their performance, to discuss

possible limitations and restrictions, and to give some practical recommendations. The scatter-plot or cluster-center representations were the basis for other operations (curve regression) applied at several application. Secondly, with simulated noise for test images, the studies showed that even the local estimates considered normal could be considerably biased. Furthermore, the weighted methods of LMS regression using cluster centres specified their advantages and what was as well important was a priori information on mixed or signal-dependent noise. The experiments were carried out: those assuming that a model of mixed noise was valid and second ones with simulated noise for i.i.d. noise. Finally, the goal of estimating mixed noise parameters was to use the obtained estimates at later stages of image processing**.**

In **chapter 1.4** Denis Benasciutti et al. developed alternative FE methods, which would allow to achieve the solution of complex three-dimensional problems through a combination of several simpler and faster one- and two-dimensional analyses, which usually require reduced computational efforts. Authors focused on mechanical and thermal problems, in which the structure was axisymmetric, but not the load. There are two aspects of this work: first is to provide a theoretical background on the use of semi-analytical FE approach in numerical analysis of axisymmetric structures loaded non-axialsymmetrically. Two original results were obtained: a plane axi-antisimmetric FE model for solving axisymmetric components loaded in torsion, and a semi-analytical approach for the analysis of plane axisymmetric bodies under non-axisymmetric thermal loadings. Authors' second aim was to explain some practical aspects in the application of semi-analytical method to engineering problems.

Sébastien Besset and Louis Jézéquel introduced in **chapter 1.5** several criteria corresponding to different vibrational propagation paths based on modal motion equations, which allowed for working with small-sized matrices. An optimization criteria founded on a multimodal description of complex structures was proposed. The modal synthesis technique presented was based on the double and triple-modal synthesis. The double modal synthesis operated by introducing generalized boundary coordinates in order to describe substructure connections. The triple modal synthesis consisted of representing the interior points of the fluid by acoustic modes, the describing of the boundary forces between the fluid and each substructure through the use of a set of loaded modes and consisted of describing the boundary forces between each substructure by introducing another set of loaded modes. To sum up, this work was mainly focused on the above mentioned triple modal synthesis method which introduced the acoustic parts of the coupled system using acoustic modes.

Chapter 1.6 authored by Kornkanok Bunwong developed the way to approximate higher order quantities and applied them to ecological problems. It was established that the new approach was suitable for a model evolving according to the transition rates affecting additionally by neighbors. The SIS epidemic model, as an example, proved that if continuous time scale is used, then two solutions of the system would be asymptotically stable or unstable depending on parameter values and stable oscillating solutions would never exist. But if discrete time scale was applied, then

various types of solution behaviors would appear such as equilibrium point solutions, period two cycles, period four cycles, period three cycles, and also chaotic solutions depending on parameter values as well.

Raquel Caballero-Águila et al. introduced in **chapter 1.7** the state estimation problem for nonlinear discrete-time systems with uncertain observations, when the evolution of the state is governed by nonlinear functions of the state and noise, and the additive noise of the observation is correlated with that of the state. In this chapter, a recursive unscented filtering algorithm for state estimation in a class of nonlinear discrete-time stochastic systems with uncertain observations was obtained. The authors propose a filtering algorithm based on the scaled unscented transformation, which provided approximations of the first and second-order statistics of a nonlinear transformation of a random vector. Furthermore, the system model was showed, the nonlinear state transition model. Apart from that, the least-squares estimation problem from uncertain observation is formulated and a brief review of the unscented transformation and the scaled unscented transformation is presented. Next, the estimation algorithm was derived using the unscented filtering procedure and the filter update accomplished by the Kalman filter equations. Finally, the performance of the proposed unscented filter was shown by a numerical simulation example, where a first order ARCH model was considered to describe the state evolution.

Ricardo Jauregui and Ferran Silva in **chapter 1.8** emphasized that all the techniques which were presented can be used not only to validate the numerical methods and simulation but in other areas that require a quantitative comparison of complex data. The significant thing, when a validation method is chosen, was that it had to provide a similar result to the expert opinion, which implied an objective analysis of the data. The emphasis is on that a perfect method to validate any kind of result did not exist. Each method presented advantages and disadvantages depending on the type of data and the type of analysis. The following items were worth considering in author's opinion: the implementation of the validation technique, the validation method should reflect human opinions, method should provide the possibility to be applied in different environments and/or applications, method should be commutative and must analyse the difference between the two data sets and always yield the same result.

In **chapter 1.9** Cruz Meneses-Fabian et al. discussed the mathematical fundamentals of parallel projection tomography and demonstrated the mathematical method for directional edge-enhancement tomography. A mathematical model was described thanks to obtaining the reconstruction of tomographic images with enhanced edges, and also experimental implementation were shown, which were applied to optical tomography of phase objects. Authors proved that the mathematical model was based on the establishment of the relation existent between the Radon transform (RT) and the 2-D directional Hilbert transform (HT). Furthermore, authors introduced a description of the experimental possibility, beginning with the relation existent between the projection and the phase of the optical wave, when it transversed a thin phase object, continuing with a description of the optical image-forming system 4*f* in

order to obtain the HT of the optical field that had been produced after crossing the object. In the end, authors added a description of the theoretical relationship between the experimental procedures used to obtain the image reconstruction with their enhanced edges in a directional manner.

The main aim of **chapter 1.10** coauthored by Emil Petre and Dan Selişteanu was to provide the mathematical tools, which were used for numerical methods, for solving PDEs and to give a brief outline of the techniques. This chapter deals with the approximation and simulations of the dynamical model for a class of nonlinear propagation bioprocesses. Furthermore, the control problem of these classes of propagation bioprocesses was analysed for which a class of nonlinear adaptive controllers was designed based on their finite order models and on the input-output linearizing techniques. At the beginning authors introduced the distributed parameter dynamical model for the class of fixed bed reactors. Apart from that, an analysis of obtained results by application of this method in the case of a fixed bed reactor without diffusion were also presented and the adaptive control strategies of propagation bioreactors. The authors introduced the performances of the designed adaptive controllers and demonstrated the simulation obtained results which the designed adaptive algorithms used in control of propagation bioreactors yield good results closely comparable to those obtained in the case when the process parameters were known.

Saeid Zahiri in **chapter 1.11** described numerical simulation with meshfree methods. Author introduced three categories and their limitations, applications, advantages and other descriptions and discussed the definition of base and shape functions and various techniques for meshfree shape function constructions. These shape functions were locally supported, because only a set of field nodes in a small local domain were used in the construction. Such a local domain was termed the support domain or influence domain. The author also discusses the point interpolation method (PIM) in detail, which was useful for creating meshfree shape functions. Author showed a scalar function defined in the problem domain that was represented by a set of scattered nodes. Polynomial basis functions and radial basis functions (RBF) were often used in meshfree methods and were also discussed by the author. The heat transfer problem as well as solid and fluid mechanics problems were solved with meshfree methods. Finally, three meshfree categories, which were used to solve the problems, were strong form methods, weak form methods and weak-strong form methods (MWS).

Part 2 (Applications) comprises eighteen chapters. In **chapter 2.1** Menglan Duan and Jinghao Chen introduce the numerical calculation for soil-riser interaction, vortex-induced vibration (VIV), fatigue, the coupling of floating vessel and riser, riser installation, etc, and provide a theoretical basis of (steel catenary riser) SCR design, which is a flexible steel pipe that conducts well fluids from the subsea wellhead to the production floating vessel. This study introduced the numerical simulation methods commonly used in offshore industry. Authors admitted that the SCR had advantages of low manufacturing cost, resistance of high temperature and high pressure, and a

good adaptability of upper floating body's motion. SCR numerical simulation demonstrated great advances, commercial software was developed for SCR design, but as authors mentioned, there were uncertainties on mechanical characteristic of SCR. In authors' opinion, the challenges for SCR design were as follows: pipe-soil interaction mechanism, turbulence and the coupled effects between hull and riser which shouldn't be neglected in the future.

Chapter 2.2 coauthored by Filote Constantin and Ciufudean Calin summarizes a comparison of the performances among three rotor flux observers, which were the vector control strategies according to the type of drive-controlled flux. The authors claim that if the rotor flux is applied as criterion in the vector control of induction motor, the value and direction of the flux needs to be known. This work analysed the performances of a conventional rotor flux simulator with a view to the temperature influence of the rotor resistance. Flux observers were used to estimate the flux. Authors analysed the performances of a robust-adaptive rotor flux observer, starting from a mathematical model and using simulation. One of part of this study presents the analysis of conventional flux simulators based on the current and tension model of the induction model. Furthermore, authors introduced the adaptive flux observer, presented simulation tests of its robustness in rotor resistance variation with temperature and closed-loop vector control system with robust-adaptive flux observer. Correct estimation methods of the rotor flux magnitude and position were checked and verified if the system oriented itself after the rotor flux direction.

Christian Maria Firrone and Stefano Zucca analysed the numerical methods currently employed to simulate the forced response of turbine bladed disks with friction interfaces in **chapter 2.3.** Furthermore, the balance equation of the bodies in contact were deduced in the frequency domain by means of the harmonic balance method and the contact elements were described due to highlight of their main features and their effect on the dynamics of the system. Authors also studied the effect of an uncoupled solution strategy based on a preliminary static analysis followed by the dynamic analysis and the critical issues arising when the methods were applied to full scale applications. The study also presents typical configurations of friction contacts in turbine bladed disks and the effect of the friction contacts on the forced response curves are computed.

In **chapter 2.4** coauthored by Jan Awrejcewicz et al., the results of stress and strain analysis of an orbital cavity are presented. The study provides an assistance for surgeons performing bony face operations. The aim of this work was to develop the numerical model of a bottom arch of an orbital cavity using a FEM. Furthermore, the model of a healthy orbit, which was based on the data obtained from computer tomography, was proposed. Modeling of an orbital cavity using finite element method and a model of a double layered pelvic bone were presented as well as some phenomena during leg flexion, extension, adduction and abduction. The authors introduced some simplifications of the model. The aim of the chapter was to show the algorithm and also to speed up the calculations. It was decided to use simple materials properties. Additionally, one part of the work dealt with oscillations of coupled DNA

bases which made substantial contribution to the process of opening DNA base pairs. Authors analyzed the dynamical behavior of the model system, investigated its stability and constructed the diagram of bifurcations.

Roxana Grigore et al. in **chapter 2.5** introduced a simplified model for a plate heat exchanger in a counter flow arrangement. They showed a model which was in concordance with the experimental results and with the results from theoretical analysis. Also, a relative degree of uncertainty was introduced by the criterial relations, which was used to calculate convection heat transfer coefficients. Numerical simulation offered a good understanding of the temperature distribution and fluid flow under turbulent motion. This study presented a theoretical and experimental study on plate heat exchanger. A numerical simulation of a counter flow plate heat exchanger was performed using finite element method. A 3D model was developed to analyze thermal transfer and fluid flow along the plate heat exchanger, using COSMOS/Flow program. The results were presented graphically and numerically and validation of the models presented was done by comparing the measured values obtained by an experimental study.

The main challenges for the R&D of aerospace plasticity technology are summarized by Lianggang Guo and He Yang in **chapter 2.6** having the unique requirements of light weight, high precision, high performance, high reliability and high efficiency for the plasticity forming manufacture of various key aerospace components. In this work, a high-end research route for aerospace plasticity technology is presented in terms of our understanding and research experiences on various metal forming processes and an application example is given for the investigation of radial-axial ring rolling technology. Furthermore, the authors discussed the involved key FE modelling technologies and reliability of the developed thermo-mechanical coupled 3D-FE model for the entire radial-axial ring rolling process, some simulation results including ring geometry evolution, stress field, strain field, temperature field, rolling forces and torques in the radial and axial directions during the process.

In **chapter 2.7** Takashi Harada proposed a new parallel link mechanism with multi drive linear motors (MDLMs) due to expansion of this limited application of PLM. The multi drive was a control method for linear motors where a number of moving parts were individually driven on one stator part. The authors investigated the kinetostatics (kinematics and static force), and dynamics characteristics of the 3D4M PLM by usage of symbolic mathematical analysis and numerical simulations. In short, in this work configurations of the 3D4M PLM on multi drive linear motors are introduced and kinematic equations, forward kinematics and derivative kinematics of the 3D4M PLM are derived. Furthermore, singularity and static forces of the 3D4M PLM are analyzed using Mathematica and the decoupled dynamical design of the 3D4M PLM are introduced.

Guang-Ping He and Zhi-Yong Geng in **chapter 2.8** present a novel mechanism for one-legged hopping robot, which is proposed on the basis of dynamic synthesis. The

proposed hopping robot mechanism is a non-SLIP model system, which shows more biological characteristics while the control problem of it is intractable, due to the complex nonlinear dynamics and the second-order nonholonomic constraints. In this study, authors introduced the novel mechanism and investigated its dynamics. Furthermore, the proposition that confirmed the nonlinear dynamics could be transformed into the strict feedback normal form. Then, a sliding mode back stepping control and the exponential stability are introduced and proved. The motion planning method for the hopping system instance phase and the feasibility of the mechanism and the stability of the control verified by some numerical simulations is presented by the authors.

In **chapter 2.9** authored by Z.Y. Jiang, a new model for rolling mechanics of thin strip in cold rolling is developed. In this work, strip plastic deformation-based model of the rolling force in the calculation is employed, and a modified semi-infinite body model is introduced to calculate the flattening between the work roll and backup roll, and the flattening between the work roll and strip. A Foppl model was employed to calculate the edge contact between the upper and down work rolls. The special rolling and strip deformation was simulated using a modified influence function method based on the theory of the slit beam. By the calculated result, author showed that the specific forces such as the rolling force, intermediate force and the shape and profile of the strip for this special rolling process were different from the forces in the cold rolling process and those from a new theory of metal plasticity in metal rolling.

Miguel Castillo-Vazquez et al. in **chapter 2.10** presented investigation of the impact of both SCR on channel characteristics. By numerical simulations, the main performance indicators of two link configurations are shown, formed by a MBT and the proposed SCR. Two points were investigated (a) the effect of transmitter spots size and ambient light sources (natural and artificial) on SNR and channel bandwidth (BW), and (b) the impact of the receivers total FOV and blockage on the transmitter power requirements. The results which were obtained by the authors in all simulations show the robustness and weaknesses of each receiver structure and prove a great potential of both SCR when operating in a multispot diffusing configuration. The study investigated the characteristics and structure of single-channel receivers, and the transmitter as well as ambient light models in the numerical analysis. Finally, the performance evaluation of receivers was carried out.

Chapter 2.11 authored by Hirohisa Kojima focused on an integrated image processing method to estimate the attitude variation of a satellite. The proposed research consisted of six steps: searching the position of a target satellite in an image using color information, extraction of feature points on the satellite using a Harris corner detector, optical flow estimation by template matching and random sample consensus, deleting incorrect optical flow using the eight-point algorithm, initial guess of the rotational axis and attitude variation from the optical flow by a heuristic approach, and an iterative method to obtain the precise rotational axis and attitude variation from the initial guess. Author proved that feature points and optical flow of rotating

target could be extracted from images taken by only one camera. The effect which was obtained by using the Harris corner detector, template matching, and RANSAC, and by removing the undesired points according to the RGB color information and the length of the optical flows and the eight-point algorithm was used to obtain a more reliable essential matrix subject to the optical flow. The studies which were introduced by the author also showed that the estimated rotational axis vector and attitude variation agree roughly with the correct values under a good lighting condition.

The aim of **chapter 2.12** authored by Tomasz Kopecki is to draw attention to gravity of the factor integrating nonlinear numerical analysis with an experiment. The author presented a methodology that could be used for assessment and current improvement of numerical models ensuring correct interpretation of results which were achieved from nonlinear numerical analyses of a structure. The author carried out experimental examination of selected crucial elements of load-carrying structures parallel with their nonlinear numerical analysis. Finally, the factors determining proper realization of adequate experiments were discussed with emphasis placed on the role which the model tests could play as a fast and economically justified research tool that could be used in the course of design work on thin-walled load-carrying structures.

Millo Federico in **chapter 2.13** presented the matter of ground transportation industry, which accepted the reality that fast, efficient, and cost effective engine and vehicle development necessitate the use of numerical simulation at every stage of the design process. Within the vehicle powertrain design and development process, three different levels of modelling were generally found and shown by the author: detailed modelling, Software in the Loop (SiL) modelling and Hardware in the Loop (HiL) modelling. Furthermore, this chapter provided a description of different methodologies, which could support engineers in each phase of the vehicle powertrain design process. Author presented the analysis of numerical models for the main powertrain subsystems. In the end, two case studies of numerical simulation applied to powertrain development were introduced, the first focused on the evaluation of vehicle efficiency, paying particular attention to the engine behavior under transient conditions, the second aimed at the assessment of the fuel economy potential of different Hybrid Electric Vehicle architectures.

In **chapter 2.14** Sebastien Roth et al. pay attention to theoretical foundations of crash analysis and show how this simulation step could be integrated in the design process. Explicit Finite Element software as Radioss might be used to the crash analysis, but many difficulties could arise during this analysis. Authors claim that problems could come from the size of the model which could generate a time consuming simulation and a particular point concerned the way to transfer CAD models towards finite element model without loss of information. Problems of standard exchange and the data management were examined. Finally, the authors assume that last decades have shown the development of numerical simulation which became essential in the design process, especially in automotive engineering.

Takafumi Suzuki and Masaki Takahashi introduced real time control method of simultaneously translational and rotational motions for an autonomous mobile robot in **chapter 2.15**. This method employed omni-directional platform for the drive system and are founded on the fuzzy potential method (FPM). The novel design method of potential membership function (PMF) is shown. In accordance to this method, the wide-robot could decide the current direction of translational motion to avoid obstacles safely by using capsule case. Through controlling the rotational motion in parallel with the translational motion in real time, the wide-robot is able to go through narrow distance between two objects. The effectiveness is specified by numerical simulations and simplified experiments. Authors have shown that the proposed method enables simultaneous control of the translational and rotational velocity within the framework of FPM.

Takafumi Suzuki and Masaki Takahashi in **chapter 2.16** summarize a real-time obstacle avoidance method introducing the velocity of obstacle relative to the robot. In his study, virtual distance function is described which is founded on distance from the obstacle and speed of obstacle, but only the projection of the obstacle velocity on the unit vector from the obstacle to the robot was considered. Authors applied the method to an autonomous mobile robot which played soccer. By correct designing of potential membership function (PMF), it proved that wheeled robots got to the goal with conveying a soccer ball and avoiding obstacles. The study showed for the purpose of avoiding the moving obstacle safely and smoothly, designed methods of the potential membership function (PMF), should consider the velocity of the obstacle relative to the robot. Numerical simulations and simplified experiments were performed.

Qingzhi Yan in **chapter 2.17** mainly studied on the models and mechanism of steel sheet pile, and proposed two kinds of instability problems: first, the supporting structure which had not enough strength or stiffness to support the load and there were several destruction forms including support buckling, pull-anchor damage, excessive deformation of the supporting structure and bending failure. The second matter was the soil instability of the foundation pit. In this work, the mechanics method was used to obtain the code formula from a reasonable discussion and systematical analysis. First, the equivalent beam method and “m” method of elastic foundation beam methods were used to reach a conclusion that the finite element method was a more ideal stability analysis method, which could be used to deal with the strength problems and deformation problems. Second, according to the different steel sheet pile supporting basic form, it put forward different form of steel sheet pile foundation pit supporting overall sliding stability analysis superposition methods. The two problems combined: focusing on the soil and steel sheet pile between interface slippage characteristics of plane strain finite element method and the realization methods in nonlinear finite element numerical analysis method software.

In **chapter 2.18** Seiya Ueno and Takehiro Higuchi confirms that there has been no research on control law to deal with uncertain information. This work proposes control law that treated uncertain information by providing new performance by

enabling the aircraft to obtain information and to check the certainty of the information. Two cases of collision avoidance control were carried out to see the effect of the information amount as parameter for control. The first case specified the problem as the uncertainty of the information changes by the relative position of the evader and the target. The second example defined the problem as the uncertainty of the information was given as absolute position. Both cases introduced smoother and safer trajectories than the conventional control laws. The simulation results showed that the control laws using information amounts did not depend on the coordinates.

I do hope that the presented book will be useful to academic researchers, engineers as well as post-graduate students. I would like to acknowledge my working visit to Darmstadt, Germany supported by the Alexander von Humboldt Award which also allowed me time to devote to the book preparation. I would like to thank Ms Ana Nikolic for her professional support and advice while preparing the book.

Jan Awrejcewicz
Technical University of Łódź,
Poland

Part 1

Theory

1

Finite Element and Finite Difference Methods for Elliptic and Parabolic Differential Equations

Aklilu T. G. Giorges
Georgia Tech Research Institute, Atlanta, GA, USA

1. Introduction

With the availability of powerful computers, the application of numerical methods to solve scientific and engineering problems is becoming the normal practice in engineering and scientific communities. Well-formed scientific theory with numerical methods may be used to study scientific and engineering problems. The numerical methods flourish where an experimental work is limited, but it may be imprudent to view a numerical method as a substitute for experimental work.

The growth in computer technology has made it possible to consider the application of partial deferential equations in science and engineering on a larger scale than ever. When experimental work is cost prohibitive, well-formed theory with numerical methods may be used to obtain very valuable information. In engineering, experimental and numerical solutions are viewed as complimentary to one another in solving problems. It is common to use the experimental work to verify the numerical method and then extend the numerical method to solve new design and system. The fast growing computational capacity also make it practical to use numerical methods to solve problems even for nontechnical people.

It is a common encounter that finite difference (FD) or finite element (FE) numerical methods-based applications are used to solve or simulate complex scientific and engineering problems. Furthermore, advances in mathematical models, methods, and computational capacity have made it possible to solve problems not only in science and engineering but also in social science, medicine, and economics. Finite elements and finite difference methods are the most frequently applied numerical approximations, although several numerical methods are available.

Finite element method (FEM) utilizes discrete elements to obtain the approximate solution of the governing differential equation. The final FEM system equation is constructed from the discrete element equations. However, the finite difference method (FDM) uses direct discrete points system interpretation to define the equation and uses the combination of all the points to produce the system equation. Both systems generate large linear and/or nonlinear system equations that can be solved by the computer.

Finite element and finite difference methods are widely used in numerical procedures to solve differential equations in science and engineering. They are also the basis for countless engineering computing and computational software. As the boundaries of numerical method applications expand to non-traditional fields, there is a greater need for basic understanding of numerical simulation.

This chapter is intended to give basic insight into FEM and FDM by demonstrating simple examples and working through the solution process. Simple one- and two-dimensional elliptic and parabolic equations are used to illustrate both FEM and FDM. All the basic mathematics is presented by considering a simplistic element type to define a system equation. The next section is devoted to the finite element method. It begins by discussing one- and two-dimensional linear elements. Then, a detailed element equation, and the forming of a final system equation are illustrated by considering simple elliptic and parabolic equations. In addition, a small number of approximations and methods used to simplify the system equation are, presented. The third section presents the finite difference method. It starts by illustrating how finite difference equations are defined for one- and two-dimensional fields. Then, it is followed by illustrative elliptic and parabolic equations.

2. Finite element method

Of all numerical methods available for solving engineering and scientific problems, finite element method (FEM) and finite difference methods (FDM) are the two widely used due to their application universality. FEM is based on the idea that dividing the system equation into finite elements and using element equations in such a way that the assembled elements represent the original system. However, FDM is based on the derivative that at a point is replaced by a difference quotient over a small interval (Smith, 1985).

It is impossible to document the basic concept of the finite element method since it evolves with time (Comini et al. 1994, Yue et al. 2010). However, the history and motivation of the finite element method as the basis for current numerical analysis is well documented (Clough, 2004; Zienkiewicz, 2004).

Finite element starts by discretizing the region of interest into a finite number of elements. The nodal points of the elements allow for writing a shape or distribution function. Polynomials are the most applied interpolation functions in finite element approximation. The element equations are defined using the distribution function, and when the element equations are combined, they yield a continuous equation that can approximate the system solution. The nodal points and corresponding functional values with shape function are used to write the finite element approximation (Segerlind, 1984):

$$\psi = N_1\psi_1 + N_2\psi_2 + \cdots + N_m\psi_m \tag{1}$$

where $\psi_1, \psi_2, \ldots \psi_m$are the functional values at the nodal points, and $N_1, N_2, \ldots N_m$ are the shape functions. Thus, the system equation can be expressed by nodal values and element shape function.

2.1 One-dimensional linear element

Before we discuss the finite element application, we present the simple characteristic of a linear element. For simplicity, we will discuss only two nodes-based linear elements. But, depending on the number of nodes, any polynomial can be used to define the element characteristics. For two nodes element, the shape functions are defined using linear equations. Fig.1 shows one-dimensional linear element.

The one-dimensional linear element (Fig. 1) is defined as a line segment with a length (l) between two nodes at x_i and x_j . The node functional value can be denoted by ψ_i and ψ_j. When using the linear interpolation (shape), the value ψ varies linearly between x_i and x_j as

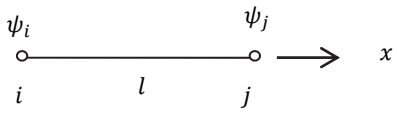


Fig. 1. One-dimensional linear element

$$\psi = mx + k \tag{2}$$

The functional value $\psi = \psi_i$ at node $i = x_i$ and $\psi = \psi_j$ at $j = x_j$. Using the functional and nodal values with the linear equation Eq. 2., the slope and the intercept are estimated as

$$m = \frac{\psi_j - \psi_i}{x_j - x_i} \text{ and } k = \frac{\psi_i x_j - \psi_j x_i}{x_j - x_i} \tag{3}$$

Substituting m and k in Eq. 2 gives

$$\psi = \left(\frac{\psi_i x_j - \psi_j x_i}{x_j - x_i}\right) + \left(\frac{\psi_j - \psi_i}{x_j - x_i}\right) x \tag{4}$$

Rearranging Eq. 4 and substituting l for the element size $(x_j - x_i)$ yields

$$\psi = \left(\frac{x_j - x}{l}\right) \psi_i + \left(\frac{x - x_i}{l}\right) \psi_j \tag{5}$$

By defining the shape functions as

$$\begin{aligned} N_i &= \left(\frac{x_j - x}{l}\right) \quad \text{(a)} \\ N_j &= \left(\frac{x - x_i}{l}\right) \quad \text{(b)} \end{aligned} \tag{6}$$

By introducing the shape function N_i and N_j in Eq. 5, the finite element equation can be rewritten as

$$\psi = N_i \psi_i + N_j \psi_j \tag{7}$$

The above equation is a one-dimensional linear standard finite element equation. It is represented by the shape functions $N_i$and N_j nodal values ψ_i and ψ_j.
The two shape functions profiles for a unit element are shown in Fig. 2. The main characters of the shape functions are depicted. These shape functions have a value of 1 at its own node and 0 at the opposing end. The two shape functions also sum up to one throughout.

2.2 Two-dimensional rectangular element

With the current computational methods and resources available, it is not clear whether or not using the FEM or modified FDM will provide an advantage over the other. However, in the early days of numerical analysis, one of the major advantages of using the finite element method was the simplicity and ease that FEM allows to solve complex and irregular two-dimensional problems (Clough 2004, Zienkiewicz, 2004, Dahlquist and Bjorck, 1974). Although several element shapes with various nodal points are used in many numerical simulations, our discussion is limited to simple rectangular elements. Our objective is to simply exhibit how two- dimensional elements are applied to define the elements and final system equation.

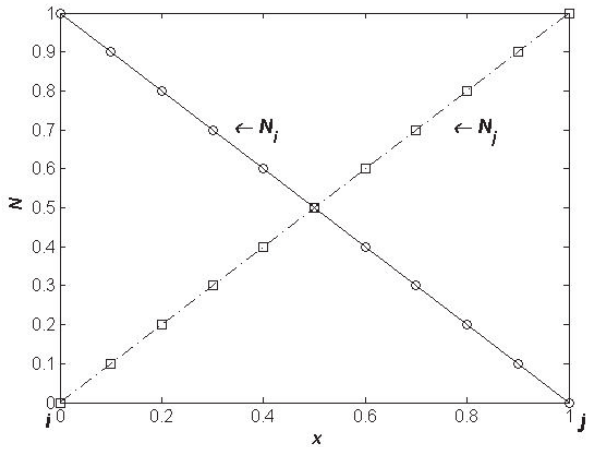


Fig. 2. Linear shape functions

Fig. 3 illustrates a linear rectangular element with four nodes. The nodes $i, j, k,$ and l have corresponding nodal values ψ_i , ψ_j, $\psi_k$and ψ_l at $(x_i, y_{i,}), (x_{j,}, y_{j,}), (x_k , y_k), and\ (x_l, y_l)$.

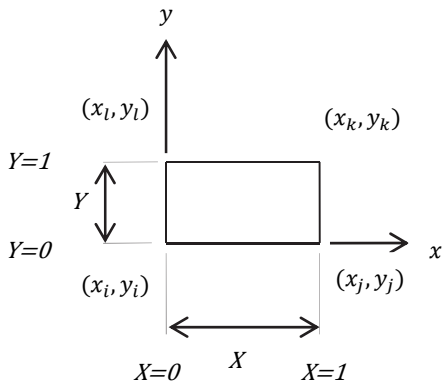


Fig. 3. Two-dimensional linear rectangular element.

The linear rectangular interpolation equation is defined as

$$\psi = k_1 + k_2 x + k_3 y + k_4 xy \tag{8}$$

Applying the nodal and functional values$(x_i, y_{i,}) = (0,0), \psi = \psi_i$,$(x_j, y_{j,}) = (X, 0), \psi = \psi_j$, $(x_k, y_k) = (X, Y)$, $\psi = \psi_k$, and $(x_l, y_l) = (0, Y), \psi = \psi_l$ in Eq. 8 yields four equations and four unknowns as

$$\begin{aligned} \psi_i &= k_1 && (a)\\ \psi_j &= k_1 + k_2 \mathrm{X} && (b)\\ \psi_k &= k_1 + k_2 \mathrm{X} + k_3 \mathrm{Y} + k_4 \mathrm{XY} && (c)\\ \psi_l &= k_1 + k_3 \mathrm{Y} && (d) \end{aligned} \tag{9}$$

Solving the unknown constants k_1 , k_2, $k_3$and k_4 in terms of the nodal values give

$$\begin{aligned} k_1 &= \psi_i && (a)\\ k_2 &= \frac{1}{\mathrm{X}}(\psi_j - \psi_i) && (b)\\ k_4 &= \frac{1}{\mathrm{XY}}(\psi_i - \psi_j + \psi_k - \psi_l) && (c)\\ k_3 &= \frac{1}{\mathrm{Y}}(\psi_l - \psi_i) && (d) \end{aligned} \tag{10}$$

Substituting the above values equations (Eq. 10.) into Eq. 8 and reorganizing in terms of nodal values give to finite element equation as

$$\psi = N_i\psi_i + N_j\psi_j + N_k\psi_k + N_l\psi_l \tag{11}$$

where

$$\begin{aligned} N_i &= \left(1 - \frac{x}{X}\right)\left(1 - \frac{y}{Y}\right) && (a)\\ N_j &= \left(\frac{x}{X}\right)\left(1 - \frac{y}{Y}\right) && (b)\\ N_k &= \left(\frac{x}{X}\right)\left(\frac{y}{Y}\right) && (c)\\ N_l &= \left(1 - \frac{x}{X}\right)\left(\frac{y}{Y}\right) && (d) \end{aligned} \tag{12}$$

Eq. 12 is two-dimensional rectangular shape functions based on element that is plotted in Fig. 3. The shape functions (Eq. 12) are plotted in Fig. 4. The shape functions satisfy the conditions: 1. the functions have a value of 1 at their own node and 0 at the other ends, 2. they vary linearly along the two adjacent edges, and 3. the shape functions sum up to one throughout.

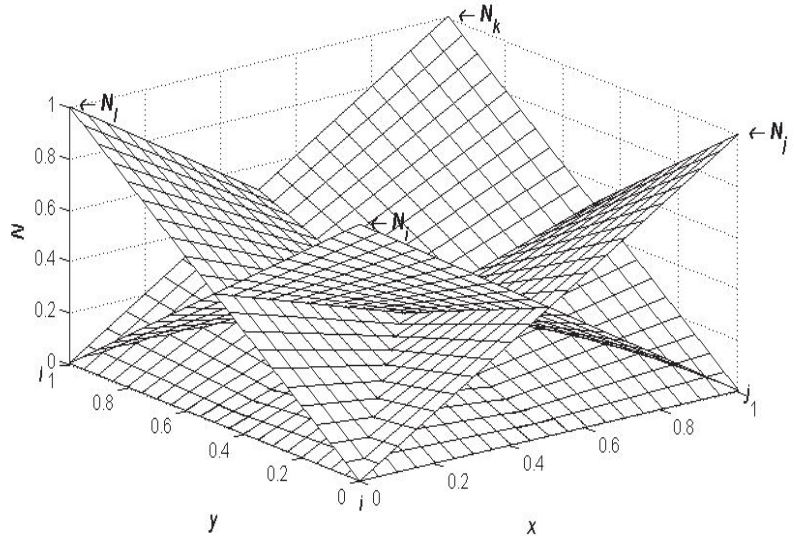


Fig. 4. Two-dimensional rectangular linear element shape functions distribution.

Finite element equation uses the element shape function to define the relationship between the nodal points. Once the element equation is defined, by assembling the element systematically the final system equation is structured. Next, we illustrate the application of the finite element method in a one-dimensional elliptic equation.

2.3 Elliptic equation in finite element method

In order to discuss the basic concept of finite element application in an elliptic equation, we start by illustrating a one-dimensional equation. A one-dimensional elliptic equation of function T can be written as:

$$D\frac{d^2T}{dx^2} + Q = 0 \tag{13}$$

This elliptic equation is used to describe the steady state heat conduction with heat generation where $D, T,$ and Q represent thermal diffusion, temperature, and heat generation. The distinctiveness of the solution of an elliptic equation is dependent on the boundary condition. Thus, it is sometimes called boundary value problem. Providing the appropriate boundary condition at the two ends, the unique solution exists for temperature distribution. The boundary condition could be a prescribed value (Dirichlet), the flux (Neumann), or a combination of both (Vichnevetsky, 1981). In order to demonstrate how the finite element method is used to solve an elliptic equation, we simplify by assuming that material has constant and uniform diffusion with heat generation. Building the finite element elliptic equation involves discretization, forming the element equation, assembling the element equation systematically, and forming the final algebraic equation of the system. Moreover, the uniqueness and the stability of the system equation depend on the specified boundary conditions, thus solving the algebraic equation requires the boundary condition to be introduced before the final equation is solved.

Before we start by forming the finite element equation of steady state heat conduction with heat generation, we have to address how the linear finite element equation is formed. One of the mathematical concepts used to generate the final system equation is called weighting residual method. In short, the weighting residual method is based on the fact that when an approximate solution is substituted in the differential equation, the error term resulted since the approximate solution does not completely satisfy the equation. Thus, the method of weighting residual is to force the product of residual and the weighting function to go to zero. In the finite element method, the weighting residual for each element nodal value is defined and the integral is evaluated using the interpolation function as

$$-\int w(x)R(x)dx = -\int W(x)\left(D\frac{d^2T}{dx^2} + Q\right)dx = 0 \tag{14}$$

where W is the weighting function and R is residual.

The major requirement to evaluate the above integral equation is that the functions that belong to the trail and weighting functions must be continuous. However, when the trail function is linear, the second derivative is not continuous and the integral cannot be evaluated as it is. Thus, in order to evaluate the integral with a lower degree of continuity by replacing the second derivative term with equivalent expression using the differentiation product rules, hence

$$\int W\frac{d^2T}{dx^2}dx = \int \frac{d}{dx}\left(W\frac{dT}{dx}\right)dx - \int \frac{dT}{dx}\frac{dW}{dx}dx \tag{15}$$

Eq. 15 shows that the degree of minimum continuity required to evaluate the integral for trail function is reduced while the continuity for weighting function is increased. The minimum continuity requirement for both weighting and trail can be fulfill with linear function and the integral can be evaluated as long as the functions are continuous within the integral interval. The finite element method is evolved from this need of finding appropriate sets of functions. The finite element method uses a systematic way of using polynomial approximate function that permits the evaluation of the integral equation. Introducing Eq. 15 in Eq. 14 gives the residual for the elliptic integral as

$$R = -\int D\frac{d}{dx}\left(W\frac{dT}{dx}\right)dx + \int D\frac{dT}{dx}\frac{dW}{dx}dx - \int QW\,dx \tag{16}$$

The finite element method uses the interpolation function as a weighting and trail functions. Even a linear element can satisfy the continuity requirement to evaluate the integral. Once we define the integral, in this case function, the next step is to evaluate the residual integral. By evaluating the integral for each element, the element contribution to the final system equation can be determined.

In order to determine the element contribution to the final system equation, we will consider linear element (e) with node i and j (Fig. 1) and evaluating the residual integral (Eq. 16) using the elements interpolation function (Eq. 12). Thus, the residual equation becomes

$$\begin{aligned} R_i^e &= -\int_i^j D\frac{d}{dx}\left(N_i\frac{dT}{dx}\right)dx + \int_i^j D\frac{dT}{dx}\frac{dN_i}{dx}dx - \int_i^j N_i Q dx \quad \text{(a)} \\ R_j^e &= -\int_i^j D\frac{d}{dx}\left(N_j\frac{dT}{dx}\right)dx + \int_i^j D\frac{dT}{dx}\frac{dN_j}{dx}dx - \int_i^j N_j Q dx \quad \text{(b)} \end{aligned} \tag{17}$$

The integral splits into two parts since the weighting functions are defined by two functions N_i and N_j. Consequently, (R_i^e) and (R_j^e) represent the two weighting functions contributions to the element nodal value residual (i) and (j), respectively. Fig. 5 shows that a system of linear interpolation functions. If we take the arbitrary element e that located anywhere in the field, except the two weighting functions N_i and N_j, all of the other weighting equations are zero contribution.

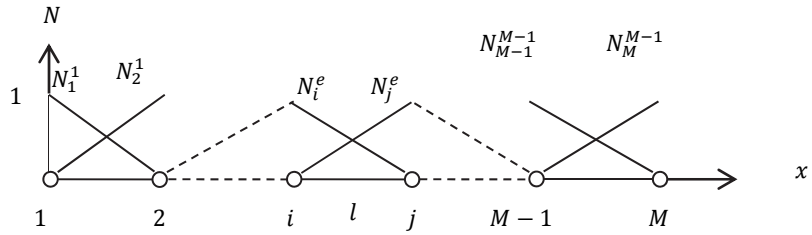


Fig. 5. System of elements shape function and nodes

The integrals on the right can be evaluated using the linear interpolation functions (Eq. 6) characteristics and the finite element equation (Eq. 7). The interpolation function characteristics are $N_i = 1$, at i and $N_i = 0$, at j. Similarly, $N_j = 1$, at j and $N_j = 0$, at i. Evaluating the first terms on the right (denote by q_i^e and q_j^e) gives

$$q_i^e = -\int_i^j D\frac{d}{dx}\left(N_i\frac{dT}{dx}\right)dx = -D\,N_i\frac{dT}{dx}\Big|_i^j = \mathrm{D}\frac{dT}{dx} \quad \text{(a)}$$
$$q_j^e = -\int_i^j D\frac{d}{dx}\left(N_j\frac{dT}{dx}\right)dx = -D\,N_j\frac{dT}{dx}\Big|_i^j = -\mathrm{D}\frac{dT}{dx} \quad \text{(b)} \tag{18}$$

These terms (Eq. 18) are the inter element contribution and vanish since the derivative terms vanished between the neighboring elements. Thus, the finite element system equation formed without these inter elements except when the flux (derivative) boundary condition is specified. When the flux boundary specified, they used to apply the flux condition at the boundaries. Furthermore, they are used to compute the flux term once the system equation is solved.

We need the first derivative of the finite element equation (Eq. 7) and the interpolation functions (Eq. 6) in order to evaluate the second integrals on the right in Eq. 17. The first derivatives of the element equation using element length (l) is

$$\frac{dT^e}{dx} = \frac{1}{l}\left(-T_i + T_j\right) \tag{19}$$

Furthermore, the derivatives of the weighting functions are

$$\frac{dN_i}{dx} = \frac{-1}{l} \quad \text{(a)}$$
$$\frac{dN_j}{dx} = \frac{1}{l} \quad \text{(b)} \tag{20}$$

Thus, the residual from the second terms (denote by k_i^e and k_j^e) in Eq. 17 become

$$k_i^e = D\int_i^j\left(\frac{dN_i}{dx}\frac{dT}{dx}\right)dx = D\int_i^j\left(\frac{-1}{l}\frac{\left(T_j - T_i\right)}{l}\right)dx = \frac{D}{l}\left(T_i - T_j\right) \quad \text{(a)}$$
$$k_j^e = D\int_i^j\left(\frac{dN_j}{dx}\frac{dT}{dx}\right)dx = D\int_i^j\left(\frac{1}{l}\frac{\left(T_j - T_i\right)}{l}\right)dx = \frac{D}{l}\left(T_j - T_i\right) \quad \text{(b)} \tag{21}$$

The last integrals in Eq. 17 are constant and evaluated using the linear weighting functions. Their contributions to the element residual are

$$f_i^e = Q\int_i^j N_i dx = \frac{Q}{l}\left(x_j x - \frac{x^2}{2}\right)\Big|_i^j = \frac{Ql}{2} \quad \text{(a)}$$
$$f_j^e = Q\int_i^j N_j dx = \frac{Q}{l}\left(\frac{x^2}{2} - x_i x\right)\Big|_i^j = \frac{Ql}{2} \quad \text{(b)} \tag{22}$$

Substituting Eqs. 18, 21, and 22 in Eq. 17 yields

$$R_i^e = D\left.\frac{dT}{dx}\right|_i + \frac{D}{l}(T_i - T_j) - \frac{Ql}{2} \quad \text{(a)}$$
$$R_j^e = -D\left.\frac{dT}{dx}\right|_j + \frac{D}{l}(-T_i + T_j) - \frac{Ql}{2} \quad \text{(b)} \tag{23}$$

For simplicity, we will introduce the matrix notation and rewrite the above terms of element contribution to residual equation in matrix form as

$$\begin{Bmatrix} R_i^e \\ R_j^e \end{Bmatrix} = \begin{Bmatrix} D\left.\frac{dT}{dx}\right|_i \\ -D\left.\frac{dT}{dx}\right|_j \end{Bmatrix} + \frac{D}{l}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} T_i \\ T_j \end{Bmatrix} - \frac{Ql}{2}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} \tag{24}$$

The matrix representation becomes very important particularly when illustrating more than one component is contributing to the element residual. Furthermore, it also comes helpful in two- and three-three dimensional spaces.
Thus, the residuals of element can represent as

$$\{R^e\} = \{q^e\} + [K^e]\{T^e\} - \{f^e\} \tag{25}$$

Eqs. 24 and 25 are representing the contribution of the element to the final system equation. The contribution are from the first right terms in Eq. 24 (denoted by $\{q\}$ at Eq.25) that are the inter element contribution and vanishes between the neighboring elements in the final system equation. The second terms, the element contribution to the final system equation, is referred as the stiffness matrix and is denoted by$[K]$. It can be easily determined from the interpolation function for each element as illustrated above and included in the final system equation. The final terms are referred as force vector and denoted by $\{f\}$. The final system equation is built by assembling the element matrices step by step or systematically. Thus the system equation becomes

$$\{R\} = \{q\} + [K]\{T\} - \{f\} = 0 \tag{26}$$

The final system equation is formed by assembling each element's contribution and adding the contribution of each element's based on the nodal points. When the system residual becomes zero, the approximate solution can be used to estimate the system. The number of elements used to define the final system equation has significant effect on the element residual. Thus, increasing the number of elements decreases the element residual and improves the approximate (FEM) solution.

2.3.1 Application of finite element for one-dimensional elliptic equation

To illustrate the application of FEM in one-dimensional elliptic equation, we will consider the temperature distribution of an insulated rod length $L = 1$ and thermal diffusivity $D = 10$. A constant heat is also being generated at the rate of $Q = 10$. The boundary conditions are specified as one end where $x = 0, T = 5$ and the opposite end where $x = 1, q = -10$. To illustrate the FEM solution this system, we uses four elements, the nodes of the elements are numbered from 1 to 5 and the element length is assumed to be uniform. Thus, the elliptic equation (Eq. 13) becomes

$$10\frac{d^2T}{dx^2} + 10 = 0 \tag{27}$$

The residual of element (Eq. 25) becomes

$$\begin{Bmatrix} R_i^e \\ R_j^e \end{Bmatrix} = \begin{Bmatrix} q_i \\ q_j \end{Bmatrix} + \frac{10}{0.25}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} T_i \\ T_j \end{Bmatrix} - \frac{10(0.25)}{2}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} \tag{28}$$

The prescribe boundary condition at the end where the temperature is fixed $T_1 = 5$ and at the opposite end where the flux boundary applied $q_5 = 10$. The assembled final system equation for four elements becomes

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{Bmatrix} T_2 \\ T_3 \\ T_4 \\ T_5 \end{Bmatrix} = \begin{Bmatrix} 5.03125 \\ 0.0625 \\ 0.0625 \\ -0.1875 \end{Bmatrix} \tag{29}$$

The finite element solution for the temperature profile produces the values $T_2 = 4.97, T_3 = 4.91, T_4 = 4.78,$ and $T_5 = 4.60$. Furthermore; using the same principle shown above in detail a computer program is developed and the computer solution for 10 and 20 elements is shown in Fig. 6. To show that the number of elements has effect in quality of the FEM solution.

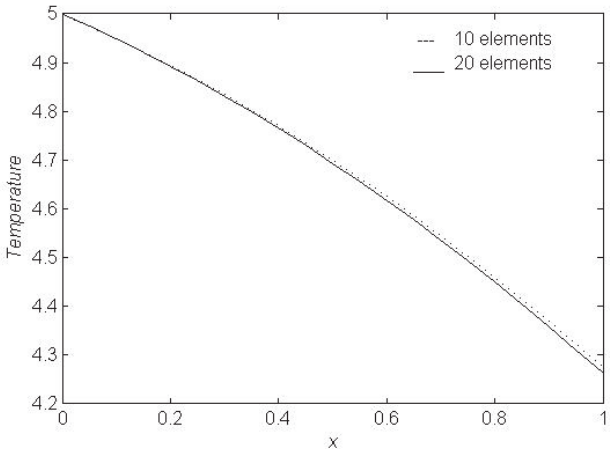


Fig. 6. Finite element approximation for steady state temperature profile for insulated rod

2.3.2 Two dimensional elliptic equation in finite element method

A two-dimensional elliptic equation is used to describe the steady state heat conduction with heat generation similar to the previous section but in a two-dimensional space. A two-dimensional steady state flow of heat in isometric material is expressed by an elliptic equation as

$$D\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + Q = 0 \tag{30}$$

It represents a bounded area. The solution uniqueness is dependent on the boundary condition. Like one-dimensional cases, the boundary condition can be specified as either the functional value or flux. However, in two-dimensional cases, the boundary values are specified at the edges while the region is an area.

In order to illustrate the finite solution of an elliptic equation, we will consider the temperature distribution in two-dimensional spaces that satisfies Eq. 30. The finite element solution satisfies the weighting integral function in two-dimensional space. For simplicity, we will use a linear rectangular element discussed in Sec. 2.2 to evaluate the integral for each elements and determine the elements contribution to the final system equation. Parallel to the one-dimensional finite element method, two-dimensional equations can be modeled by indentifying the implication of increasing dimensionality at the element integral. The interpolation functions for a linear rectangular element with four nodes are defined in (Eq. 12). For simplicity, we use a linear rectangular element with four nodes and also we use a matrix notation ($[N]^T$) to represent all nodal points of the elements instead of writing each node point contribution. Thus, the residual integral for a two-dimensional elliptic equation (Eq. 30) becomes

$$\{R_i^e\} = -\int [N]^T \left(D\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + Q \right) dA \tag{31}$$

The major difference from the one-dimensional case is that the residual integral is area integral and the boundary is line integral. Reducing the degree of continuity for the second derivative term by differentiation product rule (Eq. 15) further simplifies the element residual integral as

$$\{R_i^e\} = -D\left(\int \frac{\partial}{\partial x}\left([N]^T \frac{\partial T}{\partial x}\right) - \frac{\partial [N]^T}{\partial x}\frac{\partial T}{\partial x} + \frac{\partial}{\partial y}\left([N]^T \frac{\partial T}{\partial y}\right) - \frac{\partial [N]^T}{\partial y}\frac{\partial T}{\partial y}\right) dA - \int [N]^T Q \, dA \tag{32}$$

Substituting the element equation (quadratic linear element) $T^e = [N]\{T\}$ and rearranging the terms

$$\begin{aligned} \{R_i^e\} = -D \int \left(\frac{\partial}{\partial x}\left([N]^T \frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left([N]^T \frac{\partial T}{\partial y}\right)\right) dA \\ + D\left(\int \frac{\partial [N]^T}{\partial x}\frac{\partial N}{\partial x}\{T^e\} + \frac{\partial [N]^T}{\partial y}\frac{\partial N}{\partial y}\{T^e\}\right) dA - \int [N]^T \{Q^e\} \, dA \end{aligned} \tag{33}$$

When the derivative boundary condition is applied, the first two terms are reduced to surface integral by using Green's theorem ($\int \frac{\partial}{\partial x}\left([N]^T \frac{\partial T}{\partial x}\right) dA = \int \left([N]^T \frac{\partial T}{\partial x}\right) d\Gamma \cos\theta$). These two terms on the right can be replaced by an integral around the boundary using the outward normal. Thus,

$$-D \int \frac{\partial}{\partial x}\left([N]^T \frac{\partial T}{\partial x}\right) dA + \frac{\partial}{\partial y}\left([N]^T \frac{\partial T}{\partial y}\right) dA = -D \int [N]^T \left(\frac{\partial T}{\partial x} cos\theta + \frac{\partial T}{\partial y} sin\theta\right) d\Gamma \tag{34}$$

The integral around the boundary of the element is done in a counterclockwise direction. For the rectangular element we considered here, it is the sum of four integrals. It includes the side where the boundary condition is specified and the inter-element side. The inter-element integral vanishes due to the element continuity requirements. However, when the flux boundaries are specified, the surface integrals need to be evaluated where applicable. The general derivative boundary condition can be given as a function of the surface temperature, constant, or zero as

$$-D\frac{\partial T}{\partial n} = C_1 T + C_2 \tag{35}$$

where $\partial T/\partial n$ is the normal gradient at the surface. When the boundary condition is insulated, $\partial T/\partial n = 0$, thus $C_1 = C_2 = 0$. When the derivative is the function of the surface temperature and constant, the boundary surface integral can be evaluated along the specified surface. Therefore, introducing a relationship given by the element equation $T^e = [N]\{T\}$ where $[N]$ represent the rectangular element interpolation functions (Eq. 12) and Eq. 35 is introduced in Eq. 34 gives

$$\{q_{bc}^e\} = \int C_1([N]^T[N])\{T^e\}d\Gamma + \int [N]^T C_2\, d\Gamma \tag{36}$$

Using Eq. 12, linear quadratic element, the above integral can be evaluated. The first integral has following terms

$$\left[K_q\right] = C_1 \int \begin{bmatrix} N_i^2 & N_iN_j & N_iN_k & N_iN_l \\ N_iN_j & {N_j}^2 & N_jN_k & N_jN_l \\ N_iN_k & N_jN_k & {N_k}^2 & N_kN_l \\ N_iN_l & N_jN_l & N_kN_l & {N_l}^2 \end{bmatrix} d\Gamma \tag{37}$$

and evaluated for arbitrary side l_{ij} where N_i and N_j are the only contributing functions gives

$$[K_{bc}^e] = \frac{C_1 l_{ij}}{6}\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{38}$$

The second term in Eq. 36 for arbitrary side l_{ij} becomes

$$\{f_{bc}^e\} = C_2 \int \begin{Bmatrix} N_i \\ N_j \\ N_k \\ N_l \end{Bmatrix} dx = \frac{C_2 l_{ij}}{2}\begin{Bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{Bmatrix} \tag{39}$$

Furthermore, the middle terms integral in Eq. 33 can be evaluated using the first derivatives of rectangular shape function Eq. 12. as

$$\frac{\partial[N]^T}{\partial x}\frac{\partial N}{\partial x} = \begin{bmatrix} \frac{\partial N_i}{\partial x}^2 & \frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x} & \frac{\partial N_i}{\partial x}\frac{\partial N_k}{\partial x} & \frac{\partial N_i}{\partial x}\frac{\partial N_l}{\partial x} \\ \frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x} & \frac{\partial N_j}{\partial x}^2 & \frac{\partial N_j}{\partial x}\frac{\partial N_k}{\partial x} & \frac{\partial N_j}{\partial x}\frac{\partial N_l}{\partial x} \\ \frac{\partial N_i}{\partial x}\frac{\partial N_k}{\partial x} & \frac{\partial N_j}{\partial x}\frac{\partial N_k}{\partial x} & \frac{\partial N_k}{\partial x}^2 & \frac{\partial N_k}{\partial x}\frac{\partial N_l}{\partial x} \\ \frac{\partial N_i}{\partial x}\frac{\partial N_l}{\partial x} & \frac{\partial N_j}{\partial x}\frac{\partial N_l}{\partial x} & \frac{\partial N_k}{\partial x}\frac{\partial N_l}{\partial x} & \frac{\partial N_l}{\partial x}^2 \end{bmatrix} \quad \text{(a)} \tag{40}$$

$$\frac{\partial[N]^T}{\partial y}\frac{\partial N}{\partial y} = \begin{bmatrix} \frac{\partial N_i}{\partial y}^2 & \frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y} & \frac{\partial N_i}{\partial y}\frac{\partial N_k}{\partial y} & \frac{\partial N_i}{\partial y}\frac{\partial N_l}{\partial y} \\ \frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y} & \frac{\partial N_j}{\partial y}^2 & \frac{\partial N_j}{\partial y}\frac{\partial N_k}{\partial y} & \frac{\partial N_j}{\partial y}\frac{\partial N_l}{\partial y} \\ \frac{\partial N_i}{\partial y}\frac{\partial N_k}{\partial y} & \frac{\partial N_j}{\partial y}\frac{\partial N_k}{\partial y} & \frac{\partial N_k}{\partial y}^2 & \frac{\partial N_k}{\partial y}\frac{\partial N_l}{\partial y} \\ \frac{\partial N_i}{\partial y}\frac{\partial N_l}{\partial y} & \frac{\partial N_j}{\partial y}\frac{\partial N_l}{\partial y} & \frac{\partial N_k}{\partial y}\frac{\partial N_l}{\partial y} & \frac{\partial N_l}{\partial y}^2 \end{bmatrix} \quad \text{(b)} \tag{40}$$

Using Fig. 3 and Eq. 12, the integral of the terms above yields

$$[K^e] = \frac{DY}{6X}\begin{bmatrix} 2 & -2 & -1 & 1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{bmatrix} + \frac{DX}{6Y}\begin{bmatrix} 2 & 1 & -1 & -2 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix}\begin{Bmatrix} T_i \\ T_j \\ T_k \\ T_l \end{Bmatrix} \tag{41}$$

Using the rectangular interpolation functions, the last integral is evaluated and gives the residual as

$$\{f^e\} = \int [N]^T Q\, dA = Q\int \begin{Bmatrix} N_i \\ N_j \\ N_k \\ N_l \end{Bmatrix} dA = \frac{QXY}{4}\begin{Bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{Bmatrix} \tag{42}$$

Combining Eqs. 38, 39, 41, and 42 give all of the components contributing to the element residual integral (Eq. 33) in matrix form as

$$\{R^e\} = [K^e] + [K^e_{bc}]\{T^e\} + \{f^e_{bc}\} - \{f^e\} \tag{43}$$

2.3.3 Application of finite element for two-dimensional elliptic equation

To illustrate a two-dimensional elliptic equation, we will consider the temperature distribution of a two-dimensional rectangular region (Fig. 7) with a thermal diffusivity $D = 10$. A constant heat is being generated at the rate of $Q = 10$. Using four elements in each direction, the boundary conditions are specified where $y = 0,\ T = 5$ and at the opposite end where $y = 1,\ q = 3T - 6$ while the other regions are kept insulated. Assuming the material is isotropic and the elements are square. Thus, the elliptic two-dimensional equation (Eq. 30) becomes

$$10\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right)+10=0 \tag{44}$$

For simplicity, we use 4 elements to describe a unit square region. The numbering and the boundary conditions are show in Fig. 7. For illustrative purposes, we select element 3 and show all contribution for the system residual matrix mainly the flux boundary that is applied (l_{8-7}) top end. The element contribution becomes

$$\{R^3\}=\frac{10}{6}\begin{bmatrix}4&-1&-2&-1\\-1&4&-1&-2\\-2&-1&4&-1\\-1&-2&-1&4\end{bmatrix}+\frac{(3)(0.5)}{6}\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&2&1\\0&0&1&2\end{bmatrix}\begin{Bmatrix}T_4\\T_5\\T_8\\T_7\end{Bmatrix}-\frac{6(0.5)}{2}\begin{Bmatrix}0\\0\\1\\1\end{Bmatrix}-\frac{10(0.5)^2}{4}\begin{Bmatrix}1\\1\\1\\1\end{Bmatrix} \tag{45}$$

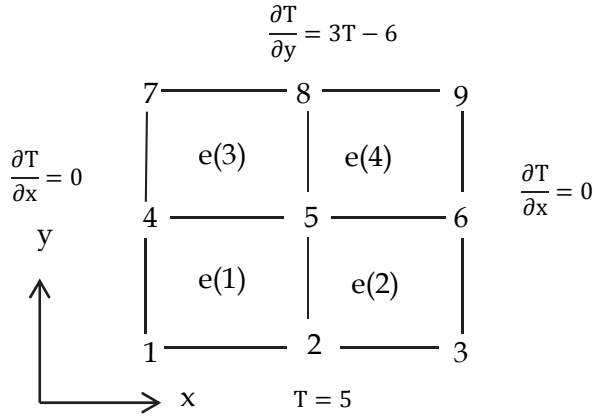


Fig. 7. Two-dimensional region divided into four square elements with boundary conditions

When combines and applied the specified boundary condition, the final system equation becomes 6 by 6 matrix as

$$\begin{bmatrix}8&-2&0&-1&-2&0\\-2&16&-2&-2&-2&-2\\0&-2&8&0&-2&-1\\1&-2&0&4.3&-0.85&0\\-2&-2&-2&-0.85&8.6&-0.85\\0&-2&-1&0&-0.85&4.3\end{bmatrix}\begin{Bmatrix}T_4\\T_5\\T_6\\T_7\\T_8\\T_9\end{Bmatrix}=\begin{Bmatrix}15.75\\31.50\\15.75\\1.275\\2.550\\1.275\end{Bmatrix} \tag{46}$$

As expected, the temperature profile is decreasing and symmetric as $T_4 = 4.97, T_5 = 4.97, T_6 = 4.97, T_7 = 4.69, T_8 = 4.69,$ and $T_9 = 4.69$.

2.4 Parabolic equation in finite element method

The major characteristics of parabolic equations are that they require boundary and initial conditions (Awrejcewicz & Krysko, 2010). The general procedure for solving parabolic equations in finite element is by evaluating the residual integral with respect to space coordinates for fixed time. Using the initial value for the new value prediction, the time history is generated. In order to illustrate the fundamental procedure in solving a parabolic equation in the finite element method, we start by discussing a one-dimensional parabolic equation followed by two-dimensional equation. The one-dimensional scheme can be modified to include a two-dimensional equation with simple two-dimensional elements substitution.

2.4.1 One-dimensional parabolic equation

The cooling and heat process of material is considered parabolic in nature. The temperature change is expressed in terms of the rate of change in time and space. The heating and/or the cooling process of an insulated bar that is subjected to the different temperature can be considered a one-dimensional parabolic equation. In order to find the temperature in time, we need to solve the governing parabolic equation

$$D\frac{\partial^2 T}{\partial x^2} + Q - \lambda\frac{\partial T}{\partial t} = 0 \tag{47}$$

where λ is a rate constant. The finite element equation that gives the element contribution to the system residual is

$$\{R^e\} = -\int [N]^T \left(D\frac{\partial^2 T}{\partial x^2} + Q - \lambda\frac{\partial T}{\partial t}\right) dx \tag{48}$$

The first integral from the above equation is similar to Eq. 14 that yields the element contribution toward the residual integral as Eq. 25. What remain is solving the time-dependent integral, we use the average value assumption that the time derivatives ($\partial T/\partial t = \dot{T}$) varies linearly between the time interval. Using the shape function relationship that

$$\{\dot{T}^e\} = [N]\{\dot{T}^e\} \tag{49}$$

Then, the second term residual integral becomes

$$\{R_c^e\} = \lambda \int [N]^T[N]\{\dot{T}^e\}\, dx \tag{50}$$

The integral above is defined as capacitance matrix ($[C^e]$) and can be evaluated using the linear element interpolation function for one-dimensional element (Eq. 6). The integral result for the linear element is

$$\{R_c^e\} = \frac{\lambda l}{6}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{Bmatrix}\dot{T}_i\\ \dot{T}_j\end{Bmatrix} \tag{51}$$

The element contribution for final system equation becomes

$$\{R^e\} = \{q^e\} + [K^e]\{T^e\} - \{f^e\} - [C^e]\{\dot{T}^e\} \tag{52}$$

When the element equation is assembled, the sums of the residual vanish. As a result, the final system equation becomes

$$-[C]\{\dot{T}\} + [K]\{T\} + \{q\} - \{f\} = 0 \tag{53}$$

A time-dependent finite element equation requires solving the equation with time first. In order to approximate the time-dependent equation, the mean value-based equation is used. The mean (Sahoo & Riedel, 1998, Segerlind, 1984) rule is based on the hypothesis that the change in function (df/dt) at a location between two points (a, b) is proportional to the average change between two values of the function (f).

$$\frac{df}{dt} = \frac{f(b) - f(a)}{\Delta t} \tag{54}$$

The value at an arbitrary point c that is between a and b can be approximated as

$$f(c) = f(a) + (c - a)\frac{f(b) - f(a)}{\Delta t} \tag{55}$$

Let $(c - a)/\Delta t$ replaced by α

$$f = (1 - \alpha)f(a) + \alpha f(b) \tag{56}$$

Parallel to the mean value approximation, the time-dependent finite element solution (Eq. 53) can be approximated by introducing the vector containing the nodal values

$$[C]\{\dot{T}\} = \frac{[K]\{T\} + \{f\}_b - [K]\{T\} + \{f\}_a}{\Delta t} \tag{57}$$

The functional value between

$$([C] + \alpha\Delta t[K])\{T\}_b = ([C] - (1 - \alpha)\Delta t[K])\{T\}_a + \Delta t((1 - \alpha)\{f\}_a + \alpha\,\{f\}_b) \tag{58}$$

Thus, the nodal value can be predicted based on the known initial value and the time scale. When $\alpha = 1/2$, it is called the center difference method and the time-dependent finite element equation becomes

$$\left([C] + \frac{\Delta t}{2}[K]\right)\{T\}_b = \left([C] - \frac{\Delta t}{2}[K]\right)\{T\}_a + \frac{\Delta t}{2}\{f\}_a + \frac{\Delta t}{2}\,\{f\}_b \tag{59}$$

The above system equation has an equal number of unknown value and equation and can be solved by linear solvers.

2.4.2 Application of FEM in one-dimensional parabolic equation

To illustrate the application of the FEM in solving a one-dimensional parabolic equation, we will consider a finite element solution of an insulated shaft that is initially at known temperature (1) and places in the environment where the ends are subjected to 0 temperatures. The material diffusivity is $D = 10$ and heat capacity $\lambda = 1$. It is assumed to be one-dimensional since the lateral temperature change is insignificant to compare with the horizontal (x) direction. The length of a bar is 1 unit and for simplicity, we use four uniform elements (0.25) be used show the temperature distribution with time. Once the boundaries conditions are applied, the stiffness and capacitance matrix become

$$[K] = \frac{10}{0.25}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \quad \text{(a)}$$

$$[C] = \frac{0.25}{6}\begin{bmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} \quad \text{(b)} \qquad (60)$$

$$[f]^T = [0\ 0\ 0]$$

Unless the material property and time step change with time, the coefficient matrix is only evaluated once. Using the center difference method, the system matrix becomes similar to Eq. 59. Thus, using the previous temperature values to estimate the new value recursively, the time cooling process may be predicted by the finite element method. Using the time step of 0.001 s, the temperature profile of $T_2 = 0.30, T_3 = 0.43, T_4 = 0.30$ estimated after 0.01 s. In addition, a computer program is written by extending the above principle for 20 elements. The cooling process is solved using 0.001 s time step. The temperature profile with several times is shown in Fig. 8. As expected the rate cooling process with time is predicted using the finite element method and the solution also improves with elements number increases.

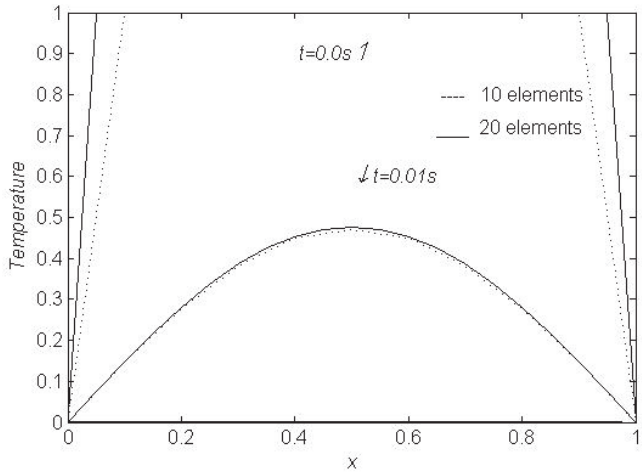


Fig. 8. The rate of cooling predicted with 10 and 20 linear elements using the finite element method.

2.4.3 Two-dimensional parabolic equation

A two-dimensional parabolic equation is represented by

$$D\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + Q - \lambda\frac{\partial T}{\partial t} \qquad (61)$$

The element contribution to the residual is

$$\{R^e\} = -\int [N]^T \left(D\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + Q - \lambda\frac{\partial T}{\partial t}\right) dA \qquad (62)$$

Parallel to the one-dimensional parabolic equation, the two-dimensional parabolic equation solution can be simply introduced by replacing the one-dimensional element integral for stiffness and capacitance matrix by the two-dimensional. In section 2.3.2, we showed that the first integral term using the linear rectangular element (Eq. 12 and the values in Fig 3.) yields the stiffness matrix and the derivative boundary conditions contribution ($[K_{bc}^e] + [K^e]$). Parallel to the one-dimensional parabolic case (Sec. 2.4.1), the time-dependent integral can be evaluated using the rectangular element (Eq. 12). This new capacitance matrix for two-dimensional element becomes

$$[C^e] = \int [N]^T \lambda [N] dA = \lambda \frac{XY}{36} \begin{bmatrix} 4 & 2 & 1 & 2 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 4 \end{bmatrix} \tag{63}$$

Thus, the two-dimensional parabolic equation is similar to Eq. 53, but the vector and the matrix are going to be larger since the four nodal values are involved per element. The vector element is 1 by 4, while the matrix is 4 by 4 except for the boundary vectors.

2.4.4 Application of FEM in two-dimensional parabolic equation

To illustrate a two-dimensional parabolic equation application, we will consider the temperature history of the two-dimensional rectangular region shown in Fig. 9. We selected this problem for simplicity and illustrative purposes. Thermal diffusivity $D = 10$ and constant heat is being generated at the rate of $Q = 10$. The boundary conditions are specified as one end where $y = 0$, $T = 5$ and $y = 1.5$, $T = 1$ while the other regions are kept insulated. Initially, the surface temperature is kept at 5 degree before it is introduced into the environment. The objective of this to show that how FEM is applied to solve this parabolic equation. The region is discretized using six square elements size of 0.5 units (2 in horizontal and 3 in vertical direction).

The terms for two-dimensional, stiffness (Eq. 41), capacitance (Eq. 63), and the applied heat (Eq. 42) and boundary conditions (Eqs. 38 and 39) become

$$[K] = \frac{10}{6} \begin{bmatrix} 8 & -2 & 0 & -1 & -2 & 0 \\ -2 & 16 & -2 & -2 & -2 & -2 \\ 0 & -2 & 8 & 0 & -2 & -1 \\ -1 & -2 & 0 & 8 & -2 & 0 \\ -2 & -2 & -2 & -2 & 16 & -2 \\ 0 & -2 & -1 & 0 & -2 & 8 \end{bmatrix} \quad (a) \tag{64}$$

$$[C] = \frac{0.25}{36} \begin{bmatrix} 8 & 4 & 0 & 2 & 1 & 0 \\ 4 & 16 & 4 & 1 & 4 & 1 \\ 0 & 4 & 8 & 0 & 1 & 2 \\ 2 & 1 & 0 & 8 & 4 & 0 \\ 1 & 4 & 1 & 4 & 16 & 4 \\ 0 & 1 & 2 & 0 & 4 & 8 \end{bmatrix} \quad (b) \tag{64}$$

$$f = [-15.75 - 31.75 - 15.75 - 3.75 - 7.25 - 3.75]^T \quad (c)$$

Using the initial condition

$$\{T_0\} = \{5\} \tag{65}$$

with center difference (Eq. 59) and the 0.001s time step, the temperature distribution of ($T_4 = 3.92, T_5 = 3.91, T_6 = 3.92, T_7 = 2.58, T_8 = 2.57,$ and $T_9 = 2.58$) is predicted after 0.1 seconds.

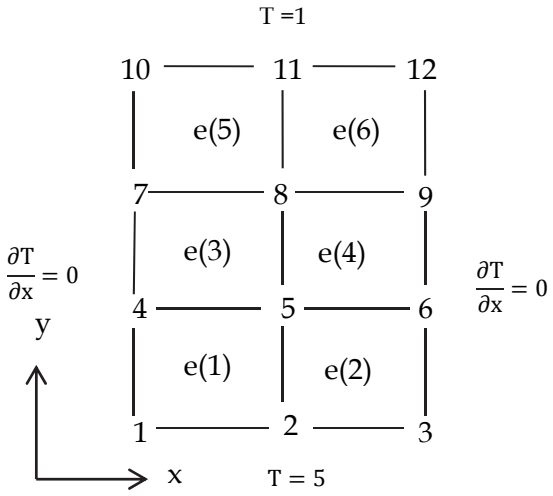


Fig. 9. Two dimensional square region using elements.

3. Finite difference method

The finite difference method is a direct interpretation of the differential equation into a discrete domain so that it can be solved using a numerical method. It is a direct representation of the governing equation $(\delta f/\delta x = (f_{i+1} - f_i)/(x_{i+1} - x_i)$. Using the discontinuous but connected regions, the governing equation is defined within the interval. In addition to direct interpretation, the deferential equation, the basic finite difference form, also can be derived from the Taylor-series expansion. Next, we will discuss the definition of one- and two-dimensional finite difference equations.

3.1 One-dimensional finite difference formulation

In order to define a finite difference representation in a one-dimensional space, we define a line space along the x-axis. The Taylor-series expansion for function f_{i+1} about point (i) is,

$$f_{i+1} = f_i + f'(x_{i+1} - x_i) + \frac{1}{2}f''(x_{i+1} - x_i)^2 + \frac{1}{6}f'''(x_{i+1} - x_i)^3 + \cdots \quad (66)$$

Let $(x_{i+1} - x_i) = \Delta x$

$$f_{i+1} = f_i + f'\Delta x + \frac{1}{2}f''\Delta x^2 + \frac{1}{6}f'''\Delta x^3 + \cdots \quad (67)$$

By rearranging

$$f' = \frac{(f_{i+1} - f_i)}{\Delta x} - \frac{1}{2}f''\Delta x - \frac{1}{6}f'''\Delta x^2 + \cdots \quad (68)$$

The above expression may be referred to as a forward difference. Furthermore, a similar expression may also be obtained by the backward difference

$$f' = \frac{(f_i - f_{i-1})}{\Delta x} \tag{69}$$

And center difference

$$f' = \frac{(f_{i+1} - f_{i-1})}{2\Delta x} \tag{70}$$

The higher order derivative of finite difference also may be derived from the Taylor-series expansion as

$$f_{i+1} = f_i + f'\Delta x + \frac{1}{2}f''\Delta x^2 + \frac{1}{6}f'''\Delta x^3 + \cdots \tag{71}$$

Using recursively and eliminate the first derivative term with previous approximation

$$\frac{1}{2}f''\Delta x^2 = f_{i+1} - f_i - \frac{(f_{i+1} - f_{i-1})}{2} - \frac{1}{6}f'''\Delta x^3 + \cdots \tag{72}$$

The final second-order derivative can be expressed as

$$f'' = \frac{1}{\Delta x^2}(f_{i+1} - 2f_i + f_{i-1}) - o(\Delta x) + \cdots \tag{73}$$

Similarly, the third-order derivative may also be defined as

$$f''' = \frac{1}{\Delta x^3}(-f_i + 3f_{i+1} - 3f_{i+2} + f_{i+3}) - o(\Delta x) \tag{74}$$

It also important to recognize that by using more points to form a discrete derivative, the error term may be minimized. For example, using five points instead of four in the equation above, the third-order derivate may be expressed as

$$f''' = \frac{1}{2\Delta x^3}(-f_{i-2} + 2f_{i-1} - 2f_{i+1} + f_{i+2}) - o(\Delta x^2) \tag{75}$$

So that, the error term is no longer a linear function but quadratic.

3.2 Two-dimensional finite difference expression

Parallel to one-dimensional forward, backward, and center difference expressions, the finite difference representation for a two-dimensional expression also can be defined. Similar to partial derivative, first, take the derivative two-dimensional space with one of the variables followed by the other as required. The first order partial derivatives of a function $f = f(x, y)$ in two- dimensional space (x, y) are expressed as

$$\begin{aligned} f_x' &= \frac{1}{2\Delta x}\left(f_{i+1,j} - f_{i-1,j}\right) \qquad (a) \\ f_y' &= \frac{1}{2\Delta y}\left(f_{i,j+1} - f_{i,j-1}\right) \qquad (b) \end{aligned} \tag{76}$$

Similarly, the second-order derivatives in x and y direction are

$$f_x'' = \frac{1}{\Delta x^2}\,(f_{i+1,j} - 2f_{i,j} + f_{i-1,j}) \qquad (a)$$
$$f_y'' = \frac{1}{\Delta y^2}\,(f_{i,j+1} - 2f_{i,j} + f_{i,j-1}) \qquad (b) \tag{77}$$

Analogous to the partial derivate in the x and y direction, the function derivative with x and y, or vice versa, can be obtained by taking the first partial in one of the directions followed by the other. The final expression of the partial derivative is the same whether the x or y direction is used first or second.

$$f_{yx}'' = \frac{1}{4\Delta y \Delta x}\,(f_{i+1,j+1} - f_{i-1,j+1}) - (f_{i+1,j-1} - f_{i-1,j-1}) \tag{78}$$

3.3 Finite difference approximation of elliptic equation

Parallel to the previous sections, we start by considering the one-dimensional elliptic equation and how the FDM equation is formed. The major difference between the finite difference and the finite element method is that the finite difference method is based on the functional value at the nodal points, while the finite element is based on using the weighting function of the element to estimate the nodal values. We start by replacing the elliptic equation (Eq. 13) with the finite difference equation as

$$\frac{D}{\Delta x^2}(T_{i+1} - 2T_i + T_{i-1}) + Q = 0 \tag{79}$$

For simplicity, $\Delta x = h$ and be uniform, and $i = 1, 2 \ldots N$, thus

$$\frac{D}{h^2}(T_{i-1} - 2T_i + T_{i+1}) + Q = 0 \tag{80}$$

The solution of the elliptic equation is required the boundary condition. Thus, the boundary condition must be specified at, $i = 1$ and $i = N$. The number of equations and the unknown depends on the boundary condition whether or not the particular value or the derivative values are specified. When the derivative boundary condition is specified, the general flux equation (Eq. 35) with all discrete finite derivative methods (Sec. 3.1) can be used to replace the derivative boundary condition. The flux at the boundary can be estimated using the forward, backward, and center difference. Forward or backward difference can be used to define the flux using the boundary point and an ideal point next to it ($dT/dn = (T_{b+1} - T_b)/h$. Moreover, for a more accurate estimate, the center difference may be used ($dT/dn = (T_{b-1} - T_{b+1})/2h$). Both methods require the introduction of a new ideal point outside the region. The ideal point is eventually eliminated by combining the equations that include the specified boundary condition and the boundary node equation.

3.3.1 Application of FDM in one-dimensional elliptic equation

To illustrate a one-dimensional elliptic equation, we will consider the temperature distribution of a one-dimensional rod that is discussed in Sec. 2.3.1. Thus, when the material properties with all the assumption applied to the finite difference elliptic equation (Eq. 72) becomes

$$\frac{10}{0.25^2}(T_{i-1} - 2T_i + T_{i+1}) + 10 = 0 \tag{81}$$

When the prescribed boundary condition at the end where the temperature is fixed applied, T_1 becomes 5. Using the forward difference, the prescribed flux boundary at the opposite end gives the equation

$$\frac{\partial T}{\partial n} = D\frac{T_6 - T_5}{h} = -10 \tag{82}$$

where $T_6 = T_5 - 10h/D$. T_6 can be illuminated between the boundary condition and the last equation. Using all the element equation, the finite difference expression of the system becomes

$$\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{Bmatrix} T_2 \\ T_3 \\ T_4 \\ T_5 \end{Bmatrix} = \begin{Bmatrix} -5.0625 \\ -0.0625 \\ -0.0625 \\ 0.1875 \end{Bmatrix} \tag{83}$$

The finite difference solution for the temperature profile produces the values $T_2 = 5.00$, $T_3 = 4.94$, $T_4 = 4.81$, and $T_5 = 4.62$. Indeed, the matrixes in Eq. 29 and Eq. 83 have some striking similarity considering they are formed from two different methods. As expected, the profiles in both FE and FD methods are the same. It is not also expected that the nodal values have differences. Furthermore, the computer solution for 10 and 20 elements is shown in Fig. 10. The profile indicates that with more nodal values, a better result can be estimated.

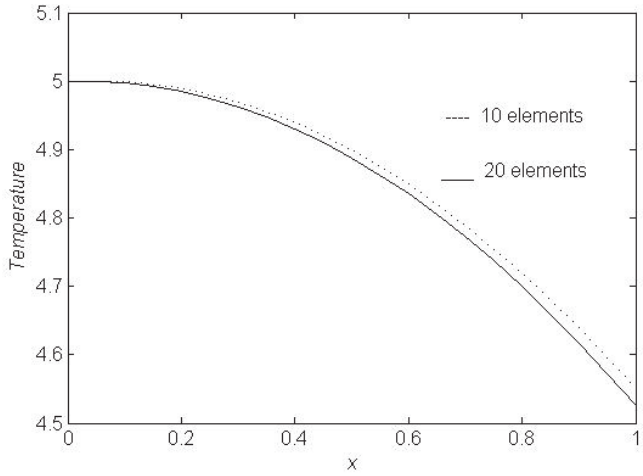


Fig. 10. Finite difference approximated temperature profile for one-dimensional elliptic equation rod.

3.3.2 Tow-dimensional elliptic finite difference solution

As previously discussed, elliptic equations are generally associated with steady state problems. The finite difference representation of two-dimensional elliptic equation (Eq. 30) for steady state temperature distribution is

$$\frac{D}{\Delta x^2}\left(T_{i+1,j} - 2T_{i,j} + T_{i-1,j}\right) + \frac{D}{\Delta y^2}\left(T_{i,j+1} - 2T_{i,j} + T_{i,j-1}\right) + Q = 0 \tag{84}$$

For simplicity, we let $\Delta x = \Delta y = h$, and assuming the diffusion is isotropic, we get

$$\frac{D}{h^2}\left(T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j}\right) + Q = 0 \tag{85}$$

As discussed earlier, the boundary condition may be a functional value and/or the flux at the edge, and the uniqueness of the solution depends on the boundary condition. There are vast methods developed over the years for direct and iterative solution methods in the form of explicit and implicit forms. As the name implies, the direct method involves a fixed number of operations to find a solution. In contrast, the iterative method starts with an approximation that successively improves (Dahlquist and Bjorck, 1974). Eq. 85 may be arranged in terms i or j so that the known value be solved using all the other known values. Furthermore by recognizing the relationship pattern among the neighboring points, the system equation can be generated as

$$T_{i+1,j} = \left(4T_{i,j} - T_{i-1,j} - T_{i,j+1} - T_{i,j-1}\right) - \frac{h^2}{D}Q \tag{86}$$

To illustrate the FD solution for two-dimensional elliptic equation, we will consider the problem discoursed before in Sec. 2.3.3 and Fig. 7. Thus, the FD system equation becomes

$$\begin{bmatrix} -4 & 2 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 2 & -4 & 0 & 0 & 1 \\ 1 & 0 & 0 & -3.15 & 2 & 0 \\ 0 & 1 & 0 & 1 & -3.15 & 1 \\ 0 & 0 & 1 & 0 & 2 & -3.15 \end{bmatrix} \begin{Bmatrix} T_4 \\ T_5 \\ T_6 \\ T_7 \\ T_8 \\ T_9 \end{Bmatrix} = \begin{Bmatrix} -5.25 \\ -5.25 \\ -5.25 \\ -0.55 \\ -0.55 \\ -0.55 \end{Bmatrix} \tag{87}$$

Solving Eq. 87 gives the temperature value for all nodal points as $T_4 = 5.0, T_5 = 5.0, T_6 = 5.0, T_7 = 4.7, T_8 = 4.7,$ and $T_9 = 4.7$. The estimated temperature profile is similar to FEM solution in Sec. 2.3.3 with some nodal value variation.

3.4 Finite difference approximation of parabolic equation

The parabolic equation is a function of space and time. Thus, it involves at least two variables, time and space. It is always expressed in partial form since at least two variables are involved. The solution requires the boundary condition and the initial condition. The finite difference parabolic equation is different from an elliptic equation since the solution starts from the known time and propagates with increases in time.

3.4.1 One-dimensional parabolic finite difference equation

To illustrate the application of the finite difference method in solving a parabolic equation, we will consider the cooling process with time as previously noted in (Sec. 2.4.1). The one-dimensional finite difference time dependent equation can be generated by replacing one of the dimensions with time in Sec. 3.1. Furthermore, for parabolic equation the first and second derivative with time and space need to be defined. Thus, one-dimensional finite element parabolic equation becomes

$$\lambda \frac{(T_{i,\tau+1} - T_{i,\tau})}{\Delta t} = D \frac{1}{\Delta x^2} \left(T_{i+1,\tau} - 2T_{i,\tau} + T_{i-1,\tau}\right) + Q \tag{88}$$

The subscript i and τ are used to reprecent the space dimension (x) and time (t), respectively. One can see that with the known boundary condition and initial condition and using the appropriate time step, the new value of T may be predicted. A very simplistic explanation of the finite difference expression above can be given by letting the boundary condition $(i+1)$ and $(i-1)$ and the initial condition $\left(T_{i,\tau}\right)$ be known. Thus, the only unknown value is $T_{i,\tau+1}$ that is a function of the time step (Δt) the element size (Δx). The time step and the stability of the system are major parts of the parabolic equation. For example, for an explicit FD equation, the value for $\Delta t/\Delta x^2 \leq 0.5$ for the system to be stable and stability is a major part of numerical solutions (Smith, 1985; Dahlquist &Bjorck, 1974).

3.4.2 Application of FDM in one-dimensional parabolic equation

In order to demonstrate the application of FDM in parabolic equation, we will consider previously discussed cooling process of a thin insulated bar that initially at some temperature is placed in the environment where the heat allows to flow from the ends (Sec. 2.4.1). Similarly, , we will use four elements where the element size becomes 0.25. And let the time step be 0.001. Thus, the finite difference expression becomes

$$\frac{(T_{i,\tau+1} - T_{i,\tau})}{0.001} = \frac{10}{0.25^2} \left(T_{i+1,\tau} - 2T_{i,\tau} + T_{i-1,\tau}\right) \tag{89}$$

The boundary conditions are $T_1 = T_5 = 0$ and the initial condition is $t = 0,$ all $T = 1$ .Thus, the system equations becomes

$$\begin{Bmatrix} T_{2,\tau+1} \\ T_{3,\tau+1} \\ T_{4,\tau+1} \end{Bmatrix} = \begin{bmatrix} 0.68 & 0.16 & 0 \\ 0.16 & 0.68 & 0.16 \\ 0 & 0.16 & 0.68 \end{bmatrix} \begin{Bmatrix} T_{2,\tau} \\ T_{3,\tau} \\ T_{4,\tau} \end{Bmatrix} \tag{90}$$

Using the time step of 0.001 s, the temperature profile of $T_2 = 0.32, T_3 = 0.45, T_4 = 0.32$ estimated after 0.01 s. Furthermore, using the same process above, we wrote the computer program using 10 and 20 elements and solved the elliptic equation using 0.001s time step. The temperature profile with time is shown in Fig. 10. As expected, the rate cooling process with time is predicted using the finite element method.

3.4.3 Two-dimensional finite difference parabolic equation

Parallel to the above one-dimensional finite difference parabolic equation, the two-dimensional equation can be simply introduced by modifying the two-dimensional notation. The notation is modified to accommodate the time variable by using the superscripts instead of the subscripts for time. Thus, the finite difference two-dimensional parabolic (Eq. 61) becomes

$$\lambda \frac{(T_{i,j}^{\tau+1} - T_{i,j}^{\tau})}{\Delta t} = \frac{D}{\Delta x^2} \left(T_{i+1,j}^{\tau} - 2T_{i,j}^{\tau} + T_{i-1,j}^{\tau}\right) + \frac{D}{\Delta y^2} \left(T_{i,j+1}^{\tau} - 2T_{i,\tau}^{\tau} + T_{i,j+1}^{\tau}\right) + Q \tag{91}$$

Parallel to the previous one-dimensional case (Sec. 3.4.1), the boundary condition may be prescribed in several ways as a functional value and/or a flux depending on the situation. Since it is a time function, it also requires the initial condition.

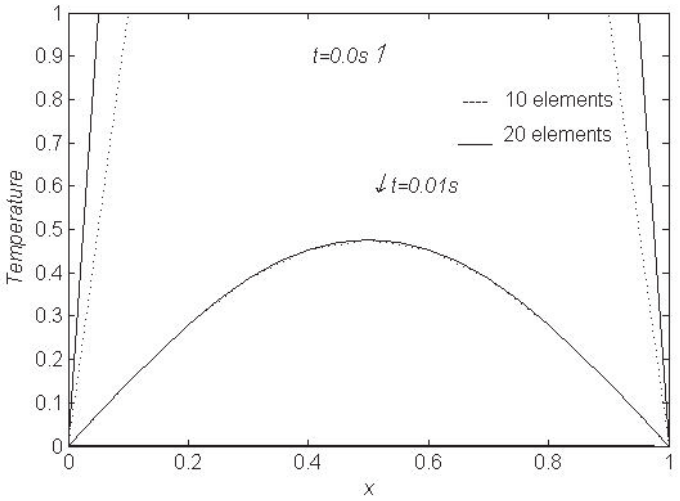


Fig. 11. The rate of cooling predicted with 10 and 20 linear elements using the finite difference method.

3.4.4 Application of FDM in two-dimensional finite difference parabolic equation

To illustrate the finite difference method application in solving two-dimensional parabolic equation, we will consider the temperature distribution of the two-dimensional rectangular region previously discussed in Sec. 2.4.4 and shown in Fig. 9. Using six elements, 2 in horizontal and 3 in vertical direction with specified boundary conditions, the two-dimensional FD parabolic equation becomes

$$\frac{(T_{i,j}^{\tau+1}-T_{i,j}^{\tau})}{\Delta t}=\frac{10}{0.5^2}\left(T_{i+1,j}^{\tau}-2T_{i,j}^{\tau}+T_{i-1,j}^{\tau}\right)+\frac{10}{0.5^2}\left(T_{i,j+1}^{\tau}-2T_{i,\tau}^{\tau}+T_{i,j+1}^{\tau}\right)+\frac{(10)0.5^2}{10} \tag{92}$$

Rearranging and forming in matrix form

$$\begin{Bmatrix}T_4\\T_5\\T_6\\T_7\\T_8\\T_9\end{Bmatrix}^{t+\Delta t}=\begin{Bmatrix}T_4\\T_5\\T_6\\T_7\\T_8\\T_9\end{Bmatrix}^{t}+\frac{10\Delta t}{0.5^2}\left(\begin{bmatrix}-4&2&0&1&0&0\\1&-4&1&0&1&0\\0&2&-4&0&0&1\\1&0&0&-4&2&0\\0&1&0&1&-4&1\\0&0&1&0&2&-4\end{bmatrix}\begin{Bmatrix}T_4\\T_5\\T_6\\T_7\\T_8\\T_9\end{Bmatrix}^{t}+\begin{Bmatrix}5.25\\5.25\\5.25\\1.25\\1.25\\1.25\end{Bmatrix}\right) \tag{93}$$

Thus, finite difference two-dimensional elliptic equation (Eq. 93) solution can be generated by solving the space equation for a fixed time first (solving the second right term in the bracket first) and using that to get the new estimate. By using the time estimate recursively, the time process may be generated. Similar to Sec. 2.4.4, the 0.001s time step, the temperature distribution of ($T_4=3.95, T_5=3.95, T_6=3.95, T_7=2.62, T_8=2.62,$ and $T_9=2.62$) is predicted after 0.1 seconds. The predicted values have similar profile and values with FEM solution in Sec. 2.4.4.

4. Concluding remarks

In this chapter, we illustrated numerical solutions of elliptic and parabolic equations using both finite element and finite difference methods. Elliptic and parabolic equations are encountered in numerous areas of engineering and science. Finite element and finite difference methods are the two most frequently applied numerical approximations, although several numerical methods are available. We illustrated how finite element method utilizes discrete elements to obtain the approximate solution of the governing differential equation. In addition, we showed how the final system equation is constructed from the discrete element equations. In addition, we also showed how finite difference method uses points over intervals to define the equation and the combination of all the points to produce the system equation. Both systems generate large linear and/or nonlinear system equations that can be solved by computer.

FEM and FDM are evolving with technology. The growth in computer technology has made it even more possible to consider using them in many science and engineering applications. In addition, more people without science and engineering backgrounds are becoming numerical simulation users. Consequently, the fundamental understanding of numerical simulation is becoming increasingly very important. Thus, this chapter intended to give some fundamental introduction into FEM and FDM by considering simple and familiar examples. We illustrated the similarity and the differences in finite difference and finite element methods by considering the simple elliptic and parabolic equations. Indeed, for the problems considered, one can see that the similarity and the difference from the final system equations and approximate solution. We designed the chapter to be introductory. By considering simple examples, we have illustrated FEM and FDM are reasonable ways of estimating solutions.

5. Acknowledgments

The author would like to thank the Food Processing Technology Division at the Georgia Tech Research Institute/Aerospace, Transportation and Advanced Systems Laboratory.

6. References

Awrejcewicz, J. & Krysko, V. A., Chaos in Structural Mechanics, 2010, Springer-Verlag, Berlin

Comini, G., Giudice,S.D. & Nonino. C. Finite Element Analysis in Heat Transfer, 1994, Taylor & Francis, Washington. DC

Clough, R.W., *Early history of the finite element method from the view point of a pioneer.* International Journal for Numerical Methods in Engineering, 2004. 60(1): p. 283-287.

Dahlquist, G. & Bjorck, A. Numerical Methods, 1974, Prentice-Hall, Englewood Cliffs, NJ

Segerlind, J.L., Applied Finite Element Analysis, 1984, John Wiley & Sins. Inc, New York.

Sahoo, P. K. & Riedel, T., Mean Value Theorems and Functional Equations, 1998, World Scientific, Singapore

Smith, G.D., Numerical Solution of Partial Differential Equations: Finite Difference Methods. 1985, Oxford: Clarendon Press.

Vichnevetsky, R., Computer Methods for Partial Differential Equations, Vol.1, 1981, Prentice-Hall, Inc, Englewood Cliffs, NJ

Yue, X., Wang, L., Wang, R., & Zhou, F. (2010). *Finite element analysis on Strains of Viscoelastic human skull and duramater,* InTech, ISBN 978-953-307-123-7

Zienkiewicz, O.C., *The birth of the finite element method and of computational mechanics.* International Journal for Numerical Methods in Engineering, 2004. 60(1): p. 3-10.

2

Data Analysis and Simulations of the Large Data Sets in the Galactic Astronomy

Eduardo B. de Amôres
SIM - Faculdade de Ciências da Universidade de Lisboa
Portugal

1. Introduction

In the last decades large data sets covering a wide range of both resolution and observed regions have been accumulated in the Astronomy. Follow-up technology involving the storage and the access to these data is also necessary to develop methods and tools that will allow their usage.
In this chapter, I will review the use of the large data sets in the galactic astronomy, most of them covering almost entire area of the sky in the several wavelengths, for both the diffuse data provided by IRAS, DIRBE/COBE, molecular and hydrogen surveys and point sources catalogues as provided by stellar large-scale surveys such as DENIS, 2MASS, SDSS, among others. A brief description of these surveys and how to access them in the context of Virtual Observatory will be also presented.
Numerical simulations have an important role in understanding and describing the nature of the observations. Particularly, large-scale surveys data can be also used to validate numerical models. I will present models and methods that describe the Galactic structure taking into account the hydrogen atomic distribution in our Galaxy, obtaining galactic parameters such as scale-height, spiral arms parameters, the co-rotation radius.
One of the biggest problems describing the spiral arms in the Galaxy from the gas distribution resides in the fact that some interpretations in the literature for ℓ-v the diagrams have not been updated in respect to rotation curves using old values for the distance from the Sun to the Galactic center. Another problem resides of the difficulty in describing the non-circular motion.
My choice in the present work is to describe the spiral structure by adopting an empirical model (section 3) which is based on the analysis of HI distribution by means its tangential directions and the observed ℓ-v. One of the aims of the present paper is to carry out self consistent inter-comparisons of our results regarding different tracers of the spiral structure.
The models of the spiral structure presented here were first introduced by Amôres & Lépine (2005, AL05) in which two models were proposed to describe the interstellar extinction in the Galaxy (see section 4 of this chapter). Model S consists in obtaining extinction predictions taking into account the spiral structure of our Galaxy. In this model, the extinction grows by steps each time a spiral arm is crossed and remains almost constant in the inter-arm regions. The models were also compared with other samples of objects and regions as pointed by Amôres & Lépine (2007).
As galactic interstellar extinction is a crucial obstacle when observing in the highly obscured regions, it is important to model it properly. I will present the recent efforts in this area. I

will present my model for interstellar extinction, its applications and comparisons with other models and maps available in the context of the Virtual Observatory called GALExtin. Star counts models are also important if one wants to estimate the counts that will be observed by large surveys (GAIA, DES, VVV, LSST, among others). A few simulations for large surveys using star counts are presented in section 5. Conclusions and final remarks are presented in section 6.

2. Large data sets

Large data sets are very important in whole Astronomy. In combination with models these can be used to construct methods to analyze and to interpretate the observed data. They can also be used in order to tune models allowing obtain the best values for a given set of parameters.

Particularly in the galactic Astronomy they can be used in order to obtain parameters of the galactic components, as the thin and thick disk, bulge, bar and galactic halo. Figure 1 shows a representation of our Galaxy structure that is composed basically of four components: disk (thin and thick), bulge, bar and halo. In the disk there is a presence of the spiral arms with young stars while in the halo there are most of the old stars (Robin et al. 2003 and references therein). The shape and properties for each component can be obtained by adjusting parameters of the galactic halo (such as eccentricity, shape) that can also allow us to identify, for instance, the existence of streams and satellite galaxies. (Majewski et al. 2003).

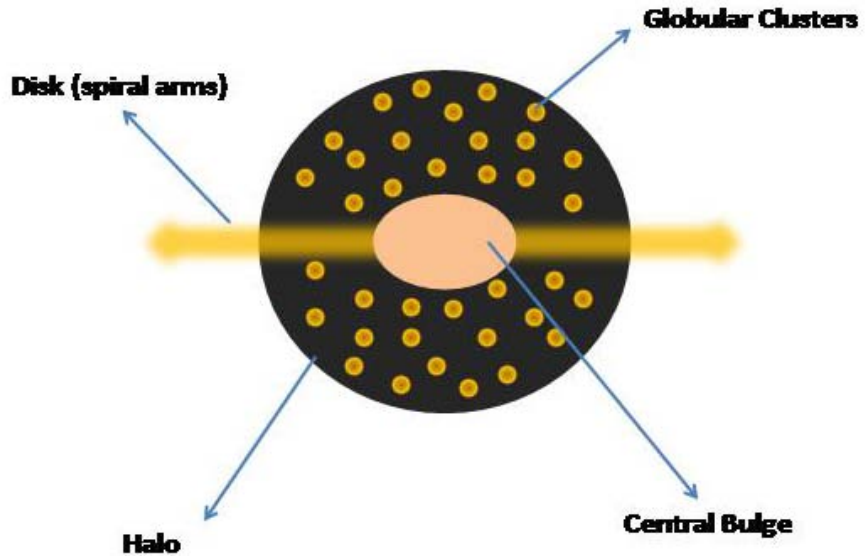


Fig. 1. Schematic Milky Way scheme

On the other hand, large data sets in the galactic Astronomy have an advantage to expand to a wide range of wavelength, from radio to high-energies. The first large surveys were obtained in the radio wavelengths in the decades of 1950-1960 (Kerr 1969) observing atomic hydrogen or HI. The author showed for the first time some interesting aspects of our Galaxy structure, as the warp, flare, spiral arms, among others and be also useful in the

extragalactic astronomy, like for instance the extinction maps elaborated by Burstein & Heiles (1978,1982) based on them. The most recent HI survey is based on the LAB (Leiden-Argentine-Bonn) survey that observed each point of the sky spaced by 0.5 and 0.25 degrees for galactic longitude and latitude producing for each one, what is called spectrum, i.e. the variation of intensity with velocity (Kalberla et al. 2005).
Figure 2 shows the distribution of the integrated intensity, i.e. the sum of the spectra for each coordinate, for all-sky as observed by LAB survey. It can be seen that most of the emission is concentrated for galactic latitudes $|b| \leq 15^o$, denoted by green, yellow and red colors but a significant part of the emission extended up to $|b| \sim 50^o$. One can see a long tail in the emission ranging from 120^o to 200^o reaching south galactic pole.

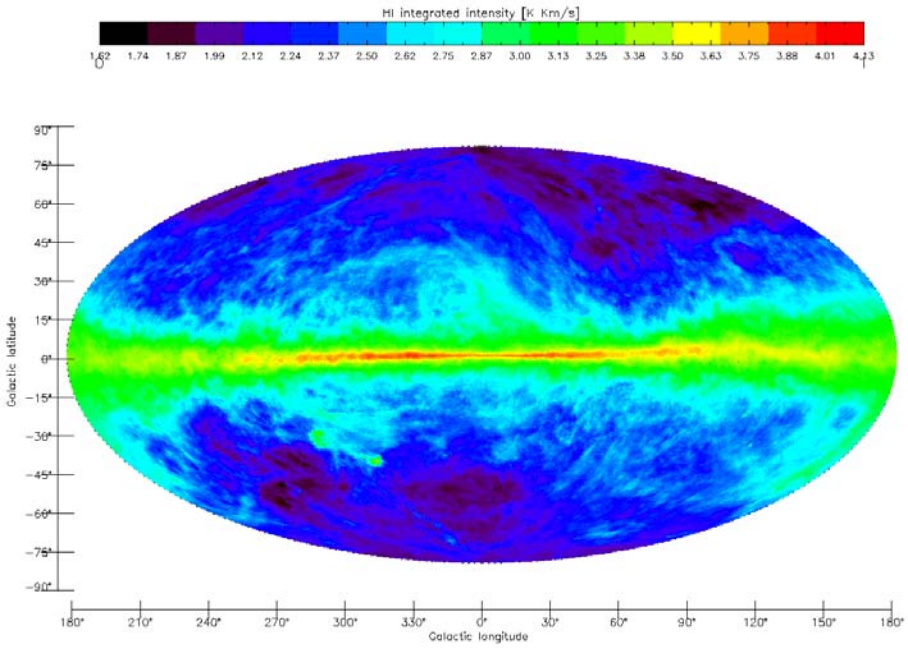


Fig. 2. Map for integrated intensity for neutral hydrogen in galactic coordinates obtained from LAB survey (Kalberla et al. 2005).

Table 1 shows some large surveys useful for galactic structure studies based on other wavelengths as CO that is particularly interesting to identify star-formation regions. Due to the distribution of the interstellar dust in the galactic plane, it obscurates the observations in some wavelengths as in the optical, for instance. Observations performed at infrared wavelengths penetrate through the dust grains and allow us to observe beyond. An extraordinary advance in the study of galactic structure was obtained with IRAS satellite (1983) that observed almost 96% of the sky in four infrared bands followed by DIRBE/COBE experiment in 1992.
Catalogs with stellar sources, as 2MASS (Cutri et al. 2003), SDSS, are also very important in the study of the galactic structure. Concerning 2MASS, Skrutskie et al. (2006) published a catalog which contains 1,647,599 observed extended sources using the classification

algorithms provided by Jarett et al. (2000). As this catalog also contains sources with bad photometry quality, contamination or confusion source due to either intrinsic 2MASS properties or high crowded fields for more detailed works is mandatory to select those sources in respect to quality criteria.
Figure 3 shows a map in galactic coordinates for 2MASS extended sources for all-sky with region observed by Vista Variable in the Via Láctea (VVV) survey (Minniti et al. 2010) region (bulge and disk) represented by white dashed line. The grid of this map is one squared degree for both longitude and latitude.

3. Simulating galactic structure from HI data

This section is organized as follows. Section 3.1 presents the data used in the present work, an analysis from the main emission peaks observed in the HI and CO galactic distribution is presented in section 3.2. Section 3.3 presents the procedure adopted for describing the spiral arms as well the galactic rotation curve used in the present work. The ℓ-v diagrams obtained from the fitting of the parameters of the spiral arms for the HI is presented in section 3.4. A discussion concerning the co-rotation point in our Galaxy is presented in section 3.5.

Wavelength	Survey	Coverage	References
HI	Berkeley	$10^o < \ell < 250^o$ ($\|b\| \leq 10^o$)	Weaver & Williams (1973)
	Parkes	$240^o < \ell < 350^o$ ($\|b\| \leq 10^o$)	Kerr et al. (1986)
	NRAO	$-11^o < \ell < 13^o$	Burton & Liszt (1983)
	LAB	All-Sky	Kalberla et al. (2005)
Near infrared	2MASS (point sources)	All-Sky	Cutri et al. (2003)
	2MASS (extended sources)	All-Sky	Skrutskie et al. (2006)
Optical	SDSS	10,000 square degrees	Adelman-McCarthy et al., (2009)
Proper motion	UCAC-3	All-Sky	Zacharias et al. (2010)
Mid and far-infrared	IRAS	All-Sky	Hauser et al. (1984)
Near/mid and far-infrared	DIRBE/COBE	All-Sky	DIRBE Explanatory Supplement
CO	Dame	$0^o < \ell < 360^o$ ($\|b\| \leq 10^o$)	Dame et al. (1987,2001)
	Stony-Brook	part of galactic plane	Clemens et al. (1986)

Table 1. List of some of the main surveys used in the galactic structure studies.

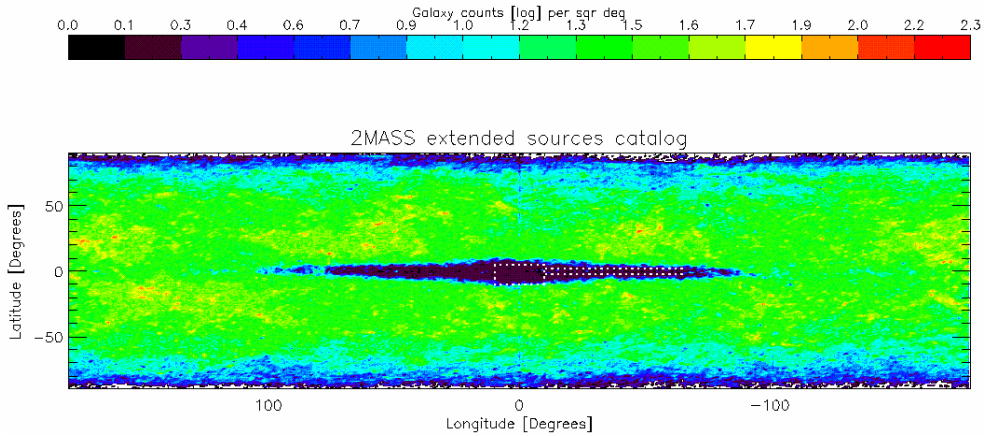


Fig. 3. Galaxy counts obtained by 2MASS. Dashed white line represents regions observed by VVV.

3.1 Data

Hartmann & Burton (1997) published a HI survey based on observations called Leiden/Dwingeloo HI survey that mapped the 21-cm spectral line emission over the entire sky above declinations of -30 degrees using a grid spacing of ~ 0.5 degree and a velocity sampling of ~ 1.03 km/s.

Kalberla et al. (2005) published the LAB (Leiden/Argentine/Bonn) survey which contains the final data release of observations of 21-cm emission from Galactic neutral hydrogen over the entire sky, merging the Leiden/Dwingeloo Survey of the sky north of -30º with the Instituto Argentino de Radioastronomia Survey of the sky south of -25º. The angular resolution of the combined material is HPBW ~ 0.6º. One of the improvements of this new survey consists of also doing corrections on the stray radiation. The LAB survey has been extensibly used in several applications as pointed by Bajaja et al. (2005), Kalberla et al. (2005), Haud & Kalberla (2007), Kalberla & Haud (2006) among others.

In the present chapter, I have employed the HI data from the LAB survey. This data comprises galactic longitudes from 0º to 360º and galactic latitudes from -90º to 90º for both the intervals the 0.5º and up to 1 km/s in velocity. The data are stored in 720 (b,v) fits file maps at longitude intervals stepped by 0.5º.

3.2 The HI and CO main emission complexes

The ℓ-v diagram constitutes an important and useful tool in studying the galactic structure. Reproducing it allows us to obtain relevant information about the distribution, the position and the gas density in the spiral arms, etc. Furthermore, the visualization and interpretation of the ℓ-v diagrams is also essential for their comparison with ones obtained from the predicted models, allowing us to analyze the structures that correspond to the spiral arms from the qualitative point of view, and also to obtain information about the velocity field for the HI and CO.

One way to perform a qualitative study of the ℓ-v diagram consists of estimating the points which correspond to the main peaks for these two gas components from their observed spectra. This task can be performed by fitting the gaussian for the observed HI and CO

spectra. This procedure allows us to identify the regions that delineate the main structures observed in the ℓ -v diagrams. The fitting of gaussians allows us to obtain the central value, the intensity and the width of the peaks that contributed most in the observed spectra.
Gaussian fits were obtained in the observed HI and CO spectra at each interval of 1 and 2º in longitude for HI and CO (in the galactic plane, i.e. b = 0º, respectively. The results of these procedures are also available since they may also be useful in other studies. Since some points with lower intensities can cause confusion in the identification of the large HI structures, for each spectrum the points were excluded for which the intensity is 30% less than the peak with the maximum intensity. This represents approximately 20% of the points obtained with this fitting procedure.
The ℓ -v diagrams obtained with this gaussian fitting procedure are presented in Figure 4. It can be seen that the HI distribution (Figure 4a) is given almost for the entire Galaxy while the CO emission is mainly predominant in the inner Galaxy (Figure 4b) since the CO is mainly concentrated in the spiral arms (see also Figure 5 of AL05).

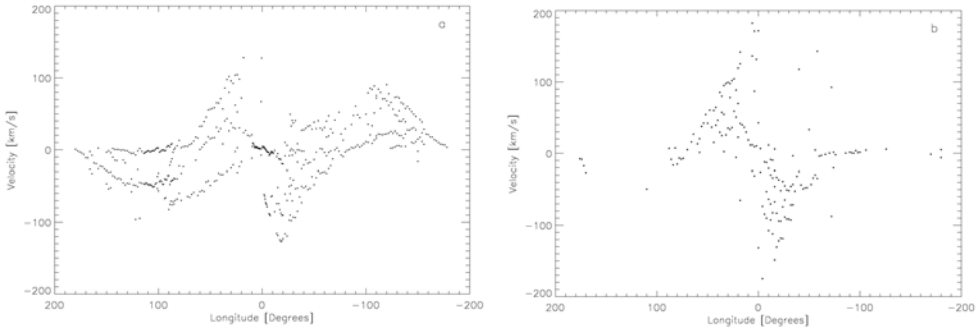


Fig. 4. ℓ -v diagram obtained from the gaussian fits performed for HI (a) and CO (b) spectra

In Figure 4a, one notes the presence of a spiral arm feature for (ℓ,v) = (~ -20,-120º) which also can be seen for HII regions (Amôres 2005). In the case of CO (Figure 4b) it is widely know that this component is distributed in the molecular cloud complexes that constitute the main place for star formation in our Galaxy. The fact that the molecular clouds and the CO emission are good tracers of the spiral structure is related to the interstellar shocks that occur in the spiral arms which produce an increase of the density transforming HI into H_2 (Marinho & Lépine 2000).
In Figure 4b, it is also possible to identify two tangential directions for $\ell \pm 30^o$ and $\ell \pm 50^o$. In the northern Galaxy, the component $\ell \sim 30^o$ splits into two other components, the first one at $\ell \sim 30^o$ and the other at $\ell \sim 25^o$. This feature was first identified by Solomon et al. (1985) from his observations in CO and it is also mentioned by Englmaier (1999). A detailed overview of the main characteristics observed in the ℓ -v diagrams for CO can be found in Fux (1999).

3.3 Description of the empirical models for the spiral arms

From the analysis of external galaxies it can be seen that the spiral logarithm (θ α lnR) fits real galaxies better than any other spiral curves. In my models the spiral arms are represented by the logarithmic spiral:

$$R = r_0 \exp[(\theta - \theta_0)\, tg(i)] \quad (1)$$

in which r_0 and are the polar coordinates which represent the initial arm radius and i is the pitch angle. In this way, to describe a spiral arm four parameters are necessary that specify the arm position in the galactic plane: r_0, θ_0, i, $\Delta\theta$ (the arm length). In addition to these variables, an extra term (δi) was added to the pitch angle in order to produce an effect of variable inclination angle.

Russell & Roberts (1992) from the morphological analysis of the galaxies NGC 5457 and NGC 1232 found an expression to describe the variation of the pitch angle in the spiral arm. This expression is given by $i = \sum A_n (r / r_o)^n$ in which A_n are the coefficients that represent the perturbation term A_0 … A_3 e r_0 corresponds to the initial radius of the spiral arm.

A comparison between the observed ℓ-v diagrams and ones obtained from fitting the parameters of the spiral arms will be presented in the next section. The main procedure consists of tracing a sequence of points in the galactic plane, i.e., in the X-Y coordinates (Figure 5a) in order to draw a hypothetical spiral arm segment. The next step consists of verifying whether this represents an observed structure (Figure 5b). If so, it is determined the distance to the galactic center for each point of this arm segment. Once the distance is obtained and its galactocentric radius.

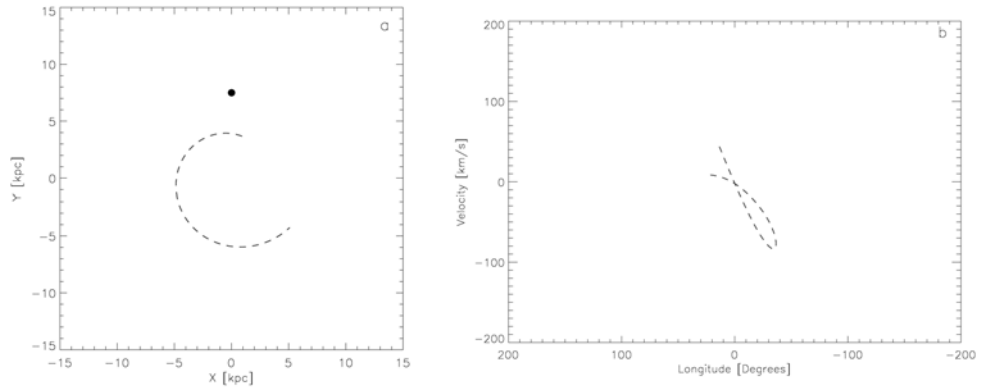


Fig. 5. a-) Arm segment represented in the galactic plane; b-) representation of the same arm in the ℓ-v diagram (the transformation from velocity into position is performed using the Clemens rotation curve and geometric properties).

With the rotation curve, it is possible to plot the longitude and velocity for this point. In short, X-Y positions were transformed into velocities. One position in the X-Y plane gives an unique point in the ℓ-v diagram without distance ambiguity.

Lépine et al. (2001) also presented a similar method to perform the representation of the ℓ-v diagram for HII regions. The main differences between that work and the present one are: i-) I reproduce the ℓ-v diagrams following an empirical scheme which describes the observed characteristics in these diagrams and not for a purely theoretical model; ii-) introduction of the velocity perturbation; iii-) presentation of the tangential directions obtained with my model; iv-) presence of arms with several positions with different pitch angle and; v-) I have used Russeil's catalog instead the Kuchar & Clark (1997) catalog for the HII regions.

In the present work, I also have used a modified Clemens rotation curve which is presented in Figure 6 in which the points represent the original data obtained by Clemens (1985) from

CO data for the interstellar medium. I introduced a modification in the original curve provided by Clemens in order to set R_0 = 7.5 kpc and to use a double exponential as presented below:

$$v_{rot} = 340\exp\left[-r/20.0 - 1.5/r\right] + 830\exp\left[-r/0.73 - 0.3/r\right] \quad (2)$$

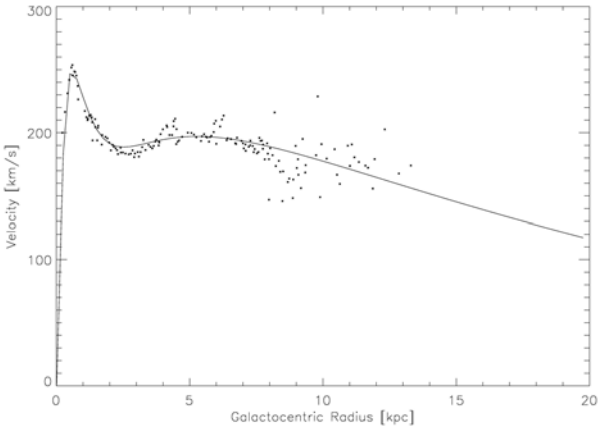


Fig. 6. Rotation curve for our Galaxy determined from the interstellar gas data (Clemens 1985). The lines represent the fit described by the expression above and the points correspond to the data obtained by Clemens.

A similar Clemens rotation curve represented as a double exponential was already presented by Lépine et al. (2001). However, in the present work, I performed a small modification in its parameters in order to improve the quality of the fit. In the present rotation curve, the *rms* errors from the comparison of the observed data and fitted expression by the first expression is equal to 5.043.

So, through the variation of parameters for the spiral arms (expression 1) were determined the best fit for the observed structures in the ℓ-v diagram which is also done verifying whether the face-on representation of the Galaxy is plausible as well whether the tangential directions are in agreement with the observed ones. It is also provide the Xi^2 total estimate calculated by the expression below:

$$\chi^2 = \frac{1}{N}\sum \min\left(\sqrt{(l_o - l_c)^2 + (v_o - v_c)^2}\right) \quad (3)$$

in which N denotes the number of fitted structures, l_0 and l_c are the observed and predicted longitude, with the same notation applying for the velocities v_0 and v_c. The function *min* represents the fact that we get only the point with the least difference. A similar formula was also used by Russeil (2003) in order to estimate the errors involved in the comparisons with her models of spiral arms with the observed HII regions.

Since a simple model of circular velocity does not adequately reproduce some of the characteristics observed in the ℓ-v diagrams, it is also necessary to introduce perturbation terms in the calculated velocities. The use and the justification of these terms were first

described by Ogorodnikov (1958) and is explained in detail by Mishurov & Zenina (1999a.b) and Lépine et al. (2001).

3.3.1 Calculating the perturbed velocities

As explained above, the procedure consists of transforming X-Y galactocentric coordinates of the arms into velocities. So each point with an X-Y coordinate has a corresponding point (ℓ,v) in the ℓ-v diagram that is calculated as explained below. The longitude is directly obtained from X-Y galactic coordinates:

$$\ell = a\tan(\frac{X}{R_0 - Y}) \tag{4}$$

in which R_0 is equal to 7.5 kpc which corresponds to the distance from the Sun to the galactic center. This value is actually largely used and one of the first mentions of its use was by Reid (1993). The final velocity is given by:

$$v = v_1 + v_{\tan y}\cos(\ell) + v_{\tan x}\sin(l) - v_{sun} \tag{5}$$

in which v_{tanx} and v_{tany} are tangential velocities projected in the X-Y plane; and v_1 is calculated from the expression below:

$$v_1 = (V_r(R) + p_{rot})\sin(\ell - \theta) - p_{rad}\cos(\ell - \theta) \tag{6}$$

$V_R(R)$ means the value of the rotation curve for a given galactocentric radius (R), the p_{rot} term corresponds to the perturbation in rotation and v_{sun} is the projected velocity of the Sun calculated by the relation:

$$v_{sun} = V_r(R_0)\sin(\ell) - v_{out}\cos(\ell) \tag{7}$$

in which v_{out} = 12.8 km/s.
The term which describes the radial and angular perturbation components is given by:

$$p_{rad} = a_2\sin(0.2 + a_1 R - 2\theta) \tag{8}$$

The term a_2 is described below:

$$a_2 = -8.0a\log(\frac{R}{r_c}) \tag{9}$$

in which r_c was set equal to 8.3 kpc and represents the co-rotation radius (section 3.5). The term a_1 = -0.4 represents the phase variation. The amplitude of the perturbation in the rotation is described by the expression:

$$p_{rot} = a_3\sin(\phi - 2\theta) \tag{10}$$

in which a_3 = 12.5*alog(R/r_c). Due to the existence of the LindBlad internal resonance (ILR, see Amaral & Lépine (1997, AL97)), the terms a_2 and a_3 are set equal to zero for r <2.5 kpc. The final values of these constants were obtained after some tests in order to reproduce the characteristics observed in the ℓ-v diagrams for both HI and HII regions.

3.4 Fitting the spiral arm parameters

In the next sub-section, the results of fitting the observed ℓ-v diagrams for the HI will be presented. For each one, it is presented the estimate (Table 2) of the Xi^2 obtained with models in comparison with the observed features as well as the same estimate with the ones obtained with a model of four arms and the superposition of 2+4 spiral arms. In order to reproduce these diagrams, I have manipulated the parameters r_0, θ_0, i, $\Delta\theta$ of the spiral arms described in expression 1.

It should be noted that in performing these fitting procedures the objective is not only to reproduce the observed ℓ-v diagram but also to ensure that the parameters obtained adequately reproduce the Galaxy face-on aspect and the observed tangential directions. In the case of the HI, they should also account for the tangential directions in the longitudinal profile for the integrated intensity (Figure 9).

Figure 7 presents the comparison between the observed ℓ-v and the one obtained after fitting the spiral arms parameters with the respective Galaxy face-on representation provided in Figure 8. In order to better visualize the results, the arms are represented by straight or dashed lines with different colors. The parameters are given in the Table 3. The final Xi^2 for this adjustment calculated by expression 3 is presented in Table 2, which also shows the Xi^2 expected with models of four and the superposition of 2 + 4 spiral arms.

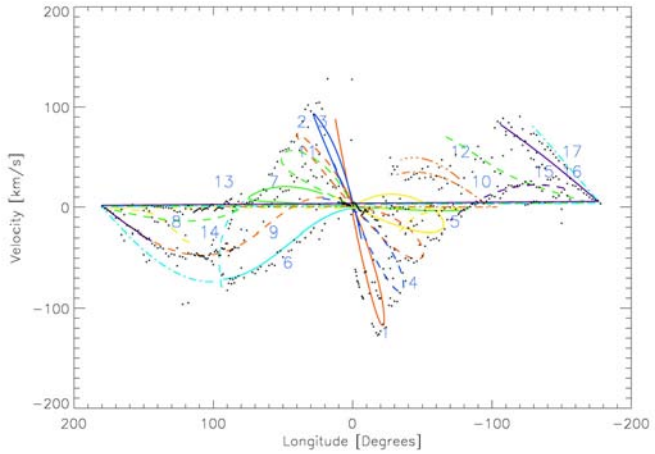


Fig. 7. ℓ-v diagram for the HI. The lines represent the spiral arms and the points the data obtained from the Gaussian fits for HI (section 3.2). The numbers correspond to the arm parameters described in the Table 3.

For the model of 2 + 4 spiral arms I have used the Model proposed by Lépine et al. (2001) with pitch angles equal to 6.8° and 13.5° for the pattern of 2 and 4, respectively. In the case of four spiral arms, the model proposed by Ortiz & Lépine (1993) was used which consists of a model with four spiral arms, all of which begin at r_0 equal to 2.3 kpc and with pitch angle equal to 13.8°, each one separated by a phase of 90°. I also add a local arm with i equal to 12.5° and size equal to 51°.

A number is presented next to each structure (according to the numbers presented in the Table 3). In total, there are 17 structures, among them: arms, bifurcations, bridges. The largest structures correspond to arms 1,2,3,4,7,11, located in the inner Galaxy with initial radius varying from 2.5 to 4.0 kpc.

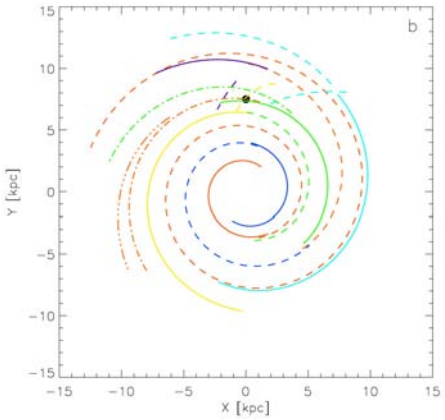


Fig. 8. Face-on Galaxy representation, where the lines represent the spiral arms, with the same colors as Figure 7.

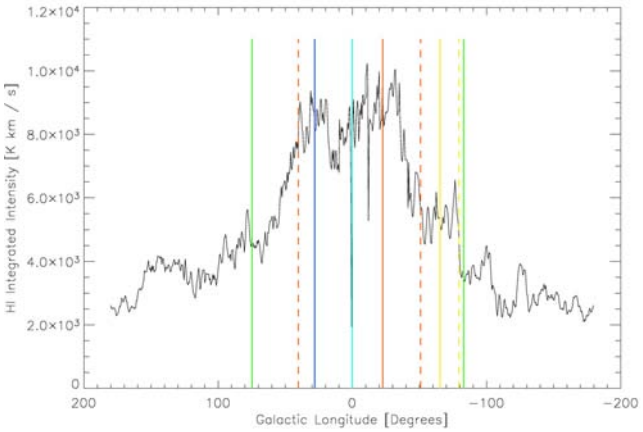


Fig. 9. Longitudinal profile for the integrated intensity for the HI where the vertical lines are the tangential directions predicted by the model.

HI model	5.265
4 arms	17.536
2+4 arms	8.189

Table 2. Comparison between the Xi^2 square estimate obtained from my model and the ones obtained from other models.

Figure 9 shows the longitudinal profile for the integrated intensity for the HI ($b = 0^o$) that was calculated from the HI surveys mentioned in section 3.1. The lines represent the tangential direction for each arm in which the color follows the same representation adopted above. Below a brief discussion is provided about each structure. Arm 7 with $r_i = 6.22$ kpc is one exception of a long arm that does not begins at low radius. Arm 5, which is represented

by the thin yellow line which can be seen in Figure 7, is a prolongation of arm 11 (green dashed). Arm 11 has r_i = 3.95 kpc with pitch angle equal to 9.40°.
However, to continue describing this arm and reproduce adequately the observed ℓ-v diagram it is necessary to introduce a new arm structure with the pitch angle slightly changed. This is done changing mainly the parameters θ and r_i in order to coincide with the arm end. Note that without this modification the loop extremity will fall in (~ -70,-10°) and not in (~-65,-20°) which represents a better description for the structures observed for the HI in this direction. In summary, structures 5 and 11 represent a unique arm with a modified pitch angle in order to reproduce the observed ℓ-v diagram. It should be noted that arms with variable pitch that can also can be seen in other galaxies, such as M 81, for example.

arm	i (°)	radius (kpc)	initial phase (°)	δi (°)	size (°)
1	6.60	2.35	-31.0	0.0	240
2	6.70	3.80	195.0	0.0	330
3	7.10	2.57	155.0	0.0	200
4	7.50	3.80	-14.0	0.0	245
5	6.35	7.50	-2.50	0.0	182
6	7.50	7.60	163.0	0.0	155
7	4.20	6.22	228.0	0.0	149
8	-20.30	9.02	345.0	-1.80	24
9	11.55	6.44	225.0	0.0	135
10	7.77	7.36	-8.40	0.05	139
11	9.40	3.95	190.0	0.0	170
12	11.50	8.60	45.40	0.00	79
13	12.80	7.35	330.0	0.00	109
14	-40.0	11.40	315.0	0.0	31
15	12.0	10.10	-10.0	0.0	48
16	9.00	5.00	-45.0	0.1	25
17	11.10	11.10	-38.0	0.0	65

Table 3. Main parameters for the arms used to fit the ℓ-v diagram for the HI. Each number represents the arm in the figure 7.

A similar characteristic is noted for arms 2 (dashed red), 6 (light blue) and 17 (dashed blue light). In this case, if we increase the size of arm 2 (Figure 10a e 10b) without changing the i and r_i we obtain an inadequate description of the HI distribution at $0° < \ell < -100°$. Arm 6 is a prolongation of arm 2, but with i = 7.5, which better represents the HI distribution in this region with the line that passes through arm 6 (light blue) providing a good match for almost all of the points in this direction. Arm 17 matches better the structures observed in $-140° < \ell < -170°$ than arm 2. The i of the arm 17 is equal to 11.1°. An interesting characteristic can be seen at the end of this structure which corresponds to a bifurcation (arm 14) as well as the beginning of the other spiral arm (arm 17).
In the Galaxy face-on representation one also notes three structures that resembles bridges (arms 8,10,14), two of them with $i < 0$. These structures were introduced in order to reproduce the observed ℓ-v diagram. Structure 14 matches most of the points from $\ell \sim -70°$ to $\ell \sim -150°$. These structures could be explained by the proximity of the Sun to the co-rotation point (see section 5). Similar characteristics were also assumed by De Simone et al. (2004).

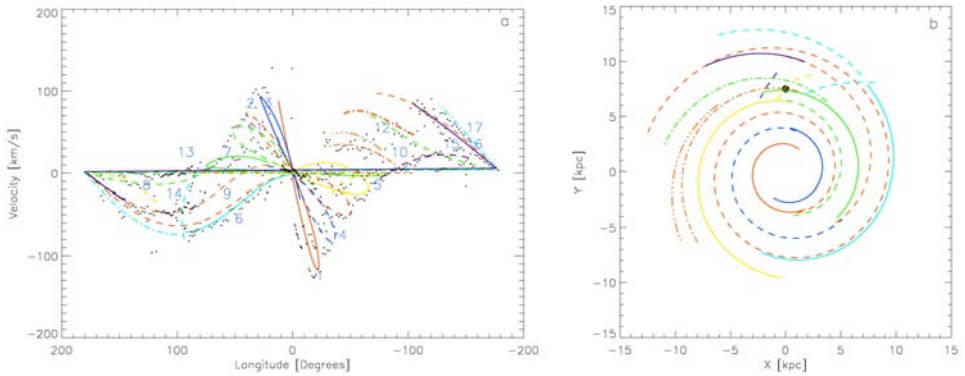


Fig. 10. a-) ℓ -v diagram with the modification of the extension for the arm 2 (red dashed) from 330 to 600º; b-) face-on aspect for this set of arm parameters. The color and the representation are the same adopted in figures 7, 8 and 9.

The violet dashed line (arm 14) in Figure 8 represents a structure that seems to be a bifurcation that begins in arm 15 (heavy green line). This arm describes the structures around ℓ < -100º in ℓ -v the diagram. Based on Figure 7, we observe that arm 9 reproduces adequately the structures observed in the ℓ -v diagram for ℓ > 80º. This arm is a prolongation of arm 4 with i = 11.55. The continuation of arm 9 is arm 16 (heavy blue line) with i = 9.0.

McClure-Griffiths et al. (2004), while analyzing ATCA data, detected a structure that extends from ℓ = 260º to 330º that they attributed to a spiral arm with i = 9.0º. This structure was also found in our results as represented by arm 12 with i = 11.5º. Kerr (1969) published latitude-velocity diagrams in which a bridge appears in the third quadrant toward positive velocities. Kerr (1969) also presented maps with this feature extending from ℓ = -150º to -70º. Davies (1972) also interpreted this region as relating to the end of a spiral arm associated with high velocity clouds.

3.5 Co-rotation radius

Since the first HI observations the existence of a hole was seen in its distribution in the Galaxy (Kerr 1969 and Burton & Gordon 1978). In these works, the authors noted a gas deficiency in the radial gas distribution for R = 11 kpc, assuming R_0 = 10 kpc. From the theoretical point of view Marochnik et al. (1972), Crézé & Mennessier (1973) and AL97 also found evidence that the co-rotation point, i.e., the point where the rotation velocity of the spiral pattern coincides with the rotation curve of the gas and stars could be located in the Solar neighborhood. The effect of the co-rotation can be understood in the following terms: there is a region where movement exists pumping the gas inside and outside resulting in a deficiency of the gas in these regions. This effect was also studied in detail by Suchkov (1978), Goldreich & Tremaine (1978), Gorkavyi & Fridman (1994). Mishurov & Zenina (1999a) also found evidence that the co-rotation point would be located near the solar radius.

Amôres, Lépine & Mishurov (2009, ALM09) analyzed the HI spectra of the whole galactic longitudes range in steps of 0.5º, with galactic latitudes in steps of one degree in the range ± 5º, plus the additional latitudes ± 10º detecting for each spectrum the velocity of the deep

minima which are present, simply by identifying the channel with the lowest value of antenna temperature.
They obtained that distribution of the density minima in the galactic plane, derived from their kinematic distances from the Sun (is shown in Figure 11). The minima observed at different latitudes are all projected on the galactic plane. The ring-shaped gap is circular and very clear. It looks like the Cassini division in Saturn's rings.

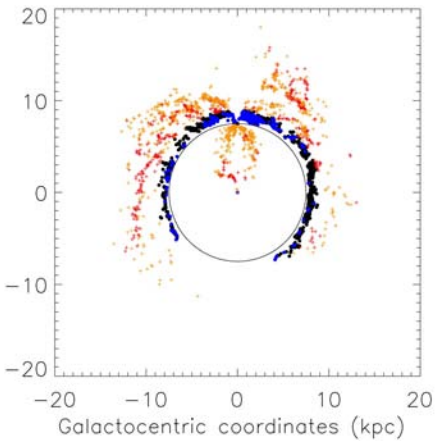


Fig. 11. Galactocentric distance distribution of the gaps for the whole sample of measurements including different latitudes. The Gaussian fit is centered at 8.25 kpc (Adapted from ALM09, figure 4).

The analysis of HI spectra allows to determine the points in which the HI has a gap in its distribution by means of an empirical method. The histogram with the galactocentric radius of each one of these gaps showed that the most of them are located nearest the solar radius which highlights the fact that the Sun is located at the co-rotation point in the Galaxy.

4. Interstellar extinction

One of the main difficulties in the study of both properties of individual objects and the galactic structure resides in the interstellar extinction determination. However, due to the clumpiness distribution of the interstellar dust, it is difficult to model it. Several works have been done in order to model the interstellar extinction distribution in the Galaxy, as presented by Arenou et al. (1992), Hakkila et al. (1997), Méndez & van Altena (1997), Drimmel et al. (2003), Amôres & Lépine (2005), Marshall et al. (2006) among others. In addition, some maps were elaborated in order to provide the integrated extinction along the line of sight, as the Burstein & Heiles (1978,1982), Schlegel et al. (1998, SFD) for the whole Galaxy, Schultheis et al. (1998) and Dutra et al. (2003) for the galactic center region, as well as Dobashi et al. (2005) map for A_V for the whole Galaxy ($|b| < 40^o$) based on the star counts method.
It is very important and fundamental for many studies to know the interstellar distribution in our Galaxy. This could be useful for estimation of distances and color corrections of

objects for which the distance can be obtained by some other method, for star counts and brightness models of the Galaxy, also for spectrum extinction correction, among other applications. In this context, we are developing a VO-Service called GALExtin (http://www.galextin.org) that provides the interstellar extinction estimate for any direction in the sky from: 2D maps and 3D models available and catalogs with extinction measure as well as diffuse emission. It is also useful to study the distribution of interstellar extinction towards star clusters. This service is very useful since most of the models and maps require the installation of programs and large files. The users may provide a list with coordinates and distances and the GALExtin (Amôres et al. 2011a) will produce as an output a list with extinction estimates for each object for a chosen model.

The determination of the interstellar extinction is fundamental for all fields of astronomy, from stellar studies in which it is important to know the extinction towards a given object from its coordinates and distances (3D Models), to extragalactic astronomy field; in this case it is mandatory to know the contribution of the extinction of our Galaxy along the line of sight.

Amôres & Lépine (2005) proposed two models to describe the interstellar extinction in the Galaxy (http://www.astro.iag.usp.br/~amores) both based on the hypothesis that gas and dust are homogeneously mixed, and that consequently, the known distribution of the gas can be used to describe the dust distribution. In the first model (model A) it was assumed that the Galaxy presents azimuthal symmetry, with the gas density depending only on the Galactic radius and on the distance from the galactic plane; as a result, in any direction the extinction increases smoothly with the distance from the Sun. In the second model (model S) the spiral structure of the Galaxy was taken into account; the parameters of the spiral arms were adjusted by fitting the observed longitude-velocity diagrams of different tracers, such as: CO, 100 μm, HI and H II regions (Amôres, 2005). In this model the integrated extinction grows by steps each time a spiral arm is crossed, and remains almost constant in the inter-arms regions (Amôres & Lépine 2007, AL07).

In this latter work, AL07 compared the extinction models with a sample of globular and open cluster and elliptical galaxies. From the comparison with elliptical galaxies the difference between Model A and Burstein & Heiles (1978) and Schlegel et al. (1998) were around 20-30 %. Figure 12 shows a map with the comparison with SFD maps for whole sky with δ E(B-V), i.e. difference between Model A and the SFD maps.

As pointed by AL07 it can be seen regions (indicated by white color) in which the Model A2 predicts less extinction than provided by SFD, i.e., the $\delta E(B-V) < -0.4$, notably around $150^o \leq \ell \leq 200^o$, in a region that extends towards negative latitudes. This region was also pointed by B03 as the region where there is a high variation in the gas-to-dust ratio.

The resolution and coverage of the different models and maps is another point that should be taken into account in the extinction studies. There are models that provide better estimates in the solar neighborhood up to distances equal to 1.0 kpc, while they fail to describe extinction outside this distance. On the other hand, there are models that provide better values at 3.0 kpc, but being imprecise at low distances. Depending on the assumptions used in their elaboration, some models can be used only for a given region in the sky.

Another interesting issue in galactic extinction is the combination of different models. In this sense, Amôres & Robin (2011) studied the variation for interstellar extinction along a line of sight joining the results from AL05 and Marshall et al. (2006, M06).

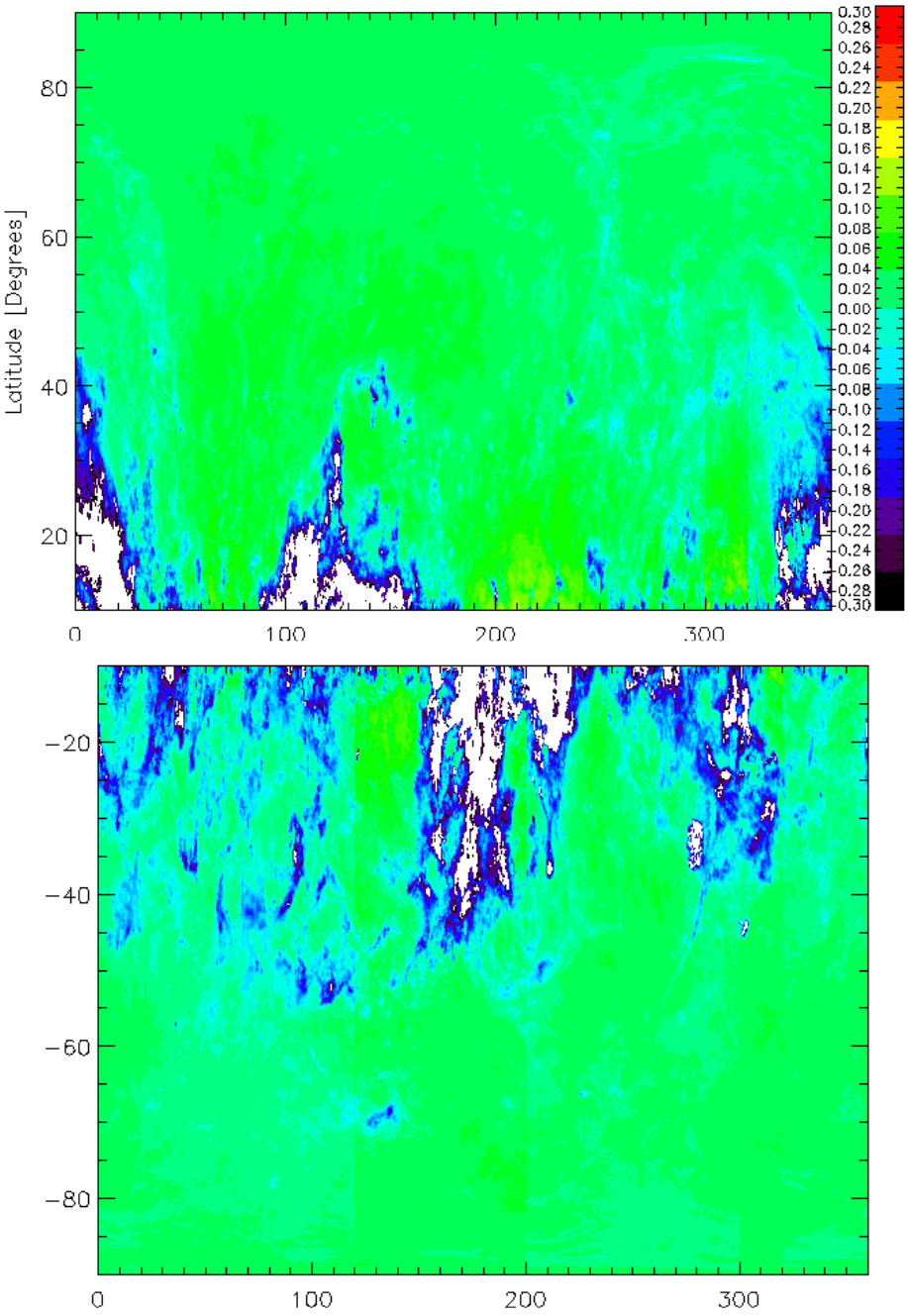


Fig. 12. Map in galactic coordinates of $\delta E(B - V)$ from the comparison of AL07 model with SFD98 for the whole Galaxy, using a grip spacing of 0.25° in l and b (Adapted from AL07, figure 3).

5. Star counts simulations

Star counts models have been largely used in astrophysics. They constitute an important and useful tool to study the galactic structure and its evolution. In particular, the Besançon Galaxy Model (BGM) provides a Galaxy description in the evolutive point of view joining both the kinematics and dynamics properties. One of the main differences in relation to others star counts models resides in the fact that the BGM is dynamically self-consistent (Bienayme et al. 1987).

The Besançon Galaxy Model (BGM) has been employed in a series of astrophysical applications (Robin et al. 2003). In the context of the GAIA mission, the BGM is the Model used to simulate the stellar content of our Galaxy. Amôres et al. (2007) presented a comparison and the improvements performed in the BGM - Java version in order to produce similar results as produced by the BGM in the Fortran version. Robin et al. (2009) also presented expected observations that will be performed by GAIA using BGM.

Amôres & Robin (2011) are using BGM in order to retrieve the parameters of spiral arms and other Galactic parameters, as warp flare by comparing BGM results with 2 MASS observations using the genetic algorithms.

On the other hand, Amôres et al. (2010) also presented simulations for the expected observations that will be performed by VVV. Vista Variables In The Via Láctea (VVV) is an ESO variability (Minniti, D., et al. 2010) survey that is performing observations in near infrared bands (Z,Y, J,H and Ks) towards the galactic bulge and part of the disk, totalizing an area of 520 square degrees. A total of 1920 observation hours (2009-2013) will be used at VISTA, within a 5 year time lapse.

In order to predict star counts distribution towards the VVV region we made use of TRI-Legal Galaxy Model (Girardi et al. 2005, G05) with galactic parameters as pointed by Vanhollebeke et al. (2009) and G05 considering VVV completeness limit, i.e. K = 20.0. For other filters, we have cut (after simulation) at other VVV limits, i.e. Z= 21.6, Y = 20.9, J = 20.6, H =19.0. The results are presented by Amôres et al. 2011b.

Since VVV observes galactic plane regions it is very important to know the interstellar extinction (Amôres et al. 2009) distribution towards them. We made use of 3D Marshall et al. (2006) interstellar extinction model. TRI-Legal predicts the star counts taking into account three galactic components, i.e. bulge, disk and halo. For each simulated star there is a set of properties as its distance, magnitude, gravity, temperature among others. From temperature we determine the spectral type.

6. Conclusions

Concerning HI model to describe spiral arms positions from HI data, the results presented in this chapter allow us to obtain the spiral arm positions based on HI distribution obtaining the spiral arm parameters (r_0, θ_0, i, $\Delta\theta$) which reproduce the main observed features in the ℓ -v diagrams for HI. The tangential directions predicted by models are also consistent with the ones predicted by other models, such as for instance the one proposed by Englmaier (1999, see his Table 1), as can be seen in the peaks existent in the longitudinal profiles for HI (Figure 9).

It should be noted that it is not being proposed a Galaxy with too many spiral arms, 17 for the atomic hydrogen since many of the structures presented in this chapter only represent prolongations of arms with different pitch angles which was necessary in order to adequately reproduce the observed ℓ -v diagram.

This is because a more simplified model would not be realistic. Many arm segments with different pitch angles are needed because real arms are not well represented by a single logarithmic spiral that follows a long path around the Galaxy. Furthermore, there are bifurcations and some short bridges. Another difficulty related to HI consists in the fact that these components are not only predominantly concentrated in the spiral arms, such as the CO, for example. Instead, there is also an amount of HI in the inter-arm regions.
In this sense, the use of the method proposed by Amôres & Lépine (2004) will also allow to reproduce not only the positions of the arms but also their density, allowing the elaboration of numerical ℓ-v diagrams, to estimate width of the arms since they are not as thin and well-resolved as presented here.
In relation to interstellar extinction important results are also being obtained from the inter-comparisons of different values of interstellar extinction obtained from Planetary Nebulae (PNe). Köppen & Amôres (2011) have used a large sample of PNe towards the galactic disk and bulge in order to obtain average values for interstellar extinction as well as to review the old values, most of the times obtained with different extinction curves.
A good estimate of interstellar extinction is also important in the study of Supernovae. Arsenijevic & Amôres (2011) have compared the distribution of interstellar extinction towards supernovae and also presented a new method to reproduce observed spectra using the Genetic Algorithms and basic parameters as galactic and galaxy host extinction and R_V (the ratio of the total to selective extinction). The use of global optimization methods can be a powerfull tool to obtain parameters from large surveys.

7. Acknowledgements

E. B. de Amôres acknowledges support from FCT under grant no. SFRH/BPD/42239/2007 and also Laboratory for Systems, Instrumentation and Modeling in Science and Technology for Space and the Environment (SIM), funded by "Fundação para a Ciência e a Tecnologia" thought Portuguese National Funds.

8. References

Adelman-McCarthy, J. K.; et al., 2009, VizieR On-line Data Catalog: II/294

Amaral, L. H. & Lépine, J. R. D. 1997, MNRAS 286, 885 (AL97)

Amôres & Lépine, 2004, Milky Way Surveys: The Structure and Evolution of our Galaxy, Proceedings of ASP Conference 317. Edited by Dan Clemens, Ronak Shah, and Teresa Brainerd. San Francisco: Astronomical Society of the Pacific, 2004, 317

Amôres , E. B., 2005, PhD Thesis, Universidade de São Paulo

Amôres, E. B. Lépine, J. R. D. 2005, AJ, 130, 679 (AL05, Paper I)

Amôres, E. B. Lépine, J. R. D. 2007, AJ, 133, 1519

Amôres, E. B. Lépine, J. R. D. 2009, Mishurov, Y., MNRAS, 400, 1768

Amôres, E. B. ; Moitinho, A. ; Arsenijevic, V, Sodré Jr., 2011, GALExtin: A VO-Service to estimate galactic interstellar extinction. In: Jenam 2011 Stars clusters in the Era of Large Scale-Surveys, 2011 Lisboa. Proceedings of Stars clusters in the Era of Large Scale-Surveys, 2011

Amôres, E. B. ; Padilla, N. ; Sodré Jr., L. ; Minniti, D. ; Barbuy, B., 2011, Simulations of star and galaxies counts towards Vista Variables in the Via Láctea survey region. In:

Environment and the formation of galaxies: 30 years later, 2011, Lisboa. Proceedings of Environment and the formation of galaxies: 30 years later, 2011.
Amôres, E. B., Robin, A. C., 2011, in preparation
Amôres, E. B., Robin, A. C., Luri, X., Masana, E., Reylé, C., Babusiaux, C. , 2007Use of the Besançon Galaxy Model for the Gaia Mission Simulator. In: VII Scientific Meeting of the Spanish Astronomical Society (SEA), Barcelona. Highlights of Spanish Astrophysics IV. Proceedings of the VII Scientific Meeting of the Spanish Astronomical Society (SEA). Dordrecht : Springer-Verlag, 2007
Amôres, E. B., Sodré, Jr., L, Barbuy, B., 2009, Low extinction windows in the VVV survey region. In: Special Session 8 at the IAU General Assembly 2009 - The Galactic Plane, in depth and across the spectrum, 2009, Rio de Janeiro. Highlights Of Astronomy, 2009. V. 15.
Arenou, F., Grenon, M. & Gomez, A. 1992, A&A, 258, 104
Arsenijevic, V., Amôres, E. B., 2011, submitted to MNRAS
Bajaja, E., Arnal, E. M., Larrarte, J. J., Morras, R., Pöppel, W. G. L., Kalberla, P. M. W., 2005, A&A, 440, 767
Bienaymé, O., Robin, A.C., Crézé, , M. 1987, A&A, 180, 94
Burstein D., Heiles, C. 1978, ApJ 225, 40
Burstein D., Heiles, C. 1982, AJ 210, 341
Burton, W. B., Gordon, M. A., 1978, A&A, 63, 7
Burton, W. B., & Liszt, H. S. 1983, A&AS, 52, 63
Burton, W. B., Liszt, H. S.,1993, A&A, 274, 765
Clemens, D. P., Sanders, 1985 ApJ 295, 422
Clemens, D. P.; Sanders, D. B.; Scoville, N. Z.; Solomon, P. M., 1986, ApJ Suppl. Series, 60, 297
Crézé, M., 1973, Mennessier, M. O., A&A, 27, 281
Cutri, R. M, et al., 2003, VizieR On-line Data Catalog: II/246
Dame, T. M., Ungerechts, H., Cohen, R. S., de Geus, E. J., Grenier, I. A., May, J., Murphy, D. C., Nyman, L.-A., Thaddeus, P., 1987,ApJ, 322, 706
Dame, T. M., Ungerechts, T. M., Cohen, R. S., Thaddeus, P., 2001, ApJ, 547, 792
Davies, R. D., 1972, MNRAS, 160, 381
De Simone, R., Wu, X., Tremaine, S., 2004, MNRAS, 350, 627-643
DIRBE Explanatory Supplement, 1998, http://lambda.gsfc.nasa.gov/product/cobe/dirbe_exsup.cfm
Dobashi, K. et al. 2005, PASJ, 57, 1
Drimmel, R., Cabrera-Lavers, A. & López-Corredoira, M. 2003, A&A, 409, 205
Dutra, C. M.; Santiago, B. X.; Bica, E. L. D.; Barbuy, B., 2003, MNRAS, 338, 253
Englmaeir, P. & Gerhard, O. 1999, MNRAS 304,512
Fux, R., 1999, A&A, 345, 787
Girardi,L.; Groenewegen,M.A.T., Hatziminaoglou,E., da Costa,L., 2005, A&A, 436, 895
Goldreich, P., Tremaine, S., 1978, Icarues, 34, 227
Gorkavyi, N. N., Fridman, A. M., 1994, Physics of Planetary Rings: Celestial Mechanics of Continiuous Medium, Moscow : Nauka
Hakkila, J. et al. 1997, AJ, 114, 2043
Hartmann D., Burton W.B., 1997, Atlas of Galactic Neutral Hydrogen, Cambridge University Press

Haud, U., Kalberla, P. M. W., 2007, A&A, 466, 555
Hauser, M., et al. 1984, ApJ, 285, 74
Jarrett, T. H., Chester, T., Cutri, R., Schneider, S., Skrutskie, M., Huchra, J. P., AJ, 2000, AJ, 119, 2498
Kalberla P.M.W., Burton W.B., Hartmann D., Arnal E.M., Bajaja E., Morras R., Poeppel W.G.L., 2005, A&A, 440, 775
Kalberla, P. M. W., Haud, U., 2006, A&A, 455, 481
Kerr, F. J., 1969, ARA&A 7, 163
Kerr, F. J., Bowers, P. F., Jackson, P. D., Kerr, M., 1986, A&A Suppl. Series, 66, 373
Köppen, J., Amôres, E. B., 2011, in preparation
Kuchar, T. A, Clark, F. O., 1997, ApJ, 488, 224
Lépine, J. R. D., Mishurov ,Y. & Dedikov, Y. 2001, ApJ, 546, 234
Majewski, Steven R.; Skrutskie, M. F.; Weinberg, Martin D.; Ostheimer, James C., 2003, ApJ, 599, 1082
Marinho, E. P., Lépine, J. R. D., 2000, A&A, 142, 165
Marochnik, L. S., Mishurov, Yu. N., Suchkov, A. A., 1972, Ap&SS, 19, 285
Marshall, D.J., Robin, A.C., Reyl\'{e}, C., Schultheis, M., Picaud, S., 2006, A&A, 453, 635
Mendez, R. A. & van Altena, W. F. 1998, A&A, 330, 910
McClure-Griffiths, N. M., Dickey, J. M., Gaensler, B. M., Green, A. J., 2004, ApJ, L127
Minniti. D., et al., 2010, New Astronomy, V. 15, Issue 5, p. 433-443
Mishurov ,Yu., Zenina, I. A., 1999a, A&A, 341, 81
Mishurov, Yu., Zenina, I. A., 1999b, Astronomy Reports, 43, 487
Ogorodnikov, K.F. 1958, The Stellar System Dynamics, Physical-Mathematical Publishing House, Moscow, USSR
Ortiz, R., Lépine, J. R. D., 1993, A&A, 279, 90
Paladini, R., Davies, R. D., DeZotti, G., 2004, MNRAS, 347, 237
Reid, M. J., 1993, ARA&A, 31, 345
Robin,A. C.,Reylé, C.,Derriere, S.,Picaud, S. 2003, A&A, 409, 523
Robin, A. C.; Reylé, C.; Grux, E.; The Gaia Dpac Consortium, 2009, SF2A-2009: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, held 29 June - 4 July 2009 in Besançon, France. Eds.: M. Heydari-Malayeri, C. Reylé and R. Samadi, p.79
Russeil, D. 2003 A&A, 397, 133
Russell, W. S., Roberts, W. W., 1992, ApJ, 398, 94
Skrutskie, M. F. et al., 2006, AJ, 131,1163
Schlegel, David J., Finkbeiner, Douglas P., Davis, Marc,1998, ApJ, 500, 525
Schultheis, M., Ganesh, S., Simon, G., Omont, A, Alard, C., Borsenberger, J., Copet, E., Epchtein, N. Fouqué, P., Habing, H., 1999, A&A, 349, L69
Solomon, P. M., Sanders, D. D., Rivolo, A. R., 1985, ApJ, 292, 19
Suchkov, A. A., 1978, AZh, 55, 972
Vanhollebeke, E, Groenewegen, M. A. T., Girardi,L., 2009, MNRAS, 498, 95
Weaver, H., Williams, D. R. W.: 1973, A&A Suppl. Series, 8, 1
Zacharias et al. 2010, AJ, 139, 2184

3

Methods for Blind Estimation of the Variance of Mixed Noise and Their Performance Analysis

Sergey Abramov[1], Victoria Zabrodina[1], Vladimir Lukin[1],
Benoit Vozel[2], Kacem Chehdi[2] and Jaakko Astola[3]
[1]*National Aerospace University,*
[2]*University of Rennes 1,*
[3]*Tampere University of Technology,*
[1]*Ukraine,*
[2]*France,*
[3]*Finland*

1. Introduction

Modern imaging systems provide a huge amount of images nowadays. These images are of different original quality. Some of them are practically ready for exploitation, e.g., visual inspection, object recognition, etc. Other ones need to be pre-processed as, e.g., by filtering, edge detection, segmentation, compression (Pratt, 2007; Bovik, 2000; Al-Shaykh&Mersereau, 1998), etc. In the latter case, it is desirable to know noise type and characteristics (Pratt, 2007; Elad, 2010). Such information is exploited by modern methods and algorithms of image denoising (Elad, 2010; Sendur&Selesnick, 2002; Donoho, 1995; Mallat, 1998), edge detection (Pratt, 2007; Touzi, 2002) for setting proper thresholds that depend on noise statistics.

In some practical situations, noise type and basic characteristics are known in advance. An example is radar imaging by synthetic aperture radar (SAR) with known number of looks and image forming mode (Oliver&Quegan, 2004). However, there are quite many practical situations where noise type and/or characteristics are not known in advance. Images acquired by digital cameras can serve as an example where noise properties are determined by camera settings, illumination conditions (Liu et al., 2008; Foi et al., 2007), etc. Then, noise characteristics are to be estimated for each particular image subject to further processing, for example, filtering or compression (Liu et al., 2008; Foi et al., 2007; Lukin et al., 2011). Similar situation holds for hyperspectral imaging where noise properties and signal-to-noise ratio (SNR) depend upon sub-band and they vary considerably in different component (sub-band) images (Curran&Dungan, 1989; Uss et al., 2011).

Note that below we mainly focus on considering multichannel images where the general term "multichannel" relates to color, multi- and hyperspectral imaging, dual and multi-polarization radar imaging, multitemporal sensing, where multiple images of the same scene or terrain are obtained. While for one or a few images it is sometimes possible to carry out manual (interactive) image analysis for the determination of noise type and characteristics, it becomes impossible or too labour-consuming to perform such actions for multichannel data, especially if estimation is to be done on-board or under conditions of fast

acquisition and processing of images. Then, one has to apply blind or automatic estimation (Vozel et al., 2009).

It is often assumed that the noise is i.i.d. and either pure additive or pure multiplicative (Petrovic et al., 2007; Ramponi&D'Alvise, 1999). Then, the task of estimating its parameters (variance or relative variance) simplifies since there exist quite many methods able to provide enough accurate estimation (Lukin et al., 2008 and references therein). However, the aforementioned assumption does not hold for a wide range of images formed by modern sensors. Recently, it has been clearly demonstrated (Liu et al., 2008; Foi et al., 2007, Lukin et al., 2011) that even for RGB colour images (for which the model of i.i.d. pure additive noise is still the most popular (Plataniotis&Venetsanopoulos, 2000), this model is not adequate enough. Similarly, for sub-band images of hyperspectral data, the presence of signal-dependent component in addition to additive component has been proved (Uss et al., 2011; Aiazzi et al., 2006) where the signal-dependent component occurs to be dominant for the last generation hyperspectral sensors.

This means that it is necessary to estimate the characteristics of mixed or signal-dependent noise. The corresponding methods are mainly based on forming scatter-plots of local estimates of noise variance (in scanning windows or blocks of a rather small size) on local mean and carrying out robust regression (polynomial curve fitting) into these scatter-plots. Among the methods that belong to this group, it is worth mentioning techniques described in (Liu et al., 2008; Foi et al., 2007; Aiazzi et. al., 2006; Abramov et. al., 2010). They are similar in basic steps as scatter-plot forming and the use of robust fitting curves, but differ in details. A question of how a curve is to be fit in the best or appropriate manner is not discussed in detail. Thus, the main goals of this paper are to consider different approaches to robust regression, to compare their performance, to discuss possible limitations and restrictions, and to give some practical recommendations. Although the problem of robust regression has been studied for several applications (DuMouchel&O'Brien, 1989), to our best knowledge, its use in robust estimation of signal-dependent noise characteristics has not been analyzed thoroughly. Also note that our intention is to attract attention to the problem statement with providing some initial practical solutions rather than developing deep theory.

2. Origins and properties of signal-dependent and mixed noise

By signal-dependent we mean here such a noise that its statistical characteristics (variance, probability density function (PDF)) depend upon information signal (image) in one or another manner. There are quite many known types of signal-dependent noise. Poisson noise (Foi et al., 2007) is the case for which noise variance is equal to the true value of image pixel and noise PDF shape also changes being almost Gaussian for large true values but considerably differing from Gaussian for rather small ones. Another example is film-grain noise commonly assumed locally Gaussian but with variance increasing with image true value (local mean in homogeneous image regions) (Öktem&Egiazarian, 1999). Speckle is one more example for which pure multiplicative model is widely exploited where noise is not Gaussian and its variance quickly increases for larger true values (proportionally to squared local mean in homogeneous image regions) (Oliver&Quegan, 2004; Touzi, 2002; Ramponi&D'Alvise, 1999). Thus, typically a dependence of signal dependent noise variance on true value $\sigma_{sd}^2 = f(I^{tr})$ is monotonically increasing (I^{tr} is true value). The examples

given above show that such properties of noise can originate from different sources, in the first order, from a method of image pixel value obtaining (photon counting, coherent processing of registered signals) or properties of material (carrier) used for data registration.
Mixed noise model (Astola&Kuosmanen, 1997) holds if there are several different sources. The most known example is, probably, mixed additive and impulse noise where the latter component can originate from coding/decoding errors at image transmission via communication channels. Another example is noise in modern sensors where there are such sources as dark noise, thermal noise, photon-counting noise (Kerekes&Baum, 2003). For some sources, noise is signal independent (as dark noise) and for other sources it is signal dependent (as photon-counting noise). However, in aggregate, under assumption of independent noise sources one gets mixed noise for which noise variance occurs to be dependent on image true value as, e.g., $\sigma_{sd}^2 = \sigma_{dc}^2 + kI^{tr}$ where σ_{dc}^2 is variance of dark current noise and k is proportionality factor (Foi et al., 2007). For side look aperture radar images (Lukin et al., 2007), slightly another model of dependence holds $\sigma_{sd}^2 = \sigma_{dc}^2 + k(I^{tr})^2$. Again, dependences $\sigma_{sd}^2 = f(I^{tr})$ are basically monotonously increasing.
However, this is not always true. Different nonlinear transformations of initial data that mainly belong to a wide class of homomorphic and variance stabilizing transforms are often used with special purposes. The intentions of using such transformations are various. For example, logarithmic type transforms can be applied to speckled radar images to convert pure multiplicative to pure additive noise with providing better pre-conditions for applying a wide set of denoising techniques (Oliver&Quegan, 2004; Solbo&Eltoft, 2004). Similarly, Anscombe transform is often used to convert Poisson noise to additive with practically constant variance (Anscombe, 1948). Modified Anscombe-like transforms have been designed to provide variance stabilization for mixed additive and Poisson-like noise in astronomy (Murtagh et al., 1995). Gamma correction is one more example of such nonlinear transformations exploited in digital cameras to improve visual perception of obtained images (Pratt, 2007).
While in the cases of logarithmic and Anscombe transforms it is assumed that original noise type is known and additive noise with constant variance is provided after image homomorphic transforms, the situation with the Anscombe-like transform (Murtagh et al.,1995) and gamma-correction is more complicated.
For carrying out Anscombe-like transform properly, it is needed to know or to estimate parameters or dependence $\sigma_{sd}^2 = \sigma_{dc}^2 + kI^{tr}$. Then, the task reduces to the model described above. In turn, what happens if the standard Anscombe transform is applied to an image corrupted by mixed additive and Poisson noise is considered in the paper (Lukin et al., 2009b). It is demonstrated that dependence of noise variance on local mean becomes monotonously decreasing.
Gamma correction (especially in combination with clipping effects (Foi et al., 2007) can lead to even more specific behaviour of dependence $\sigma_{sd}^2 = f(I^{tr})$. It occurs that in the area of small true values σ_{sd}^2 increases, then, for larger I^{tr}, there is an area of almost constant values of σ_{sd}^2 and finally, for I^{tr} close to an upper limit of dynamic range of image representation, σ_{sd}^2 starts to decrease (Liu et al., 2008; Lim, n.d.; Lukin et al., 2011).
Thus, the dependence $\sigma_{sd}^2 = f(I^{tr})$ is most often monotonously increasing but in special cases it can be also monotonously decreasing or having maximum. This means that there

should be some model of $\sigma^2_{sd} = f(I^{tr})$ to be fit in a scatter-plot with further estimation of its parameters. Polynomial models with a limited order seem to be a good choice although the use of other quite simple functions is also appropriate. If available, a priori information for model choosing is to be used. Another aspect is availability of methods and algorithms for robust fit of the corresponding curves.

One more question is introducing some restrictions. Since we deal with estimation of noise variance, it should be non-negative by definition. This means that a fitted function $\hat{f}(I^{tr})$ is to be non-negative for entire range of possible values of I^{tr}. Whilst for some functions as, e.g., exponents of different type the non-negativity condition is satisfied automatically, this is not true for polynomial fitting. Therefore, fitting with restrictions is required in practice and this additionally complicates the task. Although selection of regression function is important, below we do not concentrate on it and consider quite simple cases of polynomial fit. It is also worth mentioning here that there exist methods for blind identification of noise/degradation type (Vozel et al., 2006). Recall that these techniques allow identifying additive, multiplicative, impulsive noise, blur and all their possible combinations. However, for the general case of signal-dependent noise considered in this chapter this approach is often useless since noise statistics, as discussed above, can be specific. In particular, these methods are unable to identify Poisson and mixed Poisson and additive noise cases.

3. Robust regression approaches

Several times we have used above the terms "robust regression" and "robust fit" without explaining what is meant by "robust" and why conventional methods of curve fitting into scatter-plots (data) cannot be used. Robustness is treated here in two senses according to Huber (Huber, 1981). First, regression is to be robust with respect to outliers in data. The reasons why outliers appear will be explained in the next subsection. Second, by robustness we also mean the requirement to a method to provide reasonably accurate blind estimation of signal-dependent noise parameters for a wide set of images subject to analysis and a wide range of possible variation of noise statistics.

3.1 Properties of local estimates and ways of scatter-plot pre-processing

Any method for blind estimation of mixed noise parameters starts from obtaining local estimates of noise variance. For this purpose, square shape local windows (blocks) are commonly used. Blocks can be fully overlapping, partially or non-overlapping. In the latter case they are shifted with respect to each other by N pixels in horizontal and/or vertical directions where N denotes block side size. Accuracy of noise parameter estimation provided in the case of non-overlapping blocks is slightly worse (by about two times in the sense of estimated parameter variance), but processing is faster. The method parameter N also influences accuracy. Recommendations concerning its selection are given in (Lukin et al., 2008a) and will be briefly discussed below.

Suppose one has a set of blocks tessellating an analyzed image. Then, local estimates of noise variance are to be obtained for each block or for blocks selected for analysis. Consider first the case of estimate obtaining for all blocks. There are several ways to obtain local estimates. The most known is to calculate

$$\hat{\sigma}^2_l = (1/(N^2-1)) \sum_{i,j \in G_l} (I_{ij} - \overline{I}_l)^2 , \quad (1)$$

where $\overline{I}_l = (1/N)\sum_{i,j\in G_l} I_{ij}$ denotes the local mean for the l-th block, G_l is the area occupied by the l-th block, I_{ij} is the ij-th pixel of an analyzed image, l=1,…,N_{bl}, N_{bl} is the number of blocks that depends upon image and block size and a way of image tessellating by blocks. Here and below we consider one-channel images assuming that similar operations are carried out for each component image of multichannel data processed sequentially or in parallel).

There are also other algorithms for obtaining local variance estimates. In particular, robust estimators of data scale can be used for this purpose as, for example (Lukin et al., 2005)

$$\hat{\sigma}_l^2 = (1.483(\underset{ij\in G_l}{med}\left|I_{ij} - \underset{ij\in G_l}{med}(I_{ij})\right|))^2, \quad (2)$$

where *med(X)* denotes median value for data sample *X*. However, to apply the local estimate (2) one has to be sure that noise is Gaussian. Otherwise, biased estimates are obtained even in homogeneous image regions. Generally speaking, robust scale estimators (Crnojevic&Petrovic, 2010) can be used for local variance estimation if PDF of signal dependent noise is a priori known and it does not depend on local mean. However, this rarely happens in practice. For example, if signal-dependent noise is Poissonian or Poissonian noise is one component of mixed noise, then noise PDF changes with local mean (in homogeneous image regions).

Other ways to estimate local variance in blocks are possible as well. For example, estimators operating in orthogonal transform domain as, e.g., discrete cosine transform (DCT) can be used (Lukin et al., 2010b). However, these estimators produce biased estimates of noise variance even in image homogeneous regions if noise is spatially correlated (Lukin et al., 2008a, 2010b). Note that we intend on considering both the cases of i.i.d. and spatially correlated noise assuming that no or a limited a priori information is available on noise spatial correlation properties. In practice, noise is often spatially correlated, the reasons for this phenomenon are discussed in (Ponomarenko et al., 2011). To avoid problems dealing with possible biasedness of local estimates, below we focus on considering the estimation algorithm (1) as the basis of scatter-plot forming. Besides, we basically follow recommendations on block size setting given in (Lukin et al., 2008a). According to them, N≥5 for i.i.d. noise and N≥7 for spatially correlated noise. However, it is not worth using N>9 in both cases. Thus, N=7, 8, and 9 are good practical choices if a priori information on noise spatial correlation characteristics is not available or is limited.

Several times above it has been mentioned “in homogeneous image regions”. This is because just for blocks that belong to homogeneous image regions it is possible to obtain the so-called normal estimates of noise local variance. By normal we mean that such estimates are quite close to the corresponding true value keeping in mind that closeness is determined by block size, PDF of noise and its spatial correlation properties (Lukin et al., 2006, 2008a). Closeness can be characterized by variance of local variance estimates $\sigma^2_{\text{var}} = g(I^{tr})$. For a given I^{tr}, variance σ^2_{var} of normal estimates is directly proportional to $f(I^{tr})$ and it decreases if N increases. For spatially correlated noise, σ^2_{var} is larger than variance of local variance estimates for i.i.d. noise for the same I^{tr}, PDF of noise and N.

Let us demonstrate some of aforementioned properties of local estimates for a very simple test image corrupted by mixed noise (Lukin et al., 2009b). Consider the following case. The test image has the size 512x512 pixels and is composed of 16 horizontal strips each of width

32 pixels. For each strip, I^{tr} is constant and is equal to 20 (for the uppermost strip), 30, 40,…,170, i.e. for an n-th strip its mean is equal to 10+10n, n=1,…,16. The dependence of mixed noise variance is $\sigma_{sd}^2 = \sigma_{add}^2 + I^{tr}$, i.e. additive and Poison noise components are simulated where the latter component is dominant since additive noise variance σ_{add}^2 is set equal to 10 (then $\sigma_{add}^2 < I^{tr}$ for all strips). The noisy image is presented in Fig. 1,a.
The obtained scatter-plot is represented in Fig. 1,b (points in scatter-plot have coordinates $\hat{\sigma}_l^2$ for vertical axis and $\overline{I}_l$ for horizontal axis, N=7, non-overlapping blocks).

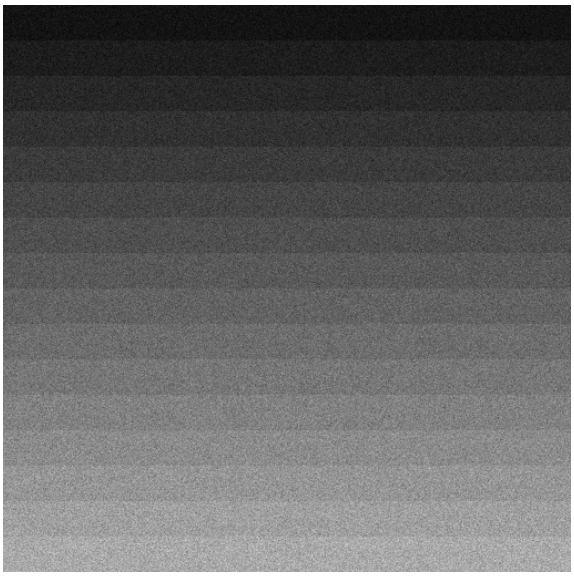


Fig. 1.a Noisy test image

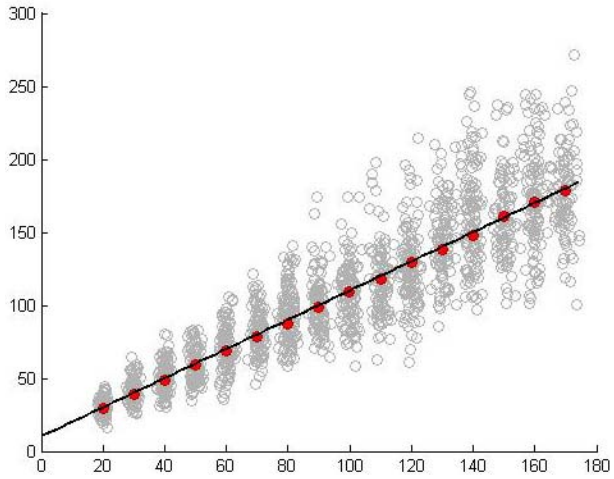


Fig. 1.b Scatter-plot of local estimates for the noisy test image in Fig. 1,a

As it is seen, the obtained points mostly concentrate along the line $\sigma_{sd}^2 = \sigma_{add}^2 + I^{tr}$ also presented at the scatter-plot in Fig. 1,b for convenience of analysis. These points, in fact, form sixteen clusters with the center coordinates approximately equal to (10+*n*, 20+*n*), *n*=1,…,16. These clusters are not of equal size. The clusters that correspond to larger *n* have larger size in both directions (especially, in vertical one) demonstrating that accuracy of estimates of local variance in blocks is worse.

The presented scatter-plot also shows that there are abnormal local estimates (outliers) of local variance placed rather far from any (including the nearest) cluster. These abnormal estimates are observed for blocks that do not fully belong to a given strip, i.e. if a given block falls into edge between strips. In a more general case, abnormal local estimates take place if a block is positioned in heterogeneous image region where by heterogeneity we mean edges, details, and textures.

Below we skip a more detailed analysis of statistics of abnormal local estimates. Only the following is worth stating. First, a percentage of abnormal local estimates basically depends upon structure (complexity) of an analyzed image and usually it is sufficiently larger than for the simple test image in Fig. 1,a (Lukin et al., 2010a). Second, according to experience of blind estimation of additive noise variance or multiplicative noise relative variance (Lukin et al., 2010a), the presence of such abnormal local estimates in any case influence final estimation even if quite robust procedures are applied to joint processing of the set of estimates. Usually, for more complex images and smaller variance of noise the provided accuracy is worse. Third, abnormal estimates give no information on noise properties and one has to rely on normal estimates. However, without special analysis (see (Lukin et al., 2010a) for details) it is difficult to predict in advance for what positions of blocks the normal or abnormal estimates will be obtained.

Intuitively, the presence of outliers should lead to worse accuracy of mixed noise parameter estimation. There are, at least, three ways to diminish the influence of abnormal estimates:

1. to apply aforementioned robust regression in curve fitting to scatter-plots;
2. to determine cluster centres and to carry out curve fitting using only them;
3. to reject introducing into a scatter-plot the estimates for blocks if they are predicted to be abnormal.

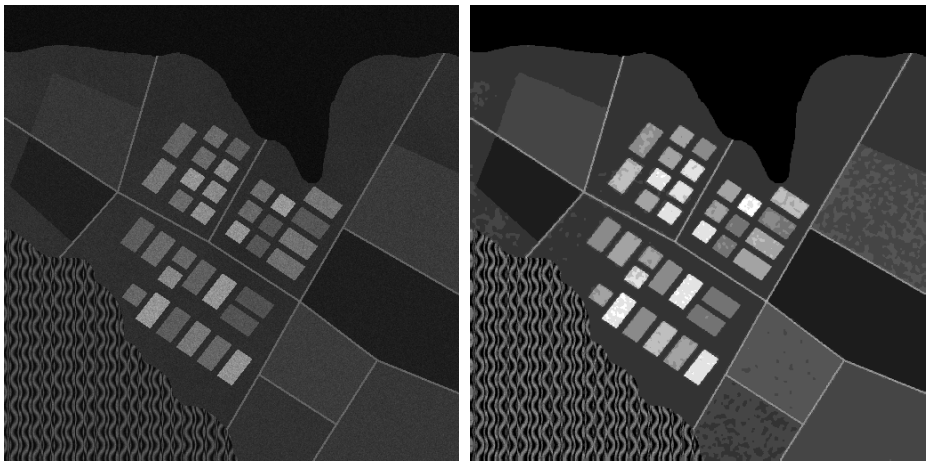
Below we concentrate on considering two former ways. For the way 2, it is possible to apply image pre-segmentation (Klaine et al., 2005). Although there are many methods for image segmentation (Yu-jin Zhang, January 2006), here one needs a method that does not exploit any a priori information on noise type and characteristics. In particular, one can use the method (Klaine et al., 2005) that allows estimating the cluster number and centers. Let us denote them as N_{cl} and $C_{cl\,m}, m = 1,...,N_{cl}$, respectively. Assume that a *k*-th block is referred to an *m*-th cluster if the corresponding central pixel of the block is referred to *m*-th cluster (level) in the segmented image. Thus, we obtain subsets of blocks for each cluster.

For each cluster, a robust method should be applied to determine its center. This method is applied to both a subset of local mean estimates and a subset of local variance estimates. Details can be found in (Lukin et al., 2008b). Thus, one gets cluster center coordinates $\hat{\sigma}_{cl\,m}^2, m = 1,...,N_{cl}$ for vertical axis and $\hat{I}_{cl\,m}, m = 1,...,N_{cl}$ for horizontal axis, respectively. $(\hat{\sigma}_{cl\,m}^2; \hat{I}_{cl\,m}^2), m = 1,...,N_{cl}$. An advantage of this approach is that after estimation of cluster centers the influence of abnormal estimates is reduced radically. The obtained estimates of

cluster centers are shown by red in Fig. 1,b and, as it is seen, they are placed close to the true curve. Then, it is possible to fit a curve (straight line) for getting the estimates $\hat{\sigma}_a^2$ and $\hat{\sigma}_\mu^2$.
Let us give an example. It is taken from the paper (Abramov et al., 2010). The noisy test image RSA is presented in Fig. 2,a. It has been corrupted by mixed additive and multiplicative Gaussian noise mimicking side-look aperture radar images:

$$\sigma_{sd}^2 = \sigma_{add}^2 + \sigma_\mu^2 (I^{tr})^2 = \sigma_{add}^2 + kX,\ X = (I^{tr})^2, \tag{3}$$

where σ_μ^2 denotes relative variance of multiplicative noise. The simulated values are $\sigma_{add}^2 = 9, \sigma_\mu^2 = 0.005$. While forming a scatter-plot, $(I^{tr})^2$ has been replaced by X for horizontal axis in order to have an opportunity to fit a first-order polynomial (such "tricks" are possible if one is confident that dependence is as (3)). The pre-segmented image is represented in Fig. 2,b.



a b

Fig. 2. Noise test image RSA (a) and the result of its pre-segmentation (b)

The obtained scatter-plot is represented in Fig. 3,a. The cluster centers are marked by red squares. For larger X, clusters are of larger size as in the previous case of signal-dependent noise with variance monotonously increasing with I^{tr}. Abnormal estimates are observed as well especially for the clusters with relatively small X. Green line shows the true dependence $\sigma_{sd}^2 = 9 + 0.005X$ whilst the red one corresponds to the curve fitted by the method (Lukin et al., 2008b) which is LMS fit using cluster centers. As seen, the curves are quite close but anyway they differ from each other.

3.2 Scatter-plot and curve fitting peculiarities for real life data

Above we have analyzed test images which are quite simple (they contain rather homogeneous image regions of rather large size). In this subsection, we consider some examples for real life data. The first example is taken from TerraSAR-X data, the image is presented in Fig. 4,a.

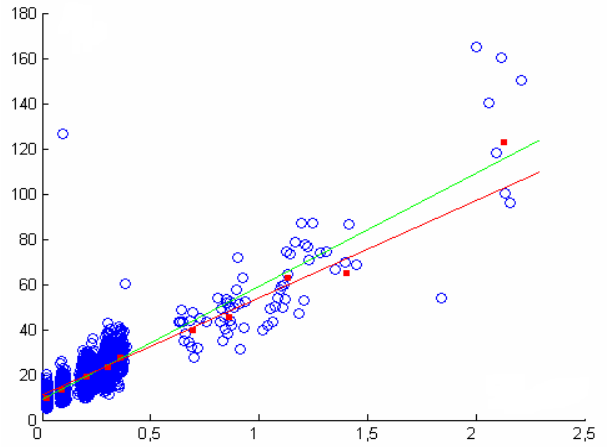
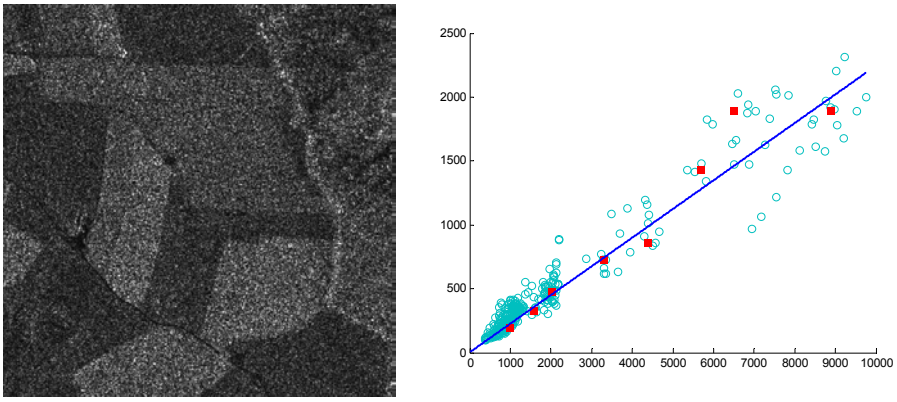


Fig. 3. a Scatter-plot of local estimates for the noisy test image in Fig. 2,a



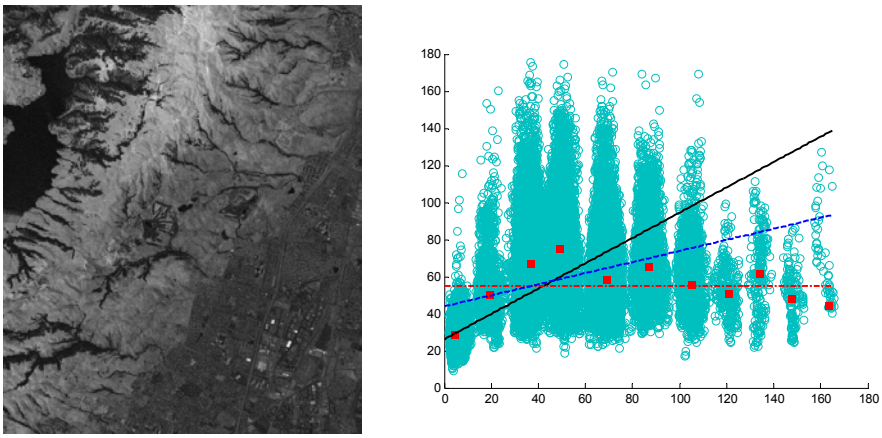
(a) (b)

Fig. 4. Real life SAR image (a) and the scatter-plot of local variance estimates after pre-segmentation (b)

This is a one-look amplitude SAR image of agricultural region in Germany (Rosenheim, http://www.infoterra.de) with fully developed speckle for which multiplicative noise is dominant and its PDF is close to Rayleigh. Since we neglect the influence of additive noise, the line that passes through the coordinate origin has been fitted. The estimated $\hat{\sigma}_{\mu}^2$ is equal to 0.22, i.e. the estimate is quite close to the value 0.273 typical for speckle with Rayleigh PDF. This example shows that, if there is reliable a priori information on noise properties, one possible practical restriction of regression when additive noise variance is supposed equal to zero.

Even more interesting results are given in Fig. 5. Fig. 5,a presents the 168-th sub-band image of hyperspectral AVIRIS data Moffett Field 1 (http://aviris.jpl.nasa.gov). Recall that it has been supposed that for hyperspectral data signal-dependent noise is characterized by the

dependence $\sigma_{sd}^2 = \sigma_{dc}^2 + kI^{tr}$. Thus, the scatter-plot of local variance estimates has been obtained (see Fig. 5,b) and regression has been carried out using three techniques. The first is Robust fit (RLMS) available in Matlab (black line). The second is standard LMS fit carried out for cluster centers (red line), and the third one is weighted LMS (WLMS) (Abramov et al., 2010) also applied to cluster centers. The estimates of additive noise variance are equal to 26.3, 54.9, and 43.9, respectively. The estimates of the parameter k are equal to 0.68, -0.00024, and 0.3, respectively. Whilst for additive noise variance estimates differ from each other but not too much, the estimates of k are very different. Moreover, for the LMS regression the obtained estimate is negative. If we suppose that the dependence is monotonically increasing, the estimate should be positive and this can be imposed as restriction to curve fitting algorithm.



(a) (b)

Fig. 5. Real life sub-band image of hyperspectral data (a) and the scatter-plot of local variance estimates with fitted lines (b)

It is interesting that for the same 168-th sub-band but for another hyperspectral data set (Lunar Lake) the estimates of additive noise variance are 27.8, 28.5 and 24.1, respectively (for the RLMS, LMS, and WLMS methods). The estimates of k are 0.07, 0.08, and 0.1. Thus, in this case, the estimates almost coincide for all three methods, and this indirectly indicates that these estimates are accurate enough. Meanwhile, these results do not give answer to a question what method is more accurate. The presented examples show only that different regression techniques can produce either quite similar or very dissimilar estimates of mixed noise parameters. Thus, numerical simulation for a set of test images of different complexity and a set of mixed noise parameters is needed to get imagination on what regression methods are better and to provide practical recommendations.

4. Numerical simulation results

To carry out numerical simulations, we have to select methods for analysis and comparisons, quantitative criteria, test images and sets of mixed noise parameters. Let us at the very beginning give brief description of regression techniques.

4.1 Considered regression techniques

In our study, we have considered the following regression techniques. The first technique is the RobustFit method realized in Matlab that is applied to scatter-plot of local estimates without their pre-processing. The second is the standard LMS fit applied to cluster centers. For this technique, robustness to outliers is provided due to robust determination of cluster centers. The third technique is weighted LMS (Abramov et al., 2010) applied to cluster centers. Its basic idea is to assign weights inversely proportional to the number of points in each cluster keeping in mind that accuracy of cluster center determination increases if the number of such points becomes larger. The fourth technique is Ransac (Fischler&Rolles, 1981) that employs lines fitting for arbitrary pairs of cluster centers, forming a confidence interval and rejecting cluster centers that occur to be out of the confidence interval. In this way, the cluster centers assumed to be determined with the worst accuracy are removed. The fifth technique is the so-called double weighted LMS (DWLMS) that has two stages. At the first stage, WLMS is applied and then, at the second stage, the cluster centers that are far from the fitted line are taken into account with smaller weights to additionally improve accuracy.

As it is seen, four the latter techniques operate with the cluster centers and this means that accuracy of cluster center determination affects the final accuracy of these techniques. To get initial imagination on the influence of cluster center estimation accuracy, let us carry out simple analysis of estimation accuracy of mixed noise parameters for the case of curve fitting for two points. It is worth mentioning here that such line can be fitted uniquely and such fitting is the main operation at the first stage of Ransac technique.

4.2 Accuracy of line fitting for two points

Suppose that we have two points (e.g., cluster centers) that have coordinates $x_1+\Delta x_1$, $y_1+\Delta y_1$ and $x_2+\Delta x_2$, $y_2+\Delta y_2$ where x_1, y_1 and x_2, y_2 denote the true values whilst Δx_1, Δy_1 and Δx_2, Δy_2 are the errors of point (cluster center) determination. Assume also that the following conditions are valid: $y_2+\Delta y_2>0, y_1+\Delta y_1>0$ and $x_2+\Delta x_2>x_1+\Delta x_1>0$. We suppose also that $\frac{\Delta y_2}{y_2}\le 1$, $\frac{\Delta y_1}{y_1}\le 1$, $\frac{\Delta x_2}{x_2}\le 1$, $\frac{\Delta x_1}{x_1}\le 1$, i.e. the errors are comparatively small (validity of these assumptions in practice will be discussed later). Then, fitted line parameters a and b are to be determined from the following simple equation system

$$\begin{cases} y_1+\Delta y_1 = a+b(x_1+\Delta x_1) \\ y_2+\Delta y_2 = a+b(x_2+\Delta x_2) \end{cases}, \tag{4}$$

Then one has

$$\hat{b}=\frac{y_2+\Delta y_2-(y_1+\Delta y_1)}{x_2+\Delta x_2-(x_1+\Delta x_1)}=\frac{y_2-y_1+(\Delta y_2-\Delta y_1)}{x_2-x_1+(\Delta x_2-\Delta x_1)}=\frac{y_2-y_1+(\Delta y_2-\Delta y_1)}{(x_2-x_1)\left(1+\frac{\Delta x_2-\Delta x_1}{x_2-x_1}\right)}, \tag{5}$$

and, under introduced assumptions on relatively small errors

$$\hat{b}\approx\frac{y_2-y_1+(\Delta y_2-\Delta y_1)}{x_2-x_1}\left(1-\frac{\Delta x_2-\Delta x_1}{x_2-x_1}\right)=\left(\frac{y_2-y_1}{x_2-x_1}+\frac{\Delta y_2-\Delta y_1}{x_2-x_1}\right)\left(1-\frac{\Delta x_2-\Delta x_1}{x_2-x_1}\right) \tag{6}$$

Let us introduce notations $\Delta x = \Delta x_2 - \Delta x_1$, $\Delta y = \Delta y_2 - \Delta y_1$. In this case

$$\begin{aligned}\hat{b} &\approx \left(b_{true} + \frac{\Delta y}{x_2 - x_1} \right)\left(1 - \frac{\Delta x}{x_2 - x_1} \right) = b_{true} - b_{true}\frac{\Delta x}{x_2 - x_1} + \frac{\Delta y}{x_2 - x_1} - \frac{\Delta x \Delta y}{x_2 - x_1}, \\ &\approx b_{true} + \frac{\Delta y - b_{true}\Delta x}{x_2 - x_1}\end{aligned} \tag{7}$$

where b_{true} is the true value of parameter *b*. Hence

$$\Delta b = \frac{\Delta y - b_{true}\Delta x}{x_2 - x_1}. \tag{8}$$

Then, assuming that $\sigma^2_{\Delta y_1}$, $\sigma^2_{\Delta y_2}$, $\sigma^2_{\Delta x_1}$, and $\sigma^2_{\Delta x_2}$ are variances of the corresponding cluster center coordinates, it is easy to obtain variance of parameter b estimate as

$$\sigma^2_{\Delta b} = \frac{\sigma^2_{\Delta y_1} + \sigma^2_{\Delta y_2} + b^2_{true}(\sigma^2_{\Delta x_1} + \sigma^2_{\Delta x_2})}{(x_2 - x_1)^2}. \tag{9}$$

This expression shows the following. Since b_{true} is usually smaller than unity and coordinates *x* are estimated with rather high accuracy, the main contribution to errors of b estimation results from Δy_1 and Δy_2. Moreover, if these errors do not have zero mean (below and in (Lukin et al., 2008a) it is demonstrated that this happens often), biased estimation of b takes place. An important conclusion is also that $\sigma^2_{\Delta b}$ is inversely proportional to $(x_2 - x_1)^2$. This is intuitively clear that more distant points "fix" a fitted line better, but this property is in no way exploited in robust regression. One more conclusion is that for smaller b_{true} variance $\sigma^2_{\Delta b}$ also decreases. Meanwhile, the ratio $\frac{\sigma^2_{\Delta b}}{b^2_{true}}$ characterizing relative error increases.

Similarly, it is possible to estimate parameter a as

$$\hat{a} = \frac{1}{2}\left(y_1 + \Delta y_1 + y_2 + \Delta y_2 - b(x_1 + \Delta x_1 + x_2 + \Delta x_2) \right). \tag{10}$$

Then one gets

$$\begin{aligned}\hat{a} &= \frac{1}{2}\left(y_1 + y_2 - b_{true}(x_1 + x_2) \right) + \frac{1}{2}\left(\Delta y_1 + \Delta y_2 - b_{true}(\Delta x_1 + \Delta x_2) - \Delta b(x_1 + \Delta x_1 + x_2 + \Delta x_2) \right) = \\ &= a_{true} + \Delta a\end{aligned} \tag{11}$$

where a_{true} is the true value of parameter a and

$$\Delta a = \frac{1}{2}\left(\Delta y_1 + \Delta y_2 - \left(b_{true}(\Delta x_1 + \Delta x_2) + \frac{\Delta y - b_{true}\Delta x}{x_2 - x_1}(x_1 + \Delta x_1 + x_2 + \Delta x_2) \right) \right). \tag{12}$$

Then variance of estimation is

$$\sigma^2_{\Delta a} = \frac{1}{4}\left(\sigma^2_{\Delta y_1} + \sigma^2_{\Delta y_2} + (\sigma^2_{\Delta y_1} + \sigma^2_{\Delta y_2})\frac{(x_1 + x_2)^2}{(x_2 - x_1)^2} + \frac{4b^2_{true}(\sigma^2_{\Delta x_1} + \sigma^2_{\Delta x_2})}{(x_2 - x_1)^2} \right). \tag{13}$$

Analysis of this expression shows the following. As can be easily predicted, $\sigma^2_{\Delta a}$ increases if variances $\sigma^2_{\Delta y_1}$, $\sigma^2_{\Delta y_2}$, $\sigma^2_{\Delta x_1}$, and $\sigma^2_{\Delta x_2}$ become larger. Thus, it is desirable to use cluster centers determined with the best accuracy. Besides, again estimation variance decreases if points are selected as distantly as possible (with larger $(x_2 - x_1)^2$ and smaller $(x_2 + x_1)^2$, i.e. if x_1 is as close to zero as possible). Finally, $\sigma^2_{\Delta a}$ is smaller if absolute values of b_{true} approach to zero (this property shows that the estimates of parameters a and b are mutually dependent). It is also easy to show that nonzero mean values of Δx_1 and Δx_2 and, especially, of Δy_1 and Δy_2 result in biased estimation.
The performed analysis shows why it is desirable to carry out weighted robust fit into pre-processed scatter-plot data. However, there are quite many ways to do this. Comparative analysis results will be presented in the next section.

4.3 Test images and accuracy criteria

The study has been performed for three test images, namely, the RSA image given in Fig. 2 and the standard test images Peppers and Goldhill, all of size 512x512 pixels. The test images Peppers and, especially, Goldhill are more complex than the RSA image. Thus, their joint and comparative analysis allows analyzing the influence of image complexity on accuracy of mixed noise parameter estimation.
Simulations, without loosing generality, have been carried out for the model for mixed additive and multiplicative noise $I_{ij} = I^{tr}_{ij} \cdot \mu_{ij} + n_{ij}$ where I^{tr}_{ij} denotes noise-free image, μ_{ij} defines multiplicative noise component obeying Gaussian distribution with unity mean and relative variance σ^2_μ, and n_{ij} describes additive noise component with zero mean Gaussian distribution with variance σ^2_a.
As a quantitative criteria, the estimation bias

$$\Delta_x = \left| \left\langle \hat{\sigma}^2_x \right\rangle - \sigma^2_x \right|, \tag{14}$$

variance

$$\theta^2_x = \left\langle \left(\hat{\sigma}^2_x - \left\langle \hat{\sigma}^2_x \right\rangle \right)^2 \right\rangle, \tag{15}$$

and aggregate error

$$\varepsilon_x = \Delta^2_x + \theta^2_x, \tag{16}$$

have been used. Sub-index x denotes belonging of the corresponding parameter for additive (a) or multiplicative (μ) noise. Notation $\langle \bullet \rangle$ means averaging by realizations. To provide statistically stable results the number of realizations was 100.

4.4 Simulation result analysis

Two sets of mixed noise parameters have been used in simulations. The first that can be called non-intensive noise is σ_a^2 = 10, σ_μ^2 = 0.005. The second case is σ_a^2 = 100, σ_μ^2 = 0.05 where for both sets multiplicative noise becomes dominant for image true values over 45.
The obtained results are given in Table 1. The optimal parameter for the RSC technique is given after /.

Let us start from considering the simplest test image RSA. As it is seen, the bias Δ_a is very small (the absolute value of the ratios Δ_a / σ_a^2 are less than 0.009 for all five considered regression techniques. Meanwhile, the values θ_a^2 considerably differ for the analyzed methods and just θ_a^2 mainly contributes to ε_a (except the results for DWLMS). The values

Image	Para-meters	Method	Δ_a	θ_a^2	ε_a	$\Delta_\mu \times 10^{-4}$	$\theta_\mu^2 \times 10^{-8}$	$\varepsilon_\mu \times 10^{-8}$
RSA	$\sigma_a^2 = 10$ $\sigma_\mu^2 = 0.005$	RLMS	0.0181	0.0080	0.0083	-1.1	0.48	1.7
		LMS	0.038	0.027	0.028	-1.5	1.0	3.0
		WLMS	0.094	0.0312	0.040	-1.8	1.0	5.0
		RSC / 7	0.0199	0.031	0.0315	-1.4	1.0	3.0
		DWLMS	-0.086	0.0069	0.014	-0.93	0.1	1.0
	$\sigma_a^2 = 100$ $\sigma_\mu^2 = 0.05$	RLMS	1.02	0.84	1.89	-23	59	590
		LMS	-1.28	2.48	4.12	-14	96	280
		WLMS	-0.93	3.64	4.50	-16	140	390
		RSC / 9	-1.27	3.84	5.45	-14	150	330
		DWLMS	-1.63	0.71	3.35	-13	38	20
Peppers	$\sigma_a^2 = 10$ $\sigma_\mu^2 = 0.005$	RLMS	15.5	2.76	243.5	-4.5	3.5	24
		LMS	19.16	1.25	368.43	-3.1	1.0	11
		WLMS	15.4	2.23	239.36	-3.5	2.0	14
		RSC / 9	20.8	39.1	469	-4.2	5.0	23
		DWLMS	12.04	0.83	145.87	-1.8	1.0	4.0
	$\sigma_a^2 = 100$ $\sigma_\mu^2 = 0.05$	RLMS	31.2	78.4	1052	-24	70	630
		LMS	18.6	83.1	429.15	-12	79	230
		WLMS	14.53	58.13	269.33	-11	49	170
		RSC / 7	13.6	46.9	232	12	40	180
		DWLMS	8.42	5.86	76.77	-8.2	22	89
Goldhill	$\sigma_a^2 = 10$ $\sigma_\mu^2 = 0.005$	RLMS	32.01	1.9	1027	-5.8	0.56	33
		LMS	27.89	1.55	779.3	-3.7	1.0	14
		WLMS	27.91	1.58	780.54	-4.8	1.0	23
		RSC / 6	28.1	8.5	801	-4.8	1.0	25
		DWLMS	18.96	0.86	360.43	-1.5	1.0	3.0
	$\sigma_a^2 = 100$ $\sigma_\mu^2 = 0.05$	RLMS	62.0	170.7	4015	-2.1	67	520
		LMS	29.74	66.11	950.8	-14	49	210
		WLMS	29.37	79.91	942.61	-13	67	240
		RSC / 8	24.0	94	671	-8.8	93	170
		DWLMS	17.42	9.95	313.46	-8.6	37	110

Table 1. Comparative results analysis for test images

ε_a and the ratios $(\varepsilon_a)^{1/2}/\sigma_a^2$ characterizing aggregate relative accuracy are small for all techniques (smaller than 0.02 that can be considered acceptable in practice).

For the multiplicative component, the bias for all methods is negative and the absolute values of the ratios $\Delta_\mu / \sigma_\mu^2$ are not larger than 0.036. The ratios $(\varepsilon_\mu)^{1/2}/\sigma_\mu^2$ do not exceed 0.048. This does not cause serious problems in practice (Abramov et al., 2004). The main contribution to ε_μ results from estimation bias although contribution of θ_μ^2 is also sufficient.

The same conclusions hold for the case $\sigma_a^2 = 100$, $\sigma_\mu^2 = 0.05$. The difference is that the absolute values of all quantitative criteria are larger. However, the ratios $(\varepsilon_a)^{1/2}/\sigma_a^2$ are still small enough (less than 0.024). The ratios $\Delta_\mu / \sigma_\mu^2$ and $(\varepsilon_\mu)^{1/2}/\sigma_\mu^2$ are of the same order. Thus, the provided accuracy is acceptable for practice (Lukin et al., 2009a).

Consider now the test image Peppers that is slightly more complex than RSA. Let us start with the case $\sigma_a^2 = 10$, $\sigma_\mu^2 = 0.005$. For additive noise component, the bias Δ_a is large and positive for all robust regression techniques. The values of the ratios Δ_a / σ_a^2 vary from 1.2 for DWLMS to about 2.1 for RSC. This shows that additive noise variance is overestimated. Estimation bias contribution to aggregate error ε_a is dominant.

On the contrary, multiplicative noise variance is underestimated (for all methods Δ_μ is negative and its contribution to ε_μ is dominant. The absolute values of the ratios $\Delta_\mu / \sigma_\mu^2$ are not larger than 0.07. The ratios $(\varepsilon_\mu)^{1/2}/\sigma_\mu^2$ do not exceed 0.1. The best results are provided by the method DWLMS.

For the case $\sigma_a^2 = 100$, $\sigma_\mu^2 = 0.05$, the values of bias Δ_a are large enough and positive for all analyzed robust regression techniques. However, the ratios Δ_a / σ_a^2 are smaller than in the previous case. Again, the influence of bias (systematic error) on ε_a is dominant. The absolute values of the ratios $\Delta_\mu / \sigma_\mu^2$ and $(\varepsilon_\mu)^{1/2}/\sigma_\mu^2$ are smaller than in the previous case. The technique DWLMS provides the best accuracy.

Finally, consider the test image Goldhill. If $\sigma_a^2 = 10$, $\sigma_\mu^2 = 0.005$, large positive valued bias is observed for estimates of additive noise variance. Multiplicative noise variance is estimated well enough, although it is slightly underestimated. Sufficient bias takes place for the estimates of additive noise variance if $\sigma_a^2 = 100$, $\sigma_\mu^2 = 0.05$. Multiplicative noise variance is again underestimated. The best accuracy is provided for the method DWLMS whilst the worst accuracy is observed for RMLS in almost all cases.

Summarizing the obtained results, it is possible to conclude the following:

1. the estimates of additive noise variance are usually biased and overestimated;
2. the estimates of multiplicative noise variance are also biased but underestimated;
3. bias contributes to aggregate error more than estimation variance, thus, its reduction is the first order task;
4. estimation accuracy is worse for more complex images; the ratios $(\varepsilon_\mu)^{1/2}/\sigma_\mu^2$ and $(\varepsilon_a)^{1/2}/\sigma_a^2$ are larger for the case of less intensive noise, thus it is more difficult to provide appropriate accuracy just for non-intensive noise situations;
5. the method DWLMS usually provides the best accuracy.

To our opinion, the main drawback of all considered techniques is overestimation (positive bias) of additive noise component. There could be three main reasons for this phenomenon. The first is self-noise in test images. The second is the influence of local image content on local variance estimates. The third is the influence of heavy tail of distributions (abnormal estimates) present in clusters. Therefore, it is worth trying to decrease this bias. However,

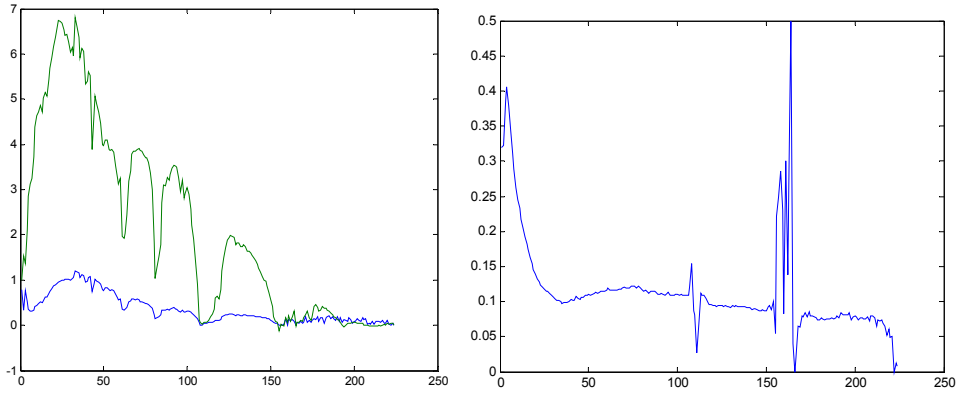
the decision to use the smallest local estimates (Liu et al., 2008) does not seem the best solution since such estimates can be by several times smaller than the true value of mixed noise variance for a given local mean.

5. Real life data testing

Testing of the methods performance has been also carried out for real life AVIRIS images for which the assumed model of mixed noise is described as $\sigma_{sd}^2 = \sigma_{add}^2 + kI^{tr}$. Thus, we have to estimate σ_{add}^2 and k . The scatter-plot example has been earlier given in Fig. 5.

We have applied the Robustfit method component-wise to two AVIRIS images, namely, Lunar Lake and Moffett Field 1. For the Moffett Field image, it has produced quite many (more than 15%) negative values of the estimates $\hat{\sigma}_{add}^2$ and a very wide range of these estimates (from -180000 to 190000). For the Lunar Lake image, no negative valued estimates $\hat{\sigma}_{add}^2$ have been obtained but the largest values have been up to 62000. Clearly, such accuracy is inappropriate since they do not agree with the estimates for other methods.

If pre-segmentation method (Klaine et al., 2005) is applied to each component image with removal of heterogeneous blocks from further consideration and then Robustfit is used, the estimates $\hat{\sigma}_{add}^2$ become in better agreement with the estimates produced by other techniques. At least, the number of negative valued estimates reduces and the limits of their variation become considerably narrower.



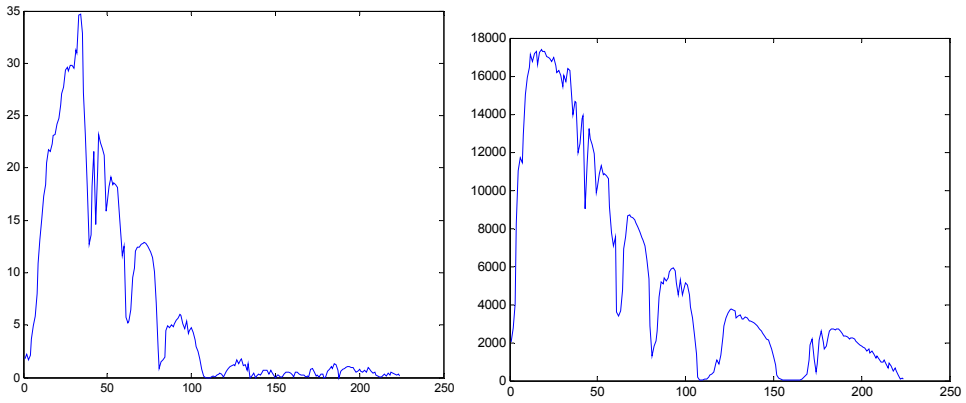
(a) (b)

Fig. 6. Estimates of *k* for Robustfit (green) and DWLMS (blue) techniques (a) and dependence of $I_{\min} / I_{\max}$ on sub-band index (b)

The estimates of k for the Lunar Lake data for the Robustfit method with pre-segmentation by the method (Klaine et al., 2005) are presented in Fig. 6,a (green color). As it is seen, mostly they are larger than unity and only for a small percentage of sub-bands, mostly with indices larger than 160, the estimates of k are smaller than unity. These results are not in agreement with results in other papers dealing with estimation of noise characteristics in AVIRIS hyperspectral images (see Uss et al., 2011, and references therein). Just overestimation of k can lead to underestimation of σ^2_{add} and negative values of its estimated mentioned earlier.

This example shows that it is worth imposing restrictions on non-negativity of both the estimates of σ^2_{add} and k. The DWLMS technique with imposed restrictions has produced quite many zero estimates of σ^2_{add}. The provided estimates of k occur to be considerably smaller (shown by blue color curve in Fig. 6,a) than for the Robustfit. Meanwhile, the shapes of these curves are very similar.

We have also analyzed dynamic range of data in sub-band images with determining the minimal and maximal values $I_{\min}$ and $I_{\max}$ in each sub-band image. The ratios $I_{\min} / I_{\max}$ for all 224 sub-bands of hyperspectral data Lunar Lake are represented in Fig. 6,b. It is seen that for most sub-bands the ratios are not zero. Thus, the histograms of sub-band image values have specific behaviour compared to most optical test images (that usually have quite many values close to zero). Then, it is difficult to expect that there are clusters that have relatively small $\hat{I}_{clm}$ and, according to analysis in subsection 4.2, the estimation accuracy reduces.



(a) (b)

Fig. 7. Estimates of *k* for DWLMS technique (a) and dependence of $I_{\max}$ on sub-band index (b)

It can be observed from Fig. 6,a that the estimates for neighbour sub-bands (close values of sub-band wavelengths) are close to each other. The estimates of k are also in correlation with sub-band image dynamic range. This is seen from analysis of plots in Fig. 7. The first plot (Fig. 7,a) shows dependence of $\hat{k}$ on sub-band index for the AVIRIS data Moffett Field 1 (the estimates have been obtained by the DWLMS method with imposed restrictions). The second plot presents dependence of $I_{\max}$ on sub-band index. It is seen that the curves have quite similar shapes.

6. Conclusions and future work

The study carried out above allows drawing a set of conclusions and giving a list of possible directions of future work.

First, estimation of mixed and signal-dependent noise parameters is only at the beginning of its development. The scatter-plot or cluster-center representations are the basis for other operations (curve regression) applied at next stages in any case. The problem of standard scatter-plot approach is that it usually contains abnormal estimates that influence estimation accuracy for any robust regression technique applied. Thus, it is desirable "to cope" with abnormal local estimates at initial stage of data processing attempting to reject them. However, it is not an easy task if parameters of mixed noise are unknown. In turn, there are also problems for cluster-center based representation. The questions that arise are: a) how to select cluster number? b) how to estimate their center positions with appropriate accuracy? c) how to predict accuracy of such estimation?

Second, the studies with simulated noise for test images have shown that even the local estimates considered normal can be considerably biased. This drawback is especially of value for additive noise component if its variance is not large. The relative bias characterized by Δ_a / σ_a^2 can be quite large. The main reason is the influence of image content. This means that one way to improve estimation accuracy is to design more accurate and robust methods for estimating local variance with diminishing the influence of image information content.

Third, the weighted methods of LMS regression using cluster centers have demonstrated their advantages. However, the potential benefits of these methods seem to be not exploited in full extent. Currently only the number of points in clusters and distances from cluster centers to initially fitted curve are used in weight adapting. It seems expedient to take other properties of clusters into account as well. One such property could be cluster size characterized in a robust manner. The positions of cluster centers and distances between them can be taken into account as well. Analysis of these aspects can be one more direction of future research.

Fourth, a priori information on mixed or signal-dependent noise is of great importance. We have carried out our experiments supposing that a model of mixed noise is a priori known and it is valid. In practice, it can be only known that noise is signal-dependent but a character (properties) of such dependence can be unknown. Then, a question arises what curve to fit? Polynomials of low order seem to be a natural choice at the first glance. However, one should keep in mind that a fitted polynomial might have intervals of negative values which, according to definition of noise local variance, is not acceptable. One simple way out is to replace negative values of a fitted polynomial by zeroes but is this the best way? Thus, the following problems and questions still remain: 1) how to impose restrictions

on fitted curves and their parameters? 2) what models of curves to apply? 3) what order of curves (polynomials) to use?

Fifth, our experiments with simulated noise have been performed only for i.i.d. noise. Spatial correlation of noise leads to several specific outcomes. In particular, statistics of local estimates changes. Certainly, this influences the performance of entire procedure of mixed noise parameters' estimation. This means that the studies for mixed noise with essential spatial correlation are to be carried out in future. Besides, if estimation is performed for hyperspectral data, considerable correlation of noise statistics in neighbour sub-bands is worth using to improve estimation accuracy.

Sixth, the goal of estimating mixed noise parameters is to use the obtained estimates at later stages of image processing. Operations used at later stages can be homomorphic transforms, edge detection, filtering, lossy compression, etc. Since any estimation is not perfect, the estimation errors influence performance of methods and algorithms applied at later stages. Degree of such negative influence is to be investigated and this will allow formulating practical requirements to accuracy of parameter estimation for mixed noise.

We also see other directions of research and studies. Whilst for particular cases of mixed and signal dependent noise there exist variance-stabilizing transforms, the general theory of such transforms is far from completeness. Note that the use of variance-stabilizing transforms simplifies applying many existing image processing methods and algorithms. In this sense, the recent studies (Foi, 2009) show perspectives and directions of future work.

7. References

Abramov S., Lukin V., Ponomarenko N., Egiazarian K., & Pogrebnyak O. (2004). Influence of multiplicative noise variance evaluation accuracy on MM-band SLAR image filtering efficiency. *Proceedings of MSMW 2004*, Vol. 1, pp. 250-252, Kharkov, Ukraine, June 2004

Abramov S., Zabrodina V., Lukin V., Vozel B., Chehdi K., & Astola J. (2010). Improved method for blind estimation of the variance of mixed noise using weighted LMS line fitting algorithm. *Proceedings of ISCAS*, pp. 2642-2645, Paris, France, June 2010

Aiazzi B., Alparone L., Barducci A., Baronti S., Marcoinni P., Pippi I., & Selva M. (2006). Noise modelling and estimation of hyperspectral data from airborne imaging spectrometers. *Annals of Geophysics*, Vol. 49, No. 1, February 2006

Al-Shaykh O.K. & Mersereau R.M. (1998). Lossy Compression of Noisy Images, *IEEE Transactions on Image Processing,* Vol. 7, No 12, (December 1998), pp. 1641-1652.

Anscombe F.J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrica*, Vol. 35, pp. 246-254

Astola J. & Kuosmanen P. (1997). *Fundamentals of nonlinear digital filtering*, CRC Press LLC, Boca Raton, USA

Bovik A. (2000). *Handbook on Image and Video Processing*. Academic Press, USA

Crnojevic V. & Petrovic N. (2010). Impulse Noise Filtering Using Robust Pixel-Wise S-estimate of Variance. *EURASIP Journal on Advances in Signal Processing*, Volume 2010, Article ID 830702, doi:10.1155/2010/830702

Curran P.J. & Dungan J.L. (1989). Estimation of signal-to-noise; a new procedure applied to AVIRIS data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 27, pp. 620-628

Donoho D.L. (1995). De-noising by Soft Thresholding. *IEEE Trans. on Information Theory,* Vol. IT-41, No 3, pp. 613-627

DuMouchel W. & O'Brien F. (1989). Integrating a Robust Option into a Multiple Regression Computing Environment in Computing Science and Statistics. *Proceedings of the 21st Symposium on the Interface*, pp. 297-301, American Statistical Association, Alexandria, VA

Elad M. (2010). *Sparse and Redundant Representations. From Theory to Applications in Signal and Image Processing*, Springer Science+Business Media, LLC

Fischler M.A. & Rolles R.C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, Vol. 24, No. 6, pp. 381-395

Foi A., Trimeche M., Katkovnik V., & Egiazarian K. (2007). Practical Poissonian-Gaussian Noise Modeling and Fitting for Single Image Raw Data. *IEEE Transactions on Image Processing*, Vol. 17, No. 10, pp. 1737-1754

Foi A. (2009). Clipped Noisy Images: Heteroskedastic Modeling and Practical Denoising. *Signal Processing*, Vol. 89, No. 12, pp. 2609-2629

http://aviris.jpl.nasa.gov

Huber P. (1981). *Robust statistics*, Wiley, New York

Kerekes J.P. & Baum J.E. (2003). Hyperspectral Imaging System Modeling. *Lincoln Laboratory Journal*, Vol. 14, No. 1, pp. 117-130

Klaine L., Vozel B., & Chehdi K. (2005). Unsupervised Variational Classification Through Image Multi-Thresholding. *Proceedings of the 13th EUSIPCO Conference*, Antalya, Turkey

Lim S.H. (2006). Characterization of Noise in Digital Photographs for Image Processing. *Proceedings of Digital Photography II*, SPIE 6069, DOI: 10.1117/12.655915

Liu C., Szeliski R., Kang S.B., Zitnick C.L., & Freeman W.T. (2008). Automatic estimation and removal of noise from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No 2, pp. 299-314

Lukin V., Koivisto P., Ponomarenko N., Abramov S., & Astola J. (2005). Two-stage Methods for Mixed Noise Removal. *CD-ROM Proceedings of EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP)*, Japan, May 2005

Lukin V., Abramov S., Ponomarenko N., Vozel B., & Chehdi K. (2006). Methods for blind evaluation of noise variance in multichannel optical and radar images. *Telecommunications and Radioengineering*, Vol. 65 (6), pp. 509-537

Lukin V., Ponomarenko N., Abramov S., Vozel B., & Chehdi K. (2007). Improved noise parameter estimation and filtering of MM-band SLAR images. *Proceedings of MSMW 2007*, Vol. 1, pp. 439-441, Kharkov, Ukraine

Lukin V., Abramov S., Vozel B., Chehdi K., & Astola J. (2008a). Segmentation-based method for blind evaluation of noise variance in images. *SPIE Journal on Applied Remote Sensing*, Vol. 2, Aug. 2008, open access paper

Lukin V., Ponomarenko N., Abramov S., Vozel B., Chehdi K., & Astola J. (2008b). Filtering of radar images based on blind evaluation of noise characteristics. *Proceedings Image and Signal Processing for Remote Sensing XIV*, Cardiff, UK, September 2008, SPIE Vol. 7109

Lukin V., Abramov S., Ponomarenko N., Uss M., Vozel B., Chehdi K., & Astola J. (2009a). Processing of images based on blind evaluation of noise type and characteristics. *Proceedings of SPIE Symposium on Remote Sensing*, Vol. 7477, Berlin, Germany, September 2009

Lukin V., Krivenko S., Zriakhov M., Ponomarenko N., Abramov S., Kaarna A., & Egiazarian K. (2009b). Lossy compression of images corrupted by mixed Poisson and additive noise. *Proceedings of LNLA*, pp. 33-40, Helsinki, August 2009

Lukin V., Abramov S., Uss M., Vozel B., & Chehdi K. (2010a). Performance Analysis of Segmentation-based Method for Blind Evaluation of Additive Noise in Images. *Proceedings of MSMW*, Kharkov, Ukraine, June 2010

Lukin V., Fevralev D., Ponomarenko N., Abramov S., Pogrebnyak O., Egiazarian K., & Astola J. (2010b). Discrete cosine transform-based local adaptive filtering of images corrupted by nonstationary noise. *Electronic Imaging Journal*, 19(2), 1, April-June 2010

Lukin V., Abramov S., Ponomarenko N., Uss M., Zriakhov M., Vozel B., Chehdi K., & Astola J. (2011). Methods and Automatic Procedures for Processing Images Based on Blind Evaluation of Noise Type and Characteristics. *SPIE Journal on Advances in Remote Sensing*, DOI: 10.1117/1.3539768

Mallat S. (1998). *A Wavelet tour of signal processing,* Academic Press, San Diego

Murtagh F., Starck J.L., & Bijaoui A. (1995). Image restoration with noise suppression using a multiresolution support, *Astron. Astrophys. Suppl. Ser*, 112, pp. 179-189

Öktem R. & Egiazarian K. (1999). Transform Domain Algorithm for Reducing the Effect of Film-Grain Noise in Image Compression. *Electronic Letters*, Vol. 35, No. 21, pp. 1830-1831

Oliver C. & Quegan S. (2004). *Understanding Synthetic Aperture Radar Images*, SciTech Publishing

Petrovic N., Zlokolica V., Goossens B., Pizurica A., & Philips W. (2007). Characterization of correlated noise in video sequences and its applications to noise removal. *Proceedings of the 3rd International Workshop on Video Processing and Quality Metrics for Consumer Eletronics*, Scottsdale, Arizona, USA, January 2007

Plataniotis K.N. & Venetsanopoulos A.N. (2000). *Color Image Processing and Applications*, Springer-Verlag, NY

Ponomarenko N., Lukin V., Egiazarian K., & Lepisto L. (2011). Color image lossy compression based on blind evaluation and prediction of noise characteristics. Accepted to *SPIE Conference Image Processing: Algorithms and Systems VII*, Vol. 7870, 12 p., San Francisco, USA

Pratt W.K. (2007). *Digital Image Processing* (Fourth Edition). Wiley-Interscience, NY, USA

Ramponi G. & D'Alvise R. (1999). Automatic Estimation of the Noise Variance in SAR Images for Use in Speckle Filtering, *Proceedings of EEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Vol. 2, pp. 835-838, Antalya, Turkey

Sendur L. & Selesnick I.W. (2002). Bivariate shrinkage with local variance estimation. *IEEE Signal Processing Letters*, Vol. 9, No. 12, pp. 438-441

Solbo S. & Eltoft T. (2004). Homomorphic Wavelet-based Statistical Despeckling of SAR Images. *IEEE Trans. on Geoscience and Remote Sensing*, Vol. GRS-42, No. 4, pp. 711-721

Touzi R. (2002). A Review of Speckle Filtering in the Context of Estimation Theory. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 11, pp. 2392-2404

Uss M., Vozel B., Lukin V., & Chehdi K. (2011). Local Signal-Dependent Noise Variance Estimation from Hyperspectral Textural Images. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 2, DOI: 10.1109/JSTSP.2010.2104312

Vozel B., Chehdi K., & Klaine L. (2006). Noise identification and estimation of its statistical parameters by using unsupervised variational classification. *Proceedings of ICASSP*, Vol. II, pp. 841-844

Vozel B., Abramov S., Chehdi K., Lukin V., Ponomarenko N., Uss M., & Astola J. (2009). Blind methods for noise evaluation in multi-component images, In: *Multivariate Image Processing,* pp. 263-295, France

Yu-jin Zhang (January 2006). *Advances in Image and Video Segmentation*, IRM Press, ISBN 1591407559

4

A Semi-Analytical Finite Element Approach in Machine Design of Axisymmetric Structures

Denis Benasciutti, Francesco De Bona and Mircea Gh. Munteanu
University of Udine
Italy

1. Introduction

In the last decades, finite element (FE) method has proven to be an efficient tool for the numerical analysis of two- or three-dimensional structures of whatever complexity, in mechanical, thermal or other physical problems. It is widely recognised that computational cost (as time and computer capability) increases greatly with structure complexity, being larger with three-dimensional analyses than with two-dimensional ones.
It is therefore desirable to devise simplified approaches that may provide a reduction in overall computational effort. An example of considerable importance is the study of bodies of revolution (axisymmetric structures) under axisymmetric loading, where a three-dimensional problem is solved by a two-dimensional analysis. Examples are vessels under internal pressure, rotating disks, foundation piles (Zienkiewicz & Taylor, 2000).
Apart from axisymmetric problems solved by a plane model, a full three-dimensional analysis is needed, in principle, whenever the structure is axisymmetric but the load is not. In such situations, often encountered in many engineering applications, it is desirable to search for simplified approaches, which may still replace (and thus avoid the computation effort needed by) the use of full three-dimensional simulations.
A particular sub-class of problems is encountered when the load applied to axisymmetric structures is exactly antisymmetric; an example is a shaft under a torsion load, which, as it will be shown, can be solved by a plane FE approach, which greatly simplifies the analysis.
Another example is represented by semi-analytical methods, which have been developed more than fifty years ago for FE analysis of axisymmetric structures loaded non-axisymmetrically (Wilson, 1965). Such methods use a Fourier series expansion to reduce a three-dimensional problem to a two-dimensional harmonic model and to compute the solution as superposition of results of every harmonic component analysis. At present, this approach is still not well established and it is rarely used in mechanical design. Even commercial FE codes including harmonic elements have found limited application, due to practical difficulties related to Fourier series conversion of external loads. Only few applications of semi-analytical methods to engineering practical cases have been reported in literature, see (Genta & Tonoli, 1996; Lai & Booker, 1991; Kim et al., 1994; Pedersen & Laursen, 1982; Taiebait & Carter, 2001; Thomas et al., 1983; Zienkiewicz & Taylor, 2000).
The goal of this work is twofold. First, it aims to provide a theoretical background on the use of semi-analytical FE approach in numerical analysis of axisymmetric structures loaded non-axialsymmetrically. In particular, two original results are developed: a plane "axi-

antisymmetric" FE model for solving axisymmetric components loaded in torsion, a semi-analytical approach for the analysis of plane axisymmetric bodies under non-axisymmetric thermal loadings. The second intent of the work is to clarify some practical aspects in the application of semi-analytical method to engineering problems. Two illustrative examples are then discussed: an axisymmetric component with shoulder fillet under axial, bending and torsion mechanical loading, followed by the more general case of a shaft; a simplified numerical approach for estimating the transient temperature field in a rotating cylinder under thermal loadings. The presented examples confirm how the proposed approach gives a high accuracy in results, with also a significant reduction in computational time, compared to classical FE analyses.

2. Theoretical background

A three-dimensional structure or solid is defined as "axisymmetric" if its geometry, material properties and boundary conditions are independent of an azimuth coordinate θ of a cylindrical reference frame (r,θ,z), where z is the component axis of symmetry and r is the radial distance from z-axis, see Fig. 1.

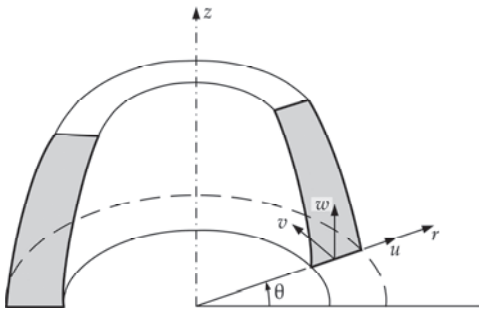


Fig. 1. An axisymmetric solid: cylindrical reference system and displacement components

Depending on the configuration of external loads, different types of analysis can be identified. For example, if also external loads are themselves axisymmetric with respect to same z-axis, the analysis is axisymmetric and mathematically two-dimensional, that is results are independent of θ and they are only function of r, z coordinates (Bathe, 1996; Cook et al., 1989; Zienkiewicz & Taylor, 2000). Examples are disks rotating at uniform speed under centrifugal forces or cylindrical vessels under internal pressure.

Another situation occurs for an axisymmetric structure under an "axi-antisymmetric" loading (this term will be clarified later on), i.e. a loading which is axial anti-symmetric with respect to z-axis and also independent of θ angle. An example is a cylindrical body under a torsion loading. As it will be shown, the analysis becomes really one-dimensional and results are only function of r and z.

A third situation, of great practical interest, occurs when a structure is axisymmetric but the loading is not, so that the analysis becomes really three-dimensional. A great simplification can be achieved by using a so-called semi-analytical approach, which adopts a harmonic model based on Fourier series method.

The next sections will address the main aspects of FE theory for the case of two- and three-dimensional axisymmetric structures under different types of both mechanical and thermal

loadings. For a mechanical analysis, the theory for axisymmetric, axi-antisymmetric and non-axisymmetric mechanical loads will be developed. For a thermal analysis, a one-dimensional finite element for dealing with axisymmetric plane body under non-axisymmetric thermal loadings will be next developed; the steady-state and transient case will be considered.

3. Mechanical analysis

This section is concerned with FE theory for mechanical analysis of axisymmetric structures under, respectively, axisymmetric, axi-antisymmetric and non-axisymmetric loads.

3.1 Plane axisymmetric finite element

Although the theory of axisymmetric mechanical analysis is well known (Cook et al., 1989; Wilson, 1965; Zienkiewicz & Taylor, 2000), it will be shortly reviewed to introduce several equations that will be used in the following sections. In the case of axisymmetric structures loaded by axially symmetric loads, by symmetry, the two displacement components u and w in any plane section of the body along its axis of symmetry completely define the state of strain and, accordingly, the state of stress (Zienkiewicz & Taylor, 2000). Thus, the circumferential (hoop) displacement v, the tangential stress components $\tau_{r\theta}$ and $\tau_{\theta z}$ and their corresponding shear strains $\gamma_{r\theta}$ and $\gamma_{\theta z}$ must be zero, see Fig. 2. The analysis then reduces to a plane FE model, characterized by only radial $u(r,z)$ and axial $w(r,z)$ displacements, where r and z denote the radial and axial coordinates of a point within the structure. For a M-nodes finite element, the vector of displacement field in the cylindrical reference system (r,θ,z) is:

$$\begin{Bmatrix} u(r,z) \\ w(r,z) \end{Bmatrix} = \begin{bmatrix} N_1 & 0 & \cdots & N_M & 0 \\ 0 & N_1 & \cdots & 0 & N_M \end{bmatrix} \begin{Bmatrix} u_1 \\ w_1 \\ \vdots \\ u_M \\ w_M \end{Bmatrix} = \sum_{i=1}^{M} \left(\begin{bmatrix} N_i & 0 \\ 0 & N_i \end{bmatrix} \begin{Bmatrix} u_i \\ w_i \end{Bmatrix} \right) \tag{1}$$

where u_i, w_i are the nodal displacements and $N_i(r,z)$ is the shape function for node i.
As already mentioned, four non-zero strain components have to be considered in an axisymmetric deformation, see Fig. 2; the strain vector in polar coordinates thus is:

$$\{\varepsilon\} = \begin{Bmatrix} \varepsilon_r \\ \varepsilon_z \\ \gamma_{rz} \\ \varepsilon_\theta \end{Bmatrix} = \begin{Bmatrix} \dfrac{\partial u}{\partial r} \\ \dfrac{\partial w}{\partial z} \\ \dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial r} \\ \dfrac{u}{r} \end{Bmatrix} = \sum_{i=1}^{M} \begin{bmatrix} B_{i\ pl\_st} \\ \dfrac{N_i}{r} \quad 0 \end{bmatrix} \begin{Bmatrix} u_i \\ w_i \end{Bmatrix} = [B] \begin{Bmatrix} u_1 \\ w_1 \\ \vdots \\ u_M \\ w_M \end{Bmatrix} \tag{2}$$

where submatrix $[B_{i,pl\_st}]$ corresponds to that of plane case and its dimension is 3x2. For example, as it is well known for 3-nodes triangular element it is:

$$\left[B_{pl\_st}\right]=\left[B_1\ \ B_2\ \ B_3\right]=\frac{1}{2A}\begin{bmatrix} z_3-z_2 & 0 & z_1-z_3 & 0 & z_2-z_1 & 0 \\ 0 & r_2-r_3 & 0 & r_3-r_1 & 0 & r_1-r_2 \\ r_2-r_3 & w_3-w_2 & r_3-r_1 & w_1-w_3 & r_1-r_2 & w_2-w_1 \end{bmatrix} \tag{3}$$

where $A=(r_1z_3+r_2z_1+r_3z_2-r_1z_2-r_2z_3-r_3z_1)/2$ is the element area. As matrix $[B]$ depends on r and z coordinates, the strains are no longer constant within an element, as in plane stress or plane strain (the strain variation is due to the ε_θ term).

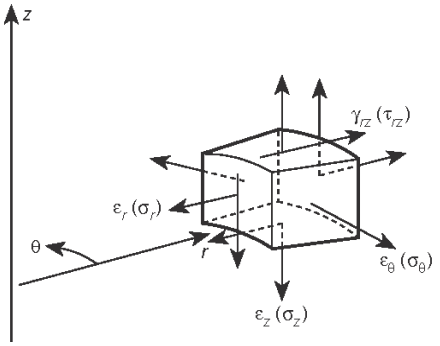


Fig. 2. Strains and stresses involved in an axisymmetric analysis

The element stiffness matrix is calculated as an integral over the element volume V_{el}, which for axial symmetry coincides with the whole ring of material:

$$[k]=\int_{V_{el}}[B]^T[D][B]d(vol)=\int_{A_{el}}\int_{-\pi}^{\pi}[B]^T[D][B]r\,\mathrm{d}\theta\,\mathrm{d}A \tag{4}$$

where A_{el} is the cross-sectional area of the element on a plane section. The elasticity or Hookean matrix $[D]$ in Eq. (4), which links the vectors of strains and stresses, in the hypothesis of isotropic material has the following form:

$$[D]=\frac{E(1-\nu)}{(1+\nu)(1-2\nu)}\begin{bmatrix} 1 & \nu/(1-\nu) & 0 & \nu/(1-\nu) \\ \nu/(1-\nu) & 1 & 0 & \nu/(1-\nu) \\ 0 & 0 & (1-2\nu)/2(1-\nu) & 0 \\ \nu/(1-\nu) & \nu/(1-\nu) & 0 & 1 \end{bmatrix} \tag{5}$$

where E is the modulus of elasticity (Young's modulus) and v is the Poisson's ratio.

3.2 Plane axi-antisymmetric finite element

An interesting application is represented by the study of axisymmetric structures subjected to axi-antisymmetric loadings (this term refers to the well known cases of symmetry and anti-symmetry). An example is a shaft of variable diameter under a torsion load applied at the ends (Timoshenko & Goodier, 1951). In this configuration, load is antisymmetric with respect to each plane passing across structure z-axis (this plane then behaves as a plane of axis-antisymmetry) and it is also independent of θ angle.

In literature, to solve the problem of an axisymmetric body under torsion by using a plane FE approach it is suggested using a variational approach, in which the solution is a stress function φ which minimises the functional (Zienkiewicz & Cheung, 1965, 1967):

$$\chi = \frac{1}{2}\iint \left\{ \left(\frac{1}{r^3}\frac{\partial \varphi}{\partial r} \right)^2 + \left(\frac{1}{r^3}\frac{\partial \varphi}{\partial z} \right)^2 \right\} dr\, dz \tag{6}$$

Instead, this work will show that it is possible to develop a simple FE theory applicable to an axisymmetric structure under torsion, by analogy with the theory of axisymmetric loading previously reviewed. In fact, in this configuration each element node has only one degree of freedom (the hoop displacement v), while radial and axial displacements u and w (warping), as well as normal stresses σ_r, σ_θ, σ_z, shear stress τ_{rz} and their related strain components vanish. By symmetry, the hoop displacement does not depend on angle θ and only two non-null strains $\gamma_{r\theta}$ and $\gamma_{\theta z}$ are present. By analogy with Eq. (1), the displacement of a point within a M-nodes element is:

$$v(r,z) = \begin{bmatrix} N_1(r,z) & \cdots & N_M(r,z) \end{bmatrix} \begin{Bmatrix} v_1 \\ \vdots \\ v_M \end{Bmatrix} = \sum_{i=1}^{M} N_i(r,z) v_i \tag{7}$$

where v_i is the nodal displacement and $N_i(r,z)$ is the shape function of node i. The related strain vector is:

$$\{\varepsilon\} = \begin{Bmatrix} \gamma_{r\theta} \\ \gamma_{\theta z} \end{Bmatrix} = \begin{Bmatrix} \dfrac{\partial v}{\partial r} - \dfrac{v}{r} \\ \dfrac{\partial v}{\partial z} \end{Bmatrix} = [B] \begin{Bmatrix} v_1 \\ \vdots \\ v_M \end{Bmatrix} \tag{8}$$

where $[B]$ is the strain-displacement matrix, with dimension 1xM. As an example, for a 3-nodes triangular finite element, matrix $[B]$ has the following expression:

$$[B] = \frac{1}{2A} \begin{bmatrix} z_3 - z_2 & z_1 - z_3 & z_2 - z_1 \\ r_2 - r_3 & r_3 - r_1 & r_1 - r_2 \end{bmatrix} - \frac{1}{r} \begin{bmatrix} N_1 & N_2 & N_3 \\ 0 & 0 & 0 \end{bmatrix}$$

$$N_1 = \left((z_3 - z_2) r + (r_2 - r_3) z + r_3 z_2 - r_2 z_3 \right) / 2A \text{ etc.} \tag{9}$$

The expression of the element stiffness matrix and of equivalent nodal loads can thus be easily evaluated using Eq. (4), with the following elasticity matrix:

$$[D] = \frac{E}{2(1+\nu)} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{10}$$

It is worth mentioning that this approach seems not to be implemented in commercial FE codes. Nevertheless, it can be implemented as a particular case of a harmonic finite element, discussed in next section.

3.3 Harmonic finite element

A third type of problem, of more practical interest, is when the structure is axially symmetric but the loading is not, so that the analysis is really three-dimensional. A great simplification can be obtained by using a semi-analytical approach, based on a harmonic FE model and Fourier series expansion of loads. As it will be shown, it can be demonstrated that, in linear analysis, a harmonic load produces a harmonic response in term of stress and displacements. The solution is then obtained by superimposing results of each harmonic, which are totally uncoupled (Cook et al., 1989; Zienkiewicz & Taylor, 2000).

To start with, the nodal loads applied to the structure can be expanded in Fourier series as:

$$\begin{Bmatrix} R(r,\theta,z) \\ T(r,\theta,z) \\ Z(r,\theta,z) \end{Bmatrix} = \begin{Bmatrix} \overline{R}_0 + \sum_{n=1}^{\infty}\left(\overline{R}_n(r,z)\cos n\theta + \overline{\overline{R}}_n(r,z)\sin n\theta\right) \\ \overline{\overline{T}}_0 + \sum_{n=1}^{\infty}\left(\overline{T}_n(r,z)\sin n\theta - \overline{\overline{T}}_n(r,z)\cos n\theta\right) \\ \overline{Z}_0 + \sum_{n=1}^{\infty}\left(\overline{Z}_n(r,z)\cos n\theta + \overline{\overline{Z}}_n(r,z)\sin n\theta\right) \end{Bmatrix} \tag{11}$$

where symbols R, T and Z indicate respectively the radial, hoop and axial load components. A similar series expansion holds also for body forces, boundary conditions, initial strains, etc. In Eq. (11), all barred quantities are amplitudes, which are functions of r, z but not of θ. Single-barred amplitudes represent symmetric load components (loads which have θ=0 as a plane of symmetry), while double-barred amplitudes represent antisymmetric load terms. The sine expansion in T load is necessary to assure symmetry, as the direction of T has to change for $\theta>\pi$. The constant terms R_0 and Z_0 permit axisymmetric load condition to be described, while the term T_0 refers to the axi-antisymmetric load. These three terms are grouped into a single vector representing the constant term of the Fourier series.

It is possible to demonstrate (Cook et al., 1989) that in a linear analysis, when loads are expanded as in Eq. (11), displacement components are described by Fourier series as well:

$$\begin{cases} u(r,\theta,z) = \sum_{n=0}^{\infty}\overline{u}_{\mathrm{n}}\cos n\theta + \sum_{n=1}^{\infty}\overline{\overline{u}}_{\mathrm{n}}\sin n\theta \\ v(r,\theta,z) = \sum_{n=0}^{\infty}\overline{v}_{\mathrm{n}}\sin n\theta - \sum_{n=1}^{\infty}\overline{\overline{v}}_{\mathrm{n}}\cos n\theta \\ w(r,\theta,z) = \sum_{n=0}^{\infty}\overline{w}_{\mathrm{n}}\cos n\theta + \sum_{n=1}^{\infty}\overline{\overline{w}}_{\mathrm{n}}\sin n\theta \end{cases} \tag{12}$$

All three displacements are needed because the physical problem is three-dimensional. The motivation of the arbitrarily chosen negative sign in the v series is that it greatly simplifies the computation of the element stiffness matrix, as it will be explained later on. As for the loads, the single- and double-barred terms refer to symmetric and antisymmetric components.

A Fourier series expansion similar to Eq. (12) can be equally used also for the nodal displacements of a finite element. Within a finite element, one can thus interpolate the amplitudes $\overline{u}_{\mathrm{n}}$, $\overline{\overline{u}}_{\mathrm{n}}$, $\overline{v}_{\mathrm{n}}$, $\overline{\overline{v}}_{\mathrm{n}}$, etc. of the displacement components in Eq. (12) from the corresponding nodal amplitudes ($\overline{u}_{\mathrm{in}}$, $\overline{\overline{u}}_{\mathrm{in}}$, $\overline{v}_{\mathrm{in}}$, $\overline{\overline{v}}_{\mathrm{in}}$, $\overline{w}_{\mathrm{in}}$, $\overline{\overline{w}}_{\mathrm{in}}$), where subscript *in* specifies that amplitude refers to node *i* and harmonic *n*. Therefore, the vector of displacement field within the element can be described in the following form:

$$\begin{Bmatrix} u(r,\theta,z) \\ v(r,\theta,z) \\ w(r,\theta,z) \end{Bmatrix} = \sum_{n=0}^{\infty}\sum_{i=1}^{M} N_{\mathrm{i}}(r,z) \begin{bmatrix} \cos n\theta & 0 & 0 \\ 0 & \sin n\theta & 0 \\ 0 & 0 & \cos n\theta \end{bmatrix} \begin{Bmatrix} \overline{u}_{\mathrm{in}} \\ \overline{v}_{\mathrm{in}} \\ \overline{w}_{\mathrm{in}} \end{Bmatrix} +$$
$$+ \sum_{n=0}^{\infty}\sum_{i=1}^{M} N_{\mathrm{i}}(r,z) \begin{bmatrix} \sin n\theta & 0 & 0 \\ 0 & -\cos n\theta & 0 \\ 0 & 0 & \sin n\theta \end{bmatrix} \begin{Bmatrix} \overline{\overline{u}}_{\mathrm{in}} \\ \overline{\overline{v}}_{\mathrm{in}} \\ \overline{\overline{w}}_{\mathrm{in}} \end{Bmatrix} = \sum_{n=0}^{\infty} \left(\left[\overline{N} \right]_{\mathrm{n}} \left\{ \overline{u}_{\mathrm{n}} \right\} + \left[\overline{\overline{N}} \right]_{n} \left\{ \overline{\overline{u}}_{\mathrm{n}} \right\} \right) \tag{13}$$

where $\left[\overline{N} \right]_{\mathrm{n}}$ and $\left[\overline{\overline{N}} \right]_{\mathrm{n}}$ are 1x*M* arrays of 3x3 submatrices (*M* is the number of element nodes); note that such matrices both depend on *n* because of the (cos*nθ*) and (sin*nθ*) terms. The strain vector in cylindrical coordinate is given by (Cook et al., 1989):

$$\{\varepsilon\} = \begin{Bmatrix} \varepsilon_r \\ \varepsilon_\theta \\ \varepsilon_z \\ \gamma_{r\theta} \\ \gamma_{\theta z} \\ \gamma_{rz} \end{Bmatrix} = \begin{Bmatrix} \dfrac{\partial u}{\partial r} \\ \dfrac{u}{r} + \dfrac{1}{r}\dfrac{\partial v}{\partial \theta} \\ \dfrac{\partial w}{\partial z} \\ \dfrac{1}{r}\dfrac{\partial u}{\partial \theta} + \dfrac{\partial v}{\partial r} - \dfrac{v}{r} \\ \dfrac{1}{r}\dfrac{\partial w}{\partial \theta} + \dfrac{\partial v}{\partial z} \\ \dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial r} \end{Bmatrix} = [\partial] \begin{Bmatrix} u \\ v \\ w \end{Bmatrix} = \sum_{n=0}^{\infty} \left(\left[\overline{B} \right]_{\mathrm{n}} \left\{ \overline{u}_{\mathrm{n}} \right\} + \left[\overline{\overline{B}} \right]_{\mathrm{n}} \left\{ \overline{\overline{u}}_{\mathrm{n}} \right\} \right) \tag{14}$$

where $[\partial]$ is a differential operator matrix, with dimension 6x3. Therefore, also strains are expanded in Fourier series and the contribution of *n*-th harmonic thus is $\{\varepsilon\}_{\mathrm{n}} = \left[\overline{B} \right]_{\mathrm{n}} \left\{ \overline{u}_{\mathrm{n}} \right\} + \left[\overline{\overline{B}} \right]_{\mathrm{n}} \left\{ \overline{\overline{u}}_{\mathrm{n}} \right\}$. Equation (14) defines, for harmonic *n*, the strain-displacement matrices $\left[\overline{B} \right]_{\mathrm{n}} = \left[\overline{B}_{1\mathrm{n}} \quad \overline{B}_{2\mathrm{n}} \quad \ldots \quad \overline{B}_{\mathrm{Mn}} \right]$ and $\left[\overline{\overline{B}} \right]_{\mathrm{n}} = \left[\overline{\overline{B}}_{1\mathrm{n}} \quad \overline{\overline{B}}_{2\mathrm{n}} \quad \ldots \quad \overline{\overline{B}}_{\mathrm{Mn}} \right]$, which are 1x*M* arrays of the 6x3 submatrices $\left[\overline{B}_{\mathrm{in}} \right] = [\partial] \left[\overline{N}_{\mathrm{in}} \right]$ and $\left[\overline{\overline{B}}_{\mathrm{in}} \right] = [\partial] \left[\overline{\overline{N}}_{\mathrm{in}} \right]$. For the *i*-th node, one submatrix is:

$$\left[\overline{B}_{\mathrm{in}}\right]=\begin{bmatrix} \dfrac{\partial N_{\mathrm{i}}}{\partial r}\cos n\theta & 0 & 0 \\ \dfrac{N_{\mathrm{i}}}{r}\cos n\theta & n\dfrac{N_{\mathrm{i}}}{r}\cos n\theta & 0 \\ 0 & 0 & \dfrac{\partial N_{\mathrm{i}}}{\partial z}\cos n\theta \\ -\dfrac{nN_{\mathrm{i}}}{r}\sin n\theta & \left(\dfrac{\partial N_{\mathrm{i}}}{\partial r}-\dfrac{N_{\mathrm{i}}}{r}\right)\sin n\theta & 0 \\ 0 & \dfrac{\partial N_{\mathrm{i}}}{\partial z}\sin n\theta & -\dfrac{nN_{\mathrm{i}}}{r}\sin n\theta \\ \dfrac{\partial N_{\mathrm{i}}}{\partial z}\cos n\theta & 0 & \dfrac{\partial N_{\mathrm{i}}}{\partial r}\cos n\theta \end{bmatrix} \tag{15}$$

while for $\left[\overline{\overline{B}}_{\mathrm{in}}\right]$ one finds that (sin$n\theta$) and (cos$n\theta$) are interchanged and, in addition, there is an algebraic sign change in the third and fourth row, that is to say that $\left[\overline{\overline{B}}_{\mathrm{in}}\right]$ can be obtained from that of $\left[\overline{B}_{\mathrm{in}}\right]$ by simply substituting (–sin$n\theta$) with (cos$n\theta$) and (cos$n\theta$) with (sin$n\theta$).

Shape functions N_{i} depend on r and z. Therefore, as shown in Eq. (15) $\left[\overline{B}_{\mathrm{in}}\right]$, as well as $\left[\overline{\overline{B}}_{\mathrm{in}}\right]$, are functions of r, z, θ and obviously of n. A unique strain-displacement matrix can be defined by assembling matrices $\left[\overline{B}_{\mathrm{in}}\right]$ and $\left[\overline{\overline{B}}_{\mathrm{in}}\right]$ as $\left[\overline{B}\right]=\left[\overline{B}_{\mathrm{n}=0}\quad \overline{B}_{\mathrm{n}=1}\quad \ldots\right]$, $\left[\overline{\overline{B}}\right]=\left[\overline{\overline{B}}_{n=0}\quad \overline{\overline{B}}_{n=1}\quad \ldots\right]$. If only H harmonics are retained in the Fourier series, matrices $\left[\overline{B}\right]$ and $\left[\overline{\overline{B}}\right]$ become 1xH arrays of the 6x3M submatrices $\left[\overline{B}_{\mathrm{in}}\right]$ and $\left[\overline{\overline{B}}_{\mathrm{in}}\right]$, respectively.

Two stiffness matrices $\left[\overline{k}_{\mathrm{n}}\right]$ and $\left[\overline{\overline{k}}_{\mathrm{n}}\right]$ have to be defined according to Eq. (4) for both single- and double-barred terms in Fourier series expansion:

$$\left[\overline{k}_{\mathrm{n}}\right]=\int_{A_{\mathrm{el}}}\left[\int_{-\pi}^{\pi}\left[\overline{B}_{\mathrm{n}}\right]^{T}[D]\left[\overline{B}_{\mathrm{n}}\right]\mathrm{d}\theta\right]dA \qquad \left[\overline{\overline{k}}_{\mathrm{n}}\right]=\int_{A_{\mathrm{el}}}\left[\int_{-\pi}^{\pi}\left[\overline{\overline{B}}_{\mathrm{n}}\right]^{T}[D]\left[\overline{\overline{B}}_{\mathrm{n}}\right]\mathrm{d}\theta\right]dA \tag{16}$$

The integrand matrix $[B]^{\mathrm{T}}[D][B]$ is a full matrix of size (3MH)x(3MH); it is composed of an HxH array of 3Mx3M submatrices [k_{n}]. The off-diagonal submatrices contain in every term the products (sin$m\theta$)(sin$n\theta$) or (cos$m\theta$)(cos$n\theta$) with $m\neq n$, which give zero when integrated from –π to +π, due to the so-called orthogonality property of trigonometric functions. The remaining H on-diagonal submatrices, with dimension 3Mx3M, contain (sin$^2n\theta$) and (cos$^2n\theta$) in every term, which integrated from –π to +π give a common factor π (or 2π for n=0). Integration on r and z variables in Eq. (16) is done, as if the problem were axially symmetric.

It should also be mentioned that, due to choice of negative sign in the second expression in Eq. (12), the stiffness matrix for double-barred terms is identical to that of single-barred terms, that is $\left[\bar{k}_{\mathrm{n}}\right]=\left[\bar{\bar{k}}_{\mathrm{n}}\right]$ (Cook et al., 1989). Stiffness matrices for single- and double-barred terms can be arranged in a single diagonal block matrix, as:

$$[k]=diag\left(\left[\bar{k}_0\right]\left[\bar{k}_1\right]\ \ldots\ \left[\bar{k}_H\right]\left[\bar{\bar{k}}_0\right]\left[\bar{\bar{k}}_1\right]\ \ldots\ \left[\bar{\bar{k}}_H\right]\right) \tag{17}$$

where each matrix $\left[\bar{k}_{\mathrm{n}}\right]$ and $\left[\bar{\bar{k}}_{\mathrm{n}}\right]$ is of size 3*M*x3*M* and subscript 0,1, …,*H* specifies the number of the Fourier harmonic. Matrix $[k_0]$ has similar dimension; it contains both symmetric and antisymmetric terms, but obviously no coupling terms are present. The above theory then shows that in linearity, due to the topology of stiffness matrix, the problem is uncoupled and *H* separate problems are solved (Bressan et al., 2009). For example, if the simple case of a cylinder is considered, in a three-dimensional analysis it would be necessary to solve about $(3p^{3/2})/(a^{1/2})$ equations (parameter *a* is the cylinder aspect ratio, defined as the ratio between the length and the radius, parameter *p* is the plane model node number), whereas in the plane case (2*H*+1) systems constituted by 3*p* equations should be solved. As *H*<<*p*, computational time would be strongly reduced.

4. Thermal analysis with a harmonic FE approach

An interesting issue here investigated is the use of a one-dimensional harmonic FE model for steady-state and transient thermal analysis of two-dimensional axisymmetric structures under non-axisymmetric thermal loadings. An example could be a rotating cylinder under imposed temperature and fluxes, which has been used as a simplified model for estimating the non-uniform transient temperature in a work roll of a hot rolling mill (Benasciutti, 2010a). Numerical FE modelling may be still very complex, even using a simplified two-dimensional modelling with a commercial FE code, as transient analysis needs large computational times and computer resources. Instead, a semi-analytical approach based on harmonic model would greatly reduce the computational burden. On the other hand, in commercial FE codes a one-dimensional harmonic thermal finite element is usually not implemented. Hence, this work will close the gap by developing the theory of a one-dimensional harmonic finite element for steady-state and transient thermal analysis of two-dimensional axisymmetric problems under non-axisymmetric thermal loadings. Other examples on thermal problems solved by a numerical approach can be found in (Awrejcewicz et al., 2007, 2009).

4.1 Steady-state thermal analysis

As discussed above, a harmonic model based on Fourier series expansion allows a three-dimensional physical problem to be reduced to a two-dimensional one. Similarly, a two-dimensional problem can be solved by a one-dimensional analysis. A significant reduction in total simulation time and also a saving of computational resources is thus achieved.

The study of two-dimensional problems by harmonic model needs a one-dimensional mesh along the radial direction. In a thermal analysis, the elements used are of "truss type" (one-dimensional) and the degree of freedom in each node is temperature. The theory of different types of harmonic finite element for thermal analysis have been formulated for a plane thermal analysis: two-node elements (with one or two Gauss points) having linear shape functions, three-node elements (with two or four Gauss points) having quadratic shape functions (Benasciutti et al., 2010b, 2011).
An example is shown in Fig. 3(a); two reference systems are used: r is the abscissa in global reference system (r=0 is the centre of the axisymmetric geometry), x is the coordinate within the element. The mesh is one-dimensional and consists of adjacent elements located along the radius of solid, see Fig. 3(b).

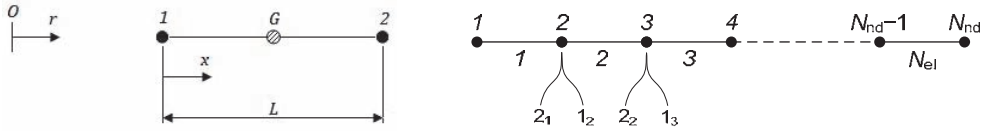


Fig. 3. (a) Two-node element with one Gauss point (r is global coordinate, x is element coordinate); (b) mesh (example of notation: 1_2 and 2_2 are nodes of element 2)

As previously discussed with reference to mechanical analysis, due to orthogonality property of trigonometric terms of Fourier series expansion the element stiffness matrix is a block diagonal matrix, see Eq. (17), where $[k_n]$ is the elementary stiffness matrix for n-th term of Fourier series and H is the number of harmonics.
Explicit expressions for shape functions and stiffness matrix have been derived for each element type mentioned above (Benasciutti et al., 2011). As an example, the stiffness matrix of the two-node element with two Gauss points is here calculated. Similarly to Eq. (12), the temperature is first expanded in Fourier series as:

$$u(r,\theta)=\sum_{n=0}^{\infty}\overline{u}_{\mathrm{n}}(r)\cos n\theta+\sum_{n=1}^{\infty}\overline{\overline{u}}_{\mathrm{n}}(r)\sin n\theta \tag{18}$$

where $\overline{u}_{\mathrm{n}}(r)$ and $\overline{\overline{u}}_{\mathrm{n}}(r)$ are the amplitudes of, respectively, symmetric (single-barred) and anti-symmetric (double-barred) terms, which both depend only on r, but not on θ. In practice, only a finite number of harmonics H is used in the summation in Eq. (18).
Within an element, amplitudes $\overline{u}_{\mathrm{n}}(r)$ and $\overline{\overline{u}}_{\mathrm{n}}(r)$ can be interpolated from *nodal* amplitudes $\overline{u}_{\mathrm{n}}^{1}$, $\overline{u}_{\mathrm{n}}^{2}$, $\overline{\overline{u}}_{\mathrm{n}}^{1}$ and $\overline{\overline{u}}_{\mathrm{n}}^{2}$ (superscript specifies the node number):

$$u(r,\theta)=\sum_{n=0}^{\infty}\left[N_1(r)\overline{u}_{\mathrm{n}}^{1}+N_2(r)\overline{u}_{\mathrm{n}}^{2}\right]\cos n\theta+\sum_{n=1}^{\infty}\left[N_1(r)\overline{\overline{u}}_{\mathrm{n}}^{1}+N_2(r)\overline{\overline{u}}_{\mathrm{n}}^{2}\right]\sin n\theta \tag{19}$$

where $N_1(r)$ and $N_2(r)$ are the shape functions, which are linear in a two-node element: $N_1(r)=(r_2-r)/L$ and $N_2(r)=(r-r_1)/L$. After some simple matrix algebra, the vector of "strains" (derivatives of temperature) in polar coordinates can be written in matrix notation as:

$$\begin{Bmatrix}\varepsilon_r\\ \varepsilon_\theta\end{Bmatrix}=\begin{Bmatrix}\dfrac{\partial u}{\partial r}\\ \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}\end{Bmatrix}=\sum_{n=0}^{\infty}\begin{bmatrix}\cos n\theta & 0\\ 0 & -\sin n\theta\end{bmatrix}\begin{bmatrix}\dfrac{\partial N_1}{\partial r} & \dfrac{\partial N_2}{\partial r}\\ \dfrac{n}{r}N_1 & \dfrac{n}{r}N_2\end{bmatrix}\begin{Bmatrix}\overline{u}_{\mathrm{n}}^{1}\\ \overline{u}_{\mathrm{n}}^{2}\end{Bmatrix}+$$
$$+\sum_{n=0}^{\infty}\begin{bmatrix}\sin n\theta & 0\\ 0 & \cos n\theta\end{bmatrix}\begin{bmatrix}\dfrac{\partial N_1}{\partial r} & \dfrac{\partial N_2}{\partial r}\\ \dfrac{n}{r}N_1 & \dfrac{n}{r}N_2\end{bmatrix}\begin{Bmatrix}\overline{\overline{u}}_{\mathrm{n}}^{1}\\ \overline{\overline{u}}_{\mathrm{n}}^{2}\end{Bmatrix} \tag{20}$$

(for more clarity, explicit dependence on variable r is omitted). Element stiffness matrix for harmonic n is calculated as in Eq. (4) or (16). Matrix product inside the integral contains terms as ($\sin^2 n\theta$) and ($\cos^2 n\theta$), which integrated from $-\pi$ to $+\pi$ give a factor π (or 2π for n=0). The following expression is then obtained (Benasciutti et al., 2011):

$$\left[\overline{k}_{\mathrm{n}}\right]=\left[\begin{matrix}\overline{\overline{k}}_{\mathrm{n}}\end{matrix}\right]=\pi h\int_0^L[B_{\mathrm{n}}]^T[D][B_{\mathrm{n}}]r\,\mathrm{d}r \quad , \quad [B_n]=\begin{bmatrix}\dfrac{\partial N_1}{\partial r} & \dfrac{\partial N_2}{\partial r}\\ \dfrac{n}{r}N_1 & \dfrac{n}{r}N_2\end{bmatrix} \quad n=1,2,... \tag{21}$$

where h is element thickness in z (axial) direction, $[D]$=diag(λ, λ) is a diagonal matrix with material thermal conductivity λ. The integral in Eq. (21) can be solved numerically (Gauss quadrature), obtaining a closed form solution. For example, for the two-node element it is:

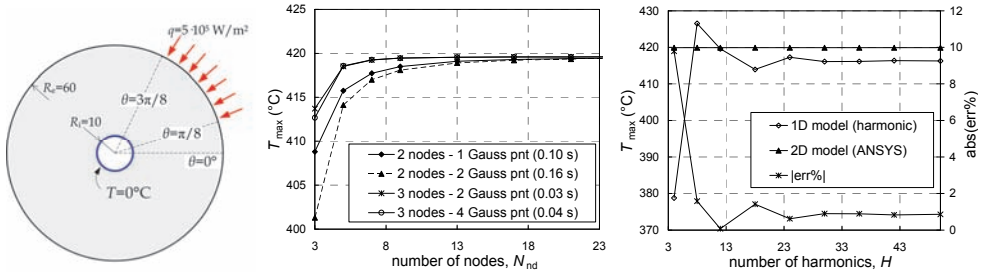
$$\left[k_{\mathrm{n}}\right]=\pi h L r_{\mathrm{G}}\cdot\begin{bmatrix}\left.\dfrac{\partial N_1}{\partial x}\right|_{x_{\mathrm{G}}} & \left.\dfrac{\partial N_2}{\partial x}\right|_{x_{\mathrm{G}}}\\ \dfrac{n}{r_{\mathrm{G}}}N_1(x_{\mathrm{G}}) & \dfrac{n}{r_{\mathrm{G}}}N_2(x_{\mathrm{G}})\end{bmatrix}^T\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\begin{bmatrix}\left.\dfrac{\partial N_1}{\partial x}\right|_{x_{\mathrm{G}}} & \left.\dfrac{\partial N_2}{\partial x}\right|_{x_{\mathrm{G}}}\\ \dfrac{n}{r_{\mathrm{G}}}N_1(x_{\mathrm{G}}) & \dfrac{n}{r_{\mathrm{G}}}N_2(x_{\mathrm{G}})\end{bmatrix} \tag{22}$$

where the two coordinates x_{G} and r_{G} are used to specify the position of Gauss point in the element and global reference system, respectively. As it can be seen, element stiffness matrix depends on both harmonic n and also on radial position of Gauss point, r_{G}. This implies that elements located at different radial positions have different stiffness matrix. Analogous expressions can be easily obtained for other elements, similarly to what done above. The increase of Gauss points would improve the accuracy of numerical integration in Eq. (21).

Theoretically, it is expected that increasing the node number would improve the numerical accuracy of results at the expense of higher computational cost. Element characterised by the highest rate of numerical convergence would allow the use of coarser meshes (with less number of elements), with a considerable decrease in number of equations to be solved and hence a significant reduction in required computational burden. Therefore, the choice of the most suitable element for a selected problem represents a crucial step in the analysis.

A test has been performed to compare the performance (accuracy and convergence rate) of different elements. A reference thermal problem is repeatedly solved by different elements, having various mesh densities. Considering that computational burden is approximately proportional to the number of equations to be solved and, in turn, to the number of nodes in the mesh, a comparable computational time would be roughly achieved by different meshes

having the same number of nodes. The comparison then assumes that different elements have an equivalent mesh if the number of nodes is the same. For example, a mesh with 10 two-node elements (21 nodes) has approximately an equivalent computational burden to a mesh with 20 three-node elements (21 nodes).
The geometry and thermal load configuration used in test is shown in Fig. 4(a). A constant thermal flux is applied over the surface angular sector (π/8÷3π/8), while a zero temperature is prescribed on the inner surface. In FE model, thermal flux is converted into equivalent nodal loads; a total of H=12 harmonics is used in Fourier series expansion.



(a) (b) (c)

Fig. 4. (a) Thermal layout used in comparative test (material thermal conductivity is λ=50 W/mK); (b) converge rate: trend of calculated maximum temperature vs. number of nodes in the mesh, for different element types; (c) results accuracy vs. number of harmonics H.

The comparison is made with reference to the maximum temperature calculated by different element and mesh types. Figure 4(b) shows the trend as a function of the number of nodes in the mesh (a uniform element density is used); figures quoted in parenthesis indicate, as a rough guide, the computational time required by each element. The asymptotic value of temperature (for an infinitely fine mesh) has to be interpreted as the true (converged) value.
At a first sight, it appears clear how three-node elements give a much faster convergence rate, compared to two-node elements. The best performance (slowest running time) is given by three-node element with two Gauss points, which has been used in all subsequent simulations (see Fig. 9(a)). Contrarily to what is expected, an increase of Gauss points (keeping the element nodes fixed) does not cause any increase in results accuracy (that is to say, in coarser meshes better results are obtained with lower Gauss points).
Another important parameter is the number of harmonics H, which should be chosen as the best balance between accuracy and simulation time. Higher H values are expected to give higher precision in Fourier expansion of applied loads (especially for step loads), as well as better accuracy and higher simulation time. A test is performed to assess the effect of the number of harmonic H on accuracy of results and computational speed. Results from semi-analytical model with different discrete values of H (i.e. 4, 8, 12, 18, etc.) have been compared with those by a two-dimensional FE model (commercial code ANSYS® has been used). A relative error between the maximum temperature calculated by harmonic model and that of plane FE model is also calculated. The comparison is shown in Fig. 4(c); with H=12 the absolute relative error is about 1%, as can be seen, while for higher harmonics the error rapidly decreases below 0.2%. On the other hand, it has been noted that simulation

time in harmonic model increases roughly linearly with H (because equilibrium equations are not coupled). Also the approximation error (Gibbs' phenomenon) in Fourier series expansion of step-wise periodic functions should be considered carefully in selecting the optimal value of H. For the example examined, the best choice would be H=24, which, however, does not actually represent a result of general validity. In fact, each problem needs to be carefully analyzed to find the best H value.
Different types of imposed boundary conditions (temperature, flux and convection) have been studied in detail. Similarly, a thermal flux applied on boundary elements is converted into equivalent nodal loads. Instead, special attention has been deserved to some particular boundary conditions, as application of different temperatures on different boundary sectors or application of convective heat exchange, see Section 4.3.

4.2 Transient thermal analysis

The differential equation of equilibrium governing a transient thermal FE analysis is:

$$[M]\{\dot{u}\}+[K]\{u\}=\{F\} \tag{23}$$

where $\{\dot{u}\}$ is the vector with derivative of temperatures, [M] and [K] are the "mass" and "stiffness" matrices (assembled from the corresponding element matrices), {F} is the vector of externally applied loads. For a one-dimensional element, the "mass matrix" is defined as:

$$[m_{\mathrm{el}}]=h\rho c\int_0^L[N]^T[N]r\,\mathrm{d}r \tag{24}$$

which depends on shape function matrix [N(x)]=[N_1(x) N_2(x)], as well as on material volumic mass ρ and specific heat c. Unlike matrix [B_n], the "mass matrix" does not explicitly depend on the harmonic n, hence it can be calculated only once for all terms of Fourier series expansion.
Explicit expressions for "mass matrix" have been calculated for two- and three-node elements (Benasciutti et al., 2011). For example, for a two-node element it is:

$$[m_{\mathrm{el}}]=h\rho c\int_0^L\begin{bmatrix}\left(1-\frac{x}{L}\right)^2 & \left(1-\frac{x}{L}\right)\frac{x}{L}\\ \left(1-\frac{x}{L}\right)\frac{x}{L} & \left(\frac{x}{L}\right)^2\end{bmatrix}r\mathrm{d}r=\frac{h\rho c}{12}L\begin{bmatrix}4r_1+L & 2r_1+L\\ 2r_1+L & 4r_1+3L\end{bmatrix} \tag{25}$$

where r_1 is radial position of node "1" in global reference system. As shown, [m_{el}] depends on the position of element in absolute reference system, which means that two elements of equal length, but different location in the mesh, would have a different "mass matrix".
In a FE model, the "mass matrices" for each element in the mesh are assembled, to get a global "mass matrix" [M], similarly to assemblage of global "stiffness matrix" [K]. Similarly to [K], also [M] is a block diagonal matrix, with non-zero terms close to the main diagonal.
Mathematically, Eq. (23) represents a system of linear differential equations of second order, which in FE procedures are usually solved by numerical methods. In fact, in FE approach time domain is represented by a discrete sequence of time instants, in which solution is calculated. The time difference between adjacent time instants is the time step Δt.

Numerical methods differ in the way the time derivative is approximated. The partial differential equation governing a transient thermal analysis can be solved by a finite difference method, implemented as an explicit or implicit numerical algorithm: explicit methods calculate the state of the system at a later time only considering the state of the system at current time, while implicit methods calculate the state of a system at a later time by solving an equation involving both the current state of the system and the later one. Implicit methods are usually slower (although more accurate) than explicit methods on single time step computation, as they require to solve a linear system at each time step. Conversely, explicit methods are usually faster, as they do not need to invert matrices.

In the present study, time integration has been performed by two different numerical methods. The first is *Forward Finite Difference* (FFD) method, which is an explicit method proven to be very quick and effective. However, it is said to be "conditionally stable", as it requires the time step Δt be smaller than a critical value to get a stable solution. This can represent a serious disadvantage, as the need of a stable solution may require a relatively small time step, which can result in a very large total simulation time.

In FFD method, the time derivative in Eq. (23) is approximated as the discrete increment $\{\dot{u}\} = \left(\{u\}_{i+1} - \{u\}_i\right) / \Delta t$, where $\{u\}_i$ and $\{u\}_{i+1}$ are the vectors of nodal temperatures calculated at consecutive time steps i and i+1. Substituting in Eq. (23) and rearranging, it follows the fundamental equation for FFD algorithm:

$$\{u\}_{i+1} = \{u\}_i - [M]^{-1} \Delta t [K] \{u\}_i + \Delta t [M]^{-1} \{F\} \tag{26}$$

Note that "mass matrix" [M] is time-independent, thus it can be inverted only once, with a considerable time saving.

An alternative and completely new original method has been developed for time integration. In analogy with the linear acceleration method (Bathe, 1996), the method here presented assumes a linear variation of the first derivative, thus it has been called *Linear Speed Method* (LSM); in symbols:

$$\{\dot{u}(\tau)\} = \{\dot{u}\}_i + \frac{\{\dot{u}\}_{i+1} - \{\dot{u}\}_i}{\Delta t} \tau \tag{27}$$

where $\{\dot{u}\}_i$ and $\{\dot{u}\}_{i+1}$ are the vectors of the derivatives of nodal solutions calculated at consecutive time steps i and i+1, while $\tau \in (0, \Delta t)$ is a dummy time variable. Integration of Eq. (27) gives $\{\dot{u}\}_{i+1} = 2\left(\{u\}_{i+1} - \{u\}_i\right) / \Delta t - \{\dot{u}\}_i$; substituting into Eq. (23) and rearranging, gives the fundamental equation of LSM method as:

$$\begin{cases} \{u\}_{i+1} = \left(\dfrac{2}{\Delta t}[M] + [K]\right)^{-1} \left(\{F\} + [M]\left(\dfrac{2}{\Delta t}\{u\}_i + \{\dot{u}\}_i\right)\right) \\ \{\dot{u}\}_{i+1} = \dfrac{2}{\Delta t}\left(\{u\}_{i+1} - \{u\}_i\right) - \{\dot{u}\}_i \end{cases} \tag{28}$$

Unlike FFD algorithm, this method is implicit, as the system in Eq. (28) has to be solved at each time step. Furthermore, LSM is also "unconditionally stable" (i.e. solution converges independently of the choice of time step). Note that to further improve the computational speed, the system in Eq. (28) can be solved by LU factorization.

A test has been performed, to compare the relative performance and accuracy of FFD and LSM algorithm in semi-analytical approach. The geometry and thermal load configuration shown in Fig. 4(a) is analyzed under a physical transient of 900 s, starting from a uniform temperature of 0°C. The comparison is made both with a plane FE and a semi-analytical model (implemented with FFD method and LSM algorithm, with and without LU factorization), with H=24. Additional material parameters used in simulation are volumic mass (ρ=7800 kg/m^3) and specific heat (c=460.5 J/kgK). Plane model adopts eight-node thermal elements (which have 3 nodes on each side), while harmonic model uses 12 elements with three nodes and two Gauss points (therefore, both models have the same number of nodes along the radial direction).

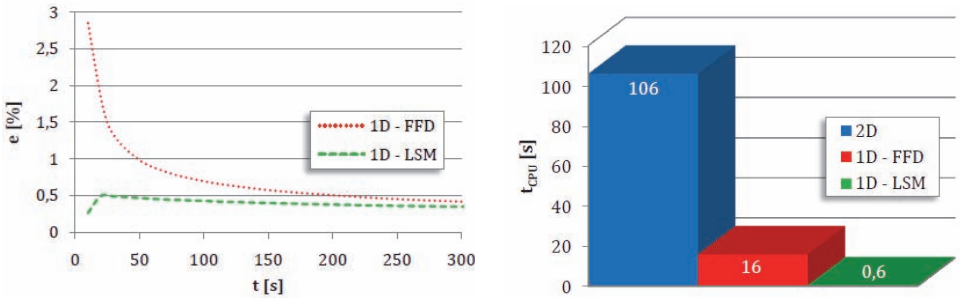


Fig. 5. (a) Errors in calculated maximum temperature vs. time; (b) comparison of computational time (the figures quoted are purely indicative).

The same number of computation steps must be imposed in all FE modes (at least, in 2D and 1D-LSM), to allow an effective comparison of simulation time. A time step Δt=1 s is chosen, so that the interval of 15 minutes (900 s) is divided into 900 time steps. Unfortunately, such a value is shown to be inadequate for 1D-FFD model; it is greater than the critical time step, thus it causes the numerical solution to not converge (numerical instability). A lower time step Δt must then be imposed; for example, an acceptable value is Δt=0.0335 s, although it gives a much greater number of time steps (26866) for the same physical transient of 15 minutes here analyzed.

In harmonic model, both algorithms (FFD, LSM) have been shown to provide almost coincident results, which are also identical to those of 2D model. A time saving for LSM of about 4% compared to FFD and even 99% with respect to a plane FE model is observed. The use of LU factorization in LSM algorithm further reduced computation time of about one third. The above test then revealed that the fastest algorithm for transient thermal analysis is LSM method with LU factorization; this has been used in the illustrative example.

4.3 Boundary conditions

4.3.1 Prescribed thermal flux and temperature

In a thermal FE analysis, a prescribed temperature in a node is equivalent to an imposed nodal displacement in a structural analysis. In a harmonic model, the temperature value imposed on a node would be expanded by the Fourier series to the *whole* circumference (nodal circle) passing through the node. Thus, it appears that in harmonic model only a constant temperature could be applied on the boundary, while in practical application it is often necessary to prescribe different temperature values along the boundary.

This particular boundary condition, however, poses particular numerical problems, due to one-dimensional nature of harmonic model. Three methods have been proposed (Benasciutti et al., 2011): two of them (fictitious thermal flux, imposed temperature) are based on Fourier series expansion of the imposed temperature, the third is based on Lagrange multipliers.

In the first method, the applied temperature is interpreted as a consequence of a fictitious thermal flux (unknown) on the same surface portion, which has to be determined. The prescribed temperature is expanded in Fourier series, similarly to thermal flux; the coefficients for thermal flux expansion are then calculated from those defining the series expansion of applied temperature. In the second method, the prescribed temperature is treated similarly to an imposed displacement (constraint) in a mechanical analysis, in which the equation corresponding to the prescribed degree of freedom is cancelled out in the system of equilibrium equations. In the harmonic model, this procedure has to be applied to every term in Fourier series expansion. In the third method with Lagrange multipliers, the "stiffness matrix" is calculated by a variational approach, which minimises the total potential energy Π. In stiffness matrix [k], formed by submatrices [k_{ni}], the generic element would be $k^{ij}=\partial^2\Pi/\partial u^i\partial u^j$, where indexes i, j range from 1 to $nF{\cdot}M$, with M total number of nodes and nF number of Fourier terms. Without going into details, it suffices to say that the method seeks the minimum of a total potential energy Π_β, constrained to prescribed temperature:

$$\Pi_\beta\left(u_i, u_j, \beta\right) = \Pi\left(u_i, u_j\right) + \beta \cdot f(u_i) \tag{29}$$

where β is the Lagrange multiplier. As a result, the stiffness matrix calculated by this third method is no longer uncoupled, as additional new rows and columns are inserted to describe the imposed temperature.

Illustrative examples showed that the performance (accuracy and computation speed) of all three methods is quite comparable, although the third one gives more flexibility in representing various boundary conditions. In particular, an imposed temperature and flux over the same boundary portion is not possible with the first two methods. The decisive advantage of the method using Lagrange multipliers is its possibility to represent any variation with angle θ for the prescribed temperature.

4.3.2 Convection

A convective flux per unit area is defined as $q_\alpha=\alpha(u-u_\infty)$, where α is convection coefficient, u the surface temperature and u_∞ a reference (bulk) temperature of surrounding medium. A convective flux then depends on the surface temperature and may also change over time. Then, algorithms for transient analysis can be used to solve convective boundary condition.

Three different methods with various approximation levels have been developed: one based on constant convective coefficient and single average surface temperature, one based on a step-wise constant convective flux and average surface temperature, the third based on trigonometric formulae (which has been implemented by two different algorithms: temperature calculated either at previous or at current time step) (Benasciutti et al., 2011).

In the first method, the constant convective coefficient α over an angular sector ($\theta_1 \div \theta_2$) is expanded in Fourier series, similarly to an imposed thermal flux. The convective heat flux is:

$$\{q_\alpha\} = h\,R\left(\theta_2 - \theta_1\right)\{\alpha\}\left(u_{\text{av}} - u_\infty\right) \tag{30}$$

where u_{av} is the average surface temperature (calculated at previous time iteration), R is the radius of the axisymmetric plane body and $\{\alpha\} = \{\bar{\alpha}_0 \quad \bar{\alpha}_1 \quad \bar{\alpha}_2 \quad \ldots \quad \bar{\bar{\alpha}}_1 \quad \bar{\bar{\alpha}}_2 \quad \ldots\}^T$ is the vector with series coefficients. Although very simple, this method assumes a constant heat flux within sector ($\theta_1 \div \theta_2$); obviously, accuracy decreases as surface temperature shows large variations over interval ($\theta_1 \div \theta_2$). An improvement can be obtained by further dividing ($\theta_1 \div \theta_2$) into sub-sectors; a step-wise constant convective coefficient and average temperature are then calculated in every sub-sector.

The third method tries to capture the real surface temperature variation within sector ($\theta_1 \div \theta_2$). Two separate Fourier series are then used, one describing surface temperature relative to bulk temperature u_∞, one modelling convective coefficient:

$$\{q_\alpha(\theta)\} = hR(\theta_2 - \theta_1)\left[\bar{\alpha}_0 + \sum_{n=1}^{\infty}\left(\bar{\alpha}_n \cos n\theta + \bar{\bar{\alpha}}_n \sin n\theta\right)\right]\left[\bar{u}_0 + \sum_{n=1}^{\infty}\left(\bar{u}_n \cos n\theta + \bar{\bar{u}}_n \sin n\theta\right)\right] \tag{31}$$

The product of the two series is further simplified through the well-known prosthaphaeresis trigonometric formulae, arriving at the following expression (for node i):

$$\left\{q_\alpha^i(\theta)\right\} = hR(\theta_2 - \theta_1)\left[\mathrm{A}^i\right]\{\bar{u}\}^i \tag{32}$$

where $[A^i]$ is a matrix including amplitudes of Fourier series of α (it can be then calculated only once) and $\{\bar{u}\}^i$ is a column vector with amplitudes of the surface temperature field u. Vector $\{q_\alpha\}$ of convective flux is finally computed from $\left\{q_\alpha^i\right\}$ according to global node numbering in the mesh. In harmonic FE model, the set of matrices $[A^i]$ are assembled to get a global sparse matrix $[A]$, with same dimension as $[K]$ and $[M]$. Equation (23) for transient thermal analysis can be updated with convective flux:

$$[M]\{\dot{u}\} + [K]\{u\} = \{F\} - [\mathrm{A}]\{u\} \tag{33}$$

where $\{u\}$ can be calculated either at current or next time step. Without going into the details, it suffices to say that a numerical test was performed to compare the relative performance (accuracy and speed) of methods. The method with constant convective coefficient was shown to be not very accurate, as it tends to over-estimate the actual flux where surface temperature is lower. A slight improvement (at the expense of higher computational time) is achieved with a step-wise constant flux, although an increase in the number of surface divisions may cause numerical instability (a smaller Δt must then be used). However, the best accuracy is reached by the method with trigonometric functions, which also has the lowest computational time (which is slightly lower to that of first method). In addition, the test example used did not reveal any difference between the algorithms which computes temperature at current or next time step.

5. Numerical examples

The numerical examples here discussed address both a mechanical and thermal analysis, with the aim of testing the performance and accuracy of semi-analytical approach described above. Two different applications are discussed: a shaft under torsion, bending and axial mechanical loadings and a rotating cylinder under surface thermal loadings. The study

confirms how semi-analytical approach is able to provide accurate results, with in addition a strong reduction in computation time, compared with classical two- or three-dimensional FE simulation.

5.1 Mechanical analysis

A first case study here presented refers to a cylindrical component with a shoulder fillet, loaded under torsion, bending and axial load applied at the ends, see Fig. 6(a). The geometry shown is a type of stress concentration frequently encountered in shafts, axle spindles and rotors machine design. The shoulder fillet is designed to have D/d=1.5, r/d=0.1, so that stress concentration factors given by literature (Peterson, 1974) are respectively $K_{t,N}$=1.88 (axial load), $K_{t,Mf}$=1.72 (bending), $K_{t,Mt}$=1.37 (torsion).
This geometry can be solved by a harmonic model; the axial, bending and torsion loads can be represented by surface loads expanded in Fourier series as:

$$\{p\} = \begin{Bmatrix} p_r \\ p_\theta \\ p_z \end{Bmatrix} = \begin{Bmatrix} 0 \\ p_z^{\text{torsion}} \\ p_{z,\text{axial}} + p_{z,\text{bend}} \end{Bmatrix} = \begin{Bmatrix} 0 \\ \tau_0 \cdot \xi \\ \sigma_{0,\text{axial}} + \sigma_{0,\text{bend}} \cdot \sin(\theta) \end{Bmatrix} \tag{34}$$

where $\xi = 0 \div 1$ is a suitable normalized radial coordinate ($\xi = 0$ at shaft centre, $\xi = 1$ at shaft the outermost radius), used to model the linear variation a torsion load with radius. Symbols $\sigma_{0,axial}$ and $\sigma_{0,bend}$ represent the maximum stress for axial and bending loading, while τ_0 is the maximum shear stress for torsion load.
The axisymmetric structure under the three different loading configurations can be easily solved by using three plane models, see Fig. 6(b)-(d): an axisymmetric model for the axial load and an axi-antisymmetric model for torsion; bending can be treated by means of a harmonic model, with only one term in Fourier series (in fact, bending stresses follow a sinusoidal variation law with respect to symmetry axis).

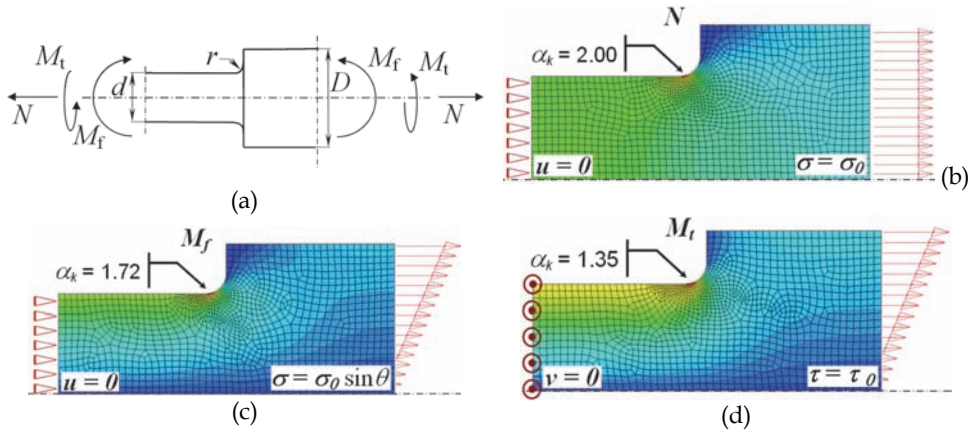


Fig. 6. (a) Analysed geometry with different applied loads (N axial, M_f bending, M_t torsion); (b), (c), (d) plane FE models used in semi-analytical approach, with loads and constraints (values of α_k indicate the numerically calculated stress concentration factors)

The figure shows the plane mesh used, as well as the applied loads and constraints. The adequacy of the mesh has been tested with a convergence analysis based on the comparison of stress concentration factors calculated by increasing levels of mesh refinement. The numerical solution given by a three-dimensional FE model with hexahedral elements is also used for comparison purposes. The stress concentration factors numerically calculated by plane models, shown in Fig. 6, have been compared with theoretical values from manuals (Peterson, 1974), as well as with numerical values from three-dimensional FE analysis. The overall comparison shows a general good agreement. Compared to three-dimensional model, the semi-analytical approach greatly reduces the overall computation time.
The semi-analytical approach here described can thus be extended to the analysis of a shaft loaded axially, in torsion and in bending, showing that the stress distribution can be accurately described (at least far from the points where loadings are applied) by using only three terms of the Fourier series, i.e. the constant and the first harmonic term.
A second case study here analysed is thus a simply supported shaft, loaded by non-planar distributed loads, see Fig. 7. A three-dimensional model and a plane harmonic model, both having similar mesh distributions, have been considered.

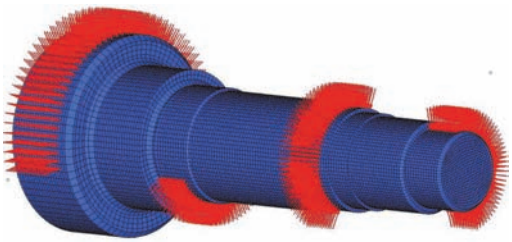


Fig. 7. Three-dimensional model of a shaft with non-planar loads

Table 1 shows the maximum tensile stress σ_z close to a fillet radius, normalised to the asymptotic stress value $\sigma_{z\infty}$ calculated by a three-dimensional FE model with a very refined mesh. Figures in parenthesis indicate a rough estimate of computational time required by each analysis.

Terms of Fourier series	**stress ratio, $\sigma_z/\sigma_{z\infty}$ (computational time)**	
	localised re-meshing	no localised re-meshing
30	1.00 (*90 s*)	0.92 (*90 s*)
7	1.00 (*21 s*)	0.92 (*21 s*)
3	0.98 (*9 s*)	0.91 (*9 s*)
3D model	1.00 (*days*)	0.84 (*12600 s*)

Table 1. Comparison of calculated stress; figures in parenthesis indicate a rough estimate of computational time

Results are reported respectively for the original mesh and after a local refinement close to stress concentrations. It is possible to observe that the plane model with only 2 harmonics and non-refined mesh gives better results with respect to those achievable with the three-dimensional model with the same mesh distribution. In the case of a plane model with

localised re-meshing, a relevant reduction in the error can be observed, without a significant increment of computational time, even if the number of degrees of freedom increases of about 10 %. It can be noticed that, with only 3 terms of Fourier series, an error lower than 2% can be obtained. Convergence can thus be achieved quite easily with the harmonic model. With respect to the harmonic case, the three-dimensional model allows to achieve a similar accuracy only at the expense of unfeasible computational times.

5.2 Thermal analysis

The second example refers to transient analysis of a rotating cylinder under thermal loadings (see Fig. 8), which has been used as a simplified model for a work roll of a hot rolling mill (Benasciutti et al., 2010a). In the simplified approach, only work roll (without the strip) was modelled in a two-dimensional analysis; thermal loadings are applied on the surface, to simulate strip heating and water jet cooling.

The thermal load configuration used in simulations is shown in Fig. 8(a): an infinitely-long cylinder, rotating at constant angular speed and subjected to constant input heat flux and convective cooling. The numerical analysis simulates a thermal transient of 3600 seconds, which corresponds to about 1690 roll revolutions. At the initial simulation time, work roll is assumed at a constant uniform temperature T_{roll}=20°C. Simulation parameters, assumed for simplicity as temperature invariant, are summarized in Table 2.

The work roll configuration in Fig. 8 is an example of axisymmetric structure under non-axially symmetric thermal loadings, which can be solved by the semi-analytical approach previously described. A two-dimensional FE model, implemented by the commercial code ANSYS®, is also used for comparison purposes, to test the accuracy of harmonic FE model.

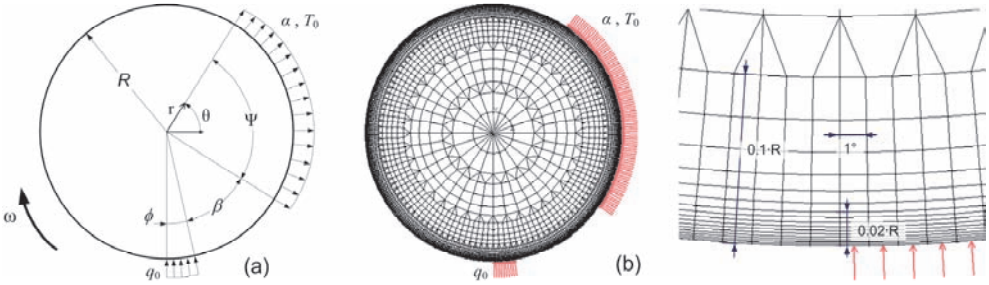


Fig. 8. (a) Thermal load configuration analyzed; (b) plane FE model (global and zoom)

Figure 8(b) shows the plane FE model of work roll. A mesh refinement is imposed near the surface, along the tangential and radial directions, to capture the thermal gradient here expected. Small elements are located for a depth of 10% of work roll radius, with even smaller elements placed immediately underneath the surface, for a depth of 2% of radius.

Since in FE analysis rigid body motion is not allowed, work roll rotation has been simulated by considering the roll at rest and by applying rotating thermal loadings. As an order of magnitude, the simulation required about 3 days of simulation. It is worth mentioning how results of thermal analysis have been validated, see (Benasciutti et al., 2010b), by an analytical solution for the stationary temperature distribution (Patula, 1981). Other details on FE model and simulation parameters can be found in (Benasciutti 2010a).

Parameter	Value
R=300 mm	Cylinder radius
ω=2,953 rad/s	Cylinder angular speed
ϕ = 10°	Heating sector
β = 45°	Angular gap between heating and cooling
Ψ = 90°	Cooling sector
q_0=13,7×10^6 W/m^2	Input thermal flux
α=10100 W/m^{2}K	Convection coefficient
T_0=20°C	Bulk temperature of cooling medium
T_{roll}=20°C	Initial temperature of work roll

Table 2. Several geometrical and thermal parameters used in simulations

In semi-analytical approach, the 2D geometry of work roll in Fig. 8(a) is represented by a 1D model. A three-node element with two Gauss points is used, see Fig. 9(a); the mesh, shown in Fig. 9 (b), has a total amount of 28 elements and 57 nodes. Note that, for an effective comparison, both 1D and 2D models have an identical element distribution in radial direction, compare Fig. 8(b) with Fig. 9 (b). In addition, the same time step in transient analysis has been chosen for both models, to allow a comparison of running times.
In harmonic model, the choice of the correct number of harmonic H is critical, especially for the load configuration in Fig. 8 (a), where a stepped flux is applied over a very small angle (ϕ=10°). In fact, in Fourier series expansion the approximation error tends to increase as step width decreases, because of Gibbs' phenomenon at jump discontinuity; the error could be minimized by an appropriate high number of harmonics. Several benchmark tests were performed (Benasciutti et al., 2011) to identify the optimal number of harmonics for the configuration here analyzed. A comparison of the maximum transient temperature, calculated by plane model and various harmonic models with different number of harmonics revealed that H=100 would be an optimal compromise between accuracy and computing time.

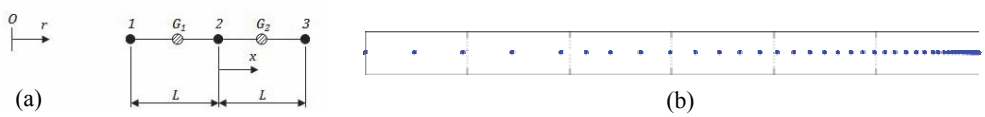


Fig. 9. (a) Three-node element with two Gauss points; (b) mesh (each dot point is a node)

For simulation of convective cooling, the choice is between two algorithms based on trigonometric functions. For H=100 harmonics, the test revealed that algorithm based on temperature computed at previous time step is faster (of about 20%) than algorithm which computes temperature at current time step, although both give comparable accuracy levels.
Conversely to plane FE approach, which needs that work roll must be fixed and thermal loads rotating, in harmonic model two options are available: work roll fixed with rotating thermal loading, or vice versa. A comparative study (Benasciutti et al., 2011) showed that both methods give similar running times, with the first one (work roll fixed) slightly faster (and also more simple to implement) than the second. Therefore, method with rotating thermal loading has been preferred.

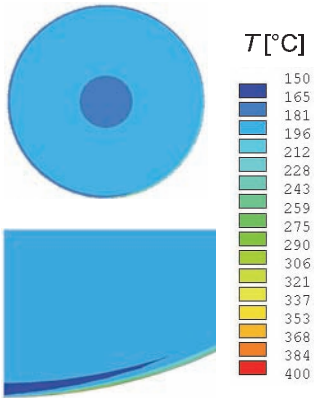


Fig. 10. Temperature distribution in work roll after 1800 s given by plane FE model.

Results for thermal simulations are presented from Fig. 10 to Fig. 13 . For example, Fig. 10 shows the temperature field in work roll after 1800 seconds calculated by plane FE model. The temperature map at other time instants (not included here) would show a progressive heating of the entire work roll, with the largest temperature gradients localised very close to the surface (which justifies the use there of very small elements in the mesh). The temperature field calculated by semi-analytical FE model is not provided, as it is very similar to that given by plane FE model.

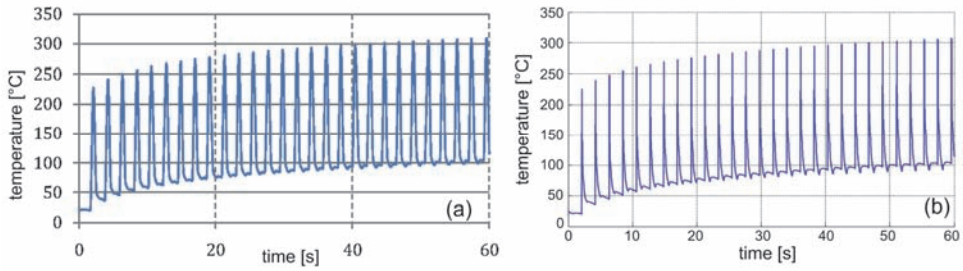


Fig. 11. Temperature history for a point on work roll surface: (a) plane and (b) harmonic FE model.

Figure 11, instead, compares the temperature time history within a 60-seconds time interval, for a point located on work roll surface. Each peak temperature occurs when the monitored point enters the heating zone, thus the series of equally-spaced peaks identifies the sequence of work roll rotations. The progressive increase of peak temperature in consecutive rotations confirms the transient nature of the thermal phenomenon here investigated. A very similar trend is observed for both FE models; the small differences may be attributed to the different rate of results saving on computer hard disk.
The result shown in Fig. 12 refers instead to the radial temperature distribution at next time instants. Only temperature distribution for plane FE model is shown, as that calculated by harmonic model would be practically coincident . The continuous temperature increase with time, especially inside the work roll, is indicated by the different curves. Both diagrams also

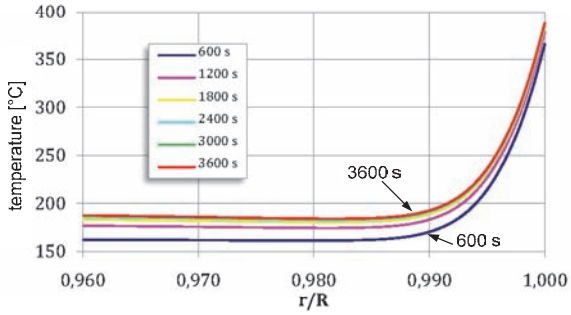


Fig. 12. Radial temperature at different time instants for plane FE model.

confirm that work roll remains at a rough uniform temperature, except for a very small portion very close to the surface (wide about 1% of roll radius), where a steep temperature gradient is observed (note that diagrams only plot radial coordinates close to roll surface). The extremely localized nature of temperature variation within this narrow region is usually called "thermal boundary layer" in technical literature.

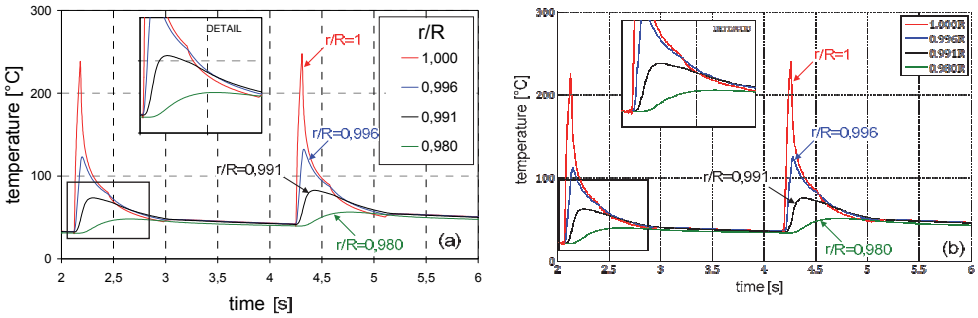


Fig. 13. Temperature time history for points at different radial depths, along the same angular position. (a) plane and (b) harmonic FE model.

Finally, Fig. 13 shows the temperature change for points at different radial depths, along a fixed angular position. A close agreement between semi-analytical and plane FE models is confirmed. The figure further emphasizes that the thermal gradient is confined in a small region close to the boundary. In fact, on work roll surface the temperature ranges of about 200°C, while at 1 mm depth from surface the variation is only 100 °C, and even negligible at 6 mm below surface. In addition, the detail in the same figure highlights a sort of "thermal inversion" phenomenon induced by forced convection cooling, in which the work roll material on the surface is at a lower temperature than material inside. The lateral expansion of surface elements, prevented by the surrounding elements at lower temperature, is the basic mechanism which explains the development of thermal stresses in work roll.

6. Conclusions

This work has investigated theory and application of simplified FE approaches for the analysis of axisymmetric structures loaded by non-axisymmetric loadings. The aim was to

develop alternative FE methods, which allow obtaining the solution of complex three-dimensional problems through a combination of several simpler and faster one- and two-dimensional analyses, which usually require reduced computational efforts. The study has focused on both mechanical and thermal problems, in which the structure is axisymmetric, but the load is not.

In the mechanical context, the classical axisymmetric stress analysis has been first reviewed from literature, as it constitutes a reference example for the subsequent discussion. Using a similar approach, an original finite element is next developed, for solving axisymmetric structures under axi-antisymmetric applied loads, as for example a shaft under a torsion load. General equations for displacement field, strain and stress have been derived.

Another and more general class of problems is that of axisymmetric structures loaded non-axisymmetrically, which are solved by a semi-analytical finite element approach based on Fourier series expansion of applied loads and displacement field. In a linear analysis, the harmonics in Fourier series are totally uncoupled due to orthogonality property of trigonometric functions. Therefore, a complex three-dimensional problem is solved by superposition of solutions of several two-dimensional analyses, or similarly a two-dimensional problem is replaced by several one-dimensional analyses. A considerable reduction in computational times and computer resources is then achieved. The theoretical framework for harmonic finite element analysis has been first presented, as the fundamental equation for stiffness matrix and the stress/strain relations. The performance of the semi-analytical approach is discussed with a numerical example: a shaft with a shoulder filled under respectively an axial, bending and torsion loading. The stress concentration factors calculated by harmonic model are compared with values provided by manuals. The study has confirmed that harmonic model is capable to predict with great accuracy the stress concentration factors of a three-dimensional geometry under three different loading conditions, by using simple two-dimensional models. A second case study showed that the case of a shaft loaded by non-planar loads can be solved quite easily by using a harmonic model with only three terms of the Fourier series. This approach gives a relevant advantage in terms of computational time compared to three-dimensional modelling.

For what concerns thermal problems, it has been developed a harmonic finite element approach for the steady-state and transient thermal analysis of two-dimensional axisymmetric structures under non-axisymmetric thermal loadings. The theory of different types of one-dimensional finite element has been derived: two-node elements (with one or two Gauss points) or three-node elements (with two or four Gauss points). The relative performance of all elements has been compared in terms of accuracy and computation speed; the best performance is provided by three-node element with two Gauss points. Also the choice of the number of harmonic is investigated with a benchmark test.

For transient thermal analysis, explicit expression of "mass matrix" for the two-node element has been presented. Furthermore, two algorithms for numerical time integration has been introduced: the *Fast Forward Difference* (FFD) and the completely original *Linear Speed Method* (LSM), also implemented by LU factorization. A benchmark test revealed that the fastest algorithm for transient analysis is LSM with LU factorization. An illustrative example is finally discussed, which refers to a rotating cylinder under thermal loadings, used as a simplified model for work roll in hot rolling mill. The transient temperature distribution is estimated by a semi-analytical approach and a plane FE model. The harmonic model has been shown to give results in very close agreement with those of plane FE model, with however a significant reduction of total simulation time.

The presented results have confirmed the great reliability offered by semi-analytical approach, in providing accurate results with at the same time a significant reduction in computational times, compared to classical FE analyses.

7. References

Awrejcewicz, J.; Krysko, V.A. & Krysko, A.V. (2007). *Thermo-dynamics of plates and shells*, Springer-Verlag, ISBN 3-540-34261-3, Berlin (Germany).

Awrejcewicz, J. & Pyryev, Yu. (2009). *Nonsmooth dynamics of contacting thermoelastic bodies*. Springer Science, ISBN 978-0-387-09652-0, New York (USA).

Bathe, K.J. (1996). *Finite element procedures in engineering analysis*, Prentice-Hall, ISBN 0-13-301458-4, New Jersey (USA).

Benasciutti, D.; Brusa, E. & Bazzaro, G. (2010a). Finite element prediction of thermal stresses in work roll of hot rolling mills. *Procedia Engineering*, Vol.2, No.1, pp. 707-716.

Benasciutti, D.; De Bona, F. & Munteanu, M. Gh. (2010b). Harmonic model for numerical simulation of thermal stresses in work roll of hot rolling mill. *Proceedings of 4th Europ. Congress on Comp. Mechanics (ECCM IV): Solids, Structures and Coupled Problems in Engineering*, Paris, May 2010.

Benasciutti, D.; De Bona, F. & Munteanu, M. Gh. (2011). Work roll in hot strip rolling: a semi-analytical approach for estimating temperatures and thermal stresses. To be included in *Proceedings of 9th International Conference on Advanced Manufacturing Systems and Technology (AMST 11)*, Mali Losinj, Croatia, June, 2011.

Bressan, F.; De Bona, F. & Munteanu, M. Gh. (2009). Semi-analytical Finite element for Shaft Design, *Proceedings of 6th International Congress of Croatian Society of Mechanics (ICCSM)*, Dubrovnik, Croatia, September-October 2009.

Cook, R.D.; Malkus, D.S. & Plesha, M.E. (1989). *Concepts and applications of finite element analysis* (3rd ed.), Wiley & Sons, ISBN 0-471-84788-7, USA.

Genta, G. & Tonoli, A. (1996). An harmonic finite element for the analysis of flexural, torsional and axial rotordynamic behaviour of discs. *J. Sound Vib.*, Vol. 196, No.1, pp. 19-43.

Kim, J.R.; Kim, S.J. & Kim, W.D. (1994). Parallel computing using semi-analytical finite element method. *AIAA Journal*, Vol.32, No.5, pp. 1066-1071.

Lai, J.Y. & Booker, J.R. (1991). Application of discrete Fourier series to the finite element stress analysis of axi-symmetric solids. *Int. J. Numer. Methods Eng.*, Vol. 31, No.4, pp. 619-647.

Patula, E.J. (1981). Steady-state temperature distribution in a rotating roll subject to surface heat fluxes and convective cooling. *J. Heat Transf.*, Vol. 103, pp. 36-41.

Pedersen, P. & Laursen, C.L. (1982). Design of minimum stress concentration by finite element and linear programming. *J. Struct. Mech.,* Vol.10, No.4, pp. 375-391.

Peterson, R.E. (1974). *Stress concentration factors*. Wiley & Sons, ISBN 0-471-68329-9, New York.

Taiebait, H.A. & Carter, J.P. (2001). A semi-analytical finite element method for three-dimensional consolidation analysis. *Comput. Geotech.*, Vol.28, No.1, pp. 55-78.

Thomas, T.J.; Nair, S. & Garg, V.K. (1983). Elasto-plastic stress analysis and fatigue life prediction of a freight car wheel under mechanical and cyclic thermal loads. *Comp. Struct.*, Vol.17, No.3, pp. 313-320.

Timoshenko, S.P. & Goodier, J.N. (1951). *Theory of elasticity* (2nd ed.), McGraw-Hill, USA.

Wilson, E.L. (1965). Structural analysis of axisymmetric solids. *AIAA Journal*, Vol.3, No.12, pp. 2269-2274.

Zienkiewicz, O.C. & Cheung, Y.K. (1965). Finite element in the solution of field problems. *The Engineer*, Vol. 24, pp. 507-510.

Zienkiewicz, O.C. & Cheung, Y.K. (1967). Stresses in shaft. *The Engineer*, Vol. 24, pp. 696-697.

Zienkiewicz, O.C. & Taylor, R.L. (2000). *The finite element method* (5th ed.), Butterworth-Heinemann, ISBN 0-7506-5049-4, Oxford.

5

Optimization of the Dynamic Behaviour of Complex Structures Based on a Multimodal Strategy

Sébastien Besset and Louis Jézéquel

École Centrale de Lyon
France

1. Introduction

The use of a modal approach to describe a structure from the standpoint of optimizing its dynamic behavior offers multiple advantages. Once modal matrices have been computed, optimization criteria can be readily defined. Both the dynamic amplification phenomena and dynamic coupling between substructures can then be described using just a small number of degrees of freedom. Furthermore, it becomes possible to link the criteria to the modal parameters used in the systemic procedure. In this chapter, we will propose optimization criteria based on a multimodal description of complex structures.

The modal synthesis technique presented herein is based on the double and triple-modal synthesis proposed by Besset & Jézéquel (2008a;d), as well as on classical component mode synthesis methods like those developed by Craig & Bampton (1968) or Hurty (1965). According to these modal synthesis techniques, many boundary degrees of freedom are capable of remaining; in such cases, numerical costs will also remain high. In order to avoid a high-cost situation, we are proposing generalized modal synthesis methods that operate by introducing generalized boundary coordinates in order to describe substructure connections: this procedure is called a "double modal synthesis".

In addition, we are proposing another procedure to analyze structures coupled with fluid. This second procedure will then be called "triple modal synthesis". The first modal synthesis is classical; it consists of representing the interior points of the fluid by acoustic modes. When considering a formulation in force, the pressure on boundary points is set equal to zero. Using a formulation in displacement, cavity modes are introduced, generating a correspondence to the free modes of a structure. The second modal synthesis consists of describing the boundary forces between the fluid and each substructure through use of a set of loaded modes. Lastly, the third modal synthesis consists of describing the boundary forces between each substructure by introducing another set of loaded modes.

Complex structures often include hollow parts and stiffeners, both of which require very accurate analysis in order to obtain satisfactory results. In this chapter, the term "hollow parts" will denote the formed steel and stiffeners that make up the skeleton of a structure. In complex structures such as automobiles, stiffeners and formed steel parts, which compose the skeleton of the structure, these parts are most responsible for overall structural behavior. To analyze these elements, the method used is the one proposed in Besset & Jezequel (2008b).

This study will focus on the acoustic parts of the coupled system using acoustic modes and based on a "triple modal synthesis method", which relies on a coupling formulation previously investigated by Morand & Ohayon (1995) and Ohayon (2001; 2003). An example of the modal analysis of a coupled system can be found in Sandberg et al. (2001). A modal analysis of the structure will yield modal mass and stiffness matrices, which can then be used to obtain effective modal parameters and in turn lead to criteria that allow optimizing the structure. These criteria will depend on the pressure values at points located in the acoustic parts of the system, e.g. inside a car, as a function of an excitation point located on a hollow part of the structure, e.g. a spar near the car engine. The criteria proposed herein allow for various vibrational propagation paths to be considered. It is thus possible to separately investigate the various noise sources within the structure.
The method proposed in this chapter may be used for any vibro-acoustic system. In fact, the ultimate goal of this approach is to define modal criteria that allow optimizing the vibro-acoustic system. Such criteria are related to the coupling terms between the systems' various substructures and are expressed as functions of the terms contained in the modal matrices. The number of criteria therefore depends on the number of substructures within the vibro-acoustic system. While our method will be described for the case of a specific system, keep in mind that it can be readily adapted to other vibro-acoustic systems.
The structure considered in this chapter is complex and comprises acoustic cavities, hollow parts and plates. The geometry of this structure is similar to that of a car. Two cavities will be considered. Consequently, this proposed method can be used to study all of the different paths through which vibrations propagate and generate noise within the car. Paths can exist through the hollow parts of the structure, as well as through the plates bounding the cavity and through the plates partitioning the two cavities.

2. Multimodal methodology

The purpose of the modal synthesis method is to describe a structure using several of its modes, most often the first modes. Many authors have already proposed such methods, based on either fixed or free interface modes. Among these authors, let's mention Hurty (1965) and Craig & Bampton (1968). This method consists of using fixed-interface modes and static deformations to describe each substructure constituting the studied system. Goldman (1968) and Hou (1969) proposed another method based on the structure's free modes.
One of the main advantages of modal synthesis is to provide a description of a structure using very few degrees of freedom, thus reducing numerical costs. This advantage is particularly useful when seeking to optimize structures involving many calculations. Although modal models can be developed through experimental testing, we focus herein on substructure models resulting from a finite element discretization.
Most structures can be divided into several substructures. After being analyzed, these substructures are most often assembled through boundary degrees of freedom. For example, Craig & Bampton (1968) proposed a method using both fixed-interface modes and constraint modes that correspond to the boundary degrees of freedom. Hence, many degrees of freedom may remain should the boundary size be large. To avoid this type of problem, Besset & Jézéquel (2008a;d) proposed a generalization of the modal synthesis method by introducing generalized boundary coordinates. This procedure is referred to as "double modal synthesis" or, if the system contains fluid parts, "triple modal synthesis". These two techniques will be described respectively in Sections 2.1 and 2.2.

2.1 Double modal synthesis method

The *double modal synthesis methode* has been proposed and described through a continuous formulation by Besset & Jézéquel (2008a). In this section, we will explain the role of the discretized formulation. In section 2.1.1, we will introduce a primal formulation based on the free modes of the structure. Section 2.1.2 will apply a dual formulation that uses fixed-interface modes. In both sections 2.1.1 and 2.1.2, both Figure 1 and the following notations and the following notations will be considered valid:

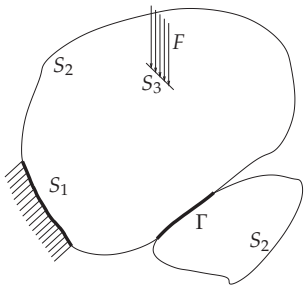


Fig. 1. Substructures and boundary conditions

- F is the force applied on the system;
- S_i are the boundaries of the system:
- Γ is the boundary between two substructures.

2.1.1 Primal formulation

The dynamic behaviour of the structure can be described through the following equation:

$$\left(-\omega^2 M + K\right) U = F \tag{1}$$

Considering the structure's free modes, U can be expressed as follows:

$$U = \phi_L q_L + \psi_r q_r + \psi_a q_a \tag{2}$$

where ϕ_L is the matrix of the structure's free modes, ψ_r the matrix of the n_r rigid body modes, and ψ_a is the attachment modes matrix associated with the free boundaries. ψ_a is obtained by imposing n_r boundary conditions, in order to suppress rigid body modes, while forces are applied on the free boundaries. If no rigid body modes are present, then ψ_a is the static flexibility matrix K^{-1}. If rigid body modes are present, then $\psi_a = A^T \psi_a^{\text{iso}} A$, where $A = I - M\psi_r M_{rr}^{-1} \psi_r^T$. ψ_a^{iso} is an isostatic flexibility matrix, obtained by imposing n_r boundary conditions on the structure and considering K_{iso}^{-1}. This method is intended to easily identify an isostatic flexibility matrix, which is corrected by adding rigid body modes, like for the finite element formulation of the Mac Neal method.

Equation 2 is then written considering the *double modal synthesis* theory, i.e. by expressing the boundary degrees of freedom as a fonction of k "branch modes" as follows:

$$U = \phi_L q_L + \psi_r q_r + \phi_b q_b \tag{3}$$

Matrix ϕ_b is the matrix of "branch modes", which are derived using the full structure equation condensed on interfaces. Let P be the matrix yielding generalized coordinates q_i as a function of displacements U (i.e. $U = P\{q_i\}$). Equation 1 then becomes:

$$\left[-\omega^2\left(P^T M P\right)+\left(P^T K P\right)\right]\left\{\begin{array}{c} q_L \\ q_r \\ q_b \end{array}\right\}=P^T F \tag{4}$$

2.1.2 Dual formulation

Like for the primal formulation described in section 2.1.1, the discretized equation of motion 1 is used herein:

$$\left(-\omega^2 M+K\right)U=F \tag{5}$$

In this section, U is expressed using fixed-interface modes:

$$U=\phi_F q_F+\psi_c q_c \tag{6}$$

As was the case for the discretization of the Craig and Bampton method Craig & Bampton (1968), matrices M and K are split into internal and boundary degrees of freedom, as follows:

$$U=\left\{\begin{array}{c} U_I \\ U_B \end{array}\right\},\quad M=\left[\begin{array}{cc} M_{II} & M_{IB} \\ M_{BI} & M_{BB} \end{array}\right],\quad K=\left[\begin{array}{cc} K_{II} & K_{IB} \\ K_{BI} & K_{BB} \end{array}\right] \tag{7}$$

The static flexibility can be written as K_{II}^{-1} and the constraint modes matrix is $\phi_c = K_{II}^{-1}K_{IB}$. The *double modal synthesis* theory is intended to express the remaining boundary degrees of freedom as a function of "branch modes" obtained by solving the following equation:

$$\left\{-\left[\omega^2\right]\left(P^T M P\right)+\left(P^T K P\right)\right\}\phi_D=0 \tag{8}$$

where $\left[\omega^2\right]$ is a diagonal matrix of eigenvalues corresponding to the columns of matrix ϕ_D. P is a transfer matrix that allows expressing U as a fuction of the generalized degrees of freedom q_F and q_c. Lastly, q_c can be expressed as $q_c = \phi_D q_b$ where q_b are the generalized boundary degrees of freedom.

2.2 Triple modal synthesis method

The *triple modal synthesis methode* has been proposed and described through a continuous formulation by Besset & Jézéquel (2008d). In this section, we will explain the discretized formulation. The structure and the notations used are recalled in figure 2.
We consider the discretized formulation of the coupled fluid-structure system, which leads to the following equations:

$$\left(-\omega^2 M+K\right)U=F \tag{9}$$

$$\left(-\omega^2 M_a+K_a\right)p=0 \tag{10}$$

Equation 9 pertains the structual part of the system, whereas Equation 10 relates to the acoustical part. This mixed formulation leads to the following equation:

$$\left(-\omega^2\underbrace{\left[\begin{array}{cc} M & 0 \\ C^T & M_a \end{array}\right]}_{\tilde{M}}+\underbrace{\left[\begin{array}{cc} K & -C \\ 0 & K_a \end{array}\right]}_{\tilde{K}}\right)\left\{\begin{array}{c} U \\ p \end{array}\right\}=\left\{\begin{array}{c} F \\ 0 \end{array}\right\} \tag{11}$$

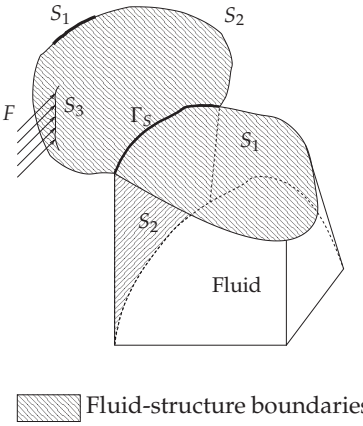


Fig. 2. Description of the triple modal synthesis

The *triple synthesis method* is performed by successively applying three modal syntheses. First, the behaviour of the fluid is expressed through "free modes" as follows:

$$p = \phi_p q_p \tag{12}$$

A second modal synthesis is then performed on the structural part of the system. Hence, the internal degrees of freedom, denoted U_i are expressed as follows:

$$U_i = \phi_i q_i + \psi_i u_j \tag{13}$$

where u_j are the boundary degrees of freedom. Lastly, a third modal synthesis is performed on the boundary degrees of freedom, i.e. degrees of freedom u_j are expressed as follows:

$$U_j = \phi_j q_j \tag{14}$$

Given Equations 12, 13 and 14, we can now define the following transfer matrix T:

$$\begin{Bmatrix} p \\ U_i \\ U_j \end{Bmatrix} = T \begin{Bmatrix} q_p \\ q_i \\ q_j \end{Bmatrix} \tag{15}$$

Equation 11 becomes:

$$\left[-\omega^2 \left(T^T \tilde{M} T\right) + \left(T^T \tilde{K} T\right)\right] \begin{Bmatrix} q_p \\ q_i \\ q_j \end{Bmatrix} = T^T F \tag{16}$$

3. Modal criteria and optimization

Optimization problems often lead to high calculation costs since the minimization function must be evaluated many times. The aim of this section is to propose modal-based criteria in order to optimize the dynamic behavior of complex structures. The examples provided herein will be applied on a schematic vehicle geometry 3. The structural case will be analyzed first in Section 3.1. Optimization criteria are linked to the "vibration paths" between substructures. The fluid-structure coupling case will then be presented in Section 3.2.

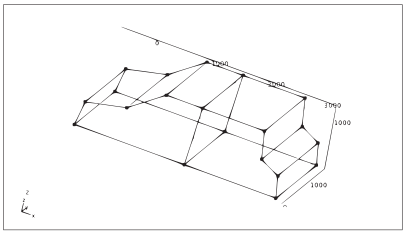


Fig. 3. Structure to be optimized

3.1 Criteria based on double modal synthesis method

This section presents the optimization criteria based on the double modal synthesis method derived in Section 2.1. These criteria are close to the modal matrices used for the modal synthesis of the structure. Once the structure has been analyzed, it therefore becomes very straightforward to calculate the optimization criteria. We will begin by explaining how to derive the modal-based criteria and indicating their use in the optimization step. Afterwards, we will demonstrate the efficiency of these parameters by optimizing the structure shown in Figure 3.

3.1.1 Modal-based criteria

The criteria proposed in this section are based on modal parameters that link degrees of freedom subjected to a displacement (so-called excited degrees of freedom) with the degrees of freedom whose displacements are to be minimized. In order to obtain these modal parameters, we will apply the double modal synthesis method proposed in Section 2.1. However, we will proceed by separating the degrees of freedom that will be excited; these degrees of freedom will remain nodal.

Let's now introduce the following notations concerning the degrees of freedom of the structure:

- u_P are the degrees of freedom relative to the plates;
- u_{Hb} are the degrees of freedom relative to the boundaries between plates.

The excitations will be relevant to the some degrees of freedom we have denoted u_{Hb}. These degrees of freedom will be denoted u_{He}. The double modal synthesis theory then leads to the expression in Equation 17:

$$\begin{cases} u_P = \Phi_P q_P + \Psi_P u_{Hb} + \Psi_{Pe} u_{He} \\ u_{Hb} = \Phi_{Hb} q_{Hb} \end{cases} \tag{17}$$

Thanks to the orthogonal properties of the modes used in the modal analysis, the motion equation can now be written as follows, in considering the damping matrix that we have assumed to be diagonal:

$$\left(-\omega^2 \begin{bmatrix} M_{EE} & M_{EHb} & M_{EP} \\ M_{HbE} & m_{Hbk} & M_{HbP} \\ M_{PE} & M_{PHb} & m_{Pk} \end{bmatrix} + i\omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_{Hbk} & 0 \\ 0 & 0 & c_{Pk} \end{bmatrix} + \begin{bmatrix} K_{EE} & K_{EHb} & 0 \\ K_{HbE} & k_{Hbk} & 0 \\ 0 & 0 & k_{Pk} \end{bmatrix} \right) \begin{Bmatrix} u_{He} \\ q_{Hb} \\ q_P \end{Bmatrix} = \begin{Bmatrix} \overline{f}_E \\ \overline{f}_{Hb} \\ \overline{f}_P \end{Bmatrix} \tag{18}$$

where matrices $[m_{Hbk}]$, $[m_{Pk}]$, $[k_{Hbk}]$, $[k_{Pk}]$, $[c_{Hbk}]$, $[c_{Pk}]$ and $[c_{Hck}]$ are diagonal matrices.
To obtain the modal parameters, u_P must be expressed as a function of f_P and u_{He}. Let's express one of the last rows in equation 34:

$$-\omega^2\left(M_{PE}^k u_{He} + M_{PHb}^k q_{Hb} + m_{Pk} q_P^k\right) + i\omega c_{Pk} q_P^k + k_{Pk} q_P^k = \overline{f}_P^k \tag{19}$$

M_{PE}^k, M_{PHb}^k are the k^{th} rows of matrices M_{PE}, M_{PHb}. $\overline{f}_P$ can be expressed as a function of f_P as follow:

$$\overline{f}_P = \Phi_P^T f_P \tag{20}$$

Equation 19 then becomes:

$$q_P^k = \frac{\Phi_P^{kT} f_P + \omega^2\left(M_{PE}^k u_{He} + M_{PHb}^k q_{Hb}\right)}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}} \tag{21}$$

where Φ_P^k is the k^{th} column of Φ_P. Equation 21 can now be written as:

$$\begin{aligned} u_P &= \sum_k \Phi_P^k q_{Pk} + \Psi_P u_{Hb} + \Psi_{Pe} u_{He} \\ &= \sum_k \left(\frac{\Phi_P^k \Phi_P^{kT}}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}}\right) f_P \\ &\quad + \left[\sum_k \left(\frac{\omega^2 \Phi_P^k M_{PE}^k}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}}\right) + \Psi_{Pe}\right] u_{He} \\ &\quad + \left[\sum_k \left(\frac{\omega^2 \Phi_P^k \tilde{M}_{PHb}^k}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}}\right) + \Psi_P\right] u_{Hb} \end{aligned} \tag{22}$$

where $M_{PHb}^k = \tilde{M}_{PHb}^k \Phi_{Hb}$.
Two modal parameters can be deduced from Equation 22. First, the dynamic flexibility matrix G is given by:

$$G(\omega) = \sum_k \left(\frac{\Phi_P^k \Phi_P^{kT}}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}}\right) \tag{23}$$

This first parameter corresponds to the relation between a force applied on a plate and the displacements it generates.
Secondly, by ignoring the static terms corresponding to the boundaries, the transmissibility matrix T is given by:

$$T(\omega) = \sum_k \left(\frac{\omega^2 \Phi_P^k M_{PE}^k}{-\omega^2 m_{Pk} + i\omega c_{Pk} + k_{Pk}}\right) \tag{24}$$

This equation corresponds to the relation between excitation and the displacements it generates on the plate. Note that this excitation is a displacement excitation. This expression can be rewritten using parameters $\tilde{G}$ and $\tilde{T}$ as follows:

$$G(\omega) = \sum_k \frac{1}{1 - \left(\frac{\omega}{\omega_k}\right)^2 + 2i\xi_k \frac{\omega}{\omega_k}} \tilde{G}_k \tag{25}$$

$$T(\omega) = \sum_k \frac{\left(\frac{\omega}{\omega_k}\right)^2}{1 - \left(\frac{\omega}{\omega_k}\right)^2 + 2i\xi_k \frac{\omega}{\omega_k}} \tilde{T}_k \tag{26}$$

where

$$\tilde{G}_k = \frac{\Phi_P^k \Phi_P^{kT}}{\omega_k^2 m_{Pk}} \tag{27}$$

$$\tilde{T}_k = \frac{\Phi_P^k M_{PE}^k}{m_{Pk}} \tag{28}$$

with the notations $c_{Pk} = 2\xi_k \sqrt{k_{Pk} m_{Pk}}$ and $\omega_k = \sqrt{\frac{k_{Pk}}{m_{Pk}}}$. Matrices $\tilde{G}_k$ and $\tilde{T}_k$ are referred to as modal parameters.

3.1.2 Example of optimization using modal parameters

In this section, we will deduce criteria from the flexibility and transmissibility matrices proposed in section 3.1.1. The sums $\sum_k ()$ appearing in these matrices correspond to a mode superposition. The optimization criteria can thus be written as follows:

$$C_G = \max_k \left| \frac{\Phi_P^k \Phi_P^{kT}}{\omega_k^2 m_{Pk}} \right| \tag{29}$$

$$C_T = \max_k \left| \frac{\Phi_P^k M_{PE}^k}{m_{Pk}} \right| \tag{30}$$

where the norm $|x|$ is the maximum component of matrix x. In considering these criteria, it is possible to optimize the structure. Moreover, obtaining the value $k_{\max}$ informs which mode is responsible for the value of the criteria.

The method proposed in this section has been tested on a complex structure that includes hollow parts and plates. The parameters we have chosen to optimize are correlated with the geometry of the hollow parts. We will in fact be optimizing D and λ, as shown in figure 4.

The optimization methods used in the next sections will introduce both D and λ as parameters. The hollow parts of the structure shown in figure 8 are split into 8 parts, and each part is optimized with optimal values of D and λ. Thus, 16 parameters are to be optimized.

The analysis of criteria C_G and C_T indicates which modes are responsible for the displacements that need to be reduced. Figure 6 shows the values of C_G^k, which are part of the criterion C_G:

$$C_G^k = \left| \frac{\Phi_P^k \Phi_P^{kT}}{\omega_k^2 m_{Pk}} \right| \tag{31}$$

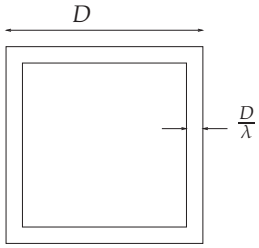


Fig. 4. Hollow part included in the structure

The same analysis can now be conducted for C_T^k:

$$C_T^k = \left| \frac{\Phi_P^k M_{PE}^k}{m_{Pk}} \right| \tag{32}$$

Figure 5 displays the values of C_T^k, which are part of the criterion C_T.

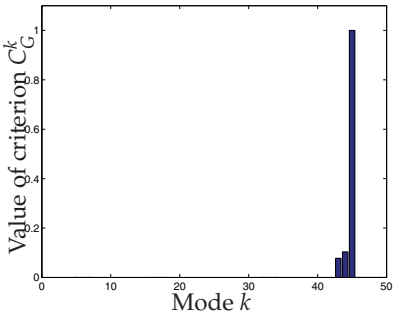


Fig. 5. Values of C_G^k

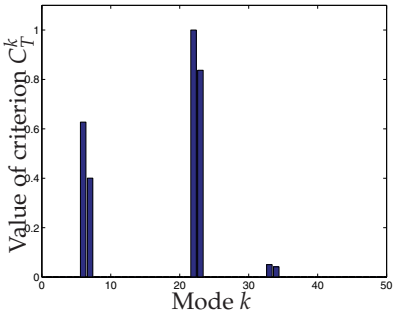


Fig. 6. Values of C_T^k

Figures 5 and 6 reveal that criteria C_G and C_T do not necessarily depend on the same modes. In this example, we only analyze the criteria for the first 50 modes of the structure. Figure 5 shows that the 45^{th} mode is mainly responsible for the value of criterion C_G, whereas Figure 6 suggests that the 6^{th}, 7^{th}, 22^{th} and 23^{th} modes are responsible for the value of criterion C_T. It is therefore necessary to take these two criteria into account in order to optimize the structure.

In this section, we will present the results obtained using a classical genetic algorithm. Figure 7 shows the Pareto diagram of the sets (λ_i, D_i). The units on x and y axes are not important since they depend on the values of c_g and c_t.

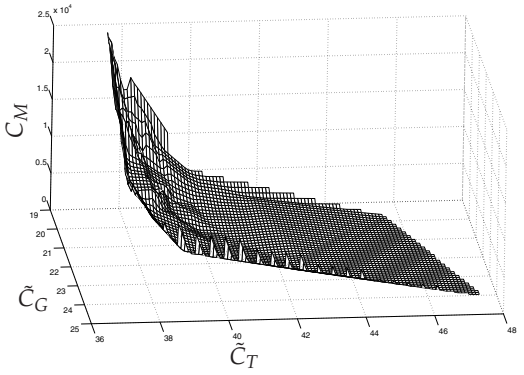


Fig. 7. Pareto points

3.2 Criteria based on triple modal synthesis method

In this section, we will condider the acoustical part inside the structure. The criteria presented here are similar to those from section 3.1, although they take into account this fluid part.

3.2.1 Modal-based criteria

The structure considered herein is complex and includes hollow parts and plates. It has been built using formed steel, which makes up its skeleton (as presented in Figure 8). The plates are fixed to this skeleton and form two cavities inside the structure (see the figure). Like in Section 3.1, the structure's geometry is similar to that of a car, for the purpose of highlighting that the proposed methods can be applied to an industrial context.

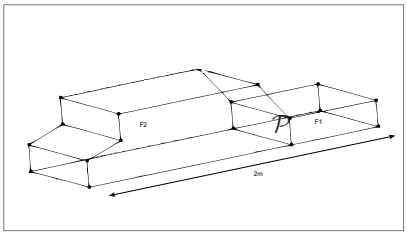


Fig. 8. The structure to be optimized

Figure 8 shows the geometry of the vibro-acoustic system under consideration. It is composed of plates and two fluid cavities. The degrees of freedom denoted u_{P_1} correspond to the plates, with the exception of the plates denoted $\mathcal{P}$ in figure 8, the degrees of freedom denoted u_{P_2} correspond to plate $\mathcal{P}$ in figure 8. Moreover, the degrees of freedom denoted p_{F_1} (resp. p_{F_2}) correspond to the pressure in the first cavity, as labeled $\mathcal{F}_1$ in figure 8 (resp. the second cavity, $\mathcal{F}_2$ in figure 8). The approach adopted in this section to study the vibro-acoustic behavior is a (u, p) formulation. Like in Section3.1.1, generalized degrees of freedom are correlated with nodal degrees of freedom through the following equations:

$$\begin{cases} u_{Hb} = \Phi_{Hb} q_{Hb} \\ u_{P_1} = \Phi_{P1} q_{P1} + \Psi_{P1} u_{Hb} + \Psi_{P1e} u_E \\ u_{P_2} = \Phi_{P2} q_{P2} + \Psi_{P2} u_{Hb} + \Psi_{P2e} u_E \\ p_{F_1} = \Phi_{F1} q_{F1} \\ p_{F_2} = \Phi_{F2} q_{F2} \end{cases} \tag{33}$$

According to the *triple modal synthesis* theory, as explained in section 2.2, the motion equation can be written as follows:

$$\left(-\omega^2 \begin{bmatrix} \overline{M}_{EE} & \overline{M}_{EHc} & 0 & \overline{M}_{EP1} & \overline{M}_{EP2} & 0 & 0 \\ \overline{M}_{HcE} & m_{Hc} & \overline{M}_{HcHb} & 0 & 0 & 0 & 0 \\ 0 & \overline{M}_{HbHc} & m_{Hb} & \overline{M}_{HbP1} & \overline{M}_{HbP2} & 0 & 0 \\ \overline{M}_{P1E} & 0 & \overline{M}_{P1Hb} & m_{P1} & 0 & 0 & 0 \\ \overline{M}_{P2E} & 0 & \overline{M}_{P2Hb} & 0 & m_{P2} & 0 & 0 \\ \overline{M}_{F1E} & 0 & \overline{M}_{F1Hb} & \overline{M}_{F1P1} & \overline{M}_{F1P2} & m_{F1} & 0 \\ \overline{M}_{F2E} & 0 & \overline{M}_{F2Hb} & \overline{M}_{F2P1} & \overline{M}_{F2P2} & 0 & m_{F2} \end{bmatrix} \right.$$

$$+ i\omega \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_{Hc} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_{Hb} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{P1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{P2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{F1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_{F2} \end{bmatrix}$$

$$\left. + \begin{bmatrix} \overline{K}_{EE} & 0 & 0 & 0 & 0 & \overline{K}_{EF1} & \overline{K}_{EF2} \\ 0 & k_{Hc} & 0 & 0 & 0 & \overline{K}_{HcF1} & \overline{K}_{HcF2} \\ 0 & 0 & k_{Hb} & 0 & 0 & \overline{K}_{HbF1} & \overline{K}_{HbF2} \\ 0 & 0 & 0 & k_{P1} & 0 & \overline{K}_{P1F1} & \overline{K}_{P1F2} \\ 0 & 0 & 0 & 0 & k_{P2} & \overline{K}_{P2F1} & \overline{K}_{P2F2} \\ 0 & 0 & 0 & 0 & 0 & k_{F1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_{F2} \end{bmatrix} \right) \begin{Bmatrix} u_E \\ q_{Hc} \\ q_{Hb} \\ q_{P1} \\ q_{P2} \\ q_{F1} \\ q_{F2} \end{Bmatrix} = \begin{Bmatrix} \overline{f}_E \\ \overline{f}_{Hc} \\ \overline{f}_{Hb} \\ \overline{f}_{P1} \\ \overline{f}_{P2} \\ 0 \\ 0 \end{Bmatrix} \tag{34}$$

The purpose of this section is to express the pressure p_{F_2} in the second cavity as a function of the structural displacements. This expression will allow the various vibrational paths to be distinguished from one another. The problem must be studied using a systemic approach. Certain criteria related to the various vibrational paths will thus be defined. Using the last row of Equation 34, the following expression is obtained:

$$-\omega^2 \left(\overline{M}^k_{F2E} u_E + \overline{M}^k_{F2Hb} q_{Hb} + \overline{M}^k_{F2P1} q_{P1} + \overline{M}^k_{F2P2} q_{P2} + m^k_{F2} q^k_{F2} \right) + i\omega c^k_{F2} q^k_{F2} + k^k_{F2} q^k_{F2} = 0 \tag{35}$$

where $\overline{M}^k_{F2E}$, $\overline{M}^k_{F2Hb}$, $\overline{M}^k_{F2P1}$ and $\overline{M}^k_{F2P2}$ are the k^{th} rows of matrices $\overline{M}_{F2E}$, $\overline{M}_{F2Hc}$, $\overline{M}_{F2Hb}$, $\overline{M}_{F2P1}$ and $\overline{M}_{F2P2}$. q^k_{F2} is the k^{th} component of vector q_{F2}.
Equation 35 leads to:

$$q_{F2}^{k} = \frac{\omega^2 \left(\overline{M}_{F2E}^{k} u_E + \overline{M}_{F2Hb}^{k} q_{Hb} + \overline{M}_{F2P1}^{k} q_{P1} + \overline{M}_{F2P2}^{k} q_{P2} \right)}{-\omega^2 m_{F2}^{k} + i\omega c_{F2}^{k} + k_{F2}^{k}} \tag{36}$$

Now let Φ_{F2}^{k} be the k^{th} column of Φ_{F2}. Combining Equations 33 and 36 yields:

$$\begin{aligned} p_{F_2} &= \sum_k \Phi_{F2}^{k} q_{F2}^{k} \\ &= \sum_k \left[\frac{\omega^2 \Phi_{F2}^{k} \overline{M}_{F2E}^{k}}{-\omega^2 m_{F2}^{k} + i\omega c_{F2}^{k} + k_{F2}^{k}} \right] u_E \\ &\quad + \sum_k \left[\frac{\omega^2 \Phi_{F2}^{k} \overline{M}_{F2Hb}^{k}}{-\omega^2 m_{F2}^{k} + i\omega c_{F2}^{k} + k_{F2}^{k}} \right] \tilde{\Phi}_{Hb} u_{Hb} \\ &\quad + \sum_k \left[\frac{\omega^2 \Phi_{F2}^{k} \overline{M}_{F2P1}^{k}}{-\omega^2 m_{F2}^{k} + i\omega c_{F2}^{k} + k_{F2}^{k}} \right] \left(\tilde{\Phi}_{P1} \left(u_{P1} - \Psi_{P1} u_{Hb} - \Psi_{P1e} u_E \right) \right) \\ &\quad + \sum_k \left[\frac{\omega^2 \Phi_{F2}^{k} \overline{M}_{F2P2}^{k}}{-\omega^2 m_{F2}^{k} + i\omega c_{F2}^{k} + k_{F2}^{k}} \right] \left(\tilde{\Phi}_{P2} \left(u_{P2} - \Psi_{P2} u_{Hb} - \Psi_{P2e} u_E \right) \right) \end{aligned} \tag{37}$$

The quantities with a "tilde" are left pseudo-inverse matrices. It is possible to define several pseudo-inverse matrices. In the case of singular systems, pseudo-inverse matrices enable finding solutions (c.f. Farhat & Géradin (1998)). In the present case, the matrices are not square and the solution obtained is a least squares approximation, due to the fact that modal synthesis does not entail fewer modes than the number of physical degrees of freedom for the system.

Equation 37 provides an approximation to the pressure field p_{F_2} in the second cavity as a function of structural displacements. A superposition of the substructural modes clearly appears in the sums $\sum_k ( )$.

Equation 37 allows for the definition of modal parameters, which can then be used to optimize the coupled system. These parameters correspond to each of the various vibrational paths, i.e.: a direct path, a path through the hollow parts of the structure, a path through the plates bounding the cavity, and a path through the plate located between the two cavities.

3.2.1.1 Direct path

The direct path is directly obtained via Equation 37, it corresponds to an excitation point on a plate located next to cavity $\mathcal{F}_2$. Let's recall that the excitations considered herein are displacement excitations.

The modal parameter corresponding to this direct path is denoted $G_E(\omega)$:

$$G_E(\omega) = \sum_k \frac{\left(\frac{\omega}{\omega_k}\right)^2}{1 - \left(\frac{\omega}{\omega_k}\right)^2 + 2i\xi_k \frac{\omega}{\omega_k}} \tilde{G}_E^{k} \tag{38}$$

where:

$$\tilde{G}_E^k = \frac{\Phi_{F2}^k \left(\overline{M}_{F2E}^k - \overline{M}_{F2P1}^k \tilde{\Phi}_{P1} \Psi_{P1e} - \overline{M}_{F2P2}^k \tilde{\Phi}_{P2} \Psi_{P2e} \right)}{m_{P2}^k} \tag{39}$$

with the notations $c_{F2}^k = 2\xi_k \sqrt{k_{F2}^k m_{F2}^k}$ and $\omega_k = \sqrt{\frac{k_{F2}^k}{m_{F2}^k}}$.

3.2.1.2 Path through the boundaries

The path through the boundaries is given by component u_{Hb} of Equation 37.
q_{Hb}^k can now be expressed according to the third row of Equation 34:

$$\begin{aligned} -\omega^2 \left(\overline{M}_{HbHc}^k q_{Hc} + m_{Hb}^k q_{Hb}^k + \overline{M}_{HbP1}^k q_{P1} + \overline{M}_{HbP2}^k q_{P2} \right) \\ + i\omega c_{Hb}^k q_{Hb}^k + k_{Hb}^k q_{Hb}^k + \overline{K}_{HbF1}^k q_{F1} + \overline{K}_{HbF2}^k q_{F2} = \overline{f}_{Hb}^k \end{aligned} \tag{40}$$

The modal parameter corresponding to the path through the hollow parts is denoted $G_H(\omega)$. Using Equations 40 and 35 together, it is possible to write:

$$G_H(\omega) = \left(\sum_k \frac{\left(\frac{\omega}{\omega_{1k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{1k}}\right)^2 + 2i\xi_{1k}\frac{\omega}{\omega_{1k}}} \tilde{G}_{H1}^k \right) \left(\sum_k \frac{\left(\frac{\omega}{\omega_{2k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{2k}}\right)^2 + 2i\xi_{2k}\frac{\omega}{\omega_{2k}}} \tilde{G}_{H2}^k \right) \tag{41}$$

where:

$$\tilde{G}_{H1}^k = \frac{\Phi_{F2}^k \overline{M}_{F2Hb}^k \tilde{\Phi}_{Hb}}{m_{F2}^k} \tag{42}$$

$$\tilde{G}_{H2}^k = \frac{\Phi_{Hb}^k \left(\overline{M}_{HbHc}^k \tilde{\Phi}_{Hc} \Psi_{He} - \overline{M}_{HbP1}^k \tilde{\Phi}_{P1} \Psi_{P1e} - \overline{M}_{HbP2}^k \tilde{\Phi}_{P2} \Psi_{P2e} \right)}{m_{Hb}^k} \tag{43}$$

with the notations $c_{F2}^k = 2\xi_{1k}\sqrt{k_{F2}^k m_{F2}^k}$, $\omega_{1k} = \sqrt{\frac{k_{F2}^k}{m_{F2}^k}}$, and $c_{Hb}^k = 2\xi_{2k}\sqrt{k_{Hb}^k m_{Hb}^k}$ and $\omega_{2k} = \sqrt{\frac{k_{Hb}^k}{m_{Hb}^k}}$.

3.2.1.3 Path through the plates

The path through the plates is given by component u_{P1} of Equation 37.
u_{P1} can then be written as a function of u_E according to the fourth row of Equation 34:

$$\begin{aligned} -\omega^2 \left(\overline{M}_{P1E}^k u_E + \overline{M}_{P1Hb}^k q_{Hb} + m_{P1}^k q_{P1}^k \right) \\ + i\omega c_{P1}^k q_{P1}^k + k_{P1}^k q_{P1}^k + \overline{K}_{P1F1} q_{F1} + \overline{K}_{P1F2} q_{F2} = \overline{f}_{P1} \end{aligned} \tag{44}$$

The modal parameter corresponding to the path through the plates is denoted $G_{P1}(\omega)$. By combining Equations 44 and 35, it is now possible to write:

$$G_{P1}^{1}(\omega) = \left(\sum_k \frac{\left(\frac{\omega}{\omega_{1k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{1k}}\right)^2 + 2i\xi_{1k}\frac{\omega}{\omega_{1k}}} \tilde{G}_{P11}^{k} \right) \left(\sum_k \frac{\left(\frac{\omega}{\omega_{2k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{2k}}\right)^2 + 2i\xi_{2k}\frac{\omega}{\omega_{2k}}} \tilde{G}_{P12}^{k} \right) \tag{45}$$

$$\begin{aligned} G_{P1}^{2}(\omega) = & \left(\sum_k \frac{\left(\frac{\omega}{\omega_{1k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{1k}}\right)^2 + 2i\xi_{1k}\frac{\omega}{\omega_{1k}}} \tilde{G}_{P11}^{k} \right) \\ & \times \left(\sum_k \frac{\left(\frac{\omega}{\omega_{2k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{2k}}\right)^2 + 2i\xi_{2k}\frac{\omega}{\omega_{2k}}} \tilde{G}_{P13}^{k} \right) \\ & \times \left(\sum_k \frac{\left(\frac{\omega}{\omega_{3k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{3k}}\right)^2 + 2i\xi_{3k}\frac{\omega}{\omega_{3k}}} \tilde{G}_{P14}^{k} \right) \end{aligned} \tag{46}$$

where:

$$\tilde{G}_{P11}^{k} = \frac{\Phi_{F2}^{k} \overline{M}_{F2P1}^{k} \tilde{\Phi}_{P1}}{m_{F2}^{k}} \tag{47}$$

$$\tilde{G}_{P12}^{k} = \frac{\Phi_{P1}^{k} \overline{M}_{P1E}^{k}}{m_{P1}^{k}} \tag{48}$$

$$\tilde{G}_{P13}^{k} = \frac{\Phi_{P1}^{k} \overline{M}_{P1Hb}^{k} \tilde{\Phi}_{Hb}}{m_{P1}^{k}} \tag{49}$$

$$\tilde{G}_{P14}^{k} = \tilde{G}_{H2}^{k} \quad \text{(see Equation 43)} \tag{50}$$

with the notations $c_{F2}^{k} = 2\xi_{1k}\sqrt{k_{F2}^{k} m_{F2}^{k}}$, $\omega_{1k} = \sqrt{\frac{k_{F2}^{k}}{m_{F2}^{k}}}$, $c_{P1}^{k} = 2\xi_{2k}\sqrt{k_{P1}^{k} m_{P1}^{k}}$, $\omega_{2k} = \sqrt{\frac{k_{P1}^{k}}{m_{P1}^{k}}}$, $c_{Hb}^{k} = 2\xi_{3k}\sqrt{k_{Hb}^{k} m_{Hb}^{k}}$ and $\omega_{3k} = \sqrt{\frac{k_{Hb}^{k}}{m_{Hb}^{k}}}$.

3.2.1.4 Path through the first cavity

The path through the first cavity is related to the plate forming the boundary between the two cavities.

The modal parameter corresponding to this path can be written according to the fifth and sixth rows of Equation 34:

$$-\omega^2\left(\overline{M}^k_{P2E}u_E+\overline{M}^k_{P2Hb}q_{Hb}+m^k_{P2}q^k_{P2}\right)+i\omega c^k_{P2}q^k_{P2}+k^k_{P2}q^k_{P2}+\overline{K}^k_{P2F1}q_{F1}+\overline{K}^k_{P2F2}q_{F2}=\overline{f}_{P2} \quad (51)$$

$$-\omega^2\left(\overline{M}^k_{F1E}u_E+\overline{M}^k_{F1Hb}q_{Hb}+\overline{M}^k_{F1P1}q_{P1}+\overline{M}^k_{F1P2}q^k_{P2}+m^k_{F1}q_{F1}\right)+i\omega c^k_{F1}q^k_{F1}+k^k_{F1}q^k_{F1}=0 \quad (52)$$

The modal parameter corresponding to the path through the plates is denoted $G_{P2}(\omega)$. By combining Equations 51, 52 and 35, it becomes possible to write:

$$G_{P2}(\omega)=\left(\sum_k \frac{\left(\frac{\omega}{\omega_{1k}}\right)^2}{1-\left(\frac{\omega}{\omega_{1k}}\right)^2+2i\xi_{1k}\frac{\omega}{\omega_{1k}}}\tilde{G}^k_{P21}\right)\times\left(\sum_k \frac{1}{1-\left(\frac{\omega}{\omega_{2k}}\right)^2+2i\xi_{2k}\frac{\omega}{\omega_{2k}}}\tilde{G}^k_{P22}\right)\times\left(\sum_k \frac{\left(\frac{\omega}{\omega_{3k}}\right)^2}{1-\left(\frac{\omega}{\omega_{3k}}\right)^2+2i\xi_{3k}\frac{\omega}{\omega_{3k}}}\tilde{G}^k_{P23}\right) \quad (53)$$

where:

$$\tilde{G}^k_{P21}=\frac{\Phi^k_{F2}\overline{M}^k_{F2P2}\tilde{\Phi}_{P2}}{m^k_{F2}} \quad (54)$$

$$\tilde{G}^k_{P22}=\frac{\Phi^k_{P2}\overline{K}^k_{P2F1}\tilde{\Phi}_{F1}}{\omega^2_{2k}m^k_{P2}} \quad (55)$$

$$\tilde{G}^k_{P23}=\frac{\Phi^k_{F1}\left(\overline{M}^k_{F1E}-\overline{M}^k_{F1P1}\tilde{\Phi}_{P1}\Psi_{P1e}-\overline{M}^k_{F1P2}\tilde{\Phi}_{P2}\Psi_{P2e}\right)}{m^k_{F1}} \quad (56)$$

with the notations $c^k_{F2}=2\xi_{1k}\sqrt{k^k_{F2}m^k_{F2}}$, $\omega_{1k}=\sqrt{\frac{k^k_{F2}}{m^k_{F2}}}$, $c^k_{P2}=2\xi_{2k}\sqrt{k^k_{P2}m^k_{P2}}$, $\omega_{2k}=\sqrt{\frac{k^k_{P2}}{m^k_{P2}}}$, $c^k_{F1}=2\xi_{3k}\sqrt{k^k_{F1}m^k_{F1}}$ and $\omega_{3k}=\sqrt{\frac{k^k_{F1}}{m^k_{F1}}}$.

3.2.1.5 Modal criteria

The modal parameters defined in the previous sections now lead to the criteria defined as follows:

$$C_E = \max_k \left| \frac{\left(\frac{\omega}{\omega_k}\right)^2}{1 - \left(\frac{\omega}{\omega_k}\right)^2 + 2i\xi_k \frac{\omega}{\omega_k}} \tilde{G}_E^k \right| \tag{57}$$

$$C_n = \max_k \left| \frac{\left(\frac{\omega}{\omega_{1k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{1k}}\right)^2 + 2i\xi_{1k} \frac{\omega}{\omega_{1k}}} \tilde{G}_n^k \right| \quad \text{where } n \text{ can be } H1,\ P11 \text{ or } P21 \tag{58}$$

$$C_n = \max_k \left| \frac{\left(\frac{\omega}{\omega_{2k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{2k}}\right)^2 + 2i\xi_{2k} \frac{\omega}{\omega_{2k}}} \tilde{G}_n^k \right| \quad \text{where } n \text{ can be } H2,\ P12, \text{ or } P13 \tag{59}$$

$$C_{P14} = \max_k \left| \frac{\left(\frac{\omega}{\omega_{3k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{3k}}\right)^2 + 2i\xi_{3k} \frac{\omega}{\omega_{3k}}} \tilde{G}_{P14}^k \right| \tag{60}$$

$$C_{P22} = \max_k \left| \frac{1}{1 - \left(\frac{\omega}{\omega_{2k}}\right)^2 + 2i\xi_{2k} \frac{\omega}{\omega_{2k}}} \tilde{G}_{P22}^k \right| \tag{61}$$

$$C_{P23} = \max_k \left| \frac{\left(\frac{\omega}{\omega_{3k}}\right)^2}{1 - \left(\frac{\omega}{\omega_{3k}}\right)^2 + 2i\xi_{3k} \frac{\omega}{\omega_{3k}}} \tilde{G}_{P23}^k \right| \tag{62}$$

Within the framework of an optimization problem, it is possible to use these criteria, for example, in order to optimize a structural geometry. However, the criteria we are proposing do not allow for derivation with respect to any of the structural parameters (e.g. geometry of the hollow parts, plate thickness…), although many optimization methods require derivatives of these criteria. For this reason, other criteria related to the original set of criteria will be defined in the next section.

As indicated below, criteria C_n (where n can be E, $H1$, $H2$, $P11$, $P12$, $P13$, $P14$, $P21$, $P22$ or $P23$) cannot be derived with respect to any parameter, even though this would be useful in most optimization problems. To remedy this shortcoming, we introduce criteria $\tilde{C}_n$, defined as follows:

$$\tilde{C}_n = \frac{1}{4} \log \sum_k \left| \lambda(\omega, k)\, \tilde{G}_n^k \right|^4 \tag{63}$$

where $\lambda(\omega)$ is a coefficient that depends on ω. This is the case, for example, of $\tilde{G}_E^k$, $\lambda(\omega) = \frac{\left(\frac{\omega}{\omega_k}\right)^2}{1 - \left(\frac{\omega}{\omega_k}\right)^2 + 2i\xi_k \frac{\omega}{\omega_k}}$.

It has been demonstrated that these criteria are very similar to the first set and have nearly the same minima and maxima Besset & Jézéquel (2008c).

3.2.2 Analysis of the criteria

In this section, the criteria developed in the previous sections will be analyzed. In order to simplify the analysis, only one excitation point placed on a hollow part bounding the first cavity $\mathcal{F}_1$ is considered.

The values of criteria C_n may change with the excitation frequency due to the coefficient $\lambda(\omega)$. It can be observed that the strength of each criterion depends on the excitation frequency. For example, it is interesting to note that criterion $CP22$ increases more strongly than the other criteria with frequency f. Therefore, the vibrational path through plate $\mathcal{P}$ becomes very significant at higher frequencies.

NLet's also note that criteria $CP13$ and $CH2$ show peaks that correspond to the global modes of the structure, since the modal matrices of the structure's hollow parts are involved in these criteria.

Modal parameters $\tilde{G}^k_{H1}$, $\tilde{G}^k_{H2}$, $\tilde{G}^k_{P21}$, $\tilde{G}^k_{P22}$, $\tilde{G}^k_{P23}$, $\tilde{G}^k_{P11}$, $\tilde{G}^k_{P13}$ and $\tilde{G}^k_{P14}$ have all been analyzed. This step will allow each of the modes responsible for the values of criteria C_n to be defined. The modal parameters must be weighted with the previously defined coefficients $\lambda(\omega)$, in order to take the excitation frequency into account. Two cases of excitation are presented, i.e. at 50 Hz and at 300 Hz.

3.2.2.1 Path through the first cavity

Considering the vibrational path through the first cavity and plate $\mathcal{P}$, three modal parameters need to be analyzed. Figures 9, 10 and 11 display the values of modal parameters $\tilde{G}^k_{P21}$, $\tilde{G}^k_{P22}$ and $\tilde{G}^k_{P23}$ as a function of the mode number k, for an excitation frequency of 50 Hz. Figures 12, 13 and 14 show these same parameters for an excitation frequency of 300 Hz.

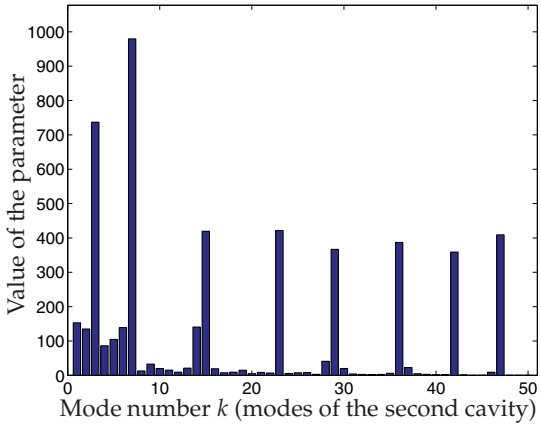


Fig. 9. Values of parameter $\left|\lambda \tilde{G}^k_{P21}\right|$ as a function of k – 50 Hz

Figure 10 shows that only one mode of plate $\mathcal{P}$ is responsible for the transmission of vibrations between the first and second cavities. The criterion associated with this figure is C_{P22}, which is related to the influence of pressure in the first cavity on plate $\mathcal{P}$. Given that the action of an acoustic fluid on a structure is not very significant when compared with that of a structure on a fluid, the parameter values given in figure 10 remain very small. For an influence to be exerted on the path through the first cavity, which is the objective of this section, one could, for example, restrict the influence of the mode presented in figure 10.

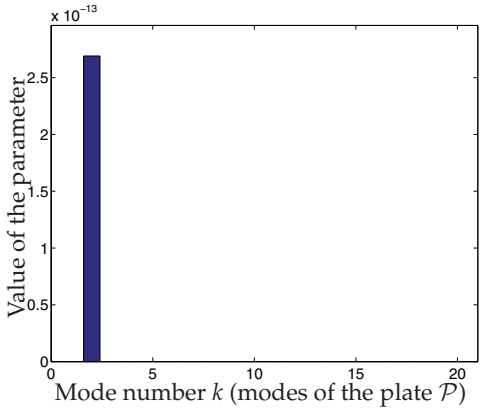


Fig. 10. Values of parameter $\left|\lambda \tilde{G}^k_{P22}\right|$ as a function of k – 50 Hz

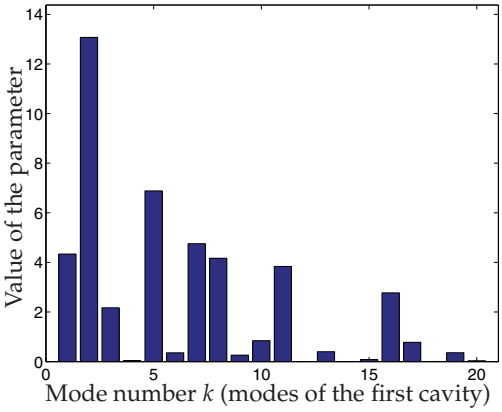


Fig. 11. Values of parameter $\left|\lambda \tilde{G}^k_{P23}\right|$ as a function of k – 50 Hz

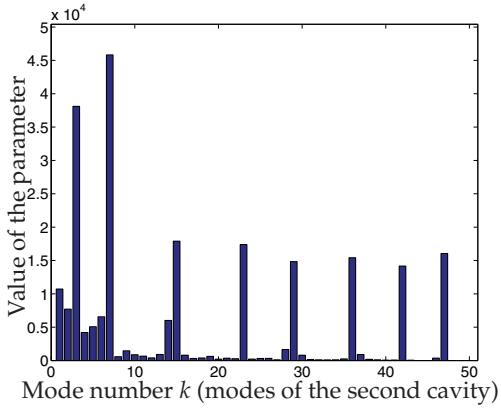


Fig. 12. Values of parameter $\left|\lambda \tilde{G}^k_{P21}\right|$ as a function of k – 300 Hz

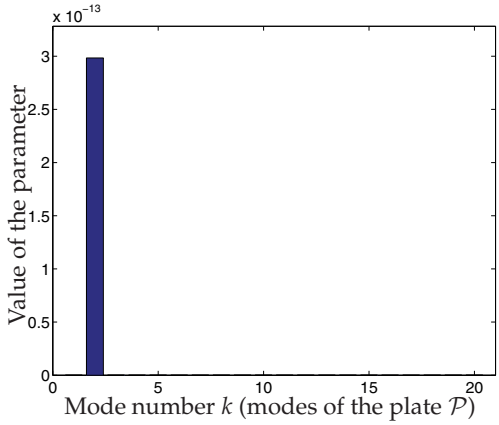


Fig. 13. Values of parameter $\left|\lambda \tilde{G}^k_{P22}\right|$ as a function of k – 300 Hz

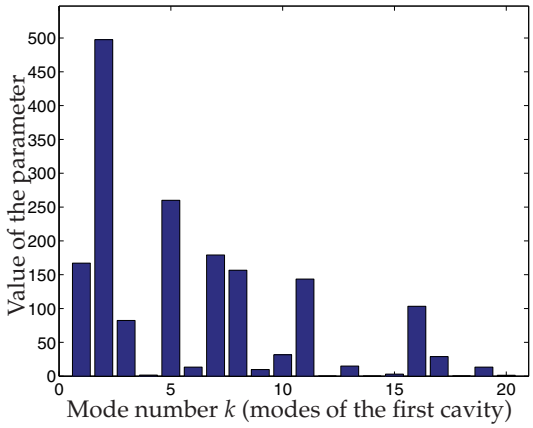


Fig. 14. Values of parameter $\left|\lambda \tilde{G}^k_{P23}\right|$ as a function of k – 300 Hz

3.2.2.2 Path through the hollow parts

For the vibrational path through the hollow parts, two modal parameters need to be analyzed. Figures 15 and 16 show the values of modal parameters $\tilde{G}^k_{H1}$ and $\tilde{G}^k_{H2}$ as a function of mode number k for an excitation frequency of 50 Hz. Figures 15 and 16 show these same parameters for an excitation frequency of 300 Hz.

In figures 15 and 16, it can be seen that many of the modes are strong. The values of these parameters increase with excitation frequency. However, the values for criterion C_{H2} are much smaller than those for criterion C_{H1}. It then becomes possible to elect to optimize the structure using C_{H1}, which seems to exert a stronger influence on the transmission of vibrations. Although the influence of criterion C_{H2} appears to be smaller, the use of this criterion in structural optimization allows studying the vibrational path through the plates, as described in the following section.

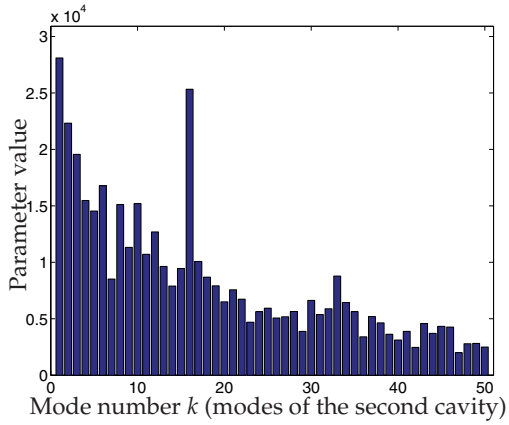


Fig. 15. Values of parameter $\left|\lambda \tilde{G}^k_{H1}\right|$ as a function of k – 50 Hz

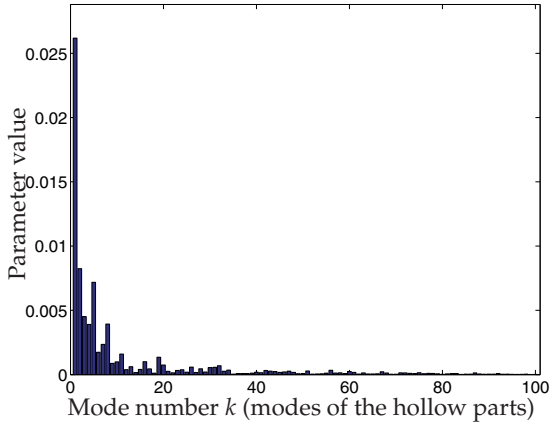


Fig. 16. Values of parameter $\left|\lambda \tilde{G}^k_{H2}\right|$ as a function of k – 50 Hz

3.2.2.3 Path through the hollow parts and the plates

For the vibrational path through the hollow parts and the plates, three modal parameters need to be analyzed. Figures 19 and 20 show the values of modal parameters $\tilde{G}^k_{P11}$ and $\tilde{G}^k_{P13}$ as a function of mode number k for an excitation frequency of 50 Hz. It should be noted that $\tilde{G}^k_{P14} = \tilde{G}^k_{H2}$, as plotted in figure 16. Figures 21 and 22 provide the values of modal parameters $\tilde{G}^k_{P11}$ and $\tilde{G}^k_{P13}$ as a function of mode number k for an excitation frequency of 300 Hz.

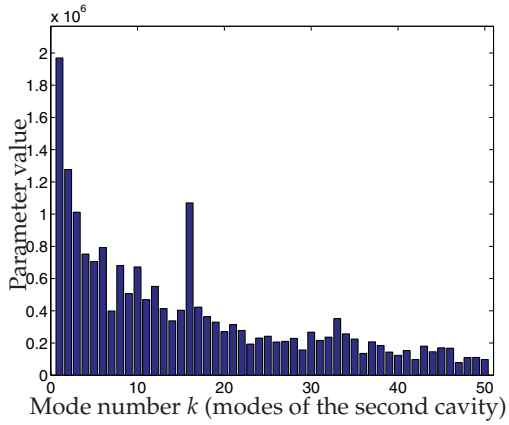


Fig. 17. Values of parameter $\left|\lambda \tilde{G}_{H1}^{k}\right|$ as a function of k – 300 Hz

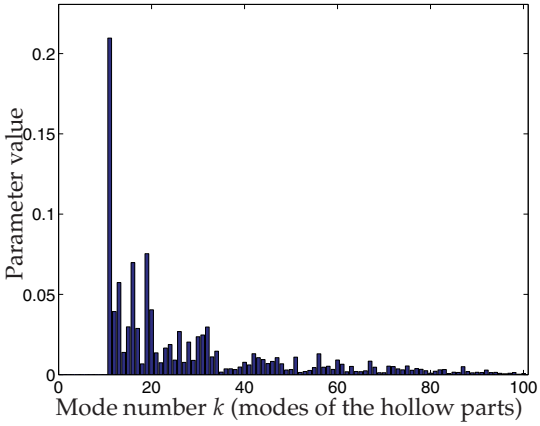


Fig. 18. Values of parameter $\left|\lambda \tilde{G}_{H2}^{k}\right|$ as a function of k – 300 Hz

As reflected in Figures 19 and 20, many modes are strong. Just as in the previous section, it is then possible to optimize the structure using criterion $C_{P14} = C_{H2}^{k}$, so as to minimize the transmission of vibrations through two different paths. It is also possible to minimize criterion C_{P11}, which is greater than C_{P13}.

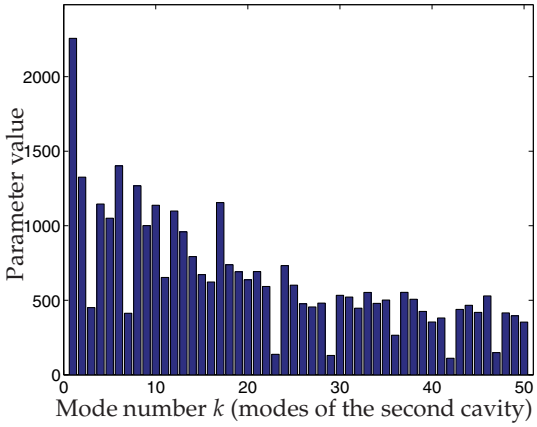


Fig. 19. Values of parameter $\left|\lambda \tilde{G}^k_{P11}\right|$ as a function of k – 50 Hz

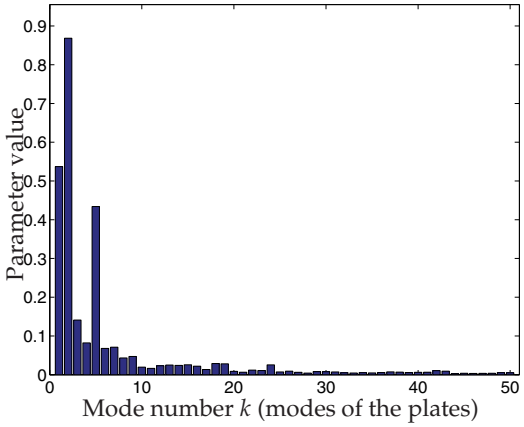


Fig. 20. Values of parameter $\left|\lambda \tilde{G}^k_{P13}\right|$ as a function of k – 50 Hz

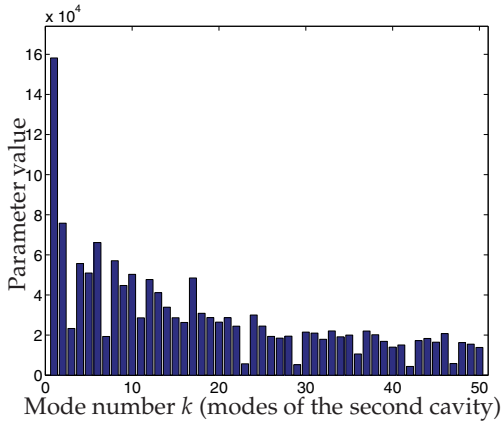


Fig. 21. Values of parameter $\left|\lambda \tilde{G}_{P11}^{k}\right|$ as a function of k – 300 Hz

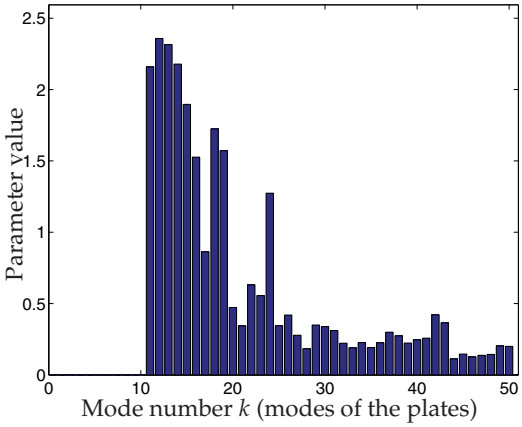


Fig. 22. Values of parameter $\left|\lambda \tilde{G}_{P13}^{k}\right|$ as a function of k – 300 Hz

4. Conclusion

In this chapter, several criteria, corresponding to different vibrational propagation paths, have been proposed. These criteria are based on modal motion equations, which allow for working with small-sized matrices. Consequently, calculation-related costs are not prohibitive during the optimization process, which is an important consideration whenever objective functions have to be evaluated many times. Moreover, a modal overview of the phenomena serves to illustrate the relative influence of each mode on noise propagation, in addition to indicating which part of the system exerts the strongest influence on noise generation inside the structure.

5. References

Besset, S. & Jézéquel, L. (2008a). Dynamic substructuring based on a double modal analysis, *J. Vib. Acoust* 130(1).

Besset, S. & Jezequel, L. (2008b). A modal analysis method to study fluid-structure coupling in hollow parts of a structure, *Journal of Computational Acoustics* 16(02): 257.

Besset, S. & Jézéquel, L. (2008c). Modal criteria for optimization of the acoustical behaviour of a vibroacoustic system based on a systemic approach, *International Journal for Numerical Methods in Engineering* 73: 1347–1373.

Besset, S. & Jézéquel, L. (2008d). Vibroacoustical analysis based on a multimodal strategy: Triple modal synthesis, *J. Vib. Acoust* 130(3).

Craig, R. R. & Bampton, M. C. C. (1968). Coupling of substructures for dynamic analysis, *AIAA Journal* 6: 1313–1321.

Farhat, C. & Géradin, M. (1998). On the general solution by a direct method on a large-scale singular system of linear equations: Application to the analysis of floating structures, *International journal for numerical methods in engineering* 41: 675–696.

Goldman, R. L. (1968). Vibration analysis by dynamic partitioning, *AIAA Journal* 7: 1152–1154.

Hou, S. N. (1969). Review of modal synthesis techniques and a new approach, *Shock and vibration bulletin* 40: 25–30.

Hurty, W. C. (1965). Dynamic analysis of structural systems using component modes, *AIAA Journal* 3: 678–685.

Morand, H. J. & Ohayon, R. (1995). *Fluid Structure Interaction*, Wiley & Sons.

Ohayon, R. (2001). Reduced symetric models for modal analysis of internal structural-acoustic and hydroelastic-sloshing systems, *Computer Methods in Applied Mechanics and Engineering* 190: 3009–3019.

Ohayon, R. (2003). Reduced models for fluid-structure interaction problems, *International Journal for Numerical Methods in Engineering* 60: 139–152.

Sandberg, G. E., Hansson, P.-A. & Gustavsson, M. (2001). Domain decomposition in acoustic and structure-acoustic analysis, *Computer Methods in Applied Mechanics and Engineering* 190: 2979–2988.

6

Numerical Simulation on Ecological Interactions in Time and Space

Kornkanok Bunwong
Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand

1. Introduction

The most formal and systematic tool to simplify a real-world phenomenon dealing with the interrelationships between organisms and their environment, including the interaction between each other is an ecological model. Of course, it is usually created for one or more purposes, for example, to gain system understanding, to forecast the future state of the system, and to develop new hypotheses. Not only that, a good model should also reach a balance between the complexities of the real-world system, which is too difficult to solve, and the need of simple formulation and valid analytic model, which can be explicitly solved. Ecological modeling has a long and rich history. So far many sophisticated ecological models have been developed in all fields of ecology such as the ecology of individuals including physiological ecology, the ecology of populations, and the study of ecosystems (classified by Hofbauer & Sigmund, 1988; Roughgarden, 1996). Ecological phenomenon can often be simplified by making some assumptions and studying it with suitable time scales and spatial interactions. It should be noted that a simple deterministic system could behave dramatically unlike when its time scale and its spatial interaction are changed. Therefore, the next coming sections aim to convince the significance of time scales and spatial interactions.

2. The classical concept of ecological modelling

The books on mathematical modeling in biology written by L. Edelstein-Keshet (1988), S. P. Ellner & J. Guckenheimer (2006), M. R. S. Kulenovic & O. Merino (2002), J. D. Murray (1993), and S. P. Otto & T. Day (2007) are the valuable teaching resources suitable for modelers who require theoretical concepts as well as desire mathematical and computational techniques to build and analyze models. In the early ecological models, their time scales were limited to either discrete or continuous. Usually, if populations of individuals have synchronized changes (i.e., reproduction, infection, recovery, migration, removal or mortality) at a regular time interval and no overlap between successive generations then the discrete time model is recommended. Otherwise, a continuous time model is preferred. Traditionally, a discrete time model is represented by using difference equation(s) whereas a continuous time model is constructed by using differential equation(s). Recently, the practical use of mathematical modeling is richly contributed. For example, L. J. S. Allen (1994), L. J. S. Allen & A. B. Burgin (2000), W. M. Getz & J. O. Lloyd-Smith (2006), and S. R.-J. Jang (2008) were interested in discrete epidemic models while K. Bunwong et al. (2009) studied nutrient removal process on a continuous time model.

According to spatial interactions, the early ecological models were based on the mass-action law, first coined in chemical reaction. Hence the system is homogenous and the spatial interaction neither exists nor plays an important role. Before the 1970s, mathematical modelers typically used ordinary differential equations, seeking equilibria and analyzing their stability (Neuhauser, 2001). Subsequently, exploring bifurcation diagrams and chaotic patterns has continuously received lots of attention, related articles were written by J. Awrejcewicz (1991), J. Awrejcewicz & C.-H. Lamarque (2003), and E. Ott (1993). In 1969, the implicit spatial model, known as metapopulation, was first introduced by R. Levins. He constructed a model to describe a population that consists of several sub-populations joined together with immigration and emigration. Levins's simple model was motivated by and applied to a pest control situation over a large region, within which local populations would fluctuate in asynchrony (Hanski & Gilpin, 1991). This new concept was closely linked with the processes of local extinctions and re-colonization. I. Hanski & M. Gilpin (1991) also provided a conceptual distinction between local, metapopulation, and geographical scales. Later on, I. Hanski (1999) applied this approach to conservation biology. However, the role of space at the individual level is still not directly mentioned.

Throughout this chapter, an SIS epidemic model, well known disease transmission model, is considered. Of course, there are lots of diseases in this world. Here we focus on the disease that does not produce immunity, for example, some STD's, the eye disease, and the common cold. Therefore, the main situation is that the population is divided into a susceptible (S) group and an infectious (I) group. The S group is infected by the I group while the I group recovers from the disease and returns to the S group. Consequently, we set $S(t)$ and $I(t)$ to represent the number of S and I individuals at time t, respectively. The total population size is assumed to be constant, $S(t)+I(t)=N$. Moreover, $(\alpha / N)S(t)I(t)$ and $\gamma I(t)$ represent the infection rate at which the S population contracts the disease and the total number of I individuals who recover per unit time at the time t, respectively. Then, the SIS epidemic model should contain two equations. With constant population size, it can be reduced to a single equation. For continuous time scale, it is

$$\frac{dS}{dt}=\frac{\alpha}{N}S^2(t)-(\alpha+\gamma)S(t)+\gamma N,\ S(t)\geq 0. \tag{1}$$

In this case, a solution of Equation (1) always behaves non-oscillatory. For discrete time scale, the SIS epidemic model becomes

$$S(t+1)=\frac{\alpha}{N}S^2(t)-(\alpha+\gamma-1)S(t)+\gamma N. \tag{2}$$

In this case, the behavior of the endemic solution can be very complicated as it can tend to an equilibrium point, to limit cycles, or show chaos. Obviously, time scale makes system behavior amazing. In the next section, an SIS model is extended with more different time scales. Finally, the qualitative structures are investigated.

3. The important of time scales

In more complex situation, flu virus can spread continuously while in season, mostly disappear for some period of time, and spread again in a new season. According to Fig. 1,

the data shows periodic outbreaks of disease (the Influenza Division of Centers for Disease Control and Prevention, 2010). Obviously, the reasonable time scale for this model should be a combination of discrete and continuous scales. Thus the model either using difference equation(s) or differential equation(s) should be improper. Consequently, the calculus on time scales has been developed. S. Hilger first introduced this theory in order to unify continuous ($\mathbb{R}$) and discrete ($\mathbb{Z}$) analysis (Agarwal et al., 2002). Since then, the theory has been extended. Nowadays, time scales theory can be used to explain system dynamics not only for continuous and discrete times but also for other types of time such as period ($\mathbb{P}$) and discrete jump with fixed length ($h\mathbb{Z}$). There are plenty of publications in theoretical results. For example, J. Hoffacker & C. C. Tisdell (2005) studied on stability and instability for dynamic equations while E. Akin et al. (2001), D. R. Anderson (2009), and Y. Xu & Z. Xu (2009) studied on oscillation and nonoscillation criteria for dynamic equations and dynamic systems. However, there are few articles in numerical results (Sae-jie & Bunwong, 2009; Siming et al., 2008). In addition, there are some applications in economics (Atici et al., 2006) and epidemiology (Sae-jie et al., 2010a, 2010b; Thomas et al., 2009). In order to contribute time scales concepts and techniques understandably, some useful notations and definition are introduced as follows.

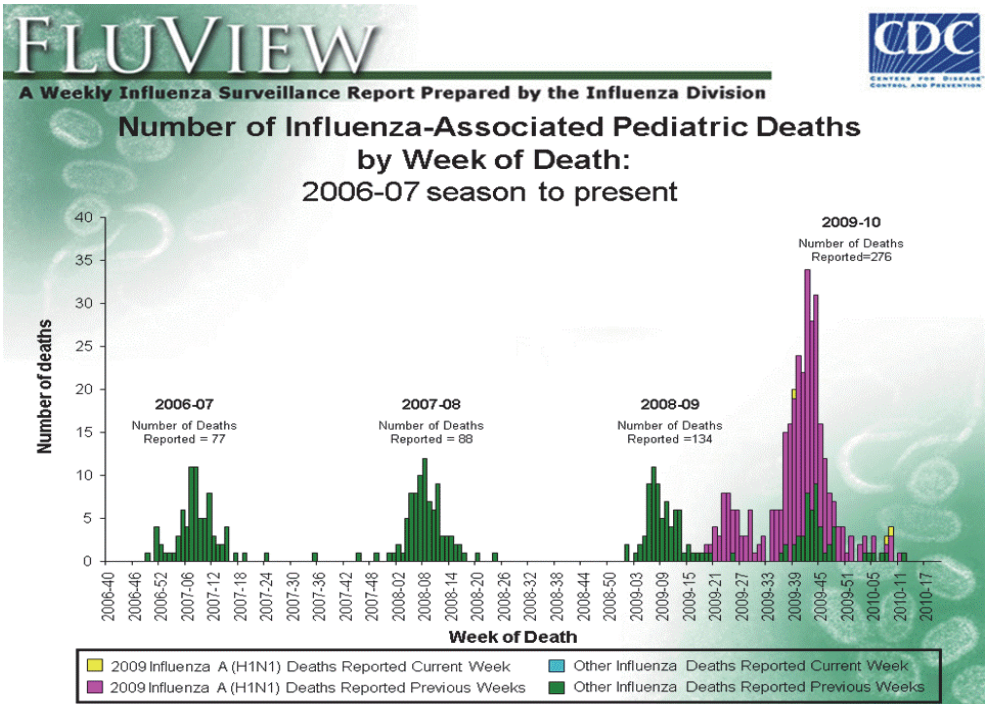


Fig. 1. The number of influenza-associated pediatric deaths.

3.1 Notations and definitions

A time scale, $\mathbb{T}$, is an arbitrary nonempty closed subset of the real numbers. Forward and backward jump operators are defined by $\sigma(t) = \inf\{s \in \mathbb{T} : s > t\}$ and $\rho(t) = \sup\{s \in \mathbb{T} : s < t\}$,

respectively where $\inf\varnothing = \sup\mathbb{T}$, $\sup\varnothing = \inf\mathbb{T}$, and $\varnothing$ denotes the empty set. A point $t \in \mathbb{T}$ is called left-dense if $t > \inf\mathbb{T}$ and $\rho(t) = t$, right-dense if $t < \sup\mathbb{T}$ and $\sigma(t) = t$, left-scattered if $\rho(t) < t$, right-scattered if $\sigma(t) > t$, isolated if $\rho(t) < t < \sigma(t)$, and dense if $\rho(t) = t = \sigma(t)$. Moreover, the set $\mathbb{T}^\kappa$ is defined by $\mathbb{T} \setminus \{m\}$ if $\mathbb{T}$ has a left-scattered maximum m. Otherwise, it is $\mathbb{T}$. Finally, the graininess function $\mu : \mathbb{T} \to [0, \infty)$ is defined by $\mu(t) = \sigma(t) - t$ (Bohner & Peterson, 2001).

3.2 Calculus on time scales

Traditionally, limits and continuity are key concepts for calculus development including calculus on time scales. A function $f : \mathbb{T} \to \mathbb{R}$ is said to be rd-continuous (right dense continuous) provided f is continuous at right-dense points and left-hand limits exist and it is finite at left-dense points in $\mathbb{T}$. Assume $f : \mathbb{T} \to \mathbb{R}$ is a function and $t \in \mathbb{T}^\kappa$. Then the following statements are equivalent:

a. The (delta) derivative of $f : \mathbb{T} \to \mathbb{R}$ at point $t \in \mathbb{T}^\kappa$, $f^\Delta(t)$, exists.

b. For all $\varepsilon > 0$, there is a neighborhood U of t (i.e., $U = (t - \delta, t + \delta) \cap \mathbb{T}$ for some $\delta > 0$) such that for all $s \in U$,

$$\left| f(\sigma(t)) - f(s) - f^\Delta(t)(\sigma(t) - s) \right| \le \varepsilon \left| \sigma(t) - s \right|. \tag{3}$$

c. $$f^\Delta(t) = \begin{cases} \lim\limits_{s \to t} \dfrac{f(t) - f(s)}{t - s} & \text{if } \mu(t) = 0 \\ \dfrac{f(\sigma(t)) - f(t)}{\sigma(t) - t} & \text{if } \mu(t) > 0. \end{cases} \tag{4}$$

d. $$f(\sigma(t)) = f(t) + \mu(t) f^\Delta(t). \tag{5}$$

Apart from discrete and continuous time scales, there are two more interesting time domain (as visualized in Fig. 2) to introduce. The first one is the combination of continuous and discrete time scales, denoted by the symbol $\mathbb{T} = \mathbb{P}_{l,h} := \bigcup_{k=0}^{\infty} [k(l+h), k(l+h) + l]$ where $l, h > 0$ and $k \in \mathbb{N}_0$. For this period time scale, *l* is the fixed length of the continuous interval while *h* is the fixed length of the discrete jump. The second one is composed of points that are equally spaced in time. Suppose that the distant between two successive points is *h*. Therefore, the symbol of this time scale is $\mathbb{T} = h\mathbb{Z} = \{hk : k \in \mathbb{Z}, h > 0\}$. Later on, forward and backward jump operators, graininess function, and (delta) derivative of function for each time scale are provided.

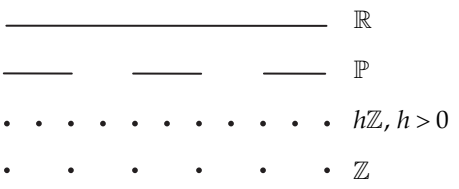


Fig. 2. Examples of time scales.

In the case $\mathbb{T}=\mathbb{R}$, we have

$$\sigma(t)=\rho(t)=t,\ \mu(t)=0,\ f^{\Delta}(t)=f'(t). \tag{6}$$

In the case $\mathbb{T}=h\mathbb{Z}$, we have

$$\sigma(t)=t+h,\ \rho(t)=t-h,\ \mu(t)=h,\ f^{\Delta}(t)=\frac{f(t+h)-f(t)}{h}. \tag{7}$$

In the case $\mathbb{T}=\mathbb{P}_{l,h}$, if $L=\bigcup_{k=0}^{\infty}[k(l+h),k(l+h)+l)$, $H_0=\bigcup_{k=0}^{\infty}\{k(l+h)+l\}$, and $H_1=\bigcup_{k=1}^{\infty}\{k(l+h)\}$ then we have

$$\sigma(t)=\begin{cases} t & if\ t\in L \\ t+h & if\ t\in H_0 \end{cases} \qquad \rho(t)=\begin{cases} t & if\ t\in L \\ t-h & if\ t\in H_1 \end{cases}$$

$$\mu(t)=\begin{cases} 0 & if\ t\in L \\ h & if\ t\in H_0 \end{cases} \qquad f^{\Delta}(t)=\begin{cases} f'(t) & if\ t\in L \\ \dfrac{f(\sigma(t))-f(t)}{h} & if\ t\in H_0 \end{cases}. \tag{8}$$

For more details on time scales, we refer the reader to Bohner & Peterson (2001,2003)

3.3 SIS epidemic model on time scales

After replacing $\frac{dS}{dt}$ in Equation (3) with $S^{\Delta}(t)$, we obtain the single equation of SIS epidemic model on time scales. Previously, one solution of continuous time model tends to one asymptotically stable equilibrium point while the solution of discrete time model displays various behaviors depending on parameter values. The following analyzes reveal system behavior on different time scale via two approaches. Firstly, we are interested in SIS epidemic model on $\mathbb{T}=h\mathbb{Z}$, i.e.,

$$S(\sigma(t))=\frac{\alpha\mu}{N}S^2(t)+(1-\alpha\mu-\gamma\mu)S(t)+\gamma N\mu. \tag{9}$$

It should be pointed out that μ and h are equivalent from now on. μ is not a graininess function. Obviously, $\mathbb{T}=\mathbb{Z}$ when $h=1$ and $\mathbb{T}=\mathbb{R}$ when h tends to zero. Secondly, we change our focus to SIS epidemic model on $\mathbb{T}=\mathbb{P}_{l,h}$, i.e.,

$$S((k+1)(l+1))=\frac{\alpha}{N}S^2(k(l+1)+l)+(1-\alpha-\gamma)S(k(l+1)+l)+\gamma N \tag{10}$$

for $t\in H_0$.

Then we explore the numerical results of previous dynamic equations individually. For $\mathbb{T}=h\mathbb{Z}$ time scale, we particularly investigate the system behavior when h or μ vary. For $\mathbb{T}=\mathbb{P}_{l,h}$ time scale, we fix $h=1$. Therefore, $\mathbb{T}=\mathbb{Z}$ when l approaches zero.

3.4 Numerical solution

Since the numerical method for ordinary differential equation and difference equation are already well-known, this section contains only some additional algorithm for computing the numerical solution of dynamic system on four time scales (Sae-jie, & Bunwong, 2009). Before running the following program, the user must define the function F and G.

```
INPUT: A                                  : starting time
       X, Y                               : initial values
       M                                  : number of step size (Case 2: M = l/H)
       H                                  : step size for continuous interval
Case 1: 𝕋 = ℝ
       FOR j = 0, 1, 2, ..., M DO
           T = A + H*j
           SAVE T, X, Y in the OUTPUT LIST
           RUNGE-KUTTA-FEHLBERG METHOD
       END FOR
Case 2: 𝕋 = ℙ_{l,h}
       P                                  : numbers of period
       l                                  : length of continuous interval
       h                                  : length of jump
       DUM3 = A
       FOR j = 0, 1, 2, ..., M DO
           T = DUM3 + H*j
           SAVE T, X, Y in the OUTPUT LIST
           DUM1 = X
           DUM2 = Y
           RUNGE-KUTTA-FEHLBERG METHOD
       END FOR
       X = DUM1 + h*F(T, DUM1, DUM2)
       Y = DUM2 + h*G(T, DUM1, DUM2)
       FOR k = 1, 2, 3, ..., P DO
           DUM3 = A + k*(l+h)
           FOR j = 0, 1, 2, ..., M DO
               T = DUM3 + H*j
               SAVE T, X, Y in the OUTPUT LIST
               DUM1 = X
               DUM2 = Y
               RUNGE-KUTTA-FEHLBERG METHOD
           END FOR
           X = DUM1 + h*F(T, DUM1, DUM2)
           Y = DUM2 + h*G(T, DUM1, DUM2)
       END FOR
Case 3: 𝕋 = hℤ
       DUM3 = A
       FOR j = 0, 1, 2, ..., M DO
           T = DUM3 + H*j
```

```
            SAVE T, X, Y in the OUTPUT LIST
            DUM1 = X
            DUM2 = Y
            X = DUM1 + H*F(T, DUM1, DUM2)
            Y = DUM2 + H*G(T, DUM1, DUM2)
        END FOR
F := X^Δ (T, X, Y)
G := Y^Δ (T, X, Y)
```

In the case $\mathbb{T} = h\mathbb{Z}$, W. Sae-jie et al. (2010a) always fixed the following parameter values $\gamma = 0.9$ and $N = 100$. For each value of α, μ was treated as the bifurcation parameter. When $\alpha = 2$, the dynamic behavior of continuous and discrete system are the same. For $\alpha = 3.4$, the solutions, however, appear as asymptotically stable, a period two cycle, a period four cycle when $\mu = 0.1, 0.9, 1.0$, respectively as shown in Fig. 3.

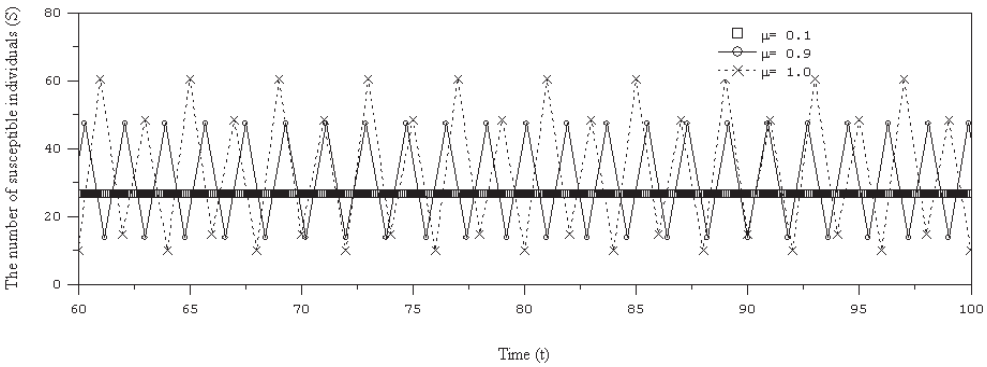


Fig. 3. The time series solution of Equation (9) when $\alpha = 3.4$.

M. R. S. Kulenovic & O. Merino (2002) and L. Tien-yien & J. A. Yorke (1975) discovered that if there exists a period three cycle, then there exists chaotic behavior. For $\alpha = 3.6$, the bifurcation diagram (as shown in Fig. 4) exhibits a period three cycle for $\mu \in (1.0476, 1.0524)$. Consequently, dynamic equation (9) can generate chaotic pattern for a proper value of μ. For more details, the non-oscillatory solution, the oscillating period two solution, and the chaos occur when $\mu = 0.1$, $\mu = 0.8$, and $\mu = 1$, respectively. The last behavior is illustrated in Fig. 5.

In the case $\mathbb{T} = \mathbb{P}_{l,h}$, this period time scale and a continuous time scale are equivalent when $\mu = h = 0$. Thus, the result appears as a non-oscillatory solution. However, W. Sae-jie et al. (2010b) always fixed the length of discrete jump together with the following parameter values $\alpha = 3.6$, $\mu = h = 1$, $\gamma = 0.9$, and $N = 100$ but varied the length of continuous interval, *l*. For sufficiently high value of *l*, the result (as shown in Fig. 6) appears as a non-oscillatory solution which is similar to the result in Fig. 3. for a continuous time scale. It disappears in some intervals because of the discrete time jump. When the length of the continuous interval decreases, a period two cycle with some continuous intervals appears as visualized in Fig. 7. If the length of the continuous interval is gradually reduced and closely zero, $\mathbb{T} = \mathbb{P}_{l,h}$ is similar to a discrete time scale.

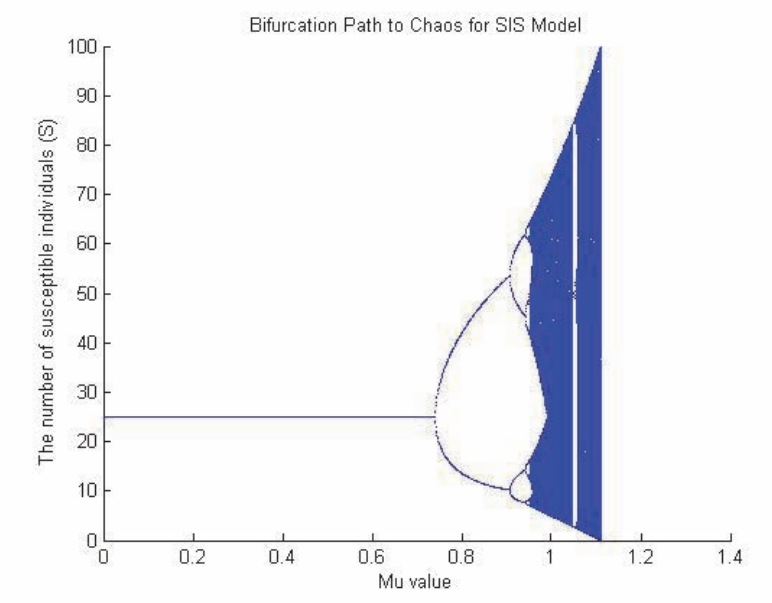


Fig. 4. The bifurcation diagram of μ when $\alpha = 3.6$.

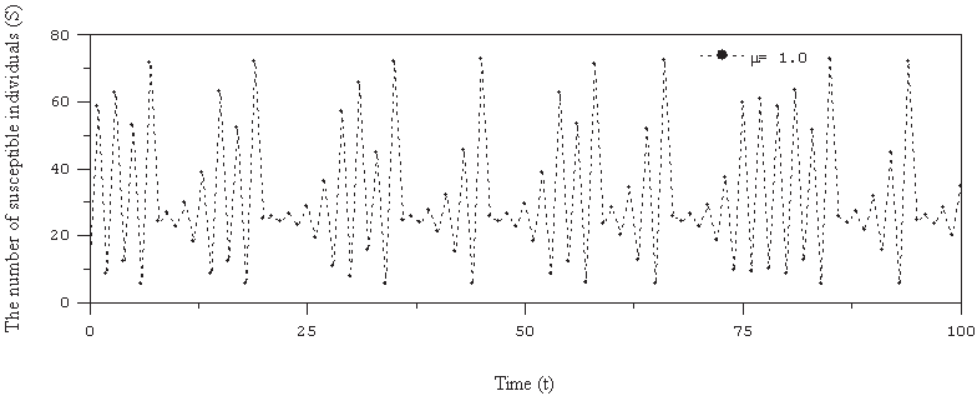


Fig. 5. The time series solution of Equation (9) when $\alpha = 3.6$ and $\mu = 1$.

4. The important of spatial interaction

In contrast to a metapopulation model, space in an explicit model should be taken into account as another variable. As for non spatial models, the time scale, space, and population state for an explicit model can be either discrete or continuous. If all of them are continuous variables, a partial differential equation has been used (for example, Fisher, 1937). However, solving partial differential equations analytically is difficult. Consequently, analysis is based on computer simulation. Then we face another limitation because solving them numerically

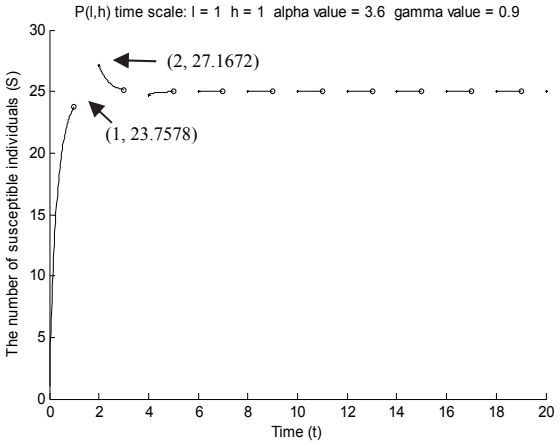


Fig. 6. The time series solution of Equation (10) on time scale $\mathbb{P}_{1,1}$.

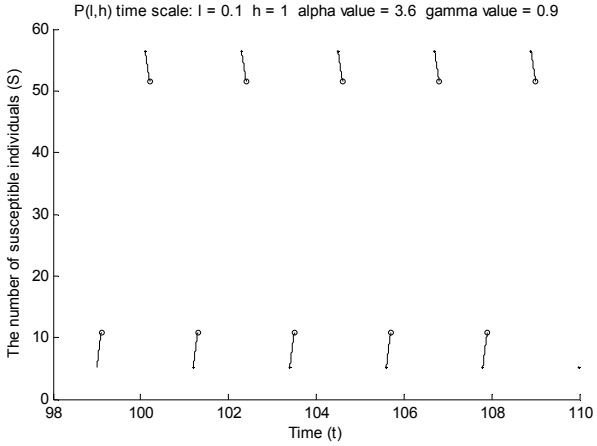


Fig. 7. The time series solution of Equation (10) on time scale $\mathbb{P}_{0.1,1}$.

involves discretizing space and time. In principle, this reduces them to coupled map lattice model. On the other hand, if space is treated as discrete, a lattice model is proposed. With respect to state variables, lattice models can be divided into three subgroups, these are coupled map lattice (Morris, 1997), cellular automaton (Wolfram, 1986), and network model (Sole & Manrubia, 1996, 1997; Verdasca et al., 2005). A major problem with the lattice model is that analysis is often restricted to direct computer simulation which consumes expensive time when it is a stochastic model on a reasonably large lattice.

Fortunately, in microscopic point of view, there is an alternative way, so-called pair approximation, that can help us not only to produce the numerical solutions of spatially explicit lattice models by using ordinarily numerical method but also to analyze them. The Japanese researchers were pioneers to apply this idea to ecological systems and have

continuously produced theoretical result and applications (Harada et al., 1995; Iwasa et al., 1998; Kubo et al., 1996; Matsuda et al., 1992; Sato et al., 1994). They referred to the pair approximation as the doublet decoupling approximation and represent their system as coupled system of ordinary differential equation for the density of singletons and pairs. However, Rand's colleagues (Bauch, 2000; Bunwong 2006; Keeling, 1995; Morris, 1997; van Baanlen & Rand, 1998) used pair approximation as a moment closure approximation where the number of pairs and the number of singletons are the only state variables. Higher order correlations are neglected. A more advantage of this method is that it keeps spatial correlations. There are varieties of applications in ecological interactions (Dieckmann et al., 2000; Ellner, 2001; Ellner et al., 1998), epidemiology (Benoit et al., 2006; de Aguiar et al., 2004 ; Dieckmann et al., 2002 ; Elliott et al., 2000 **;** Joo & Lebowitz, 2004), and forest dynamics (Schlicht & Iwasa, 2007). However, the mathematical formulas are still limited. Thus, we attempt to broaden the pair approximation idea by calculating configuration averages. We call our method as a new approach and previous method as an original approach. In order to contribute our techniques understandably, some useful notations and definition are introduced as follows.

4.1 Notations and definitions

Under a given configuration $\underline{\sigma} = (\sigma k)$ where $k \in \{x, e\}$, the following Rand's notations are defined (Rand, 1998).

$\sigma x, \sigma e$	are the state of the site x and the edge e, respectively,
$\sigma x = i$	is that the state of the individual x is i,
$\sigma e = ij$	is that one end of the edge e is in state i, e_i, while the other is in state j, e_j,
$[i],[ij],[ijk]$	are the number of sites, edges, and triples in state i, ij, and ijk, respectively,
$Q_x(i), Q_{e_j}(i)$	are the number of i-state neighbors of the sites x and e_i, respectively,
$\langle Q_x(i) \rangle_{\sigma x = j}$	is the average value of the number of i-state neighbors of a j-state site,
$\left\langle Q_{e_j}(i) \right\rangle_{\sigma e = jk}$	is the average value of the number of i-state neighbors of a j-state site in a jk-state edge,
q_i	equals $[i]/N$ where N is the total population size, and
$q_{i\|j}$	equals $[ij]/Q[j]$ where Q is the average number of neighbors.

In this framework, space is represented by a network of sites. Each site can either be occupied by an individual or remains as an empty site that is still available for an individual to occupy. Two sites are neighbors when they regularly interact with each other. Joining these two neighboring sites performs an edge or pair. A line is used for this interaction. Fig. 8 provides an example when the state of site x is i and the state of site y is j. Both sites are neighbors. Moreover, there is an edge e.

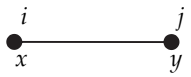


Fig. 8. $\sigma x = i, \sigma y = j, \sigma e = ij$.

4.2 SIS Master and correlation equations

Our correlations are microcorrelations which can be measured on the scale of the interactions of individuals. After approximating higher order terms in master equations, we obtain a system of ordinary differential equation which composed of density of lower order terms, known as correlation equations. Moreover, the approximation technique is called the moment closure approximation. The differential equation for the single numbers involves pair numbers, triple numbers, etc. The differential equation for the pair numbers involves triple numbers, etc. So we get an infinite hierarchy of equations. However, we need to truncate the hierarchy at some point. For instant, pair approximation, the first order of moment closure approximation, truncates triples and higher order terms as functions of singletons and pairs only (Rand, 1998).

Let f be a real-valued function of the state of the network at time t, which can be approximated as continuous. The equation f is derived by summing over all events in the population which affect f and the total change produced by those events is

$$\frac{df}{dt} = \sum_{\varepsilon \in events} r(\varepsilon)\Delta f_{\varepsilon} \tag{11}$$

where $r(\varepsilon)$ is the rate of event ε and Δf_{ε} is the change produced in f by event ε. It is called the master equation.

For our case study, the state of each site and edge will change over time as a consequence of two major types of events – infection and recovery. Infection changes the state $\sigma e = SI$ of the edge e into the state $\sigma' e = II$ at rate β and recovery changes the state $\sigma x = I$ of a site x into the state $\sigma' x = S$ at rate δ. Therefore, the SIS spatial model becomes

$$\frac{d[S]}{dt} = \sum\nolimits_{\sigma x = I} \delta - \sum\nolimits_{\sigma e = SI} \beta = -\frac{d[I]}{dt}$$

$$\frac{d[SI]}{dt} = -\sum\nolimits_{\sigma x = I} \delta Q_x(S) + \sum\nolimits_{\sigma e = SI} \beta (Q_{e_S}(S) - Q_{e_S}(I)) + \sum\nolimits_{\sigma x = I} \delta Q_x(I)$$

$$\frac{d[SS]}{dt} = 2\sum\nolimits_{\sigma x = I} \delta Q_x(S) - 2\sum\nolimits_{\sigma e = SI} \beta Q_{e_S}(S)$$

$$\frac{d[II]}{dt} = 2\sum\nolimits_{\sigma e = SI} \beta Q_{e_S}(I) - 2\sum\nolimits_{\sigma x = I} \delta Q_x(I) . \tag{12}$$

If δ and β are constant, the original approach of pair approximation is still valid. However, the real-world situation is more complicated. For example, the human-to-human transmission of Swine Flu occurs by inhalation of infectious droplets and droplet nuclei, and by direct contact, which is facilitated by air and land travel and social gatherings (Sinha, 2009). Therefore, the transmission rate and the recovery rate could vary depending on the surrounding infectious people. Consequently, we are able to assume that the infection rate and the recovery rate are $\beta = b_0 + b_1 Q_{e_S}(I)$ and $\delta = d_0 - d_1 Q_x(I)$, respectively where b_0, b_1, d_0, d_1 are constant. Then the formulation of pair approximation is in trouble. Using the fact that

$$\sum_{\sigma x=j} Q_x(i) = [j]\langle Q_x(i)\rangle_{\sigma x=j} \text{ and } \sum_{\sigma e=jk} Q_{e_j}(i) = [jk]\left\langle Q_{e_j}(i)\right\rangle_{\sigma e=jk}, \tag{13}$$

we, then, can use average forms instead of summation terms in the master equation as follows,

$$\frac{d[S]}{dt} = d_0[I] - d_1[I]\langle Q_x(I)\rangle_{\sigma x=I} - b_0[SI] - b_1[SI]\left\langle Q_{e_S}(I)\right\rangle_{\sigma e=SI}$$

$$\begin{aligned}\frac{d[SI]}{dt} &= -d_0[I]\langle Q_x(S)\rangle_{\sigma x=I} + d_1[I]\langle Q_x(I)Q_x(S)\rangle_{\sigma x=I} + b_0[SI]\left\langle Q_{e_S}(S)\right\rangle_{\sigma e=SI} \\ &\quad + b_1[SI]\left\langle Q_{e_S}(I)Q_{e_S}(S)\right\rangle_{\sigma e=SI} - b_0[SI]\left\langle Q_{e_S}(I)\right\rangle_{\sigma e=SI} - b_1[SI]\left\langle Q_{e_S}(I)Q_{e_S}(I)\right\rangle_{\sigma e=SI} \\ &\quad + d_0[I]\langle Q_x(I)\rangle_{\sigma x=I} - d_1[I]\langle Q_x(I)Q_x(I)\rangle_{\sigma x=I}\end{aligned}$$

$$\begin{aligned}\frac{d[SS]}{dt} &= 2d_0[I]\langle Q_x(S)\rangle_{\sigma x=I} - 2d_1[I]\langle Q_x(I)Q_x(S)\rangle_{\sigma x=I} - 2b_0[SI]\left\langle Q_{e_S}(S)\right\rangle_{\sigma e=SI} \\ &\quad - 2b_1[SI]\left\langle Q_{e_S}(I)Q_{e_S}(S)\right\rangle_{\sigma e=SI}\end{aligned}$$

$$\begin{aligned}\frac{d[II]}{dt} &= 2b_0[SI]\left\langle Q_{e_S}(I)\right\rangle_{\sigma e=SI} + 2b_1[SI]\left\langle Q_{e_S}(I)Q_{e_S}(I)\right\rangle_{\sigma e=SI} - 2d_0[I]\langle Q_x(I)\rangle_{\sigma x=I} \\ &\quad + 2d_1[I]\langle Q_x(I)Q_x(I)\rangle_{\sigma x=I}\end{aligned}. \tag{14}$$

4.3 Calculating configuration averages

It should be pointed out that there are two types of average values in this new approach. Previously, all mentioned average values are called space or population average value because they are averages of the quantity over subsets of population. Now we introduce another way to calculate an average value. That is the expected value with respect to probability distribution. This seems reasonable to assume that if population size is large enough, then the configuration averages approximate probability expectations. Under some necessary and sufficient conditions, the space average and probability average are identical. K. Bunwong (2006, 2010a, 2010b) developed more formulas under multinomial distribution with parameters Q and p_i where $p_i = q_{i|j}$ and Poisson distribution for coordination numbers, respectively. In this chapter, the SIS spatial model is based on the framework that each site connects to a fixed number of neighbors, Q. The following Bunwong's formulas (2006) are used.

$$\langle Q_x(i)\rangle_{\sigma x=j} = \frac{[ij]}{[j]}$$

$$\langle Q_x(i_1)Q_x(i_2)\rangle_{\sigma x=j} = \begin{cases} Qq_{i_1|j} + \dfrac{Q!}{(Q-2)!}q_{i_1|j}^2 & ;i_1 = i_2 \\ \dfrac{Q!}{(Q-2)!}q_{i_1|j}q_{i_2|j} & ;i_1 \neq i_2 \end{cases}$$

$$\left\langle Q_{e_j}(l)\right\rangle_{\sigma e=ij}=\begin{cases}(Q-1)q_{l|j} & ;l\neq i\\ (Q-1)q_{l|j}+1 & ;l=i\end{cases}$$

$$\left\langle Q_{e_j}(l_1)Q_{e_j}(l_2)\right\rangle_{\sigma e=ij}=\begin{cases}(Q-1)q_{l_1|j}+\dfrac{(Q-1)!}{(Q-3)!}q_{l_1|j}^2 & ;l_1=l_2\\ \dfrac{(Q-1)!}{(Q-3)!}q_{l_1|j}q_{l_2|j} & ;l_1\neq l_2\end{cases} \quad (15)$$

4.4 Numerical solution

The numerical method for spatial model (14) is as same as for ordinary differential equation. Here, the number of pairs and the number of singletons are state variables. K. Bunwong (2010a) mainly investigated the density of infected individuals, defined by $q_i=[I]/N$, along the time series. We always fixed $b_0=0.3, d_0=0.2$. In case that the infection rate and the recovery rate are not affected by the surrounding infectious individuals ($b_1=0, d_1=0$), Fig. 9 reveals that the more nearby neighbors, the higher density of infected individuals at the equilibrium point. Moreover, in case that only the infection rate is affected by the surrounding infectious individuals ($b_1\neq 0, d_1=0$), the stronger effect of the surrounding infectious individuals on the infection rate, the higher density of infected individuals at the equilibrium point. In case that only the recovery rate is affected by the surrounding infectious individuals ($b_1=0, d_1\neq 0$), the stronger effect of the surrounding infectious individuals on the recovery rate, the higher density of infected individuals at the equilibrium point. The illustrations can be seen in Bunwong (2010a).

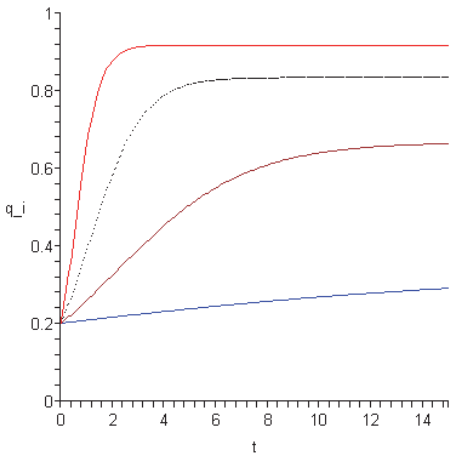


Fig. 9. Time evolution of the density of infected individuals q_i. Parameters: $b_0=0.3, b_1=0, d_0=0.2, d_1=0$ and $Q=8,4,2,1$ (from top to bottom, respectively).

5. Conclusion

Obviously, the assumption on timescale affects the system behavior. In order to simplify real-world correctly, choosing a suitable time scale is important for model formulation. Our example, SIS epidemic model, proves that if continuous time scale is used then two solutions of the system are asymptotically stable or unstable depending on parameter values and stable oscillating solutions have never existed. In contrast, if discrete time scale is applied then there are various types of solution behaviors such as equilibrium point solutions, period two cycles, period four cycles, period three cycles, and also chaotic solutions depending on parameter values as well. Consequently, the predicted behaviors from a model can be qualitatively very different for different time scales. The hard to answer question is "What is the proper model to understand observed data in a variety of time measurements?". Of course, the observed data is usually discrete. Can we use differential equation(s)? With theory of time scales, the models are more varieties. Bifurcation diagram shows that the discrete jump and the continuous interval are essential. The differential equation(s) can be described observed data if the jump distant of data is sufficiently small. The distant also depends on other parameters. Moreover, numerical verification can visualize other interesting behavior pattern and guide other mathematician to consider theoretically.

In recent years, the effect of spatial structure is often taken into account in ecological interactions. Pair approximation is one of the powerful tools to understand human to human interactions. We have developed the way to approximate higher order quantities and applied to ecological problems. Particularly, our new approach is suitable for a model evolving according to the transition rates affecting additionally by neighbors. For example, people infect flu virus easily from their nearby neighbors. The health organization usually suggests infectious people to have some rest and be away from public places. It implies that if we surrounding with more infectious people, then we have higher chance to infected and/or lower chance to recover as shown in the numerical results.

Further studies on time scales should involve more numerical techniques and applications while future works on spatial interaction should concern development of formulation.

6. Acknowledgment

I would like to thank Prof. David Rand for introducing me to spatial world, Prof. Yoh Iwasa for sharing his experience, Prof. Ravi Agarwal for introducing me to time scale world, Dr. Elvin Moore for intuitive ideas, and Dr. Wichuta Sae-jie for collaboration. I am grateful to the Development and Promotion of Science and Technology Talent Project (DPST) and Thailand Research Fund (TRF) for financial supports under contact number MRG5180246.

7. References

Agarwal, R.; Bohner, M.; O'Regan, D. & Peterson, A. (2002). Dynamic Equations on Time Scales: A Survey. *Journal of Computational and Applied Mathematics*, Vol. 141, No. 1-2, (April 2002), pp. 1-26, ISSN 0377-0427.

Akin, E.; Erbe, L.; Peterson, A. & Kaymakçalan, B. (2001). Oscillation Results for a Dynamic Equation on a Time Scale. *Journal of Difference Equations and Applications*, Vol. 7, No. 6, pp. 793-810, ISSN 1023-6198.

Allen, L. J. S. (1994). Some Discrete-Time SI, SIR, and SIS Epidemic Models. *Mathematical Biosciences*, Vol. 124, No. 1, (November 1994), pp. 83-105, ISSN 0025-5564.

Allen, L. J. S. & Burgin, A. B. (2000). Comparison of Deterministic and Stochastic SIS and SIR Models in Discrete Time. *Mathematical Biosciences*, Vol. 163, No. 1, (January 2000), pp. 1-33, ISSN 0025-5564.

Anderson, D. R. (2009). Oscillation and Nonoscillation Criteria for Two-Dimensional Time-Scale Systems of First-Order Nonlinear Dynamic Equations. *Electronic Journal of Differential Equations*, Vol. 2009, No. 24, pp. 1-13, ISSN 1072-6691.

Atici, F. M.; Biles, D. C. & Lebedinsky, A. (2006). An Application of Time Scales to Economics. *Mathematical and Computer Modelling*, Vol. 43, No. 7-8, (April 2006), pp. 718-726, ISSN 0895-7177.

Awrejcewicz, J. (1991). *Bifurcation and Chaos in Coupled Oscillators*, World Scientific Publishing, ISBN 9-8102-0579-1, Singapore.

Awrejcewicz, J. & Lamarque, C.-H. (2003). *Bifurcation and Chaos in Nonsmooth Mechanical Systems*, World Scientific Publishing, ISBN 9-8123-8459-6, Singapore.

Bauch, C. T. (2000). Moment Closure Approximations in Epidemiology. *Warwick PhD Thesis*.

Benoit, J.; Nunes, A. & da Gama, M. T. (2006). Pair Approximation Models for Disease Spread. *European Physical Journal B*, Vol. 50, No. 1-2, (March 2006), pp. 177-181, ISSN 1434-6028.

Bohner, M. & Peterson, A. (2001). *Dynamic Equations on Time Scales: An Introduction with Applications*, Birkhäuser, ISBN 0-8176-4225-0, Boston, USA.

Bohner, M. & Peterson, A. (2003). *Advances in Dynamic Equations on Time Scales*, Birkhäuser, ISBN 0-8176-4293-5, Boston, USA.

Bunwong, K. (2006). Spatial Modeling in Evolutionary Ecology. *Warwick PhD Thesis*.

Bunwong, K.; Sae-jie, W. & Lenbury, Y. (2009). Modelling Nitrogen Dynamics of a Constructed Wetland: Nutrient Removal Process with Variable Yield. *Nonlinear Analysis: Theory, Methods & Applications*, Vol. 71, No. 12, (December 2009), pp. e1538-e1546, ISSN 0362-546X.

Bunwong, K. (2010). Pair Approximations for Ecological Models with Additional Neighborhood Effects. *Proceedings of the 4th International Conference on Applied Mathematics, Simulation, Modelling (ASM '10),* pp. 165-169, ISBN 978-960-474-210-3. Corfu Island, Greece, July 22-25, 2010.

Bunwong, K. (2010). A New Approach to Pair Approximation Method for Spatial Model in Ecology. *WSEAS Transactions on Mathematics*, Vol. 9, No. 10, (October 2010), pp. 768-777, ISSN 1109-2769.

de Aguiar, M. A. M.; Rauch, E. M. & Bar-Yam, Y. (2004). Invasion and Extinction in the Mean Field Approximation for a Spatial Host-Pathogen Model. *Journal of Statistical Physics*, Vol. 114, No. 5-6, (March 2004), pp.1417-1451, ISSN 0022-4715.

Dieckmann, U.; Law, R. & Metz, J. A. J. (2000). *The Geometry of Ecological Interactions: Simplifying Spatial Complexity,* Cambridge University Press, ISBN 0-5216-4294-9, Cambridge, UK.

Dieckmann, U.; Metz, J. A. J.; Sabelis, M. W. & Sigmund, K. (2002). *Adaptive Dynamics of Infectious Diseases: In Pursuit of Virulence Management*, Cambridge University Press, ISBN 0-5217-8165-5, Cambridge, UK.

Edelstein-Keshet, L. (1988). *Mathematical Models in Biology*, Random House, ISBN 0-3943-5507-5, New York, USA.

Elliott, P.; Wakefield, J. C.; Best, N. G. & Briggs, D. J. (2000) *Spatial Epidemiology: Methods and Applications*, Oxford University Press, ISBN 0-1926-2941-7, New York, USA.

Ellner, S. P. (2001). Pair Approximation for Lattice Models with Multiple Interaction Scales. *Journal of theoretical biology*, Vol. 210, No. 4, (June 2001), pp. 435-447, ISSN 0022-5193.

Ellner, S. P. & Guckenheimer, J. (2006). *Dynamic Models in Biology,* Princeton University Press, ISBN 0-6911-1843-4, New Jersey, USA.

Ellner, S. P.; Sasaki, A.; Haraguchi, Y. & Matsuda, H. (1998). Speed of Invasion in Lattice Population Models: Pair-Edge Approximation. *Journal of Mathematical Biology*, Vol. 36, No. 5, pp. 469-484, ISSN 0303-6812.

Fisher, R. A. (1937). The Wave of Advance of Advantageous Genes. *Annals of Human Genetics*. Vol. 7, No. 4, pp. 355-369.

Getz, W. M. & Lloyd-Smith, J. O. (2006). Basic Methods for Modeling the Invasion and Spread of Infectious Diseases. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Vol. 71, pp. 1-23, ISSN 1052-1798.

Harada, Y.; Ezoe, H.; Iwasa, Y.; Matsuda, H. & Sato, K. (1995). Population Persistence and Spatially Limited Social Interaction. *Theoretical population Biology*, Vol. 48, No. 1, (August 1995), pp. 65-91, ISSN 0040-5809.

Hanski, I. (1999). *Metapopulation Ecology*, Oxford University Press, ISBN 0-1985-4065-5, New York, USA. (Reprinted 2005).

Hanski, I. & Gilpin, M. (1991). Metapopulation Dynamics: Brief History and Conceptual Domain. *Biological Journal of the Linnean Society*, Vol. 42, No. 1-2, (January 1991), pp. 3-16. ISSN 1095-8312. (Reprinted in Gilpin, M. & Hanski, I. (eds). (1991). *Metapopulation Dynamics: Empirical and Theoretical Investigations*, Academic Press, ISBN 0-12-284120-4, UK.)

Hofbauer, J. & Sigmund, K. (1988). The Theory of Evolution and Dynamical Systems, Cambridge University Press, ISBN 0-5213-5838-8, Cambridge, UK.

Hoffacker, J. & Tisdell, C. C. (2005). Stability and Instability for Dynamic Equations on Time Scales. *Computers and Mathematics with Applications*, Vol. 49, No. 9-10, (May 2005), pp. 1327-1334, ISSN 0898-1221.

Iwasa, Y.; Nakamaru, M. & Levin, S. A. (1998). Allelopathy of Bacteria in a Lattice Population: Competition between Colici-Sensitive and Colicin-Producing Strains. *Evolutionary Ecology*, Vol. 12, pp. 785-802, ISSN 0269-7653.

Jang, S. R.-J. (2008). Backward Bifurcation in a Discrete SIS Model with Vaccination. *Journal of Biological Systems*, Vol. 16, No. 4, (December 2008), pp. 479-494, ISSN 0218-3390.

Joo, J. & Lebowitz, J. L. (2004). Pair Approximation of the Stochastic Susceptible-Infected-Recovered-Susceptible Epidemic Model on the Hypercubic Lattice. *Physial Review E.*, Vol. 70, No. 3, (April 2004), ISSN 1539-3755,

Keeling, M. J. (1995). The Ecology and Evolution of Spatial Host-Parasite Systems. *Warwick PhD Thesis*.

Kubo, T.; Iwasa, Y. & Furumoto, N. (1996). Forest Spatial Dynamics with Gap Expansion: Total Gap Area and Gap Size Distribution. *Journal of Theoretical Biology*, Vol. 180, No. 3, (June 1996), pp. 229-246, ISSN 0022-5193.

Kulenovic, M. R. S. & Merino, O. (2002). *Discrete Dynamical Systems and Difference Equations with Mathematica*. Chapman and Hall/CRC, ISBN 1584882875, Florida, USA.

Matsuda, H.; Ogita, N.; Sasaki, A. & Sato, K. (1992). Statistical Mechanics of Population-the Lattice Lotka-Voltera Model. *Progress of Theoretial Physics*, Vol. 88, pp. 1035-1049, ISSN 0033-068X.

Morris, A. J. (1997). Representing Spatial Interactions in Simple Ecological Models. *Warwick PhD Thesis*.

Murray, J. D. (1993). *Mathematical Biology*. Springer-Verlag, ISBN 3-5405-7204-X, Berlin, German.

Neuhauser, C. (2001). Mathematical Challenges in Spatial Ecology. *Notices of the AMS*. Vol. 48, No. 11, (December 2001), pp. 1304-1314.

Ott, E. (1993). *Chaos in Dynamical Systems*. Cambridge University Press, ISBN 0-5214-3799-7, Cambridge, UK.

Otto, S. P. & Day, T. (2007). *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press, ISBN 978-0-6911-2344-8, New Jersey, USA.

Rand, D.A. (1998). Correlation Equations and Pair Approximations for Spatial Ecologies. *Warwick Preprint*.

Roughgarden, J. (1996). *Theory of Population Genetics and Evolutionary Ecology: an Introducation*. Prentice Hall, ISBN 0-1344-1965-0, New Jersey, USA.

Sae-jie, W. & Bunwong, K. (2009). Numerical Verification of Certain Oscillation Results on Time Scales. *Neural, Parallel & Scientific Computations*, Vol. 17, pp. 317-338, ISSN 1061-5369.

Sae-jie, W.; Bunwong, K. & Moore, E. J. (2010). Qualitative Behavior of SIS Epidemic Model on Time Scales. *Proceedings of the 4th International Conference on Applied Mathematics, Simulation, Modelling (ASM '10)*, pp. 159-164, ISBN 978-960-474-210-3. Corfu Island, Greece, July 22-25, 2010.

Sae-jie, W.; Bunwong, K. & Moore, E.J. (2010). The Effect of Time Scales on SIS Epidemic Model. *WSEAS Transactions on Mathematics*, Vol. 9, No. 10, (October 2010), pp. 757-767, ISSN 1109-2769.

Sato, K.; Matsuda, H. & Sasaki, A. (1994). Pathogen Invasion and Host Extinction in Lattice Structured Populations. *Journal of Mathematical Biology*, Vol. 32, No. 3, pp. 251-268, ISSN 0303-6812.

Schlicht, R. & Iwasa, Y. (2007). Spatial Pattern Analysis in Forest Dynamics: Deviation from Power Law and Direction of Regeneration Waves. *Ecology Research*, Vol. 22, No. 2, pp. 197-203, ISSN 0912-3814.

Siming, Z.; Yanhui, L. & Zhixin, R. (2008). Region Qualitative Analysis of Predator-Prey System on Time Scales. *Annals of Differential Equations,* Vol. 24, pp. 121-126, ISSN 1002-0942.

Sinha, M. (2009). Swine Flu. *Journal of Infection and Public Health*, Vol. 2, pp. 157-166, ISSN 1876-0341.

Sole, R. V., and Manrubia, S. C. (1996). Extinction and Self-Organized Criticality in a Model of Large-scale Evolution. *Physical Review E*, Vol. 54, No. 1, (July 1996), pp. R42-R45. ISSN 1063-651X.

Sole, R. V., and Manrubia, S. C. (1997). Criticality and Unpredictability in Macroevolution. *Physical Review E*, Vol. 55, No. 4, (April 1997), pp. 4500-4507.

Takenaka, Y.; Matsuda, H. & Iwasa, Y. (1997). Competition and Evolutionary Stability of Plants in a Spatially Structured Habitat. *Researches on Population Ecology*, Vol. 39, No. 1, pp. 67-75, ISSN 0034-5466.

The Influenza Division of Centers for Disease Control and Prevention. (May 2010). Fluview: A Weekly Influenza Surveillance Report, In: *Seasonal Influenza (Flu)*, 07.04.2011, Available from: http://www.cdc.gov/flu/weekly/weeklyarchives2009-2010/weekly20.htm

Thomas, D.; Weedermann, M.; Billings, L.; Hoffacker, J. & Washington-Allen, R. A. (2009). When to Spray: a Time-Scale Calculus Approach to Controlling the Impact of West Nile Virus. *Ecology and Society*, Vol. 14, No. 2, Article 21, ISSN 1708-3087.

Tien-yien, L. & Yorke, J. A. (1975). Period Three Implies Chaos. *The American Mathematical Monthly*, Vol. 82, No. 10, (December 1975), pp. 965-992, ISSN 0002-9890.

van Baanlen M. & Rand, D. A. (1998). The Unit of Selection in Viscous Populations and the Evolution of Altruism. *Journal of Theoretical Biology*, Vol. 193, No. 4, (August 1998), pp. 631-648, ISSN 0022-5193.

Verdasca, J.; Da Gama, M. M. T.; Nunes, A.; Bernardino, N. R.; Pacheco, J. M. & Gomes, M. C. (2005). Recurrent Epidemics in Small World Networks. *Journal of Theoretical Biology*, Vol. 233, No. 4, (April 2005), pp. 553-561, ISSN 0022-5193.

Wolfram, S. (1986). *Theory and Applications of Cellular Automata*. World Scientific Publishing. ISBN 9-9715-0123-6, Singapore.

Xu Y. & Xu, Z. (2009). Oscillation Criteria for Two-Dimensional Dynamic Systems on Time Scales. *Journal of Computational and Applied Mathematics*, Vol. 225, No. 1, (March 2009), pp. 9-19, ISSN 0377-0427.

7

Unscented Filtering Algorithm for Discrete-Time Systems with Uncertain Observations and State-Dependent Noise

R. Caballero-Águila[1], A. Hermoso-Carazo[1] and J. Linares-Pérez[2]
[1]*Dpto. de Estadística. Universidad de Jaén. Paraje Las Lagunillas. 23071. Jaén*
[2]*Dpto. de Estadística. Universidad de Granada. Avda. Fuentenueva. 18071. Granada*
Spain

1. Introduction

Most conventional filtering algorithms address situations in which the signal to be estimated is always present in the observations. However, in many real situations, usually the measurement device or the transmission of such measurements can be subject to random failures, generating observations which may consist of noise only. More specifically, there is a positive probability (*false alarm probability*) that the signal to be estimated is not present in the corresponding observation; that is, the observations may be only noise (*uncertain observations*). Since it is not generally known whether the observation used for estimation contains the signal or it is only noise, and only the probabilities of occurrence of such cases are available to the estimation, the observation equation is designed by including a random multiplicative noise described by a sequence of Bernoulli random variables, whose values - one or zero - indicate the presence or absence of the signal in the observations, respectively.

The least-squares optimal estimation problem in systems with uncertain observations is not easily treatable in general, due to the fact that the multiplicative noise perturbing the observations causes that the joint distribution of the signal and the observations is not gaussian (even if the signal and additive noises are gaussian processes). For this reason, the research on the estimation problem in these systems has been focused on the search of suboptimal estimators for the signal that can be easily derived. Nahi (1969) was the first who described this observation model and analyzed the linear least-squares estimation problem in linear systems with independent uncertainty. After that, numerous studies have been developed in this context, assuming different hypotheses on the Bernoulli random variables modelling the uncertainty when the state-space model is known and, also, when only covariance information is available (see Nakamori et al. (2005) and references therein).

On the other hand, there are many practical applications in communication theory (phase modulation of analog communication systems, object tracking in video sequences, robot navigation, location tracking, navigation sensors, etc.) where the observations are not linear function of the signal to be estimated. Although the estimation problem in discrete-time systems from uncertain observations has been extensively studied in linear systems, the literature on nonlinear filtering with uncertainty, which is the focus of this chapter, is fairly

limited, with the exception of a few results such as those reported in NaNacara & Yaz (1997) and, more recently, in Hermoso & Linares (2007) and Nakamori et al. (2009).

Nonlinear filtering is an interesting research area in which many approaches have been developed, the most popular being the extended Kalman filter (see e.g. Simon (2006), among others), which approximates the optimal estimator by linearizing the nonlinear system equations around the last state estimate to generate a linear system to which the Kalman filter equations can be applied. This technique provides approximations of the mean and covariance of the signal which are accurate, at least, up to the first terms of their Taylor series expansions. Assuming full knowledge of the state-space model of the signal to be estimated, the extended Kalman filter has been widely applied by different authors. For example, in Angrisani et al. (2006) the discrete extended Kalman filter is used to estimate the shape factors of ultrasonic echo envelopes. Boussak (2005) addressed the speed and rotor position estimation problem of interior permanent magnet synchronous motor drive through an extended Kalman filter algorithm. The node localization problem in a delay-tolerant sensor network is studied in Pathirana et al. (2005) using an estimation technique based on the robust extended Kalman filter. Routray et al. (2002) applied an extended Kalman filter to the frequency estimation problem of distorted signals in power systems. When the state equations are unknown and only the covariance functions of the processes involved in the observation equation are available, Nakamori (1999) derived filtering and fixed-point smoothing algorithms for discrete-time systems with nonlinear observation mechanism, by using a similar idea to the extended Kalman filter.

Although the extended Kalman filter has been successfully applied to numerous nonlinear discrete systems, the use of truncated Taylor expansion yields some important drawbacks involving, on the one hand, the evaluation of the Jacobian matrices and, on the other, its instability. Among other nonlinear techniques, the unscented Kalman filtering (see e.g. Julier & Uhlmann (2004)), which does not require the calculation of Jacobian matrices, is a relatively new one that improves the extended one, providing approximations of the mean which are accurate up to the second term of its Taylor expansion.

Different generalizations of the extended and the unscented Kalman filters have been proposed in Hermoso & Linares (2007) for a class of nonlinear discrete-time systems with additive noises, using uncertain observations; from comparison between both techniques, superior performance of the unscented filter is also found for this class of systems.

The current chapter is concerned with the state estimation problem for nonlinear discrete-time systems with uncertain observations, when the evolution of the state is governed by nonlinear functions of the state and noise, and the additive noise of the observation is correlated with that of the state. The random interruptions in the observation process are modelled by a binary white noise taking either the value one (when the measurement is the current system output) or the value zero (when only noise is available). A filtering algorithm is designed using the scaled unscented transformation, which provides approximations of the first and second-order statistics of a nonlinear transformation of a random vector. This algorithm extends to that proposed in Hermoso & Linares (2007) in two directions. On the one hand, we consider a more general state transition model in which the noise is not necessarily additive and, on the other, the independence between the state and observation noises is removed, thus addressing those situations in which the observation noise is correlated with the state.

The chapter is organized as follows: in Section 2 the system model is described; more specifically, we introduce the nonlinear state transition model, perturbed by a white noise, and the observation model, governed by nonlinear functions of the state affected by an

additive white noise correlated with that of the state and a multiplicative noise describing the uncertainty. In Section 3 the least-squares estimation problem from uncertain observation is formulated and a brief review of the unscented transformation and the scaled unscented transformation is presented. Next, in Section 4, the estimation algorithm is derived using the unscented filtering procedure, which acts in the prediction and update steps. The filter update is accomplished by the Kalman filter equations, which require the conditional statistics of the observation; hence, the correlation between the state and observation noise must be taken into account in this phase. Finally, the performance of the proposed unscented filter is illustrated in Section 5 by a numerical simulation example, where a first order ARCH model is considered to describe the state evolution.

Keywords

Nonlinear stochastic systems, Uncertain observations, Unscented Kalman filter.

2. Nonlinear model: system description and assumptions

In some practical situations, there exist random failures in the observation mechanism, accidental loss of some measurements, or data inaccessibility during certain times; this causes that the measurements may be either the current system output or only noise. This occurs, for instance, in tracking systems where the observations may either contain actual output contaminated with noise or be noise alone, and only the probabilities of occurrence of such cases are available to the estimation.

Our aim is to estimate an n-dimensional discrete-time state process, $\{x_k;\ k \geq 0\}$, whose evolution is perturbed by a q-dimensional white noise, $\{w_k;\ k \geq 0\}$, and governed by known functions of the state and noise; that is:

$$x_{k+1} = f_k(x_k, w_k), \quad k \geq 0, \tag{1}$$

where $f_k : \mathbb{R}^{n+q} \to \mathbb{R}^n$ is assumed to be continuously differentiable with respect to x_k and w_k. Consider that the nonlinear observation, y_k, is either the current system output (with probability p_k) or only noise (with probability $1 - p_k$), and assume that this occurs independently at different sampling times. So, considering independent random variables $\gamma_k \in \{0, 1\}$, $k \geq 1$, with the understanding that $\gamma_k = 1$ means that the measurement at time k is the current system output and $\gamma_k = 0$ means that only noise is available, and assuming that $P[\gamma_k = 1] = p_k$, the observation model is specified by nonlinear functions of the state perturbed by additive white noise, $\{v_k;\ k \geq 1\}$, and multiplicative noise, $\{\gamma_k;\ k \geq 1\}$, describing the uncertainty; that is:

$$y_k = \gamma_k h_k(x_k) + v_k, \quad k \geq 1, \tag{2}$$

where $h_k : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable functions.

The first and second-order moments of the processes determining the evolution of the state and describing the observations are specified by the following hypotheses.

(H1) The initial state, x_0, is a random vector with mean $\overline{x}_0$ and covariance P_0.

(H2) The state and observation white noises, $\{w_k;\ k \geq 0\}$ and $\{v_k;\ k \geq 1\}$, respectively, are correlated zero-mean processes with covariance matrices

$$\begin{aligned} E[w_k w_k^T] &= Q_k, \quad k \geq 0 \\ E[v_k v_k^T] &= R_k, \quad k \geq 1 \\ E[w_j v_k^T] &= S_k \delta_{j,k-1}, \quad k \geq 1, \end{aligned}$$

where δ denotes the Kronecker delta function.

(H3) The multiplicative noise $\{\gamma_k;\ k \geq 1\}$ describes the uncertainty in the observations and is a sequence of independent Bernoulli random variables with known probabilities $P[\gamma_k = 1] = p_k$. The probability $1 - p_k$, named *false alarm probability*, represents the probability that the observed value at time k does not contain the signal.

(H4) x_0, $\Big(\{w_k;\ k \geq 0\}, \{v_k;\ k \geq 1\}\Big)$ and $\{\gamma_k;\ k \geq 1\}$ are mutually independent.

3. Least-squares estimation problem from uncertain observations

The least-squares estimator of the state x_k from the observations $Y^k = \{y_1, \ldots, y_k\}$ is the conditional expectation of x_k given Y^k,

$$E[x_k/Y^k] = \int x_k g(x_k/Y^k) dx_k,$$

and, hence, the computation of this conditional mean requires the knowledge of $g(x_k/Y^k)$, the conditional density function of x_k given Y^k.

Due to the uncertainty in the observations, this conditional density function is a mixture or weighted sum of 2^k conditional density functions (corresponding to the different values of $\gamma_1, \ldots, \gamma_k$) and, moreover, computation of these conditional densities is generally difficult (even if the distributions of the processes involved in the system are known) due to the nonlinearity of the functions f_k and h_k. These severe drawbacks have motivated the search of suboptimal estimators based on approximations of the conditional mean to address the estimation problem in systems with uncertain observations and, more generally, in nonlinear systems.

One of the most frequently used methods to address the estimation problem in nonlinear systems without uncertainty in the observations (i.e. system models like (1)-(2) with $\gamma_k = 1,\ \forall k$) is the well-known *extended Kalman filter* (Simon (2006)), based on the linearization of the state and observation equations. However, as indicated in Julier & Uhlmann (2004), the extended Kalman filter has serious handicaps, which should be kept in mind when it is used; in particular, the performance of this filter can be very poor if the functions f_k and h_k present intense nonlinearities. In such cases, alternative approximations improving the estimation must be used. Among others, the *unscented Kalman filter* is a superior alternative to the extended one in a great variety of application domains, including state estimation.

Among other advantages, the unscented Kalman filter overcomes the deficiencies of linearization of the extended Kalman filter by providing an algorithm based on a direct, explicit mechanism for transforming the mean and covariance information when a nonlinear function is considered (Julier & Uhlmann (2004)). Unlike the linearization operation of the extended Kalman filter, the unscented Kalman filter uses the nonlinear models directly; it captures the posterior mean and covariance accurately up to the terms corresponding to the third-order moments in the Taylor series expansions, for the Gaussian distribution, and at least up to second-order for an arbitrary distribution.

Various generalizations of the extended and unscented Kalman filters were proposed in Hermoso & Linares (2007) for a class of nonlinear discrete-time systems with uncertain observations, when the state and observation noises are additive and the Bernoulli variables modelling the uncertainty are independent; from comparison of both techniques, superior performance of the unscented filter is also found in this uncertainty case.

We propose a modification of the unscented Kalman filter for estimating the state of the nonlinear system model with uncertain observations described in Section 2. This filter provides an approximation to the conditional mean $E\left[x_k/Y^k\right]$ based on the use of the unscented transformation; more precisely, we will use an extension of this transformation, called scaled unscented transformation. Both transformations are briefly described below (see Julier & Uhlmann (2004) for details).

3.1 Unscented Transformation (UT)

This is a method for approximating the mean and covariance matrix of a random vector, $Y = g(X)$, from the mean and covariance of X. The idea is to choose deterministically a fixed number of points and weights which capture the mean and covariance of X exactly; then, the mean and covariance of $g(X)$ are approximated by the weighted sample mean and covariance of the transformed points.

Specifically, if X is an N-dimensional random vector with mean $\widehat{X}$ and covariance P_X, the UT considers $2N+1$ points and weights, $\{(\xi_i, \psi_i),\ i = 0,\dots,2N\}$, called *sigma points*, which are defined as follows:

$$\begin{aligned}
&\xi_0 = \widehat{X}\\
&\xi_i = \widehat{X} + \left(\sqrt{(N+\kappa)P_X}\right)_i,\ \ i = 1,\dots,N\\
&\xi_i = \widehat{X} - \left(\sqrt{(N+\kappa)P_X}\right)_{i-N},\ \ i = N+1,\dots,2N\\
&\psi_0 = \frac{\kappa}{N+\kappa}\\
&\psi_i = \frac{1}{2(N+\kappa)},\ \ i = 1,\dots,2N,
\end{aligned} \tag{3}$$

where κ is a scaling parameter which can be used to capture additional information on the distribution of X, and $(A)_j$ denotes the j-th column of a matrix A.

Although $\sum_{i=0}^{2N} \psi_i = 1$, the sigma points $\{(\xi_i, \psi_i),\ i = 0,\dots,2N\}$ do not necessarily define a probability distribution since κ can be a negative number (the only condition is $N+\kappa > 0$); however, its moments can be defined as in a discrete probability distribution, and it is easy to prove that the first and second-order moments of the sigma points are equal to those of X.

To approximate the statistics of a transformation $Y = g(X)$, each point ξ_i is propagated through the function g, and the first and second order moments of Y are approximated by those of the transformed sigma points $g(\xi_i)$, $i = 0,\dots,2N$. Therefore, the mean of Y is approximated by the weighted average of the transformed points,

$$\widehat{Y} \approx \sum_{i=0}^{2N} \psi_i g\left(\xi_i\right),$$

the covariance of Y is approximated by

$$P_Y \approx \sum_{i=0}^{2N} \psi_i \left(g\left(\xi_i\right) - \widehat{Y}\right)\left(g\left(\xi_i\right) - \widehat{Y}\right)^T$$

and the cross-covariance of X and Y is approximated by

$$P_{XY} \approx \sum_{i=0}^{2N} \psi_i \left(\xi_i - \widehat{X}\right)\left(g\left(\xi_i\right) - \widehat{Y}\right)^T.$$

If g is an analytic function, the approximations of the mean and covariance of $g(X)$ are accurate up to the second and first term of their Taylor expansion series, respectively. However, the approximations are inaccurate if the higher-order sample moments have a great effect on the Taylor expansions; this can occur, for example, if the dimension, N, of vector X is very large, since the radius of the sphere containing the sigma points increases with N. This drawback can be avoided by using the scaled unscented transformation.

3.2 Scaled Unscented Transformation (SUT)

The SUT considers a set of sigma points, $\chi_i = \xi_0 + \alpha(\xi_i - \xi_0)$, $i = 0, \ldots, 2N$, where α is a scaling parameter which can be arbitrarily small. The points χ_i have basically the same form as in (3), just replacing κ by $\lambda = \alpha^2(N + \kappa) - N$; the associated weights, calculated in order to capture the mean and covariance of X, are now

$$W_0 = \frac{\psi_0}{\alpha^2} + (1 - 1/\alpha^2)$$

$$W_i = \frac{\psi_i}{\alpha^2},\ i = 1, \ldots, 2N.$$

Besides reducing the dispersion of the sigma points considered, the SUT allows to modify them in order to prevent nonpositive semidefinite approximated covariances (which can occur if $W_0 < 0$), as well as to incorporate additional information on the fourth-order moments of X; this is achieved by modifying the weight of χ_0 in the approximation of the covariance, which improves the precision in this approximation. Thus, the sigma points and weights in the SUT are specified as follows:

$$\begin{aligned}
&\chi_0 = \widehat{X}\\
&\chi_i = \widehat{X} + \left(\sqrt{(N+\lambda)P_X}\right)_i,\ \ i = 1, \ldots, N\\
&\chi_i = \widehat{X} - \left(\sqrt{(N+\lambda)P_X}\right)_{i-N},\ \ i = N+1 \ldots, 2N\\
&W_0^{(m)} = \frac{\lambda}{N+\lambda}\\
&W_0^{(c)} = W_0^{(m)} + (1 - \alpha^2 + \beta)\\
&W_i^{(m)} = W_i^{(c)} = \frac{1}{2(N+\lambda)},\ \ i = 1, \ldots, 2N\\
&\lambda = \alpha^2(N + \kappa) - N,
\end{aligned}$$

where α is the scaling parameter (usually a small value), and κ and β are used to incorporate prior information on the distribution of X ($\kappa = 3 - N$ and $\beta = 2$ provide the optimal values if X has a Gaussian distribution).

The mean and covariance of a transformation $g(X)$ are approximated, respectively, by the sample mean and covariance of the transformed values, $g(\chi_i)$, $i = 0, \ldots, 2N$, with weights $W_i^{(m)}$ for the mean and $W_i^{(c)}$ for the covariance. The cross-covariance of X and $Y = g(X)$ is approximated by the sample cross-covariance of χ_i, $i = 0, \ldots, 2N$ and the transformed values, $g(\chi_i)$, $i = 0, \ldots, 2N$, with weights $W_i^{(c)}$.

4. Unscented filtering algorithm

The aim is to obtain an estimator of x_k, the system state at time k described in (1), based on the observations given in (2) up to that time, $Y^k = \{y_1, \ldots, y_k\}$; for this purpose we compute an approximation, $\widehat{x}_{k/k}$, of the conditional mean $E\left[x_k/Y^k\right]$. As usual, the estimator $\widehat{x}_{k/k}$ will be obtained from the estimator at the previous time, $\widehat{x}_{k-1/k-1}$, through the following prediction and update steps:

(i) *Prediction:* Taking into account the relationship (1), approximations $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$ of $E\left[x_k/Y^{k-1}\right]$ and $Cov\left[x_k/Y^{k-1}\right]$, respectively, are obtained by applying a SUT to the nonlinear transformation $x_k = f_{k-1}(x_{k-1}, w_{k-1})$; then, this step requires us to work jointly with the state and noise vectors x_{k-1} and w_{k-1}.

(ii) *Update:* When the predictor $\widehat{x}_{k/k-1}$ is available, it is updated with the new observation y_k to obtain an approximation of $E[x_k/Y^k]$ and $Cov[x_k/Y^k]$; this is achieved using the following expression, with a similar structure to those of the Kalman filter:

$$E[x_k/Y^k] \approx \widehat{x}_{k/k} = \widehat{x}_{k/k-1} + Cov\left[x_k, y_k/Y^{k-1}\right]\left(Cov\left[y_k/Y^{k-1}\right]\right)^{-1}\left(y_k - E\left[y_k/Y^{k-1}\right]\right).$$

This expression require the conditional statistics of y_k; specifically, it is necessary to approximate the conditional mean and covariance, $E\left[y_k/Y^{k-1}\right]$ and $Cov\left[y_k/Y^{k-1}\right]$, as well as the conditional cross-covariance $Cov\left[x_k, y_k/Y^{k-1}\right]$.

Hence, in view of (2), the correlation between x_k and v_k must be taken into account in this step. More specifically, since $x_k = f_{k-1}(x_{k-1}, w_{k-1})$, the correlation between and w_{k-1} and v_k must be taken into account.

These reasons lead us to work jointly with the vectors x_{k-1}, w_{k-1} and v_k and hence, we define the following $(n+q+m)$-dimensional augmented vectors:

$$X_k = \begin{pmatrix} x_k \\ w_k \\ v_{k+1} \end{pmatrix}, \quad k \geq 0.$$

The problem is then reformulated as that of finding the filter of this augmented vector, $\widehat{X}_{k/k}$, whose first n-dimensional block-component provides the filter for the original state.

The prediction and update steps are detailed in the following subsections.

4.1 Unscented algorithm: prediction step

The starting points of the proposed algorithm are the filter and the covariance matrix at the initial state X_0 which, from the model hypotheses, are given by:

$$\widehat{X}_{0/0} = E[X_0] = \begin{pmatrix} \overline{x}_0 \\ 0 \\ 0 \end{pmatrix}, \quad P^{XX}_{0,0/0} = Cov[X_0] = \begin{pmatrix} P_0 & 0 & 0 \\ 0 & Q_0 & S_1 \\ 0 & S_1^T & R_1 \end{pmatrix}.$$

For each $k > 1$, we start with approximations $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$ of the conditional mean and covariance of X_{k-1} given Y^{k-1} which, from the independence between (w_{k-1}, v_k)

and Y^{k-1}, the conditional independence between x_{k-1} and (w_{k-1}, v_k) and hypothesis (H2), are given by:

$$\widehat{X}_{k-1/k-1} = \begin{pmatrix} \widehat{x}_{k-1/k-1} \\ 0 \\ 0 \end{pmatrix}, \qquad P^{XX}_{k-1,k-1/k-1} = \begin{pmatrix} P^{xx}_{k-1,k-1/k-1} & 0 & 0 \\ 0 & Q_{k-1} & S_k \\ 0 & S_k^T & R_k \end{pmatrix}.$$

The aim is to find approximations $\widehat{X}_{k/k-1}$ and $P^{XX}_{k,k/k-1}$ for the conditional mean and covariance of X_k given Y^{k-1} which, reasoning similarly, are

$$\widehat{X}_{k/k-1} = \begin{pmatrix} \widehat{x}_{k/k-1} \\ 0 \\ 0 \end{pmatrix}, \qquad P^{XX}_{k,k/k-1} = \begin{pmatrix} P^{xx}_{k,k/k-1} & 0 & 0 \\ 0 & Q_k & S_{k+1} \\ 0 & S_{k+1}^T & R_{k+1} \end{pmatrix}. \tag{4}$$

Hence, we only need the conditional statistics $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$ of $x_k = f_{k-1}(x_{k-1}, w_{k-1})$, which are approximated from $\widehat{X}_{k-1/k-1}$, and $P^{XX}_{k-1,k-1/k-1}$ using the SUT, as follows:

- We consider a set of sigma-points

$$\left\{ \chi_{i,k-1/k-1} = \left(\chi^{xT}_{i,k-1/k-1} \mid \chi^{wT}_{i,k-1/k-1} \mid \chi^{vT}_{i,k/k-1} \right)^T,\ i = 0, \ldots, 2N \right\} \quad (N = n + q + m),$$

  whose mean and covariance are exactly $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$:

$$\begin{aligned} \chi_{0,k-1/k-1} &= \widehat{X}_{k-1/k-1} \\ \chi_{i,k-1/k-1} &= \widehat{X}_{k-1/k-1} + \left(\sqrt{(N+\lambda) P^{XX}_{k-1,k-1/k-1}} \right)_i, \quad i = 1, \ldots, N \\ \chi_{i,k-1/k-1} &= \widehat{X}_{k-1/k-1} - \left(\sqrt{(N+\lambda) P^{XX}_{k-1,k-1/k-1}} \right)_{i-N}, \quad i = N+1, \ldots, 2N \end{aligned} \tag{5}$$

  and their associated weights, $W_i^{(m)}$ for the mean and $W_i^{(c)}$ for the covariance:

$$\begin{aligned} W_0^{(m)} &= \frac{\lambda}{N+\lambda} \\ W_0^{(c)} &= \frac{\lambda}{N+\lambda} + (1 - \alpha^2 + \beta) \\ W_i^{(m)} &= W_i^{(c)} = \frac{1}{2(N+\lambda)}, \quad i = 1, \ldots, 2N \\ \lambda &= \alpha^2 (N + \kappa) - N \end{aligned}$$

  where α is a scaling parameter determining the spread of the sigma-points around $\widehat{X}_{k-1/k-1}$, and κ and β are tuning parameters.

- Then, by defining $f^a_{k-1}(X_{k-1}) = f_{k-1}(x_{k-1}, w_{k-1}) = x_k$, the mean and covariance of x_k given Y^{k-1} are approximated by the corresponding sample statistics of the transformed sigma-points, $f^a_{k-1}(\chi_{i,k-1/k-1}) = f_{k-1}\left(\chi^x_{i,k-1/k-1}, \chi^w_{i,k-1/k-1} \right)$:

$$\begin{aligned}\widehat{x}_{k/k-1} &= \sum_{i=0}^{2N} W_i^{(m)} f_{k-1}^a(\chi_{i,k-1/k-1}) \\ P_{k,k/k-1}^{xx} &= \sum_{i=0}^{2N} W_i^{(c)} \left(f_{k-1}^a(\chi_{i,k-1/k-1}) - \widehat{x}_{k/k-1}\right)\left(f_{k-1}^a(\chi_{i,k-1/k-1}) - \widehat{x}_{k/k-1}\right)^T.\end{aligned} \tag{6}$$

The conditional mean and covariance of X_k given Y^{k-1} are then approximated by (4) with $\widehat{x}_{k/k-1}$ and $P_{k,k/k-1}^{xx}$ given in (6).

4.2 Unscented algorithm: update step

The approximations $\widehat{X}_{k/k-1}$ and $P_{k,k/k-1}^{XX}$ given in (4) and (6) are now updated with the new observation, y_k, by using the Kalman filter equations

$$\widehat{X}_{k/k} = \widehat{X}_{k/k-1} + Cov\left[X_k, y_k/Y^{k-1}\right]\left(Cov\left[y_k/Y^{k-1}\right]\right)^{-1}\left(y_k - E\left[y_k/Y^{k-1}\right]\right)$$

$$P_{k,k/k}^{XX} = P_{k,k/k-1}^{XX} - Cov\left[X_k, y_k/Y^{k-1}\right]\left(Cov\left[y_k/Y^{k-1}\right]\right)^{-1} Cov\left[y_k, X_k/Y^{k-1}\right].$$

For this purpose, we need to approximate the conditional mean, $E\left[y_k/Y^{k-1}\right]$, and covariance, $Cov\left[y_k/Y^{k-1}\right]$, of y_k given Y^{k-1}, as well as the conditional cross-covariance of X_k and y_k given Y^{k-1}, $Cov\left[X_k, y_k/Y^{k-1}\right]$.

In systems with uncertain observations, the conditional distribution of $\gamma_k h_k(x_k)$ given Y^{k-1} has a mixture type whose components are the conditional distributions corresponding to $\gamma_k = 1$ and $\gamma_k = 0$, with mixture parameters $P\left[\gamma_k = 1/Y^{k-1}\right]$ and $P\left[\gamma_k = 0/Y^{k-1}\right]$, respectively. Since $P\left[\gamma_k = 1/Y^{k-1}\right] = p_k$ (which follows from (H3) and (H4)), the approximations of the conditional statistics of $\gamma_k h_k(x_k)$ are directly obtained using this mixture type, and, taking into account (2), the statistics of y_k given Y^{k-1} are expressed in terms of those corresponding to $z_k = h_k(x_k)$ and v_k as follows:

$$\begin{aligned}E[y_k/Y^{k-1}] &= p_k E[z_k/Y^{k-1}] \\ Cov[y_k/Y^{k-1}] &= p_k Cov[z_k/Y^{k-1}] + p_k(1-p_k)E[z_k/Y^{k-1}]E[z_k^T/Y^{k-1}] \\ &\quad + p_k Cov[z_k, v_k/Y^{k-1}] + p_k Cov[v_k, z_k/Y^{k-1}] + R_k \\ Cov[X_k, y_k/Y^{k-1}] &= p_k Cov[X_k, z_k/Y^{k-1}] + Cov[X_k, v_k/Y^{k-1}].\end{aligned} \tag{7}$$

Moreover, since z_k and v_k are conditionally independent of w_k and v_{k+1}, the conditional cross-covariances $Cov[X_k, z_k/Y^{k-1}]$ and $Cov[X_k, v_k/Y^{k-1}]$ require only the conditional cross-covariances of x_k with z_k and v_k, respectively; that is:

$$Cov[X_k, y_k/Y^{k-1}] = \begin{pmatrix} p_k Cov[x_k, z_k/Y^{k-1}] + Cov[x_k, v_k/Y^{k-1}] \\ 0 \\ 0 \end{pmatrix} \tag{8}$$

Then, we proceed to approximate the conditional statistics appearing in (7) and (8), which correspond to the vectors z_k, x_k and v_k.

- The first two vectors, z_k and x_k, are both functions of x_k and, consequently, their conditional statistics can be approximated from $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$ by considering a set of sigma-points, $\left\{\chi^x_{i,k/k-1},\ i=0,\ldots,2n\right\}$, whose mean and covariance are exactly $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$:

$$\begin{aligned}
\chi^x_{0,k/k-1} &= \widehat{x}_{k/k-1} \\
\chi^x_{i,k/k-1} &= \widehat{x}_{k/k-1} + \left(\sqrt{(n+\lambda)P^{xx}_{k,k/k-1}}\right)_i, \quad i=1,\ldots,n \\
\chi_{i,k-1/k-1} &= \widehat{x}_{k/k-1} - \left(\sqrt{(n+\lambda)P^{xx}_{k,k/k-1}}\right)_{i-n}, \quad i=n+1,\ldots,2n
\end{aligned} \tag{9}$$

  and their associated weights, $\omega_i^{(m)}$ for the mean and $\omega_i^{(c)}$ for the covariance:

$$\begin{aligned}
\omega_0^{(m)} &= \frac{\lambda}{n+\lambda} \\
\omega_0^{(c)} &= \frac{\lambda}{n+\lambda} + (1-\alpha^2+\beta) \\
\omega_i^{(m)} &= \omega_i^{(c)} = \frac{1}{2(n+\lambda)}, \quad i=1,\ldots,2n \\
\lambda &= \alpha^2(n+\kappa) - n.
\end{aligned}$$

  Then the statistics of $z_k = h_k(x_k)$ are approximated by those of the transformed sigma-points, $h_k(\chi^x_{i,k/k-1})$:

$$\begin{aligned}
E[z_k/Y^{k-1}] &\approx \widehat{z}_{k/k-1} = \sum_{i=0}^{2n} \omega_i^{(m)} h_k(\chi^x_{i,k/k-1}) \\
Cov[z_k/Y^{k-1}] &\approx P^{zz}_{k,k/k-1} = \sum_{i=0}^{2n} \omega_i^{(c)} \left(h_k(\chi^x_{i,k/k-1}) - \widehat{z}_{k/k-1}\right)\left(h_k(\chi^x_{i,k/k-1}) - \widehat{z}_{k/k-1}\right)^T \\
Cov[x_k, z_k/Y^{k-1}] &\approx P^{xz}_{k,k/k-1} = \sum_{i=0}^{2n} \omega_i^{(c)} \left(\chi^x_{i,k/k-1} - \widehat{x}_{k/k-1}\right)\left(h_k(\chi^x_{i,k/k-1}) - \widehat{z}_{k/k-1}\right)^T.
\end{aligned} \tag{10}$$

- The vector v_k, however, cannot be expressed in terms of $X_k = (x_k^T \mid w_k^T \mid v_{k+1}^T)^T$, but it is a function of X_{k-1}; so its conditional statistics must be approximated from those of X_{k-1}. Thus, expressing $z_k = h_k\left(f^a_{k-1}(X_{k-1})\right) = h_k\left(f_{k-1}(x_{k-1}, w_{k-1})\right)$ and using the sigma-points

$$\chi_{i,k-1/k-1} = \left(\chi^{xT}_{i,k-1/k-1} \mid \chi^{wT}_{i,k-1/k-1} \mid \chi^{vT}_{i,k/k-1}\right)^T, \quad i=0,\ldots,2N$$

associated to $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$, the following approximations are used:

$$\begin{aligned} Cov[z_k, v_k/Y^{k-1}] &\approx P^{zv}_{k,k/k-1} = \sum_{i=0}^{2N} W_i^{(c)} h_k\left(f_{k-1}(\chi^x_{i,k-1/k-1}, \chi^w_{i,k-1/k-1})\right)\chi^{vT}_{i,k/k-1} \\ Cov[x_k, v_k/Y^{k-1}] &\approx P^{xv}_{k,k/k-1} = \sum_{i=0}^{2N} W_i^{(c)} f_{k-1}(\chi^x_{i,k-1/k-1}, \chi^w_{i,k-1/k-1})\chi^{vT}_{i,k/k-1}. \end{aligned} \tag{11}$$

Finally, these statistics are substituted in (7) and (8) to obtain approximations $\widehat{y}_{k/k-1}$, $P^{yy}_{k,k/k-1}$ and $P^{Xy}_{k,k/k-1}$ of the conditional statistics of y_k,

$$\begin{aligned} &\widehat{y}_{k/k-1} = p_k \widehat{z}_{k/k-1} \\ &P^{yy}_{k,k/k-1} = p_k P^{zz}_{k,k/k-1} + p_k(1-p_k)\widehat{z}_{k/k-1}\widehat{z}^T_{k/k-1} + p_k P^{zv}_{k,k/k-1} + p_k P^{vz}_{k,k/k-1} + R_k, \\ &P^{Xy}_{k,k/k-1} = \begin{pmatrix} p_k P^{xz}_{k,k/k-1} + P^{xv}_{k,k/k-1} \\ 0 \\ 0 \end{pmatrix}. \end{aligned} \tag{12}$$

These approximations are used in the following equations providing the filter of X_k and the corresponding filtering error covariance matrix:

$$\begin{aligned} \widehat{X}_{k/k} &= \widehat{X}_{k/k-1} + P^{Xy}_{k,k/k-1}\left(P^{yy}_{k,k/k-1}\right)^{-1}(y_k - \widehat{y}_{k/k-1}), \quad k \geq 1 \\ P^{XX}_{k,k/k} &= P^{XX}_{k,k/k-1} - P^{Xy}_{k,k/k-1}\left(P^{yy}_{k,k/k-1}\right)^{-1} P^{yX}_{k,k/k-1}, \quad k \geq 1, \end{aligned} \tag{13}$$

with initial conditions

$$\widehat{X}_{0/0} = \begin{pmatrix} \overline{x}_0 \\ 0 \\ 0 \end{pmatrix}, \qquad P^{XX}_{0,0/0} = \begin{pmatrix} P_0 & 0 & 0 \\ 0 & Q_0 & S_1 \\ 0 & S_1^T & R_1 \end{pmatrix}. \tag{14}$$

Computational summary
In summary, given $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$, the above results suggest the following recursive computational procedure to obtain the proposed unscented filter:

(I) Compute the sigma-points given in (5), whose mean and covariance are $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$, respectively, and, with them:

(Ia) Compute $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$ by (6).

(Ib) Compute $P^{zv}_{k,k/k-1}$ and $P^{xv}_{k,k/k-1}$ by (11).

(II) Compute the sigma-points given in (9), whose mean and covariance are $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$, respectively, and, with them, compute $\widehat{z}_{k/k-1}$, $P^{zz}_{k,k/k-1}$, and $P^{xz}_{k,k/k-1}$ by (10).

(III) From *(Ia)*, compute $\widehat{X}_{k/k-1}$ and $P^{XX}_{k,k/k-1}$ by (4).

(IV) From *(Ib)* and *(II)*, compute $\widehat{y}_{k/k-1}$, $P^{yy}_{k,k/k-1}$ and $P^{Xy}_{k,k/k-1}$ by (12).

(V) From *(III)* and *(IV)*, compute $\widehat{X}_{k/k}$ and $P^{XX}_{k,k/k}$ by (13).

The initial conditions of the proposed algorithm, $\widehat{X}_{0/0}$ and $P^{XX}_{0,0/0}$, are given in (14).

Finally, by extracting the first n-dimensional block-components of $\widehat{X}_{k/k}$ and $P^{XX}_{k,k/k}$, the filter of the original state vector, x_k, and the filtering error covariance matrix are obtained, respectively.

Remark 1: Although the derivation of the algorithm does not requires that the functions f_k and h_k are continuously differentiable, these hypotheses guarantee that the approximations of the conditional mean and covariances are accurate, at least, up to the first and second terms of their Taylor series expansions, respectively.

Remark 2: The proposed algorithm reduces to that in Hermoso & Linares (2007) when the functions f_k are linear in the noise, and the state and observation noises are uncorrelated. Moreover, the unscented filter agrees with the optimal linear one when the functions f_k and h_k are linear.

5. Numerical simulation results

In this section, a numerical simulation example is presented to illustrate the application of the proposed unscented filter. The application deals with a first order autoregressive conditional heteroscedastic model (ARCH (1)); these models, introduced by Engle in 1982 and widely known in volatility modelling in finance (Peiris & Thavaneswaran (2007)), have been considered in Tanizaki (2000) as an example to compare the performance of various nonlinear filters when the observed variables consist of a sum of the ARCH (1) process and an independent error term.

Here, according to the theoretic study, we assume that the measurements can be only the error term with a known probability, and that the noise process is correlated with the ARCH (1) process.

Let us consider that the evolution of the state is described by the following discrete-time multiplicative transition equation

$$x_{k+1} = \sqrt{a + bx_k^2}\, w_k, \;\; k \geq 0,$$

where the initial state x_0 is a Gaussian variable with zero mean and unity variance, the noise $\{w_k;\, k \geq 0\}$ is a zero-mean Gaussian process with variance $Q_k = 1$ and $a = 1 - b$ is taken to normalize the unconditional variance of x_k to be one.

Uncertain observations of the state with additive noise are considered for the estimation:

$$y_k = \gamma_k x_k + v_k, \;\; k \geq 1,$$

where $\{v_k;\, k \geq 1\}$ is a zero-mean white process with variance $R_k = 1$, and the multiplicative noise, $\{\gamma_k;\, k \geq 1\}$, is a sequence of independent Bernoulli variables with constant known probability $P[\gamma_k = 1] = p$.

The state and additive observation noises $\{w_k;\, k \geq 0\}$ and $\{v_k;\, k \geq 1\}$ are assumed to be joint Gaussian processes with known and constant cross-covariance $S_k = S, \forall k$.

We have implemented a MATLAB program that simulates the state x_k for $b = 0.5$, and the uncertain measurements, y_k, for $k = 1, \ldots, 50$, for different values of S and p, and provides the unscented filtering estimates of x_k.

The root mean square error (RMSE) criterion was used to quantify the performance of the estimates. Considering 1000 independent simulations and denoting by $\{x_k^{(s)},\; k = 1, \ldots, 50\}$

the s-th set of the artificially simulated states and by $\widehat{x}_{k/k}^{(s)}$ the filtering estimate at time k in the s-th simulation run, the RMSE of the filter at time k is calculated by

$$\mathrm{RMSE}_k = \left(\frac{1}{1000} \sum_{s=1}^{1000} \left(x_k^{(s)} - \widehat{x}_{k/k}^{(s)} \right)^2 \right)^{1/2}.$$

Let us first examine the performance of the algorithm for different values of S; since analogous results are obtained for opposite correlations S and $-S$, only nonnegative values are considered in the simulations shown here.

Figure 1 displays the RMSE_k when the uncertainty probability is $p = 0.5$ and different values of S are considered; specifically, $S = 0,\ 0.3,\ 0.5,\ 0.7$ and $S = 0.9$; this figure shows, as expected, that the higher the value of S (which means that the correlation between the state and the observations increases) the smaller that of RMSE_k and, consequently, the performance of the estimators is better. Analogous results are obtained for other different values of p and S.

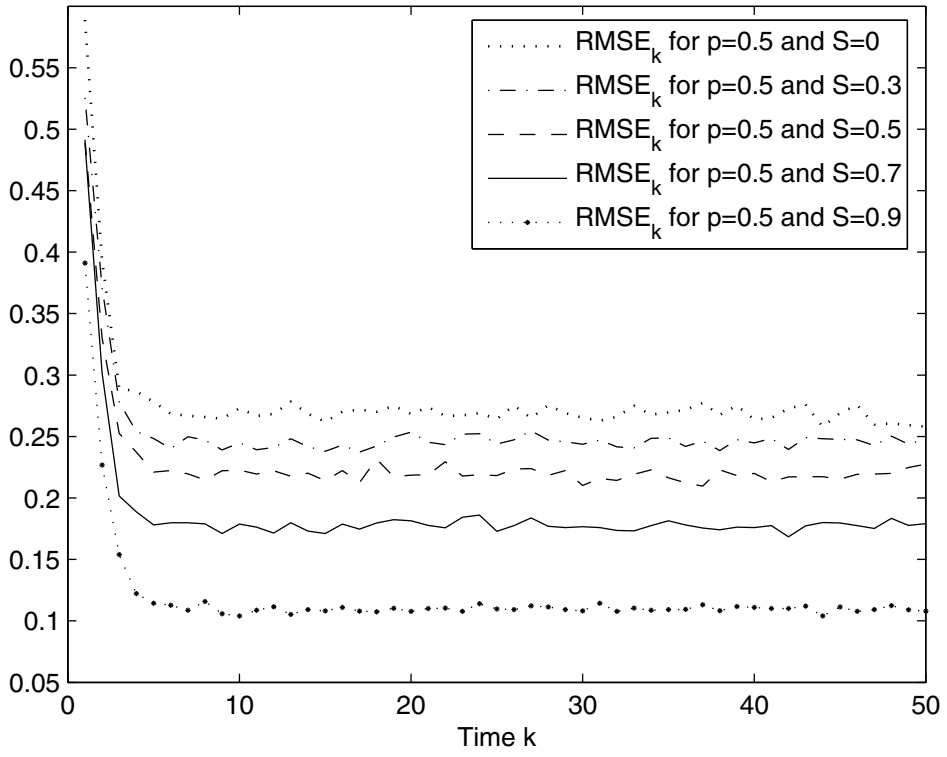


Fig. 1. RMSE_k for the unscented filtering estimates when $p = 0.5$ and $S = 0,\ 0.3,\ 0.5,\ 0.7,\ 0.9$.

Moreover, in order to compare the performance of the estimators as a function of the uncertainty probability p, the means of RMSE_k corresponding to the 50 iterations were

calculated for the different values of S considered in Figure 1 and $p = 0.1,\ 0.2, \ldots, 0.9$. The results are shown in Figure 2, from which it is apparent that the means decrease when p increases (that is, when the probability that the observations contain the state is greater) and consequently, as expected, the performance of the estimators deteriorates as the probability p falls. From this figure, it is also inferred that, for each fixed value of p, the means decrease as S increases, which extends the result in Figure 1 to different values of p.

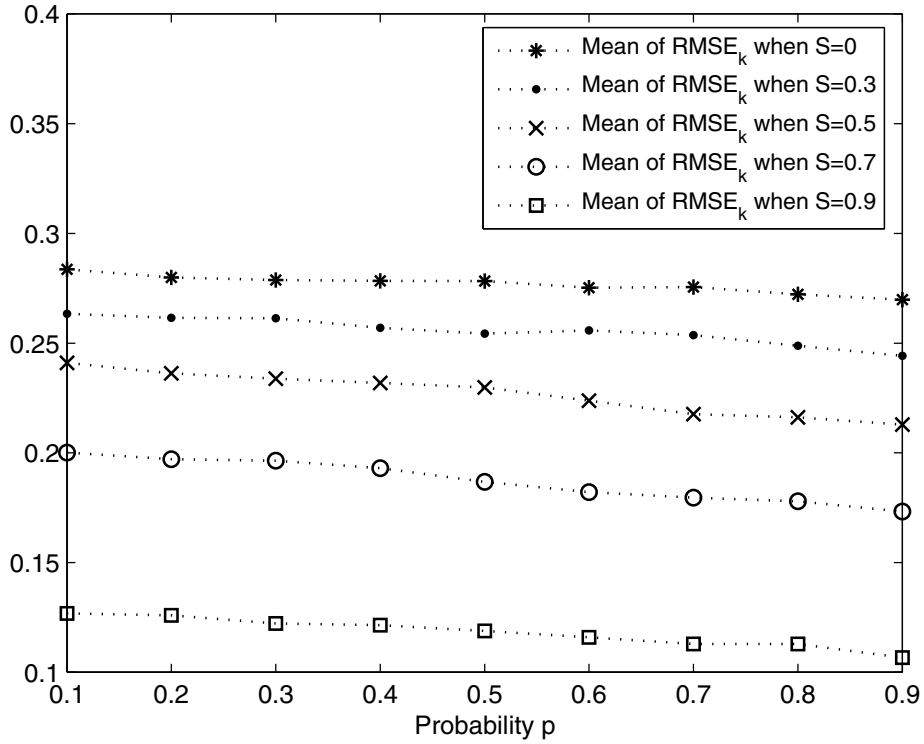


Fig. 2. Mean of RMSE_k for the unscented filtering estimates when $S = 0,\ 0.3,\ 0.5,\ 0.7,\ 0.9$, versus p.

6. Conclusion

In this chapter, a recursive unscented filtering algorithm for state estimation in a class of nonlinear discrete-time stochastic systems with uncertain observations is obtained. The uncertainty is modelled by a binary white noise taking the value one (when the measurement is the current system output) or zero (when only noise is observed), and the additive noise of the observation is correlated with that of the state.

We propose a filtering algorithm based on the scaled unscented transformation, which provides approximations of the first and second-order statistics of a nonlinear transformation of a random vector.

This algorithm extends to that in Hermoso & Linares (2007) in two directions. On the one hand, we consider a more general state model in which the noise is not necessarily additive

and, on the other, the independence between the state and observation noises is removed, thus covering those situations in which the observation noise is correlated with the state. The algorithm performance is illustrated with a simulation example in which a first-order ARCH model is considered to describe the state evolution.

7. Acknowledgments

This research is supported by Ministerio de Educación y Ciencia (grant No. MTM2008-05567) and Junta de Andalucía (grant No. P07-FQM-02701).

8. References

Angrisani, L.; Baccigalupi, A. & Lo Moriello, R. S. (2006). A measurement method based on Kalman filtering for ultrasonic time-of-flight estimation. *IEEE Transactions on Instrumentation and Measurement*, Vol. 55, No. 2, pp. 442–448, ISSN 0018-9456

Boussak, M. (2005). Implementation and experimental investigation of sensorless speed control with initial rotor position estimation for interior permanent magnet synchronous motor drive. *IEEE Transactions on Power Electronics*, Vol. 20, No. 6, pp. 1413–1422, ISSN 0885-8993

Hermoso-Carazo, A. & Linares-Pérez, J. (2007). Different approaches for state filtering in nonlinear systems with uncertain observations. *Applied Mathematics and Computation*, Vol. 187, pp. 708–724, ISSN 0096-3003

Julier, S.J. & Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, Vol. 92, No. 3, pp. 401–422, ISSN 0018-9219

Nahi, N. (1969). Optimal recursive estimation with uncertain observation. *IEEE Transactions on Information Theory*, Vol. 15, pp. 457–462, ISSN 0018-9448

Nakamori, S. (1999). Design of estimators using covariance information in discrete-time stochastic systems with nonlinear observation mechanism. *IEICE Transactions on Fundamentals of Electronics, Comunications and Computer Sciences*, Vol. E82-A, No. 7, pp. 1292–1304, ISSN 0916-8508

Nakamori, S.; Caballero-Águila, R.; Hermoso-Carazo, A. & Linares-Pérez, J. (2005). Fixed-interval smoothing algorithm based on covariances with correlation in the uncertainty. *Digital Signal Processing*, Vol. 15, pp. 207–221, ISSN 1051-2004

Nakamori, S.; Caballero-Águila, R.; Hermoso-Carazo, A.; Jiménez-López, J.D. & Linares-Pérez, J. (2009). Signal estimation with nonlinear uncertain observations using covariance information. *Journal of Statistical Computation and Simulation*, Vol. 79 , No. 1, pp. 55–66, ISSN 0094-9655

NaNacara, W. & Yaz, E.E. (1997). Recursive estimator for linear and nonlinear systems with uncertain observations. *Signal Processing*, Vol. 62, pp. 215–228, ISSN 0165-1684

Pathirana, P. N.; Bulusu, N.; Savkin, A. V. & Jha, S. (2005). Node localization using mobile robots in delay-tolerant sensor networks. *IEEE Transactions on Mobile Computing*, Vol. 4, No. 3, pp. 285–296, ISSN 1536-1233

Peiris, S. & Thavaneswaran, A. (2007). An introduction to volatility models with indices. *Applied Mathematics Letters*, Vol. 20, pp. 177–182, ISSN 0893-9659

Routray, A.; Pradhan, A. K. & Rao, K. P. (2002). A novel Kalman filter for frequency estimation of distorted signals in power systems. *IEEE Transactions on Instrumentation and Measurement*, Vol. 51, No. 3, pp. 469–479, ISSN 0018-9456

Simon, D. (2006). *Optimal State Estimation: Kalman, H_∞ and Nonlinear Approaches*, John Wiley & Sons, Inc., ISBN 9780471708582, New Jersey, U.S.A.

Tanizaki, H. (2000). Nonlinear and non-Gaussian estimation: a quasi-optimal estimator. *Communications in Statistics, Theory and Methods*, Vol. 29, No. 12, pp. 2805–2834, ISSN 0361-0926

8

Numerical Validation Methods

Ricardo Jauregui and Ferran Silva
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

1. Introduction

In the last years, numerical simulation has seen a great development thanks to costs reduction and speed increases of the computational systems. With these improvements, the mathematical algorithms are able to work properly with more realistic problems. Nowadays, the solution of a problem using numerical simulation is not just finding a result, but also to ensure the quality. However, can we say that the model results are correct regarding the behaviour of the system? In other words, how could we quantify the similarity between reality and simulations? To answer these questions, it is necessary to establish a validation criterion that allows an objective quantification of the difference between the results and the reality. Another way to say this is, how "true" our results are.
In the case of numerical methods, the main objective is to replicate as closely as possible the behaviour of the "real" world through numbers. Normally, the results of the numerical methods are expressed in terms of graphics, pictures, etc. These results represent the view of reality that the chosen method provides (Oñate, 1998). In order to affirm that the result of a numerical solution is fully consistent with the reality, it must be satisfied that:

a. The mathematical model must incorporate all aspects of the real world.
b. The numerical method has to solve exactly the equations of the mathematical modelling.

The problem starts with these two conditions that guarantee the "truth" of the results, since none of them are fully accomplished and it must be admitted that the numerical prediction never completely matches the "real" world behaviour. Then you can only be sure that the numerical solution is a good approximation of the reality. Now, new questions arise: How much does the result obtained by a numerical method resemble the reality? How can we objectively quantify this similarity? The answers to these questions are those that give rise to the validation methods.

2. Types of validation

All validation is done through a comparison of a pattern or a reference model with the model under study. There are many ways to make a validation, but in general they are usually classified according to the pattern used in the comparison (Godoy & Dardati, 2001), (Archambeault & Connor, 2008):

a. **Validation using other numerical solutions.** This technique compares the results to be validated with the results obtained through other numerical methods previously

validated. In other words, one technique has been validated before it can be used as a reference to validate the second method.

Another way to use this technique is using more than one numerical method to solve the problem. If the physics of the problem is properly modelled in all the techniques used, the results should have a clear trend and similarity; therefore, knowing the advantages and disadvantages of each technique it will be possible to perform a validation of our technique.

b. **Validation using analytical solutions**. This type of comparison can be used when the researcher knows the analytical theory behind the problem and makes a direct comparison of the simulation results with the analytical solution. One of the main problems of this technique is that it can only be used in extremely simple cases, because trying to find the analytical solution of real problems is almost impossible (in fact, this is the reason of making numerical simulations). However, this technique is useful when you want to validate the code of the numerical method. Through the analytical solution it is possible to obtain the exact value of the problem, thereby reducing the external variables that can affect the results.

c. **Validation using experimental results.** This technique is the most popular of all; this is due mainly to the fact that the measurement shows the consistency of the model with the reality. However, one cannot forget that whenever you perform a measurement you should introduce a measuring instrument and this directly or indirectly affect the system being measured (Archambeault & Connor, 2008). For this reason, it is essential to have the greatest similarity between the measurement and simulation configurations. For the real environment (measurement) one should take into account the possible limitations of the laboratory and the equipment required to perform the measurement. The most important issue is to narrow down to a minimum any device that cannot be fully simulated such as cables, connectors, etc. On the other hand, for the computing environment (simulation) one should try to model the entire possible setup or at least, include the most important characteristics. Otherwise, it runs the risk that the simulation results do not represent the reality faithfully, causing a validation error.

d. **Validation using intermediate results.** This technique compares the intermediate results of the numerical model with experimental or theoretical known values, although these results are not the final objective of our comparison. The major drawback of this method is to find an intermediate result that is really close related with the final result under study. On the other hand, it is very easy to lose sight of the factors that could affect the intermediate variable making the comparison of the final result not valid. However, this technique is frequently used to monitor some parameters of the numerical simulations, but it is rarely used alone or as a main validation method.

A good example of this technique can be found in electromagnetic simulations. Imagine that it is required to compare the far-field simulations and measurements produced by a source inside an airplane. In this case, to make far-field measurements in a structure as big as an aircraft can be very expensive and complicated. However, it is possible to measure some near-field values at specific points near the aircraft and to compare them with simulations results calculated at the same points. Based on the direct relationship between the far-field and near–field, the similitude between simulation and existing measurements in the near field will be proportionately the same than in the far-field. As

can be seen in this example, the validation is not done with the final results (far-field), but an intermediate result (near-field) is used to set a criterion of similarity between simulation and measurement.

Validation using convergence. This type of validation is based on a comparison of the convergence of the numerical model with the pattern or the reference results. This comparison is done knowing that the solution found is not the best, but assuming that the model results converge.

Another situation where this type of validation can be used is when it is impossible to get a "pattern" to do the comparison. A good example of this type of validation can be found in chaos in structural mechanics area (Awrejcewicz & Krysko, 2008), where the validation is used in two ways; first, using a method to solve the system individually and observing if it converges or not. The second one is to use various methods to analyze the system at the same time and consider whether they all converge towards the same result and which of them do it faster.

In the electromagnetic area this technique of validation is often used when we want to know a general behaviour of the system in a very short time. Normally, a very simple model with very coarse meshes is simulated. Then, the most important resonances and the general behaviour are observed to analyze whether or not we are on track. The only drawback of this technique is that it is not recommended for use as a final method of validation, because convergence of the system cannot be guaranteed.

Regardless of the type of comparison that is performed (analytical, numerical, experimental, etc) at the end, the validation process is reduced to the comparison of the results and, in many cases, to the comparison of a pair of graphs. After that, once the type of validation to be performed in our model is chosen, the problem is how we can compare our results with the pattern in a quantitative mode. In many fields of research, simple visual inspection is used as the validation method when faced with the need to compare their results with the established models or patterns.

Although the visual comparison is used in different environments with apparent reliability, it has potential limitations. Among the most common problems are:

- The eye concentrates on peak positions and ignores the poor correlation of intensities (D. E. Coleby & A. P. Duffy, 2002).
- Due to the potentially subjective nature of the visual comparison, the results produced cannot be used with confidence (D. E. Coleby & A. P. Duffy, 2002),(A. Duffy, D. Coleby, A. Martin, M. Woolfson, & T. Benson, 2003).
- Comparing and quantifying the results objectively between different groups of experts can be difficult (Williams, M. S. Woolfson, T. M. Benson, & A. P. Duffy, 1997), (A. Duffy, D. Coleby, A. Martin, M. Woolfson, & T. Benson, 2003).
- The data may be too large (either a high volume of data or a very complex topography) to be compared visually with ease (Williams et al., 1997).

These limitations force the need to investigate reliable and objective computational techniques to compare the differences of the results and evaluate their quality.

3. Validation methods

Numerous studies show that a direct comparison point by point is not feasible when large amounts of data are compared (D. E. Coleby & A. P. Duffy, 2002), (Drozd, 2005), (Archambeault & Connor, 2008). Therefore, this method is not recommended to validate the

results and much less to assign an absolute value of accuracy. This approach makes sense only for simple models, but in the numerical simulations, the results are often very complex. Today there are several methods of validation. Among the most used are:

3.1 Correlation

This is a widely used method for its ease implementation and interpretation and it is intended for quantitative variables for which there is a linear relationship. The correlation between two variables is perfect when output value is closest to 1 or -1 and gets worse as it approaches to 0 (D. E. Coleby & A. P. Duffy, 2002). The sign indicates the direction of the association: a value equal +1 indicates a perfect positive linear relationship. When this case happens the relationship between two variables has exactly the same behaviour: when one of them increases, the other increases too. If instead of that, the value is -1, it is said that there is a perfect negative relationship and implies that both signals have a linear relationships, one will decrease as the other increases.

The most popular type of correlation is called the "Pearson correlation coefficient" and is usually used to measure the strength of the relationship between two variables when there is a linear relationship between them.

The Pearson correlation coefficient is defined by the following expression:

$$R_{pe} = \frac{\sum_{i=0}^{n} Y1_{(i)} \cdot Y2_{(i)} - \frac{\left(\sum_{i=0}^{n} Y1_{(i)} \cdot \sum_{i=0}^{n} Y2_{(i)}\right)}{n}}{\sqrt{\left(\sum_{i=0}^{n} Y1_{(i)}^{2} - \frac{\left(\sum_{i=0}^{n} Y1_{(i)}\right)^{2}}{n}\right) \cdot \left(\sum_{i=0}^{n} Y2_{(i)}^{2} - \frac{\left(\sum_{i=0}^{n} Y2\right)^{2}}{n}\right)}} \tag{1}$$

Where $Y1_{(i)}$ is the dataset 1, $Y2_{(i)}$ the dataset 2 and n is the total number of points in both data sets.

The major limitation of this correlation technique is that it only can be used when the relationship between the variables is linear. This means that when variables are closely related, but not linearly, the validation results can not reflect the expert opinion. There are very few cases that have a linear relationship between variables so this method is only used for extremely basic cases.

An additional problem with this method, even when the data sets to compare have a linear relationship, is the interpretation a determined coefficient value. Or how does one know if a value is high or low? The answers to these questions depend largely on the nature of the investigation and the sample size used. For example, a correlation of 0.01 may be significant in a sufficiently large sample and a 0.9 may not be in a small sample. The law of large numbers is fulfilled, being that the weak trends are very unlikely from a null hypothesis and large amounts of data, while strong trends may be relatively likely in the small data size.

3.2 Reliability factor

This method is known as the R-Factor (The Reliability Factor) and it is one of the main criteria accepted in the validation area. The R factor could be considered a type of correlation; it is an objective method that provides with a single number, the similarity between two data. This method was created mainly to compare the intensities between the experimental and theoretical results in structural determinations of X-rays. A variation of different R-factors have been proposed: The first of these was introduced by Zanazzi and

Jona (Zanazzi & Jona, 1977), followed by Van Hove (Van Hove, 1977) and Pendry (Pendry, 1980).

3.2.1 Zanazzi and Jona R-factor

The R-factor of Zanazzi and Jona is also known as "R_{ZJ} -factor" (Zanazzi & Jona, 1977), and was planned to study the similarity of X-Ray diffraction and surface crystallography. Further studies modified their original equations and applied them in the numerical simulations area (Williams et al., 1997). This method was designed to accentuate the maximum slopes rather than the heights. This is accomplished by comparing the gradients of the two signals you want to compare. It basically made the differences between the signals for the first and second derivatives, accentuating the features present; making the R-Factor sensitive to positional changes in the data. The equation used to calculate this R-Factor is as follows:

$$\mathrm{R_{ZJ}} = \frac{\sum_{\mathrm{i}=0}^{\mathrm{n}} \mathrm{W(i)} \,.\, \mathrm{F(i)}}{\sum_{\mathrm{i}=0}^{\mathrm{n}} \mathrm{Y1(i)}} \tag{2}$$

$$W(f) = \frac{\left|Y1''_{(i)} - C.Y2''_{(i)}\right|}{\left|Y1'_{(i)}\right| + \left|\max(Y1'_{(i)})\right|} \tag{3}$$

$$F(f) = \left|Y1'_{(i)} - C.Y2'_{(i)}\right| \tag{4}$$

$$C = \frac{\sum_{i=0}^{n} Y1_{(\mathrm{i})}}{\sum_{i=0}^{n} Y2_{(\mathrm{i})}} \tag{5}$$

Where $Y1_{(i)}'$ and $Y1_{(i)}''$ is the first and the second derivative of dataset 1. $Y2_{(i)}'$ and $Y2_{(i)}''$ is the first and the second derivative of dataset 2. *Y1* is normally used with the experimental value *Y2* obtained and the theoretical value or reference pattern. *C* is used to adjust the intensity in both dataset.

This technique is useful when comparing sharp signals with peaks where the importance lies in their peaks or valleys, but does not offer the same reliability when you want to compare noisy signals where the peaks are not as marked and the variations between valleys is very fast.

3.2.2 Pendry R-factor

The Pendry R-Factor (Pendry, 1980) is used generally to measure the degree of correlation between two signals that have many variations in their maximum positions. This method uses derivatives in place of their intensities; this is attributed to all peaks of the same weight, regardless of the height of each one. The idea of this technique is that any maximum contains structural information due a constructive interference. Thus the maximum occurring at high energies are generally lower than those obtained at low. The equation used to calculate this R-factor is as follows:

$$R_p = \frac{\sum_{i=0}^{n} (Y_{SET1} - Y_{\mathrm{SET2}})^2}{\sum_{i=0}^{n} (Y_{SET1}^2 + Y_{SET2}^2)} \tag{6}$$

$$Y_{SET} = \frac{L_{SET}^2}{L_{SET}} \tag{7}$$

$$L = \frac{SET'}{SET} \tag{8}$$

Where L is the intensity and L' is the differentiated intensity in each dataset. $Y1_{(i)}$ is the dataset 1, $Y2_{(i)}$ the dataset 2 and n is the total number of points in both data sets.
Unlike "Zanazzi and Jona" factor where it is necessary to calculate the second derivative, the Pendry R-Factor only requires the first derivative, making it less susceptible to small or rapid changes. This feature makes it a useful tool for analyzing very noisy signals. However, the main problem with the Pendry method is that it requires finding an adjustable parameter that is not constant (Robertson et al., 1990), which seriously restricts the use of this technique.

3.2.3 Van Hove R-factor

The technique of Van Hove (Van Hove, 1977) is the most widespread of all the R-Factor ones. This technique uses five different equations (9)-(15) to compare the position and width of the signal peaks; the shape of the peaks and the troughs, their number and their heights (D. E. Coleby & A. P. Duffy, 2002). The different indicators of this method are calculated using the following equations:

$$R_1 = \frac{\sum_{i=0}^{n} \left| Y1_{(\mathrm{i})} - C.Y2_{(\mathrm{i})} \right|}{\sum_{i=0}^{n} \left| Y1_{(i)} \right|} \tag{9}$$

$$C = \frac{\sum_{i=0}^{n} Y1_{(\mathrm{i})}}{\sum_{i=0}^{n} Y2_{(\mathrm{i})}} \tag{10}$$

$$R_2 = \frac{\sum_{i=0}^{n} \left(Y1_{(\mathrm{i})} - C.Y2_{(\mathrm{i})} \right)^2}{\sum_{i=0}^{n} \left(Y1_{(i)} \right)^2} \tag{11}$$

Where $Y1_{(i)}$ is the dataset 1 and $Y2_{(i)}$ is the dataset 2. Both indicators (R1&R2) show the similarity in positions, heights and widths of peaks and troughs.

$$R_3 = \frac{Nº\ slopes^{+}(Y1)}{Nº\ slopes^{-}(Y1)} - \frac{Nº\ slopes^{+}(Y2)}{Nº\ slopes^{-}(Y2)} \tag{12}$$

Where $N^o slopes^{+}$ is the number of positive slopes and $N^o slopes^{+}$ is the number of negative slopes for each dataset. The R3 indicator compares the number of positive slopes with the negative slopes of the opposite graph.

$$R_4 = \frac{\sum_{i=0}^{n} \left| Y1'_{(\mathrm{i})} - C.Y2'_{(\mathrm{i})} \right|}{\sum_{i=0}^{n} \left| Y1'_{(i)} \right|} \tag{13}$$

$$R_5 = \frac{\sum_{i=0}^{n}\left(Y1'_{(i)} - C.Y2'_{(i)}\right)^2}{\sum_{i=0}^{n}\left(Y1'_{(i)}\right)^2} \tag{14}$$

Where Y1′ is the first derivate of dataset 1 and Y2′ is the first derivate of dataset 2. In this case the R4 and R5 indicators are used to compare the gradient of the data sets.
Finally, the Van Hove Factor has a very useful indicator that combines all indicators to calculate the total difference between the two graphs, it is called: "R_T" (15). This indicator allows to quickly and accurately getting an overall idea of how good a result is with regard to the pattern.

$$R_T = \sqrt{R_1^2 + R_2^2 + R_3^2 + R_4^2 + R_5^2} \tag{15}$$

3.3 Integrated with logarithmic frequency error (IELF)

The idea behind IELF (Integrated against Error Log Frequency) is based on the premise that when comparing data with a very high feature density, the overriding factor to be assessed is a function of the difference between the two traces (Simpson, Jones, MacDiarmid, A. Duffy, & D. Coleby, 2005). Basically, this method is the difference between two traces in logarithmic axis and in frequency domain. Then the result is integrated (summing) to get a single value. The IELF equation is given in (16):

$$IELF = \frac{\sum_{i=0}^{n-1}|error_i|.\left[\ln(f_{(i+1)}) - \frac{\ln(f_{(i-1)})}{2}\right]}{\ln(f_{(n)}) - ln(f_{(0)})} \tag{16}$$

Where *f* are the frequency points being compared (from point 0 to point n, resulting in n+1 discrete frequencies), $|error|$ is the difference between the two data sets at the nth data point.
There is an improved version of IELF method known as IELF modified ($IELF_{MOD}$). This modification involves summing the elements halfway between the data points in order to improve the approximation of the difference in the measured data. This modification is given in equation (17):

$$IELF_{MOD} = \frac{\sum_{i=1}^{n-1}\left|error_{(i)}\right|.\left[\frac{\ln(f_{(i+1)} + f_i)}{2} - \frac{\ln(f_{(i-1)} + f_{i-1})}{2}\right]}{\ln(f_{(n)}) - ln(f_{(0)})} \tag{17}$$

Today, the ELF_{MOD} method is widely used to validate large volumes of data. This method is used in circumstances in which the data to be compared has a high visual density, ie, it is impossible to differentiate visually (Knockaert, Catrysse, & Belmans, 2006), (A. Duffy, D. Coleby, A. Martin, M. Woolfson, & T. Benson, 2003). The main disadvantage of this method is that it has very few tools to perform a good validation; only one indicator is very little to interpret all the aspects present in the validation process. Another important disadvantage is that it is defined only for the frequency domain in logarithmic axis and there are some weaknesses with abrupt changes in graphics.

3.4 Feature Selective Validation Method (FSV)

The method of Feature Selective Validation (FSV) was developed by Anthony Martin and Alistair Duffy in 1999 (A. Martin, 1999) and today is the method most widely used because of its versatility and simplicity. This method is widespread and is currently being developed as a standard of validation for Computational ElectroMagnetics (CEM) within the project by IEEE 1597.1/1597.2 (Standard IEEE, 2008). The Feature Selective Validation (FSV) method was developed with the specific aim to reflect the approach taken by engineers when assessing data presented visually during the validation of computational electromagnetic simulation. Today it is possible to find two FSV free online software, the first one was developed by the Aquila University in Italy (Orlandi, 2006). The second one has been developed by the Electromagnetic Compatibility Group of the Universitat Politècnica de Catalunya (GCEM, 2011). GCEM-UPC has built and developed new tools for the traditional FSV allowing the user to evaluate the graphics in a very quick and easy way (for more information visit: http://www.upc.edu/web/gcem/files/FSV.exe).

The FSV method is based on the decomposition of the results into two groups; the first one discusses the difference in amplitude (Amplitude Difference Measure, ADM) and the second one the difference between the characteristic of the signals (Feature Difference Measure, FDM). The combination of these two indicators (ADM and FDM) is a measurement of the overall difference (Global Difference Measure, GDM) (A. P. Duffy et al., 2006; Orlandi et al., 2006).

All indicators ADM, FDM and GDM have the ability to be configured to perform a point-to-point analysis. The advantage of relying on a point-to-point data is to know which areas of the data sets have the major differences. A subscript "i" is added to consider this point-by-point feature (ADMi, FDMi and GDMi).

Another way to qualitatively analyse the FSV indicators is represented by a probability density function. It is useful for a rapid and comprehensive analysis of the results. This indicator uses a histogram that can be divided into six categories: excellent, very good, good, fair, poor, very poor.

Finally, a technique that has proved useful in presenting and interpreting FSV data, particularly the confidence histograms, is a "Grade and Spread" (G/S) diagram (Archambeault, A. P. Duffy, & Orlandi, 2009; Archambeault & Yu, 2009).The Spread serves a similar purpose to variance or standard deviation in statistical methods and is a measurement of the spread of a distribution. The Grade is a measurement of the quality of the results and serves a similar purpose to skew measurements in statistics. It is important to remember that Grade and Spread must be used together, since if only one is used, the interpretation can be inaccurate.

The FSV method requires a serie of steps to obtain each indicator, a brief summary of some of them are following described:

a. The first step is to interpolate the two sets of data to be compared to having the same number of samples for comparison.
b. Once both datasets have the same number of samples, the Fourier Transformer is applied. Then, a high pass filter is applied in the dataset obtained the "Hi" data. The same procedure with a band pass filter is done to obtain the "Lo" data. An important aspect to consider is that these two new dataset (Hi and Lo) are separated by the breaking point, which is chosen with 40% of all data.
c. Knowing all values of "Lo" data, the ADM indicator is calculated according to:

$$\mathrm{ADMi}_{(i)} = \frac{(|\mathrm{Lo}_1(i)| - |\mathrm{Lo}_2(i)|)}{\frac{1}{n}\sum_{j=1}^{n}|(|\mathrm{Lo}_1(j)| + |\mathrm{Lo}_2(j)|)|} + \mathrm{ODM}_{(i)} \cdot e^{\mathrm{ODM}(i)} \tag{18}$$

$$ODM_{(i)} = \frac{x(i)}{\delta(i)} \tag{19}$$

$$\mathrm{x}_{(i)} = (|\mathrm{DC}_1(i)| + |\mathrm{DC}_2(i)|) \tag{20}$$

$$\delta_{(i)} = \frac{1}{n}\sum_{j=1}^{n}(|\mathrm{DC}_1(j)| + |\mathrm{DC}_2(j)|) \tag{21}$$

d. The next step calculated is the FDM. It is composed of three parts based on the derivatives calculated in the last step. The numerical values in the equations are parts of the heuristic and have been determined empirically.

$$\mathrm{FDM}_{1(i)} = \frac{(|\mathrm{Lo}_1'(i)| - |\mathrm{Lo}_2'(i)|)}{\frac{2}{n}\sum_{j=1}^{n}|(|\mathrm{Lo}_1'(j)| + |\mathrm{Lo}_2'(j)|)|} \tag{22}$$

$$\mathrm{FDM}_{2(i)} = \frac{(|\mathrm{Hi}_1'(i)| - |\mathrm{Hi}_2'(i)|)}{\frac{6}{n}\sum_{j=1}^{n}|(|\mathrm{Hi}_1'(j)| + |\mathrm{Hi}_2'(j)|)|} \tag{23}$$

$$\mathrm{FDM}_{3(i)} = \frac{(|\mathrm{Hi}_1''(i)| - |\mathrm{Hi}_2''(i)|)}{\frac{7.2}{n}\sum_{j=1}^{n}|(|\mathrm{Hi}_1''(j)| + |\mathrm{Hi}_2''(j)|)|} \tag{24}$$

$$\mathrm{FDM}_{(i)} = 2\left(\left|\mathrm{FDM}_{1(i)} + \mathrm{FDM}_{2(i)} + \mathrm{FDM}_{3(i)}\right|\right) \tag{25}$$

e. The GDMi indicator is calculated using the ADM and FDM indicators, as shown in equation (26).

$$\mathrm{GDMi}_{(i)} = \sqrt{\mathrm{ADM}_{(i)}^2 + \mathrm{FDM}_{(i)}^2} \tag{26}$$

f. Calculation of the mean value (XDMtot). After the ADM, FDM and GDM point-by-point values are calculated, it is possible to find the average value (27). These indicators are very useful to evaluate the quality of the results with one number.

$$\mathrm{XDMtot} = \frac{\sum_{i=1}^{n}\mathrm{XDMi}(i)}{n} \tag{27}$$

g. Calculation of the confidence histogram. This is the term that is most often used in the descriptions of the quality of comparisons. The determination of the histogram (Fig 1.) is simply a case of counting the proportion of points that fall into one of the categories, according to the rule base in Table 1.

XDMc value (X=A,F,G)	
XDMc≤0.1	Excellent
0.1≤XDMc≤0.2	Very Good
0.2≤XDMc≤0.4	Good
0.4≤XDMc≤0.8	Fair
0.8≤XDMc≤1.6	Poor
1.6≤XDMc	Very Poor

Table 1. XDMc interpretation scale.

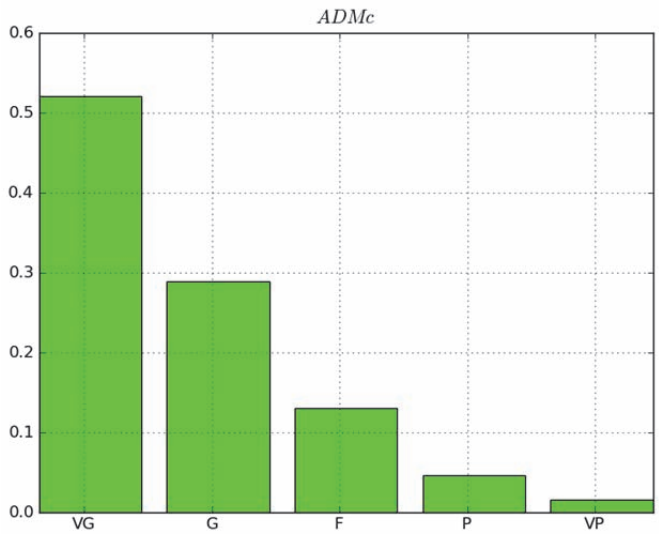


Fig. 1. Confidence histogram example (ADMc).

It is quite possible that in some cases there is significant spread between the different categories (Excellent (Ex), Very Good (VG), Good (G), Fair (F), Poor (P), and Very Poor (VP)). Often this spread is caused when the datasets are very noisy. A quick and easy solution is to use a wider window width which produces a greater degree of smoothing.

The "Grade" and the "Spread" are calculated based on the ADMc and FDMc. The "Grade" is calculated by taking the number of category, starting from the best (Excellent) to the worst (very poor), which include a user defined amount (named "threshold", an 85 % is recommended) of total samples of data sets to being compared. The "Spread" is similar to a typical standard deviation since it also determines how many categories are required to include 85% of the data, but the starting category is the highest rated (instead of the excellent category, as in the Grade).

Despite all the benefits offered by the FSV, in the validation processes, some studies found problems when trying to use it to analyze transient signals (R Jauregui, Riu, & Silva, 2010; Ri Jauregui, Silva, Orlandi, Sasse, & A. Duffy, 2010). In particular, the main problem was in the ADM indicator.

The problem with the ADM indicator lies in the way it calculates the dataset "Lo" (R Jauregui, Rojas-Mora, & Silva, 2011). This dataset is obtained through the breakpoint (IBP),

which is calculated with the 40% of the signal, because it assumes that most of the signal energy content is within this range. The problem is that many times in a transient signal up to 90% of the energy can be contained in the first peaks. Therefore, it is highly probable that only the first peak of the transient is considered, and the other low level differences are not taken into account when comparing with FSV method. For this reason, before using this method to validate a new type of signal that has not been previously tested, it is recommend making a small review and analyzing the consistency of results.

3.5 Validation Transient Signals in Time Domain (VTTD)

One of the most interesting signals from the viewpoint of the numerical simulation is the impulsive noise, also known as transient phenomenon. These types of signals are used in the time domain for analyzing large frequency bands. On the other hand, these same signals offer a challenge in areas such as electromagnetism or resistance of structures in building.

The transient signals could be described as a signal that varies between two consecutive steady states during a short period of time compared to the time scale of interest. In other words, there must be a "momentary" change of the magnitude seen for a very short time in the sense that this short interval of time should be much less than one cycle of the signal. Thanks to the different derivatives performed, it only takes into account the changing intervals in the graphs without giving attention to the level differences affected. This particular feature makes it difficult to analyze the effect using common methods of validation (R Jauregui, 2009). For this reason, a special method was developed to perform the Validation of the Transient in Time Domain (VTTD). It is proposed to use five indicators to assess the different parameters of the transient data sets:

a. **Feature Difference Measure (FSV-FDM).** The calculation of this indicator is made using the equations of the FDM specified in the FSV method (see equation (25)). Unlike the amplitude indicator (ADM), the FDM does not present any problem when it is used to analyse the transient in time domain. Thanks to the different derivatives performed in this indicator, it only takes into account the changing intervals in the graphs without attention to the level differences.

   This indicator is applied before taking into account any other indicators, because its value determines whether or not to continue the validation process. Some studies (R Jauregui, Pous, Fernández, & Silva, 2010; R Jauregui, Silva, & Riu, 2007) determined that the optimal limit for a correct interpretation is when the FDM_{TOT} indicator is equal or less than 0.8. This value ensures that the two data sets (numerical simulation and measurement, for example) have a similarity that is within the acceptable margin.

b. **Amplitude Pulse Level (APL).** This indicator measures the difference between the maximum amplitude of the signals. The maximum level of a transient signal is very important because it can produce several types of problems. Thus, the APL indicator aims to assess the maximum amplitude level difference between the two data sets.

   According to the equations (28) the APL calculates the difference of the maximum of each data set in absolute value to guarantee that the analysis is independent of the polarity.

$$APL = \frac{|\max(Y1) - \max(Y2)|}{|\max(Y1)|} \qquad |\max(Y1)| \geq |\max(Y2)| \tag{28}$$

$$APL = \frac{|\max(Y2) - \max(Y1)|}{|\max(y2)|} \qquad |\max(Y1)| < |\max(Y2)| \tag{29}$$

Where max (Y1) is the maximum magnitude for the first dataset and max (Y2) is the corresponding one for the second dataset. An APL range result is from 0 to 1. When the APL is equal to 0 the similarity is perfect, but as they increase, the result moves to 1.

c. **Maximum Rise Time (MRT)**. One important issue in a transient signal is the rise time. The lower it is, the more contents of the disturbance are on the high-frequency band, which is usually a problem in validation analysis. The calculation of this indicator is very similar to the one used in APL; the only difference is that it calculates the first derivative (30) and then applies the equations (31) or (32).

$$D_i^j = \left(\frac{y_i^j - y_{i-1}^j}{x_i^j - x_{i-1}^j}\right) \qquad \begin{matrix} j = \{Y1, Y2\} \\ i = \{1,2,3 \dots n\} \end{matrix} \tag{30}$$

$$MRT = \frac{|\max(D^{Y1}) - \max(D^{Y2})|}{|\max(D^{Y1})|} \qquad |\max(D^{Y1})| \geq |\max(D^{Y2})| \tag{31}$$

$$MRT = \frac{|\max(D^{Y2}) - \max(D^{Y1})|}{|\max(D^{Y2})|} \qquad |\max(D^{Y1})| < |\max(D^{Y2})| \tag{32}$$

Where *i* is the number of the point (from 1 to n). *j* is the set of the graph that we want to analyze (*Y1* is the first one and *Y2* the second one). D_i is the derivative for each point (1 to n) for both dataset. $|max(D^{Y1})|$ is the absolute maximum value of the derivative of the first dataset and $|\max(D^{Y2})|$ is the derivative of the second one. Similar to the APL indicator, the equation (31) or (32) is applied in order to ensure that MTR varies from 0 to 1.

d. **Energy Contained in the Signals (ECS).** This indicator measures the energy contained in the transients. In many cases, a transient energy could be very significantly affecting the behaviour of a system; therefore, it is important to evaluate. Applying the equations (34) or (35), the difference of energy between both datasets can be determined for the same interval of time.

$$E^j = \int_{t1}^{tn} U^j(t)\, dt \qquad \begin{matrix} j = \{Y1, Y2\} \\ t_1 < t_n \end{matrix} \tag{33}$$

$$ECS = \frac{|(E^{Y1}) - E^{Y2}|}{|E^{Y1}|} \qquad |E^{Y1}| \geq |E^{Y2}| \tag{34}$$

$$ECS = \frac{|(E^{Y2}) - E^{Y1}|}{|E^{Y2}|} \qquad |E^{Y1}| < |E^{Y2}| \tag{35}$$

Where "E" represent the energy and "U" is the magnitude recorded for each dataset, both datasets must be defined from t = 0 to $t = t_n$.

e. **The Total Error Average (TEA)**. As seen in previous validation methods, it is very useful to have an indicator that reflects the overall quality of results. This indicator

allows a quick and simple way to have a general idea of the quality of results. The indicator TEA meets this objective quickly and easily.

The calculation of TEA is based on finding the squared error of the indicators FDM, APL, MRT and ECS as shown in equation (36). In this equation, a weighting factor for each indicator can be defined by "α", "β", "γ" to highlight the importance of a particular indicator in a particular situation.

$$\mathrm{TEA} = \sqrt{\frac{\alpha.(\mathrm{APL}^2) + \beta.(\mathrm{MRT}^2) + \gamma.(\mathrm{ECS}^2)}{\alpha + \beta + \gamma}} \tag{36}$$

The intervals of this indicator may vary depending on the transient and the type of problem being analyzed. It is the user who must define the scales and values that define it. For example, in the case of Electromagnetic Compatibility (EMC) area, a useful scale to analyze the transients is:

Good: from 0 to 0.3

Regular: from 0.3 to 0.5.

Bad: from 0.5 to 1.

This method allows rapid and objective quantification of the simulation results, but it is important to note that this validation method is valid only to study the transient in time domain.

4. Validation method application examples

In order to show the application of the different validation methods previously presented, two real cases are chosen to compare their results. Before making any comparison, it is necessary to normalize all validation methods to ensure that all of them are within the same scale (from 0 to 1). As it is usual, we use different categories to help us to identify the quality of the results: excellent (from 0 to 0.16), very good (from 0.17 to 0.34), good (from 0.35 to 0.5), fair (from 0.51 to 0.65), poor (from 0.66 to 0.80), very poor (from 0.81 to 1). Finally, a survey among some experts has been done to compare their opinion with the different validation methods under test.

The VTTD method was not used in these examples, as it is defined only for transient analysis in time domain. If you need more information about the use or implementation of this method, it is recommended to view documents (R Jauregui, 2009; R Jauregui, Riu, & Silva, 2010; Riu, R Jauregui, Silva, & Fernandez, 2007).

4.1 First case of study

The aim of this first example is to examine the efficiency of each validation method to analyze the similarity between two signals. These signals were obtained by the measurement (Fig. 2-blue) and the simulation (Fig. 2-red) and show the transfer function between two electromagnetically short monopoles inside a resonant cavity. As one can observe, the simulation and the measurement have a similar behaviour for the entire frequency range. However, some minor differences can be found in the negative resonances. In general, the results were classified by a panel of experts with 0.25, which is in the "Very good" category.

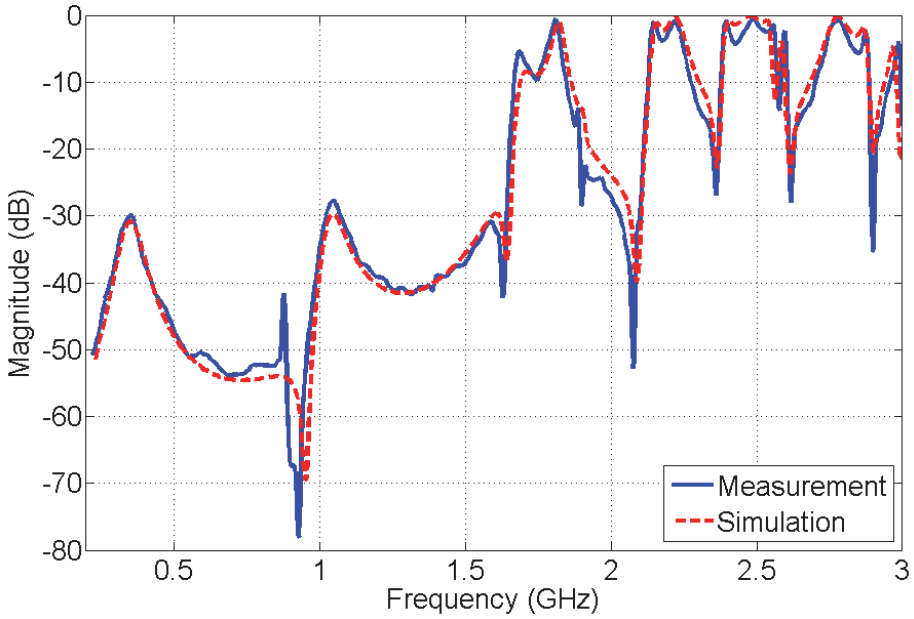


Fig. 2. Comparison between measurement and simulation.

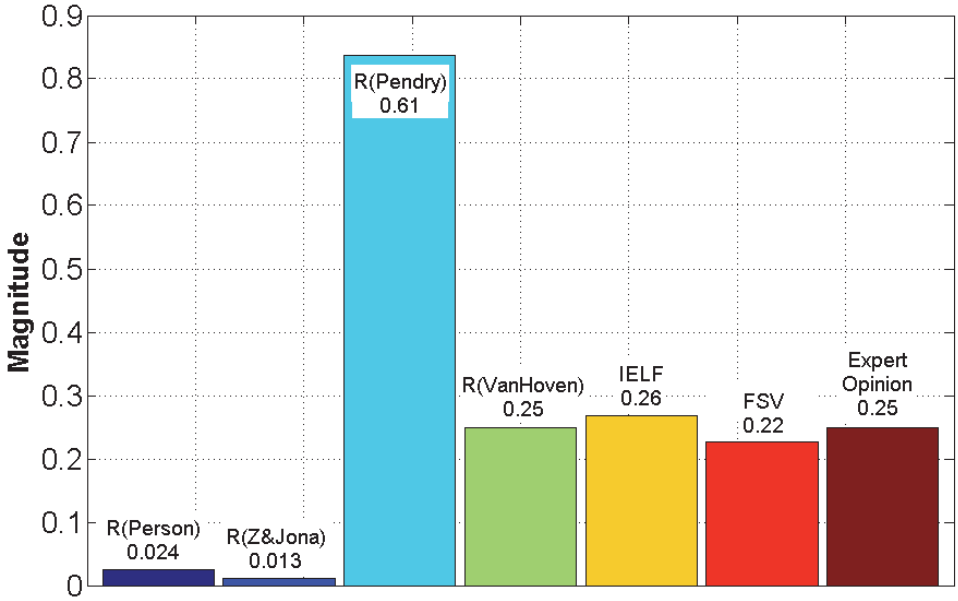


Fig. 3. Results from the different validation methods used and the experts opinion.

Applying the different validation methods studied before, we obtain the results shown in the Fig 3. It is possible to observe that each method gives different results that lead to different interpretations. In the cases of the Pearson correlation (0.024) and the R_{ZJ}-Factor (0.013) methods, the results show that there is a almost a perfect match between the measurement and the simulation, and this does not agree with the expert's opinion.
The main reason for Pearson and R_{ZJ}-Factor results is that the correlation method is unable to evaluate the differences caused by rapid changes in slopes. This method only analyzes the correspondence in amplitude over the signals but no other feature is considered. Because of that reason, those methods are not suitable to analyze signals with abrupt changes or, in particular, with noise.
If the focus is now on the third method (R_{Pendry}-Factor), the results show a poor similarity between the simulation and measurement. Again, this result does not match with the expert's opinion. The main limitation of this method is that it is very sensitive to sudden changes in the signals and the indicators are directly affected.
Finally, we have the methods of Van Hoven factor, FSV and IELF which results are very close to expert's opinion. This result is not surprisingly, since these three methods are particularly robust and have been tested against different types of behaviours. Therefore, as it has been explained in the preceding paragraphs, they are ideal for the numerical simulations output validation process.

4.2 Second case of study

The aim of this second example is to use the validation methods in a more realistic application for the field of numerical electromagnetic simulations. In this case, we compare a measurement (Fig. 4-Blue) with two different simulations (Fig. 4-red & black). Each

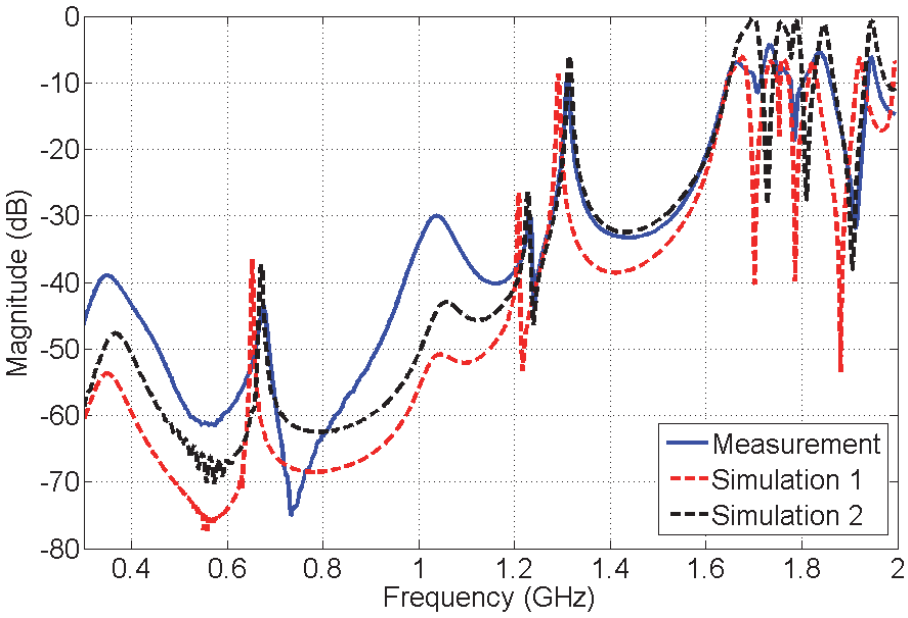


Fig. 4. Comparison between measurement and simulations.

simulation was performed using the same Finite Difference in Time Domain (FDTD) algorithm but with different settings like meshing and time step. The target is to decide, by using IELF and FSV methods, which is the simulation that better fits the measurement from the point of view of a panel of experts in this topic.

At a first sight by doing a quick comparison between the signals, anyone can deduce that both simulations can be improved, but more difficult is to decide which one is better. When we want to study the influence of a particular simulation parameter such as time, mesh, etc. it is very important to identify which simulation has a greater similarity with the measurement of the real setup which it is supposed to be the right one.

Some people can realize that first simulation (Fig. 4-red) has a greater similarity than the second (Fig. 4-red); since it seems to have a close behaviour with the measurement at high-frequencies. However, the simulation number one has important differences at the low-frequency band that should not be forgotten because for another set of people this could be an important feature where to focus the comparison. Furthermore, the second simulation seems to have the opposite behaviour to the first one: a closer similarity at lower frequencies, but a significant amplitude difference at higher frequencies. Therefore, in this case, it is not an easy task to take an overall decision without the help of an experts group or an appropriate validation method.

Fig 5 shows the results when the IELF and FSV are applied to compare the simulations 1 and 2 with the measurement. Observing all the indicators of the used methods, one can see how both methods are quite close to the experts. These results show that the worst simulation is the second one (black) or in other words, the first simulation has more similarity with the measurement considering whole plot.

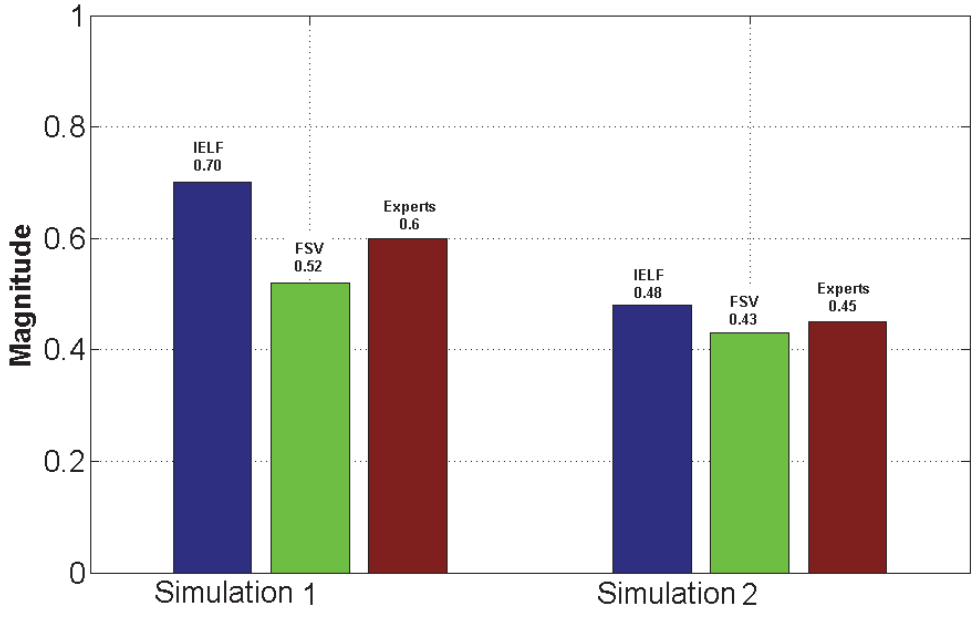


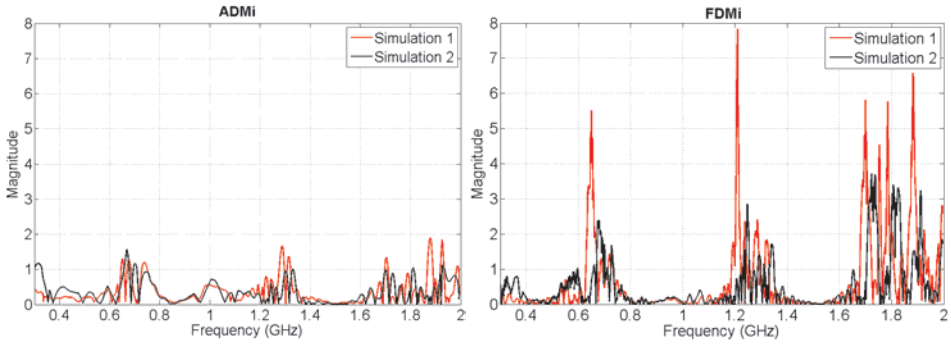
Fig. 5. Results of different validation methods used and the expert panel opinion.

Even though the two validation methods have a unified approach on which simulation is better, each one produces a different value for the final indicator. This is a very common problem when several types of validation methods are compared. It is therefore very important to always use one single method throughout one validation process.
Another point of vital importance, when a validation method is chosen, is to have tools that help us to analyze the different features present in the signals. This is one of the main limitations of the method IELF, because it concentrates all the comparison information in a single number. With this method, it is not possible to obtain more information about the validation process.
The FSV method has, as noted above, several analysis tools that can help to establish a more comprehensive comparison in the validation process. With the mean value of each indicator (Table 2), one can see that the largest difference between the simulations is in the shape indicator (FDM_{TOT}) and not in the amplitude (ADM_{TOT}) as one might think.

FSV indicators			
Simulation 1		*Simulation 2*	
ADM_{TOT}	0.28	ADM_{TOT}	0.35
FDM_{TOT}	0.66	FDM_{TOT}	0.65
GDM_{TOT}	0.80	GDM_{TOT}	0.82

Table 2. FSV mean values indicators.

One way to analyze in greater detail what happens in each comparison is using the point-by-point indicators. It is important to recall that, in this case, the indicators ADMi and FDMi correspond to the point-by-point analysis between each simulation and the measurement. Fig. 6 shows that the indicator most affected in these comparison is the FDMi indicator and of course for the first simulation. Now we can see, very clearly, that the problem is mainly in the shape and not in amplitude.



(a) (b)

Fig. 6. Point-by-point indicators results. (a) ADMi indicator results for simulation 1 & 2 with respect to the measurement. (b) FDMi indicator results for simulation 1 & 2 with respect to the measurement.

Another powerful use of these indicators is to identify where, and the exact value, the major differences over all the data set (frequencies in this case) is produced. For both analysis

made in this example, it is clear that the major difference is found at frequencies of 0.6, 1.2 and 1.8 GHz. In the worst case, we can see (Fig. 6b) that the FDMi indicator reaches values near to 8, showing an important difference between the first simulation and the measurement.

With these two short examples, it became clear that it is very important to choose a proper method of validation that objectively represents the opinion of experts. We have also seen how important it is to have the necessary tools to interpret these results.

It should be noted that all the techniques presented can be used not only to validate the numerical methods and simulation; its use can be extended to other areas that require a quantitative comparison of complex data. The only important thing when a validation method is chosen is that it must provide a similar result to the expert opinion, which implies an objective analysis of the data.

On the other hand, it is needed to take into account that a perfect method to validate any kind of result does not exist. Each one of the methods presents advantages and disadvantages depending on the type of data and the type of analysis desired. The most essential thing at the time to apply the validation is to consider the following items:

a. The implementation of the validation technique should be as simple as possible; this will avoid confusion and data clouding.
b. The validation method should reflect human opinions. Any technique which leads to conflict with the views of the user will fall rapidly into disuse.
c. The validation method should provide the possibility to be applied in different environments and/or applications.
d. The validation method should be commutative. The results of the comparison should always be the same regardless of which is used as a reference or pattern. In other words, the user satisfaction and credibility of the method can be affected if the quality of the technique varies depending on which data is used as a pattern for comparison.
e. The validation method must analyse the difference between the two data sets and always yield the same result, regardless of the user and number of times the comparison is made.

5. Acknowledgment

Part of the work described in this paper and the research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013, under grant agreement no 205294, HIRF SE project and was supported (in part) by the Spanish Ministerio de Ciencia e Innovación under project DPI2010-16093 ("Probes for the Simultaneous Measurement of High-Intensity Electric and Magnetic Fields in the Near Field and Time Domain").

6. References

Archambeault, B., & Connor, S. (2008). Proper model validation is important for all EMI/EMC applications. *Proc. IEEE International Symposium on Electromagnetic Compatibility EMC 2008* (pp. 1-8). doi: 10.1109/ISEMC.2008.4652152.

Archambeault, B., Duffy, A. P., & Orlandi, A. (2009). Using the Feature Selective Validation Technique to Compare Data Sets. *Electrical Engineering*, 248-253.

Archambeault, B., & Yu, Z. (2009). Application of the Feature Selective Validation Method to radio path loss measurements. *2009 IEEE International Symposium on Electromagnetic Compatibility*, 259-263. Ieee. doi: 10.1109/ISEMC.2009.5284666.

Awrejcewicz, J., & Krysko, V. A. (2008). *Chaos in structural mechanics* (1st ed.). Springer. doi: 10.1007/978-3-540-77676-5.

Coleby, D. E., & Duffy, A. P. (2002). Analysis of techniques to compare complex data sets. *The current issue and full text archive of this journal is available*, *21*, 540-553.

Drozd, A. L. (2005). Selected methods for validating computational electromagnetic modeling techniques. *Electromagnetic Compatibility, 2005. EMC 2005. 2005 International Symposium on* (Vol. 1, p. 301--306Vol.1). doi: 10.1109/ISEMC.2005.1513518.

Duffy, A. P., Martin, A. J. M., Orlandi, A., Antonini, G., Benson, T. M., & Woolfson, M. S. (2006). Feature selective validation (FSV) for validation of computational electromagnetics (CEM). part I-the FSV method. *Electromagnetic Compatibility, IEEE Transactions on*, *48*(3), 449-459. doi: 10.1109/TEMC.2006.879358.

Duffy, A., Coleby, D., Martin, A., Woolfson, M., & Benson, T. (2003). Progress in quantifying validation data. *Electromagnetic Compatibility, 2003 IEEE International Symposium on* (Vol. 1, p. 323--328vol.1).

GCEM. (2011). FSV Software. Retrieved from http://www.upc.edu/web/gcem/cat/default.htm.

Godoy, L., & Dardati, P. (2001). Validación de modelos en mecánica computacional. *Mecanica Computacional*, *20*, 663-670.

Jauregui, R. (2009). *Tesis Doctoral Comportamiento Electromagnético en el Dominio del Tiempo de Estructuras Complejas Mediante FDTD*. UPC. Universitat Politècnica de Catalunya.

Jauregui, R, Pous, M., Fernández, M., & Silva, F. (2010). Transient Perturbation Analysis in Digital Radio. *2010 IEEE EMC Symposium, USA, 2010* (pp. 2-7).

Jauregui, R, Riu, P., & Silva, F. (2010). Transient FDTD Simulation Validation. *2010 IEEE EMC Symposium, USA, 2010*.

Jauregui, R, Rojas-Mora, J., & Silva, F. (2011). Study of transient phenomena with Feature Selective Validation Method. *Progress in Electromagnetics Research Symposium (PIERS)* (pp. 2-6).

Jauregui, R, Silva, F., & Riu, P. (2007). FDTD Analysis of the distribution of a transient field inside a car. *EMC Europe 2007 Paris Francia*.

Jauregui, Ri, Silva, F., Orlandi, A., Sasse, H., & Duffy, A. (2010). Factors influencing the successful validation of transient phenomenon modelling. *Asia-Pacific Electromagnetic Compatibility Symposium and Technical Exhibition. Beijing, 2010.* (pp. 2-5).

Knockaert, J., Catrysse, J., & Belmans, R. (2006). Comparison and validation of EMC-measurements by FSV and IELF. *IEEE*.

Martin, A. (1999). *Feature Selective Validation*. De Montfort University.

Orlandi, A., Duffy, A. P., Archambeault, B., Antonini, G., Coleby, D. E., & Connor, S. (2006). Feature selective validation (FSV) for validation of computational electromagnetics (CEM). part II- assessment of FSV performance. *Electromagnetic Compatibility, IEEE Transactions on*, *48*(3), 460-467. doi: 10.1109/TEMC.2006.879360.

Orlandi, A. (2006). stand-alone FSV application. Retrieved from http://ing.univaq.it/uaqemc/public_html/.

Oñate, E. (1998). Limite de los métodos numéricos. *Centro internacional de metodos numericos de ingenieria (CIMNE), N191*.

Pendry, J. B. (1980). Reliability factors for LEED calculations. *Journal Physics C: Solid State Physics*.

Riu, P., Jauregui, R, Silva, F., & Fernandez, M. (2007). Transient Electromagnetic Field Computation in Automotive Environments using FDTD. *Proc. IEEE International Symposium on Electromagnetic Compatibility EMC 2007* (pp. 1-4). doi: 10.1109/ISEMC.2007.8.

Robertson, A. W., Somers, J. S., Ricken, D. E., Bradshaw, A. M., Kilcoyne, A. L. D., & Woodruff, D. P. (1990). "Photoelectron diffraction study of Cu(110)-(2 £ 1)-0. *Surface Science*, *227*, 237-45.

Simpson, R. J., Jones, C. R., MacDiarmid, I., Duffy, A., & Coleby, D. (2005). The integrated error against log frequency (IELF) method for CEM validation. *Electromagnetic Compatibility, 2005. EMC 2005. 2005 International Symposium on* (Vol. 1, p. 296--300Vol.1). doi: 10.1109/ISEMC.2005.1513517.

Standard IEEE. (2008, June). IEEE 1597.1 Standard for Validation of Computational Electromagnetics Computer Modeling and Simulations.

Van Hove, M. A. (1977). Surface structure refinement of 2H-MoS2 via new reliability factors for surface crystallography. *Surface Science*, *64*, 85.

Williams, A. J. M., Woolfson, M. S., Benson, T. M., & Duffy, A. P. (1997). Quantitative validation of numerical technique results against experimental data. *Antennas and Propagation, Tenth International Conference on (Conf. Publ. No. 436)* (Vol. 1, p. 532--535vol.1).

Zanazzi, E., & Jona, F. (1977). Reliability factor for surface structure determination. *Surface Science*, *62*, 61.

9

Edge Enhancement Computed Tomography

Cruz Meneses-Fabian, Gustavo Rodriguez-Zurita, and Areli Montes-Pérez
Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias Físico-Matemáticas, Puebla,
México

1. Introduction

The term tomography comes from the Greek words *tomos*, which means to cut or to divide, and graphos, which is a graphic representation. Thus, tomography is a technique aimed to obtain a cut section (a slice) image of an object. This technique is extensively used in Medicine, Astrophysics, Archaeology, Biology, Geophysics, Oceanography, Material Sciences, and other scientific disciplines. Although cut sections can be obtained using several alternative or possible schemes, many of them are based on a mathematical procedure called tomographic reconstruction [Deans, 1983]. The mathematical formulae that are used to reconstruct 2D slices from several 1D so called projections were developed by the Austrian mathematician Johann Radon [1887-1956] for slices with no particular symmetry. His work were not widely known and it was rediscover in connection with radar and X-ray applications X-ray equipment was not able to make several cross-sections, neither were there the computers with the necessary capacity that is required for the automatic calculations. In order to apply them to Medicine, it was necessary to wait until enough computer power was developed. It was also required the equipment able to make several axial images separated by small distances, to electronically store the results, and finally to analyze them. All this was done separatelly by A. M. Cormack and G. H. Hounsfield in the 70´s [Byer & Garbuny, 1973]. The computerized optical tomography (OT) of X rays, substituting X rays by visible light, was born as an extension of the Computerized Tomography (CT) of X rays. The CT is considered as an non-invasive method, which make possible the cross section of objects or bodies, which have certain internal properties, without having to cut through them o to damage the sample. Basically, it consists of the sequential detection of the changes suffered by some wave when it travels through a slice with different projection angles [Byer & Shepp, 1979] and to obtain the information with adequate techniques of inversion and reconstruction algorithms [Kak & Slaney, 1987].
The radiation source is found in the visible range of the electromagnetic spectrum, and the object that will be analyzed has transparency characteristics [Byer & Shepp, 1979] in which it is possible to observe at least one of the possible effects, such as changes in the polarization, absorption, diffraction [Brown, 1966] [Wolf, 1969], refraction, birefringence, phase, etc. This paper is focused on the analysis of transparent objects that present phase effects only.
In this context, the analysis is begun considering a probe that travels as a plane wave, or in interpretative terms, as a set of parallel rays. Afterwards, the action exerted on the object after crossing is considered indifferent respect to its path, but it is affected in its amplitude and/or phase. This type of interaction is known as refractionless limit (diffractionless, no polarization

changes, no dispersion, etc.), and it is known as parallel ray tomography [Stanley, 1981]. Thus, the Radon transform is simplified and reduced to projected path integrals or to the central theorem of the Fourier transform [Stanley, 1981]. It is noteworthy that this type of approximation can be used in several applications, such as X-ray tomography [Stanley, 1981], where the amplitude of an X-ray is absorbed by the human body without deviating significantly in its path; positron-emission tomography [Kak, 1979], in which a radioactive isotope that is inside the body generates a pair of high energy particles whose paths are straight lines; optical tomography of phase objects [Byer, 1979], where thin objects in phase or absorption are considered [Brown, 1966]; and there are also interferometric techniques of spatial filtering that have been suggested [Rodríguez-Zurita *et al.,* 2005], in which a phase operation is made and a modification over the reconstruction is obtained [Philipp, *et. al*., 1993]. There are also studies of thick phase objects, but it is necessary to make a correction of the projections before applying the reconstruction algorithm [Oppenheim, *et. al*., 1973].

It is important to note that in the discussion mentioned before, it is generally assured that the image reconstruction shows information of the slice of the object under consideration. Nevertheless, in some specific cases, it can result of interest to know only the interface line that divides two media, in which a technique that reconstructs images with enhanced edges is useful. This is particularly true for phase objects, where the detection of phase is an issue to be solved in the first place. This can be done with interferometry, or by optical derivatives, for example. In either case, an inversion procedure is needed in order to quantitatively recover the phase distribution, such as determination of the inverse tangent for the first example, or an integration for the second example. But techniques rendering edge-enhancement of phase can permit qualitative edge visualization of phase slices without such an inversion stage. This would be also true for optical amplitude derivatives, of course, but simpler schemas are possible with similar results. Hilbert transform filtering is one of them.

In this chapter, a mathematical model is described in order to obtain the reconstruction of tomographic images with enhanced edges, and there is an experimental implementation shown which is applied to optical tomography of phase objects. In the first place, there is a proof that the mathematical model is based on the establishment of the relation existent between the Radon transform (RT) and the 2-D directional Hilbert transform (HT). Afterwards, there is a description of the experimental possibility, beginning with the relation existent between the projection and the phase of the optical wave when it transverses a thin phase object, continuing with a description of the optical image-forming system 4*f* in order to obtain the HT of the optical field that is produced after crossing the object [Meneses-Fabian, 2011]. Finally, there is a description of the theoretical relationship between the experimental procedures used to obtain the image reconstruction with their enhanced edges in a directional manner whose mathematical expression accepts the directional HT of the object slice as such, and whose direction can be selected with the position of the filter in the Fourier plane of the 4*f* system respect to the projection angle [Montes-Perez *et. al.,* 2011]. Moreover, the possibility of image reconstruction with isotropic edge-enhancement is shown [Montes-Perez *et. al.,* 2011]

2. Parallel-ray tomography and the Radon transform

In tomography, the probe must be able to cross the sample and carry the information of the structure to the outside. The action exerted upon the object due to the interaction with the medium is known as the projection or profile $\check{f}$ and the predominant physical property in

the sample is denoted by the function *f*. This physical effect can be mathematically related with the Radon transform [Deans, 1983] by

$$\overset{\vee}{f} = \Re\{f\}, \tag{1}$$

where $\Re\{\cdots\}$ denotes the RT operator. Since the profile is the physical quantity that can be measured experimentally, the inverse problem consists of knowing *f* from the projections $\overset{\vee}{f}$, which implies the establishment of an inverse transform relationship.

Specifically, considering only one ray of the test probe that is crossing the sample, the path that it describes in the object can be affected in its direction, and this deviation can be caused by the effects of the refraction, but in any case, the path must follow a geodesic, because the ray will follow the path in the minimum time. Therefore, the Radon transform is described by a path integral,

$$\overset{\vee}{f} = \Re\{f\} = \int_C f dl, \tag{2}$$

where C represents the geodesic path of the ray, and *dl* is an arc differential over *C*. In this study, we will only analyze the particular case in which the geodesic is a straight line and when the lines contained in the probe are parallel, so we be discussing parallel ray tomography, and unless mentioned otherwise in this chapter, we will always refer to this type of tomography. In the experimental context, the parallel-ray tomography can be obtained when there is a medium in which the sample is immersed is very similar respect to the test object. This effect is known as refractionless limits, and there are sever examples of test probes that satisfy this condition: the X-rays that cross the human body [Bates & Peters, 1983], the positrons that emerge from an animal due to a radioactive isotope that has been introduced into it [Kak, 1979], or a light wave crossing a phase object that has been submerged in oil to equalize de refraction index [Goodman, 1985].

This discussion is focused on the study of only one slice of the object. The retrieval of the construction of a volume can be done making cuts from different positions and afterwards interpolating this set of reconstructions [Deans, 1983]. This is the reason why the definition of the Radon transform in a plane is stated in the following form:

Definition: Consider the coordinates (x, y) as points in the plane and an arbitrary function f which is defined in a domain *D* of R^2. If *L* is a straight line in the plane, then the path defined by the projection or the line integral of *f* over all the possible lines *L* is the Radon transform of *f*. In other words,

$$\overset{\vee}{f} = \Re\{f\} = \int_L f(x, y) dl, \tag{3}$$

where *dl* is a differential element of length along *L,* as shown in Fig. 1. The domain *D* can include the entire plane or it can be some region. If the Radon transform exists for every possible line *L* in the domain *D*, *f* must be continuous. Fig. 1 shows a straight line characterized by its distance from the origin and the angle that it forms with the horizontal axis, so the equation of the line *L* is given in terms of p and ϕ by the following expression:

$$p = x\cos\phi + y\sin\phi \tag{4}$$

the points (x,y) that satisfy (4) are all the points that lie along the straight line L. Rotating the reference system by an angle ϕ, where $(p,p_\perp)$ are the rotated axes, then p is given by Eq. (4) and $p_\perp$ is given by

$$p_\perp = -x\sin\phi + y\cos\phi \tag{5}$$

In the rotated system, the result is that the line L is described by $p = K$, where K is constant. Therefore, $dl = dp_\perp$ is always parallel to $p_\perp$, which indicates the propagation direction of the probe. This point of view is useful when it is necessary to numerically implement the Radon transform.

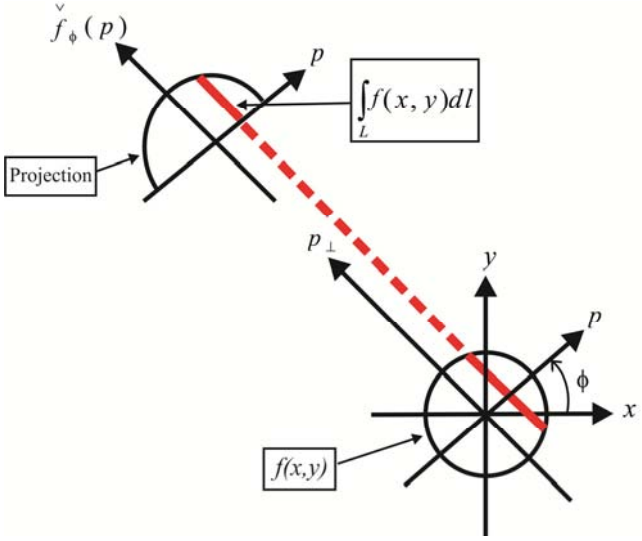


Fig. 1. The set of all the projections for every angle constitute the Radon transform.

The line L expressed in Eq. (4) can be used to rewrite the line integral described in (3) using the following equation,

$$\check{f}(p,\phi) = \Re\{f(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(p - x\cos\phi - y\sin\phi)dxdy\,, \tag{6}$$

where $\check{f}(p,\phi)$ is the RT, the coordinates (p,ϕ) are the points that lie in the Radon space, where each point defines a line in the object space, and this point has a value given by the integral of the function $f(x,y)$ through this line. The argument of the Dirac delta function restricts the integral in the plane to the line L. The Dirac delta function selects the line $p = x\cos\phi + y\sin\phi$ from the rest of the points in the $\Re^2$ space. In this manner, the Radon transform has been defined over all the $\Re^2$ .space.

When ϕ remains constant and p is varied over all its possible values, the integral stated in Eq. (6) becomes a special case

$$\check{f}_\phi(p) = \Re_\phi\{f(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(p - x\cos\phi - y\sin\phi)dxdy\,, \tag{7}$$

which is a sample of the RT, and it is an one-dimensional function of the variable p known as the parallel projection of the function $f(x,y)$ at an angle ϕ, p is the projection variable, and ϕ is the projection angle. Fig. 2 shows three parallel projections $\check{f}_{\phi}(p)$ at different projection angles.

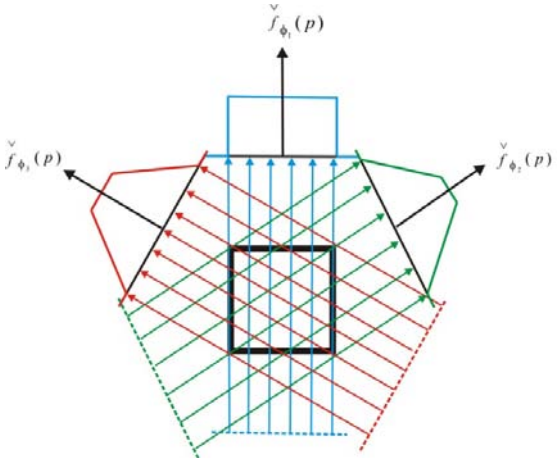


Fig. 2. Parallel projections are taken by measuring a set of parallel rays for three projection angles.

3. Properties of parallel projections

The RT defined in the last section can be extended to R^n. If the reader is interested, he can review the references [Bates & Peters, 1983], and this can also be generalized to any probe the test probe follows through the medium. In this case, the reader can review [Bates & Peters, 1983]. It is interesting to note that the RT and the projection have important properties, which can be stated as follows:

1. They are periodic from 0 to 2π.
2. They are symmetric respect to π.
3. The zero-momentum of the Radon transform is constant for any projection angle.

These properties are necessary to be sure that the Radon inverse and the unicity of the function $f(x,y)$ exist univocally in a semiperiod $\phi \in (0,\pi)$.

3.1 Proof of the properties

First property: Periodicity with a period of 2π, $\check{f}_{\phi+2\pi}(p) = \check{f}_{\phi}(p)$. The proof is trivial.

Second property: Symmetry respect to a half period π, $\check{f}_{\phi+\pi}(p) = \check{f}_{\phi}(-p)$. In this case, in order to prove it, it is only necessary to substitute $\phi = \phi + \pi$ in Eq. (7):

$$\begin{aligned}\check{f}_{\phi+\pi}(p) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdyf(x,y)\delta[p - x\cos(\phi+\pi) - y\sin(\phi+\pi)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdyf(x,y)\delta[p + x\cos\phi - y\sin\phi]\end{aligned}, \tag{8.1}$$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdyf(x,y)\delta[(-p)-x\cos\phi-y\sin\phi] \\ &= \overset{\vee}{f}_{\phi}(-p) \end{aligned} \tag{8.2}$$

which proves the property.

Third property: The property that the area is constant for every projection, which is every general, and is related with the zero-momentum of the RT defined as the integral of a tomographic projection $\int_{-\infty}^{\infty} dp \overset{\vee}{f}_{\phi}(p)$, which represents the probe as it is crossing the sample. Substituting Eq. (7), it is possible to prove that:

$$\int_{-\infty}^{\infty} dp \overset{\vee}{f}_{\phi}(p) = \int_{-\infty}^{\infty} dp \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(p-x\cos\phi-y\sin\phi)dxdy$$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy \int_{-\infty}^{\infty} dp\delta(p-x\cos\phi-y\sin\phi) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy \qquad , \\ &= \Im\{f(x,y)\}_{\mu=\nu=0} = \tilde{f}(0,0) \end{aligned} \tag{9}$$

which is the volume under the function $f(x,y)$, or the value of (0,0) order of its bidimensional Fourier-transform (FT) $\tilde{f}(0,0)$. The volume of $f(x,y)$ does not depend on ϕ. Therefore, the zero-order $\tilde{f}(0,0)$ is the only common point of the FT (one-dimensional, from p to w) of every projection, so it has to have the same value, not depending on ϕ. Interpreting the area of every projection, which must have the same value, notwithstanding the angle of incidence on the object. It is suggested that if this were carried to the expermental area, this property can be useful as a criterion for discarding experimental data if it is contradicted [Ornelas-Rodríguez *et al.*, 1999]. This property is also useful in the design of reconstruction algorithms based on iterative and/or statistical methods [Kak & Slaney, 1987].

3.2 Central slice Fourier theorem

The central slice Fourier theorem, also known as the projection theorem, describes the relation between the RT and the FT of a function. This relationship is the cap stone of parallel projection tomography, because it is the fundamental bridge to find the inverse RT and to theoretically make possible the retrieval of the internal information of the sample. The projection theorem predicts in what measure can a parallel projection to the angle ϕ can be used in the retrieval of the internal information of the object.

We begin by defining the bidimensional Fourier transform of the object function as

$$\tilde{f}(\mu,\nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(\mu x+\nu y)}dxdy \ , \tag{10}$$

and the FT of the parallel projection $\overset{\vee}{f}_{\phi}(p)$ to the angle ϕ can be expressed in the following form:

$$\overset{\sim}{\overset{\vee}{f}}_{\phi}(w) = \int_{-\infty}^{\infty} \overset{\vee}{f}_{\phi}(p) e^{-i\pi wp} dp \,. \tag{11}$$

A simple example of the slice Fourier theorem is given by a projection in $\phi = 0$. Let us consider first that the FT of the object along a line in the frequence domain is given by $\nu = 0$. The FT described in Eq. (10) is

$$\tilde{f}(\mu,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) e^{-i\pi\mu x} dxdy \,, \tag{12}$$

This result is due to the fact that in the phase factor, the dependence on y has been eliminated, so we can divide the integral in two parts; i. e.,

$$\tilde{f}(\mu,0) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(x,y)dy\right] e^{-i\pi\mu x} dx \,. \tag{13}$$

From the definition of the parallel projection, it is possible to recognize that what lies between the square brackets is an equation for the projection angle along a line with x constant. Thus, if we substitute this

$$\tilde{f}(\mu,0) = \int_{-\infty}^{\infty} \overset{\vee}{f}(x,0) e^{-i2\pi\mu x} dx \,. \tag{14}$$

In this way, the right member of the equation represents the one-dimensional FT of the projection when $\phi = 0$, since in this case, $p = x$ and $w = \mu$, Therefore, in this way we have found a relationship between the vertical projection and the 2-D transform of the object function in the following manner

$$\tilde{f}(\mu,0) = \overset{\sim}{\overset{\vee}{f}}(\mu,0) \,. \tag{15}$$

Up to this moment, it has only been proven for a particular and simple example of the Fourier slice. Nevertheless, it can be stated that this result is independent of the orientation that exists between the object and the coordinate system.

In order to deduce the Fourier slice theorem, let us consider the following rectangular coordinate axes system $(p, p_{\perp})$ and $(w, w_{\perp})$ obtained through the rotation of the axes (x, y) and (μ, ν) by an angle ϕ (as shown in Fig. 3).

$$\begin{pmatrix} p \\ p_{\perp} \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}; \quad \begin{pmatrix} w \\ w_{\perp} \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \mu \\ \mu \end{pmatrix}. \tag{16}$$

It is possible to mathematically describe a sampling of the transform along a line that passes through the origin, with a slope given by ϕ:

$$\tilde{f}(\mu,\nu)\delta(-\mu\sin\phi + \nu\cos\phi) = \tilde{f}(\mu,\nu)\delta(w_{\perp}) \,, \tag{17}$$

this equation describes a mapping in the frequence Fourier space, as shown in the right-handed side of the diagram in Fig. 3, where $\delta(w_\perp)$ denotes the sampling of the transform along the line $w_\perp = 0$; i. e., along the axis w. The sampling that is obtained corresponds to the object plane with a convolution between the inverse transforms of each factor. If we now consider the fact that $\Im^{-1}\{\delta(w_\perp)\} = \delta(p)$, then for every value of ϕ, it is possible to obtain the following identities:

$$\Im^{-1}\left\{\tilde{f}(\mu,\nu)\delta(w_\perp)\right\} = \Im^{-1}\left\{\tilde{f}(\mu,\nu)\right\} \otimes\otimes \Im^{-1}\left\{\tilde{f}\,\delta(w_\perp)\right\} = f(x,y)\otimes\otimes\delta(p) = \check{f}_\phi(p), \quad (18)$$

where $\otimes\otimes$ indicates the convolution operation in two dimensions, one with respect to x and the other one with respect to y. Eq. (18) is known as the slice Fourier theorem, or the projection theorem [Haykin, 1985], and it can be stated as establishing the fact that the one-dimensional Fourier transform of the projection to the angle ϕ is identical to the sampling along a line which has a slope with an angle ϕ of the bidimensional Fourier transform of the slice function. In this proof, it has been assumed that $f(x,y)\otimes\otimes\delta(p) = \check{f}_\phi(p)$, and in order to verify this, it is possible to rewrite Eq. (7) changing the integration variables (x,y) by (ξ,η) and substituting $p = x\cos\phi + y\sin\phi$, so

$$\begin{aligned}\check{f}_\phi(p) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi,y)\delta(x\cos\phi + y\sin\phi - \xi\cos\phi - \eta\sin\phi)d\xi d\eta \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi,y)\delta\left[(x-\xi)\cos\phi + (y-\eta)\sin\phi\right]d\xi d\eta\end{aligned}, \quad (19)$$

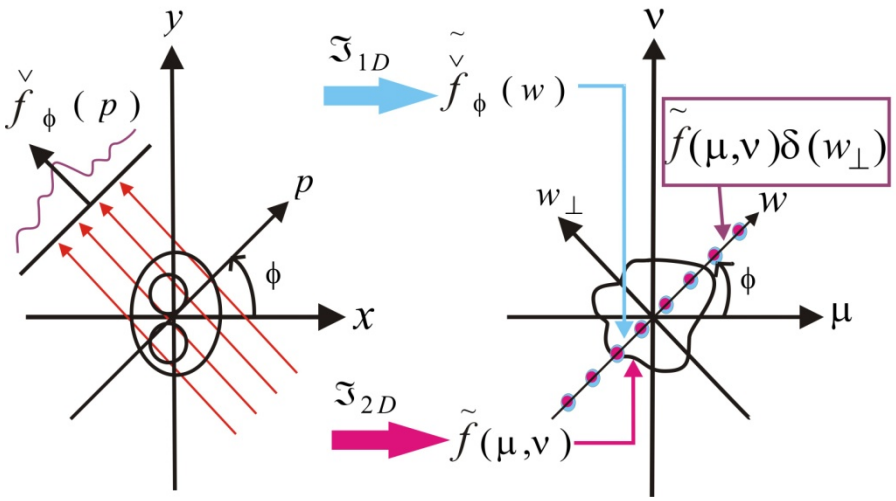


Fig. 3. Coordinate system for parallel projections. From the plane of the slice of object *f*, to the Fourier plane $\tilde{f}$.

this can be seen as a convolution of the form $f(x,y)\otimes\otimes\delta(x\cos\phi+y\sin\phi)$, which proves the last statement, since $\delta(x\cos\phi+y\sin\phi)=\delta(p)$. It can be concluded that the Radon transform can be symbolically written as a convolution of the slice function $f(x,y)$ with a Dirac delta function

$$\overset{\vee}{f}(p,\phi)=\Re\{f(x,y)\}=f(x,y)\otimes\otimes\delta(p), \tag{20}$$

hence, the slice Fourier theorem, expressed in Eq. (18), can finally be written as

$$\tilde{f}(\mu,\nu)\delta(w_\perp)=\Im\left\{\overset{\vee}{f}_\phi(p)\right\}=\overset{\sim}{\overset{\vee}{f}}_\phi(w). \tag{21}$$

It is important to note that the convolution of the two functions in order to find the FT of Eq. (20) is identical to Eq. (21), which is the projection theorem. Therefore, the same properties of symmetry and the zero-moment of the RT must satisfied in the frequency space.

3.3 Inverse Radon transform

The inverse Radon transform can be solved interpreting the parallel projection $\overset{\vee}{f}_\phi(p)$ as an unidimensional inverse transform $\Im_{1D}^{-1}\{...\}$ of a sampling of the bidimensional transform $\tilde{f}=\Im_{2D}\{f(x,y)\}$ made along the line $w_\perp=0$. Thus, if $\overset{\vee}{f}_\phi(p)$ is known, then for every ϕ it is possible to state that

$$f(x,y)=\Im_{2D}^{-1}\left\{\tilde{f}(\mu,\nu)\right\}=\Im_{2D}^{-1}\left\{B\Im_{1D}\left[\overset{\vee}{f}_\phi(p)\right]\right\}, \tag{22}$$

where it is implied that $\tilde{f}(\mu,\nu)$ is constructed with all $\overset{\vee}{f}_\phi(p)$ conventionally accomodated. Here $\Im_{1D}\{...\}$, is from $p\to w$ and the accomodation is denoted by B [Deans, 1983].
If we start from the definition of the inverse Fourier transform of the slice function,

$$f(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d\mu d\nu\tilde{f}(\mu,\nu)e^{i2\pi(\mu x+\nu y)}, \tag{23}$$

changing to polar coordinates: $\mu=w'\cos\phi$ and $\nu=w'\sin\phi$, with $w'^2=\mu^2+\nu^2$, then $d\mu d\nu=w'dw'd\phi$, and by substituting this in the equation of the inverse bidimensional transform, the following expression is obtained:

$$f(x,y)=\int_0^{2\pi}d\phi\int_0^{\infty}w'dw'\tilde{f}(w',\phi)e^{i2\pi w'(x\cos\phi+y\sin\phi)}, \tag{24}$$

where $\tilde{f}(\mu,\nu)=\tilde{f}(w',\phi)$ and this last equation can be rewritten as:

$$f(x,y)=\int_0^{\pi}d\phi\int_0^{\infty}w'dw'\tilde{f}(w',\phi)e^{i2\pi w'(x\cos\phi+y\sin\phi)}+\int_{\pi}^{2\pi}d\phi\int_0^{\infty}w'dw'\tilde{f}(w',\phi)e^{i2\pi w'(x\cos\phi+y\sin\phi)}\,, \quad (25)$$

But if $\varphi=\varphi+\pi$ is substituted in the second term of the right-hand side of the equation,

$$f(x,y)=\int_0^{\pi}d\phi\int_0^{\infty}w'dw'\tilde{f}(w',\phi)e^{i2\pi w'(x\cos\phi+y\sin\phi)}+\int_0^{\pi}d\phi\int_0^{\infty}w'dw'\tilde{f}(w',\phi+\pi)e^{-i2\pi w'(x\cos\phi+y\sin\phi)}\,. \quad (26)$$

In the rotated coordinate system, the function $\tilde{f}(\mu,\nu)$ is described by $\tilde{f}(w,w_{\perp})$. Nevertheless, considering in particular one projection angle ϕ, it is possible to establish the fact that

$$\tilde{f}(w',\phi)=\tilde{f}(w,w_{\perp})\delta(w_{\perp})u(w)\,, \quad (27)$$

where $u(w)$ is the unitary step function or the *Heaviside* function defined under the Direclet criterion [Hwei, 1970], which selects only the positive part of the sampling of $\tilde{f}(w,w_{\perp})$ over the line $w_{\perp}=0$ along w, which is caused by $\delta(w_{\perp})$. Substituting $\tilde{f}(w,w_{\perp})=\tilde{f}(\mu,\nu)$ and applying the Fourier slice theorem,

$$\tilde{f}(w',\phi)=\tilde{\check{f}}_{\phi}(w)u(w)\,, \quad (28)$$

Selecting one half of the projection for a positive w, it is possible to prove in a similar manner that

$$\tilde{f}(w',\phi+\pi)=\tilde{\check{f}}_{\phi}(w)u(-w)\,, \quad (29)$$

which selects the other half of the projection for negative w. Making the change of variable $w'=wu(w)$, then $dw'=dwu(w)$ so substituting this in Eq. (28), in the integral of the first term of Eq. (26), and if another change of variable $w'=-wu(-w)$ is done, then $dw'=-dwu(-w)$, so by substituting Eq. (29) in the second term of Eq. (26), it is possible to arrive to the conclusion that

$$f(x,y)=\int_0^{\pi}d\phi\int_{-\infty}^{\infty}wu(w)dw\tilde{\check{f}}_{\phi}(w)e^{i2\pi wp}-\int_0^{\pi}d\phi\int_{-\infty}^{\infty}wu(-w)dw\tilde{\check{f}}_{\phi}(w)e^{i2\pi wp}\,, \quad (30)$$

where it has been assumed that $\exp\left[i2\pi wu(\pm w)p\right]=u(\mp w)+u(\pm w)\exp(i2\pi wp)$, where $u^n(\pm w)=u(\pm w)$ with n real, and $u(w)u(-w)=0$. By grouping the integrals, we have

$$f(x,y)=\int_0^{\pi}d\phi\int_{-\infty}^{\infty}w\left[u(w)-u(-w)\right]dw\tilde{\check{f}}_{\phi}(w)e^{i2\pi wp}\,, \quad (31)$$

and by rewriting the *signum* function in terms of the *Heaviside* function in the form $\operatorname{sgn}(w) = u(w) - u(-w)$ [Ronald, 2000]; it is possible to conclude that $|w| = w\left[u(w) - u(-w)\right]$, where the bars indicate an absolute value, so we finally arrive to the following expression

$$f(x,y) = \int_0^{\pi} d\phi \int_{-\infty}^{\infty} |w| dw \tilde{\check{f}}_{\phi}(w) e^{i2\pi wp} , \tag{32}$$

If the filtered projection is defined with a spatial filter of the form $\check{g}_{\phi}(p) = \int_{-\infty}^{\infty} dw |w| \tilde{\check{f}}_{\phi}(w) e^{i2\pi wp}$, then

$$\check{g}_{\phi}(p) = \Im^{-1}\left\{ |w| \tilde{\check{f}}_{\phi}(w) \right\} = \int_{-\infty}^{\infty} dw |w| \tilde{\check{f}}_{\phi}(w) e^{i2\pi wp} , \tag{33}$$

so the inverse Radon transform can be written as

$$f(x,y) = \int_0^{\pi} d\varphi \check{g}_{\phi}(p) = \int_0^{\pi} d\phi \check{g}_{\phi}(x\cos\phi + y\sin\phi) . \tag{34}$$

This result establishes the existence of a possible inverse Radon transform operation which is mediated by a spatial filter $|w|$. This is the reason why it is known as a filtered projection, and Eq. (34) is known as the retroprojection of the filtered projections. It is important to note that the resulting filter is reminiscent of the derivative (by a linear factor w) [Deans, 1983].

4. Numerical simulation

A numerical simulation is a mathematical recreation of a natural process. By using numerical simulations, it is possible to study physical processes. Thus, the field of numerical simulations represents a rich field of interdisciplinary research. Some of the scientific problems are studied principally through the use of numerical simulations, such as those scientific fields that are governed by non-linear simultaneous equations, or those that are not easily reproducible in the laboratory. The use of these to study a problem normally requires a careful study of the numerical methods and algorithms that will be used and the fundamental process that will be included. A numerical simulation differs from a mathematical model in the sense that the first one is a representation in every instant of the process that will be simulated, while the second one is a mathematical abstr5action of the fundamental equations necessary to analyze the phenomenon. Normally, the use of a numerical simulation in the analysis of a given problem requires a careful planning of the mathematical model that will be used and the necessary algorithms that will be employed.

The numerical simulation presented here will have the following steps:

1. Definition of the slice function of the object $f(x,y)$,
2. Calculation of the profiles or parallel projections of the object $\check{f}_{\phi}(p)$ for every ϕ in the range $[0,2\pi]$, and

3. Calculation of the reconstruction of the slice of the object through the filtered retroprojection algorithm.

Fig. 4(a) presents a slice of an object in a finite domain in gray levels and in 3D, where in the sake of simplicity, a unitary value has been assigned inside, and it is null outside this domain. The image shown has a resolution of 200x200 pixels and is encoded in 256 gray levels. The slice of the object generally is a cross section at a constant height of the study object. Once the slice function $f(x,y)$ has been defined, it is necessary to calculate the projections through the implementation of the Radon transform. This sample is identified with a sinogram, as can be seen in Fig. 4(b) in 2D and 3D. A sinogram is a data matrix, where the number of rows represents the number of projections taken between 0 and 2π radians; i. e. ϕ, and the number of columns is the number of samples considered in the projection coordinates *p*. Each row is identified with a projection angle and the data in these can be seen as a profile in this angle. The angular step $\Delta\phi = \pi/100$ was used, where 200 was the number of projections calculated, while the number of data for each projection was of 200, so the sinogram has a resolution of 200 x 200 pixels, and it is encoded in 256 levels of grays.

Respect to the reconstruction of the slice of the object, in this chapter, the filtered retroprojection algorithm is used [Deans, 1983], which is a discrete implementation of the inverse Radon transform. The filtered retroprojection algorithm was first proposed by Bracewell and Riddle [Bracewell & Riddle, 1967] and was also stated independently by Ramachandran and Lakshminarayanan [Ramachandran & Lakshminarayanan, 1971]. The superiority of the filtered retroprojection algorithm when compared to the algebraic techniques was first proven by Shepp and Logan [Shepp & Logan, 1974]. This was developed for data with a form of a fan, and it was done by Lakshminarayanan [Lakshmnarayanan, 1975] for the case of equidistant collinear detectors, and was later extended by Herman and Naparstek [Herman & Naparstek, 1977] for the case of equiangular rays. Several authors [Baba & Murata, 1977] [Kenue & Greenleaf, 1979] [Kwoh *et al.*, 1977] [Lewitt, 1979] [Tanaka & Iinuma, 1975] have proposed variations to the filter function of the filtered retroprojection algorithm. The images can be reconstructed from data from beams that have a fan form. The filtered retroprojection algorithm was used for reconstructions that were done from data generated by the use of very narrow angles in a fan form and beams that rotate and cross continuously the surroundings of the object. The projection algorithm is also known for non-uniform data of the sample [Horn, 1978] [Horn, 1979] [Lewitt & Bates, 1978] [Oppenheim, 1975] [Sato *et al.*, 1980] [Tam & Perez-Mendez, 1981] and for reconstructions obtained from incomplete and limited projections.

Complete reconstructions in three dimensions are discussed in [Chiu *et al.*, 1979] [Chiu *et al.*, 1980] [Smith, 1985]. It is also possible to obtain tomographic images with the use of the direct inverse Fourier transform, although it has less precision, instead of using the filtered retroprojection method. This was proven first by Bracewell [Bracewell, 1956] in radio astronomy, and afterwards it was also done independently by DeRosier and Bracewell [DeRosier & Klug, 1968] for electronic microscopy and Rowley [Rowley, 1969] in optic holography. Several authors apply this method to the radiography, such as Tretiak [Tretiak *et al.*, 1969]. In order to use 2-D FFT algorithms for image formation, they focus specially on the direct Fourier approximation. For some methods that are more recent that reduce to the minimum the resultant interpolation error, you can see [Stark *et al.*, 1981]. Recently, Wernecke and D'Addario [Wernecke & D'Addario, 1977] have proposed an approach of

maximum entropy to direct the Fourier inverse. Their procedure is especially useful if for some reason, the projection of the data is insufficient. A more extensive discussion about these algorithms, as well as the study of their characteristic artifacts is beyond the present study.

The implementation of the filtered retroprojection, in the first place, requires the calculation of the filtered projection (Eq. 34), which is implemented in an approximated manner, introducing a limited bandwidth,

$$g_\phi(p) = \int_{-\Gamma}^{\Gamma} dw|w| \tilde{\check{f}}_\phi(w) e^{i2\pi wp} \ . \tag{35}$$

The projection with a bandwidth Γ must satisfy the Nyquits criterion, $\Gamma = 1/2\varepsilon$ where ε is the sampling range of the projection. Under this condition, the original ramp function $|w|$ is multiplied by a window function $rect(w/2\Gamma)$, $H(w) = |w| rect(w/2\Gamma)$. Another type of window, such as the Hanning, hamming, etc, have also been used [Deans, 1983], so under these considerations, it is possible to describe the impulse response that validates a data sample [Deans, 1983]

$$h(n\varepsilon) = \begin{cases} \frac{1}{4\varepsilon^2} & n = 0 \\ 0 & n = par \\ -\frac{1}{n\pi\varepsilon^2} & n = impar \end{cases} \tag{36}$$

If the discrete number of the projection is denoted by $\check{f}_\phi(k\varepsilon)$ $(k = 0,...N-1)$ at the angle ϕ, the filtered projection described by Eq. (34) can be expressed as a convolution in the spatial domain as

$$g_\phi(n\varepsilon) = \varepsilon \sum_{k=0}^{N-1} h([n-k]\varepsilon) \check{f}_\phi(\mathrm{k}\varepsilon), \quad n = 0,1,...,N-1. \tag{37}$$

Continuing with the procedure in order to obtain the object reconstruction, the filtered projection is adequately introduced in the retroprojection sum

$$b(x,y) = \sum_{m=0}^{M-1} g_\phi\left(x\cos(m\Delta\phi) + y\sin(m\Delta\phi)\right), \tag{38}$$

this comes from the suggesting integral of the inverse Radon transform of *f*, Eq. (34), where ϕ increases by steps $\Delta\phi$, and for some point $p = n\varepsilon$ and ϕ=m$\Delta\phi$ the transformation points (x,y) must satisfy $n\varepsilon = x\cos(m\Delta\phi) + y\sin(m\Delta\phi)$. If we choose $\varepsilon = 1$, in the simulation and applying successively Eqs. (36-38), the sinogram obtained in Fig. 4-(b), the reconstruction of the defined object in Eq. (22) is obtained, as shown in Fig. 4(c) [Deans, 1983].

5. Tomography of enhanced edges using the Hilbert transform

As was mentioned before, the internal information of the object under study can be retrieved from the projections measured with adequate inversion techniques with having to

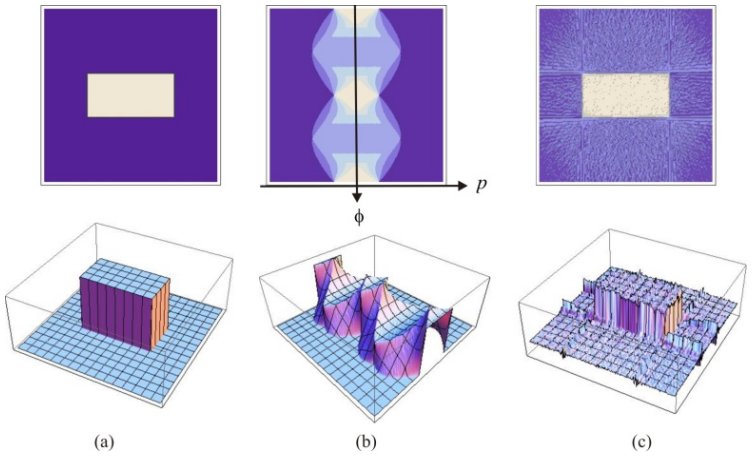


Fig. 4. (a) Slice of the object $f(x,y)$ in 2-D and 3-D; (b) sinogram corresponding to the slice of the object in 2-D and 3-D; (c) reconstruction of the slice of the object in 2-D and 3-D.

cut or damage the sample. In this section, a mathematical model is shown to retrieve the Hilbert transform of the slice function instead of only the slice function. It will be proven that this transform can be obtained in a directional manner in 2-D, and this effect can be accomplished in a spatial filtering process in the frequency domain of the projection. It will be shown that this idea, applied to optical tomography of phase objects can find, for the case of thin phase objects, the directional Hilbert transform of the refraction index distribution, which results in a directional edge enhancement of the reconstructed image, and in the case of thick phase objects, it will be shown that an isotropic edge enhancement of the refraction index can be found. This is based on the relationship that the Hilbert transform has with the Radon transform, as will be stated below.

5.1 The Hilbert transform

It is known that the Hilbert transform is a useful mathematical tool for the description of the complex envelope of the modulated signal for a real carrier. Mathematically, in one dimension, the transform is defined in the following manner [Abhilash, 2006]

$$\overset{\frown}{s}(x) = H\{s(x)\} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(x')}{x-x'}dx' = \frac{1}{\pi x}\otimes s(x)\,, \tag{39}$$

where $s(x)$ is the input function, $\overset{\frown}{s}(x)$ is its corresponding Hilbert transform, and $H\{\cdots\}$ is the respective operator. The integral is evaluated in its principal value. The last term indicates that the Hilbert transform can be written in a symbolic manner as the convolution of the input signal with the function $1/\pi x$.

The typical procedure commonly used for the calculation of the Hilbert transform of a function is done using Fourier analysis [Poularikas, 2000]

$$\overset{\frown}{s}(x) = H\{s(x)\} = \Im^{-1}\left\{-i\,\mathrm{sgn}(\mu)\tilde{s}(\mu)\right\}, \tag{40}$$

where the convolution property of the Fourier transform has been applied, with $s(\mu)=\Im\{s(x)\}$ and $-i\,\mathrm{sgn}(\mu)=\Im\{1/\pi x\}$ [Abhilash, 2006], where $\mathrm{sgn}(\mu)$ is known as the *signum* function, which is a special function defined as

$$-i\,\mathrm{sgn}(\mu)=\begin{cases}-i, & \mu>0\\ 0, & \mu=0\\ -i, & \mu<0\end{cases}, \tag{41}$$

in general, the *signum* function is undefined at $\mu=0$, but here, it has been defined as equal to zero under the Direchlet criterion [Seeley, 1970].

The expression enclosed between brackets in Eq. (40) can be interpreted as a filter operation in a linear system, with $-i\,\mathrm{sgn}(\mu)$ having the role of the filter in the spectral domain of the signal $s(x)$, and its corresponding Fourier inverse $1/\pi x=\Im^{-1}\{-i\,\mathrm{sgn}(\mu)\}$ as a consequence of the input. The operation can be seen as a phase change of $+\pi/2$ radians for each one of the positive frequency components of the signal and a phase change of $-\pi/2$ radians for each one of the negative frequency components of the signal, without changing its amplitude. For this reason, an experimental implementation is viable in areas in which the signal treatment is possible; for example, in communication systems with electric signals [Almeida, 1994] or in optical data processing in Fourier optics [Philipp *et al.*, 1992].

In two dimensions, the HT has been defined in several ways, being one of the most used [Abhilash, 2006]

$$\overset{\frown}{s}(x,y)=H\{s(x,y)\}=\frac{1}{\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{s(x',y')}{(x-x')(y-y')}dx'dy'=\frac{1}{\pi^2 xy}\otimes\otimes s(x,y), \tag{42}$$

Again, the integrals are evaluated in their respective principal Cauchy value. One definition in n dimensions can be seen in [Poularikas, 2000]. On the other hand, the partial Hilbert transforms have also been defined [Poularikas, 2000]. For example, the partial transform with respect to x has the form

$$\overset{\frown}{s}^{0}(x,y)=H^{0}\{s(x,y)\}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(x',y)}{x-x'}dx', \tag{43}$$

and with respect to y, it has the form

$$\overset{\frown}{s}^{\frac{\pi}{2}}(x,y)=H^{\frac{\pi}{2}}\{s(x,y)\}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(x,y')}{y-y'}dy', \tag{44}$$

where the superindices are angles in radians, that indicate the direction in which the transform operation is done, although for the x direction, the notation $\overset{\frown}{s}_x(x,y)$ is more commonly used, and for the y direction, $\overset{\frown}{s}_y(x,y)$ is more extensively employed. In these expressions, the angles 0 and $\pi/2$ have been used as superscripts. This notation has been selected for simplicity, and in order to avoid a possible ambiguity with the notation used in tomography. In a symbolical form, for x this can be written as

$$\overset{\cap}{s}^{0}(x,y)=\frac{\delta(y)}{\pi x}\otimes\otimes s(x,y)\,, \tag{45}$$

and for the variable y

$$\overset{\cap}{s}^{\frac{\pi}{2}}(x,y)=\frac{\delta(x)}{\pi y}\otimes\otimes s(x,y)\,. \tag{46}$$

In a similar manner, we define the directional Hilbert transform in the form

$$\overset{\cap}{s}^{\alpha}(x,y)=H^{\alpha}\{s(x,y)\}=\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta\left[-(x-x')\sin\alpha+(y-y')\cos\alpha\right]}{(x-x')\cos\alpha+(y-y')\sin\alpha}s(x',y')dx'dy'\,, \tag{47}$$

and in a symbolic form,

$$\overset{\cap}{s}^{\alpha}(x,y)=\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\otimes\otimes s(x,y)\,, \tag{48}$$

where the angle α is expressed in radians, and it denotes the direction at which the Hilbert transform is done. It is important to note that the partial Hilbert transform with respect to x and y as defined in Eqs. (45) and (46) can be derived from the definition of the directional Hilbert transform when $\alpha=0$ and when $\alpha=\pi/2$, respectively.
Using Fourier analysis, the directional transform can be defined using the convolution property in two dimensions through

$$\overset{\cap}{s}^{\alpha}(x,y)=\Im^{-1}\left\{-i\,\mathrm{sgn}(\mu\cos\alpha+\nu\sin\alpha)\tilde{s}(\mu,\nu)\right\}, \tag{49}$$

where $\tilde{s}(\mu,\nu)=\Im\{s(x,y)\}$ and
$-i\,\mathrm{sgn}(\mu\cos\alpha+\nu\sin\alpha)=\Im\{\delta(-x\sin\alpha+y\cos\alpha)/\pi(x\cos\alpha+y\sin\alpha)\}$, which can be proved in the following manner. Using the integral definition of the Fourier transform,

$$\Im\left\{\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\right\}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\exp\left[-i2\pi(\mu x+\nu y)\right]dxdy\,, \tag{50}$$

rotating by an angle α both coordinate systems (x,y) and (μ,ν), it is possible to state that

$$\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}r\cos\alpha-r_{\perp}\sin\alpha\\r\sin\alpha+r_{\perp}\cos\alpha\end{pmatrix}\text{ and }\begin{pmatrix}\mu\\\nu\end{pmatrix}=\begin{pmatrix}\sigma\cos\alpha-\sigma_{\perp}\sin\alpha\\\sigma\sin\alpha+\sigma_{\perp}\cos\alpha\end{pmatrix}, \tag{51}$$

where the rotated coordinates have been denoted by $(r,r_{\perp})$ and $(\sigma,\sigma_{\perp})$, respectively and thus, $dxdy=drdr_{\perp}$, $\mu x=\sigma r$, and $\nu y=\sigma_{\perp}r_{\perp}$. Substituting Eq. (51) in Eq. (50),

$$\Im\left\{\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\right\}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\delta(r_{\perp})}{\pi r}\exp\left[-i2\pi(\sigma r+\sigma_{\perp}r_{\perp})\right]drdr_{\perp}\,. \tag{52}$$

This equation can be written as a separable integral

$$\Im\left\{\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\right\}=\int_{-\infty}^{\infty}\frac{1}{\pi r}\exp(-i2\pi\sigma r)dr\int_{-\infty}^{\infty}\delta(r_{\perp})\exp(-i2\pi\sigma_{\perp}r_{\perp})dr_{\perp}, \tag{53}$$

where each factor is a one-dimensional Fourier transform that is known, the first being $\Im\{1/\pi r\}=-i\,\mathrm{sgn}(\sigma)$ and the second one is $\Im\{\delta(r_{\perp})\}=1$, so the solution is given by

$$\Im\left\{\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\right\}=-i\,\mathrm{sgn}(\sigma), \tag{54}$$

Finally, by direct substitution of $\sigma=\mu\cos\alpha+\nu\sin\alpha$ the following expression is obtained:

$$\Im\left\{\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\right\}=-i\,\mathrm{sgn}(\mu\cos\alpha+\nu\sin\alpha), \tag{55}$$

which proves the point. It is thus important to note that the corresponding expressions for the definitions of the partial Hilbert transform can be directly obtained from Eq. (49) using Fourier analysis by substituting $\alpha=0$ for the partial derivative with respect to x:

$$\overset{\cap}{s}^{0}(x,y)=\Im^{-1}\left\{-i\,\mathrm{sgn}(\mu)\tilde{s}(\mu,\nu)\right\}, \tag{56}$$

and $\alpha=\pi/2$ for the partial with respect to y:

$$\overset{\cap}{s}^{\frac{\pi}{2}}(x,y)=\Im^{-1}\left\{-i\,\mathrm{sgn}(\nu)\tilde{s}(\mu,\nu)\right\}, \tag{57}$$

The definition of the directional Hilbert transform and the slice theorem will serve as a basis to establish the relationship that the Hilbert transform has with the Radon transform, and afterwards to establish how the directional Hilbert transform can be obtained of the slice function of the object study [Meneses-Fabian *et. al.*, 2011].

5.2 Tomography and the Hilbert transform

Starting from the projection data $\overset{\vee}{f}_{\phi}(p)$, in the first place it is necessary to calculate the Hilbert transform for every projection angle, to prove that this new projection data is still valid from the point of view of tomography. Then the proofs of symmetry and zero-momentum of the Radon transform, and it is proven that the partial Hilbert transform of the slice of the object with respect to y is reconstructed. In the second place, after finding the directional Hilbert transform of the slice function $f(x,y)$, we proceed to establish the filter function in the frequency space of the projection in order to reconstruct the directional Hilbert transform of $f(x,y)$ as if we had begun from the projection data $\overset{\vee}{f}_{\phi}(p)$.

The Hilbert transform of the projection $\overset{\vee}{f}_{\phi}(p)$ using the definition given in Eq. (49) is

$$\overset{\frown}{\underset{\vee}{f}}_{\phi}(p) = H\{\overset{\vee}{f}_{\phi}(p)\} = \overset{\vee}{f}_{\phi}(p) \otimes \frac{1}{\pi p}. \tag{58}$$

In order to prove the symmetry property, it is necessary to substitute $\phi = \phi + \pi$ and we proceed in an analogous manner to the one followed in section (3.1),

$$\begin{aligned} \overset{\frown}{\underset{\vee}{f}}_{\phi+\pi}(p) &= \overset{\vee}{f}_{\phi+\pi}(p) \otimes \frac{1}{\pi p} \\ &= -\overset{\vee}{f}_{\phi}(-p) \otimes \frac{1}{\pi(-p)}, \\ &= -\overset{\frown}{\underset{\vee}{f}}_{\phi}(-p) \end{aligned} \tag{59}$$

the negative sign that is outside the function reveals that the property is not satisfied. One half of the projections is multiplied by +1 and the other one by -1. In order to compensate this, it has been observed that the function $\operatorname{sgn}[\sin(\phi)]$ can successfully compensate the negative sign, because $\sin(\phi)$ is positive when $\phi \in (0, \pi)$, and it is negative when $\phi \in (\pi, 2\pi)$. In this manner, the new projection is given by

$$\operatorname{sgn}[\sin\phi]\overset{\frown}{\underset{\vee}{f}}_{\phi}(p) = \operatorname{sgn}[\sin\phi]\overset{\vee}{f}_{\phi}(p) \otimes \frac{1}{\pi p}. \tag{60}$$

This projection satisfies the symmetry property. In order to prove the zero momentum of the RT,

$$\begin{aligned} \int_{-\infty}^{\infty} dp \operatorname{sgn}[\sin\phi]\overset{\frown}{\underset{\vee}{f}}_{\phi}(p) &= \int_{-\infty}^{\infty} dp \operatorname{sgn}[\sin\phi]\overset{\vee}{f}_{\phi}(p) \otimes \frac{1}{\pi p} \\ &= \operatorname{sgn}[\sin\phi] \int_{-\infty}^{\infty} \mathrm{dp} \int_{-\infty}^{\infty} \mathrm{d}q \, \overset{\vee}{f}_{\phi}(q) \frac{1}{\pi(p-q)} \quad , \\ &= \operatorname{sgn}[\sin\phi] \int_{-\infty}^{\infty} \mathrm{d}q \, \overset{\vee}{f}_{\phi}(q) \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dp}{(q-p)} = 0 \end{aligned} \tag{61}$$

where it has been assumed that $\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dp}{(q-p)}$ is the Hilbert transform of the unity, which is equal to zero, $H\{1\} = 0$. In this manner, it is proven that it is a constant equal to zero and independent respect to the projection angle, so the zero-momentum property of the RT is also satisfied. In this moment, it is possible to state that a set of modified projection data, as those indicated in Eq. (60) is considered enough to reconstruct the slice function. Up to this moment, it is not possible to analytically describe this function. In order to deduce it, the modified projection indicated in Eq. (60) can be expressed using Fourier analysis:

$$\operatorname{sgn}(\sin\phi)\overset{\frown}{\underset{\vee}{f}}_{\phi}(p) = \operatorname{sgn}(\sin\phi)\Im^{-1}\left\{-i\operatorname{sgn}(w)\tilde{\overset{\vee}{f}}_{\phi}(w)\right\}, \tag{62}$$

Applying the slice Fourier theorem and rearranging the terms,

$$\operatorname{sgn}(\sin\phi)\overset{\cap}{\underset{\vee}{f}}_{\phi}(p)=\Im^{-1}\left\{-i\operatorname{sgn}(w\sin\phi)\tilde{f}(\mu,\nu)\delta(w_{\perp})\right\}, \tag{63}$$

Using Eq. (16), and taking into account the filter property of the delta function, it is possible to find that $\operatorname{sgn}(w\sin\phi)=\operatorname{sgn}(v)$, and obtaining the inverse Fourier assuming that the convolution property for the triple product is given by the convolution of each one of the individual inverses, it can be proven that

$$\operatorname{sgn}(\sin\phi)\overset{\cap}{\underset{\vee}{f}}_{\phi}(p)=\left[\frac{\delta(x)}{\pi y}\otimes\otimes f(x,y)\right]\otimes\otimes\delta(p)=\overset{\pi}{\overset{\cap}{f}}{}^{2}(x,y)\otimes\otimes\delta(p)=\overset{\pi}{\underset{\cap}{\underset{\vee}{f}}}{}^{2}_{\phi}(p), \tag{64}$$

where it has been assumed that what is enclosed between brackets is the definition of the partial Hilbert transform respect to y, and in this manner, the following relationship can be stated,

$$\operatorname{sgn}(\sin\phi)H\left\{\Re_{\phi}\left\{f(x,y)\right\}\right\}=\Re_{\phi}\left\{H^{\frac{\pi}{2}}\left\{f(x,y)\right\}\right\}, \tag{65}$$

which means that the Hilbert transform of a projection at an angle ϕ mediated by a sign factor from the sine of the projection angle equal to the projection of the partial Hilbert transform with respect to y of the slice function. This result predicts that the partial Hilbert transform respect to y will be reconstructed when each parallel projection will be calculated. Their Hilbert transform will have as a result an edge enhancement in the vertical direction.

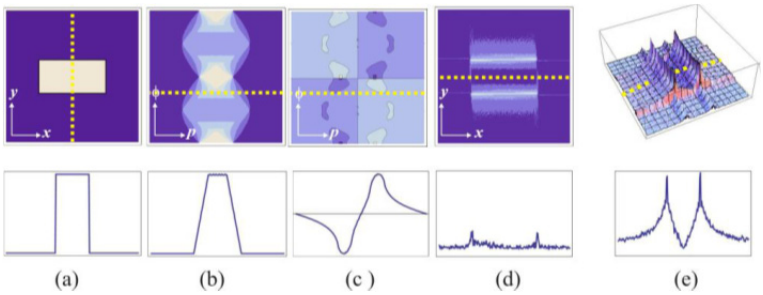


Fig. 5. Numerical simulation. Edge-enhancement along the vertical direction of a unitary rectangle used as the slice function of a thin phase object: (a) slice, (b) sinogram, (c) Hilbert-sinogram, (d-e) vertical edge enhancement reconstruction. The second row shows a line or column data corresponding to each image at first row as is indicated with the yellow dotted line.

Fig. 5 shows a numerical simulation of the theory mentioned above. Column (a) shows the slice of the object $f(x,y)$, while the lower image shows the plot of the profile of the function along the dotted vertical line that crosses the rectangle in the center. Column (b) shows the

sinogram of $f(x,y)$. The image of the upper part of column (c) shows the convolution for each one of the projections in the sinogram that are consequence of the input of the system $1/\pi p$ with the corresponding sign compensation, as indicated in Eq. (60). Observe that the signs of the projections change around $\phi = \pi$, as a result of the symmetry property of the Radon transform. We have denominated this modified sinogram as a Hilbert-sinogram. The lower image shows a sketch of the Hilbert transform of the projection along the same angle used for the last profile. Moreover, with the use of a filtered retroprojection algorithm, the tomographic reconstruction is obtained. The images of the column (d-e) present the reconstruction of the slice function in 2-D and 3D, respectively, where the edge enhancement in the vertical direction is shown, as was expected. The images in the lower part show the profile of the dotted line in the reconstructed image. The line in (d) selects a row of the reconstructed image, as shown in the lower image, and the line in (e) selects one profile of the image, presenting an edge enhancement on the vertical direction.
In this point, and with the purpose of generalizing this discussion, the directional Hilbert transform is calculated using the definition (48) of the slice function.

$$\overset{\frown}{f}^{\alpha}(x,y)=\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\otimes\otimes f(x,y)\,. \tag{66}$$

The projection at an angle ϕ is symbolically given by

$$\overset{\vee}{\overset{\frown}{f}}{}^{\alpha}_{\phi}(p)=\frac{\delta(-x\sin\alpha+y\cos\alpha)}{\pi(x\cos\alpha+y\sin\alpha)}\otimes\otimes f(x,y)\otimes\otimes\delta(p)\,. \tag{67}$$

Applying the convolution property for the triple product of functions in the Fourier space, Eq. (67) will be given by

$$\overset{\sim}{\overset{\vee}{\overset{\frown}{f}}}{}^{\alpha}_{\phi}(w)=-i\,\mathrm{sgn}(\mu\cos\alpha+\nu\sin\alpha)\,\tilde{f}(\mu,\nu)\delta(w_{\perp})\,. \tag{68}$$

Rotating the coordinates using Eq. (16), it is possible to demonstrate the equation $\mu\cos\alpha+\nu\sin\alpha=w\cos(\phi-\alpha)-w_{\perp}\sin(\phi-\alpha)$, and by taking this into account, the last expression can be simplified to the following expression:

$$\overset{\sim}{\overset{\vee}{\overset{\frown}{f}}}{}^{\alpha}_{\phi}(w)=-i\,\mathrm{sgn}[w\cos(\phi-\alpha)]\overset{\sim}{\overset{\vee}{f}}(w)\,, \tag{69}$$

where the filter property of the Dirac delta function and the slice Fourier theorem have been applied. Calculating the inverse Fourier transform for both sides of the equation,

$$\overset{\vee}{\overset{\frown}{f}}{}^{\alpha}_{\phi}(p)=\mathrm{sgn}\left[\cos(\phi-\alpha)\right]\frac{1}{\pi p}\otimes\overset{\vee}{f}_{\phi}(p)=\mathrm{sgn}\left[\cos(\phi-\alpha)\right]\overset{\frown}{\overset{\vee}{f}}_{\phi}(p)\,, \tag{70}$$

where it has been assumed that the relationship $\operatorname{sgn}\left[w\cos\left(\phi-\alpha\right)\right]=\operatorname{sgn}\left(w\right)\operatorname{sgn}\left[\cos\left(\phi-\alpha\right)\right]$ is satisfied. Therefore, the projection of the Hilbert transform of the slice function $f\left(x,y\right)$ along the direction indicated by α is identical to a Hilbert transform of the projection at an angle ϕ mediated by a sign factor by the cosine of the projection angle and the director angle. It is important to note that when $\alpha=\pi/2$, Eq. (70) is reduced to the particular case stated in Eq. (65). Expressing the last expression in terms of operators,

$$\Re_{\phi}\left\{H^{\alpha}\left\{f\left(x,y\right)\right\}\right\}=\operatorname{sgn}\left[\cos\left(\phi-\alpha\right)\right]H\left\{\Re_{\phi}\left\{f\left(x,y\right)\right\}\right\}, \tag{71}$$

i. e., the Radon transform of the directional Hilbert transform of the slice function is identical to the Hilbert transform of the projection mediated by the *signum* function of the cosine of the projection angle and the angle that is indicated by the operation.

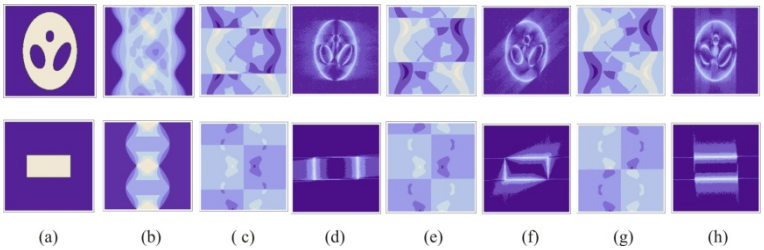


Fig. 6. Reconstruction of the directional Hilbert transform of the slice function. Two examples of slices are presented: (a) phatom of Shepp-Logan, (b) sinograms, (c, e, g) Hilbert-sinograms of irradiance with $\alpha=0,\pi/4,\pi/2$, respectively, (d, f, g); direction reconstruction of the Hilbert-sinograms of (c, e, g).

Fig. 6 presents a numerical simulation of directional edge-enhancement tomography for two slice functions and three particular directions. In column (a) slice functions, Shepp-Logan (upper image) and a rectangle (lower image) are shown. It is important to mention that the first row shows everything related with the Shepp-Logan, while the second row presents everything that is related to the rectangle. The column (b) shows the sinograms for each one of the slices indicated in (a); in the columns (d), (f), and (h), there are reconstructions of the Hilbert-sinograms shown in (c), (e), and (f). In this simulation, the reconstruction of the directional Hilbert transform of the slice function is obtained, just in the manner stated above.

5.3 Optical tomography of edge-enhancement

When the diffraction and refraction effects are significant, it becomes impossible to predict the path of a ray. If the algebraic techniques are applied under these conditions, it is very common to obtain insignificant results. If the refraction and diffraction effects are small (in inhomogeneous media, they are less than 2 for the 3% of the average value and the correlation width for these inhomogeneities is much greater that the wavelength), in some cases it is possible to combine algebraic techniques with digital ray tracing [Andersen & Kak, 1982], who designed iterative procedures that first construct the corresponding transmittance and receive the positions using this distribution, and finally use these rays to construct a more exact series of algebraic equations. Experimentally, this iterative procedure

has been verified for low refraction objects. In particular, when a light beam crosses an object and only suffers phase changes, it is considered as a phase object. In the study of these kinds of objects, in terms of optical tomography, a minimum refraction is assumed (refractional limits). In this approximation, the light rays travel in straight lines and cross the object without changing their path, so it is possible to not to take into account the changes of diffraction, dispersion, polarization, refraction, etc.
In the last section, the theory required for reconstructing the directional Hilbert transform of the slice function was stated, instead of only the slice function, which was proved with a numerical simulation and taking tomographic images with edge enhancement. In this section, we will present how this idea can be implemented in an experiment applied to optics. In particular, thin phase objects will be presented, where well described analytic relations are obtained, and an approximated description for the case of non-thin phase objects will be stated, in which there is an isotropic enhancement instead of the directional enhancement. This experimental implementation is based on obtaining the Hilbert transform in the optical field, which is done through the implementation of an optical correlator or a 4*f* image forming system of a double Fourier transform.

5.3.1 Hilbert transform using a 4*f* system

It is well known that the implementation of the Hilbert transform using an optical system can be efficiently implemented employing a double Fourier transform 4*f* image system, also known as an optical correlator. Its functioning is based on the optical Fourier theory. There are several proposals to optically implement the two-dimensional Hilbert transform. Some have proposed to obtain the transform under a rectangular system [Davis *et. al.,* 2000] with a sign filter and others say that it is possible to find the radial Hilbert transform with a spiral step filter [Davis *et. al.,* 2000]. There are also others who suggest the implementation of a fractional Hilbert transform in rectangular coordinates with a variable phase step filter [Davis *et. al.,* 2002]. It has already been proven that the Hilbert transform of the optical field shows an edge enhancement in terms of the image, and it is noteworthy to state that this fact has been used in optical image processing. In the implementation discussed here, a phase step filter of π radians will be used to obtain the partial Hilbert transform respect to the projection coordinate, and to obtain the one-dimensional Hilbert transform of the field that goes out of the object at a height z. On the other hand, it is important to mention that the implementation of the 4*f* system to obtain the partial Hilbert transform of the field will constitute the bases to construct an optical tomographer used to get tomographic images with experimental edge-enhancement.

5.3.1.1 Statement of the problem

In order to begin the discussion, let us assume that there is a homogeneous plane wave without slope and with amplitude A, linearly polarized, which propagates in the positive direction of $p_{\perp}$, whose spatial part is described by $t_A(p,z)=A$. If this wave crosses the phase object described in a right-handed coordinate system, whose refraction index is a point of the sample denoted by $f(x,y,z)$, then the wave that exits the object is given by

$$A_{\phi}(p,z)=A\exp\left[i\frac{2\pi}{\lambda}\overset{\vee}{f}_{\phi}(p,z)\right], \tag{72}$$

where λ is the wavelength of the light used and $\overset{\vee}{f}_{\phi}(p,z)$ is the accumulated optical path of the light ray that exits the object at a height *z*, which in general can be described by [Deans, 1983]

$$\overset{\vee}{f}_{\phi}(p,z)=\int_{C} f(x,y,z)dl\,. \tag{73}$$

Under the minimum refraction approximation, the path *C* described by a light ray would be very similar to a straight line, as can be seen in Fig. 71, and it is noteworthy to state that this situation can be obtained if the object is submerged in oil [Goodman, 1985]. In this case, the optical path can be interpreted as the parallel projection at an angle ϕ as described in Eq. (73):

$$\overset{\vee}{f}_{\phi}(p,z)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)\delta(p-x\cos\phi-y\sin\phi)dxdy\,, \tag{74}$$

where *z* is constant and indicates the height at which the cut is done, or the plane that defines the slice of the object that will be reconstructed.
In order to obtain the projection from the optical field (Eq. 72), it is necessary to apply an indirect method because an optical detector such as a CCD camera would observe something proportional to the square module of the amplitude of the field, being the phase undetectable, so the projection would also be undetectable. There are several proposals to obtain a projection of the optic filed, such as the interferometric techniques, which basically consist in making an interference of the field that exits the object with a reference wave, from which an interference pattern is obtained. The phase information is obtained from this pattern with some phase extraction technique, such as the phase shifting interferometry introduced by Brunning [Bruning *et. al.,* 1974] or by the Fourier transform suggested by Takeda [Takeda *et al.*, 1982], among others [Meneses-Fabian *et. al.,* 2011]. This type of technique is known as interferometric tomography, and has been applied in the study of temperature gradients in flames [Braslavsky *et al.*, 1998], in radiators and electronic chips [Wu & He, 1999] or in the measurement of concentration profiles in layers formed in the boundaries of cathodes in an electrolytic cell of $ZnCl_2$ [Kujawinska & Kniazewski, 2006], and in the study of phase objects, such as glasses, oils, acetates [Philipp *et al.*, 1992]. Nevertheless, image processing was always necessary to increase the processing quantity and the retrieval time of the image. Moreover, with these methods, it is possible to only retrieve the information related with the slice function of the object, but not of any operation over it. An alternative to retrieve some operation of the slice function has been proposed by finding the angular derivative, being made using an ESPI system which is sensible out of the plane to obtain the difference between two adjacent projections [Meneses-Fabian *et. al.,* 2003]. There are also other proposals, such as the use of a 4*f* system to obtain the directional derivative of the slice function, using a square root filter to obtain the semi derivative of the field, and by using it, to obtain something proportional to the derivative of the parallel projection [Rodríguez-Zurita *et al.*, 1997]. In this article, the proposal consists in using the 4*f* system to obtain the Hilbert transform of the field to detect information of the phase and the parallel projection.

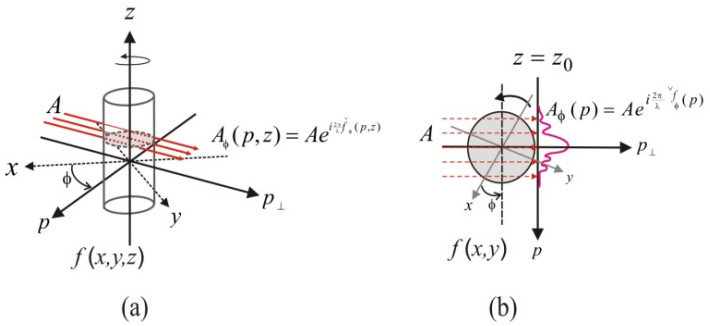


Fig. 7. Phase objects crossed by a plane wave, with a rotated reference system: (a) 3D object rotated around the z axis to generate the projection angle and (b) the slice of the object with z constant.

5.3.1.2 The Hilbert transform from the optical point of view

In this section, a tomographic system is proposed to detect the projection data from the irradiance, where the detection time for each projection data is reduced in time and cost. Moreover, not only are the projection data detected, but it is also possible to do an edge enhancement on the projection data emerging from the objects under study, and to retrieve the slice function of the object enhancing its characteristics, which cannot be detected with the unaided eye.

If the field that exits the object is considered as a function of the input transmittance of the image forming optical system, as schematically presented in Fig. 8, the field at the exit of the system is altered. The optical correlator consists of three planes and two lenses with the same focal distance *f*. As can be seen in Fig. 8, the distance between one lens and the adjacent plane is the focal distance of the lens *f*. The input plane is known as the object plane. The first lens finds the optical Fourier transform of the input field and this distribution is obtained in the Fourier plane. The second lens finds the optical Fourier transform in the Fourier plane, and distributes it in the output plane, known as the image plane. Then, at the exit of the system, we have the transform of the Fourier transform of the input transmittance function, which is the same function as the input, but inverted. Since it has also been convened that it is necessary to invert the axes in the image plane, the second lens makes the inverse Fourier transform. If a pupil function is placed on the Fourier plane, also known as frequency space filter, the exit function is the convolution of the input function with the result of the impulse of the filter [Abhilash, 2006]. In this case, the filtered used is a phase step of π radians, which is made depositing over one half of a slide a thin metallic layer calculating its thickness to provoke an expected delay for the wavelength used. The filter can be modeled by the *signum* function, and without loss of generality, the filter function can be written as

$$\tilde{h}_{\phi}(w,\zeta) = -i\,\mathrm{sgn}(w), \tag{75}$$

where ζ is the spatial frequency variable of z. In this manner, the optical field at the exit of the system can be mathematically described as

$$\overset{\frown}{A}_{\phi}(p,z)=\frac{\delta(z)}{\pi p}\otimes\otimes A_{\phi}(p,z), \tag{76}$$

which is the partial Hilbert transform respect to p of the input transmittance function, where it has been assumed that the relation $\delta(z)/\pi p=\Im^{-1}\{-i\,\mathrm{sgn}(w)\}$ is the response of the system. In what respects to the detection of the field, it is only possible to observe its irradiance, because an optical detector such as a CCD camera is used,

$$I_{\phi}(p,z)=\left|\frac{\delta(z)}{\pi p}\otimes\otimes A_{\phi}(p,z)\right|^{2}=\frac{\delta(z)}{\pi p}\otimes\otimes A_{\phi}(p,z)\left[\frac{\delta(z)}{\pi p}\otimes\otimes A_{\phi}^{*}(p,z)\right], \tag{77}$$

the symbol "*" indicates the complex conjugate. In the following sections, it will be shown how this intensity can be used to obtain the Hilbert transform of the projection data under the approximation of thin phase objects, and in the case of thick phase objects, it will be shown that this irradiance is taken directly as projection data, to reconstruct tomographic images with directional and isotropic edge-enhancement, respectively.

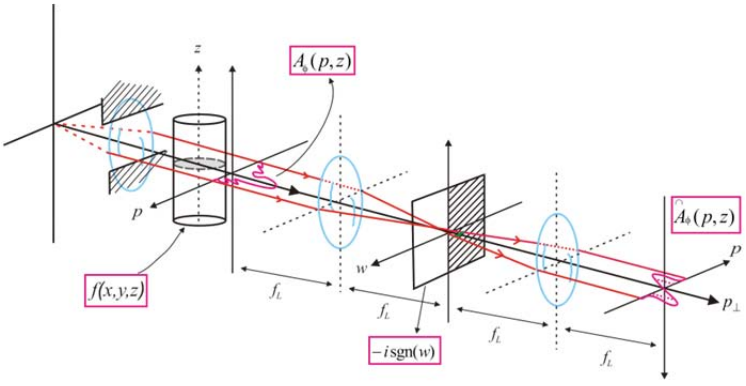


Fig. 8. Optical 4*f* image forming system

5.3.2 Edge-enhancement in optical tomography

5.3.2.1 Thin phase objects

In the thin phase object approximation, the exponential in Eq. (72) expresses the optical field that exits the object, and it is approximately equal to the initial terms of the Taylor series [Zernike, 1955], which is given by

$$A_{\phi}(p,z)=1+i\frac{2\pi}{\lambda}\overset{\vee}{f}_{\phi}(p,z), \tag{78}$$

where for the sake of simplicity, the amplitude of the field A has been approximated to 1. Substituting Eq. (72) in Eq. (78), the field exiting the system is given by:

$$\overset{\frown}{A}_{\phi}(p,z)=H^{0}\left\{1+i\frac{2\pi}{\lambda}\overset{\vee}{f}_{\phi}(p,z)\right\}=i\frac{2\pi}{\lambda}\overset{\frown}{\overset{\vee}{f}}{}^{0}_{\phi}(p,z), \tag{79}$$

where it has been assumed that $H\{1\}=0$. This equation indicates that the Hilbert transform of the field that is exiting the object is proportional to the Hilbert transform of the projection data for every height z, while its irradiance is expressed by:

$$I_\phi(p,z)=\frac{4\pi^2}{\lambda^2}\left(\overset{\cap}{\underset{\vee}{f_\phi^0}}(p,z)\right)^2. \tag{80}$$

In this manner, it is possible to obtain the module of the Hilbert transform of the projection, using:

$$\left|\overset{\cap}{\underset{\vee}{f_\phi^0}}(p,z)\right|=\frac{\lambda}{2\pi}\sqrt{I_\phi(p,z)}\,, \tag{81}$$

where the bars indicate absolute value. It is important to mention that the field at the exit of the system is proportional to the directional Hilbert transform of the input field with respect to p. Once a slice with height z has been selected, the Hilbert transform of the parallel projection at an angle ϕ is given. If this modified projection were extracted from the image plane, the reconstruction obtained would be extracted with the theory that is presented in section 5. But since this field is observed with a detector such as a CCD camera, then the result obtained would be proportional to the module of the Hilbert transform of the projection.
In order to validate the data indicated in Eq. (81) as projection data, it is necessary to use two properties of the Radon transform. In the case of the symmetry property,

$$\left|\overset{\cap}{\underset{\vee}{f_{\phi+\pi}^0}}(p,z)\right|=\left|\overset{\vee}{f_{\phi+\pi}^0}(p,z)\otimes\otimes\frac{\delta(z)}{\pi p}\right|=\left|-\overset{\vee}{f_\phi^0}(-p,z)\otimes\otimes\frac{\delta(z)}{\pi(-p)}\right|, \tag{82}$$

where it has been substituted in the symmetry property of the projection (Eq. 1). The first negative sign on the right-handed side of this equation can be eliminated for the sake of convenience due to the presence of the bars, so

$$\left|\overset{\cap}{\underset{\vee}{f_{\phi+\pi}^0}}(p,z)\right|=\left|\overset{\vee}{f_\phi^0}(-p,z)\otimes\otimes\frac{\delta(z)}{\pi(-p)}\right|=\left|\overset{\cap}{\underset{\vee}{f_\phi^0}}(-p,z)\right|, \tag{83}$$

which proves this property. For the zero momentum property, it is possible to state that

$$\int_{-\infty}^{\infty}dp\left|\overset{\cap}{\underset{\vee}{f_\phi^0}}(p,z)\right|=\int_{-\infty}^{\infty}dp\left|\overset{\vee}{f_\phi^0}(p,z)\otimes\otimes\frac{\delta(z)}{\pi(p)}\right|. \tag{84}$$

What lies in the parallel bars is the Hilbert transform of the projection for any height z. Nevertheless, it has been proved in section 6.2 that the zero momentum for this function is

null (Eq. 61). Therefore, in general, the area under the curve in the ranges in which the function is greater or equal to zero is identical to the area under the curve for the ranges in which the function is negative. This means that the positive area is equal to the negative, being the total area equal to zero. An intuitive proof of the fact that the areas in Eq. (84) remains constant, independently of how the projection angle is oriented, in the sense o the energy captured by the detector. Since the filter used in the Fourier plane of the 4*f* system has only one phase, this implies that there is no change in the energy of the light wave at the exit in the image plane when compared with the input in the object plane. Therefore, the irradiance in both the object plane, as well as in the image plane, is conservative. Hence, it is possible to intuitively conclude that the zero momentum of the Radon transform is satisfied for the projection that was obtained from the absolute value of the Hilbert transform of the projection. A more formal demonstration of this property remains beyond the purpose of this chapter.

Let us suppose that the slice function $f(x,y)$ with a rectangular domain in the way defined in Fig. 4a is shown in Fig. 9a1. The accumulated optical path of a light beam that crosses this slice $\check{f}_{\phi}(p)$ is calculated using the projection integral for different angles in the range $\phi \in (0,2\pi)$, in the way explained in section 6.3.1.1. The image obtained is shown in Fig. 9a2. The column (b) of this figure shows the optical field $A_{\phi}(p)$ that exits the slice of the object $f(x,y)$. The real and imaginary parts are shown in Fig. 8b1 and 8b2, respectively. The column (c) present the Hilbert transform of the field that exits the slice of the object. The real and imaginary parts are in Fig. 9c1 and 9c2, respectively. The upper row of the following column show the irradiance of $\hat{A}_{\phi}(p,z)$, which is considered as a modified sinogram: a Hilbert-sinogram of the irradiance. In Fig. 9d2, the zero momentum of the Radon for the HIS is presented. It is important to note that the symmetry property and the zero momentum of the Radon transform are demonstrated, as was expected. Under the filtered retroprojection algorithm, the tomographic reconstruction is obtained. Fig. 9e shows the reconstruction obtained in 2D and 3D. In this algorithm, the rectangular filter is used, while the impulse response to this filter is used to filter the projection data using a numerical convolution.

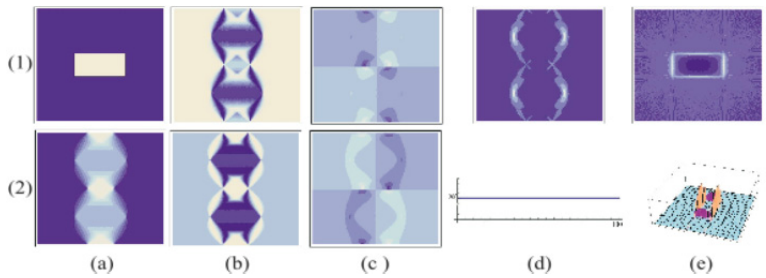


Fig. 9. Numerical simulation: (a) slice of the object (1) and sinogram (2); (b) field exiting the slice of the object: real (1) and imaginary (2); (c) Hilbert transform exiting the slice of the object, (d) Hilbert-sinogram of the irradiance (1) zero momentum of the Radon transform (2); and (e) reconstruction of the edge enhancement: en 2D (1) and 3D (2).

5.3.2.2 Thick phase objects

In the last section, the application of the tomographic theory of edge enhancement to thin phase objects was presented, where the first two terms in the Taylor series were used. In this section, the thick phase objects are considered, in which the field that exits the object cannot be approximated with only a few terms of the series. In this case, the irradiance of the partial Hilbert transform respect to p of the field that goes out of the object is considered as the projection data and it is proven that these data satisfies the two properties of the Radon transform required for the reconstruction of a slice function.
As was mentioned before, a CCD camera would observe the irradiance of the field existent in the image plane of the 4*f* system, as indicated in Eq. (80), which is now considered as the projection data:

$$I_\phi(p,z)=\left|\overset{\frown}{A}_\phi(p,z)\right|^2=\overset{\frown}{A}_\phi(p,z)\overset{\frown}{A}_\phi^*(p,z). \tag{85}$$

which satisfies the symmetry property. Now, let us verify this in the image plane, where the field is described by Eq. (76):

$$\overset{\frown}{A}_{\phi+\pi}(p,z)=\frac{\delta(z)}{\pi p}\otimes\otimes A_{\phi+\pi}(p,z)=\frac{\delta(z)}{\pi p}\otimes\otimes A_\phi(-p,z)=-\overset{\frown}{A}_\phi(-p,z). \tag{87}$$

The negative sign in the last term indicates that the partial Hilbert transform respect to p of the field at the exit of the object does not satisfy the property. Similarly, for the conjugated amplitude, it is possible to prove it:

$$\overset{\frown}{A}_{\phi+\pi}^*(p,z)=-\overset{\frown}{A}_\phi^*(-p,z). \tag{88}$$

Finally, for the irradiance in the image plane (Eq. 85), we have

$$I_{\phi+\pi}(p,z)=\overset{\frown}{A}_{\phi+\pi}(p,z)\overset{\frown}{A}_{\phi+\pi}^*(p,z)=\left[-\overset{\frown}{A}_\phi(-p,z)\right]\left[-\overset{\frown}{A}_\phi^*(-p,z)\right]=I_\phi(-p,z), \tag{89}$$

which proves the property. For the zero momentum of the Radon transform,

$$\int_{-\infty}^{\infty}dpI_\phi(p,z)=\int_{-\infty}^{\infty}dp\,\overset{\frown}{A}_\phi(p,z)\overset{\frown}{A}_\phi^*(p,z). \tag{90}$$

Applying Parseval's theorem,

$$\int_{-\infty}^{\infty}dpI_\phi(p,z)=\int_{-\infty}^{\infty}dw\,\overset{\sim}{\overset{\frown}{A}}_\phi(w,z)\overset{\sim}{\overset{\frown}{A}}_\phi^*(w,z)=\int_{-\infty}^{\infty}dw\left[-i\,\mathrm{sgn}(w)\tilde{A}_\phi(w,z)\right]\left[i\,\mathrm{sgn}(w)\tilde{A}_\phi^*(w,z)\right], \tag{91}$$

where it has been assumed that $\overset{\sim}{\overset{\frown}{A}}_\phi(w,z)=-i\,\mathrm{sgn}(w)\tilde{A}_\phi(w,z)$ is the partial Fourier transform respect to p of the field in the image plane of the 4*f* system, and $A_\phi(w,z)=\Im\{A_\phi(p,z)\}$ is the partial Fourier transform respect to p of the field that exits the object. Accepting the identity $\mathrm{sgn}^2(w)=1$, it is possible to simplify this expression, resulting in:

$$\int_{-\infty}^{\infty} dpI_{\phi}(p,z) = \int_{-\infty}^{\infty} dw \tilde{A}_{\phi}(w,z)\tilde{A}_{\phi}^{*}(w,z), \tag{92}$$

and by using again Parseval's theorem, Eq. (92) can be written in the following way:

$$\int_{-\infty}^{\infty} dpI_{\phi}(p,z) = \int_{-\infty}^{\infty} dpA_{\phi}(p,z)A_{\phi}^{*}(p,z). \tag{93}$$

This expression states that the intensity of the object plane and in the image plane are the same, which is consistent, due to the fact that the filter used has only one phase. Now, if Eq. (72) is substituted in Eq. (93), the following expression is finally obtained:

$$\int_{-\infty}^{\infty} dpI_{\phi}(p,z) = \int_{-\Gamma/2}^{\Gamma/2} dpA\exp\left[i\frac{2\pi}{\lambda}\overset{\vee}{f}_{\phi}(p,z)\right]A\exp\left[-i\frac{2\pi}{\lambda}\overset{\vee}{f}_{\phi}(p,z)\right] = A^2\Gamma, \tag{94}$$

where Γ is the width of the detector, where the irradiance is zero outside of this width. Considering the irradiance in the image plane as the projection data, the zero momentum of the Radon transform is the total energy in the finite detector, and is independent of ϕ. Therefore, the symmetry property and the zero momentum of the Radon transform have been proved in a satisfactory way, so the irradiance in the image plane can be considered as the projection data, and thus, the tomographic reconstruction must present a consistent image.
In Fig. 10, there are two examples: in the row 1, a uniform ring is considered as the slice of the object, and in row 2, a non-symmetrical slice of the object is considered as the slice of the object. In column (a), we present the slices of the object. The set of data in the image plane for every possible projection angle form a modified sinogram known here as a Hilbert-sinogram of the irradiance (HIS) is displayed in column (b). In column (c), the zero momentum of the Radon transform is displayed, and finally, in columns (d) and (e), both of the reconstructions in 2D and 3D are presented, respectively.

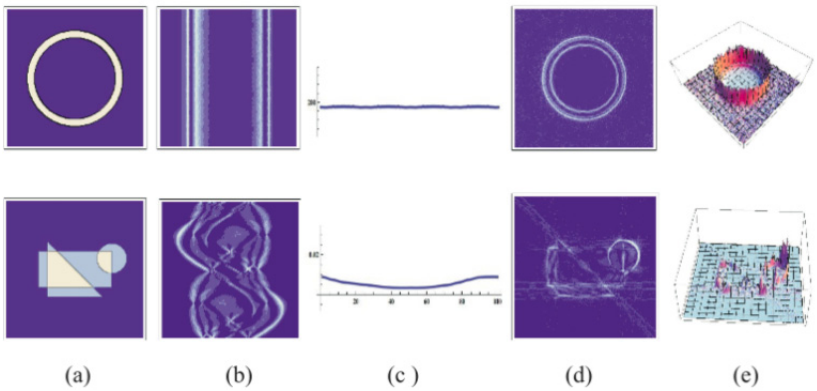


Fig. 10. Numerical simulation: (a) Two test slices: a uniform ring and a non-uniform object, (b) HIS; (c) the zero momentum of the Radon transform (2); (d, e) the reconstruction of the edge enhancement in 2D (1) and 3D (2).

The irradiance in the image plane of the 4*f* optical system using a phase filter of π radians has been considered as the projection data in optical tomography of phase objects, and this has been mathematically demonstrated, that the irradiance satisfies the symmetry property, as well as with the zero momentum of the Radon transform. It has been proven that for every possible projection angle, it is possible to obtain a modified sinogram, which is called a Hilbert-sinogram of the irradiance. As a consequence of the direct use of this HIS, the reconstruction obtained consists of images that show an isotropic edge enhancement for the numerical simulation. Thus, the filtered Hilbert transform not only serves to detect the phase projections, but it also is capable to reinforce tomographic images as an additional characteristic.

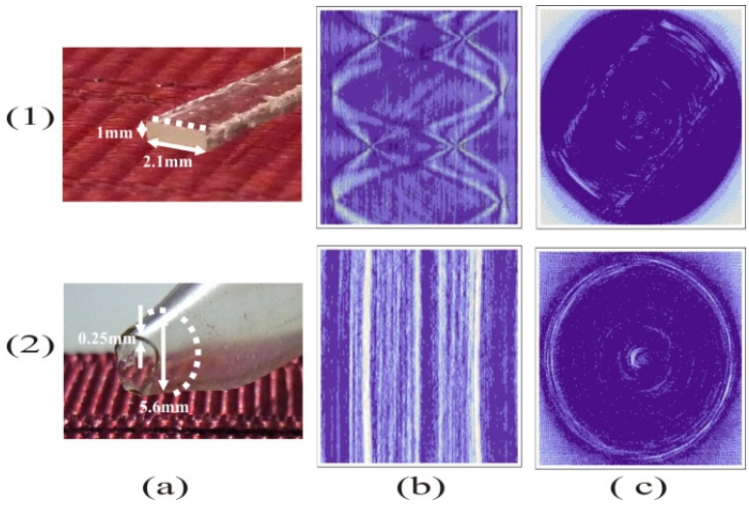


Fig. 11. Experimental results: (a1) a rectangular piece, (a2) pipette, both made of Pyrex, with a refraction index of 1.52; (b) their respective Hilbert-sinogram of irradiance obtained experimentally; and (c) the reconstruction of a slice of the object, respectively for each of the objects under study, where the slice is indicated by the dotted line in (a).

5.3.2.3 Experimental results

In this section, the experimental results for the tomography of phase objects with an isotropic edge enhancement are presented. The 4*f* system shown in Fig. 8 is a telecentric system that can be implemented in laboratory table, in order to achieve the optical phase object tomography. In the experimental implementation, the lenses used have a focal distance of $f_L = 479mm$; the step filter of π radians was found adequate for the wavelength used of $\lambda = 632.8nm$. In the experimental setup, the object under study is submerged on one of its ends in a liquid opening with immersion oil. This oil has a nominal index of 1.515. In this manner, the refractional limit condition is satisfied, while the other end is suspended by the axis of a step-wise motor, with a resolution of 1.8º for every step, which is controlled by a program designed in LabVIEW with an interface for a PC, which also controls the CCD camera, and selects the line of interest in the object, defining the slice of the object.

With this experimental array, and with the control managed from the PC, it is possible to place the sample at the desired projection angle. In the initial position of the sample, it is considered in a rectangular coordinate system, which rotates around the z axis, and where ϕ is the azimuthal rotation angle respect to the x axis, for z constant. In order to experimentally obtain the Hilbert-sinogram of the irradiance, the CCD camera is placed in the image plane of the 4*f* system and is controlled through the visual interface, synchronizing the acquisition of the projection data. In order to obtain the retroprojection data that will form the Hilbert sinogram of the irradiance, it is necessary to select a data line, and it is maintained fixed during the acquisition of all the projections. In the initial position of the object, it is considered that $\phi = 0$ as the first projection angle. The data observed in this row in the image plane using the CCD camera are placed in the new image where the Hilbert-sinogram of the irradiance will be constructed. Once these procedure has been finished, the step-wise motor rotates the sample one step of 1.8º, generating a second projection angle, and it is noteworthy to state that due to this rotation, the distribution of the field has been modified. Once again, the same row of data is selected to be observed by the CCD camera and they are arranged in the second row of the image where the Hilbert-sinogram of the irradiance is being constructed. Repeating this for each one of the angular steps of the motor until a complete rotation of 360º has been accomplished, the Hilbert-sinogram of irradiance has been completely constructed for one slice of the object, so this will have 200 rows. The reconstruction of the slice with an isotropic edge-enhancement is obtained submitting the HIS to data processing, using the filtered retroprojection. It is important to state that due to the symmetry property of the parallel projections, it was only necessary to use the projections that lie in the range $(0^0, 180^{0'})$.

In Fig. 11, there are experimental results for two phase objects. Fig. 11a presents the photographs of the transparent objects that were studied. The sample number 1 is an object that is a rectangular piece of 1x3x20 mm, (1) a pipette of 5.6 mm of external diameter and 0.25 mm of thickness. Both objects are made of Pyrex glass with a nominal refraction index of 1.52. Fig. 11b show the HIS experimentally obtained of a slice of each one of the objects. The slices that were selected for each one of the objects are shown in Fig. 11a by a dotted line. Fig. 11c presents their respective tomographic reconstruction using the filtered retroprojection algorithm. The tomographic images that were obtained show edge enhancement. These enhancements are related principally to the refraction index over some slice in particular.

6. Conclusions

In this chapter, there was a discussion of the mathematical fundamentals of parallel projection tomography, and there was also a demonstration of the mathematical method for directional edge-enhancement tomography. It is important to state that this technique is possible thanks to the mathematical relationship existent between the Radon transform and the Hilbert transform. Numerical simulations were done, and it was proven how this technique can be applied in phase object tomography, presenting reconstructed images with isotropic edge enhancement. Note that an analytic proof of the isotropic edge enhancement continues being an open problem.

Three conditions were required in order to make possible the implementation of the edge-enhancement tomography:

1. The existence of a relationship between the Hilbert transform and the Radon transform.
2. The verification of the symmetry property and the zero momentum of the Radon transform of the Hilbert transform of the projections.
3. The possibility of an experimental implementation of the Hilbert transform

Due to its analytic form, the third one can be implemented in an invariant and linear system because the spectrum is the product of two functions in Fourier-space, where one of the functions can be interpreted as the filter function. This is why the optical implementation of the Hilbert transform is possible, where the step phase filter is used because it is modeled through the *signum* function. Moreover, in terms of detection, this method is more direct and faster than the usual methods used in interferometric tomography, because there is no data preprocessing implied in the obtaining of the projection data.

Speaking in general terms, this same detection method could be extended using another type of filter that can make edge-enhancement, such as the full or fractional directional derivative. This technique could be employed using another type of probe that can make possible the implementation of a linear and invariant system. We think that the tomography through electric capacitance can be done in this way, because it is possible to implement a system of this type, in the way done in communication systems.

7. Acknowledgment

This work was partially supported by PROMEP under grant PROMEP/103.5/09/4544 and by Vicerrecoría de Investigación y Estudios de Posgrado under grant MEFC.

8. References

Abhilash, G., (2006). Hilbert transform: A simple, signal-theoretic formulation, *Article series of the Signals Research Laboratory*

Almeida, L.B., (1994). The fractional Fourier transform and time-frequency representation. IEEE Trans. Sig. Proc., 42:3084–3091

Andersen, A. H. and Kak, A. C., (1982). Digital ray tracing in two-dimensional refractive fields, *Journal of the Acoustical Society of America*, Vol. 72, pp. 1593-1606

Baba, N. and Murata, K., (1977). Filtering for image reconstruction from projections, *Journals Optics Society America*, pp.662-668

Bates, R. H. T. and Peters, T. M., (1983). Overview of computerized tomography with emphasis on future developments, *Proceedings IEEE*, Vol. 71, pp. 356-372

Bracewell, R. N. and Riddle, A. C., (1967). Inversion of fan-beam scans in radio astronomy, *The Astrophysical Journals*, Vol.150, pp. 427-434

Bracewell, Ronald N.; McGraw-Hill (2000). *The Fourier transform and its application*, Stanforf University

Bracewell, R. N., (1956). Strip integration in radio astronomy, *Australian Journal of Physics*, Vol. 9, pp. 198-217

Braslavsky, I. and Lipson Technion, S. G., (1998). Interferometric Tomography Measurement of the Temperature Field in the Vicinity of a Dendritic Crystal Growing from a

Supercooled Melt. Transactions of Optical methods and data processing in heat and fluid flow, IMECHE, London, pp. 423-432.

Brown, W. P., Jr., (1966). Validity of the Rytov Approximation in Optical Propagation Calculations. *Journals Optics Communications*, Vol. 56, pp.1045-1052

Bruning, J. H., Herriott, D.R., Gallagher, J. E., Rosenfeld, D. P., White, A. D., and Brangaccio, D. J., (1974). Digital Wavefront Measurement Interferometer for Testing Optical Surfaces and Lenses, Applied Optics, Vol. 13, pp. 2693

Byer, R. L., Garbuny, M., (1973). Pollutant detection by absorption using Mie scattering and topographic targets as reflectors, *Applied Optics,* Vol. 12, No.7, pp. 1496-1505

Byer, R. L., Shepp, L. A., (1979). Two-dimensional remote air-pollution monitoring via tomography, *Optics Letters*, Vol.4, pp. 375-377

Chiu, M. Y., Barrett, H. H., Simpson, R. G., Chou, C., Arendt, J. W., and Gindi, G. R., (1979). Three dimensional radiographic imaging with a restricted view angle, *Journal of the Optical Society of America* ,Vol. 69, pp. 1323-1330

Chiu, M. Y., Barrett, H. H., Simpson, R. G., (1980). Three dimensional reconstruction from planar projections, *Journal of the Optical Society of America*, Vol. 70, pp. 755-762

Davis, Jeffrey A. and Nowak, Maria D., (2001). Selective edge enhancement of images with an acousto-optic light modulador, *Applied Optics*, Vol. 23, pp. 4835-4839

Davis, Jeffrey A., McNamara, D. E., Cottrell, D. M., and Campos, J., (1998). Analysis of the fractional hilbert transform, *Applied Optics*, Vol. 37, pp. 6911-6913

Davis, Jeffrey A., McNamara, D. E., Cottrell, D. M., and Campos, J., (2000). Image processing with the radial Hilbert transform: theory and experiments, *Optics Letters*, Vol. 25, pp.99-101

Davis, Jeffrey A., and Nowak, M. D., (2002). Selective-edge enhancement of images with an acousto-optic light modulator, *Applied Optics*, Vol. 41, pp. 4835-4839

Deans, Stanley R.; John Wiley & Sons (1983). *The Radon Transform and Some of its Applications,* New York

DeRosier, D. J. and Klug, A., (1968). Reconstruction of three dimensional structures from electron micrographs, *Nature*, Vol. 217, pp. 130-134

Eden, O. Tretiak. M., and Simen, M., (1969). International structures dor three dimensional images, in *Proc. 8th Int. Conference on Medicine Biological Engineering, Chicago, IL*

Fatemi, M. and Kak, A. C., (1980). Ultrasonic B- scan imaging: Theory of image formation and a technique for restoration, *Ultrasonic Imaging on ScienceDirect*, Vol. 2, pp. 1-47

Glover, G. H., (1982). Compton scatter effects in CT CT reconstructions, *Medical Physics*, Vol. 9, pp. 860-867

Goodman, Jooseph W.; J. Wiley & Sons (1985). *Introduction to Fourier optics*

Haykin, S., Ed. Englewood Cliffs, NJ; Prentice-Hall (1985). *Tomographic imaging with diffracting and non-diffracting sources, in Array Signal Processing*

Herman, G. T. and Naparstek, A., (1977). Fast image reconstruction based on a Radon inversion formula appropriate for rapidly collectes data, *SIAM Journals on Applied Mathematics*, Vol. 33, pp. 511-533

Horn, B. K. P., (1978). Density reconstruction using arbitrary ray sampling schemes, *Proccedings of the IEEE*, Vol. 66, pp. 551-562

Horn, B. K. P., (1979). Fan-beam reconstruction methods, *Proccedings of the IEEE*, Vol. 67, pp. 1616-1623

Hsu, Hwei P.; Addison-Wesley Iberoamericana (1970). *Analisis de Fourier*, Nueva York

Kak, A. C., Slaney, M.; New York (1987). *Principles of Computarized Tomographic Imaging*, IEEE Press

Kenue, S. K. and Greenleaf, J. F., (1979). Efficient convolution kernels for computerized tomography, *Ultrason Imaging*, Vol. 1, pp.232-244

Kenue, S. K. and Greenleaf, J. F., (1979). Efficient convolution kernels for computerized tomography, *Ultrason Imaging*, Vol. 1, pp.232-244

Kujawinska, Malgorzata, Kniazewski, Pawel, (2006). Enhanced interferometric and photoelastic tomography for 3D studies of phase photonics elements, Proceedings of the Symposium on Photonics Technologies for 7th Framework Program Wroclaw 12-14 October, pp. 467-471

Kwoh, Y. S., Reed, I. S., and Truong, T. K., (1977). A generalized $|w|$-filter for 3-D reconstruction, *IEEE Transactions on Nuclear Science*, Vol. NS-24, pp. 1990-1977

Lakshminarayanan, A. V., (1975). Reconstruction from divergent ray data, Tech. Rep. 92, Dep. Of Computer Science, State Univ. of New York at Buffalo

Lewitt, R. M. and Bates, R. H. T., (1978). Image reconstruction from projections, *Optik*, Vol. 50, pp. 19-33

Lewitt, R. M., (1979). Ultra-fast convolution approximation for computerized tomography, *IEEE Transactions on Nuclear Science*, Vol. NS-26, pp. 2678-2681

Meneses-Fabian, Cruz, Montes-Perez, Areli, and Rodriguez-Zurita, Gustavo, (2011). Directional edge enhancement in optical tomography of thin phase objects, *Optics Express*, Vol. 19, pp. 2608-2618

Meneses-Fabian, Cruz, Rodriguez-Zurita, Gustavo, Rodriguez-Vera, Ramon, Vazquez-Castillo, Jose F., (2003).Optical tomography with parallel projection differences and Electronic Speckle Pattern Interferometry, Optics Communications, Vol. 228, pp. 201-210

Montes-Perez, Areli, Meneses-Fabian, Cruz, Rodriguez-Zurita, Gustavo, (2011). Optical Hilbert-transform to isotropic edge-enhancement in phase object tomography, *Optics Express*, Vol. 19, pp. 5350-5356

Oppenheim, B. E., Ter Pogossian, M. M. et. al., Eds. Baltimore, MD ; (1975). *Reconstruction tomography from incomplete projections, in reconstruction Tomography in Diagnostic Radiology and Nuclear Medicine*, University Park Press

Ornelas-Rodríguez, F.J., Rodríguez-Zurita, G., Rodríguez-Vera, R., Pastrana Sánchez , R., de la Rosa-Miranda, E., (1999). Zero-order moment of the Radon transform in tomography: some further remarks, Optics Communications, Vol. 161, pp. 19-24

Poularikas, A. D.; CRC PRESS, IEEE PRESS (2000). The transforms and applications handbook, United States of Ameirca

Philipp, H., Neger, T., Jäger, H. and Woisetschläger, J., (1992). Optical tomography of phase objects by holographic interferometry, Elsevier Ltd. Vol. 10, pp. 170-181

Ramachandran, G. N. and Lakshminarayanan, A. V., (1971). Three dimensional reconstructions from radiographs and electron micrographs: Application of convolution instead of Fourier transforms, *Proceedings of the National Academy of Sciences*, Vol. 68, pp. 2236-2240

Rodríguez-Zurita, G., Meneses-Fabián, C., Pérez-Huerta, J.-S., Vázquez- Castillo, J.-F., (2005). Tomographic directional derivative of phase objects slices using 1-D derivative spatial filtering of fractional order 1/2, *20th Congress of the International Commission for Optics*, *Proceedings SPIE*, Vol. 6027, pp. 0405-056

Rodríguez-Zurita, G. and Pastrana-Sánchez, R., (1997). Tomographyc phase-distribution edge enhancement with differences projectionsa proposal, 4 Revista Mexicana de Física, Vol. 43, pp. 167-174

Rowley, P. D., (1969). Quantitative interpretation of three dimensional weakly refractive phase objects using holographic interferometry, *Journal of the Optical Society of America*, Vol 59, pp. 1496-1498

Sato, T., Norton, S. J., Linzer, M., Ikeda, O., and Hirama, M., (1980). Tomographic image reconstruction from limited projections using iterative revisions in image and transform spaces, *Applied Optics*, Vol. 20, pp. 395-399

Seeley, R., Ed Reverte (1970). *Introducción a las Series e Integrales de Fourier*

Shepp, L. A. and Logan, B. F., (1974). The Fourier reconstruction of a head section, *IEEE Transactions on Nuclear Science*, Vol. NS-21, pp. 21-43

Smith, B. D., (1985). Image reconstruction from cone-beam projections; Necessary and sufficient conditions and reconstruction methods, *IEEE Transactions on Medical Imaging*, Vol. MI-4, pp. 14-25

Stark, H., Woods, J. W., Paul, I., and Hingorani, R., (1981). Direct Fourier reconstruction in computer tomography, *IEEE Transactions Acoustics and Speech Signal Processing*, Vol. ASSP-29, pp. 237-244.

Takeda, M., Ina, H., and Kobayashij, S., (1982). Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry, Optics Society American, Vol. 72, pp. 156-160

Tam, K. C. and Perez-Mendez, V., (1981). Tomographical imaging with limited angle input, *Journal of the Optical Society of America*, Vol. 71, pp. 582-592

Tanaka, E. and Iinuma, T. A., (1975). Correction functions for optimizing the reconstructed image in transverse section scan, *Physics in Medicine Biology*, Vol. 20, pp. 789-798

Vest, C. M., and Radulovic, P. T.; Pergamon (1977). *Measurement of three-dimensional temperature fields by holographic interferometry, in Applications of Holography and Optical Data Processing*, E. Marcom, A. A. Friesem, and E. Wiener-Avnear, pp. 241-249, Oxford

Wernecke, S. J. and D'Addario, L. R., (1977). Maximum entropy image reconstruction, *IEEE Transactions Computers*, Vol. C-26, pp. 351-364

Wolf, E., (1969). Three-dimensional structure determination of semi-transparent objects from holographic data. *Optics Communications*, Vol.1, pp.153-156

Wu, Donglou and He, Anzhi, (1999). Measurement of three-dimensional temperature fields with interferometric tomography, Applied Optics, Vol. 38, pp. 3468 -3473

Zernike, F., (1955). How I discovered phase contrast, *Science*, Vol. 121, pp. 345-349

10

Model Approximation and Simulations of a Class of Nonlinear Propagation Bioprocesses

Emil Petre and Dan Selişteanu
Department of Automatic Control, University of Craiova
Romania

1. Introduction

It is well known that the biotechnology is one of the fields that over the last three decades has a very high and quick development. Therefore, due to their advantages, the control of industrial bioprocesses has been an important practical problem attracting wide attention. The main motivation in applying control methods to such living systems is to improve their operational stability and production efficiency. The operation in Stirred Tank Reactors (STR) has been and it is still a widely used technology in fermentation bioprocesses. But, other new technologies such as fixed bed, fluidized bed or air lift reactors, are considered for bioprocesses operation. These reactors present several advantages over the "classical" STRs. For instance, the fixed bed and fluidized bed reactors are characterized by higher production performance, i.e. larger production capacity and higher productivity (Bastin & Dochain, 1990; Bastin, 1991; Bouaziz & Dochain, 1993).

From mathematical point of view, the dynamics of these processes are characterized by partial differential equations and therefore are classified as distributed parameter systems (Bastin & Dochain, 1990; Bouaziz & Dochain, 1993; Christofides, 2001; Dochain et al., 1992). For instance, the concentrations of the reactants and products are not anymore homogeneous in the whole reactor, like in STRs, but are characterized by a spatial profile along the reactor. It is clear that the distributed parameter feature of these systems makes the control problem even more difficult (Bouaziz & Dochain, 1993; Christofides, 2001; Dochain et al., 1992; Petre & Selişteanu, 2007a; Slotine & Li, 1991). Therefore, the modelling and simulation of them, which are the objectives of this chapter, are associated with the formulation of the process model using partial differential equations (PDEs), in general nonlinear, and with the computation of the solution. Since in the most of cases for this kind of systems, with the exception of a few simple cases, there are no analytical methods for finding the solution of the involved equations, the following alternatives are commonly employed (Christofides, 2001; Petre & Selişteanu, 2007a; Slotine & Li, 1991; Vilas, 2008):

- To assume that these processes behave like lumped parameter systems (the states are only time dependent).
- To use classical numerical methods like finite differences, finite elements or finite volumes.

These methods are based on discretization techniques which allow us to approximate the infinite set of numbers that represent a continuous function by means of a finite set of

parameters (Bouaziz & Dochain, 1993; Christofides, 2001; Dochain et al., 1992; Petre & Selişteanu, 2007a; Slotine & Li, 1991; Vilas, 2008).
The first option is only valid when the spatial distribution is negligible as compared with the time evolution, for instance in reactors where the homogenization of the medium is achieved by means of stirring devices (Dochain & Vanrolleghem, 2001; Vilas, 2008). Nevertheless, in the remaining cases it is necessary to use the second alternative. Its main inconvenience is that the numerical solution is computationally involved (especially in 2D or 3D spatial domains) making the approach unsuitable for real time tasks like control or online optimization (Bouaziz & Dochain, 1993; Christofides, 2001; Dochain et al., 1992; Petre & Selişteanu, 2007a; Slotine & Li, 1991; Vilas, 2008).
An alternative to these classical numerical methods is the development of some techniques for the projection of the PDEs onto a low dimensional subspace. In accordance to these techniques, the original PDEs are transformed into a set of ordinary differential equations (ODEs) known as reduced order model (Aksikas et al., 2007; Americano da Costa Filho et al., 2009; Bouaziz & Dochain, 1993; Christofides, 2001; Dochain et al., 1992; Hoo & Zheng, 2001; Petre et al., 2007; Shvarstman et al., 2000).
As a result, the first objective of this chapter is to provide the mathematical tools, which are used for most of numerical methods, for solving PDEs and, on this basis, to give a brief outline of the most commonly employed techniques. Among the different alternatives, some of them based on Garlerkin scheme will be described and used in this chapter. In particular, the finite element method will be chosen on the basis of its flexibility and reduced order models since they are the most efficient (Vilas, 2008). These reduced order models obtained in this way can be used either for the process simulation or computation of their solution.
As we mentioned above, over recent years, a considerable research effort was concentrated on the design of control policies for distributed process systems (Christofides, 2001). Standard approaches to control this kind of systems are based on the spatial discretization of the original set of PDEs to obtain a set of ODEs. This allows us to employ standard finite-dimensional methods just described above to construct the controller (Christofides, 2001; Dochain et al., 1992; Hoo & Zheng, 2001; Shvarstman et al., 2000). However, this approach can result in a set of ODEs of high dimensionality which could make the approach unsuitable for real time applications. Also, the controllability and observability properties would depend on the number of discretization points as well as its location and may lead to a poor control quality (Christofides, 2001). Due to these disadvantages, new methods based on spectral decomposition techniques, which take into account the spatially distributed nature of these systems, have developed (Aksikas et al., 2007; Shi et al., 2006). This approach uses the Galerkin method so as to approximate the system by a low-dimensional set of ODEs to design the controller (Aksikas et al., 2007; Shvarstman et al., 2000). In (Americano da Costa Filho et al., 2009) this approach is used in combination with the Lyapunov's direct method to derive stabilizing controllers applied in the case of chemical processes that are carried out in tubular reactors.
This chapter is an extended work of the research achieved in some works of the authors: (Petre, 2003; Petre & Selişteanu, 2005, 2007a, 2007b; Petre et al., 2007, 2008; Petre, 2008; Selişteanu & Petre, 2004), and deals with the approximation and simulations of the dynamical model for a class of nonlinear propagation bioprocesses.
First, the dynamics of a class of propagation bioprocesses involving n components and m reactions that are carried out in fixed bed reactors without dispersion is analyzed. Since the dynamics of these bioprocesses are described by partial differential equations, either for

simulation but especially for their controlling, one method consists of approximation of these infinitely order models by finite order models. These approximate models are in fact a set of ordinary differential equations obtained here by orthogonal collocation method. More exactly, infinitely dimension of the initial parameter distributed model will be reduced by approximating the partial derivative equation of each reaction component by a finite number, equal to p+1, of ordinary differential equations at p+1 discrete spatial positions along the bioreactor. These points are chosen as zeros of some orthogonal polynomials. Since it is difficult to know the connections between the original distributed parameter model and its approximate version (Christofides, 2001), our objective is to analyze the behaviour of both models to observe their intrinsic dynamical properties. This is realized by simulations conducted in the case of a fixed bed bioreactor without diffusion (Dochain et al., 1992; Petre et al., 2007; Petre & Selişteanu, 2007a).

In the following the control problem of these classes of propagation bioprocesses is analyzed. Since the biotechnological processes have a nonlinear nature, to control these processes some nonlinear control techniques will be used. These techniques not only improve the linear control methods and allow the analysis of strong nonlinearities but it also allow us to deal with model uncertainties and even the controller design may result simpler than in its linear counterpart (Slotine & Li, 1991). A widely extended nonlinear control technique is the feedback linearization (Isidori, 1995; Khalil, 2002; Slotine & Li, 1991), which makes use of algebraic transformations to obtain a closed loop linear system in which the conventional control techniques can be applied. The main inconveniences of this technique are two: firstly, the tracking control problem may lead to complex transformations and secondly, model uncertainty may affect the control performance. Therefore, it is necessary to apply adaptive and robust-adaptive control techniques able to drive the system to the desired reference despite the presence of uncertainties.

Consequently, by using the obtained results in (Petre & Selişteanu, 2005, 2007a; Petre et al., 2007, 2008; Petre, 2008; Selişteanu & Petre, 2006), to control the mentioned propagation bioprocesses, in this chapter a class of nonlinear adaptive controllers are designed based on their finite order models. The nonlinear controller design is based on the input-output linearizing technique. The information required about the process is the measurements of the state variables and its relative degree. It must be noted that if for the analyzed process there are no accessible state variables, these will be estimated by using an appropriate state observer.

Numerical simulations conducted in the case of a fixed bed reactor are included to illustrate the performances of the presented adaptive control strategies.

All simulations are achieved by using the development, programming and simulation environment MATLAB (registered trademark of The MathWorks, Inc., USA).

The chapter is organized as follows. Section 2 introduces the distributed parameter dynamical model for the class of fixed bed reactors. Its reduction to an ordinary differential equation system by orthogonal collocation method is presented in Section 3. A detailed analysis of obtained results by application of this method in the case of a fixed bed reactor without diffusion is presented in Section 4. The adaptive control strategies of propagation bioreactors are developed in Section 5, the performances of the designed adaptive controllers being presented in Section 6. Finally, concluding remarks and further research directions are presented.

2. Dynamical model of fixed bed bioreactors

A fixed bed bioreactor is a reactor where the biomass is immobilized on fixed carriers such as polymers, porous glass or ceramics. Consider a fixed bed bioreactor without dispersion operating in plug flow conditions as shown in Fig. 1, in which takes place a single autocatalytic growth reaction $\phi = \mu X$, where μ is the specific growth rate, with one limiting substrate S and one biomass population X.

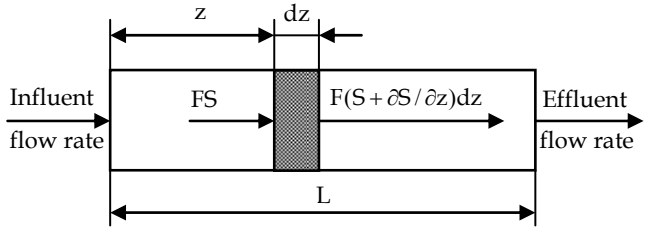


Fig. 1. A schematic view of a fixed bed bioreactor

To achieve the model of this bioreactor consider a section of the reactor with length dz located at a distance z ($0 \le z \le L$) from the bioreactor input. Assuming that along the length L of the bioreactor the cross section is constant and equal to A, then, the volume of this section is dV = Adz.
The mass balance of substrate concentration S around this section is given by:

$$\frac{\partial}{\partial t}(SA\,\mathrm{d}z) = FS - F\left(S + \frac{\partial S}{\partial z}\mathrm{d}z\right) - k_1 \mu XA\,\mathrm{d}z, \tag{1}$$

where the term $\frac{\partial}{\partial t}(SA\,\mathrm{d}z)$ is the time variation of the amount of substrate in the elementary volume dV, FS and $F\left(S + \frac{\partial S}{\partial z}\mathrm{d}z\right)$ are the influent and the effluent substrate into, respectively from the volume dV, where F is the hydraulic flow rate, and the term $k_1 \mu XAdz$ represents the amount of substrate consumed by the biomass in the volume dV, where k_1 is the yield coefficient.
Similarly, the mass balance of the biomass concentration X in the volume dV is given by:

$$\frac{\partial}{\partial t}(XA\,\mathrm{d}z) = FX - F\left(X + \frac{\partial X}{\partial z}\mathrm{d}z\right) + \mu XA\,\mathrm{d}z, \tag{2}$$

where the term $\frac{\partial}{\partial t}(XA\,\mathrm{d}z)$ is the time variation of the amount of biomass in the elementary volume dV and term $\mu XAdz$ represents the amount of biomass produced in the volume dV. Since the biomass is immobilized, the hydraulic terms FX and $F\left(X + \frac{\partial X}{\partial z}\mathrm{d}z\right)$ have to disappear in this equation.
After some calculus, the relations (1) and (2) lead to:

$$\frac{\partial X}{\partial t}=\mu X;\quad \frac{\partial S}{\partial t}=-\frac{F}{A}\cdot\frac{\partial S}{\partial z}-k_1\mu X. \tag{3}$$

The equations (3) constitute the *distributed parameter dynamical model* of the analyzed fixed bed bioreactor. For completeness, we must to define the limit and initial conditions as:

$$S(t,z=0)=S_{in}(t),\quad X(0,z)=X_0(z). \tag{4}$$

where $S_{in}(t)$ is the influent substrate concentration and $X_0(z)$ is the initial immobilized biomass concentration.

Assume now that two reactions take place in bioreactor: (i) an autocatalytic growth reaction with one limiting substrate S and one biomass population X with a reaction rate $\phi=\mu X$, as in the previous case; (ii) a death reaction of microorganisms $X \rightarrow X_d$, where X_d is the non-active biomass. If we assume that the non-active biomass leaves the bioreactor, *the distributed parameter dynamical model* of this fixed bed bioprocess will be described as:

$$\frac{\partial X}{\partial t}=\mu X-k_d X;\quad \frac{\partial S}{\partial t}=-\frac{F}{A}\cdot\frac{\partial S}{\partial z}-k_1\mu X;\quad \frac{\partial X_d}{\partial t}=-\frac{F}{A}\cdot\frac{\partial X_d}{\partial z}+k_d X, \tag{5}$$

where k_d is the death coefficient. The limit and initial conditions are defined as:

$$S(t,z=0)=S_{in}(t);\quad X_d(t,z=0)=0;\quad X(0,z)=X_0(z). \tag{6}$$

Let us define the state vector $\zeta=[X\ S\ X_d]^T$ with the following partitions:

$$\xi_1=X;\quad \xi_2=[S\ X_d]^T. \tag{7}$$

If we denote by $\tilde{r}=[\tilde{r}_1\ \tilde{r}_2]^T=[\mu X\ k_d X]^T$ the reaction rate vector, the model (5) can be rewritten as:

$$\frac{\partial \xi_1}{\partial t}=\tilde{K}_1\tilde{r}(\xi_1,\xi_2);\quad \frac{\partial \xi_2}{\partial t}=-\frac{F}{A}\cdot\frac{\partial \xi_2}{\partial z}+\tilde{K}_2\tilde{r}(\xi_1,\xi_2), \tag{8}$$

where

$$\tilde{K}_1=[1\ \ -1],\quad \tilde{K}_2=\begin{bmatrix}-k_1 & 0\\ 0 & 1\end{bmatrix}, \tag{9}$$

with the limit and the initial conditions:

$$\xi_2(t,z=0)=\xi_{2,in}(t)=[S_{in}\ 0]^T,\quad \xi_1(0,z)=\xi_{10}(z). \tag{10}$$

From the two above examples, one can deduce that in the case of a fixed bed bioreactor in which m biochemical reactions with n reactants take place, among which n_1 are microorganisms fixed on some supports and which remain within the reactor, and n_2 other components flow through the reactor, the distributed parameter dynamical model will be described as:

$$\frac{\partial \xi_1}{\partial t}=\tilde{K}_1\tilde{r}(\xi_1,\xi_2);\quad \frac{\partial \xi_2}{\partial t}=-\frac{F}{A}\cdot\frac{\partial \xi_2}{\partial z}+\tilde{K}_2\tilde{r}(\xi_1,\xi_2), \tag{11}$$

with the following limit conditions:

$$\xi_2(t, z=0) = \xi_{2,in}(t), \quad \xi_1(0, z) = \xi_{10}(z). \tag{12}$$

Usually in (11) and (12), $\xi_1 \in \Re^{n_1}$ is the fixed biomass concentration vector, $\xi_2 \in \Re^{n_2}$ is the other concentration vector, $\xi_{2,in} \in \Re^{n_2}$ is the influent concentration vector, $\tilde{r}(\xi_1, \xi_2) \in \Re^{m}$ is the reaction rate vector, $\tilde{K}_1 \in \Re^{n_1 \times m}$ and $\tilde{K}_2 \in \Re^{n_2 \times m}$ are the yield coefficient matrices having appropriately structures and dimensions.

3. Approximation of the dynamical model via orthogonal collocation

Since the model (11) is infinitely dimensional, in this section we will reduce the model order by approximating it by a set of ordinary differential equations. From (11) one can see that the state variables ξ_1 and ξ_2 are functions of time and space, that is $\xi_1 = \xi_1(t, z)$ and $\xi_2 = \xi_2(t, z)$. We shall reduce the dimension of the model (11) by approximating the partial derivative equation of each component of ξ_1 and ξ_2 by a finite number, equal to p+1, of ordinary differential equations at p+1 discrete spatial positions along the bioreactor. To do this, we expand each variable as a finite sum of products of some time functions and space functions as:

$$\xi_1(t, z) \cong \sum_{i=0}^{p+1} \beta_i(z) \cdot \xi_{1,i}(t); \quad \xi_2(t, z) \cong \sum_{i=0}^{p+1} \beta_i(z) \cdot \xi_{2,i}(t), \tag{13}$$

where $\xi_{k,i}(t) = \xi_k(t, z = z_i)$, $k = 1, 2$; $i = 0, 1, \ldots, p+1$ are the values of $\xi_k(t, z)$ at some discrete spatial positions along the bioreactor, called collocation points, and the basis functions $\beta_i(z)$ are chosen as orthogonal functions (e.g. Lagrange polynomials) such as:

$$\beta_i(z) = \frac{\prod_{k=0, k\neq i}^{p+1} (z - z_k)}{\prod_{k=0, k\neq i}^{p+1} (z_i - z_k)}, \quad i = 0, 1, \ldots, p+1, \text{ with } \beta_i(z_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \tag{14}$$

The integer p in (13) corresponds to the number of interior collocation points determined by *collocation method*. The points $z = z_0$ and $z = z_{p+1}$ correspond to the input (z = 0) and the output (z = L) of the reactor.

It must be noted that the collocation method offers two important advantages: first, its implementation is easier and second, the nature and physical dimension of the state variables remain unchanged after the reduction procedure. Moreover, the orthogonal methods preserve mass balances (Christofides, 2001; Dochain et al., 1992).

According to *collocation method*, the partial derivative of ξ_2 with respect to z appearing in (11) can be written as:

$$\frac{\partial \xi_2}{\partial z} = \sum_{i=0}^{p+1} b_{ji} \cdot \xi_{2,i}(t), \text{ with } b_{ji} = \left. \frac{\mathrm{d}\beta_i(z)}{\mathrm{d}z} \right|_{z=z_j}, i = 0, 1, \ldots, p+1, \; j = 1, \ldots, p+1. \tag{15}$$

By introducing (13)-(15) into (11) and (12), each partial derivative equation is transformed into p+1 differential equations at the p interior collocation points and at the output of the reactor. Thus it is obtained the following n(p+1) order system of one order ordinary differential equations:

$$\dot{x}_1 = K_1\, r(x_1, x_2)\,; \quad \dot{x}_2 = -(F/A)Bx_2 + F_R + K_2\, r(x_1, x_2), \tag{16}$$

where:

$$x_k = \begin{bmatrix} \xi_{k,1} \\ \xi_{k,2} \\ \vdots \\ \xi_{k,p+1} \end{bmatrix}; K_k = \begin{bmatrix} \tilde{K}_k & 0 & \cdots & 0 \\ 0 & \tilde{K}_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{K}_k \end{bmatrix}; r(x_1, x_2) = \begin{bmatrix} \tilde{r}(\xi_{1,1}, \xi_{2,1}) \\ \tilde{r}(\xi_{1,2}, \xi_{2,2}) \\ \vdots \\ \tilde{r}(\xi_{1,p+1}, \xi_{2,p+1}) \end{bmatrix}, k = 1, 2,$$

with

$$x_k \in \Re^{n_k \times (p+1)}, \quad K_k \in \Re^{n_k (p+1) \times m(p+1)}, \quad r \in \Re^{m \times (p+1)};$$

$$B = [B_{ji}], i, j = 1, 2, \ldots, p+1, \text{ with } B_{ji} = diag\{b_{ji}\}, B_{ji} \in \Re^{n_2 \times n_2};$$

$$F_R = \frac{F}{A} \cdot [\tilde{b}_1\ \tilde{b}_2 \cdots \tilde{b}_{p+1}]^T \cdot \xi_{2,in}(t), \tilde{b}_j = diag\{-b_{j0}\}, j = 1, \ldots, p+1, \tilde{b}_j \in \Re^{n_2 \times n_2}.$$

4. Application to a fixed bed bioreactor

In this section, the above presented collocation method will be applied in the case of the distributed parameter dynamical model of the fixed bed bioreactor described by (5)-(6). We will consider four interior collocation points, i.e. p = 4. Firstly, we define the values of the concentration of X, S and X_d and of the specific growth rate μ at each interior collocation point and at the output of the reactor, z_i, $i = \overline{1,5}$ as: $X_i = X(z = z_i)$, $S_i = S(z = z_i)$, $X_{di} = X_d(z = z_i)$, $\mu_i = \mu_i(X_i, S_i, X_{di})$.
For p = 4, the relations in (13) are particularized as follows:

$$X(t,z) \cong \sum_{i=0}^{5} \beta_i(z) \cdot X_i(t), \quad S(t,z) \cong \sum_{i=0}^{5} \beta_i(z) \cdot S_i(t), \quad X_d(t,z) \cong \sum_{i=0}^{5} \beta_i(z) \cdot X_{di}(t), \tag{17}$$

where $\beta_i(z)$ are given by (14). Since $S_0(t) = S_{in}(t)$, $X_{d0}(t) = X_{d,in}(t) = 0$ and $X_0(t) = 0$, from (15) one obtains:

$$\left.\frac{\partial S(t,z)}{\partial z}\right|_{z=z_j} = \sum_{i=0}^{5} b_{ji} S_i(t), \quad \left.\frac{\partial X_d(t,z)}{\partial z}\right|_{z=z_j} = \sum_{i=1}^{5} b_{ji} X_{di}(t), \; j = 1, \ldots, p+1. \tag{18}$$

Let us define the state vectors at collocation points as:

$$x_1 = [X_1\ X_2\ X_3\ X_4\ X_5]^T, \quad x_2 = [S_1\ X_{d1}\ S_2\ X_{d2} \cdots S_5\ X_{d5}]^T. \tag{19}$$

Then, the reduced order model which approximates the exactly infinitely dimensional model (5)-(6) will be described by the following ordinary differential equations:

$$\dot{x}_1 = K_1\, r(x_1, x_2)\,; \quad \dot{x}_2 = -(F/A)Bx_2 + F_R + K_2\, r(x_1, x_2), \tag{20}$$

where:

$$r(x_1, x_2) = [\mu_1 X_1 \;\; k_d X_1 \;\; \mu_2 X_2 \;\; k_d X_2 \cdots \mu_5 X_5 \;\; k_d X_5]^T,$$

$$K_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \; K_2 = \begin{bmatrix} -k_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -k_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B = \begin{bmatrix} B_{11} & B_{12} & B_{13} & B_{14} & B_{15} \\ B_{21} & B_{22} & B_{23} & B_{24} & B_{25} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ B_{51} & B_{52} & B_{53} & B_{54} & B_{55} \end{bmatrix}, \quad \text{with } B_{ji} = \begin{bmatrix} b_{ji} & 0 \\ 0 & b_{ji} \end{bmatrix}, i, j = 1,2,\ldots,5,$$

$$F_R = -\frac{F}{A} \cdot \begin{bmatrix} b_{10} & 0 & \cdots & b_{50} & 0 \\ 0 & b_{10} & \cdots & 0 & b_{50} \end{bmatrix}^T \cdot \begin{bmatrix} S_{in} \\ X_{d,in} \end{bmatrix} = -\frac{F}{A} \cdot [b_{10} \;\; 0 \;\; \cdots \;\; b_{50} \;\; 0]^T \cdot S_{in}\,.$$

Numerous simulation experiments have been performed on the example presented above to illustrate the dynamical behaviour of the two classes of models (exactly and reduced). The values of bioreactor dimensions and of process parameters used in simulation are (Petre & Selişteanu, 2007a; Petre et al., 2008): L = 1 m, A = 0.02 m^2, k_1 = 0.4, k_d = 0.05 h^{-1}. For the specific growth rate μ we have chosen a Contois model:

$$\mu(S, X) = \mu_{\max} \frac{S}{K_C X + S} \tag{21}$$

with $\mu_{\max}$ = 0.35 h^{-1} and K_C = 0.4.

In fact, in simulations we compared the reduced order model obtained by orthogonal collocation method with another approximation of the original model (5) obtained by a finite difference method, where the derivatives of variable ξ with respect to time and space are approximated by finite differences as:

$$\left.\frac{\partial \xi}{\partial t}\right|_{t=k\Delta t} \cong \frac{\xi_{k+1} - \xi_k}{\Delta t}, \quad \left.\frac{\partial \xi}{\partial z}\right|_{z=j\Delta z} \cong \frac{\xi_{j+1} - \xi_j}{\Delta z}, \tag{22}$$

where Δt and Δz are discretization intervals. The choice of the discretization intervals is a very important problem (Bouaziz & Dochain, 1993). In order to have a satisfactory accuracy,

a high number of discretization points may be necessary, but this choice requires excessive computer time.

The graphics in Fig. 2 show the response of the bioprocess to a step of the influent substrate concentration S_{in} from 7.5 to 10 g/l at time t = 5 s, for three values of space discretization points (**1**: M = 100, **2**: M = 200, **3**: M = 400).

To choose the best approximation, for the simulation of the reduced model (16) different solutions have been tested. So, as orthogonal basis functions have been chosen Lagrange polynomials (14), which are assumed to be reliable and easy to compute. The collocation points have been chosen as the zeros of the Jacobi polynomials and of the Legendre polynomials. The Jacobi polynomials can be computed by using the following recursive expression (Corduneanu, 1981; Dochain et al., 1992):

$$P_n^{(\alpha,\beta)}(z)=\left[z-g_n(\alpha,\beta,n)\right]P_{n-1}^{(\alpha,\beta)}(z)-h_n(\alpha,\beta,n)P_{n-2}^{(\alpha,\beta)}(z),\quad n=1,2,\ldots,p, \tag{23}$$

with:

$$\alpha,\beta>-1,\quad P_{-1}(z)=0,\quad P_0(z)=1,\quad h_1=0,$$

$$h_n=\frac{(n-1)(n+\alpha-1)(n+\beta-1)(n+\alpha+\beta-1)}{(2n+\alpha+\beta-1)(2n+\alpha+\beta-2)^2(2n+\alpha+\beta-3)},\quad \text{for n} > 1,$$

$$g_n=\frac{1}{2}\left[1-\frac{\alpha^2-\beta^2}{(2n+\alpha+\beta-1)^2-1}\right],\quad \text{for n} > 0.$$

The Legendre polynomials can be computed by using the following recursive expression (Corduneanu, 1981):

$$(n+1)P_{n+1}(z)=(2n+1)zP_n(z)-nP_{n-1}(z),\quad P_0(z)=1,\quad P_1(z)=z,\quad n=1,2,\ldots,2p. \tag{24}$$

Different values of the parameters α and β of the Jacobi polynomials and different numbers of interior collocation points have been considered. We found out that the best choice of α and β is $\alpha=0$ and $\beta=4$. We also found out that a number p = 4 of interior collocation points are sufficient for correctly simulating the process.

The simulation results performed in the same conditions as in the first experiment are presented in Fig. 3. These graphics show the behaviour of reduced order process model for four different sets of values of α and β of the Jacobi polynomials of order 4 as follows:

1: $\alpha=0,\beta=0$; **2**: $\alpha=0,\beta=2$; **3**: $\alpha=0,\beta=3$; **4**: $\alpha=0,\beta=4$.

Thus, we obtain a set of four Jacobi polynomials, whose zeros are given by the following four sets of values:

1: $z_1=0.0694,\quad z_2=0.3300,\quad z_3=0.6699,\quad z_4=0.9306;$

2: $z_1=0.2019,\quad z_2=0.4755,\quad z_3=0.7488,\quad z_4=0.9414;$

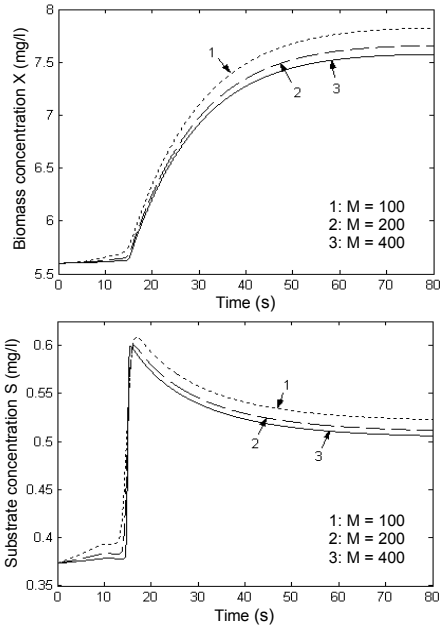


Fig. 2. The response of the original model

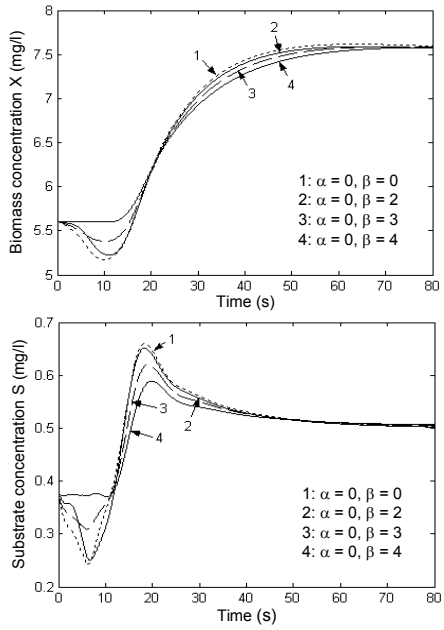


Fig. 3. The behaviour of the reduced order model

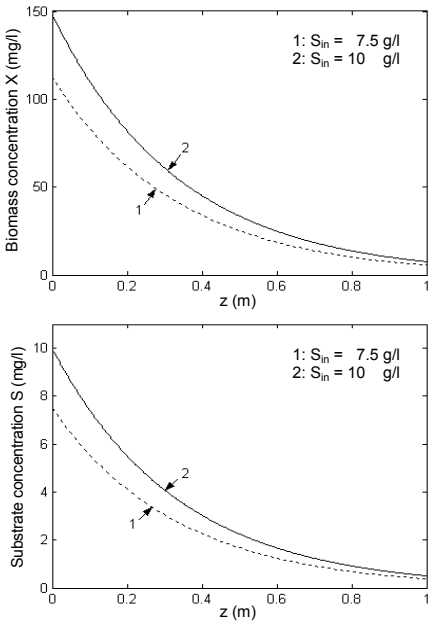


Fig. 4. Profiles of steady state regime of biomass X and substrate S

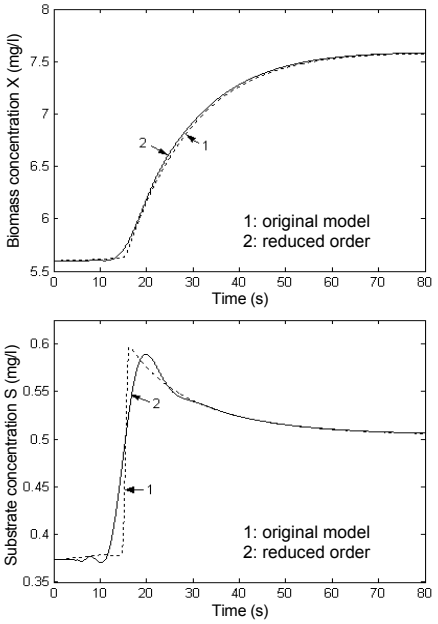


Fig. 5. The behaviour of the two classes of models (original and reduced order model)

$$\mathbf{3}:\ z_1 = 0.2597,\quad z_2 = 0.5307,\quad z_3 = 0.7815,\quad z_4 = 0.9497;$$

$$\mathbf{4}:\ z_1 = 0.3121,\quad z_2 = 0.5789,\quad z_3 = 0.8130,\quad z_4 = 0.9627.$$

The initial simulation conditions have been chosen so as to correspond to a steady state obtained from:

$$\mu X - k_d X = 0,\quad -\frac{F}{A}\cdot\frac{\partial S}{\partial z} - k_1 \mu X = 0,\quad -\frac{F}{A}\cdot\frac{\partial X_d}{\partial z} + k_d X = 0, \tag{25}$$

with S(t, z = 0) = S_{in}(t), X_d(t, z = 0) = 0. From the first two equations in (25), the steady state regime for biomass X and substrate S is shown in Fig. 4.
Note that in all the simulations, the influent flow rate is constant, F = 2 l/s.
From Fig. 2 and 3 one can observe that the reduced order model performed by orthogonal collocation method using only p = 4 interior collocation positions obtained as solutions of the Jacobi polynomial with $\alpha = 0$ and $\beta = 4$ constitutes the best approximate of initial model given by our simulations. This can be better observed from graphics in Fig. 5 where the comparative behaviour of the two classes of models is presented.

5. Adaptive control of the propagation bioprocesses

In this section, the control problem of a class of propagation bioprocesses that are carried out in fixed bed reactors without dispersion is presented. The nonlinear adaptive controllers are designed based on the finite order model (16) obtained from exactly model (11) by using the orthogonal collocation method. It can be see that the model (16) may be rewritten as (Bastin & Dochain, 1990; Petre, 2008):

$$\dot{\zeta}(t) = K\, r(\zeta) - D\zeta + F_v \tag{26}$$

where $\zeta = [x_1^T\ x_2^T]^T$ is the state vector, $K = [K_1^T\ K_2^T]^T$ is the yield coefficient matrix, $r(\zeta) = r(x_1, x_2)$ is the reaction rate vector, $F_v = [0\ F_R^T]^T$ is the influent flow rate vector, and $D = \begin{bmatrix} 0 & 0 \\ 0 & (F/A)B \end{bmatrix}$ is the dilution matrix.

5.1 Problem statement

For the bioreactors described by the model (16) the control objective is to regulate the concentration of a single component at the bioreactor output, under the following conditions:

i. The control input is the influent flow rate F.
ii. The controlled variable is measured not only at the bioreactor output, but also at every interior collocation point and at the reactor input (only in the case of external substrate).
iii. The yield coefficients are positive constants (some of them being unknown).
iv. $m_1 \le m$ reaction rates are unknown.

For simplicity, we will denote by y the concentration of the controlled component, by y_i the value of y at each interior collocation point $z = z_i$, $i = 1,\ldots,p$, i.e. $y_i(t) = y(t, z = z_i)$, and by

y_{p+1} the value of the controlled component at the output of bioreactor $y_{p+1}(t) = y(t, z = z_{p+1})$. Using these notations, y_{p+1} may by expressed as a linear combination of state variables x_1 and x_2 as:

$$y_{p+1} = c^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{27}$$

where $c^T = [c_1^T \; c_2^T]$ is a vector with appropriately dimension, used to select the controlled variable.
Using (16), the dynamics of y_{p+1} in (27) is given by:

$$\dot{y}_{p+1}(t) = c_1^T K_1 r(\cdot) + c_2^T K_2 r(\cdot) - (F / A) \cdot c_2^T B x_2 + c_2^T F_R . \tag{28}$$

Consider that for the bioprocess described by the model (20), the controlled variable is the substrate concentration at the output of the bioreactor, that is $y_{p+1} = S_5$. Since the state vector ζ is now given by

$$\zeta = [x_1^T \; x_2^T]^T = [X_1 \, X_2 \ldots X_5 \,|\, S_1 \, X_{d1} \ldots S_4 \, X_{d4} \, S_5 \, X_{d5}]^T ,$$

then the entries of the vector c in (27) will be: $c^T = [0\;0\ldots0\,|\,0\;0\ldots0\;0\;1\;0]$ with $c_1^T = [0\;0\;0\;0\;0]$, $c_2^T = [0\;0\,|\,0\;0\,|\,0\;0\,|\,0\;0\,|\,1\;0]$ or $c_2^T = [c_{21}\, c_{22}\, c_{23}\, c_{24}\, c_{25}]$.
The dynamics of the concentration S_5 is given by:

$$\dot{S}_5 = -(F / A) \cdot \sum_{i=1}^{5} b_{5i} S_i - (F / A) \cdot b_{50} S_0 - k_1 \mu_5 (X_5, S_5) X_5 . \tag{29}$$

Using (29), the dynamics of y_{p+1} in (28) can be written as:

$$\dot{y}_{p+1}(t) = -\frac{F}{A} c_2^T B x_2 - \frac{F}{A} \tilde{c}_2^T \tilde{b}_{p+1} \xi_{2,in} + \tilde{c}_2^T \tilde{K}_2 \tilde{r}(\xi_{1,p+1}; \xi_{2,p+1}) , \tag{30}$$

where:

$$\tilde{c}_2^T = c_{25} = [1\;0], \quad \xi_{2,in} = [S_{in} \; X_{d,in}]^T = [S_{in} \; 0]^T . \tag{31}$$

It is easy to verify that the term Bx_2 in (30) is a linear combination only of variables y_i at the interior collocation points $z_i, i = 1, \ldots, p$. The term $\xi_{2,in}$ contains the influent concentrations at the input of the bioreactor. With the condition (iv), the last term in (30) can be rewritten as (Dochain et al., 1992):

$$\begin{aligned} &\tilde{c}_2^T \tilde{K}_2 \tilde{r}(\xi_{1,p+1}, \xi_{2,p+2}) = K_{21} \tilde{r}_1 + K_{22} \tilde{r}_2^T \\ &= K_{21} \tilde{r}_1 + [\theta_2 \; \theta_3 \ldots \theta_{m-m_1+1}][\tilde{r}_{21} \; \tilde{r}_{22} \ldots \tilde{r}_{m-m_1}]^T = \theta^T \Phi \end{aligned} \tag{32}$$

where $\tilde{r}_1$ and $\tilde{r}_2$ contain the unknown and known reaction rates respectively, and θ and Φ are given by:

$$\theta^T = [K_{21}\tilde{r}_1 \;\; \theta_2 \;\; \theta_3 \ldots \theta_{m-m_1+1}], \quad \Phi^T = [1 \;\; \tilde{r}_{21} \;\; \tilde{r}_{22} \ldots \tilde{r}_{m-m_1}]. \tag{33}$$

As a conclusion, θ contains all the unknown parameters and Φ contains the known reaction rates. Then, the dynamics of output y_{p+1} takes the form:

$$\dot{y}_{p+1}(t) = -(F/A)c_2^T B x_2 - (F/A)\tilde{c}_2^T \tilde{b}_{p+1}\xi_{2,in} + \theta^T \Phi \,. \tag{34}$$

5.2 Exactly linearizing controller

As it was mentioned above, the control objective is to regulate the concentration of variable y_{p+1} at the output of the bioreactor at a desired value y_{p+1}^* by acting on the feeding substrate flow rate F.

Controller design is achieved by using the input-output linearizing technique. Remember that the input-output linearizing principle (Isidori, 1995) consists in the calculus of a nonlinear control law such that the behaviour of closed loop system (controller + process) is the same as the behaviour of a linear stable system. Assume that for the closed loop system we wish to have the following first-order linear stable dynamics:

$$\frac{d}{dt}(y_{p+1}^* - y_{p+1}) + \lambda_1(y_{p+1}^* - y_{p+1}) = 0, \quad \lambda_1 > 0 \,. \tag{35}$$

Firstly, we consider the ideal case, where maximum prior knowledge concerning the process is available. In particular we suppose that the parameters θ in (34) are known and all the state variables are available for on-line measurements. It can be seen that equation (34) has the relative degree equal to 1 (Isidori, 1995). Then, from (34) and (35), the above closed-loop dynamics will be achieved by implementing the following exactly linearizing nonlinear control law:

$$F = A\frac{\dot{y}_{p+1}^* + \lambda_1(y_{p+1}^* - y_{p+1}) - \theta^T \Phi}{-\tilde{c}_2^T \tilde{b}_{p+1}\xi_{2,in} - c_2^T B x_2} \tag{36}$$

The control law (36) leads to the following linear error model:

$$\dot{e}(t) = -\lambda_1 e(t) \tag{37}$$

with $e(t) = y_{p+1}^*(t) - y_{p+1}(t)$, which for $\lambda_1 > 0$ has an asymptotic stable point at $e = 0$. But the use of control law (36) requires the complete knowledge of the process. It is well known that because of the reaction rates, the bioprocesses are characterized by highly nonlinear dynamics and furthermore the kinetic and process parameters are often partially or completely unknown; as a consequence an accurate model for these processes is difficult to develop. Therefore, in recent years, it has been noticed a great progress in adaptive and robust adaptive control of bioprocesses, due to their ability to compensate for parametric uncertainties (Bastin & Dochain, 1990; Petre & Selişteanu, 2005; Petre, 2008; Selişteanu & Petre, 2006). Consequently, in the following section, for the class of the presented fixed bed bioreactors, we will develop an adaptive control algorithm considering some real realistic conditions.

5.3 Adaptive control of fixed bed bioreactors

If the parameters θ in (36) are assumed unknown (see the conditions (iii) and (iv)), these will be replaced by their on-line estimates $\hat{\theta}$. Then the control law (36) becomes an adaptive control law given by:

$$F_a = A\frac{\dot{y}^*_{p+1} + \lambda_1(y^*_{p+1} - y_{p+1}) - \hat{\theta}^T\Phi}{-\tilde{c}_2^T\tilde{b}_{p+1}\xi_{2,in} - c_2^T B x_2}. \tag{38}$$

The estimates $\hat{\theta}$ can be on-line calculated by using, for example, a linear regressive parameter estimator (Bastin & Dochain, 1990; Petre & Selişteanu, 2005; Petre, 2008), described here by the following equations:

$$\begin{aligned}
&\dot{\Psi} = -\omega\Psi + \Phi \\
&\dot{\Psi}_0 = -\omega\Psi_0 - \omega y_{p+1} - \frac{F}{A}\left(c_2^T B x_2 + \tilde{c}_2^T\tilde{b}_{p+1}\xi_{2,in}\right) \\
&\dot{\hat{\theta}} = \Gamma\Psi(y_{p+1} - \Psi_0 - \Psi^T\hat{\theta}) \\
&\dot{\Gamma} = -\Gamma\Psi\Psi^T\Gamma + \lambda\Gamma, \quad \Gamma(0) > 0,\ 0 < \lambda \le 1
\end{aligned} \tag{39}$$

where Φ stands for regressor matrix, Γ is a positive and symmetric gain matrix, and $\lambda \in (0,1]$, named forgetting coefficient, and $\omega > 0$ are design parameters to control the stability and convergence properties of the estimator (Petre & Selişteanu, 2005; Selişteanu & Petre, 2006).

6. Simulation results

The performances of the designed nonlinear adaptive controllers were verified by several simulation experiments performed upon the fixed bed bioreactor described by the model (5). The values of bioreactor dimensions and process parameters used in simulation are the same as in Section 4 (Petre & Selişteanu, 2007a; Petre et al., 2008). Also, for the specific growth rate μ we have chosen a Contois model (21), with $\mu_{\max} = 0.35\ \text{h}^{-1}$ and $K_C = 0.4$.
The interior collocation points of the reduced model (20) have been chosen as zeros of the Jacobi polynomials given in Section 4. For $p = 4$, $\alpha = 0$ and $\beta = 4$, the abscises of the four interior collocation points are: $z_1 = 0.3121$, $z_2 = 0.5789$, $z_3 = 0.8130$, $z_4 = 0.9627$. Of course, these values will determine the values of the entries b_{ji} in the matrices B_{ji} and b_j.
The control objective is to regulate the substrate concentration S_5 at the output of the bioreactor, i.e. $y_{p+1} = S_5$. From (19) and (20) the dynamics of S_5 is obtained as:

$$\dot{S}_5 = -\frac{F}{A}\sum_{i=1}^{5} b_{5i}S_i - \frac{F}{A}b_{50}S_0 - k_1\mu_5(X_5, S_5)X_5. \tag{40}$$

The exactly linearizing control law (36) takes the form:

$$F = A\frac{\dot{S}^*_5 + \lambda_1(S^*_5 - S_5) + k_1\mu_5(\cdot)X_5}{-b_{50}S_{in} - \sum_{i=1}^{5} b_{5i}S_i}. \tag{41}$$

The behaviour of the closed loop system in the ideal case, when all the parameters are completely known, is presented in Fig. 6.

The initial simulation conditions correspond to a process steady state regime. So, for the interior collocation points, the used values are: $X_1(0)$ = 44.1051 mg/l, $X_2(0)$ = 19.8101 mg/l, $X_3(0)$ = 9.8169 mg/l, $X_4(0)$ = 6.2634 mg/l; $X_5(0)$ = 5.6010 mg/l; $S_1(0)$ = 2.9403 g/l, $S_2(0)$ = 1.3207 g/l, $S_3(0)$ = 0.6545 g/l, $S_4(0)$ = 0.4176 g/l, $S_5(0)$ = 0.3734 g/l, and $X_0(0)$ = 0 mg/l, $S_0(0)$ = $S_{in}(0)$ = 7.5 g/l.

To verify the regulation properties of the controller, for the reference variable a piece-wise constant variation was considered as:

$$S^* = S_5^* = \begin{cases} 0.35\, g/l, & 0 \le t < 80s \\ 0.30\, g/l, & 80 \le t < 175s \\ 0.25\, g/l, & 175 \le t < 215s \\ 0.30\, g/l, & 215 \le t < 250s \end{cases} \tag{42}$$

The value of the gain parameter λ_1 in (41) is $\lambda_1 = 1$. The system evolves in open loop from the time $t = 0$ to time $t_1 = 10\,s$, after which the system is closed by using the control law (41). The influent substrate concentration S_{in} acts as a perturbation given by $S_{in}(t) = S_{in0} \cdot (1 + 0.2 \cdot \sin(\pi t / 25) - 0.05 \cdot \cos(\pi t / 5))$ with $S_{in0} = 7.5$ g/l for $0 \le t < 125$ s and $S_{in0} = 15$ g/l for $t > 125$ s.

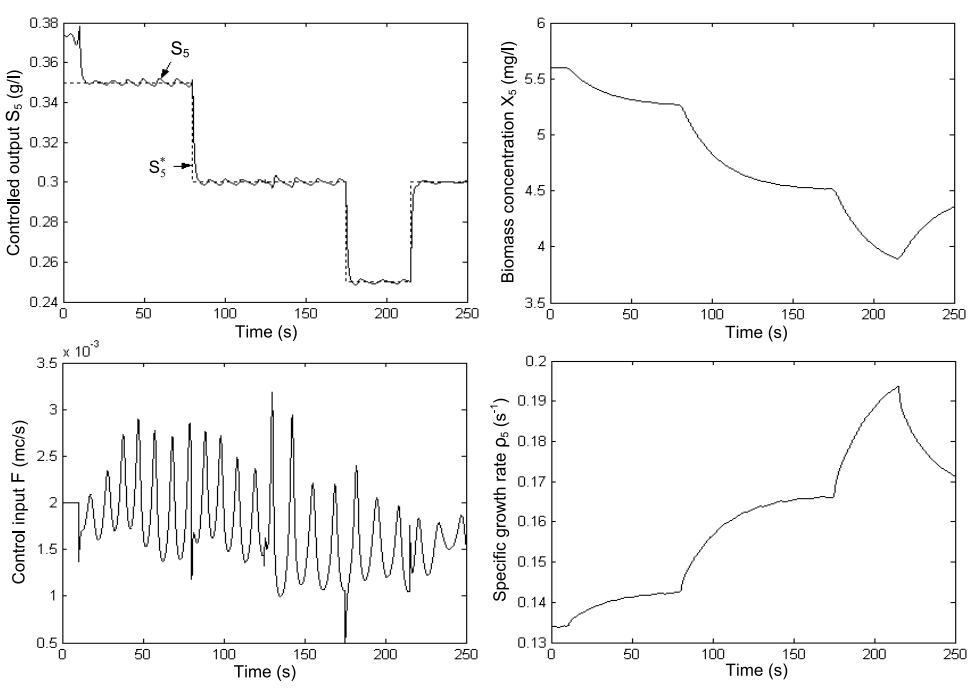


Fig. 6. The behaviour of the closed loop system with the exactly linearizing controller

From Fig. 6 one can observe that the controller (41) is efficiently both in regulation of controlled variable and in rejection of the perturbation S_{in}.
Assume now that the death parameter k_d is known, and the specific growth rate μ is unknown. Assume also that $\mu_5(\cdot)$ in (41) can be rewritten as:

$$\mu_5(\cdot) = \rho_5 S_5 , \tag{43}$$

where ρ_5 is considered as an unknown positive parameter. It is clear that if $\mu_5(\cdot)$ should be known, then ρ_5 is a function of bioreactor state given by:

$$\rho_5(X_5, S_5) = \mu_{\max} \frac{1}{K_C X_5 + S_5} . \tag{44}$$

Assume also that at the output of the bioreactor the only measured variable is the substrate concentration S_5. It can be seen that the practical implementation of the control law (41) requires the knowledge of the state X_5, and of the specific reaction rate μ_5.
Since in (41) the variable X_5 is not directly measurable, this will be substituted by its estimate $\hat{X}_5$. For the estimation of X_5, independent of the unknown specific reaction rate $\mu_5(\cdot)$, we use an asymptotic state observer (Petre & Selişteanu, 2005), which can be derived as follows. Let us define the auxiliary state z as:

$$z = S_5 + k_1 X_5 . \tag{45}$$

The dynamic of z deduced from model (20) is expressed by the following linear stable equation:

$$\dot{z} = \frac{F}{A}\left(-b_{50} S_{in} - \sum_{i=1}^{5} b_{5i} S_i \right) - k_d (z - S_5) . \tag{46}$$

From (45) and (46), the estimated value $\hat{X}_5$ of the state X_5 is given by $\hat{X}_5 = (1/k_1)\cdot(z - S_5)$.
Using the definitions of ρ_5 in (43) and the definitions of z in (45), the dynamics of output S_5 in (40) takes the form:

$$\dot{S}_5 = \frac{F}{A}\left(-b_{50} S_{in} - \sum_{i=1}^{5} b_{5i} S_i \right) - \rho_5 S_5 (z - S_5) . \tag{47}$$

Then, the adaptive version of the control law (41) is given by:

$$F_a = A \frac{\dot{S}_5^* + \lambda_1 (S_5^* - S_5) + \hat{\rho}_5 S_5 (\hat{z} - S_5)}{-b_{50} S_{in} - \sum_{i=1}^{5} b_{5i} S_i} , \tag{48}$$

where the estimates $\hat{\rho}_5$ of ρ_5 are on-line calculated by using the regressive parameter estimator (37), where $\theta = \rho_5$, and $\hat{z}$ is calculated by using (46) rewritten as:

$$\dot{\hat{z}} = \frac{F}{A}\left(-b_{50} S_{in} - \sum_{i=1}^{5} b_{5i} S_i \right) - k_d (\hat{z} - S_5) . \tag{49}$$

The adaptive algorithm given by (48), (39) and (49) was implemented under the same conditions as in the first case. The values of the controller design parameters used in simulations are: $\lambda_1 = 1$, $\lambda = 0.6$, $\omega = 10$, and the initial conditions are: $\hat{\theta}(0) = \hat{\rho}_5(0)$ = 0.12, $\Gamma(0) = 0.1$, $\hat{z}(0)$ = 2.75 g/l. Much more, in order to test the behaviour of the adaptive controlled system in more realistic circumstances, we considered that the measurements of controlled variable in all interior collocation points and also at the input and the output of the bioreactor are corrupted with an additive white noise with zero average (5% from their nominal values).

The simulation results are shown in Fig. 7. As in the first case, the system evolves in open loop starting from $t = 0$ to time $t_1 =$ 10 s, after that the system is closed by using the above adaptive algorithm. The perturbation S_{in} has the same evolution as in the ideal case.

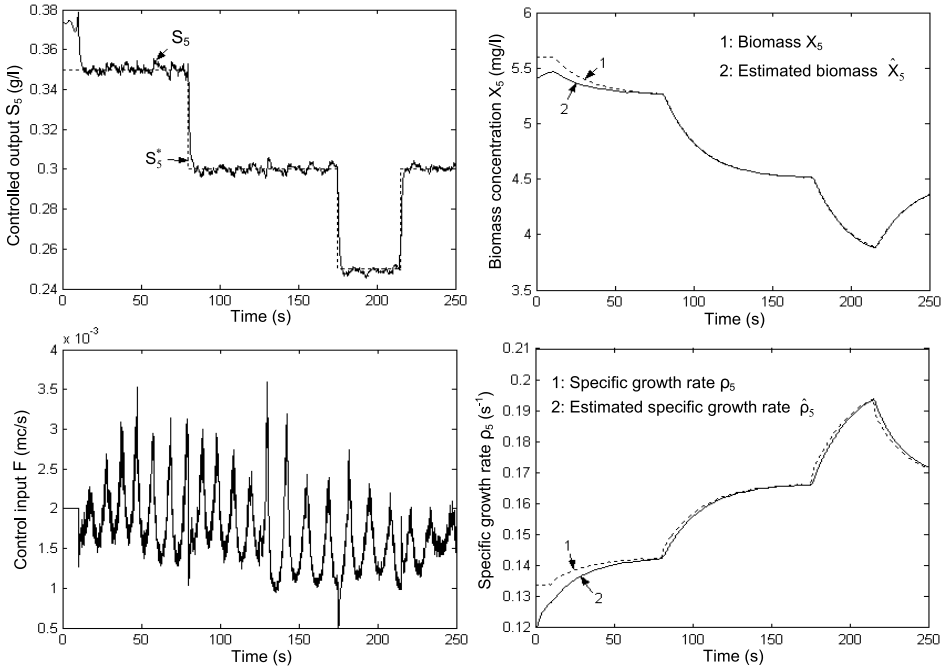


Fig. 7. The behaviour of the closed loop system with the adaptive controller

From the graphics in Fig. 7 one can deduce that even if the initialization of $\hat{z}$ and $\hat{\rho}_5$ are different from their ideal values (given by $\hat{z}(0) = S_5(0) + k_1 X_5(0)$ and $\hat{\rho}_5(0) = \mu_{\max} / (K_M X_5(0) + S_5(0))$, the adaptive controller is efficiently both in regulation of controlled variable and in rejection of the perturbation S_{in} despite the very high load variations of S_{in}. The behaviour of controlled variables and of control inputs is comparable with the results obtained in the free noise simulation. One can observe also a good behaviour both of the proposed state observer (49), (45) and parameter estimator (39).

Moreover, it was proven that the adaptive algorithm given by (48), (39) and (49) is robust, that is even though the process model (5) has uncertainty parameters, the behaviour of closed loop system is good. It was verified that if the death coefficient k_d suffers variations by comparison to its nominal value (e.g. $k_d = 0.04 \div 0.06$ h^{-1}) the obtained results are still good.

7. Conclusion

The approximation of the infinitely order dynamical model for a class of nonlinear propagation bioprocesses described by partial differential equations was examined. These approximate models consist of a set of ordinary differential equations obtained by orthogonal collocation method. The results obtained by application of this method in the case of a fixed bed reactor showed that by an appropriately choosing of the collocation points along the reactor, the behaviour of the reduced order model is very close to the behaviour of original infinitely order model.
After that, the obtained reduced order model was used to design some control algorithms for these types of reactors. The controller design is based on the input-output linearization technique. The obtained algorithm was tested in the controlling problem of substrate concentration for a propagation bioprocess that is carried out in a fixed bed reactor.
The simulation obtained results demonstrated that the designed adaptive algorithms used in control of propagation bioreactors yield good results closely comparable to those obtained in the case when the process parameters are completely known and/or time invariable.
Moreover, these algorithms prove to be robust as well yielding good results even though the measurable variables are affected by noises and/or the model parameters suffer variations between wide limits. It must be also noted that these algorithms can relatively easily be extended to other types of distributed parameters bioreactors: fluidized bed, air lift reactors.

8. Acknowledgment

This work was supported by CNCSIS–UEFISCDI, Romania, project number PNII–IDEI 548/2008.

9. References

Aksikas, I.; Winkin J. & Dochain D. (2007). Optimal LQ-feedback regulation of a nonisothermal plug flow reactor model by spectral factorization. *IEEE Trans. Autom. Control*, Vol. 52, No. 7, pp. 1179–1193, ISSN 0018-9286

Americano da Costa Filho, M.V.; Barbosa Monteiro Julieta; Magazoni, F.C. & Colle, S. (2009). Modeling, simulation and analysis of ethanol fermentation process with control structure in industrial scale. In: *Proceedings of ECOS 2009 - 22nd Int. Conf. on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems*, Aug 31 – Sept 3, 2009, Foz do Iguaçu, Paraná, Brazil

Bastin, G. & Dochain, D. (1990). *On-line Estimation and Adaptive Control of Bioreactors*, Elsevier, ISBN 978-0-444-88430-5, Amsterdam, Netherlands

Bastin, G. (1991). Nonlinear and adaptive control in biotechnology: a tutorial. *European Control Conference ECC'91*, pp. 2001–2012, Grenoble, France, July 1991

Bouaziz, B. & Dochain, D. (1993). Control analysis of fixed bed reactors: a singular perturbation approach. *European Control Conference ECC'93*, Vol. 3, pp. 1741–1746, Groningen, Netherlands, June 28 – July 1, 1993

Christofides, P.D. (2001). *Nonlinear and robust control of PDE systems: method and applications to transport-reaction processes*, Birkhauser, ISBN 0-8176-4156-4, Boston

Corduneanu, A. (1981). *Differential equations with application in electrotechnics* (in Romanian), Ed. Facla, Timisoara, Romania

Dochain, D.; Babary, J.P. & Tali-Maamar, N. (1992). Modelling and adaptive control of nonlinear distributed parameter bioreactors via orthogonal collocation. *Automatica*, Vol. 28, No. 5, pp. 873–883, ISSN 0005-1098

Dochain, D. & Vanrolleghem, P. (2001). *Dynamical Modelling and Estimation in Wastewater Treatment Processes*, IWA Publishing, ISBN 1900222507

Hoo, K. & Zheng, D. (2001). Low-order control-relevant models for a class of distributed parameter systems. *Chemical Engineering Science*, Vol. 56, No. 23, pp. 6683-6710, ISSN 0009-2509

Isidori, A. (1995). *Nonlinear Control Systems* (3rd Edition), ISBN 3540199160, Springer-Verlag, New York

Khalil, H. K. (2002). *Nonlinear Systems* (3rd Edition), Prentice Hall, ISBN 0-13-067389-7 Upper Saddle River, New Jersey

Petre, E. (2003). Adaptive Control Strategies for a Class of Time Delay Nonlinear Bioprocesses. *Rev. Roum. Sci. Techn.- Électrotechn. et Énerg.*, Vol. 48, No. 4, pp. 567-582, ISSN 0035-4066, Bucarest, Romania

Petre, E. & Selişteanu, D. (2005). *Modelling and Identification of Depollution Bioprocesses* (in Romanian), Universitaria, ISBN 973-742-232-5, Craiova, Romania

Petre, E.; Popescu, D. & Selişteanu, D. (2007). Finite dimensional models of nonlinear distributed parameter bioreactors via orthogonal collocation. In: *Proceedings of the 8th Int. Carpathian Control Conference*, pp. 548-551, ISBN 978-80-8073-805-1, Štrbské Pleso, High Tatras, Slovak Republic, May 2007

Petre, E. & Selişteanu, D. (2007a). Approximation of the dynamical model for a class of nonlinear propagation bioprocesses. *Rev. Roum. Sci. Techn.- Électrotechn. et Énerg.*, Vol. 52, No. 3, pp. 371-381, ISSN 0035-4066, Bucarest, Romania

Petre, E. & Selişteanu, D. (2007b). Some aspects concerning the modelling of chlorine residuals in drinking water distribution networks. *CEAI – Control Engineering and Applied Informatics*, Vol. 9, No. 2, pp. 48-58, ISSN 1454-8658, Bucharest, Romania

Petre, E.; Popescu, D. & Selişteanu, D. (2008). Adaptive control strategies for a class of nonlinear propagation bioprocesses. *Acta Montanistica Slovaca*, Vol. 13, No. 1, pp. 118-126, ISSN 0018-9286

Petre, E. (2008). *Nonlinear Control Systems - Applications in Biotechnology* (in Romanian, 2nd Edition), Universitaria, ISBN 973-8043-140-0, Craiova, Romania

Selişteanu, D. & Petre, E. (2004). On Adaptive Control of a Wastewater Biodegradation Bioprocess. *CEAI-Control Engineering and Applied Informatics*, Vol. 6, No. 3, pp. 48-56, ISSN 1454-8658, Bucharest, Romania

Selişteanu, D. & Petre, E. (2006). *Control Methods Control of Depollution Bioprocesses* (in Romanian), Universitaria, ISBN 978-973-742-543-0, Craiova, Romania

Shi, D.; El-Farra, N.H.; Li, M.H.; Mhaskar, P. & Christofides, P.D. (2006). Predictive control of particle size distribution in particulate processes. *Chemical Engineering Science*, Vol. 61, No. 1, pp. 268-281, ISSN 0009-2509

Shvarstman, S.Y.; Theodoropoulos, C.; Rico-Martinez, R.; Kevrekidis, I.G.; Titi, E.S. & Mountziaris, T.J. (2000). Order reduction for nonlinear dynamic models of distributed reacting systems. *Journal of Process Control*, Vol. 10, pp. 177-184, ISSN 0959-1524

Slotine, J.-J.E. & Li, W. (1991). *Applied Nonlinear Control*, Prentice-Hall, ISBN 0-13-040890-5, Englewood Cliffs, New Jersey

Vilas, C. (2008). Modelling, Simulation and Robust Control of Distributed Processes: Application to Chemical and Biological Systems. PhD thesis, University of Vigo, Vigo, Spain

11

Meshfree Methods

Saeid Zahiri
Shiraz University,
Islamic Republic Of Iran

1. Introduction

In this chapter we will describe about numerical simulation with meshfree methods. We know; phenomena in nature, whether physical, geological, mechanical, electrical, or biological, can often be describe by means of algebraic, differential, or integral equations. Obtaining exact solutions for these equations is ideal. Unfortunately, we can only obtain exact ones for limited practical problems because most of these problems are complex. Therefore using numerical procedure to obtain approximate solutions is inevitable. One of the most important tools in the field of numerical methods that has been developed newly is meshfree or meshless methods.

A meshfree method is a method used to establish system algebraic equations for the whole domain of problem without using a predefined mesh for the domain discretization. This infant method uses a set of scattered nodes, called field nodes, to establish the problem domain and boundaries, which do not require any priori information on the relationship between the nodes for the interpolation or approximation of the unknown functions of field variables. In the FEM, a continuum with a complicated shape is divided into elements, *finite elements*. The individual elements are connected together by a topological map called a *mesh*.

Meshfree methods have been proposed and achieved remarkable progress over the past few years. According to the formulation procedure, meshfree methods fall into three categories: meshfree weak form methods (like: EFG, MLPG, LRPIM,…), meshfree strong form methods (like: SPH, Collocation method,…) and meshfree weak-strong form methods based on the combination of both weak form and strong form (like: MWS method). These three categories and their limitations, applications, advantages and other descriptions will be introduced. In seeking for an approximate solution to the problem governed by PDEs and boundary conditions, one first needs to approximate the unknown field function using shape (trial or base) functions before any formulation procedure can be applied to establish the discretized system equations. In this chapter definition of base and shape functions and various techniques for meshfree shape function constructions are discussed. These shape functions are locally supported, because only a set of field nodes in a small local domain are used in the construction and the shape function is not used or regarded as zero outside the local domain. Such a local domain is termed the support domain or influence domain. The concept and kinds of support domain and determination of the dimension of the support domain will be described.

After introducing the concept of support domain, the point interpolation method (PIM) in detail will be discussed. Point interpolation method is one of the series representation

methods for the function approximation, and useful for creating meshfree shape functions. A scalar function defined in the problem domain that is represented by a set of scattered nodes will be shown. There are two types of PIM shape functions have been developed so far using different forms of basis functions Polynomial basis functions and radial basis functions (RBF) have often been used in meshfree methods. These two types of PIMs will be discussed in the following chapter.
For satisfying the boundary conditions, penalty method, direct method, lagrange multiplier method and direct interpolation method can be used to enforce essential boundary conditions. One of these methods due to using meshfree method can be elected and they will be explained and compared.
To simulate some problems, the partial differential equations and boundary conditions for two dimensional solid mechanics and fluid mechanics problem and heat transfer problem especially thermodynamics of plates and shells will be given in sub-sections. These problems are solved with meshfree methods.

2. Meshfree methods categories

According to the formulation procedure, meshfree methods fall into three categories: meshfree weak form methods (like: EFG, MLPG, LRPIM,…), meshfree strong form methods (like: Collocation method, SPH,…) and meshfree weak-strong form methods based on the combinations of both weak forms and strong forms (like: MWS method).

2.1 Strong form methods

Regarding to formulate the governing equations, the direct approximate solution from the differential equations is used. It means the strong form of governing equations for boundary conditions are directly discretized at the field nodes to obtain a set of discretized system equations. If Taylor series is used and the differentiations are replaced, the method is the strong form method. The strong form method does not need the numerical integration. Thus The background mesh even locally is not needed for the strong form methods.
Meshfree strong form methods have some attractive advantages: a simple algorithm, computational efficiency, and truly meshfree. However, Meshfree strong form methods are often unstable, not robust, and inaccurate, especially for problems with derivative boundary conditions. Several strategies may be used to impose the derivative (Neumann) boundary conditions in the strong form methods, such as the use of fictitious nodes, the use of the Hermite-type Meshfree shape functions, the use of a regular grid on the derivative boundary.

2.2 Weak form methods

In Meshfree weak form methods, the governing partial differential equations (PDEs) with derivative boundary conditions are first transformed to a set of so called weak form integral equations using different techniques. The weak forms are then used to derive a set of algebraic system equations through a numerical integration process using sets of background cells that may be constructed globally or locally in the problem domain. Meshfree weak form methods were relatively under developed before 1990, but there has been a substantial increase in research effort since then.
There are now many different versions of Meshfree weak form methods. Meshfree weak form methods based on the global weak forms are called Meshfree global weak form methods, and

those based on local weak forms are called Meshfree local weak form methods. Meshfree global weak form methods are based on the global Galerkin weak form for equations of problems and the Meshfree shape functions. Two typical Meshfree global weak form methods: the element free Galerkin (EFG) method (Belytschko et al., 1994a) and the radial point interpolation method (RPIM) (GR Liu and Gu, 2001c; Wang and GR Liu, 2000; 2002a).
Another typical Meshfree global weak form method is the reproducing kernel particle method (RKPM) proposed by Liu and coworkers in 1995 (Liu et al., 1995). The main idea of RKPM is to improve the SPH approximation to satisfy consistency requirements using a correction function. RKPM has been used in nonlinear and large deformation problems (Chen et al., 1996; Chen et al., 1998; Liu and Jun, 1998), inelastic structures (Chen et al., 1997), structural acoustics (Uras et al., 1997), fluid dynamics (Liu and Jun et al., 1997), et cetera. Meshfree local weak form methods were developed by Atluri and coworkers based on the local Petrov-Galerkin weak form , and the Meshfree shape functions. Some other Meshfree weak form methods have also been developed, such as the hp-cloud method (Armando and Oden, 1995), the partition of unity finite element method (PUFEM) (Melenk and Babuska, 1996; Babuska and Melenk, 1997), the finite spheres method (De and Bathe, 2000), the free mesh method (Yagawa and Yamada, 1996), et cetera.

2.3 Weak-strong form methods

These Meshfree methods are called Meshfree weak-strong (MWS) form methods in this book because are based on the combination of weak and strong form methods. The MWS method was developed by GR Liu and Gu (2002d, 2003b). The key idea of the MWS method is that in establishing the discretized system equations, both the strong form and the local weak form are used for the same problem, but for different groups of nodes that carries different types of equations/conditions. The local weak form is used for all the nodes that are on or near boundaries with derivative (Neumann) boundary conditions. The strong form is used for all the other nodes. The MWS method uses least background cells for the integration, and it is currently the almost ideal Meshfree method that can provide stable and accurate solutions for mechanics problems.
There are also Meshfree methods based on the integral representation method for function approximations, such as the Smooth Particle Hydrodynamics (SPH) methods (Lucy, 1977; Gingold and Monaghan, 1977; GR Liu and Liu, 2003, etc.). In the standard SPH method, the function approximation is performed in a weak (integral) form, but strong form equations are directly discretized at the particles.

2.4 Comparisons between three meshfree categories

Each meshfree method has features with advantages and defects. With these properties, the appropriate method can be selected to solve the problem. The features of methods are presented in sub-sections.

2.4.1 Advantages and disadvantages

Convergence rate and highest accuracy are important properties in numerical methods. When the problems include Dirichlet boundary conditions, the strong form methods are the best but in case of Neumann boundary conditions, weak form methods are optimum and when the both of Dirichlet and Neumann boundaries are used in problems, the weak-strong form methods are useful.

The strong form methods are with good convergence rate and they are truly meshless. The procedure is straightforward, and the algorithms and coding are simple. They are computationally efficient, and the solution is accurate when there are only Dirichlet boundary conditions.

However, Meshfree strong form methods have disadvantages: they are often unstable and less accurate, especially for problems governed by PDEs[1] with derivative boundary conditions. Derivative boundary conditions (DBCs) involve a set of separate differential equations defined on the boundary; these are different from the governing equations defined in the problem domain. These DBCs require special treatments. Unlike integration, which is a smoothing operator, differentiation is a roughening operator; it magnifies errors in an approximation. This magnified error is partially responsible for the instability of the solution of PDEs. Hence, Meshfree strong form methods are often unstable. Special treatments are employed to implement the derivative boundary conditions in Meshfree strong form methods. However, such treatments cannot always control the error. A technique suitable for one problem may not work for another, even one of the same types. A set of parameters tuned for one problem may not work for another.

The common feature of Meshfree weak form methods is that the PDE of a problem is first replaced by or converted into an integral equation (global or local) based on a principle (weighted residual methods, energy principle). Weak form system equations can then be derived by integration by parts. A set of system equations of Meshfree weak form methods can be obtained from the discretization of the weak form using meshfree interpolation techniques. There are four features of the local weak form . The integral operation can smear the error over the integral domain and, therefore improve the accuracy in the solution. It acts like some kind of regularization to stabilize the solution. The requirement of the continuity for the trial function is reduced or weakened, due to the order reduction of the differential operation resulting from the integration by parts. The force (derivative) boundary conditions can be naturally implemented using the boundary integral term resulting from the integration by parts. The system equations in the domain and the derivative boundary conditions are conveniently combined into one single equation.

These features give Meshfree weak form methods the following advantages. They exhibit good stability and excellent accuracy for many problems. The derivative (Neumann) boundary conditions can be naturally and conveniently incorporated into the same weak form equation. No additional equations or treatments are needed and no errors are introduced in the enforcement of traction boundary conditions. A method developed properly using a weak form formulation is applicable to many other problems. A set of parameters tuned for one method for a problem can be used for a wide range of problems. This robustness of the weak form methods have been demonstrated through many practical problems. It is this robustness that makes the weak form methods applicable to many practical engineering problems.

However, Meshfree global weak form methods are meshfree only in terms of the interpolation of the field variables. Background cells have to be used to integrate a weak form over the global domain. The numerical integration makes them computationally expensive, and the background mesh for the integration means that the method is not truly meshless.

In the Meshfree local weak form methods, the local integral domain in the interior of the problem domain is usually of a regular shape. It can be as simple as possible and can be

[1]Partial Differential Problems

automatically constructed in the process of computation. The Meshfree local weak form methods have obtained satisfactory results in solid mechanics and fluid mechanics (Atluri and Shen, 2002; GR Liu, 2002).

Although the Meshfree local weak form methods made a significant step in developing ideal meshfree methods, the numerical integration is still burdensome, especially for nodes on or near boundaries with complex shape. The local integration can still be computationally expensive for some practical problems. It is therefore desirable to minimize the need for numerical integrations.

The Mesh Weak-Strong method is designed to combine the advantages of strong form and weak form methods and to avoid their shortcomings. This can be performed only after a thorough examination of the features of both types of methods, presented in the above sentences. An Meshfree weak-strong (MWS) form method was proposed recently by GR Liu and Gu (2002d); it aimed to remove the background mesh for integration as much as possible, and yet to obtain stable and accurate solutions even for PDEs with derivative boundary conditions. The MWS method has been successfully developed and used in solid mechanics (Gu and GR Liu, 2005; GR Liu and Gu, 2003b) and fluid mechanics (GR Liu and Wu et al., 2004; GR Liu and Gu et al., 2003c).

The convergence of the MWS method is studied numerically by comparison with other methods. The weak form method treats the Neumann boundary condition naturally and easily. In addition, the accuracy achieved by meshfree methods based on the weak form equations are generally much better than those based on strong form equations. However, the efficiency is a big problem for the weak form methods because of the need for weak form integration.

The MWS method proposed by Liu and Gu was based on both collocation and local radial point interpolation formulation. In the present MWS method, the strong form of meshfree collocation method is applied to the internal nodes and the nodes on the essential boundaries, while the local radial point interpolation weak form is applied to the nodes on the natural boundaries. The advantages of this MWS method are:

1. The Neumann boundary condition can be imposed straightforwardly and accurately with arbitrary nodal distributions.
2. Stable and accurate solution can be obtained with high efficiency.

2.4.2 Applications of each category

Strong form methods are suitable for Dirichlet boundary conditions problems and weak form methods are used more with problems that have Neumann boundary conditions. Weak-strong form methods are appropriate for problems with both of Dirichlet and Neumann boundary conditions.

3. Shape functions

In seeking for an approximate solution to a problem governed by PDEs and boundary conditions, one first needs to approximate the equation variables using shape functions, befor any formulation procedure can be applied to establish the descretized system equations.

This section discusses various techniques for MFree shape function constructions. These shape functions are locally supported, because only a set of field nodes in a small local domain are used in the construction and the shape function is not used or regarded as zero

outside the local domain. Such a local domain is termed the support domain or influence domain or smoothing domain.

3.1 Point interpolation methods shape functions

The point interpolation method (PIM) is one of the series representation methods for the function approximation, and is useful for creating Meshfree shape functions. Consider a scalar function $T(x)$ defined in the problem domain Ω that is represented by a set of scattered nodes. The PIM approximates $T(x)$ at a point of interest x in the form of

$$\mathrm{T(x)}=\sum_{i=1}^{m} B_i(x)a_i \tag{1}$$

where the $B_i(x)$ are the basis function defined in the space Cartesian coordinates $X^T=[x, y]$, m is the number of basis functions, and the a_i are the coefficients.

For function approximation, a local support domain is first formed for the point of interest at x which includes a total of n field nodes. For the conventional point interpolation method (PIM), $n=m$ is used that results in the conventional PIM shape functions that pass through the function values at methods. The RPIM interpolation augmented with each scattered node within the defined support domain.

For the weighted least square (WLS) approximation or the moving least squares (MLS) approximation, n is always larger than m. There are two types of PIM shape functions have been developed so far using different forms of basis functions. Polynomial basis functions (GR Liu and Gu, 1999; 2001a) and radial basis functions (RBF) (Wang and GR Liu, 2000; GR Liu, 2002) have often been used in Meshfree methods.

3.1.1 Conventional polynomial PIM

Using polynomials as the basis functions in the interpolation is one of the earliest interpolation schemes. It has been widely used in establishing numerical methods, such as the FEM. Consider a continuous function $u(x)$ defined in a domain Ω , which is represented by a set of field nodes. The $u(x)$ at a point of interest x is approximated in the form of

$$\mathrm{u(x)}=\sum_{\mathrm{i=1}}^{\mathrm{m}} p_i(\mathrm{x})a_i = \{p_1(\mathrm{x})\ p_2(\mathrm{x})\ \ \ldots\ldots\} \begin{Bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \end{Bmatrix} = P^T a \tag{2}$$

$$P^T(x)=(1\ \ x\ \ \ y\ \ x^2\ \ \ xy\ \ y^2) \qquad for\ m=6,\ p=2\ (2\text{-}D) \tag{3}$$

where $p_i(x)$ is a given monomial in the polynomial basis function in the space coordinates $x^T=[x,y]$, m is the number of monomials, and a_i is the coefficient for $p_i(x)$ which is yet to be determined. The $p_i(x)$ in Equation is built using Pascal's triangles, and a complete basis is usually (but not always)

3.1.2 Radial point interpolation shape functions

In order to avoid the singularity problem in the polynomial PIM, the radial basis function (RBF) is used to develop the radial point interpolation method (RPIM) shape functions for Meshfree weak form methods (GR Liu and Gu, 2001c; Wang and Liu, 2000; 2002a,c). The RPIM shape functions will be used for both Meshfree weak form and strong form polynomials can be written as

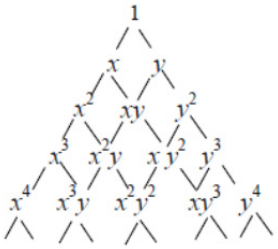


Fig. 1. Pascal-khayyam triangle of monomials for two dimensional domain

$$u(x,y)=\sum_{j=1}^{n} R_j(x,y)b_j + \sum_{i=1}^{m} p_i(x,y)a_i = R^T B + P^T A \tag{4}$$

Where $R_j(x)$ is a radial basis function (RBF), n is the number of RBFs, $p_i(x)$ is monomial in the space coordinates $x^T=[x, y]$, and m is the number of polynomial basis functions. When $m=0$, pure RBFs are used. Otherwise, the RBF is augmented with m polynomial basis functions. Coefficients a_i and b_j are constants. r is the distance between the point of interest (x,y) and a node (x_i,y_i) at

$$r=\sqrt{(x-x_i)^2+(y-y_i)^2} \tag{5}$$

There are a number of types of radial basis functions (RBF), and the characteristics of RBFs have been widely investigated (Kansa,1990; Sharan et al.,1997; Franke and Schaback, 1997; etc). Four often used RBFs, the multi-quadrics (MQ) function, the Gaussian (Exp) function, the thin plate spline (TPS) function, and the Logarithmic radial basis function, are listed in Table.1.

	Name	†Expression	Shape Parameters
1	Multi-quadrics (MQ)	$R_i(x,y)=(r_i^2+(\alpha_c d_c)^2)^q$	$\alpha_c \geq 0\,,\, q$
2	Gaussian (EXP)	$R_i(x,y)=\exp[-\alpha_c(\frac{r_i}{d_c})^2]$	α_c
3	Thin Plate Spline (TPS)	$R_i(x,y)=r_i^\eta$	η
4	Logarithmic	$R_i(x,y)=r_i^\eta \log r_i$	η

Table 1. Typical radial basis functions with dimensionless shape parameters

Note: d_c is a characteristic length that relates to the nodal spacing in the local support domain of the point of interest x, and it is usually the average nodal spacing for all the nodes in the local support domain.

$$R_0=\begin{bmatrix} R_1(r_1) & R_2(r_1) & \cdots & R_n(r_1) \\ R_1(r_2) & R_2(r_2) & & R_n(r_2) \\ \vdots & & & \vdots \\ R_1(r_n) & R_2(r_n) & & R_n(r_n) \end{bmatrix} \tag{6}$$

the polynomial moment matrix is

$$P_m^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ P_m(x_1) & P_m(x_2) & & P_m(x_n) \end{bmatrix} \tag{7}$$

$$r_k = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{8}$$

$$\tilde{U}_S = \begin{bmatrix} U_S \\ 0 \end{bmatrix} = \begin{bmatrix} R_0 & P_m \\ P_m^T & 0 \end{bmatrix} \begin{Bmatrix} a \\ b \end{Bmatrix} = G\,a_0 \tag{9}$$

$$u(x) = \left\{ R^T(x) \quad P^T(x) \right\} G^{-1} \tilde{U}_S = \tilde{\Phi}^T(x) \tilde{U}_S \tag{10}$$

$$\begin{aligned} \tilde{\Phi}^T(x) &= \left\{ R^T(x) \quad P^T(x) \right\} G^{-1} = \\ &= \{\varphi_1(x) \quad \varphi_2(x) \cdots \varphi_n(x) \quad \varphi_{n+1}(x) \cdots \varphi_{n+m}(x)) \end{aligned} \tag{11}$$

The above equations are brought to show the procedure of shape function produce. The shape functions Φ are obtained and then the discretized derivatives can be used to governing equations and the parameters are shown with the equation

$$u(x) = \Phi^T(x) U_S = \sum_{i=1}^{n} \phi_i u_i \tag{12}$$

The derivatives of u(x) are easily obtained as

$$u_{,l}(x) = \Phi_{,l}^T(x) U_S \tag{13}$$

where l denotes either the coordinates x or y.

3.2 Support domain

The accuracy of interpolation for the point of interest depends on the nodes in the support domain as shown in Fig.2. Therefore, a suitable support domain should be chosen to ensure an efficient and accurate approximation. For a point of interest at x_q, the dimension of the support domain d_c is determined by

$$\begin{aligned} r_{sx} &= \alpha_s . d_{cx} \\ r_{sy} &= \alpha_s . d_{cy} \end{aligned} \tag{14}$$

α_s is the dimensionless size of the support domain, and d_c is the nodal spacing near the point at x_q. If the nodes are uniformly distributed, d_c is simply the distance between two neighboring nodes. When nodes are non uniform and where α is a constant of shape parameter, dc can be defined as an average nodal spacing in the support domain of x_q. The exponential function of the support domain α_s controls the actual dimension of the support domain.

Rectangular support domains (r_{sx} and r_{sy}: dimensions of the support domain in x and y directions). The support domain is centred by x_Q.

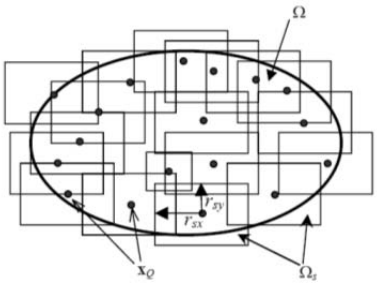


Fig. 2. Support domains of points of interest at x_Q in Meshfree models.

The actual number of nodes, n, can be determined by counting all the nodes included in the support domain. Generally, an α_s=2.0~3.0 leads to good results for many problems that we have studied. Note that the support domain is usually centered by a point of interest at x_Q.

4. Satisfying boundary conditions

For the Dirichlet boundary condition, the essential boundary conditions for u can be simply given as follows: (when node is on the boundary)

$$u=\overline{T} \tag{15}$$

The essential boundary condition can be directly imposed using the direct interpolation method. another method is the Penalty method has been used to enforce essential boundary conditions in the MLPG and LRPIM Methods. Since RPIM shape functions possess the Kronecker delta function property, the essential boundary conditions can be easily enforced as in the FEM (see, e.g., GR Liu and Quek, 2003).
The natural boundary conditions can be satisfy automatically when we use weak-strong form method and no additional equation or treatment is needed.

4.1 Direct method

The ith component is prescribed by setting

$$u_i=\overline{T}_i \tag{16}$$

Such an essential boundary condition can then be enforced directly into the system Equation through the following modifications to the global matrix and the global right vector. The global matrix, K, is changed to

$$K=\begin{bmatrix} k_{11} & \cdots & k_{1(i-1)} & 0 & k_{1(i+1)} & \cdots & k_{1n} \\ \vdots & & & 0 & & & \vdots \\ k_{(i-1)1} & & & 0 & & & k_{(i-1)n} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ k_{(i+1)1} & & & 0 & & & k_{(i+1)n} \\ \vdots & & & 0 & & & \vdots \\ k_{n1} & \cdots & k_{n(i-1)} & 0 & k_{n(i+1)} & \cdots & k_{nn} \end{bmatrix} \tag{17}$$

The components in the global right vector are changed to

$$F_i = \begin{cases} \overline{T}_i & i = j \\ F_j - k_{ji}\overline{T}_i & i \neq j \end{cases} \tag{18}$$

The direct method can exactly enforce essential boundary conditions, but changing matrices and vectors needs additional computational operations. In addition, the algorithm of the direct method is also complicated.

4.2 Penalty method

The penalty method is a convenient alternative for enforcing the essential boundary conditions, in which the diagonal entry k_{ii} in the stiffness matrix, is changed to

$$k_{ii} = \alpha . k_{ii} \tag{19}$$

where α is the penalty coefficient that is the much larger number than the components of the global matrix K. In the global right vector F, only the component F_i is changed as follows

$$F_i = \begin{cases} \alpha . k_{ii} . \overline{T}_i & i = j \\ F_j & i \neq j \end{cases} \tag{20}$$

The penalty method has some advantages: there are only two changes of matrices, and the algorithm is very simple. However, the penalty method can only approximately satisfy the essential boundary conditions. In addition, the accuracy is affected by selection of the penalty coefficient.
the global matrix, K, is then changed to

$$K = \begin{bmatrix} k_{11} & \cdots & k_{1(i-1)} & k_{1i} & k_{1(i+1)} & \cdots & k_{1n} \\ \vdots & & & \vdots & & & \vdots \\ k_{(i-1)1} & & & k_{(i-1)i} & & & k_{(i-1)n} \\ k_{i1} & \cdots & k_{i(i-1)} & \alpha k_{ii} & k_{i(i+1)} & \cdots & k_{in} \\ k_{(i+1)1} & & & k_{(i+1)i} & & & k_{(i+1)n} \\ \vdots & & & \vdots & & & \vdots \\ k_{n1} & \cdots & k_{n(i-1)} & k_{ni} & k_{n(i+1)} & \cdots & k_{nn} \end{bmatrix} \tag{21}$$

5. Examples for numerical simulations

In this section, some problems are brought to show the abilities of meshfree methods for solving the heat, solid and fluid mechanics problems.

5.1 Heat conduction

Meshfre methods are used to solve the heat transfer problem. For example, heat conduction in the plate is solved.

5.1.1 Formulation of heat transfer in the plate

Formulation of heat transfer in Cartesian coordinate is

$$\frac{\partial}{dx}(k\frac{\partial T}{\partial x})+\frac{\partial}{dy}(k\frac{\partial T}{\partial y})+q'=\rho C_p(\frac{\partial T}{\partial t}) \tag{22}$$

If k(conductivity coefficient) is constant and for steady state and without any energy generation we have:

$$(\frac{\partial^2 T}{\partial x^2})+(\frac{\partial^2 T}{\partial y^2})=0 \tag{23}$$

T is the Temperature and q' is the rate of energy generation and ρ is the density and C_p is the specific heat in the formula.

5.1.2 Numerical results and discussion Domain representation for heat transfer

First, the temperature distribution in square plate is obtained. In problem 1, the bottom wall is in temperature T_0 and other walls are in temperature 0 and in problem 2, the up wall has Neumann boundary condition. To check the validity of the method, three different problems are considered. Fig.3 shows the domain representation for problems 1 and 2 by the scattered nodes. The essential and natural boundary conditions should be satisfied on the boundary nodes.

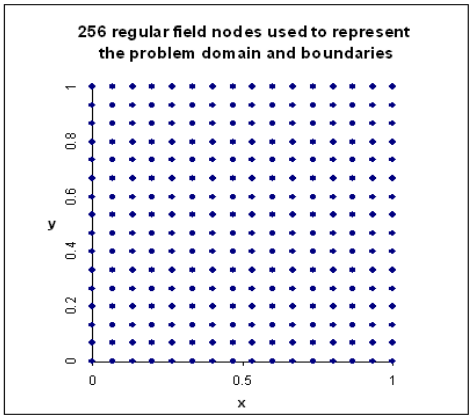


Fig. 3. The problem is represented by 256(16x16) regular nodes

5.1.3 Problem 1 with essential boundary conditions

Fig.4 shows the problem 1 and its boundary conditions. The temperature distribution in the plate obtained by MWS method presented in this chapter is given in Fig. 5.

The constant temperature lines in this figure are shown by solid lines. Table.2 compares the results obtained by collocation method and LRPIM method with those obtained by the analytical method.

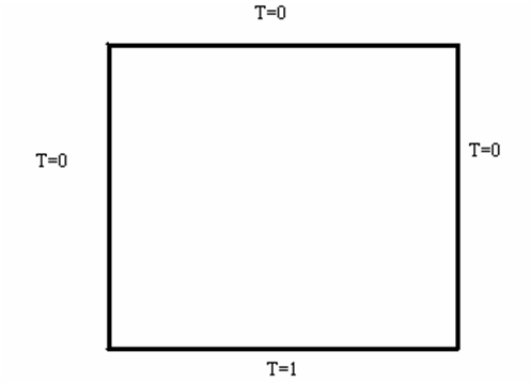


Fig. 4. Problem 1 and its boundary conditions

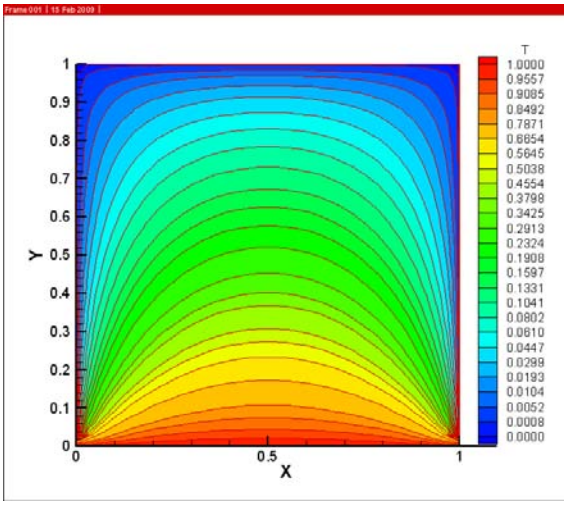


Fig. 5. Temperature distribution with essential boundary conditions (problem1 solved with MWS method)

the analytical solution of the problem can be written as

$$T(x,y)=\frac{2}{\pi}\sum_{n=1}^{\infty}\frac{-1^{n+1}+1}{n}\sin(n\pi x)(-\tanh(n\pi).\cosh(n\pi y)+\sinh(n\pi y)) \tag{24}$$

(x,y) are the coordinate of points in the plate. *T* is the temperature. We showed the difference between three meshfree methods and the difference between using different number of nodes to give the better results.
We used the error norm

$$Err(j)=\frac{\left|T(j)-T_{analytic}(j)\right|}{T_{analytic}(j)} \tag{25}$$

and Total Error defined as

$$Total\,Err = \frac{(\sum_{j=1}^{n} Err(j))}{n} \tag{26}$$

x=0.5		256 nodes			1156 nodes	
y	Analytical	Collocation	LRPIM	MWS	LRPIM	MWS
0	1	1	1	1	1	1
0.1	0.8017	0.7978	0.7980	0.7976	0.8010	0.8013
0.2	0.6208	0.6137	0.6142	0.6134	0.6198	0.6203
0.3	0.4679	0.4624	0.4632	0.4620	0.4667	0.4671
0.4	0.3449	0.3390	0.3401	0.3386	0.3445	0.3448
0.5	0.2500	0.2462	0.2473	0.2458	0.2493	0.2495
0.6	0.1765	0.1728	0.1737	0.1724	0.1760	0.1761
0.7	0.1194	0.1174	0.1181	0.1171	0.1190	0.1191
0.8	0.0737	0.0720	0.0725	0.0717	0.0735	0.0736
0.9	0.0351	0.0345	0.0419	0.0342	0.0350	0.0350
1	0	0	0	0	0	0

Table 2. Comparison between Mfree methods and analytical method in problem 1

x=0.5		256 nodes			1156 nodes		
y	Analytical Err	Collocation Err	LRPIM Err	MWS Err	Collocation Err	LRPIM Err	MWS Err
0.1	0	0.0051	0.0046	0.0051	0.0015	0.0009	0.0004
0.2	0	0.0119	0.0106	0.0119	0.0029	0.0016	0.0008
0.3	0	0.0126	0.0100	0.0126	0.0045	0.0026	0.0009
0.4	0	0.0183	0.0139	0.0183	0.0038	0.0012	0.0009
0.5	0	0.0168	0.0108	0.0168	0.0056	0.0028	0.0008
0.6	0	0.0232	0.0159	0.0232	0.0062	0.0028	0.0006
0.7	0	0.0193	0.0109	0.0193	0.0067	0.0034	0.0008
0.8	0	0.0271	0.0163	0.0271	0.0068	0.0027	0.0014
0.9	0	0.0256	0.1937	0.0256	0.0085	0.0028	0.0001
Total Err		0.018	0.0319	0.0178	0.005	0.0023	0.0007

Table 3. Errors in problem 1

5.1.4 Problem 2 with natural boundary condition

The MWS method is used to solve the same problem with both essential and natural boundary conditions by 256(16x16) and 1156(34x34) nodes. The temperature distributions for those problems are given in Fig.7. It should be noted that the essential boundary conditions are satisfied exactly whereas the natural (Neumann) boundary conditions are satisfied in the weak form formulation.

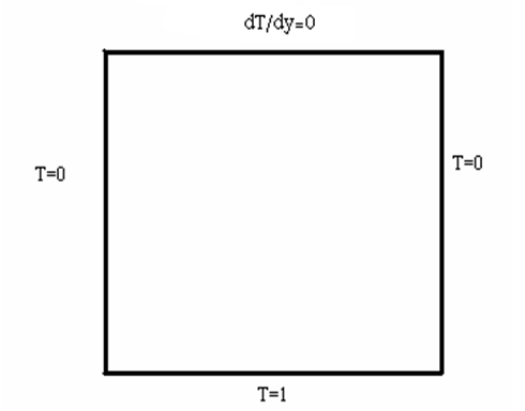


Fig. 6**.** Problem 2 and its boundary conditions

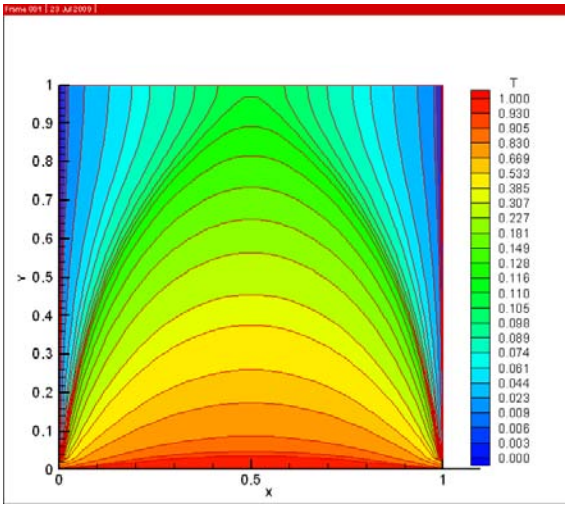


Fig. 7. Temperature distribution with 1156 nodes (Problem2 solved with MWS method)

In Tables 4 and 5 the LRPIM and MWS methods are compared with the analytical method. The numerical values for the temperature distributions with 256 and 1156 nodes are also given in Tables 4 and 5. The defined error equations (25 and 26) are used to show the accuracy of MWS and LRPIM.

5.2 Lid driven cavity problem

In this section, the lid driven cavity problem is solved by meshfree method. In this problem, the cavity is full of fluid and the upper plate in the cavity drive horizontally. It is shown that the moving boundary conditions in the top wall are easily applied and natural boundary condition can be satisfied.

x=0.5		256 nodes		1156 nodes	
y	analytical	MWS	LRPIM	MWS	LRPIM
0	1	1	1	1	1
0.1	0.8047	0.8006	0.8009	0.8035	0.8041
0.2	0.6271	0.6196	0.6202	0.6254	0.6262
0.3	0.4782	0.4723	0.4732	0.4762	0.4771
0.4	0.3606	0.3538	0.3550	0.3589	0.3598
0.5	0.2718	0.2675	0.2688	0.2705	0.2712
0.6	0.2071	0.2026	0.2039	0.2059	0.2066
0.7	0.1617	0.1591	0.1603	0.1608	0.1614
0.8	0.1320	0.1293	0.1304	0.1314	0.1318
0.9	0.1152	0.1137	0.1147	0.1147	0.1151
1	0.1098	0.1081	0.1089	0.1093	0.1097

Table 4. Comparison between Meshfree methods and analytical method in problem 2

x=0.5		256 nodes		1156 nodes	
y	Analytical Err	MWS Err	LRPIM Err	MWS Err	LRPIM Err
0.1	0	0.0051	0.0047	0.0015	0.0007
0.2	0	0.0120	0.0110	0.0027	0.0014
0.3	0	0.0123	0.0105	0.0042	0.0023
0.4	0	0.0189	0.0155	0.0047	0.0022
0.5	0	0.0158	0.0110	0.0048	0.0022
0.6	0	0.0217	0.0155	0.0058	0.0024
0.7	0	0.0161	0.0087	0.0056	0.0019
0.8	0	0.0205	0.0121	0.0045	0.0015
0.9	0	0.0130	0.0043	0.0043	0.0009
1	0	0.0158	0.0082	0.0046	0.0009
Total Err		**0.0162**	**0.0108**	**0.0046**	**0.0017**

Table 5. Errors in problem 2

5.2.1 Formulation and boundary conditions of driven cavity problem

The application of Navier-Stokes equation in solving fluid flow has evolved in the past few decades with meshfree method as one of the most adopted techniques. In this section the Navier-Stockes equation is solved:

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=-\frac{\partial P}{\partial x}+1/\mathrm{Re}\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \tag{27}$$

The boundary conditions are shown in fig.8 and it is shown that three walls are without motion and the upper wall move with the fix speed.

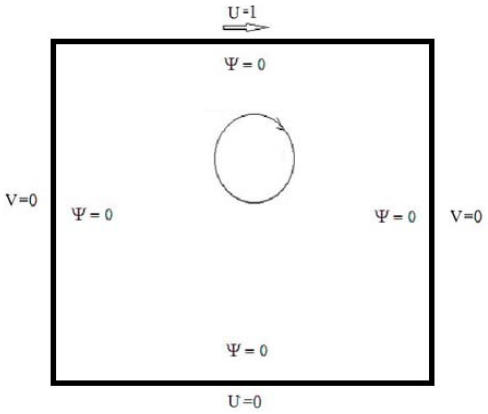


Fig. 8. Boundary conditions for driven cavity problem

5.2.2 Numerical results

The results from the solution of driven cavity problem when Reynolds number is 100 are shown below:

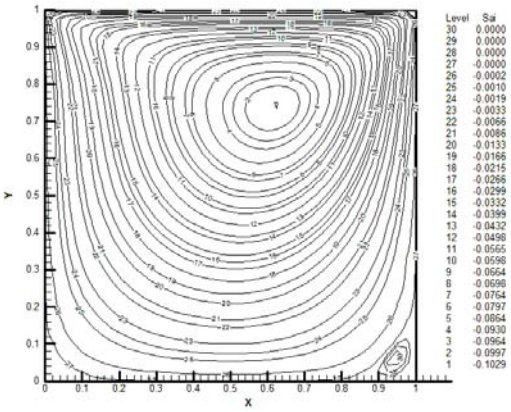


Fig. 9. Stream line contours for Reynolds 100

The vortex on the corner is created and is related to Reynolds number and the corner vortex will grow if Reynolds number increases. The vorticity and velocity contours are shown:

Location of large vortex	Minimum stream function	Reynolds	Reference
(0.6172,0.7344)	-0.0103	100	Ghia and Shin
(0.6196,0.7373)	-0.0103	100	Hou and Doolen
(0.6231,0.7460)	-0.0129	100	Present
(0.5547,0.6055)	-0.114	400	Ghia
(0.5500,0.6125)	-0.113	400	Gupta and Kalita
(0.5411,0.6005)	-0.114	400	Present

Table 6. Minimum stream function and location of large vortex

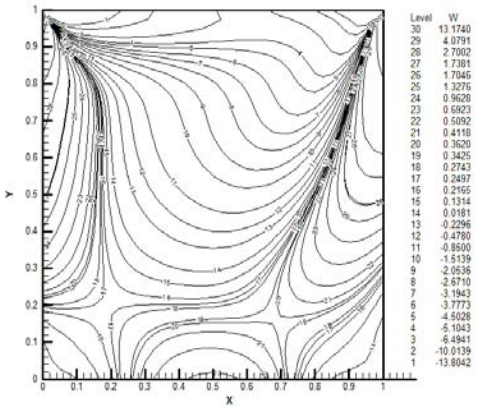


Fig. 10. Vorticity contours for driven cavity problem for Reynolds 100

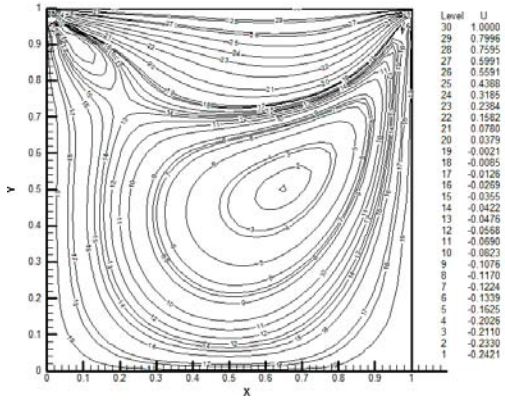


Fig. 11. Horizontal velocity contours for driven cavity problem for Reynolds 100

In Table 6 the results are compared with the minimum stream function and the location of large vortex. The results show the ability of meshfree methods to simulate the fluid mechanics problems.

5.3 Cantilever beam problem

Numerical studies are conducted for a cantilever beam that is often used for benchmarking numerical methods because the analytic solution for this problem is known. This problem is a sample of solid mechanics.

5.3.1 Formulation of cantilever beam problem

The equilibrium equation is used with the formula:

$$\sigma_{ij,j} + b_i = 0 \tag{28}$$

σ is the stress vector and b_i is the body force vector components. The strain-displacement relations are another formula that are brought in two directions:

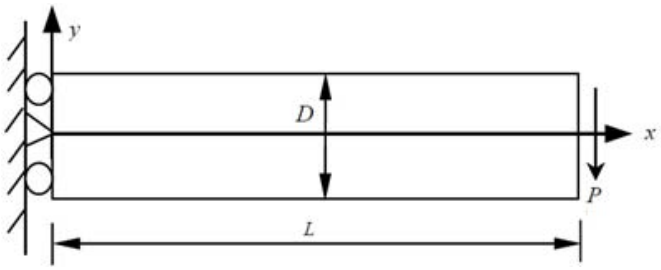


Fig. 12. The beam problem

$$\varepsilon_{xx} = \frac{\partial u}{\partial x} \tag{29}$$

ε_{xx} is the strain component and u is the displacement in the horizontal direction.

$$\varepsilon_{yy} = \frac{\partial v}{\partial x} \tag{30}$$

ε_{yy} is the strain component and v is the displacement in the vertical direction.
The last equation is Hook's law:

$$\sigma = D_e \varepsilon \tag{31}$$

D_e is the matrix of elastic constant.

5.3.2 Numerical results

The analytical solution is obtained for displacement of points of beam:

$$u = \frac{Py}{6EI}\left[(6L - 3x)x + (2 + \upsilon)(y^2 - \frac{D^2}{4})\right] \tag{32}$$

u is the displacement of points in horizontal direction and P is the force at the end of the beam. E is the elasticity modulus and υ is the poisson ratio and moment of inertia is I, D is the height and L is the length

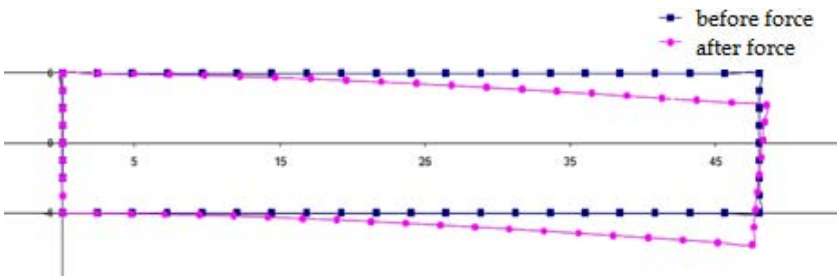


Fig. 13. The beam after effect of force

The energy norm is defined to compare the results.

$$e_n = \sqrt{\frac{1}{2}(\int_{\Omega} (\varepsilon_{numer} - \varepsilon_{exact}) D_e (\varepsilon_{numer} - \varepsilon_{exact}) d\Omega} \quad (33)$$

Ω is the problem domain and ε_{num} and ε_{exact} are the strain vector with numerical and analytical solutions.

Energy norm	Solution method	No.
0.0258	*MWS*	1
0.026	*LRPIM*	2

Table 7. Energy norm for beam problem

These results show the capacity of MWS and LRPIM methods to solve the solid mechanics problems. The Table 7 shows the errors are minimums with comparison with the analytic solution.

6. Acknowledgement

This research has been mainly financed by Shiraz university. The author appreciate the support from Shiraz university.

7. Conclusion

The meshfree methods are numerical methods that can be used to solve the many different and complicated problems. The heat transfer problems, solid and fluid mechanics problems have been solved with meshfree methods.

Three categories are used to solve the problems. Strong form methods, weak form methods and weak-strong form methods(MWS) are meshfree categories. They can be used to solve the problems with Dirichlet and Neumann boundary conditions. For examples the heat conduction problem and lid driven cavity and cantilever beam are solved that they have different type of boundary conditions. Solutions are related to many parameters: the selected meshfree method, number of nodes, shape function parameters et cetera.

Nowadays, many changes are employed to different types of meshfree methods. The advantages are improved and the high convergence rate and high accuracy are accessible.

8. References

Hou, S. , Doolen, G. & Cogley, A. (1995). Simulation of cavity flows by lattice boltzmann method, *Journal of computational physics,* Vol.118, pp. 329-347

Gu, Y.T. & Liu, G.R. (2005). A meshfree weak-strong (MWS) form method for time dependent problems, *Computational Mechanics*, Vol. 35, No.2 , pp. 134-145

Gu, Y.T & Liu, G.R. (2001). A Local Point Interpolation Method (LPIM) For Static And Dynamic Analysis Of Thin Beams, *Computer Methods in Applied Mechanics and Engineering*. Vol. 190

Hong, W.u. & Quan, W. (2007) Meshless method based on local weak form s for steady-state heat conduction problems, *International Journal of Heat and Mass Transfer* Vol.51, (2008) pp. 3103-3112

Incropera, Frank. & Witt, David.P. (2002). *Introduction to heat transfer,* 4th edition, springer

Liu, G.R. & Gu, Y.T. (2005). *An Introduction to Meshfree Methods and Their Programming*, springer

Liu, G.R. , Wu, Y.L. & Ding, H. (2005). Meshfree weak-strong(MWS) form method and its application to imcompressible flow problems, *International Journal for Numerical Methods in Fluids*, Vol.46, pp. 1025-1047

Liu, G.R. & Gu, Y.T. (2003). A meshfree method: Meshfree Weak-Strong (MWS) form method, for 2-D solids, *Computational Mechanics*, Vol.33, No.1, pp. 2-14

Liu, G.R. , Yan, L. , Wang, J.G. & Gu, Y.T. (2002). Point Interpolation Method Based On Local Residual Formulation Using Radial Basis Functions, *Structure Engineering Mechanic*, Vol.14, No.6, pp. 713-732

Liu, G.R. & Gu, Y.T. (2001). A local radial point interpolation method (LRPIM) for free vibration analyses of 2-D solids, *Journal of Sound and Vibration,* Vol.246, No.1, pp. 29–46

Liu, G.R. & Gu, Y.T. (2003). A meshfree formulation of local radial point interpolation method (LRPIM) for incompressible flow simulation, *Computational Mechanics* Vol.30, pp. 355–365

Rao, S. (2004), *The Finite Element Method in Engineering*, Elsevier Science & Technology Books.

Reddy, J.N. (2006), An *introduction to finite element method*, Third edition, Springer, McGraw-Hill Publishing Corporation

Zahiri, S. , Daneshmand, F. and Akbari, M.H. (2009). Using meshfree weak-strong form method for 2-D heat transfer problem, *Proceedings of ASME 2009 International Mechanical Engineering Congress and Exposition*, IMECE2009-12525, Lake Buena Vista, Florida, USA, November 13-19, 2009

Part 2

Application

12

Mechanics of Deepwater Steel Catenary Riser

Menglan Duan[1], Jinghao Chen[1] and Zhigang Li[2]
[1]*Offshore Oil/Gas Research Center, China University of Petroleum, Beijing,*
[2]*Offshore Oil Engineering Co., Ltd., Tianjian,*
P. R. China

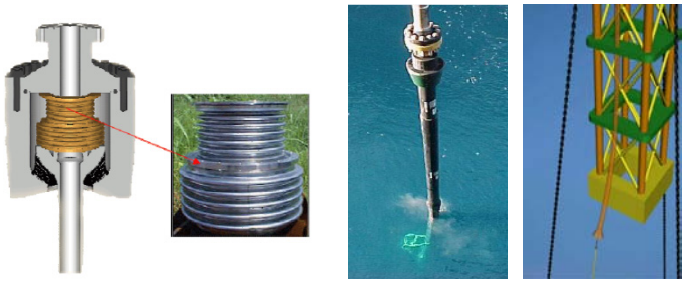
1. Introduction

With the exploration and development of oil and gas expanded to deepwater area, many new floating structures are developed to reduce the cost. Steel catenary Riser (SCR) is a flexible steel pipe that conducts well fluids from the subsea wellhead to the production floating vessel. SCR has the advantages of low manufacturing cost, resistance of high temperature and high pressure, good adaptability of upper floating body's motion, etc, and is widely used in the development of deepwater oil and gas fields. Because of the complicated marine environment, SCR theory research involves fluid dynamics, nonlinear mechanics, soil mechanics and other disciplines. This chapter presents the numerical calculation for soil-riser interaction, vortex-induced vibration (VIV), fatigue, the coupling of floating vessel and riser, riser installation, etc, and provides a theoretical basis of SCR design.

1.1 Configuration of SCR

The static configuration of SCR is shown in figure 1.1(Kavanagh et al., 2004). SCR connects floating structure with some kinds of devices, such as flexjoint, J-tube, tapered stress joint, etc. Three kinds of connecting devices are shown in figure 1.2. In order to reduce Vortex-induced vibration (VIV), the segment of SCR is equipped with VIV suppression device(Boubenider, 2008; Taggart and Tognarelli, 2008). Figure 1.3 presents two main suppression devices of helical strake and fairing.

Fig. 1.1 The static configuration of SCR



a. Flexjoint b. TSJ c. J-Tube

Fig. 1.2 Three kinds of connecting devices

a. Helical Strake b. Fairing

Fig. 1.3 VIV suppression devices

1.2 History of SCR development

SCRs were initially installed on fixed platforms. Until 1994, SCRs were firstly installed on a floating platform-Auger TLP, and since then have been widely used in deepwater oil/gas fields. The main application of SCR is in Brazil, the Gulf of Mexico and West Africa on TLP, Spar and semi-submersible. Recent years, SCR is widely installed on FPSO in West Africa. This section presents some installation instances(Bai and Bai, 2005).

1994 The first SCRs were installed on Auger TLP in Gulf of Mexico.

1997 The first SCRs were installed on semi-submersible in Marlim Field.

2001 The first SCRs were installed on truss spars locate at Boomvang and Nansen fields, in Gulf of Mexico.

2004 The first SCRs were installed on FPSO vessel in West Africa.

1.3 Analysis software for Risers

Various riser analysis tools can be divided into two classes. One is general finite element software(Abaqus Analysis User's Manual 6.9; ANSYS ASAS Brochure), the other is professional software(Bai and Bai, 2005; Orcaflex Manual Version 9.3c).

1. General finite element software

ABAQUS

ABAQUS is developed by world famous computer simulation software company SIMULIA. It is a powerful finite element software which can solve the problem scoped from simple linear analysis to complex nonlinear analysis.

ABAQUS/Aqua is a module used in offshore industry. It includes jacket and riser analysis, J-tube pull simulations, bottom-bending calculations, and floating structure studies.

Structures can be subjected to drag, buoyancy, and fluid inertia force under steady current and wave loadings. Wind loading is available for riser above the surface of the water.

ANSYS

ANSYS is a general purpose finite element software. It can solve structure, fluid, electric field, magnetic field, sound field and multi-physical coupling problems.
ANSYS/ASAS is a structural finite element system containing the features to meet the needs of offshore and marine engineers. ANSYS/ASAS provides the capabilities to analyze the global structures of most types of marine structures, including jackets, jack-ups, risers, offshore wind farms, and floating systems such as FPSOs, SPARs and semi-submersibles.

2. Professional software

Orcaflex

OrcaFlex is 3D non-linear time-domain finite element software developed by Orcina for static and dynamic analysis of a wide range of offshore systems, including marine risers, moorings and towed systems. It also provides some modeling elements, such as line, vessel, buoy, winch and seabed, etc. It also provides extensive graphics to assist understanding.

Shear7

Shear7 is one of the leading modeling tools for the prediction of vortex-induced vibration (VIV), developed by Professor J. Kim Vandiver in MIT. Shear7 is a mode-superposition software, which evaluates the modes likely to be excited by vortex shedding and estimates VIV response in uniform or sheared flows. It is capable of analyzing multi-mode and non-lock-in response as well as single mode lock-in response.
Other software includes Offpipe, Riflex, VIVA, Flexcom, etc.

1.4 Overview on present research of SCR

Mechanics of deepwater steel catenary riser is a cross-disciplinary subject that involves soil mechanics, fluid mechanics, wave mechanics and vibration mechanics. The challenge of SCR design mainly reflects three aspects: pipe-soil interaction, VIV, and coupled riser and hull.

1. Pipe-soil interaction

The touchdown zone (TDZ) is one of the key locations where the fatigue damage happens. Pipe-soil interaction affects the assessment of fatigue damage. A lot of work has been done to discover the pipe-soil interaction mechanism. Two models have been established: non-degradation model and degradation model. Non-degradation model has been applied to SCR design. Degradation model takes consideration of soil plastic deformation, water mixed, soil reconsolidation, etc. These factors affect the deformation of the trench. The establishment of degradation model is still a challenge.

2. Vortex-induced vibration

So far, the numerical simulation of VIV does not have perfect solution. The key issue is the determination of hydrodynamic force. Empirical models are obtained from experiments, which are related to empirical coefficients. The experiments at high Reynolds number should be carried out to show what new phenomena appear. Another method is CFD. Directly solving the Navier-Strokes equations requires very refined grid and micro time step. Researchers seek an approximation model. Reynolds averaged Navier-Strokes (RANS) and Large Eddy Simulation (LES) mesh more rough and save computation time, which are widely used in engineering.

3. Coupled hull and riser

The importance of coupled analysis between riser and hull has been recognized in deepwater engineering. Considering the complexity of the coupled system, each part should be modeled mathematically. To make it more efficient, proper simulation and numerical methods should be used. Recently, the hull is taken as a 6 DOF rigid body and a slender rod theory is applied to simulating the riser and mooring line. Spring and damper are used to simulate the interaction and connection. The main difference is how they solve the coupled equation. Several engineering software have been compiled based on current research results.

The numerical simulation of SCR involved many theoretical systems and each has many branches. This chapter does not carry out a thorough study on mechanics of SCR. It only briefly introduces the numerical simulation methods commonly used in offshore industry.

2. Pipe-soil interaction

When the SCR is subjected to oscillating movement, there is a complex interaction between SCR and seabed. The touchdown zone (TDZ) is also the key location where SCR fatigue damage happens. Pipe-Soil interaction is the important factor that should be considered in SCR strength and fatigue analysis. Owing to the complex nonlinear behavior of soil, it's hard to establish a precise model. How to accurately simulate this interaction is still a challenge and has been a hot academic research. It affects the calculation of fatigue damage and the prediction of fatigue life.

The interaction between SCR and the seabed is affected by many factors, such as SCR properties, water entrainment, soil consolidation time, soil erosion and the development of trenching. A linear or nonlinear spring is used to model the seabed soil interacting with SCR. But linear spring does not represent the real behavior of the soil. So, recently researches on the behavior of soil interacting with SCR have focused on P-y curves of soil (where P stands for the resistance force of soil and y for the vertical penetration of the SCR). Many researchers acquired the empirical equations from experiments. It can be classified into two kinds: non-degradation model and degradation model. This section presents a typical non-degradation Pipe-Soil interaction models.

2.1 Process of pipe-soil interaction

A typical cycle of loading-unloading-reloading is presented in Figure 2.1(Nakhaee, 2010). The P-y curve can be divided into four different paths. As shown in Figure2.2, the process of pipe-soil interaction is introduced form (a) to (e), corresponding to the stage 1-5.

Stage 1. The pipe is initially laid on virgin seabed. And there is no penetration.

Stage 2. It describes the initial penetration following along the backbone curve from Point 0 to 1. The penetration displacement is determined by W and P. W is the vertical force acting on the soil (usually the submerged weight of SCR per unit length) and P is the soil resistance.

Stage 3. The pipe moves upwards and soil resistance is reduced quickly. After the soil resistance reaches zero, the loading is soil suction and increases to the peak rapidly. This process is described by P-y curve from point 1 to 2 on figure 2.1.

Stage 4. When the pipe is going on to uplift, the soil suction gradually diminishes and reaches zero from point 2 to 3.

Stage 5. The pipe penetrates again into the soil. The resistance force follows along the curve from point 3 to 1, which is lower than backbone curve.

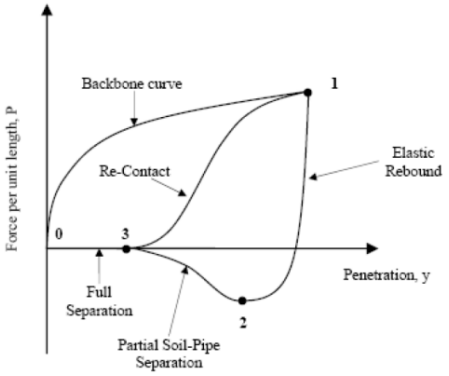
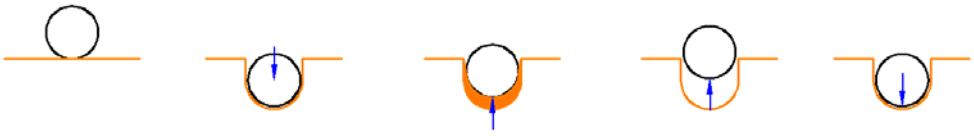


Fig. 2.1 Typical P-y curve



(a) (b) (c) (d) (e)

Fig. 2.2 Process of pipe-soil interaction

If the riser continues to experience the periodic loading cycle, the P-y relation will repeat the loop enclosed by the second, third and fourth path under the assumption of a non-degradation model. It should be noted that the loop area is greatly exaggerated in Figure2.1 for the purpose of demonstrating.

2.2 Pipe-soil interaction model

Earlier pipe-soil models are too simplified to simulate the complex interaction between pipe and soil. They assumed the seabed as rigid flat or linearly elastic spring. Many researchers investigate the mechanism of pipe-riser interaction by experiments and give various experiment equations. One famous experiment is STRIDE JIP's full scale harbor experiment(Willis and West, 2001). This experiment experiences 3 months at a harbor in the west of England. Figure 2.3 presents this test. Bridge developed advanced non-degradation models using published data and pipe-soil interaction experiments conducted by the STRIDE and CARISIMA JIP(Bridge and Laver, 2004). Bridge's models have been widely used in many Gulf of Mexico deepwater projects involving SCRs. Bridge's models are typical non-degradation model.

The process of pipe-soil interaction is introduced in 2.1. The soil force can be divided into 2 phases: soil resistance and soil suction. This part introduces the related concepts of pipe-soil interaction model.

2.2.1 Backbone curve

The backbone curve shows the relation between soil resistance per unit length and penetration depth for the first time. The backbone curve is typically governed by equation (2.1).

Fig. 2.3 Full scale harbor test

$$Q_U = B(N_C S_U + \gamma Z) \tag{2.1}$$

Where,

Q_U Ultimate bearing load per unit length of pipe

B Bearing width of pipe

N_C Non-dimensional shape and depth of factor

S_U Undrained shear strength of soil

γ Submerged unit weight of soil

Z The depth of pipe penetration

2.2.2 Soil resistance

As the pipe moves vertically downward and contacts the soil, the soil presents resistance to the pipe's movement. This soil resistance is determined by soil stiffness. There are three types soil stiffness used for modeling pipe-soil interaction: static stiffness, large displacement dynamic stiffness and small displacement dynamic stiffness, as shown in figure 2.4.

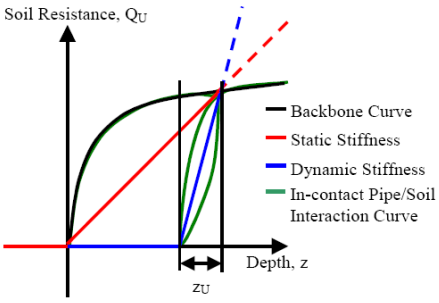


Fig. 2.4 Three soil stiffness

1. Static stiffness

Static stiffness is used for initial penetration into the virgin seabed. The soil reaction force can be calculated using equation (2.2).

$$R_C = m\sqrt{\frac{EI}{H}} \tag{2.2}$$

Where,

R_C Reaction force

m Submerged weight of SCR per unit length

E Elastic modulus

I Moment of inertia

H Tension at the TDP

2. Large displacement dynamic stiffness

Large displacement dynamic stiffness is a secant stiffness which accounts for the initial deformation of the soil. It is used to model the pipe-soil interaction where the pipe breaks out the soil. So the large displacement dynamic stiffness is calculated by equation (2.3).

$$K = \frac{Q_U}{Z_U} = \frac{Q_U}{\Lambda D} \tag{2.3}$$

Where,

Z_U Mobilisation distance, equating ΛD

Λ Non-dimensional parameter coming from the test

3. Small displacement dynamic stiffness

Small displacement dynamic stiffness is a hyperbolic model simulating the in-contact pipe-soil interaction. So the small displacement dynamic stiffness is calculated by equation (2.4).

$$Q = \frac{Z_D}{(1-X)\Lambda D + XZ_D} Q_U \tag{2.4}$$

Where,

Q Reaction force per unit length

Z_D Dynamic displacement, the maximum value is Z_U

X Soil parameter

2.2.3 Soil suction

When the pipe is lifted from the seabed, the soil has suction force to the pipe. Owing to the complex nonlinear behavior and multi-influence factor of soil, the suction model is mainly based on experiment. A soil suction model is based on STRIDE and CARISIMA JIP 's testing as shown in figure 2.5. This soil suction model defines two parameters: maximum soil suction force $Q_{S,MAX}$ and break-out displacement Δ_B .

Maximum soil suction force can be got from equations (2.4)-(2.6)

$$Q_{S,MAX} = K_C \times K_V \times K_T \times N \times D \times S_U \tag{2.4}$$

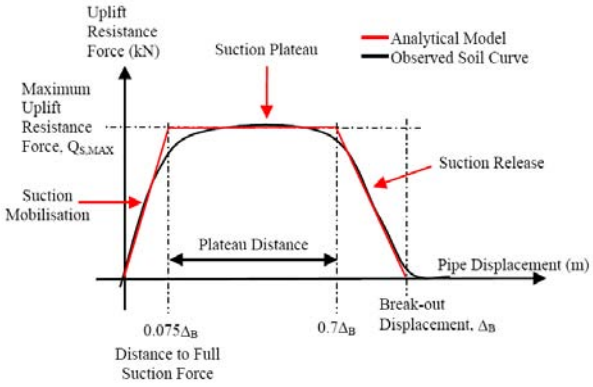


Fig. 2.5 Soil suction model

$$K_V = K_F\left(\frac{V}{D}\right)^{n_F} \tag{2.5}$$

$$K_T = K_{TF}\frac{F_C\sqrt{C_V t}}{LD^2} + C_{TF} \tag{2.6}$$

Where

K_C : Cyclic loading factor

K_V : An empirical pull-out velocity factor

V: Pull-out velocity

F_C : Consolidate force

C_V : Coefficient of consolidation

t : consolidation time

K_F , n_F , K_{TF} , C_{TF} : Empirical constant from test data

Break-out displacement can be got from equations (2.7)-(2.9):

$$\Delta_B = K_{DV} \times K_{DT} \times D \tag{2.7}$$

$$K_{DV} = K_D \times V^{n_D} \tag{2.8}$$

$$K_{DT} = K_{DTF}\frac{F_C\sqrt{C_V t}}{LD^2} + C_{DTF} \tag{2.9}$$

Where

K_{DV} : An empirical break-out displacement factor

K_{DT} : Consolidation time factor

K_D , n_D , K_{DTF} , C_{DTF} : Empirical constant from test data

Other researchers (Aubeny and Biscontin, 2008; Nakhaee and Jun Zhang, 2007) have also done this work and established the non-degradation models. Degradation models are still ongoing in laboratory and have not yet been applied in engineering. The process of degradation models are also developed (Fontaine, et al., 2004; Nakhaee and Jun Zhang, 2009; Hodder and Byrne, 2009).

2.3 Conclusion

Although many pipe-soil interaction models have been established, there is still uncertainty on the mechanism of pipe-soil interaction in TDZ. This part introduces a vertical non-degradation model developed by Bridge, which is used in offshore industry. Soil degradation, trench formation and lateral pipe-soil interaction are the problems that have never been solved. Pipe-soil interaction is the key issue for assessment of fatigue damage to which the mechanism of pipe-soil interaction is of importance.

3. Coupling of hull and riser

In the past, the coupling effect between the hull and mooring/riser was neglected. An uncoupled analysis which only takes the static restoring force of slender structures into account and neglects the inertial force of riser and the hydrodynamic load acting on the riser. In this way, mooring/riser is not coupled with the hull to conduct dynamic computation. But researches showed that such uncoupled analysis of TLPs, spars and FPSOs(Paulling and Webster, 1986;Zhang et al., 2008; Tahar and Kim, 2008) may be inaccurate when used in deepwater. Because as the water depth gets deeper and deeper, the inertia effect increases. So, the interaction effect greatly influences hull and line motions. The complete coupled analysis is necessary for the analysis of riser system.

The main objective of coupled analysis is to give a good estimation of floater motions. Detailed slender structure response such as the riser is secondary. The main procedure of the coupled analyses of hull and mooring/risers is as follows:

1. Establish the equation to describe the motions of hull.
2. Calculate the loads acting on the hull, including the environmental loads.
3. Establish the equation of the moor/risers. Nowadays, a discretization method is used to deal with the equation to make computation more accurate.
4. Calculate the loads acting on the mooring/risers.
5. Analyze the boundary condition and connection point of the whole system, including the connection between the platform and mooring/risers and the interaction between the risers and seabed.
6. Apply the numerical method to solve the coupled equations of hull and mooring/risers.

3.1 Coupled hull and riser models

A floating production system includes three parts: (1) the hull; (2) risers and mooring lines; (3) connections. To establish the coupled function of the system, every part should be modeled from mathematics.

3.1.1 The hull model

The hull is usually taken as a rigid body with 6 degrees of freedom. To simulate its motion, two coordinate systems are needed. One is the space fixed system $\hat{o}\hat{x}\hat{y}\hat{z}$, and the other is the body- fixed system $oxyz$. See figure 3.1.(Chen, 2002)

According to the Newton second law, the equations of the rigid body can be written as:

$$m\vec{a}_g = \vec{F} \tag{3.1}$$

$$I_g \frac{d\vec{\omega}}{dt} + \vec{\omega} \times I_g \vec{\omega} = \vec{M}_g \tag{3.2}$$

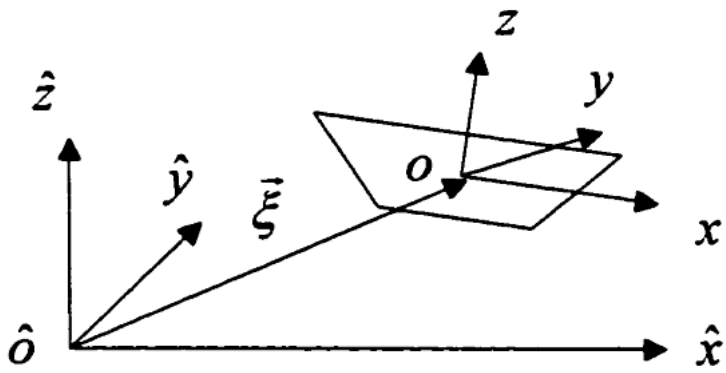


Fig. 3.1 2 coordinate systems of the hull

Where, $\vec{a}_g$ is the acceleration at the center of gravity, I_g is the moment of inertia matrix expressed in the body-fixed coordinate system $oxyz$ and $\vec{\omega}$ is the angular velocity also in the coordinate system $oxyz$. $\vec{F}$ is the resultant applied force and M_g is the resultant applied moments.

To obtain a more specific equation, the acceleration $\vec{a}_g$ can be expressed in the space fixed system $\hat{o}\hat{x}\hat{y}\hat{z}$ as:

$$\vec{a}_g = \vec{a}_o + T^t(\frac{d\vec{\omega}}{dt} \times \vec{r}_g + \vec{\omega} \times (\vec{\omega} \times \vec{r}_g)) \tag{3.3}$$

While the resultant moment in the body fixed system can be written as:

$$\vec{M} = \vec{M}_o - \vec{r}_g \times T\vec{F} \tag{3.4}$$

Inserting equation (3.3) and (3.4) into equation (3.1) and (3.2) respectively, equation (3.1) can be expressed in the space fixed system and equation (3.2) can be expressed in the body fixed system:

$$m\frac{d^2\xi}{dt^2} + mT^t(\frac{d\omega}{dt} \times r_g) + mT^t(\omega \times (\omega \times r_g)) = \hat{F} \tag{3.5}$$

$$I_o\frac{d\omega}{dt} + \omega \times I_o\omega + mr_g \times (T\frac{d^2\xi}{dt^2}) = M_o \tag{3.6}$$

Where, superscript t represents transpose of a matrix.

$\hat{a}_o = \frac{d^2\xi}{dt^2}$, is the acceleration at point o of the body in $\hat{o}\hat{x}\hat{y}\hat{z}$

$\vec{\xi} = (\xi_1, \xi_2, \xi_3)^t$, is the displacement at point o of the body in $\hat{o}\hat{x}\hat{y}\hat{z}$

$\vec{\omega} = (\omega_1, \omega_2, \omega_3)^l$, is the angular velocity of the body in oxyz

$\vec{r}_g = (x_g, y_g, z_g)^l$, is the vector of the center of gravity (mass) of the body in oxyz

I_o is the moment of inertia of the body with respect to point o in oxyz
$\vec{F}$ is the total forces applied on the body in $\hat{o}\hat{x}\hat{y}\hat{z}$
$\vec{M}_o$ is the total moments with respect to point o of the $oxyz$ coordinates
T is a transfer matrix between the body-fixed coordinate system and the space-fixed coordinate system:

$$T = \begin{bmatrix} \cos a_3 \cos a_2 & \sin a_3 \cos a_1 + \cos a_3 \sin a_2 \sin a_1 & \sin a_3 \sin a_1 - \cos a_3 \sin a_2 \cos a_1 \\ -\sin a_3 \cos a_2 & \cos a_3 \cos a_1 - \sin a_3 \sin a_2 \sin a_1 & \cos a_3 \sin a_1 + \sin a_3 \sin a_2 \cos a_1 \\ \sin a_2 & -\cos a_2 \sin a_1 & \cos a_2 \cos a_1 \end{bmatrix}$$

where, a_1, a_2, a_3 are the Euler angles expressing the roll, pitch and yaw motion of the hull.
The forces acting on the hull consist of wind, wave and current loads. To be more specific, the general forces can be written as:

$$F = F_R + F_W + F_{WD} + F_{HS} + F_M + F_{Wind} + F_{Current} \tag{3.7}$$

where, F_R represents radiation forces, F_W exciting forces, F_{WD} wave drift damping forces, F_{HS} hydrostatic restoring forces, F_M mooring/riser/tendon system forces, F_{Wind} wind forces, and $F_{Current}$ is current forces.
The wave forces acting on a floating vessel are well documented in the literature. Corresponding computation methods can be referred to related thesis. For the wind and current forces, the force coefficients are usually needed which come from experiments. If the wind has mean and slowly varying components, a suitable spectrum is in need. Data from the OCIMF is useful for the FPSO analysis.

3.1.2 Riser model

The riser can be taken as a flexible system. It can make large displacement and angle movement with the hull. As a result, the geometric nonlinearity is a big problem for the risers. Here are some calculation methods which are used to solve the problem.

1. Linear spring method. The riser is taken as a linear spring. This method is the main method used to make frequency domain analysis. Because this method neglects too many details of the riser, the result becomes less and less accurate with the water depth increases.
2. Catenary method. This method uses the catenary equation to transform the 3-dimention to 2-dimention and fulfill the static conditions. As in method (1), the dynamic details of risers and bending stiffness are neglected.
3. Slender rod theory. Risers can be modeled by using the elastic slender rod theory. It can analyze the condition of the riser after deformation in the original coordinate system. The position of every element is expressed by a vector. Higher-order function is used to simulate the deformation of rod. But because the effect of torsion is neglected, non-vector problem of the large rotation angle deformation is not taken into account.

Method (3) is the most widely used in the offshore industry, and is illustrated in detail as follows.(Garrett, 1982; Garrett 2005)
The equation of slender rod can be written as:

$$-(Br'')'' + (\lambda r')' + q = \rho \ddot{r} \tag{3.8}$$

If the stretch of rod is assumed to be linear and small, the inextensibility condition can be approximated as:

$$\frac{1}{2}(r'\bullet r'-1)=\frac{T-T_0}{AE} \tag{3.9}$$

Where, B is the bending stiffness (EI)
r : the vector represents the centerline of the rod
$\lambda = T - Br''r''$: Lagrange multiplier
q : the applied load(weight, drag, etc)
T : the tension
T_0 : the unstretched tension
AE : the axial stiffness
The external forces applied on the riser include gravity, hydrostatic and hydrodynamic forces.
The gravity can be expressed as a distributed load as:

$$q_t(s,t)=-\rho_t g A_t e_y \tag{3.10}$$

Where, ρ_t is riser density, A_t is section area.
The hydrodynamic forces consist of added-mass force, drag force and Froude-Krylov force. The first two forces can be predicted by Morison's equation that will be discussed in section 4. Froude-Krylov force due to sea water outside the riser is:

$$\vec{q}_f^{F-k}(s,t)=\rho_f(g\vec{e}_y+\ddot{\vec{u}})A_f+(P_f A_f \vec{r}')' \tag{3.11}$$

Where,
P_f is the pressure of sea water
A_f is the outer cross-section area of the riser
ρ_f is the mass density of the sea water
Due to the internal fluid, Froude-Krylov force for the riser is:

$$\vec{q}_i^{F-K}(s,t)=-\rho_i g A_i \vec{e}_y-(P_i A_i \vec{r}')' \tag{3.12}$$

Where, P_i is the pressure from the internal fluid; ρ_i is the mass density of the internal fluid. A_i is the inner area of riser.

3.1.3 The connection

The connection between the riser and platform can be established by using spring and damper. A linear spring can define the translational motion between platform and the top of riser. A rotational spring can define the relation between the rotation of platform and the tangential direction of the riser. For the spring, one end is at the rod element, the other one is on the rigid body. These assumptions of the connections can easily couple the motion of platform and risers numerically and simulate different kinds of connections by using different stiffness of the spring.
With the assumption of small-angle rotation, the connector forces (Q_i) and moments (L_i) of the spring on the end node are written as:(Arcandra, 2001)

$$Q_i=K_i^L(X_i+p_i+\theta_j\times C_{ji}-r_i) \tag{3.13}$$

$$L_i = K^{\theta}(e_i + \theta_j D_{ji} - \frac{r'_i}{(r'_m r'_m)^{0.5}} - \frac{r_i' r_j'}{(r_n' r_n')^{3/2}}) \tag{3.14}$$

Where, K_i^L and K^{θ} are the stiffness matrix of linear and rotation spring
X_i is the translational motion of the rigid body
θ_i is the rotational motion of the rigid body
p_i is position vector of the point on the platform where the springs are attached
r_i is the position vector of the node of the mooring/riser which is attached by the spring
C_{ji} and D_{ji} are defined to make it easier to do the numerical calculation with the position vector p_i and the unit vector e_i as:

$$[C] = \begin{bmatrix} 0 & -p_3 & p_2 \\ -p_3 & 0 & -p_1 \\ p_2 & -p_1 & 0 \end{bmatrix}$$

$$[D] = \begin{bmatrix} 0 & -e_3 & e_2 \\ -e_3 & 0 & -e_1 \\ e_2 & -e_1 & 0 \end{bmatrix}$$

The resultant force F_i^S and moment M_i^S transferred to the body are defined as follows:

$$F_i^S = -Q_i^S \tag{3.15}$$

$$\begin{aligned} M_i^S &= M_i^L + M_i^{\theta} \\ &= Q_k^S C_{ki} + L_k^S D_{ki} \end{aligned} \tag{3.16}$$

Where, $M_i^L = Q_k^S \times p_j$ is the moment resulting from the linear spring, and $M_i^{\theta} = L_k^S \times e_j$ is the moment resulting from the rotational spring. The force F_i^S and the moment M_i^S act on the body.
The damper is used to control the excessive resonance of the high frequency vibration of the tensioned line like the tether or the riser in the TLP. The damper is simulated by using linear damping force proportional to the vibratory velocity of the line on the top connection node of the hull and the mooring/riser. The damping force N_i^D on the connection node of the line is given by (Ran, 2000):

$$N_i^D = C_d\left(\dot{X}_i + \dot{\theta}_j C_{ji} - \dot{r}_i\right) \tag{3.17}$$

Where C_d is the damping coefficient, $\dot{X}$ and $\dot{\theta}$ are the translational and rotational velocity of the rigid body, $\dot{r}$ is the velocity of the attached node of the line to the body.
The force acts on the rigid body as reaction force by:

$$F_i^D = -N_i^D \tag{3.18}$$

3.2 Integrated hull and riser equations

Compared with time domain analysis, frequency domain analysis can demonstrate directly the wave response of platform to a certain degree. So it is widely used in the theory

computation. But in the frequency domain analysis, all nonlinearities must be linearized which may require a perturbation assumption. The time domain analysis can confirm influence of current and previous course on the mechanical state of riser at every time step. As a result to deal with all the nonlinearities in the coupled analysis, time domain is obviously the most effective way in the practical project.
The assembled equation of the coupled system of the rigid body and mooring/risers can be expressed as (HARP manuals):

$$\begin{bmatrix} [\,K^L\,] & [K^C] \\ \text{----------} & \\ [(K^C)^T] & [K^B] \end{bmatrix} \begin{Bmatrix} U^L \\ \text{---} \\ U^B \end{Bmatrix} = \begin{Bmatrix} F^L \\ \text{---} \\ F^B \end{Bmatrix} \tag{3.19}$$

where $[K^L]$ is the stiffness matrix of the mooring/riser and the connector springs which has $n_L \times [8 \times (n_E + 1) - 1]$ rows where n_L is the total number of lines, $[K^B]$ is the stiffness matrix of the rigid body, $[K^C]$ and $[(K^C)^T]$ are the coupled stiffness matrices and its transpose matrix including the coupling terms of the rigid body and the mooring/riser. $[U^L]$ and $[U^B]$ are the displacement matrices of the lines and the body, $[F^L]$ and $[F^B]$ are the force and moment terms acting on the lines and the body. The size of $[K^B]$ is 6×6. $[K^C]$ has the size of $[8 \times (n_E + 1) - 1]$ rows and 6 columns per line. n_E is the number of elements per line.
To calculate the terms of stiffness matrix, further analysis is required. And the time domain analysis consists of static analysis and dynamic analysis.

3.2.1 Static analysis

Before a dynamic analysis is conducted, the static problem should be solved first. Considering the geometric nonlinearity of the riser and moorings, the static analysis should be solved iteratively (Low and Langly, 2006). Here the Newton-Raphson method is applied. The hull, mooring/riser and connection are combined to form the stiffness matrix in the equation (3.19).
At each iteration step, the coupled assembly system equations are solved to obtain the behaviors for the body and lines simultaneously, and the iteration continues until the norms of the solutions reach a specified tolerance.

3.2.2 Time domain analysis

The time domain analysis requires integration. Several implicit methods have been developed to do it, for example, Newmark-B method, Runge-Kuta method and the Adams-Moulton method. Because the last one can solve the coupled equations of the hull and mooring/riser at every time step, it is used to integrate the nonlinear force. In this way, the terms of the stiffness matrix can be obtained from the integration.

3.3 Conclusion

The coupled effect between hull and risers is usually included in the whole coupled analysis of floating production system. The whole system mainly has two components: the hull and mooring/risers. Meanwhile, the connection between platform and mooring/riser and the interaction between riser and seabed also play a part in the coupled analysis. At present, the hull is described as a 6 DOF rigid body. Risers are modeled by using a finite element representation of an elastic rod. The connection can use a combination of springs and dampers. At last, a static analysis and dynamic analysis are conducted.

4. Fatigue

The assessment of fatigue damage is important in SCR design. Cyclic loading can cause the fatigue damage of SCR, including wave, vortex induced vibration (VIV), hull motion, etc. This Section mainly discusses the fatigue induced by wave and VIV, and introduces fatigue assessment methods.

4.1 Wave loading

The method of calculating wave load on cylinder is chosen according to the size of cylinder and wavelength. SCR is a slender structure, and it mainly uses Morison's equation which has been presented by Morison etc (Morison, et al., 1950; Burrows et al., 1997). The basic assumption of this method is that the diameter of cylinder D compared with wavelength L is small, D/L<0.2. The cylinder cannot affect the wave field.

One cylinder stands vertically on the seabed, and the depth of water is d. Wave height is H and spreads along the coordinate x. The intersection point coordinate of cylinder axis and seabed is (x, z), as shown in figure 4.1. Morison etc assumed that the horizontal wave force f_H acting on cylinder at any height z consists of two parts: Water particle horizontal velocity u_x induced force-drag force f_D; Water particle horizontal acceleration $\dot{u}_x$ induced force-inertia force f_I.

Morison's equation for wave loading on unit height of cylinder at any height z:

$$\begin{aligned} f_H &= f_D + f_I \\ &= \frac{1}{2}\mathrm{C_D}\rho A u_x |u_x| + \rho V_0 \frac{du_x}{dt} + C_m \rho V_0 \frac{du_x}{dt} \\ &= \frac{1}{2}\mathrm{C_D}\rho A u_x |u_x| + C_M \rho V_0 \frac{du_x}{dt} \end{aligned} \tag{4.1}$$

Where, u_x is the water particle velocity at the axis center of cylinder

$\frac{du_x}{dt}$ is the water particle acceleration at the axis center of cylinder

A is the projected area of unit height of cylinder normal to the direction of wave:

V_0 is the tonnage of unit height of cylinder

ρ is seawater density

C_m is an added mass coefficient

C_M is mass coefficient

$\mathrm{C_D}$ is drag force coefficient

For a cylinder, equation (4.1) can be written as following:

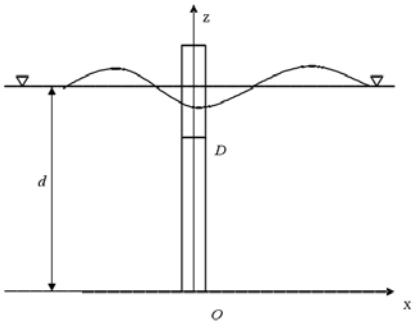


Fig. 4.1 Small scale straight cylinder coordinate system

$$f_H = \frac{1}{2}\mathrm{C_D}\rho D u_x |u_x| + C_M \rho \frac{\pi D^2}{4}\frac{du_x}{dt} \tag{4.2}$$

Equation (4.2) is applied for the fixed cylinder. If the cylinder moves under wave loading, the equation (4.2) should be written as following.

$$f_H = \frac{1}{2}C_D \rho D(u_x - \dot{x})|u_x - \dot{x}| + C_M \rho \frac{\pi D^2}{4}\frac{\partial u_x}{\partial t} - C_m \rho \frac{\pi D^2}{4}\ddot{x} \tag{4.3}$$

Where, x is the horizontal displacement at z
$\dot{x}$ is the horizontal velocity at z
$\ddot{x}$ is the horizontal acceleration at z
Equation (4.2) and equation (4.3) are used for calculating the wave force of straight cylinder. But SCR is slant cylinder, and equation (4.2) must be modified. Assuming two-dimensional wave spreads along the direction x. At any point of cylinder, U_n is orthogonal component of water particle velocity and U_t is the tangent component of water particle velocity. $\dot{U}_n$ and $\dot{U}_t$ are corresponding acceleration. As a two-dimensional problem, the direction of velocity and acceleration are general not in the same line. So the Morison's equation must be written as vector. Equation (4.4) is Morison's equation for slant cylinder.

$$f = \frac{1}{2}C_D \rho D U_n |U_n| + C_M \rho \frac{\pi D^2}{4}\dot{U}_n \tag{4.4}$$

Where, f is the vector of wave force at z .
The projection of U_n to three coordinate is $\{U_x, U_y, U_z\}$. Assuming e is the unit vector along the axial of cylinder.

$$e = e_x i + e_y j + e_z k \tag{4.5}$$

The parameters used in Morison's equation are as following:

$$\begin{cases} U_x = u_x - e_x(e_x u_x + e_z u_z) \\ U_y = -e_y(e_x u_x + e_z u_z) \\ U_z = u_z - e_z(e_x u_x + e_z u_z) \end{cases} \tag{4.6}$$

$$\begin{cases} \dot{U}_x = (1-e_x^2)\dfrac{\partial u_x}{\partial t} - e_z e_x \dfrac{\partial u_z}{\partial t} \\ \dot{U}_y = -e_x e_y \dfrac{\partial u_x}{\partial t} - e_z e_y \dfrac{\partial u_z}{\partial t} \\ \dot{U}_z = -e_x e_z \dfrac{\partial u_x}{\partial t} + (1-e_z^2)\dfrac{\partial u_z}{\partial t} \end{cases} \tag{4.7}$$

$$|U_n| = (u_x^2 + u_z^2 - (e_x u_x + e_z u_z)^2)^{1/2} \tag{4.8}$$

If slant cylinder is moving under wave load, it similar to equation (4.3).

Drag force coefficient C_D and mass coefficient C_M are empirical coefficients. They are from the tests, and the value of them are also depends on wave theory. The value of C_D and C_M can refer relevant standards or criterion of Classification Societies. Morison equation is the main method for calculating the wave load on small size cylinder.

4.2 VIV

If a cylindrical structure is subjected to the current, alternately vortex shedding will be taking place at each side of the cylinder. Vortex shedding will produce periodic force which makes cylinder vibration. When the frequency of vortex shedding is close to the nature frequency of cylinder, cylinder generates resonate. Despite decades of intensive numerical simulation and experiment research, there is not a model that can accurately simulate VIV (Williamson and Govardhan, 2008). The problem of VIV is how to determine the hydrodynamic force. Hydrodynamic force is from experiment or computational fluid dynamics (CFD), and VIV response prediction model can be classified into two categories: empirical model and CFD model.

4.2.1 Basic concept

VIV is affected by some parameters, such as Reynolds number, Strouhal number etc (Pan et al., 2005; Klamo, 2007). Detailed explanations of these parameters are as follows.

1. Reynolds number

The Reynolds number is a ratio of inertial force to viscous force. Its value affects the type of flow as the fluid goes through the bluff body, as depicted in figure 4.2.

$$\mathrm{Re} = \frac{UD}{v} \tag{4.9}$$

Where,

U Fluid velocity

v Coefficient of kinematic viscosity

D The outer diameter of cylinder

2. Strouhal number

Strouhal number depends on the Reynolds number, which comes from experiments. Equation (4.10) is used to calculate the vortex shedding frequency.

$$St = \frac{f_v D}{U} \tag{4.10}$$

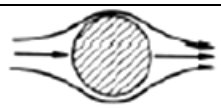
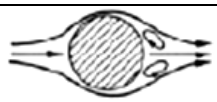
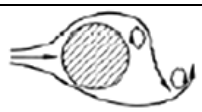
	$R_e < 5$	No separation phenomenon
	$(5 \sim 15) \le R_e < 40$	A pair of fixed small eddy behind cylinder
	$40 \le R_e < 150$	Periodic vortex shedding in laminar flow
	$300 \le R_e < 3\times10^5$	Periodic vortex shedding in turbulent flow, which can be extended to 50D (cylinder outer diameter)
	$3\times10^5 \le R_e < 3.5\times10^6$	Transition section. The separation point is backward and vortex shedding is aperiodic
	$3.5\times10^6 \le R_e$	Recovering periodic vortex shedding in turbulent flow

Fig. 4.2 The relation between Re and vortex shedding

Where

f_v Vortex shedding frequency

3. Reduced velocity

Reduced velocity is convenient for measuring the flow velocity. Reduced velocity is defined as:

$$U_r = \frac{U}{f_n D} \tag{4.11}$$

f_n The nature frequency of cylinder

Other parameters, such as mass ratio, damping factor and aspect ratio etc are not introduced here.

4.2.2 Empirical model

VIV experiments involve forced or vibration of a cylinder. The empirical models are different varying with experiments. It can be divided into time domain models and frequency domain models.

Time domain model

A well known time domain model is wake oscillators which satisfy van der Pol or Rayleigh equation. Wake oscillator model couples with structure vibration and fluid wake oscillate.

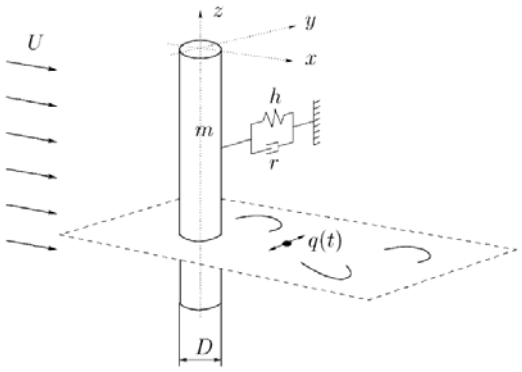


Fig. 4.3 Wake oscillator model

The 2D wake oscillator model is assumed to be an elastically supported cylinder, as shown in figure 4.3. The coupled structure and wake oscillator are described by (Facchinetti et al., 2003; Xu et al., 2010):

$$\begin{aligned} &\ddot{y} + (2\xi\delta + \frac{\gamma}{\mu})\dot{y} + \delta^2 y = s \\ &\ddot{q} + \varepsilon(q^2 - 1)\dot{q} + q = f \end{aligned} \tag{4.12}$$

Where, $\delta = \dfrac{\Omega_S}{2\pi St(U/D)}$, $s = S\dfrac{D}{4\pi^2 St^2 U^2 m}$, $f = F\dfrac{D}{4\pi^2 St^2 U^2}$,

The parameters of wake oscillator equations are explained as follows:

$y : y = Y/D : Y$ is cross-flow displacement, D is outer diameter.

q : Dimensionless wake variable.

St : Strouhal number.

U : Fluid velocity.

m : Mass accounting the mass of structure and fluid-added mass.

ξ : Structure reduced damping.

Ω_S : Structural angular frequency.

γ : Stall parameter.

Frequency domain model

A famous VIV analysis software Shear7 is frequency domain tool, which is based on mode-superposition method (Vandiver and Li, 2005). Its theoretical background is briefly described in the section. The governing equation is given by:

$$m_t\ddot{y} + R\dot{y} - Ty'' = P(x,t) \tag{4.13}$$

Where, m_t is mass per unit length, $\ddot{y}$ is the acceleration of the structure, R is the damping per unit length. $\dot{y}$ is the velocity of the structure, T is the tension, y'' is the second derivative of the displacement of the structure with respect to the spatial variable. $P(x,t)$ is the excitation force per unit length.

The structure displacement response can be written as the modal superposition.

$$y(x,t) = \sum_r Y_r(x) q_r(t) \tag{4.14}$$

Where, $Y_r(x)$ is the rth mode shape of the structure. Substituting equation (4.14) into equation (4.13) and performing modal analysis:

$$M_r \ddot{q}_r + R_r \dot{q}_r(t) + K_r q_r(t) = P_r(t) \tag{4.15}$$

Where, M_r is modal mass, R_r is modal damping, K_r is modal stiffness, $P_r(t)$ is modal force. For pined-pined beam with varying tension, the nth natural frequency is defined by:

$$\int_0^L \sqrt{-\frac{1}{2}\frac{T(s)}{EI(s)} + \frac{1}{2}\sqrt{(\frac{T(s)}{EI(s)})^2 + 4\frac{m_t(s)\omega_n^2}{EI(s)}}}\, ds = n\pi \tag{4.16}$$

Where, $T(s)$ is tension, $EI(s)$ is the bending stiffness, $m_t(s)$ is the mass per unit lengh, ω_n^2 is the nth natural frequency of the structure.
nth mode shape is written as:

$$Y_n(x) = \sin(\int_0^L \sqrt{-\frac{1}{2}\frac{T(s)}{EI(s)} + \frac{1}{2}\sqrt{(\frac{T(s)}{EI(s)})^2 + 4\frac{m_t(s)\omega_n^2}{EI(s)}}}\, ds) \tag{4.17}$$

4.2.3 CFD model

The CFD method obtains hydrodynamic force by solving Navier-Stokes equations directly. Laminar flow can be solved by CFD model that is identified by testing. The key issue is the establishment of turbulence model. There are 3 primary methods including Reynolds averaged Navier-Strokes (RANS), Large Eddy Simulation (LES) and direct numerical simulation (DNS). The framework of turbulent flow simulation methods is presented in figure 4.4. This section only introduces the RANS model and LES model commonly used in offshore industry.

1. RANS model

In order to solve the details of turbulent fluctuations, time-average method is usually applied. So RANS is coming up. RANS does not solve directly instantaneous Navier--Stokes equations. It decomposes flow velocity into 2 components: mean flow and fluctuating component. Time-average Navier-Strokes equations can be written as equation (4.18) (Pan et al., 2007). It describes incompressible fluid. Comparing to DNS and LES, RANS is less time-consuming and widely used in commercial work.

$$\begin{cases} \dfrac{\partial u_i}{\partial x_i} = 0 \\ \dfrac{\partial}{\partial t}(\rho u_i) + \dfrac{\partial}{\partial x_j}(\rho u_i u_j) = -\dfrac{\partial p}{\partial x_i} + \dfrac{\partial}{\partial x_j}(\mu S_{ij} - \rho\overline{u_i' u_j'}) \end{cases} \tag{4.18}$$

Where,

u , p : Time-average velocity and pressure

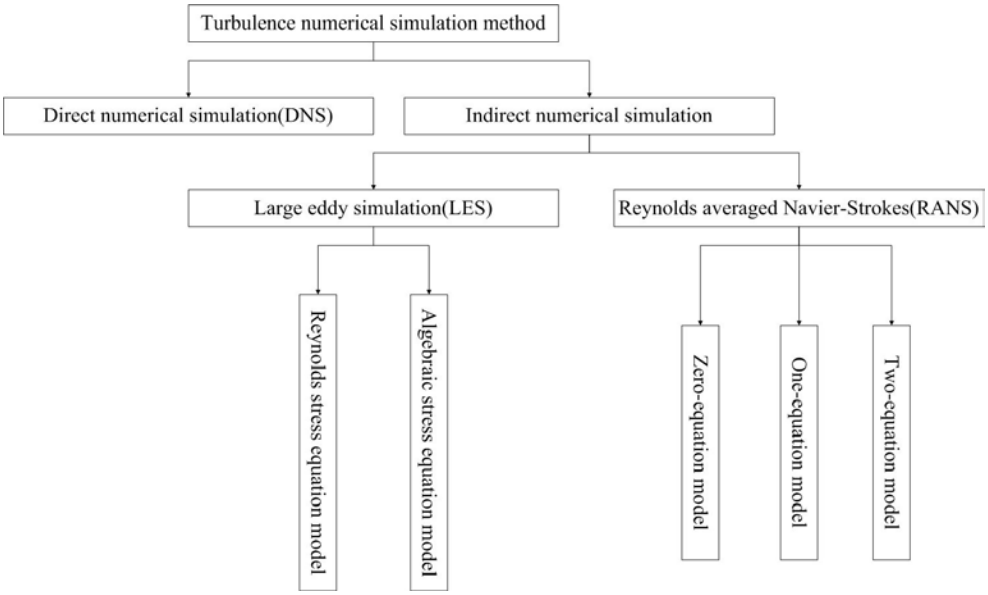


Fig. 4.4 turbulent models

μ : Molecular viscosity
S_{ij} : Mean stress tensor
ρ : Fluid density
The time-average equation (4.18) has addition item: $-\rho\overline{u_i'u_j'}$, which is called Reynolds stresses, namely:

$$\tau_{ij} = -\rho\overline{u_i'u_j'} \tag{4.19}$$

Reynolds stresses is a new unknown item. According to different assumptions, RANS can be mainly divided into 2 types: Reynolds stresses model and eddy viscosity model, as shown in figure 4.4. Recently two-equation model is extensively used in engineering. The basic two-equation model is standard $k-\varepsilon$ model, which brings in k (turbulence kinetic energy) and ε (turbulent dissipation rate) (Dixon and charlesworth, 2006).
Eddy viscosity model indirectly solve Reynolds stress. It is expressed as a function of turbulent viscosity. The key issue of this method is the calculation of turbulent viscosity. Turbulent viscosity is represented via the Boussinesq assumption. This assumption establishes the relationship between Reynolds stress and mean gradients of the velocity:

$$-\rho\overline{u_i'u_j'} = \mu_t\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)-\frac{2}{3}\left(\rho k+\mu_t\frac{\partial u_i}{\partial x_i}\right)\delta_{ij} \tag{4.20}$$

Where, μ_t : turbulent viscosity, u_i : mean velocity, δ_{ij} : "Kronecker delta" symbol (i=j, δ_{ij} =1; $i \neq j$, δ_{ij} =0), k : turbulent kinetic energy.

$$k = \frac{\overline{u_i' u_i'}}{2} = \frac{1}{2}(\overline{u'^2} + \overline{v'^2} + \overline{w'^2}) \tag{4.21}$$

Adding turbulent dissipation rate equation, it establishes the standard $k-\varepsilon$ model.

$$\varepsilon = \frac{\mu}{\rho}\overline{\left(\frac{\partial u_i'}{\partial x_k}\right)\left(\frac{\partial u_i'}{\partial x_k}\right)} \tag{4.22}$$

In standard $k-\varepsilon$ model, the transport equations about incompressible fluid are as follows:

$$\begin{cases} \dfrac{\partial(\rho k)}{\partial t} + \dfrac{\partial(\rho k u_i)}{\partial x_i} = \dfrac{\partial}{\partial x_i}\left[\left(\mu + \dfrac{\mu_t}{\sigma_k}\right)\dfrac{\partial k}{\partial x_j}\right] + G_k - \rho\varepsilon \\ \dfrac{\partial(\rho\varepsilon)}{\partial t} + \dfrac{\partial(\rho\varepsilon u_i)}{\partial x_i} = \dfrac{\partial}{\partial x_i}\left[\left(\mu + \dfrac{\mu_t}{\sigma_\varepsilon}\right)\dfrac{\partial \varepsilon}{\partial x_j}\right] + \dfrac{C_{1\varepsilon}}{k}G_k - C_{2\varepsilon}\rho\dfrac{\varepsilon^2}{k} \end{cases} \tag{4.23}$$

Where, G_k: generation of turbulence kinetic energy due to the mean velocity gradients. $C_{1\varepsilon}$ and $C_{2\varepsilon}$ are constant. σ_k and σ_ε are turbulence Prandtl numbers.

2. LES model

LES directly solves the instantaneous Navier-Strokes equation. Unlike DNS, LES resolves only large eddies directly, while small eddies are modeled. The effect of small eddies is represented by adding an item to LES equations. So, LES can use much coarser grid than DNS. Commercial software Fluent has a LES module for user (Fluent User's Guide).

LES governing equations are obtained by using filtered function that effectively filters out the small eddies. A filter function is defined by:

$$\overline{\phi} = \int_D \phi G(x, x')dx' \tag{4.24}$$

Where D is fluid domain, $G(x,x')$ is the filter function that determines the scale of eddies. $G(x,x')$ has variable expressions. For finite volume method, the filter function can be defined by:

$$G(x,x') = \begin{cases} 1/V, x' \in v \\ 0, x' \notin v \end{cases} \tag{4.25}$$

Where V is the volume of computational cell. So equation (4.24) can be written as followed:

$$\overline{\phi} = \frac{1}{V}\int_D \phi dx' \tag{4.26}$$

LES equations for incompressible fluid are obtained by using equation (4.26) to filter the Navier-Strokes equations:

$$\begin{cases} \dfrac{\partial \overline{u_i}}{\partial x_i} = 0 \\ \dfrac{\partial}{\partial}\left(\rho\overline{u_i}\right) + \dfrac{\partial}{\partial}\left(\rho\overline{u_i}\,\overline{u_j}\right) = \dfrac{\partial}{\partial x_j}\left(\mu\dfrac{\partial \delta_{ij}}{\partial x_j}\right) - \dfrac{\partial \overline{p}}{\partial x_j} - \dfrac{\partial \tau_{ij}}{\partial x_j} \end{cases} \tag{4.27}$$

Where τ_{ij} is subgrid-scale stress which is defined by equation (4.28). δ_{ij} is stress tensor.

$$\tau_{ij} = \rho\overline{u_i u_j} - \rho\overline{u_i}\,\overline{u_j} \tag{4.28}$$

The Subgrid-scale stress also complies the Boussinesq hypothesis:

$$\tau_{ij} - \frac{1}{3}\tau_{kk}\delta_{ij} = -2\mu_t\overline{S_{ij}} \tag{4.29}$$

Where μ_t is subgrid-scale turbulent viscosity. $\overline{S_{ij}}$ is the rate-of-strain tensor. In Smagorinsky-Lilly model, μ_t and $\overline{S_{ij}}$ are defined by:

$$\mu_t = \rho L_s^2 \left|\overline{S}\right| \tag{4.30}$$

$$\overline{S_{ij}} = \frac{1}{2}\left(\frac{\partial\overline{u_i}}{\partial x_i} + \frac{\partial\overline{u_i}}{\partial x_j}\right) \tag{4.31}$$

Where, $\left|\overline{S}\right| = \sqrt{2\overline{S_{ij}}\,\overline{S_{ij}}}$, $L_s = \min\left(\kappa d, C_s V^{\frac{1}{3}}\right)$

κ is von karman constant, d is the distance to the closest wall, C_s is Smagorinsky constant.

4.3 Fatigue assessment methods

Cyclic loading can cause the fatigue damage of SCR. The S-N curve and linear cumulative damage law (Palmrgen-Miner rule) are used to calculate the fatigue life of SCR in offshore industry (DNV-RP-C203, 2010).

The basic S-N curve is defined as:

$$\log N = \log a - m\log\Delta\sigma \tag{4.32}$$

Where N is predicted number of cycles to failure for stress range $\Delta\sigma$. m is negative inverse slope of S-N curve. $\log\bar{a}$ is intercept of long N-axis by S-N curve.

The fatigue strength of welded joints is partly dependent on plate thickness. The S-N curve accounting for the effect of thickness is defined by:

$$\log N = \log\bar{a} - m\log\left(\Delta\sigma\left(\frac{t}{t_{ref}}\right)^k\right) \tag{4.33}$$

Where t_{ref} is reference thickness. t is thickness through which a crack will mostly grow. k is thickness exponent on fatigue strength.

According to Palmrgen-Miner rule, the fatigue damage is calculated by (ABS, 2003):

$$D = \sum_{i=1}^{k}\frac{n_i}{N_i} \le \eta \tag{4.34}$$

Where D is accumulated fatigue damage. n_i is the number of stress cycles stress block i. η is allowable damage ratio.

The value of parameters in Equation (4.26) can refer to the standards published by each Classification Society, such as DNV, ABS and CCS.

4.4 Conclusion

This section introduces the factors inducing SCR fatigue and mainly introduces the calculation methods of wave force and VIV that are commonly used in offshore industry. Wave force on SCR is obtained by Morsion's equation, which is widely used in calculation of wave loading on slender structure. VIV prediction model can be classified into empirical model and CFD model. Empirical models have been developed for past decades. Commercial software is developed, such as Shear7, VIVA and VIVNA. With the development of computer technology and increased storage, CFD application is growing faster recently. Although the decades of research on VIV and wave induced fatigue, there are uncertainties about the prediction of SCR fatigue life. It is usually to select larger safety coefficient in SCR design.

5. SCR installation

SCR installation needs to consider the effects of vessel's motion, wave, current and the interface with other structures. Some factors are discussed above. So this section only introduces the SCR installation methods and equipments. Giving a mechanical model of SCR can be used in installation.

5.1 The introduction of SCR installation

SCR installation is a procedure that connects to the pipe laying. Three pipe laying methods are S-lay, J-lay and Reel-lay. These three pipelay ships are shown in figure 5.1. And some SCR installation equipments are shown in figure 5.2, including tensioner, A&R winch and integrated riser lifting system (Mao et al. , 2010; Duan et al. , 2011).

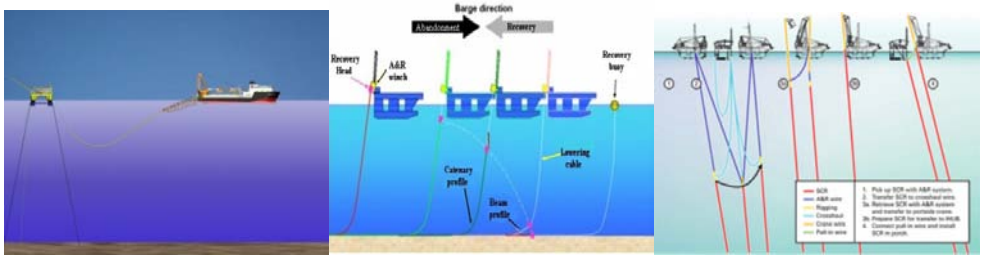
(a) S-lay ship (b) J-lay ship (c) Reel-lay ship

Fig. 5.1 Three pipelay ships

The installation method can be classified into 2 broad categories: 1st installation and 2nd installation.1st installation is that the pipelay ship starts pipe laying from the platform and lays in the opposite direction to platform, as shown in figure 5.3 (a). 2nd installation has 2 methods: pre-lay and post-lay. Pre-lay method means laying the pipe on the seabed after pipe laying is finished, while the platform is not in position. Then installation vessel lifts the pipe from the seabed and installs the pipe on the platform when the platform is in position. Post-lay method (Smith, 2007) means directly install the pipe on the platform after the pipelay ship finished pipe laying when the platform is in position. Figure 5.3(b) and (c) show the procedure of pre-lay and post-lay.

(a) tensioner (b) A&R winch (c) Integrated riser lifting system

Fig. 5.2 SCR installation equipments



(a) 1st installation (b) pre-lay method (c) post-lay method

Fig. 5.3 SCR installation methods

5.2 Mechanical model of SCR

SCR installation need to consider some factors: installation vessel's movement, environment loadings and the interface with other offshore structures etc. This section only introduces a model of SCR during installation due to the other factors mainly discussed in above sections.

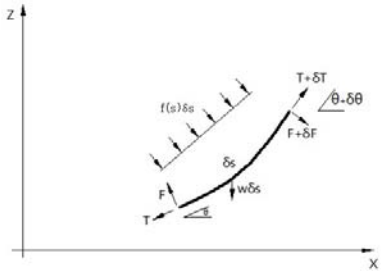


Fig. 5.4 The forces acting on a short segment of SCR

SCR is a large deformation slender structure during installation. It belongs to large deformation and small strain problem. Slender rod and straight beam method can't model catenary riser. This section introduces a Large-Angle deflection beam model (Sparks, 2007). Figure 5.4 shows the forces acting on a short segment of SCR. According to the force balance, the differential equations can be deduced:

$$\begin{cases} EI\dfrac{d^3\theta}{ds^3} - T\dfrac{d\theta}{ds} + w\cos\theta + f(s) = 0 \\ \dfrac{dT}{ds} = w\sin\theta \end{cases} \tag{5.1}$$

Where, $f(s)$ is external loading. EI is the constant bending stiffness. θ is a angle measured from horizontal. T is axial tension.

5.3 Conclusion

This section introduces SCR installation methods and equipments. Giving a mechanical model of SCR can be used in SCR installation. The factors affecting SCR installation are discussed in the above sections.

6. Conclusion

SCR plays an important role in connecting the floating structure and submarine production facilities. SCR has advantages of low manufacturing cost, resistance of high temperature and high pressure, good adaptability of upper floating body's motion, etc. SCR numerical simulation has made great advances through the decades of research. Some commercial software was developed for SCR design. Despite decades of researches, there are uncertainties on mechanical characteristic of SCR. These are challenges for SCR design and future endeavor is suggested as follows:

1. Pipe-soil interaction mechanism. Pipe-soil interaction models are obtained from experiments. Empirical models have been established. Non-degradation models have been applied in SCR strength analysis and the prediction of fatigue damage. From the field observation, it will develop a trench beneath the SCR under cyclic loading. The development of trench is a difficulty. The formation of trench is affected by complex factors, such as soil plastic deformation, water mixed, soil reconsolidation and soil erosion. Only infering the trench shape will not reach to precise calculation results.
2. Turbulence is recognized as a difficult problem in the world. Turbulent flows exist in our lives: Smoke comes from a chimney, water flows in a river and a strong wind strikes the structure. Turbulent flow is unsteady, irregular, random and chaotic. So, despite decades of intensive research, there are no models that can accurately simulate the turbulent flow. The challenge of VIV of SCR is the calculation of hydrodynamic force under turbulent flow. Empirical models have been used for SCR design. Some commercial software has been developed, such as Shear7, VIVA. Another approach is CFD technology. In theory, hydrodynamic force can be accurately calculated by directly solving Navier-Strokes equations. This method is also called Direct Numerical Simulation (DNS). But DNS requires very refinement grid and micro time step, and the requirement of storage and computation speed is very high. So RANS and LES are widely applied in engineering. The future works should be done to establish more exact model.
3. With the increase of water depth, the coupled effects between hull and riser cannot be neglected. Any uncoupled or simple coupled analysis can't reflect the coupling effect. Now there exists systematic method to conduct the couple analysis. But how to deal with the nonlinear wave load and geometric nonlinearity of the slender structure is still a problem. Although frequency domain analysis can demonstrate the wave response of

platform to a certain degree and its calculation efficiency is relative high, time domain analysis is the better way to reflect all the nonlinearities. As a result, some software whose main task is the coupled analysis such as the HARP and Cable3D are based on time domain analysis.

7. Acknowledgement

This chapter is financially supported by the National 863 Program of China (granted number 2006AA09A105), and funded by the National Natural Science Foundation of China (granted number 50979113). Sincerely thanks go to the colleagues in the COOEC Ltd. and in the Offshore Oil/Gas Research Centre of China University of Petroleum, who are involved in this wide range of researches.

8. References

[1] Yong Bai and Qiang Bai. Subsea Pipelines and Risers. Elsevier Science Ltd, 2005.
[2] Rafik Boubenider. Effectiveness of Polyethylene Helical Strakes in Suppressing VIV Responses after Sustaining High Roller Load Deformation during S-Lay Installation. Offshore Technology Conference, 2008.
[3] Samuel Taggart and Michael A.Tognarelli. Offshore Drilling Riser VIV Suppression Devices-What's Available to Operators. Offshore Mechanics and Arctic Engineering, 2008.
[4] W.K.Kavanagh, K.R.Farnsworth and P.G.Griffin. Matterhorn Steel Catenary Riser: Critical Issues and Lessons Learned for Reel-Layed SCRs to a TLP. Offshore Technology Conference, 2004.
[5] Abaqus Analysis User's Manual 6.9.
[6] ANSYS ASAS Brochre.
[7] Orcaflex Manual Version 9.3c.
[8] Ali Nakhaee. Study of the fatigue fatigue life of steel catenary risers in interaction with the seabed. Ph.D dissertation of Texas A&M University, 2010.
[9] M.S.Hodder and B.W.Byrne. 3D experiments investigating the interaction of a model SCR with the seabed. Applied Ocean Research, 2009.
[10] N.R.T.Willis and P.T.J.West. Interaction between Deepwater Catenary Risers and a Soft Seabed: Large Scale Sea Trials. Offshore Technology Conference, 2001.
[11] C.Aubeny and G.Biscontin. Interaction Model for Steel Compliant Riser on Soft Seabed. Offshore Technology Conference, 2008.
[12] Christoper Bridge and Katherine Laver. Steel Catenary Riser Touchdown Point Vertical Interaction Models. Offshore Technology Conference, 2004.
[13] Ali Nakhaee and Jun Zhang. Trenching effects on dynamic behavior of a steel catenary riser. Ocean Engineering, 2009.
[14] E.Fontaine, J.F.Nauroy, P.Foray, A.Roux, H.Gueveneux. Pipe– soil interaction in soft kaolinite: vertical stiffness and damping. In: Proceedings of the 14th International Offshore and Polar Conference, 2004.
[15] Ali Nakhaee and Jun Zhang. Dynamic Interaction Between SCR And Seabed. Offshore Mechanics and Arctic Engineering, 2007.
[16] Paulling, J.R, and Webster, W.C. A Consistent, large-amplitude analysis of the coupled response of a TLP and Tendon system. Proceedings Fifth International Mechanics and Arctic Engineering Symposium,1986.

[17] Arcandra Tahar and M.H.Kim. Hull/mooring/riser coupled dynamic analysis and sensitivity study of a tanker-based FPSO. Applied Ocean Research. 2003.
[18] Zhang Fan, Yang Jian-min, Li Run-pei, Chen Gang. Coupling effects for cell-truss spar platform: Comparison of frequency- and time-domain analysis with model tests. 2008.
[19] Xiaohong Chen. Studies on dynamic interaction between deep-water floating structures and their mooring/tendon systems. Ph.D dissertation of Texas A&M University, 2002.
[20] D.L.Garrett. Dynamic analysis of slender rods. Journal of energy resources technology, 1982.
[21] D.L.Garrett. Coupled analysis of floating production systems. Ocean Engineering, 2005.
[22] Arcandra. Hull/Mooring/Riser coupled dynamic analysis of a deepwater floating platform in with polyester lines. Ph.D dissertation of Texas A&M University, 2001.
[23] Zhihuang Ran. Coupled dynamic analysis of floating structures in waves and currents. Ph.D dissertation of Texas A&M University, 2000.
[24] HARP manuals user's manual.
[25] Y.M.Low and R.S.Langley. Time and frequency domain coupled analysis of deepwater floating production systems. Applied Ocean Research, 2006.
[26] Morison J R, O'Brien M D, Johnson J W, and Schaaf S A. The force exerted by surface waves on piles. Petrol Trans AIME, 1950.
[27] R.Burrows, R.G.Tickell, D. Hames and G.Najafian. Morison wave force coefficients for application to random seas. Applied Ocean Research, 1997.
[28] C.H.K. Williamson and R. Govardhan. A brief review of recent results in vortex-induced vibrations. Journal of Wind Engineering and Industrial Aerodynamics, 2008.
[29] Pan Zhi-yuan, Cui Wei-cheng and Zhang Xiao-ci. An Overview on VIV of Slender Marine Structures. Journal of Ship Mechanics, 2005.
[30] M.L.Facchinetti, E.de Langre and F.Biolley. Coupling of structure and wake oscillators in vortex-induced vibrations. Journal of fluids and structures, 2003.
[31] Xu Wan-hai, Wu Ying-xiang and Yu Jian-xing. A new wake oscillator model for predicting vortex induced vibration of a circular cylinder. Journal of Hydrodynamics, 2010.
[32] J.Kim Vandiver and Li Li. SHEAR7 V4.4 PROGRAM THEORTICAL MANUAL. 2005.
[33] Z.Y.Pan, W.C.Cui and Q.M.Miao. Numerical simulation of vortex-induced vibration of a circular cylinder at low mass-damping using RANS code. Journal Fluids and Structures, 2007.
[34] Joseph Thomas Klamo. Effects of Damping and Reynolds Number on Vortex-Induced Vibrations. Ph.D dissertation of California Institute of Technology, 2007.
[35] M.Dixon and D.charlesworth. Application of CFD for Vortex-Induced Vibration Analysis of Marine Risers in Projects. Offshore Technology Conference, 2006.
[36] Fluent User's Guide.
[37] DNV-RP-C203. FATIGUE DESIGN OF OFFSHORE STEEL STRUCTURES. 2010.
[38] ABS. Guide for the fatigue assessment of offshore structures, 2003.
[39] Charles P. Sparks. Fundamentals of Marine Riser Mechanics: Basic Principle and Simplified Analyses. Penn Well, 2007.
[40] Christopher E. Smith. SPECIAL REPORT: Independence Hub sees record SCR installation. Oil & Gas Journal, 2007.
[41] DUAN Meng-lan, WANG Yi, Segen ESTEFEN, HE Ning, LI Li-na and CHEN Bang-min. An Installation System of Deepwater Riser by An S-Lay Vessel. China Ocean Engineering, 2011.
[42] Mao Dongfeng, Duan Menglan, Wang Yi, He Ning, Chen Bangmin and Zhang Yingjie. Model test investigation on an innovative lifting system for deepwater riser installation. PETROLEUM SCIENCE, 2010.

13

Robust-Adaptive Flux Observers in Speed Vector Control of Induction Motor Drives

Filote Constantin[1,2] and Ciufudean Calin[1]
[1]*Stefan cel Mare University of Suceava*
[2]*SC Germaro Electronics SRL*
Romania

1. Introduction

The speed control of induction motors can be divided into two distinct strategies, depending on the type of dynamics that is required: scalar control (static control of the torque) and vector control (dynamic control of the torque) (Blaabjerg et al. 2005; Bose, 2000, 2006; Filote et al. 2009; Holtz, 2002; Leonhard, 1990; Umanand & Bhat, 1995; Vas, 1998).
Since the couple and flux levels depend on the motor behavior to frequency and voltage applied to it, maintaining the flux constant is strongly required in scalar control (V/f control). Despite the simplicity and the low cost implementation of the control method, it still presents the disadvantage of poor torque dynamics.
Vector control reassesses one of the advantages of direct current (dc) drives, which is the separation of speed and couple loops. According to its structure and to its own functioning principle, the dc drive is naturally field oriented, hence, the separation of the speed and current loops.
In the case of vector controllers of induction motor, the magnitude and phase stator current are controlled in accordance with the flux vector. There are three vector control strategies according to the type of drive-controlled flux: stator flux, rotor flux and air gap flux.
We can determine the magnitude and position of the rotor flux by using flux sensors (direct field-oriented control), estimators or observers (indirect field-oriented control) and by measuring some electrical and mechanical measurable states of electrical drives induction motor systems.
When the rotor flux estimation is performed after measuring only the electrical measurable states (voltages and stator currents), in this case we have a sensorless vector control system.
The implementation of the direct field-oriented control requires the measurement or calculation of the flux space vector (magnitude and position). The measurement of the air gap flux requires the use of some Hall sensors (Umanand & Bhat, 1995), which requires specially constructed induction motor. On one hand, they are very sensitive to temperature and mechanical vibrations. The flux signal is highly distorted by slot harmonics (Kreindler, 1994), whose spectrum and amplitude depend on the rotor speed (hence, the difficulty to be filtered).
The implementation of the indirect field-oriented control requires the identification of rotor flux instantaneous position, and the calculus of stator current prescribed value in sensorless vector control system (Gadue et. al. 2009; Griva et al. 1997; Lascu et al. 2004). The

transformation of stator axes and calculus of commands have to be applied to the inverter to obtain this current. The flux estimation is dependent on the induction machine parameters.
In this chapter, we present a comparison of the performances among three rotor flux observers. If the rotor flux is applied as criterion in the vector control of induction motor, the value and direction of the flux needs to be known. Starting from two induction motor mathematical models, this section analyses theoretically and in terms of simulation, the performances of a conventional rotor flux simulator with a view to the temperature influence of the rotor resistance. Flux observers were used to estimate the flux, since classic methods do not seem to provide acceptable performances. This section analyses the performances of a robust-adaptive rotor flux observer, starting from a mathematical model and using simulation.
Section 2 presents the analysis of conventional flux simulators based on the current and tension model of the induction model. The numerical simulation results of the two simulators generate conclusions regarding their implementation in applications. In section 3, we introduce the adaptive flux observer and present simulation tests of its robustness in rotor resistance variation with temperature. Closed-loop vector control system with robust-adaptive flux observer is introduced in section 4. The correct estimation methods of the rotor flux magnitude and position are checked and we verify if the system orients itself after the rotor flux direction.

2. Analysis of the conventional flux simulators

The structure of the adaptive observers is based on the combination of a simulator for the estimated magnitude with a corrector for the estimation error (Schauder, 1989). For the asynchronous motor, two types of such simulators can be inferred:
- simulator based on a current model;
- simulator based on a tension model.

2.1 Non-linear current model

An asynchronous motor model with a random orthogonal reference is going to be built to be used for flux observer and speed estimator simulation, in rotor resistance identification in sensorless vector control.
When the motor is sinusoidal voltage fed and it functions in a random reference, the equivalent classic single-phase scheme of asynchronous motor is built according to Fig. 1.

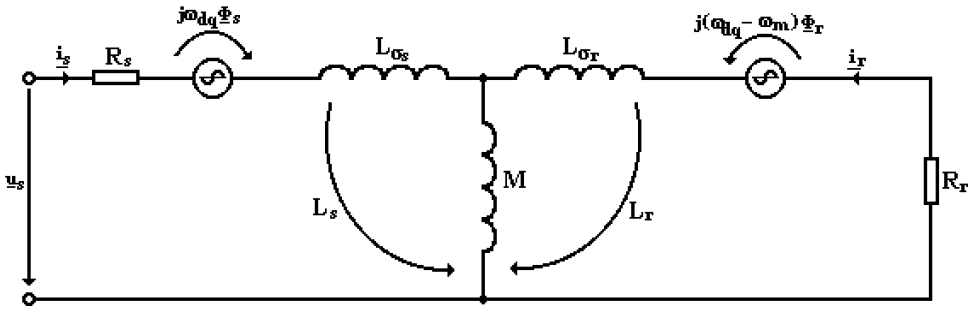


Fig. 1. Equivalent single-phase scheme of asynchronous motor in a random orthogonal reference

In a randomly oriented orthogonal reference, the vector equations of the asynchronous motor, considering d axes as the real axes (1), and q axes as the imaginary direction (j), can be written as follows:

$$\begin{cases} \underline{u}_s = R_s \cdot \underline{i}_s + \dfrac{d\underline{\Phi}_s}{dt} + j \cdot \omega_{dq} \cdot \underline{\Phi}_s \\ \underline{u}_r = 0 = R_r \cdot \underline{i}_r + \dfrac{d\underline{\Phi}_r}{dt} + j \cdot \left(\omega_{dq} - \omega_m\right) \cdot \underline{\Phi}_r \end{cases} \quad (1)$$

where:

$$\begin{cases} \underline{u}_s = u_{sd} + j \cdot u_{sq} \\ \underline{u}_r = u_{rd} + j \cdot u_{rq} \end{cases} \quad \begin{cases} \underline{i}_s = i_{sd} + j \cdot i_{sq} \\ \underline{i}_r = i_{rd} + j \cdot i_{rq} \end{cases} \quad \begin{cases} \underline{\Phi}_s = \Phi_{sd} + j \cdot \Phi_{sq} \\ \underline{\Phi}_r = \Phi_{rd} + j \cdot \Phi_{rq} \end{cases} \quad (2)$$

The expressions of stator and rotor fluxes, according to Fig. 1, can be written as follows:

$$\begin{cases} \Phi_{sd} = L_s \cdot i_{sd} + M \cdot i_{rd} \\ \Phi_{sq} = L_s \cdot i_{sq} + M \cdot i_{rq} \end{cases} \quad \begin{cases} \Phi_{rd} = L_s \cdot i_{rd} + M \cdot i_{sd} \\ \Phi_{rq} = L_s \cdot i_{rq} + M \cdot i_{sq} \end{cases} \quad (3)$$

The non-linear model of the asynchronous motor has as input data the stator orthogonal voltages that resulted from the three-phased voltages by means of a Park system transform [*P(0)*], and as output data, the stator currents and the rotor fluxes in the two orthogonal axes.
The expressions of rotor currents are obtained from relations (3), according to model input values:

$$\begin{cases} i_{rd} = \dfrac{1}{L_r}(\Phi_{rd} - M \cdot i_{sd}) \\ i_{rq} = \dfrac{1}{L_r}(\Phi_{rq} - M \cdot i_{sq}) \end{cases} \quad (4)$$

The derived rotor flux expressions are obtained from rotor voltage equations (1):

$$\begin{cases} \dfrac{d\Phi_{rd}}{dt} = -R_r \cdot i_{rd} + (\omega_{dq} - \omega_m) \cdot \Phi_{rq} \\ \dfrac{d\Phi_{rq}}{dt} = -R_r \cdot i_{rq} - (\omega_{dq} - \omega_m) \cdot \Phi_{rd} \end{cases} \quad (5)$$

Relations (4) and (5) are replaced and the expressions of two model output are obtained according to the model input:

$$\begin{cases} \dfrac{d\Phi_{rd}}{dt} = \dfrac{R_r \cdot M}{L_r} \cdot i_{sd} - \dfrac{R_r}{L_r} \cdot \Phi_{rd} + (\omega_{dq} - \omega_m) \cdot \Phi_{rq} \\ \dfrac{d\Phi_{rq}}{dt} = \dfrac{R_r \cdot M}{L_r} \cdot i_{sq} - \dfrac{R_r}{L_r} \cdot \Phi_{rq} - (\omega_{dq} - \omega_m) \cdot \Phi_{rd} \end{cases} \quad (6)$$

The stator flux expressions (3) of the two orthogonal axes are replaced in the stator voltage equations (1) and there results:

$$\begin{cases} \dfrac{di_{sd}}{dt} = \dfrac{1}{L_s}(-M\dfrac{di_{rd}}{dt} - R_s \cdot i_{sd} + \omega_{dq} \cdot L_s \cdot i_{sq} + \omega_{dq} \cdot M \cdot i_{rq} + u_{sd}) \\ \dfrac{di_{sq}}{dt} = \dfrac{1}{L_s}(-M\dfrac{di_{rq}}{dt} - R_s \cdot i_{sq} - \omega_{dq} \cdot L_s \cdot i_{sd} - \omega_{dq} \cdot M \cdot i_{rd} + u_{sq}) \end{cases} \tag{7}$$

The rotor current expressions obtained in (4) are replaced, and then, the relations (6) are replaced in (7) and we obtain:

$$\begin{cases} \dfrac{di_{sd}}{dt} = \dfrac{1}{1-\dfrac{M^2}{L_s L_r}}\left\{-(\dfrac{R_s}{L_s} + \dfrac{R_r M^2}{L_s {L_r}^2}) \cdot i_{sd} + (1-\dfrac{M^2}{L_s L_r})\omega_{dq} \cdot i_{sq} + \dfrac{R_r M}{L_s {L_r}^2} \cdot \Phi_{rd} + \dfrac{M}{L_s L_r}\omega_m \cdot \Phi_{rq}\right\} \\ \dfrac{di_{sd}}{dt} = \dfrac{1}{1-\dfrac{M^2}{L_s L_r}}\left\{-(1-\dfrac{M^2}{L_s L_r})\omega_{dq} \cdot i_{sd} - (\dfrac{R_s}{L_s} + \dfrac{R_r M^2}{L_s {L_r}^2}) \cdot i_{sq} - \dfrac{M}{L_s L_r}\omega_m \cdot \Phi_{rd} + \dfrac{R_r M}{L_s {L_r}^2} \cdot \Phi_{rq}\right\} \end{cases} \tag{8}$$

The total dispersion coefficient is defined noted by:

$$\sigma = 1 - \frac{M^2}{L_s L_r}, \tag{9}$$

value which simplifies considerably the writing of the equations characterizing the non-linear model:
Relations (8) noted with (9) become:

$$\begin{cases} \dfrac{di_{sd}}{dt} = -(\dfrac{R_s}{\sigma L_s} + \dfrac{R_r(1-\sigma)}{\sigma L_r}) \cdot i_{sd} + \omega_{dq} \cdot i_{sq} + \dfrac{R_r(1-\sigma)}{\sigma M L_r} \cdot \Phi_{rd} + \dfrac{1-\sigma}{\sigma M} \cdot \omega_m \cdot \Phi_{rq} \\ \dfrac{di_{sd}}{dt} = -\omega_{dq} \cdot i_{sd} - (\dfrac{R_s}{\sigma L_s} + \dfrac{R_r(1-\sigma)}{\sigma L_r}) \cdot i_{sq} - \dfrac{1-\sigma}{\sigma M} \cdot \omega_m \cdot \Phi_{rd} + \dfrac{R_r(1-\sigma)}{\sigma M L_r} \cdot \Phi_{rq} \end{cases} \tag{10}$$

Differential equations (6) and (10) describe a system of "electrical" differential equations which can be expresses as a state vector system:

$$\begin{cases} \dfrac{d}{dt} X_e = A_e \cdot X_e + B_e \cdot U_e \\ Y_e = C_e \cdot X_e \end{cases} \tag{11}$$

where:

- state vector is:

$$X_e = \begin{bmatrix} i_{sd} & i_{sq} & \Phi_{rd} & \Phi_{rq} \end{bmatrix}^T \tag{12}$$

- excitation vector is:

$$U_e = \begin{bmatrix} u_{sd} \\ u_{sq} \end{bmatrix} \tag{13}$$

- state matrix is:

$$A_e = \begin{bmatrix} -\left(\frac{R_s}{\sigma L_s} + \frac{R_r(1-\sigma)}{\sigma L_r}\right) & \omega_{dq} & \frac{R_r(1-\sigma)}{\sigma M L_r} & \frac{1-\sigma}{\sigma M}\cdot\omega_m \\ -\omega_{dq} & -\left(\frac{R_s}{\sigma L_s} + \frac{R_r(1-\sigma)}{\sigma L_r}\right) & -\frac{1-\sigma}{\sigma M}\cdot\omega_m & \frac{R_r(1-\sigma)}{\sigma M L_r} \\ \frac{MR_r}{L_r} & 0 & -\frac{R_r}{L_r} & \omega_{dq}-\omega_m \\ 0 & \frac{MR_r}{L_r} & -(\omega_{dq}-\omega_m) & -\frac{R_r}{L_r} \end{bmatrix} \tag{14}$$

- state matrix, vector and matrix of initial conditions are:

$$B_e = \begin{bmatrix} \frac{1}{\sigma L_s} & 0 \\ 0 & \frac{1}{\sigma L_s} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$Y_e = \begin{bmatrix} i_{sd} \\ i_{sq} \end{bmatrix} \qquad C_e = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{15}$$

A more simplified form can be used with respect to the notations if stator and rotor time constants are noted with:

$$\tau_s = \frac{L_s}{R_s} \qquad \tau_r = \frac{L_r}{R_r}. \tag{16}$$

A state matrix becomes:

$$A = \begin{bmatrix} -\left(\frac{1}{\sigma\tau_s} + \frac{1-\sigma}{\sigma\tau_r}\right) & \omega_{dq} & \frac{1-\sigma}{\sigma M\tau_r} & \frac{1-\sigma}{\sigma M}\cdot\omega_m \\ -\omega_{dq} & -\left(\frac{1}{\sigma\tau_s} + \frac{1-\sigma}{\sigma\tau_r}\right) & -\frac{1-\sigma}{\sigma M}\cdot\omega_m & \frac{1-\sigma}{\sigma M\tau_r} \\ \frac{M}{T_r} & 0 & -\frac{1}{\tau_r} & \omega_{dq}-\omega_m \\ 0 & \frac{M}{\tau_r} & -(\omega_{dq}-\omega_m) & -\frac{1}{\tau_r} \end{bmatrix} \tag{17}$$

This generalised non-linear model of the induction motor applies to a random reference both in transient and permanent regimes.

In a stator related reference, $(d\text{-}q) \equiv (d_s\text{-}q_s)$, the condition for the equation system (11) is:

$$\omega_{dq} = 0 \tag{18}$$

Under these conditions, the differential equation system is:

$$\frac{d}{dt}\begin{bmatrix} i_{sd} \\ i_{sq} \\ \Phi_{sd} \\ \Phi_{sq} \end{bmatrix} = \begin{bmatrix} -(\frac{1}{\sigma\tau_s}+\frac{1-\sigma}{\sigma\tau_r}) & 0 & \frac{1-\sigma}{\sigma M\tau_r} & \frac{1-\sigma}{\sigma M}\cdot\omega_m \\ 0 & -(\frac{1}{\sigma\tau_s}+\frac{1-\sigma}{\sigma\tau_r}) & -\frac{1-\sigma}{\sigma M}\cdot\omega_m & \frac{1-\sigma}{\sigma M\tau_r} \\ \frac{M}{\tau_r} & 0 & -\frac{1}{\tau_r} & -\omega_m \\ 0 & \frac{M}{\tau_r} & \omega_m & -\frac{1}{\tau_r} \end{bmatrix} \cdot \begin{bmatrix} i_{sd} \\ i_{sq} \\ \Phi_{sd} \\ \Phi_{sq} \end{bmatrix} + \begin{bmatrix} \frac{1}{\sigma L_s} & 0 \\ 0 & \frac{1}{\sigma L_s} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} u_{sd} \\ u_{sq} \end{bmatrix} \tag{19}$$

To this "electrical" differential equation system we add a "mechanical" equation system (*m* indices) which can be written as:

$$\begin{cases} \frac{d}{dt}X_m = A_m \cdot X_m + B_m \cdot U_m \\ Y_m = C_m \cdot X_m \end{cases} \tag{20}$$

where,

$$X_m = \begin{bmatrix} \omega_m \\ m_{rez} \end{bmatrix} \qquad A_m = \begin{bmatrix} -\frac{F_f}{J} & -\frac{p}{J} \\ 0 & 0 \end{bmatrix} \qquad U_m = m_e \tag{21}$$

and

$$B_m = \begin{bmatrix} \frac{p}{J} \\ 0 \end{bmatrix} \qquad C_m = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{22}$$

The expression of the electromagnetic torque which is expressed according to model output values is:

$$m_e = \frac{3}{2}\cdot p \cdot \frac{M}{L_r}\cdot(\Phi_{rd}\cdot i_{sq} - \Phi_{rq}\cdot i_{sd}) \tag{23}$$

Relations (19), (20) and (23) underlay the achievement of current non-linear model of asynchronous motor (**I_mas_md)** that was used in all estimators and identification algorithms of induction motor parameters.

2.2 Voltage non-linear model

This model has two orthogonal axes and load torque as input values, and stator and rotor currents between the two axes as output values.

For a random reference *d-q*, the derivative expressions in relation with the rotor current times are calculated on the basis of the general relations (1), (3), thus resulting the following differential "electrical" system, expressed by means of state vectors:

$$\begin{cases} \dfrac{d}{dt} X_e = A_e \cdot X_e + B_e \cdot U_e \\ Y_e = C_e \cdot X_e \end{cases} \tag{24}$$

where:

- state vector is:

$$X_e = \begin{bmatrix} i_{sd} & i_{sq} & i_{rd} & i_{rq} \end{bmatrix}^{tT} \tag{25}$$

- excitation vector is:

$$U_e = \begin{bmatrix} u_{sd} \\ u_{sq} \end{bmatrix} \tag{26}$$

- state matrix is:

$$A_e = \begin{bmatrix} -\dfrac{R_s}{\sigma L_s} & -\dfrac{(\omega_{dq}-\omega_m)\cdot M^2}{\sigma L_s L_r} + \dfrac{\omega_{dq}}{\sigma} & \dfrac{R_r M}{\sigma L_s L_r} & \dfrac{M}{\sigma L_s}\cdot\omega_m \\ \dfrac{(\omega_{dq}-\omega_m)\cdot M^2}{\sigma L_s L_r} - \dfrac{\omega_{dq}}{\sigma} & -\dfrac{R_s}{\sigma L_s} & -\dfrac{M}{\sigma L_s}\omega_m & \dfrac{R_r M}{\sigma L_s L_r} \\ \dfrac{R_s}{\sigma L_s} & -\omega_m \dfrac{M}{\sigma L_r} & -\dfrac{R_r}{\sigma L_r} & \dfrac{\omega_{dq}-\omega_m}{\sigma} - \dfrac{\omega_{dq} M^2}{\sigma L_s L_r} \\ \omega_m \dfrac{M}{\sigma L_r} & \dfrac{R_s}{\sigma L_s} & -\dfrac{\omega_{dq}-\omega_m}{\sigma} + \dfrac{\omega_{dq} M^2}{\sigma L_s L_r} & -\dfrac{R_r}{\sigma L_r} \end{bmatrix} \tag{27}$$

- excitation matrix, vector and matrix of the initial conditions are:

$$B_e = \begin{bmatrix} \dfrac{1}{\sigma L_s} & 0 \\ 0 & \dfrac{1}{\sigma L_s} \\ -\dfrac{M}{\sigma L_s L_r} & 0 \\ 0 & -\dfrac{M}{\sigma L_s L_r} \end{bmatrix} \qquad Y_e = \begin{bmatrix} i_{sd} \\ i_{sq} \end{bmatrix} \qquad C_e = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{28}$$

To this differential system we add the "mechanical" state equations system (20), forming together voltage model **V_mas_md**.

2.3 Simulator based on the current model

Next, it will be demonstrated why the simulators cannot be used as flux observers, solution which will lead to the corrector's removal from the adaptive calculus methods. The matrix state equations describing the working of an induction motor, in orthogonal co-ordinates *d-q*

(Filote & Graur,1998; Filote et al. 2007, 2009; Marchensoni et al. 1994; Pană, 1995; Schauder, 1989), can be written in complex as:

$$\begin{bmatrix} \dot{\underline{i}}_s \\ \dot{\underline{\Phi}}_r \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} \underline{i}_s \\ \underline{\Phi}_r \end{bmatrix} + \begin{bmatrix} b_1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{u}_s \end{bmatrix} \tag{29}$$

where:

$$\begin{cases} a_{11} = -\dfrac{R_s}{\sigma L_s} - \dfrac{R_r(1-\sigma)}{\sigma L_r}; \\ a_{12} = \dfrac{M}{\sigma L_s L_r} \cdot \left(\dfrac{R_r}{L_r} - j\omega_m \right); \end{cases} \quad \begin{cases} a_{21} = \dfrac{MR_r}{L_r} \\ a_{22} = -\dfrac{R_r}{L_r} + j\omega_m \end{cases} \quad \begin{cases} b_1 = \dfrac{MR_r}{L_r} \\ \sigma = 1 - \left(\dfrac{M^2}{L_s L_r} \right) \end{cases} \tag{30}$$

The current model can be obtained if we consider the second equation from relation (29) that allows us to touch the estimated flux as:

$$\dot{\hat{\underline{\Phi}}}_r = a_{21} \cdot \underline{i}_s + a_{22} \cdot \hat{\underline{\Phi}}_r = \left(\frac{MR_r}{L_r} \right) \cdot \underline{i}_s + \left(-\frac{R_r}{L_r} + j\omega_m \right) \cdot \hat{\underline{\Phi}}_r \tag{31}$$

The estimation error is given by the difference between the value of the estimated flux and the real one:

$$\underline{\varepsilon} = \hat{\underline{\Phi}}_r - \underline{\Phi}_r \,. \tag{32}$$

When the motor parameters are constant, the derivative of that error, which localizes the poles of the observer in complex plane, becomes:

$$\dot{\underline{\varepsilon}} = a_{22} \cdot \underline{\varepsilon} = \left(-\frac{R_r}{L_r} \pm j\omega_m \right) \cdot \underline{\varepsilon} \,. \tag{33}$$

The relation (33) shows that the derivative of the estimation error depends only by motor variable parameters (rotor resistance and inductance, motor speed), so it can be controlled. Because the poles in complex plane depend on the motor speed, the observer stability will decrease while the speed increases.

2.4 Simulator based on the tension model

The expression of the tension model can be obtained from the current model of the induction motor (29), written as follows:

$$\begin{cases} \dot{\underline{i}}_s = a_{11} \cdot \underline{i}_s + a_{12} \cdot \underline{\Phi}_r + b_1 \cdot u_s \\ \dot{\underline{\Phi}}_r = a_{21} \cdot \underline{i}_s + a_{22} \cdot \underline{\Phi}_r \end{cases} \tag{34}$$

If we replace the expression for the rotor flux in the first equation, from the relation (34), with that obtained from the second equation, there results the simulator based on the tension model, defined with the expression:

$$\hat{\underline{\Phi}}_r = \left(a_{21} - \frac{a_{22} \cdot a_{11}}{a_{12}} \right) \cdot \underline{i}_s + \frac{a_{22}}{a_{12}} \cdot \dot{\underline{i}}_s - \frac{a_{22} \cdot b_1}{a_{12}} \cdot \underline{u}_s = \frac{L_r}{M} \cdot \left(\underline{u}_s - R_s \cdot \underline{i}_s \right) - \left(\sigma \cdot \frac{L_r L_s}{M} \right) \cdot \dot{\underline{i}}_s \tag{35}$$

In that case, the derivative of the estimation error is:

$$\dot{\varepsilon} = 0 \,. \tag{36}$$

Therefore, the simulator based on tension model is a pure integrator. Because of this reason, the simulator based on model tension does not allow reducing the initial estimation error and, at the same time, it is very sensitive at low speeds.

The advantage that the estimation error variation does not depend on R_r results in the fact that the simulator is totally insensitive to the variation of rotor resistance. Consequently, the observations made above do not have to be taken into account.

2.5 Results of numerical simulator simulations

Next, the dependence of the rotor fluxes by the variation of the rotor resistance with the temperature, at the output of a conventional simulator based on the current model of the induction motor, will be presented.

As it is well known, the rotor resistance grows up with the temperature (Leonhard, 1990) and, because of this, the results obtained for a variation with +25% will be shown.

Analyzing the dependence between the orthogonal components, the rotor flux module and the rotor resistance variation with temperature in transitory starting conditions (low speeds), the following conclusions can be inferred:

- the rotor flux doesn't remain constant with the rotor resistance variation by temperature;
- that variation concerns only the magnitude of both orthogonal components and the module of the rotor flux (Figures 2 and 3);

Consequently, that model of the induction motor cannot be used as a flux observer into a pretentious vector control system as it is the orientation system after the rotor flux.

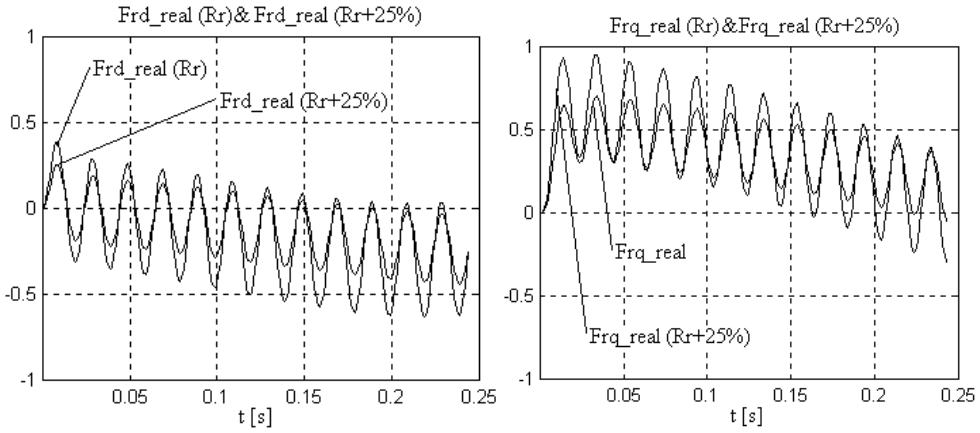


Fig. 2. Evolution of the flux orthogonal components with rotor resistance variation (low speeds).

It can be concluded that none of these simulators can be used with satisfactory results in rotor flux estimation for the asynchronous motor because of:

- the sensibility of the performances at rotor's resistance variation, for the simulator based on a current model (Fig. 2);
- the unsatisfactory low speeds working, for the simulator based on the tension model;

The satisfactory solution for indirect determination of the rotor flux is that of a robust-adaptive flux observer. Such an observer (Bose, 2006; Vas, 1998) offers remarkable robustness, with respect to the variation of the rotor's resistance, even for high values.

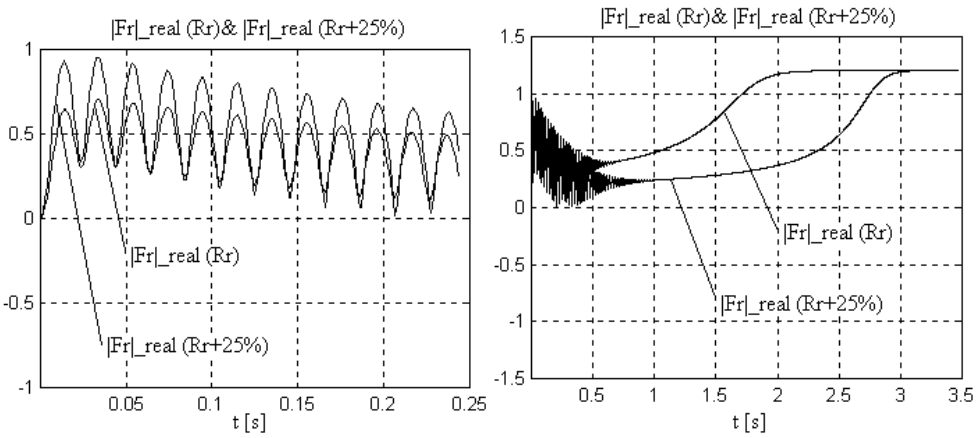


Fig. 3. Evolution of the rotor flux module due to the growth of the rotor resistance at low and high speeds.

3. Rotor flux observer

The above-mentioned disadvantages can be eliminated if flux observers are used. The latter can calculate the phasor of the rotor flux (amplitude and orientation) on the basis of some electrical units (u_s tensions and i_s stator currents) and mechanical ones (ω_m motor speed).

3.1 The structure of the flux observer

The structure of adaptive flux observers can be deduced from the linear model of induction motor (constant rotor speed) which is represented in a *d-q* system of axes, fixed on the stator and having a complex written form. This way of presentation is preferred because the study of the stability of these observers and the on-line determination of the parameters of the "matrix gate" require the fixation of the position of the poles in the complex plane.

The basic structure (Filote & Graur, 1998; Kubota et al. 1993; Pană, 1995; Schauder, 1989) of the adaptive observers consists of two essential functional elements:

- a simulator which emulates estimated values in a reference system;
- a corrector of estimated values based on an adjustable model, which is in fact a reaction loop which uses a gate to amplify, the error between the estimated value and the real one.

$$\hat{\Phi}_r = simulator + gate * (reaction\ value - estimated\ corrector\ value) \tag{37}$$

In the case of asynchronous motor, there are two possible simulators, mainly based on:
- Current model;
- Voltage model.

These simulators, taken separately, can use two types of correctors:
- Stator equations;
- Motor voltage model.

For each corrector type, we can use three types of reaction signals:
- Stator voltage;
- Stator current;
- Stator current derivative.

Combining the above presented types of simulators, correctors and reaction signals leads us to physically achievable observer structures (Pană, 1995).
In Fig. 4, it is presented the GOPINATH type flux observer which is formed of:
- **Simulator** (reference model) based on the current model of asynchronous motor;
- **Corrector** (adjustable model) based on motor stator equation
- **Model reaction value** is stator current derivative (di_s/dt), which is obtained from the corrector
- **Gate** (auto-tuning) noted with g complex value.

From a mathematical point of view, this structure of the Gopinath flux observer can be represented as relation (37).

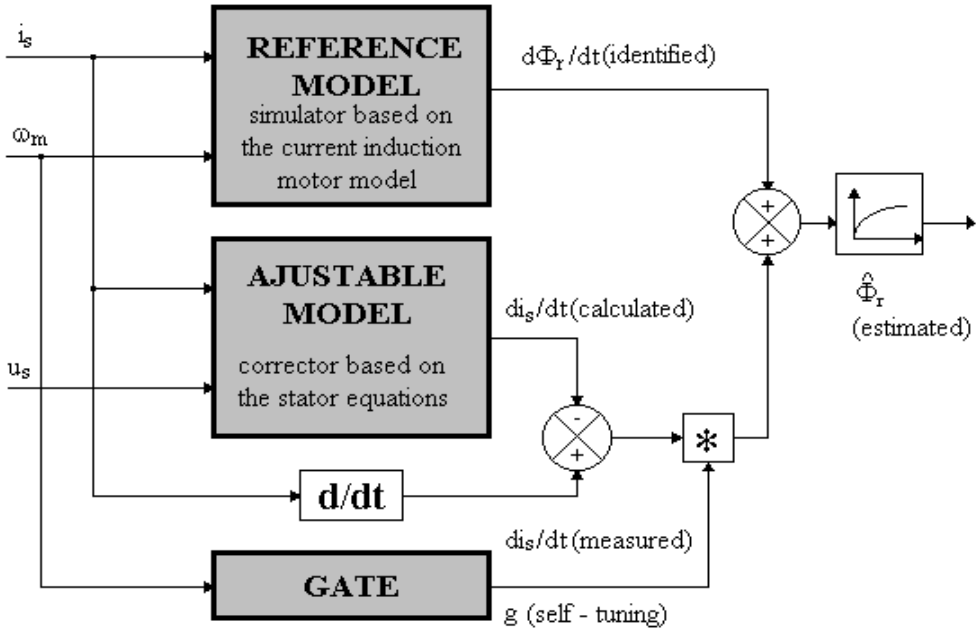


Fig. 4. The structure of the adaptive flux observer

Following the relation (37) and in accordance with Figure 4, we can obtain the **mathematical model of the estimator** (38):

$$\begin{aligned}\dot{\hat{\underline{\Phi}}}_r &= a_{21}\cdot\underline{i}_s + a_{22}\cdot\hat{\underline{\Phi}}_r + \underline{g}\cdot\left[\dot{\underline{i}}_s - \left(a_{11}\cdot\underline{i}_s + a_{12}\cdot\hat{\underline{\Phi}}_r + b_1\cdot\underline{u}_s\right)\right] = \\ &= \left(a_{22} - \underline{g}\cdot a_{12}\right)\cdot\hat{\underline{\Phi}}_r + \left(a_{21} - \underline{g}\cdot a_{11}\right)\cdot\underline{i}_s - \underline{g}\cdot b_1\cdot\underline{u}_s + \underline{g}\cdot\dot{\underline{i}}_s\end{aligned} \tag{38}$$

3.2 The stability analysis

The essential element, which determines the stability of the flux observer, as well as its insensitivity to the motor parameters variation, is gate "$\underline{g}$", a complex number:

$$g = g_a + jg_b \tag{39}$$

When motor parameters are constant, we define the estimation error, expressed by the difference between the estimated and the real rotor flux:

$$\underline{\varepsilon} = \hat{\underline{\Phi}}_r - \underline{\Phi}_r \tag{40}$$

Dynamics of estimation error (system stability) is given by the first derivative of estimation error:

$$\dot{\varepsilon} = \left(a_{22} - g\cdot a_{12}\right)\cdot\varepsilon = -h\cdot\varepsilon\ . \tag{41}$$

Relation (41) provides the position of the two complex joined poles of flux estimator; their coordinates are *Re(-h)* şi *+/-Im(-h)*. As it can be noticed, the position of these two poles depends on rotor speed, by means of a_{22} şi a_{12} coefficients (relation 30).

According to stability criteria, the system is stable if and only if the poles are in the negative complex semi plane. Here, for each value of rotor speed we have to determine g_a şi g_b estimator gate coefficients, such as the two poles fulfil the stability condition. Performing the calculus in relation (41), we obtain:

$$-\frac{R_r}{L_r} + j\cdot\omega_m - \frac{M}{\sigma L_s L_r}\cdot\left(g_a + jg_b\right)\cdot\left(\frac{R_r}{L_r} - j\omega_m\right) = -\alpha + j\beta \tag{42}$$

where α and β represent the coordinates imposed to the two observer poles. Balancing the real and imaginary parts leads to a two equation system providing gate coefficients equal to:

$$g_a = \left(\frac{\frac{R_r}{L_r}\cdot\alpha + \omega_m\beta}{\left(\frac{R_r}{L_r}\right)^2 + {\omega_m}^2} - 1\right)\cdot\frac{\sigma L_s L_r}{M} \qquad g_b = \left(\frac{\omega_m\cdot\alpha - \frac{R_r}{L_r}\cdot\beta}{\left(\frac{R_r}{L_r}\right)^2 + {\omega_m}^2}\right)\cdot\frac{\sigma L_s L_r}{M} \tag{43}$$

The best position of the two poles, α and β respectively, can be obtained from the analysis of the sensitivity of the observer to the rotor resistance variation with the temperature. In the simplest case, we can adopt the position of the poles in the negative semi plane precisely on the negative axis, which leads to the following relations:

$$\begin{cases}\alpha = k\cdot\sqrt{\left(\frac{R_r}{L_r}\right)^2 + {\omega_m}^2} & (k>0)\\ \beta = 0\end{cases} \tag{44}$$

One can notice that the values of the gate coefficients depend on the rotor speed consequently they must be calculated in real time (100-200 μs).

3.3 The mathematical model of the flux observer

Explaining the coefficients from expression (6.3.10) offers the final form of the GOPINATH flux observer model:

$$\underline{\hat{\dot{\Phi}}}_r = \left(a_{22} - \underline{g}\cdot a_{12}\right)\cdot\underline{\hat{\Phi}}_r + \left(a_{21} - \underline{g}\cdot a_{11}\right)\cdot\underline{i}_s - \underline{g}\cdot b_1\cdot\underline{u}_s + \underline{g}\cdot\underline{\dot{i}}_s \tag{45}$$

$$\begin{aligned}\left(\dot{\Phi}_{rd} + j\dot{\Phi}_{rq}\right) &= \left[-\frac{R_r}{L_r} + j\omega_m - \left(g_a + jg_b\right)\cdot\frac{M}{\sigma L_s L_r}\cdot\left(\frac{R_r}{L_r} - j\omega_m\right)\right]\cdot\left(\Phi_{rd} + j\Phi_{rq}\right) + \\ &+\left[\frac{MR_r}{L_r} - \left(g_a + jg_b\right)\cdot\left(-\frac{R_s}{\sigma L_s} - \frac{R_r(1-\sigma)}{\sigma L_r}\right)\right]\cdot\left(i_{sd} + ji_{sq}\right) - \\ &-\left(g_a + jg_b\right)\cdot\frac{1}{\sigma L_s}\cdot\left(u_{sd} + ju_{sq}\right) + \left(g_a + jg_b\right)\cdot\left(\dot{i}_{sd} + j\dot{i}_{sq}\right)\end{aligned} \tag{46}$$

Generally, it is advisable that the mathematical model be tested as a matrix state equation, which can be simulated or implemented in software environments from MATLAB or SIMULINK.

Matrix form of flux observer obtained from (46) is:

$$\begin{aligned}\begin{bmatrix}\dot{\Phi}_{rd}\\ \dot{\Phi}_{rq}\end{bmatrix} &= \begin{bmatrix}\left(-\frac{R_r}{L_r} - \frac{MR_r}{\sigma L_s {L_r}^2}\cdot g_a - \frac{M\omega_m}{\sigma L_s L_r}\cdot g_b\right) & -\left(\omega_m + \frac{M\omega_m}{\sigma L_s L_r}\cdot g_a - \frac{MR_r}{\sigma L_s {L_r}^2}\cdot g_b\right)\\ \left(\omega_m + \frac{M\omega_m}{\sigma L_s L_r}\cdot g_a - \frac{MR_r}{\sigma L_s {L_r}^2}\cdot g_b\right) & \left(-\frac{R_r}{L_r} - \frac{MR_r}{\sigma L_s {L_r}^2}\cdot g_a - \frac{M\omega_m}{\sigma L_s L_r}\cdot g_b\right)\end{bmatrix}\cdot\begin{bmatrix}\Phi_{rd}\\ \Phi_{rq}\end{bmatrix} + \\ &+\begin{bmatrix}-\frac{g_a}{\sigma L_s} & \frac{g_b}{\sigma L_s} & \left(\frac{MR_r}{L_r} + \frac{R_s}{\sigma L_s}g_a + \frac{R_r(1-\sigma)}{\sigma L_s}g_a\right) & -\left(\frac{R_s}{\sigma L_s}g_b + \frac{R_r(1-\sigma)}{\sigma L_r}g_b\right) & g_a & -g_b & 0\\ -\frac{g_b}{\sigma L_s} & -\frac{g_a}{\sigma L_s} & \left(\frac{R_s}{\sigma L_s}g_b + \frac{R_r(1-\sigma)}{\sigma L_r}g_b\right) & \left(\frac{MR_r}{L_r} + \frac{R_s}{\sigma L_s}g_a + \frac{R_r(1-\sigma)}{\sigma L_r}g_a\right) & g_b & g_a & 0\end{bmatrix}\cdot\begin{bmatrix}u_{sd}\\ u_{sq}\\ i_{sd}\\ i_{sq}\\ \dot{i}_{sd}\\ \dot{i}_{sq}\\ \omega_r\end{bmatrix}\end{aligned} \tag{47}$$

3.4 Results of numerical rotor flux observer simulation

The presented model has been simulated with a SIMULINK for MATLAB, which allows the inclusion of the differential system equations (47) in a simulation scheme, as shown in Fig. 5.

In accordance with the simulation scheme in Fig. 5 the values of the orthogonal components and the estimated rotor flux modulus will be compared (Fig. 6), from the output of the adaptive flux observer, with the real ones obtained at the output of the induction motor represented by the model in current (Li et al. 2005; Zhang et al. 2006).

The success of design flux observers is determined by pole assigning. The observers' sensitivity can be adjusted using gain coefficient (*k*).

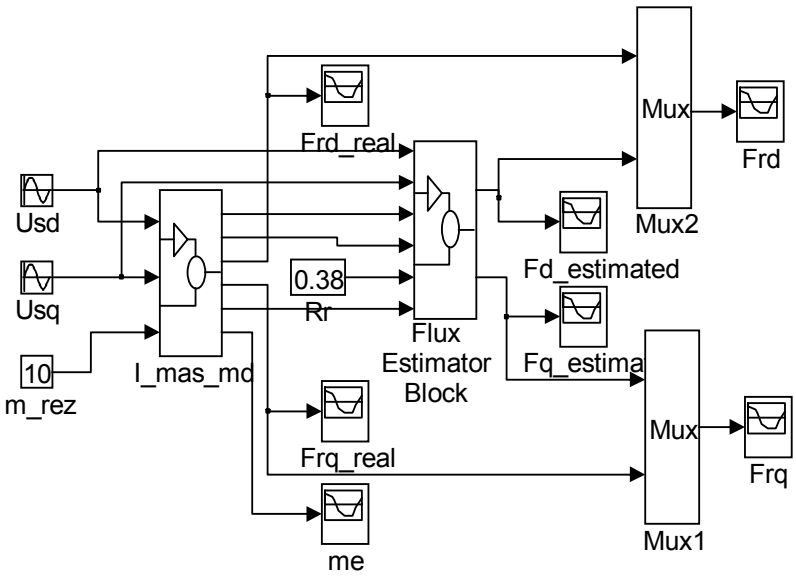


Fig. 5. The simulation scheme of the observer-motor assembly.

The very good results obtained in the flux estimation; even for important variations of the rotor, resistance with temperature (100%, in Fig. 6) demonstrates the robustness of such an observer and recommends its utilization in the applications of vector control for induction motor.

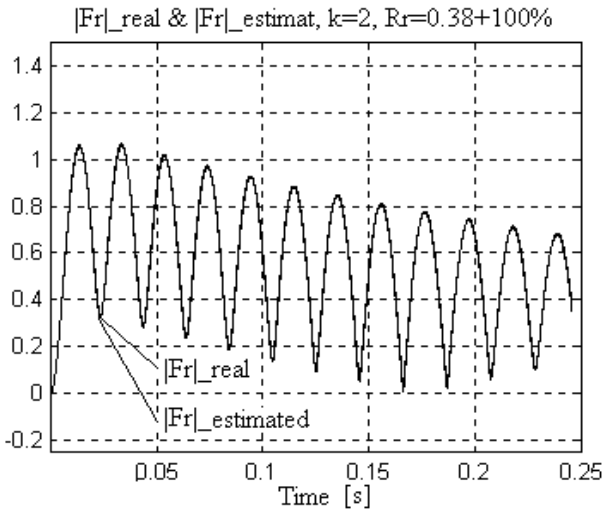


Fig. 6. Comparative presentation of the estimated (forward of the output filter) and the real rotor flux modulus.

4. Closed-loop vector control system

Figures 7 and 8 presents the bloc diagram of a rotor flux vector control system, which contains: robust-adaptive rotor flux observer, which estimates the instantaneous position and rotor flux module; vector flux analyzer that turns after rotor flux spatial vector; control algorithm and control voltage decupling block. Algorithm block is PI type and it is implemented on two control loops: closed-loop flux control and closed-loop torque control (Zhang et al. 2006).

The measured feedback values are the orthogonal components of the stator current and voltage, as well as the rotor's rotational speed. All these are applied on the robust-adaptive observer's input.

The flux analyzer calculates the modulus and the instantaneous position of the rotor flux vector, using the orthogonal components of the estimated flux. The modulus of the rotor flux and the rotor speed are feedback values in the two independent control loops of the vector control system.

The compensatory voltage block uses these two values in order to calculate the prescribed control voltage values for PWM inverter that powers the induction machine (Alexa et al. 2008; Lascu et al. 2004).

The inputs in the voltage decoupling block are the components of the stator control current, $i^*_{sd\lambda r}$ and $i^*_{sq\lambda r}$ which are provided by the two **PI** controls. The fitting of the transfer functions of the two PI controls in speed and flux is based on the relations obtained from the rotor flux orientation strategy for the reactive component $i^*_{sd\lambda r}$ and from the expression of the electromagnetic torque for the active component of the stator current $i^*_{sq\lambda r}$.

Results of real time close-loop simulation of vector control system of robust-adaptive flux observer asynchronous motor are presented by the following wave forms.

Analysis of wave forms in Fig. 9 reveals their co-sinusoidal variation and the maintenance of the two orthogonal voltages. The maximal value stabilizes fast at a value corresponding to a ratio voltage/ frequency which is dynamically calculated for an approximate $n_N/2$ speed.

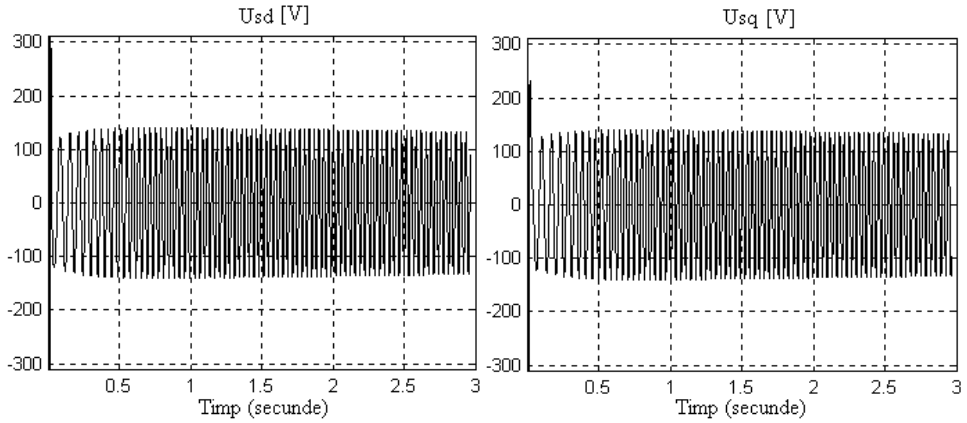


Fig. 9. Evolution of orthogonal stator voltages in a vector control system with rotor flux estimator.

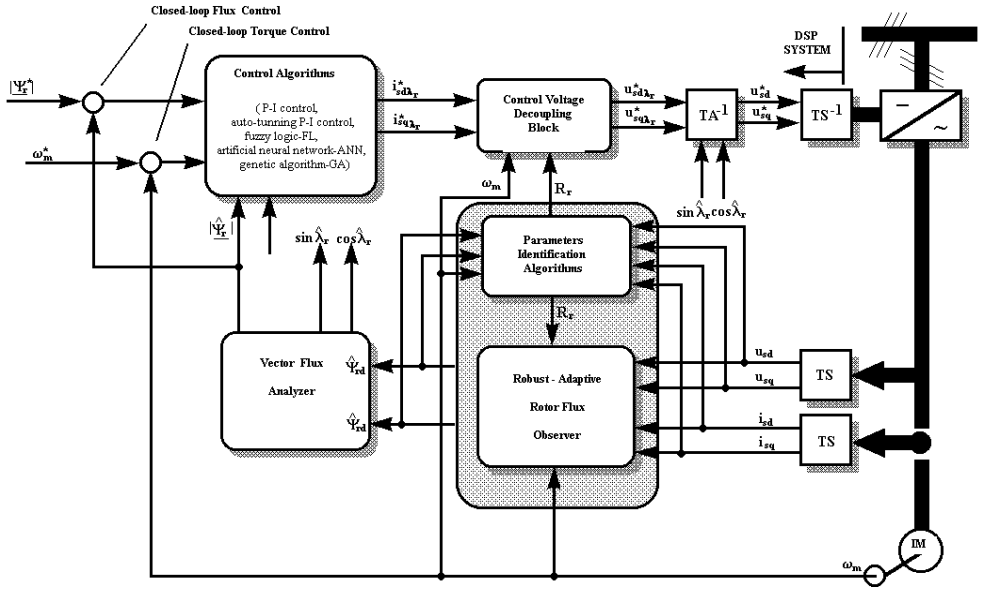


Fig. 7. Vector control of induction motor with speed rotor measurement [8].

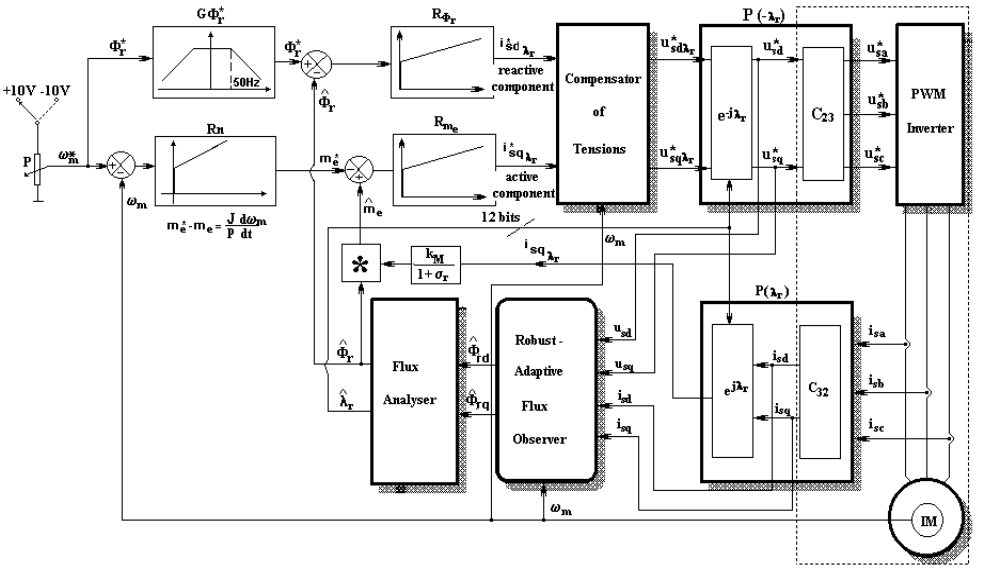


Fig. 8. Modelling of the closed-loop vector control system

The current increase in transitory regime (Fig. 10) is not as important anymore as in the case of direct start from industrial power supply, although during the simulation, the response to step signal was monitored

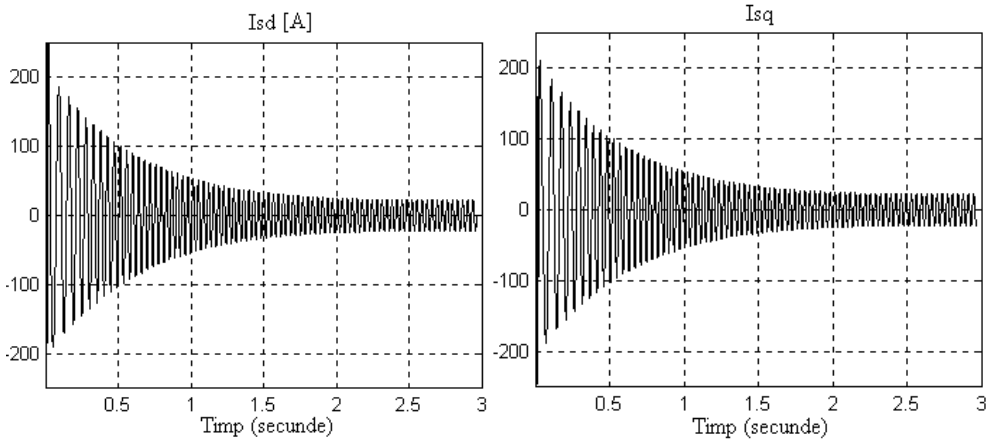


Fig. 10. The evolution of orthogonal stator currents, **a.c. values**, for vector control with robust-adaptive rotor flux estimator ($\omega_m^* = 157$ *rad/s*).

Oriented stator tensions (Fig. 11), from tension compensator output, are already d.c. values (slowly variable) after a very short time (0.01 - 0.02 s).

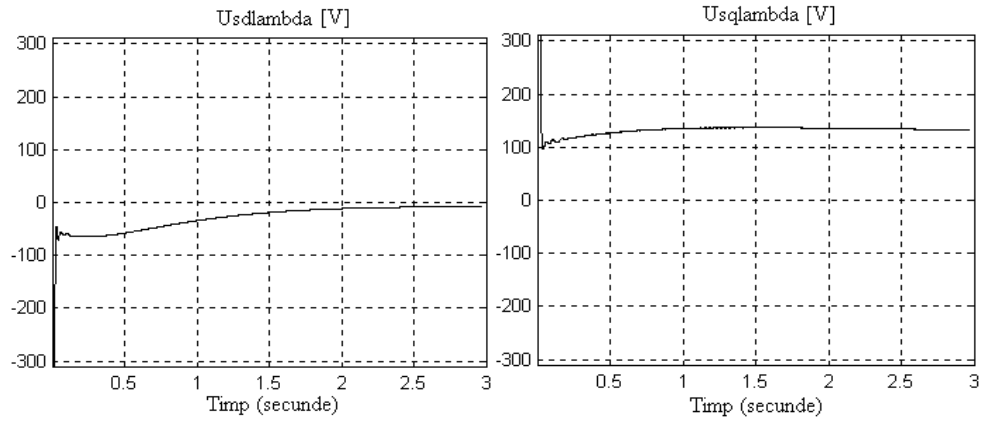
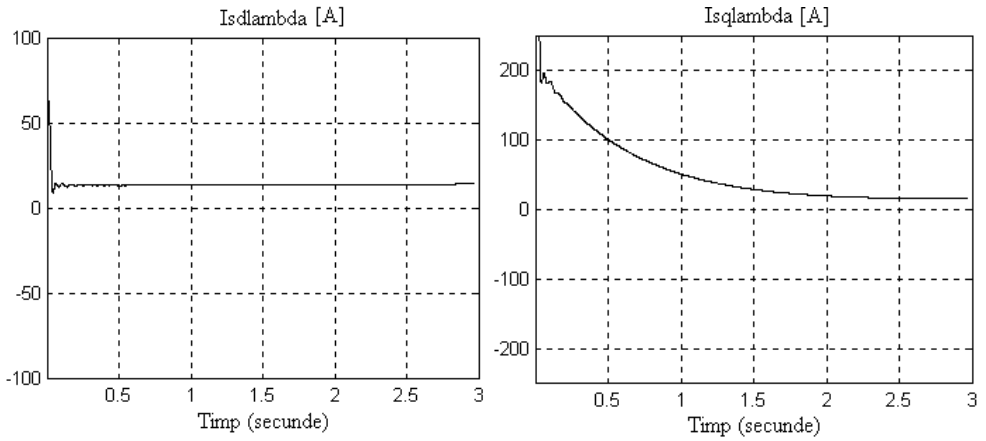


Fig. 11. The evolution of oriented orthogonal stator tensions (**d.c. values**), for vector control with robust-adaptive rotor flux estimator ($\omega_m^* = 157$ *rad/s*).

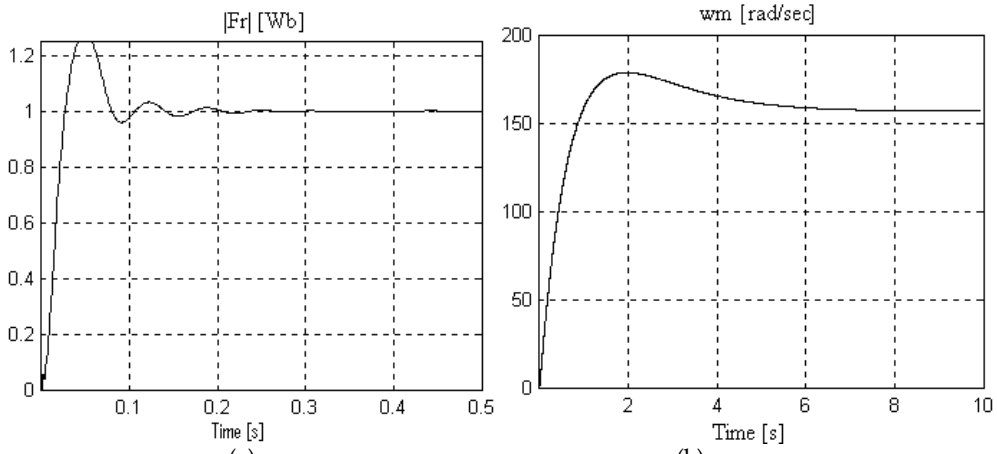
Same observation can be made for the evolution of oriented control stator currents (Fig. 12) from the output of torque and flux regulators.

The response to step signal leads to obtaining the flux (Fig. 13a) and speed (Fig. 13a) prescribed values, after the transitory regime ends. Adjusting the two regulators can be improved, thus obtaining a better indicial response.



(a) (b)

Fig. 12. The evolution of oriented orthogonal stator currents (d.c. values), for vector control with robust-adaptive rotor flux estimator ($\omega_m^* = 157$ *rad/s*).



(a) (b)

Fig. 13. The evolution of rotor speed (a) and rotor modulus flux (b) for a prescribed value ω_m*=157 rad/s.

5. Conclusion

The vector control principle re-establishes the analogy with the dc motor drive. In the case of rotor flux vector control, where the control achieved by using a reference system oriented after the rotor flux direction, the two control loops of the flux and of the electromagnetic torque get decoupled and do not influence each other.

In this reference system depending on the rotor flux, the control strategy for induction motor becomes identical with the one used for the dc motor drive. Consequently, the flux becomes stable by means of the reactive component of the stator current ($i_{sd\lambda}$) and the torque by means of its active component ($i_{sq\lambda}$).

On the orientation direction after the rotor flux, the orthogonal oriented reactive stator current component ($i_{sd\lambda}$) is similar to the excitation current of dc motor drive. If we perform the calculation on λ_r direction, we obtain the following value for $i_{sd\lambda}$:

$$i_{sd\lambda} = i_{mr} = \frac{\Phi_r}{M} = \frac{1\,Wb}{0.0617\,H} \cong 16A \tag{48}$$

which is also confirmed by the results obtained during the simulation (Fig. 12a) for the orthogonal oriented reactive stator current component ($i_{sd\lambda}$). This current is similar to an excitation current i_{ex} of an dc motor drive with ϕ_{ex} = *1Wb* and with an inductance L_{ex} equal to 61,7 mH.

$$i_{ex} = \frac{\Phi_{ex}}{L_{ex}} = \frac{1\,Wb}{0.0617\,H} \cong 16A \tag{49}$$

Before the implementation of the control algorithm on DSP systems, the simulation was the only method to prove that the system orients itself after the correct direction of the rotor flux.

6. Acknowledgements

The authors would like to acknowledge the financial support of the Romanian National Council for Project Management under Grant PN II No. 71-065 / 2007 and No. 22-137 / 2008.

7. References

Alexa, D.; Goraş, T.C. ; Sârbu, A. ; Pletea, I.V.; Filote, C. ; Ionescu, F. (2008). An Analysis of the Two-Quadrant Convertor with RNSIC, *IET Power Electronics*, Vol. 1, No. 2, (June 2008), pp. 224–234, ISSN 1755-4535

Blaabjerg, F.; Consoli, A.; Ferreira, J.A. (2005). The Future of Electronics Power Processing and Conversion, *IEEE Transactions on Industry Applications*, Vol. 41(1), 2005, pp. 3-8, ISSN 0093-9994, 2005

Bose, B. K. (2006). *Power Electronics and Motion Drives: Advances and Trends*, Elsevier, ISBN 13: 978-0-12-088405-6, USA

Bose, B. K. (2000). Energy, Environment and Advances in Power Electronics, *IEEE Transaction on Power Electronics*, Vol. 15(4), pp. 688-701, ISSN 0885-8993, 2000

Filote, C.; Graur, A. (1998). *Electrical Drives Systems. Induction machine,* Vol. 1, Editura Universităii Stefan cel Mare (Ed.), ISBN 973-98389-8-7, Suceava, Romania

Filote, C.; Alexa, D.; Pletea, I.V.; Micea, M.; Ciufudean, C.; Cozgarea, A.M. (2009). Robust-Adaptive Flux Observers in Induction Motor Drive Systems, *Proceeding of the Twelfth IASTED International Conference Intelligent Systems and Control*, ActaPress (Ed.), pp. 127-134, ISBN 978-0-88986-814-4, Cambridge, Massachusetts, USA, November 2-4, 2009

Filote, C.; Ciufudean, C.; Graur, A.; Cozgarea, A.M.; Amarandei, D.; Petrescu, C. (2008). Robust-Adaptive Flux Observer in High Performance Vector Control in Induction Motors, *International Symposium on Power Electronics, Electrical Drives, Automation*

and Motion-SPEEDAM, pp. 1097-1102, ISBN 978-1-4244-1664-6, Ischia, Italy, June 11-13, 2008

Filote, C.; Amarandei, D.; Ciufudean, C.; Graur, A.; Mandici, L. (2007). Vector Speed Control Modelling of Induction Motors Using Robust-Adaptive Flux Observer, *Proc. of The 26th IASTED International Conference Applied Simulation and Modelling – ASM 2007*, pp.479-484, ISBN 978-0-88986-687-4, Palma de Mallorca, Spain, September 1-3, 2007

Gadoue, S.M.; Giaouris, D.; Finch, J.W. (2009). Sensorless Control of Induction Motor Drives at Verry Low and Zero Speeds Using Neural Network Flux Observers, *IEEE Transations on Industrial Electronics*, Vol. 56, No. 8, pp. 3029-3039, ISSN 0278-0046, August 2009

Griva, G.; Ilas, C.; Eastham, J.F.; Profumo, F.; Vranka, P. (1997). High Performance Sensorless Control of Induction Motor Drives for Industry Applications, *Power Conversion Conference*, Vol.2, pp. 535-539, Nagaoka, Japan, 1997

Holtz, J. (2002), Sensorless Control of induction Motor Drives, Proceedings of the IEEE, Vol. 90, No. 8, pp. 1359-1394, ISSN 0018-9219, August, 2002

Kreindler, L.; Moreira, J.C.; Testa, A.; Lipo, T.A. (1994). Direct Field Orientation Controller Using the Stator Phase Voltage Thrid Harmonic, *IEEE Transaction on Industry Applications*, vol. 30, No. 2, pp. 441-447, ISSN 0093-9994, March/April, 1994

Kubota, H.; Matsuse, K.; Nakano, T. (1993). DSP-Based Speed Adaptive Flux Observer of Induction Motor, *IEEE Transactions on Industry Applications*, vol.29 , no. 2, pp. 344-348, ISSN 0093-9994, 1993

Lascu, C.; Boldea, I.; Blaabjerg, F. (2004). Direct Torque Control of Sensorless Induction Motor Drives: A Sliding-mode Appoach, *IEEE Transaction on Industrial Applications,* Vol. 40, No. 2, pp. 582-590, ISSN 0093-9994, March/April 2004.

Leonhard, W. (1990). *Control of Electrical Drives*, Springer-Verlag, ISBN 3-540-13650-9, Berlin, Germany

Li, J.; Xu, L.; Zhang, Z. (2005). An Adaptive Sliding-Mode Observer for Induction Motor Sensorless Speed Control, *IEEE Transactions on Industry Applications*, Vol. 41, No. 4, pp. 1039-1046, ISSN 0093-9994, July/August, 2005

Marchensoni, M.; Segarich, P.; Soressi, E. (1994). A Simple Approach to Flux and Speed Observation in Induction Motor Drives*, IECON*, pp. 305-310, Bologna, Italy, 1994

Schauder, C. (1989). Adaptive Speed Identification for Vector Control of Induction Motors Without Rotational Transducers, *Proc. IEEE Industry Applications Society Annual Meeting,* , pp. 493-499, San Diego, California, USA, Octomber 1-5, 1989

Pană, T. (1995). *Matlab in Electrical Drive Systems,* Universitatea Tehnica Cluj Napoca (Ed.), Romania

Umanand, L.; Bhat, S.R. (1995). On Line Estimation of the Stator Resistence of an Induction Motor for Speed Control Applications, *IEE Proceedings, Electric Power Applications*, vol.142, (March 1995), pp. 97-103, ISSN 1350-2352

Vas, P. (1998). *Sensorless vector and direct torque control,* Oxford University Press, ISBN 0-19-856465-1, Oxford, U.K.

Zhang, Z.; Xu, H.; Xu, L.; Heiman, L. (2006). Sensorless Direct Field-Oriented Control of Three-Phase Induction Motors Based on "Sliding Mode" for Washing-Machine Drive Applications, *IEEE Transaction on Industry Applications,* Vol. 42(3), pp. 694-701, ISSN 0093-9994, 2006

14

Modelling Friction Contacts in Structural Dynamics and its Application to Turbine Bladed Disks

Christian Maria Firrone and Stefano Zucca
Politecnico di Torino, Department of Mechanics, Italy

1. Introduction

Modelling the effect of friction contacts in structural dynamics (Awrejcewicz & Pyryev, 2009) has become a major issue in the design of machines (Awrejcewicz et al, 2008a) and structures. In order to obtain reliable predictions of the stress levels in vibrating structures, one of the key points is dry friction modelling (Awrejcewicz et al, 2008b).
In the dynamic design of turbine bladed disks (Srinivasan, 1997; Szwedowicz, 2008), friction damping generated at sliding contacts plays a relevant role, since due to the high modal density of these components and to the wide spectrum of the dynamic excitation , the complete detuning of the bladed disk is not often a feasible design option and resonances may occur. In order to prevent high cycle fatigue failures due to large response levels at resonance, friction dampers are commonly designed and added to bladed disks in order to reduce the vibration amplitude.
The commonest sources (Fig. 1) of friction damping in turbine bladed disks are the blade-disk interfaces (Petrov & Ewins, 2005; Charleux et al, 2006; Allara et al., 2007), the shrouds (Yang & Menq, 1998c; Petrov & Ewins, 2003; Siewert et al, 2009) located at the blade tip in order to connect adjacent blades by interference, and underplatform dampers (Csaba, 1998; Yang & Menq, 1998a-b; Sanliturk et al, 2001; Panning et al, 2003; Petrov & Ewins, 2007; Szwedowicz et al., 2008; Zucca et al, 2008; Cigeroglu et al., 2009; Firrone et al., 2011), metal devices located under the blade platforms and pressed against them during rotation by the centrifugal force.
In order to compute the forced response of bladed disks with friction contacts, commercial finite element codes are not suitable since they are based on the time integration method of the non-linear differential balance equations and they require very large calculations times, which make unfeasible any parametric analysis typical of the design phase.
For this reason, ad hoc numerical codes must be developed in order to compute the forced response in the frequency domain. These codes are based on the Harmonic Balance Method (HBM) (Cardona et al. 1998): the periodic variables (displacements and forces) are expressed as a superposition of harmonic terms by Fourier analysis and then the balance of each harmonic is imposed, turning the original nonlinear differential equations in a set of nonlinear algebraic equations.

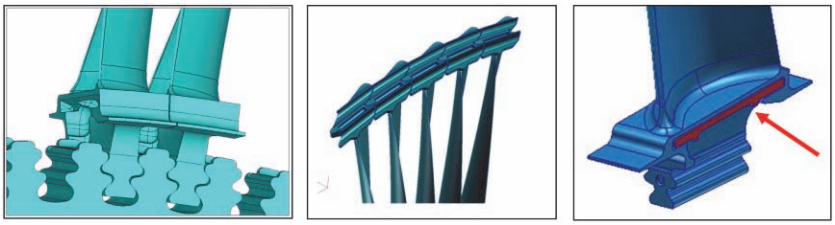


Fig. 1. Different types of friction contacts in turbine bladed disks: blade root joints, blade shrouds and underplatform dampers

In order to couple the bodies in contact when operating in the frequency domain, several contact elements have been developed. The main 4 contact elements available in the literature are:

- 1D tangential relative displacement and constant normal load (Griffin, 1980),
- 1D tangential relative displacement and variable normal load (Yang et al, 1998; Petrov & Ewins, 2003),
- 2D tangential relative displacement and constant normal load (Sanliturk & Ewins, 1996; Menq & Yang, 1998),
- 2D tangential displacement and variable normal load (Yang & Menq, 1998c).

Due to the complex kinematics of friction contacts, contact models A and C, neglecting the variation of normal loads during the system vibration, are not suitable for implementation, since they do not allow modelling partial lift-off during vibration, while the contact models B and D have established themselves as the references to model friction contacts.

All the above mentioned contact models have a common feature: for the periodical contact forces to be computed, the value of a static normal pre-load must be provided, as a result of a preliminary static analysis of the system. In (Firrone et al, 2011), it has been demonstrated that the preliminary static analysis is not necessary and that the simultaneous calculation of static and dynamic contact forces is possible.

In this chapter, the numerical methods currently employed to simulate the forced response of turbine bladed disks with friction interfaces are analyzed; in detail:

- The balance equation of the bodies in contact are deduced in the frequency domain by means of the harmonic balance method.
- The main contact elements available in the literature are described in order to highlight their main features and their effect on the dynamics of the system.
- The effect of an uncoupled solution strategy based on a preliminary static analysis followed by the dynamic analysis are highlighted and the coupled static/dynamic approach for the simultaneous calculation of the contact forces is described in detail, describing also the effect of multiple harmonics.
- The critical issues arising when the methods are applied to full scale applications are discussed, with emphasis on reduction techniques and modelling distributed contacts.
- Case studies, representing typical configurations of friction contacts in turbine bladed disks are presented and the effect of the friction contacts on the forced response curves are computed and discussed.
- The numerical methods necessary to solve the set of non-linear algebraic equations are analyzed and different Continuation strategies are discussed.

2. Balance equations and Harmonic Balance Method (HBM)

The starting point in the forced response calculation of a mechanical system with friction contacts is the development of the finite element (FE) model of the system, whose balance equations are

$$M \cdot \ddot{Q} + C \cdot \dot{Q} + K \cdot Q = F_E + F_{NL}\left(Q, \dot{Q}\right) \tag{1}$$

where M, C and K are the mass, damping and stiffness matrices of the system, Q is the displacement vector of degrees of freedom (dofs), F_E are the periodical external forces acting on the system and F_{NL} are the nonlinear forces, generated at the friction contacts by the relative displacements of the contact nodes.
In order to reduce the calculation times typical of numerical integration of non-linear systems, the harmonic balance method (HBM) can be used to compute the steady-state response of the system (Cardona et al, 1998; Griffin, 1980; Petrov & Ewins 2003).
In detail, due to the periodicity of the external excitation, also the displacements Q and the non-linear forces F_{NL} are periodical at steady-state. Therefore they can be expressed as a truncated series of harmonic terms

$$Q = Q^{(0)} + \Re\left(\sum_{n=1}^{N_H} Q^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right) \tag{2}$$

$$F_E = F_E^{(0)} + \Re\left(\sum_{n=1}^{N_H} F_E^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right) \tag{3}$$

$$F_{NL} = F_{NL}^{(0)} + \Re\left(\sum_{n=1}^{N_H} F_{NL}^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right) \tag{4}$$

where N_H is the maximum number of harmonics and ω is the fundamental frequency of the excitation forces acting on the system. The generic harmonic coefficient $X^{(n)}$ (n>1) is a complex quantity. If Equations (2)-(4) are replaced into the balance equation (1), the following sets of algebraic complex equations are deduced

$$D^{(n)} \cdot Q^{(n)} = F_E^{(n)} + F_{NL}^{(n)} \text{ with n = 0..}N_H \tag{5}$$

where $D^{(n)} = -(n\omega)^2 M + in\omega C + K$ is the n^{th} dynamic stiffness matrix of the system and the 0th order represents the static balance equation.
Since the non-linear contact forces $F_{NL}^{(n)}$ depend on the relative displacement of contact dofs and since the number of contact dofs is typically much lower than the number of total dofs, it is convenient to rearrange equations (5) in order to decouple the solution of the non-linear part of the system from its linear part. To do this, the receptance matrix $R^{(n)}$, inverse of $D^{(n)}$ matrix, can be computed and the set of balance equations can be written in the receptance form

$$Q^{(n)} = Q_E^{(n)} + R^{(n)} \cdot F_{NL}^{(n)} \tag{6}$$

where the first term at the right hand side of equation (6) is the linear response due to the external excitation. i.e. $Q_E^{(n)} = R^{(n)} \cdot F_E^{(n)}$, while the second terms takes into account the contribution of the non-linear forces.

Then, the displacement vector $Q^{(n)}$ can be split in the non-linear dofs $Q_{NL}^{(n)}$ where the nonlinear contact forces act and all the other linear dofs $Q_{LN}^{(n)}$, and the balance equations (6) become

$$\begin{Bmatrix} Q_{NL}^{(n)} \\ Q_{LN}^{(n)} \end{Bmatrix} = \begin{Bmatrix} Q_{E,NL}^{(n)} \\ Q_{E,LN}^{(n)} \end{Bmatrix} + \begin{bmatrix} R_{NL,NL}^{(n)} & R_{NL,LN}^{(n)} \\ R_{LN,NL}^{(n)} & R_{LN,LN}^{(n)} \end{bmatrix} . \begin{Bmatrix} F_{NL}^{(n)} \\ 0 \end{Bmatrix} . \tag{7}$$

Since the non-linear contact forces $F_{NL}^{(n)}$ only depend on the displacement $Q_{NL}^{(n)}$ of non-linear dofs, only the equation

$$Q_{NL}^{(n)} = Q_{E,NL}^{(n)} + R_{NL,NL}^{(n)} \cdot F_{NL}^{(n)} \tag{8}$$

is non-linear and must be solved iteratively with a non-linear solver, while the response of the linear dofs $Q_{LN}^{(n)}$ can be computed with the equation

$$Q_{LN}^{(n)} = Q_{E,LN}^{(n)} + R_{LN,NL}^{(n)} \cdot F_{NL}^{(n)} \tag{9}$$

once the non-linear forces $F_{NL}^{(n)}$ are known. It must be observed that the balance equations (8) are coupled to each other, because the arbitrary harmonic component $F_{NL}^{(n)}$ of the non-linear contact forces depends on all the harmonic components of the displacement of non-linear dofs Q_{NL}. In order to solve the non-linear balance equations (8) a contact model is necessary to compute the harmonic components $F_{NL}^{(n)}$ of the periodical contact forces from a given set of harmonics components $Q_{NL}^{(n)}$ of the nodal displacements of non-linear dofs.

3. Contact elements

In the technical literature, the problem of modeling periodical contact forces at friction contacts for their implementation in numerical solvers for the forced response of mechanical systems with friction contacts has been addressed by several authors, leading to several contact models.

All the models are based on the flow-chart show in Fig. 2 and based on the following steps:

- For given Fourier coefficients $Q_{NL}^{(n)}$ of the displacement of contact nodes of equation (8), relative displacements $\Delta Q_{NL}^{(n)}$ are computed from the contact kinematics.
- Inverse Fast Fourier Transform (IFFT) is applied to $\Delta Q_{NL}^{(n)}$ to compute the periodical relative displacements ΔQ_{NL} in the time domain.
- Constitutive laws of the contact model are used to compute the periodical non-linear contact forces F_{NL}.
- Fast Fourier Transform is applied to F_{NL} to compute the harmonic components $F_{NL}^{(n)}$ of the periodical contact forces to be used in the balance equations (8).

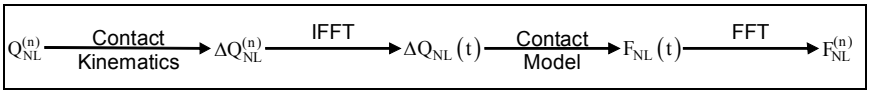


Fig. 2. Typical flow-chart of contact models.

In the literature 4 main contact models exist for the calculation in the frequency domain of the forced response of mechanical systems with friction contacts:

- 1D tangential relative displacement and constant normal load (Griffin, 1980),
- 1D tangential relative displacement and variable normal load (Yang et al, 1998; Petrov & Ewins, 2003),
- 2D tangential relative displacement and constant normal load (Sanliturk & Ewins, 1996; Menq & Yang, 1998),
- 2D tangential displacement and variable normal load (Yang & Menq, 1998c).

In this section the above listed contact models are described and used to simulate the forced response calculation of a single dof system in order to highlight their effect on the system dynamics and to compare their features and their performance.

3.1 1D tangential relative displacement and constant normal load

This contact model (Fig. 3) has been described and used for the first time in (Griffin, 1980). It is able to model 1D relative displacement in the tangential direction and to take into account the effect of a constant normal load acting on the contact.

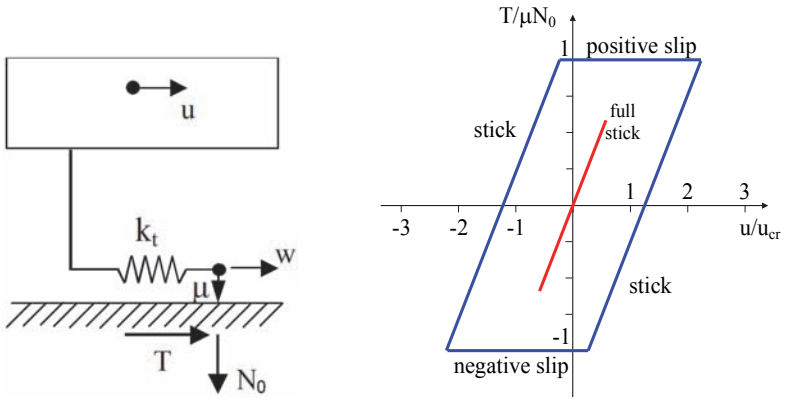


Fig. 3. 1D relative displacements and constant normal load: contact model (left) and typical hysteresis cycles (right)

The tangential contact stiffness is modelled by a spring of stiffness k_t, a coefficient of friction μ is assumed between the contact surfaces while constant static normal load N_0 keeps the bodies in contact. For a given periodic relative displacement of contact nodes u(t) the periodic tangential force T(t) is computed. The amount of tangential slip between the contact surfaces is w(t). Two contact states can be modelled: stick and slip. In the sticking mode, the contact is elastic, no slip occurs, and the tangential force is

$$T = k_t \cdot (u - w) \;\; with \;\; \dot{w} = 0 \,. \tag{10}$$

In the slipping mode, the modulus of the tangential force is equal to the Coulomb limit value and its versus depends on the versus of the slipping velocity according to the equation

$$T = \mathrm{sgn}(\dot{w}) \cdot \mu \cdot N_0 \,, \tag{11}$$

where sgn(x) is the sign function whose values are

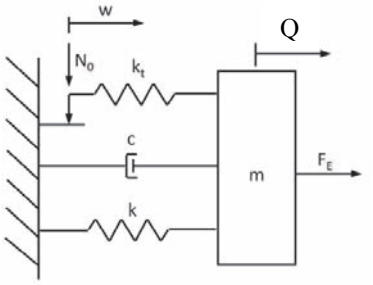
$$\mathrm{sgn}(\dot{w}) = \begin{cases} -1 \;\; if \;\; \dot{w} < 0 \;\; (negative\ slip) \\ 1 \;\; if \;\; \dot{w} > 0 \;\; (positive\ slip) \end{cases} \tag{12}$$

Stick and slip alternate each other according to the transition criteria shown in Table 1

Contact state	Transition criteria
Stick-to-slip	$\lvert T \rvert = \mu \cdot N_0$
Slip-to-stick	$\dot{w} = 0$

Table 1. 1D contact with constant normal load: transition criteria.

If the amplitude of the tangential relative displacement u(t) is lower than the critical value $u_{cr}=\mu \cdot N_0/k_t$ the contact is in full stick conditions (red line in Fig. 3) and vibration energy is not dissipated by friction. If the critical value u_{cr} is exceeded alternating stick and slip occur (blue line in Fig. 3) and the area of the hysteresis cycle is the energy dissipated per cycle.



k = 3.55e4
c = 0.3
m = 1
k_t = 1e4
μ = 0.5

Fig. 4. Single dof description

In order to demonstrate the effect of this contact element on the dynamics of a vibrating system, a single dof system (Fig. 4) is here analyzed. The vibrating system is in contact with a fixed wall and therefore, due to the simple contact kinematics, the relative displacement u(t) in the tangential direction is equal to the absolute displacement Q(t) of the mass.
The forced response is computed for different values of the static normal load N_0 (Fig. 5 – left) and of external excitation F_E (Fig. 5 – right) in order to show the effect of friction contacts on a vibrating structure, using only the first order of the Fourier series.

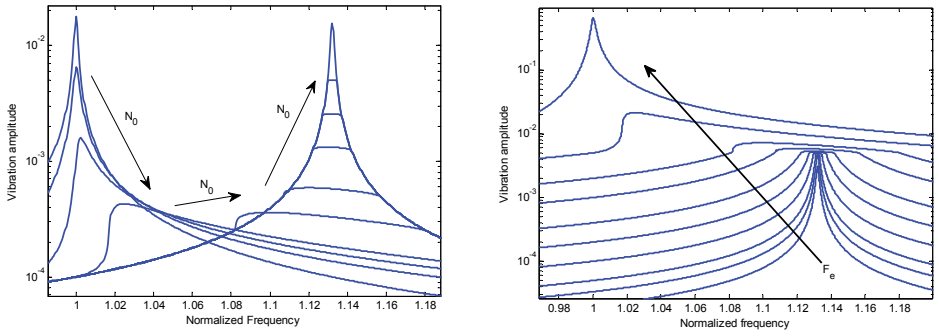


Fig. 5. Effect of N_0 (left) and of F_E (right) on the dynamics of the system

In the first case, at $N_0 = 0$ the response of the linear system is computed since the tangential force T is null. As N_0 growing larger, the maximum response of the system decreases, reaches a minimum value and then increases again as the contact tends towards the full sticking mode. The contact affects not only the damping of the system but also its stiffness since the resonance frequency of the system becomes larger as N_0 increases.
The effect of the external excitation is the opposite. For small values of F_E the contact is in the full sticking mode, while as the excitation force becomes larger, the system tends towards the linear condition.
The effect of N_0 and F_E on the system dynamics can be shown by means of two compact diagrams (Fig. 6) called the optimization and the performance plot, respectively. The optimization plot (Yang & Menq, 1998a,Yang & Menq 1998b) can be used to identify the optimum value of N_0 corresponding to the minimum vibration amplitude of the system.

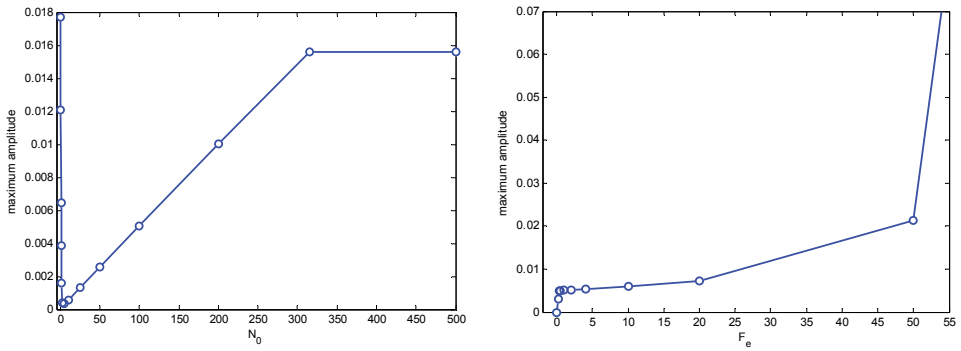


Fig. 6. Optimization and performance plots.

On the contrary, the performance plot can be used to identify the range of external excitation for a robust friction contact design, corresponding to the flat part of the performance curve, where large variations of the external force produce small variations of the system response.

3.2 1D tangential relative displacement and variable normal load.

The contact model described in Section 3.1 is not able to model the effect of the variation of the contact normal load, due to a periodical normal relative displacements of contact points, on the hysteresis cycle and on the energy dissipated by friction.
This important feature can be taken into account by means of the contact model originally developed in (Yang et al., 1998) for the single HBM and then extended in (Petrov & Ewins, 2003) for the multi-harmonic balance method (MHBM).
According to this model, shown in Fig. 7, the tangential and normal contact stiffness are modeled by springs of stiffness k_t and k_n, respectively, a coefficient of friction μ is assumed between the contact surfaces.
The relative displacements in the tangential and normal directions are u(t) and v(t) respectively, while the amount of tangential slip between the contact surfaces is w(t).
The normal contact force N(t) is defined as

$$N = \max\left(N_0 + k_n \cdot v, 0\right) \tag{13}$$

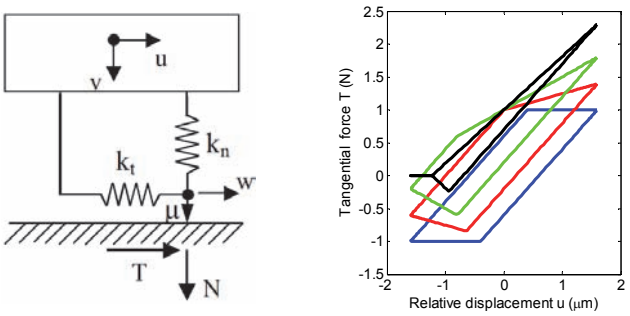


Fig. 7. 1D relative displacements and variable normal load: contact model (left) and typical hysteresis cycles (right)

where N_0 is the static normal load. If N_0 is positive, the bodies are in contact before vibration starts, while if N_0 is negative an initial gap exists between the two bodies. According to the value of the normal relative displacement v(t) three conditions are possible: full contact, partial lift-off and full-lift off. According to equation (13), when lift-off occurs, the normal contact load is set equal to 0, since negative values are not acceptable.

$$T=\begin{cases} k_t\cdot(u-w) & \text{sticking mode} \\ \operatorname{sgn}(\dot{w})\cdot\mu\cdot N & \text{slipping mode} \\ 0 & \text{lift-off mode} \end{cases} \tag{14}$$

stick, slip and lift-off (see Equation (14) for the constitutive equations) may alternate each other during the periodic vibration, according to the transition criteria shown in Table 2.

Initial state	Final state	Criteria
Stick	Slip	$\lvert T\rvert = \mu\cdot N$
	Lift-off	$N=0$ *and* $\dot{N}<0$
Slip	Stick	$\dot{w}=0$
	Lift-off	$N=0$ *and* $\dot{N}<0$
Lift-off	Stick	$N=0$ *and* $\dot{N}>0$ and $\lvert\dot{T}\rvert<\mu\cdot\dot{N}$
	Slip	$N=0$ *and* $\dot{N}>0$ and $\lvert\dot{T}\rvert>\mu\cdot\dot{N}$

Table 2. 1D contact with variable normal load: transition criteria.

The effect of the variable normal contact load N(t) on the hysteresis cycle of the tangential contact load T(t) is shown in Fig. 7, where a sinusoidal v(t) is in-phase with a sinusoidal u(t). If v(t) is null, the contact model is coincident with the model of Section 3.1 (blue line). As the amplitude of v(t) growing larger, the shape and the area of the cycle change. In case of full contact (blue, red and green lines) the alternating stick-slip phenomenon occurs, but when the value of v(t) becomes so large that partial lift-off occurs (black line), then a horizontal segment at T(t)=0 appears, and the cycle becomes a triangle.
Fig. 8 shows the effect of the variable normal contact load on the dynamic behavior of the vibrating structures.

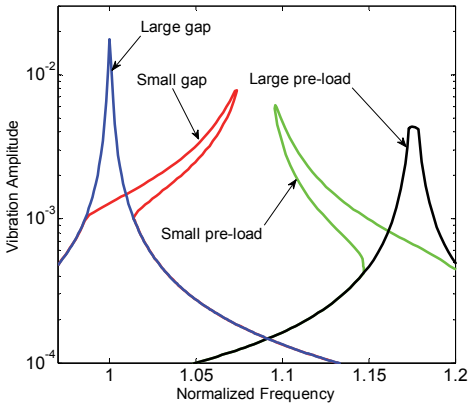


Fig. 8. Effect of N_0 on the dynamics of the system with variable normal contact load.

The blue line refers to such a large gap (N_0<<0) that no contact occurs during the vibration and the system response corresponds to the response of the free system without contact. The red line, partly overlapping the blue line, represents a case of small gap (N_0<0) which is closed during vibration due to the normal relative displacements of the contact points. The stiffening effect of the closing contact on the resonance peak is clearly visible, as well as the existence of multiple solutions for a certain range of vibration frequencies. At the same way, the green curve refers to a positive pre-load (N_0>0), not large enough to prevent lift-off during vibration. In this case, the softening effect of the opening contact on the resonance peak is visible. Finally, the black curve represents a case of a large pre-load (N_0>>0), which imposes full contact conditions during the periodic vibration.

3.3 2D tangential relative displacement and constant normal load.

The contact models described in Sections 3.1 and 3.2 are able to simulate the behavior of friction contact in case of 1D tangential relative displacements. However, in several applications, 2D tangential relative displacements of the contact points occur and modeling these contact assuming a linear trajectory can lead to an underestimation of the friction damping (Griffin & Menq, 1991). In order to model this feature, two different modeling techniques are available:

1. to use two 1D contact elements like those described in Sections 3.1 and 3.2, placed orthogonal to each other, in order to take into account the 2D trajectory of the contact points on the contact plane;
2. to use a 2D contact element able to take into account the coupling between the two orthogonal components of the tangential relative displacements.

The second approach is possible using a contact model developed independently in (Sanliturk & Ewins, 1996) and (Menq & Yang, 1998) characterized by 2D tangential relative displacement and constant normal load (Fig.9, left).

Being (x,y) the contact plane, the tangential contact stiffness is modeled by two springs of stiffness k_{tx} and k_{ty} along the two orthogonal directions, a coefficient of friction μ is assumed between the contact surfaces while constant static normal load N_0 keeps the bodies in contact.

For a given periodic relative displacement of contact nodes $u(t) = \{u_x(t); u_y(t)\}$ the periodic tangential force $T(t) = \{T_x(t); T_y(t)\}$ is computed. The amount of tangential slip between the contact surfaces is $w(t) = \{w_x(t); w_y(t)\}$.
Two contact states can be modeled: stick and slip. In the sticking mode, the contact is elastic, no slip occurs, and the tangential force is

$$\begin{Bmatrix} T_x \\ T_y \end{Bmatrix} = \begin{bmatrix} k_{tx} & 0 \\ 0 & k_{ty} \end{bmatrix} \cdot \left(\begin{Bmatrix} u_x \\ u_y \end{Bmatrix} - \begin{Bmatrix} w_x \\ w_y \end{Bmatrix} \right) with \begin{Bmatrix} w_x \\ w_y \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{15}$$

In the slipping mode, the modulus of the tangential force is equal to the Coulomb limit value

$$\sqrt{T_x^2 + T_y^2} = \mu \cdot N_0 \tag{16}$$

On the contact plane, Equation (16) represents the equation of a circumference of radius equal to $\mu \cdot N_0$. In slipping conditions, the friction contact force is oriented along the circumference radius (see Fig.9, right). Stick and slip may alternate each other according to the transition criteria fully described in (Sanliturk & Ewins, 1996; Menq & Yang, 1998). In Fig. 9 (right), different trajectories of the tangential contact force are shown over the contact plane. All the curves refer to sinusoidal relative displacements $u_x = u_{x0} \sin(2\pi t)$ and $u_y = u_{y0} \sin(2\pi t + \phi)$, with $u_{x0}/u_{y0} = 2.5$ and $\phi = \pi/2$. The red curves refer to cases where the contact is in full sticking conditions and no slip occurs between the contact points. The blue curves, on the contrary represent three cases where the vibration amplitude is large enough to induce alternating stick-slip phenomena. In detail, the slipping parts of the cyclic force are those where the blue lines are overlapped over the black circle which represents the Coulomb limit which cannot be exceeded by the tangential force T(t).
If the vibration amplitude increases further, the blue line will overlap completely the black circle, having a full slip condition, which cannot occur in case of 1D tangential relative displacement.

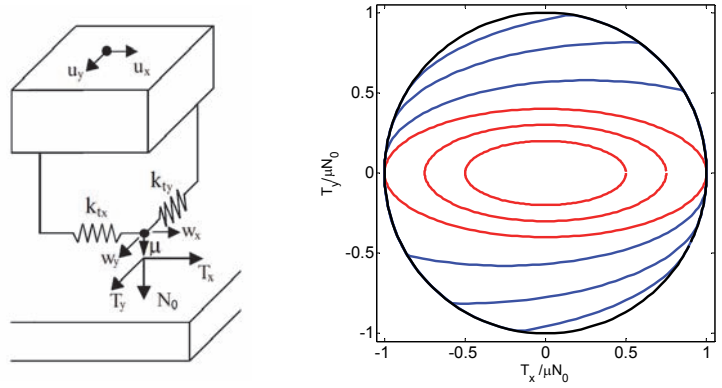


Fig. 9. 2D relative displacements and constant normal load: typical trajectories of the tangential contact forces.

In order to show the effect of the 2D contact kinematics on the behaviour of vibrating structures, a simple dynamic system is here analyzed. A mass vibrates over an (x,y) plane.

Q_x and Q_y are the modal displacement components, they are harmonic quantities $Q_x = Q_{x0} \sin(\omega t)$ and $Q_y = Q_{y0} \sin(\omega t+\pi/2)$, the ratio $Q_{x0}/Q_{y0} = 1$, so that a circular motion of the system is obtained.

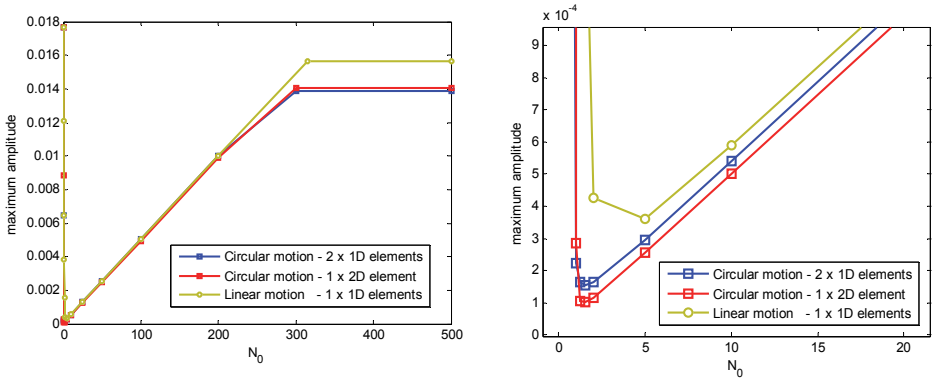


Fig. 10. Optimization curves: full view (left) and zoom around the optimum value (right)

The system dynamics is analyzed for different values of the normal pre-load N_0 and the resulting optimization curves are computed. The contact is modeled in three different ways. In the first way, the y-component of the trajectory is neglected and the contact is modeled with a 1D contact element oriented along the x direction. In the second case, two 1D contact elements (see Section 3.1) are oriented along the x and y directions respectively, while in the third case one 2D contact element described in the current section is used. The optimization curves of the friction contact are shown in Fig. 10.

If the trajectory of the contact point is modeled as linear (1st case) neglecting the y component, the friction damping is not correctly predicted (Griffin & Menq, 1991) and the minimum vibration amplitude is overestimated with respect to the 2nd and the 3rd case. The comparison between the 2nd and the 3rd case does not show significant differences, indicating that both the uncoupled approach based on two orthogonal 1D contact elements and the coupled approach based on one 2D contact element catch the fundamental features of the bi-dimensional trajectory of the contact point.

3.4 2D tangential displacement and variable normal load.

The most advanced contact model available in the literature, consider both the 2D trajectory of the contact points on the contact plane and the periodical variation of the normal load (Yang & Menq, 1998c), see Fig.11, left.

Being (x,y) the contact plane, the tangential contact stiffness is modeled by two springs of stiffness k_{tx} and k_{ty} along the two orthogonal directions, a coefficient of friction μ is assumed between the contact surfaces while variable static normal load defined as

$$N = \max\left(N_0 + k_n \cdot v, 0\right) \tag{17}$$

keeps the bodies in contact.

For a given periodic relative displacement of contact nodes $u(t) = \{u_x(t); u_y(t)\}$ the periodic tangential force $T(t) = \{T_x(t); T_y(t)\}$ is computed. The amount of tangential slip between the contact surfaces is $w(t) = \{w_x(t); w_y(t)\}$.

Three contact states can be modeled: stick, slip and separation. In the sticking mode, the contact is elastic, no slip occurs, and the tangential force is

$$\begin{Bmatrix} T_x \\ T_y \end{Bmatrix} = \begin{bmatrix} k_{tx} & 0 \\ 0 & k_{ty} \end{bmatrix} \cdot \left(\begin{Bmatrix} u_x \\ u_y \end{Bmatrix} - \begin{Bmatrix} w_x \\ w_y \end{Bmatrix} \right) with \begin{Bmatrix} w_x \\ w_y \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \text{ and } N \geq 0. \tag{18}$$

In the slipping mode, the modulus of the tangential force is equal to the Coulomb limit value and its directed parallel to the slip velocity $\dot{w}$.

$$\begin{Bmatrix} T_x \\ T_y \end{Bmatrix} = \mu N \Big/ \left\| \dot{w} \right\| \cdot \begin{Bmatrix} \dot{w}_x \\ \dot{w}_y \end{Bmatrix} \quad \text{with } \left\| \dot{w} \right\| = \sqrt{\dot{w}_x^2 + \dot{w}_y^2} \tag{19}$$

When separation of the contact points occurs, both the normal and the tangential contact forces are equal to zero. Transitions between contact states are governed by transition criteria described in (Yang & Menq, 1998c).

Four examples of contact force trajectories over the contact plane are shown in Fig. 11, where the blue lines represent the contact force and the black line the Coulomb limit curve. In detail, they represent:

a. A case with constant normal load $N=N_0$, which could be also simulated with the contact model described in Section 3.3
b. A case with a small variation of N, with clearly visible alternating stick-slip states
c. A case with a large variation of N, with suppression of one of the stick states
d. A case with a very large variation of N, inducing partial separation of the contact points, corresponding to the point where the trajectory passes through the origin.

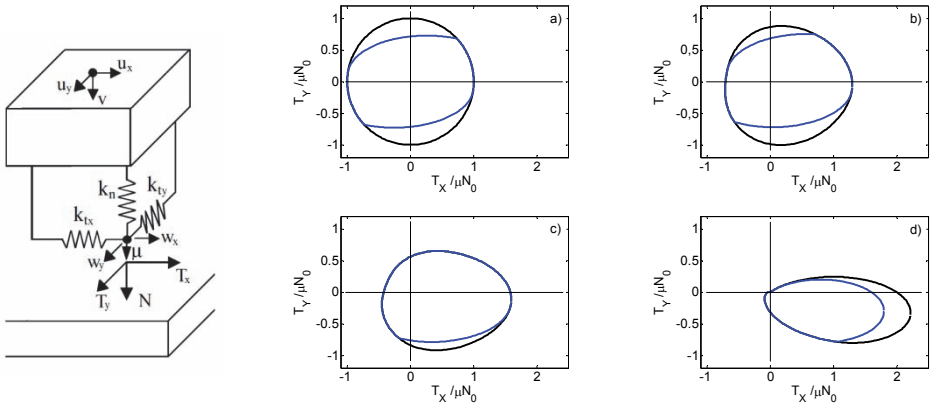
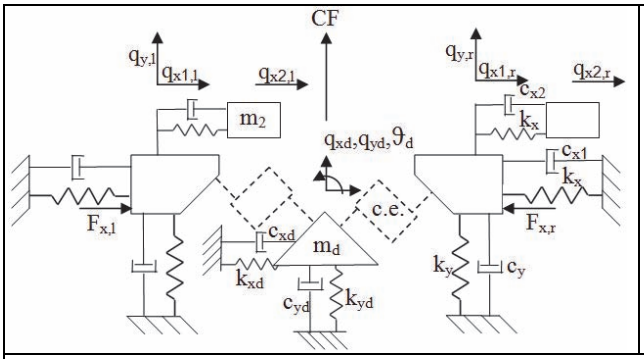


Fig. 11. Tangential contact force trajectories on the contact plane

4. Static/dynamic coupling and higher order harmonics

The contact models introduced in Section 3 share a common feature, i.e., they all need a normal load N_0 as an input parameter together with the variable relative displacements u(t) and v(t) in order to calculate the hysteresis loop of the tangential force T(t). It is easy to understand that an accurate calculation of N_0 is of primary importance when those contact

models are applied to practical problems to calculate the forced response of structures with friction joints, since N_0 determines the amount of stick and slip of the contact, therefore the amount of friction damping. The value of N_0 cannot be chosen in an arbitrary way, since it represents the static normal pre-load acting on the friction contact. The commonest strategy is to use the pre-loads obtained by the solution of the static equilibrium of the structure, assuming that the oscillating response is superimposed to the static deformed shape. In fact, the static equilibrium can take into account the influence of the weight of the bodies in contact or the interference and the presence of the centrifugal force for rotating components.
In order to address the issue of the choice of N_0 over the contact and its consequence on the calculation of the forced response of the structure, the simple structure shown in Fig. 12 is modelled and the differential equilibrium equations of the system are integrated in the time domain (DTI, Direct Time Integration). The 2D structure of Figure 12 is a simplified representation of two blades of a blade array of a turbine rotor and a blade-to-blade friction damper (the so called underplatform damper, see introduction and Fig. 1), i.e., a metal component pressed against the blades platforms during rotation by the centrifugal force. Each blade is modelled with two masses (m_1 the platform and m_2 the blade airfoil) and has three dofs (q_{x1}, q_{x2}, q_y). The underplatform damper is modelled as a rigid body with 3 dofs, whose center of mass is connected to ground by a set of linear springs and viscous dampers. Moreover, the underplatform damper has a triangular shape and the contact with the two platforms is modeled with a node-to-node contact element allowing for a 1D relative displacement and a variable normal load (contact model B, Section 3.2, see Fig. 7). The positive axis for the tangential relative displacement u for both the contact elements is oriented toward the apex of the underplatform damper (Fig. 12b) while, accordingly to Fig. 7, v is positive during contact. The system has nine dofs and the values of the structural parameters are listed below the Fig. 12. The centrifugal load CF is applied to the center of mass of the underplatform damper while two symmetrical harmonic horizontal forces $F_x(t)$ act on the blades.



$k_x = k_y = 3.95e5$	$k_{xd}=k_{yd}=k_{\vartheta d}=3.95e3$	k_t = 1e5	$F_{x,r}=F_{x,l}=15$
$c_{x1}=56.2$; $c_{x2}=28.1$; $c_y=62.8$	$c_{xd}=c_{yd}=c_{\vartheta d}=62.8$	k_n = 1e5	CF=30
$m_1=m_2=m_d=1$		μ = 0.5	

Fig. 12. Multiple dof description, blade-to-blade friction damper application

4.1 Integration of the balance equations in the time domain

The second order differential equilibrium equations of the nine degrees of freedom of the system of Figure 12 are:

$$
\begin{aligned}
&m_1\ddot{q}_{x1,l} + c_{x1}\dot{q}_{x1,l} + k_{x1}q_{x1,l} + c_{x2}\left(\dot{q}_{x2,l} - \dot{q}_{x1,l}\right) + k_{x2}\left(q_{x2,l} - q_{x1,l}\right) = F_{x,l} - F_{Cx,l}\\
&m_1\ddot{q}_{y,l} + c_y\dot{q}_{y,l} + k_y q_{y,l} = -F_{Cy,l}\\
&m_2\ddot{q}_{x2,l} - c_{x2}\left(\dot{q}_{x2,l} - \dot{q}_{x1,l}\right) - k_{x2}\left(q_{x2,l} - q_{x1,l}\right) = 0\\
&m_1\ddot{q}_{x1,r} + c_{x1}\dot{q}_{x1,r} + k_{x1}q_{x1,r} + c_{x2}\left(\dot{q}_{x2,r} - \dot{q}_{x1,r}\right) + k_{x2}\left(q_{x2,r} - q_{x1,r}\right) = F_{x,r} - F_{Cx,r}\\
&m_1\ddot{q}_{y,r} + c_y\dot{q}_{y,r} + k_y q_{y,r} = -F_{Cy,r}\\
&m_2\ddot{q}_{x2,r} - c_{x2}\left(\dot{q}_{x2,r} - \dot{q}_{x1,r}\right) - k_{x2}\left(q_{x2,r} - q_{x1,r}\right) = 0\\
&m_d\ddot{q}_{xd} + c_{xd}\dot{q}_{xd} + k_{xd}q_{xd} = F_{Cx,l} + F_{Cx,r}\\
&m_d\ddot{q}_{yd} + c_{yd}\dot{q}_{yd} + k_{yd}q_{yd} = F_{Cy,l} + F_{Cy,r} + CF\\
&I_d\ddot{\vartheta}_d + c_{\vartheta d}\dot{\vartheta}_{yd} + k_{\vartheta d}\vartheta_d = -F_{Cx,l}b_l + F_{Cy,l}a_l - F_{Cx,r}b_r + F_{Cy,r}a_l
\end{aligned}
\tag{20}
$$

where $F_{C\,x,l}$, $F_{C\,y,l}$, $F_{C\,x,r}$, $F_{C\,y,r}$ are the contact forces acting on the damper projected over the horizontal and vertical direction respectively for the left and right side. If α_L and α_R are the left and right angle between the left and right inclined platform and the vertical axis, it is:

$$
\begin{aligned}
&F_{Cx,l} = T_l\sin(\alpha_L) + N_l\cos(\alpha_L);\quad F_{Cy,l} = T_l\cos(\alpha_L) - N_l\sin(\alpha_L)\\
&F_{Cx,r} = -T_r\sin(\alpha_R) - N_r\cos(\alpha_R);\quad F_{Cy,r} = T_r\cos(\alpha_R) - N_r\sin(\alpha_R)
\end{aligned}
\tag{21}
$$

Equilibrium equations (20) are rewritten in terms of a first order differential problem and are solved using the explicit Runge-Kutta formula, Dormand-Prince pair (ODE45) in MATLAB code. The external loads (centrifugal force and blade excitations) are applied according to two load-steps shown in Fig. 13a and Fig. 13b: first a quasi-static application of the centrifugal force determines the static deformation of the system and the static contact loads over the contacts. During this load step the forces on the blades are set to zero ($F_{x,l}$= $F_{x,r}$=0). Second, once the static equilibrium of the system is achieved, the harmonic excitations on the two blades $F_{x,l}$ and $F_{x,r}$ are applied by linearly increasing their amplitude (Fig. 13b). The last set of displacements of the first load step is used as first set for the second load step, while the centrifugal force CF is kept constant to 30 N. The excitation frequency is equal to the first natural frequency of the blade (61.8 Hz). The contact forces are calculated once the relative displacements at the two contacts are calculated from the absolute displacements of the system:

$$
\begin{aligned}
&u_L = (q_{x1,l} - q_{xd,l})\cdot\sin\alpha_L + (q_{y,l} - q_{yd,l})\cdot\cos\alpha_L;\quad u_R = -(q_{x1,r} - q_{xd,r})\cdot\sin\alpha_R + (q_{y,r} - q_{yd,r})\cdot\cos\alpha_R\\
&v_L = (q_{x1,l} - q_{xd,l})\cdot\cos\alpha_L - (q_{y,l} - q_{yd,l})\cdot\sin\alpha_L;\quad v_R = -(q_{x1,r} - q_{xd,r})\cdot\cos\alpha_R - (q_{y,r} - q_{yd,r})\cdot\sin\alpha_R
\end{aligned}
\tag{22}
$$

where $q_{xd,l}$, $q_{yd,l}$, $q_{xd,r}$, $q_{yd,r}$ are the local absolute displacement of the damper at the left and right side respectively:

$$
\begin{Bmatrix} q_{xd,i} \\ q_{yd,i} \end{Bmatrix} = \begin{bmatrix} 1 & 0 & -b_i \\ 0 & 1 & a_i \end{bmatrix} \cdot \begin{Bmatrix} q_{x,d} \\ q_{y,d} \\ \vartheta_d \end{Bmatrix} \quad i = l, r
\tag{23}
$$

being a_i and b_i the horizontal and vertical coordinates of the two contact points from the center of mass of the damper. Then the system reacts under the effect of the external loads and of the contact forces whose transitions are calculated according to the criteria in Table 2.

Since the structure is symmetric and the loads are applied symmetrically, also the response of the two blades is symmetric, therefore only the results referred to the left blade are shown. Fig. 13c and Fig. 13d show the tip displacement of the blade (mass m_2) respectively for the first and second load step. It can be noted that in the first load step, as expected, the left blade is pushed toward negative values of the x-axis (symmetrically, the right blade is pushed toward the positive x-axis of the same amount). The tangential contact force T acting on the left side of the damper is displayed in Fig. 13e-f together with the boundaries ±μN.

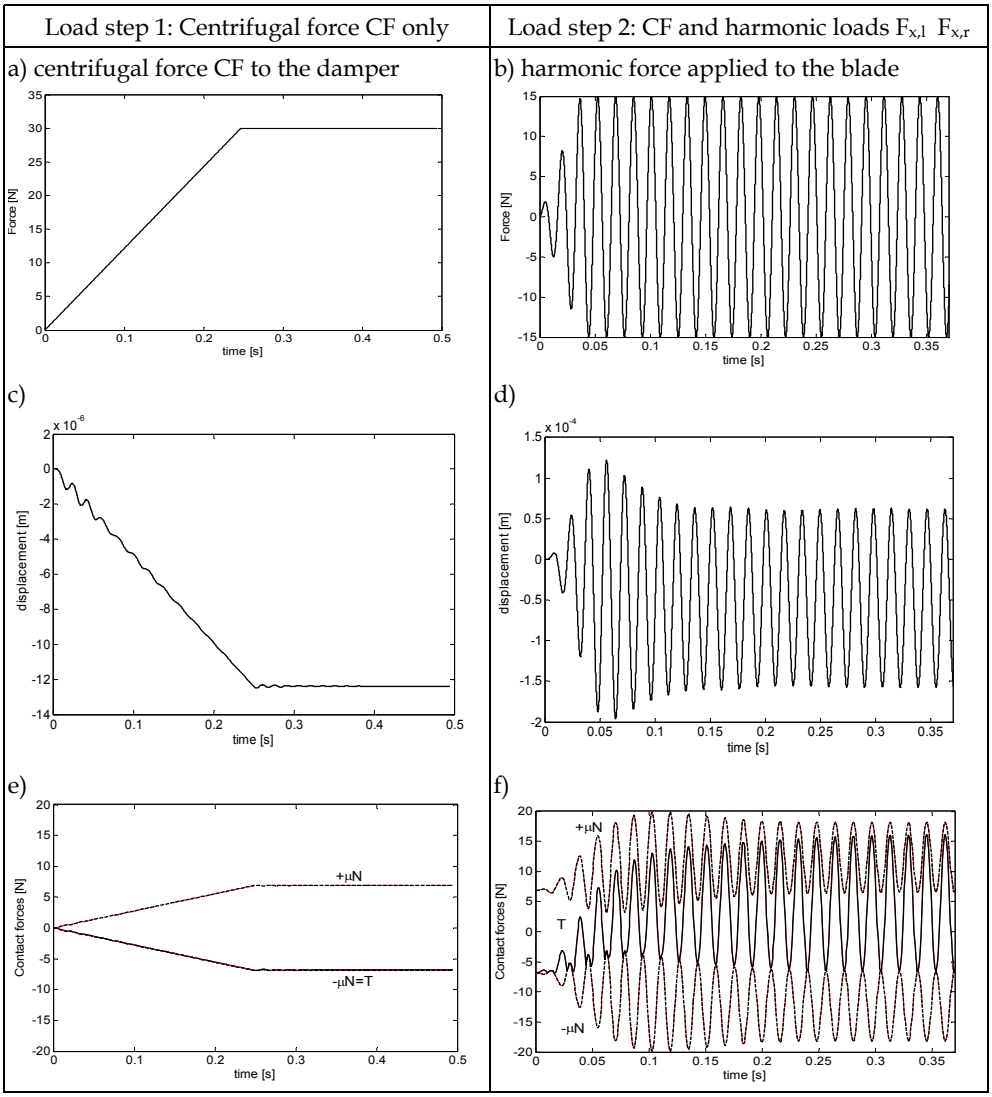


Fig. 13. DTI of the blade-to-blade damper system: load step 1 (left) and load step 2 (right).

The quasi-static application of the centrifugal force in the first load step gives a value of the tangential contact force T equal to –μN at the end of the simulation. The force distribution corresponding to the static equilibrium at the end of the first load step is shown in Fig. 14. At the end of the second load step it is possible to see that the underplatform damper is fully stuck since the tangential relative motion at the contacts is not large enough to cause the slip of the surfaces. The stick state is visible in Fig. 13f since the tangential contact force T is not large enough to cross the boundary limit ±μN. In particular, T is tangent only to the lower boundary limit –μN. It must be finally highlighted that, when the steady-state response is reached at the end of the second load step, the static value of the normal and tangential forces changed with respect to the static values at the end of the first load step: the static normal load nearly doubled changing from 14N to 25N while the static tangential force increased from -7N to 4N.

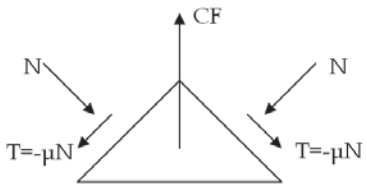


Fig. 14. Forces distribution on the damper after load step 1

This is a clear example of how the static contact forces over the contacts, both normal and tangential, are strongly influenced by the dynamics of the system and in general are not equal to the contact forces that can be carried out by the calculation of a static equilibrium of the structure.

4.2 Forced response in the frequency domain: the static/dynamic coupling of the balance equations and comparison with the uncoupled approach

The solution of the equilibrium equations using the direct time integration is usually a prohibitive approach for industry due to the large number of dofs used to model the complexity of a real component, in particular when the time is compared with the time-to-market constraints. For these reasons the HBM is a valuable tool based on the approximation of the time-depending quantities to their first harmonics according to the Fourier series approximation (see Section 2). The commonest approach that is found in literature and in the industry applications to solve iteratively the set of Equations (8) is based on the algorithm shown in Fig. 15a.

It is possible to see that the contact model employed within the algorithm needs a static normal load as an input parameter as already explained in Section 3. The relative motion between the mating surfaces is then processed within the contact model with the static normal load chosen a priori in order to calculate the contact forces.

Unfortunately, this approach may result not reliable if the choice of N_0 is based on a static equilibrium of the structure as proved in Section 4.1. Sometimes, as demonstrated in (Zucca et al., 2008), the range of possible solutions that can be obtained by assuming different conditions for the static equilibrium is so large that the simulation is not useful to assess the damping effectiveness. In order to avoid a preliminary static calculation of the contact forces, the authors presented a refined contact model (Firrone et al, 2011) which couples the static and dynamic equilibrium of the structure. The basic idea is to link the normal and

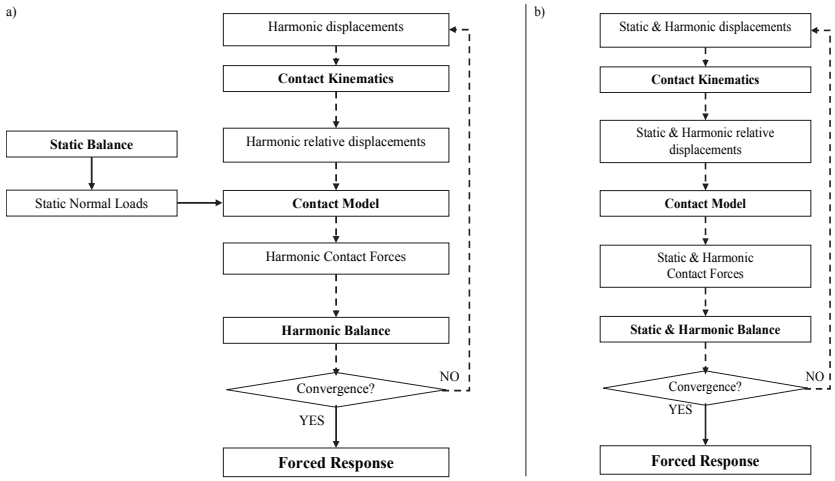


Fig. 15. Uncoupled (a) and coupled (b) algorithms for the forced response calculation

tangential static contact forces directly to the normal and tangential static relative displacements $u^{(0)}$, $v^{(0)}$ at the contacts. In this case the quantities u(t), v(t), $u^{(0)}$, $v^{(0)}$ are the input data of the contact model instead of u(t), v(t) and N_0. In detail, instead of using Equation (13) based on a value of N_0 that is chosen a priori, N(t) is defined as

$$N = \max\left[\, k_n \cdot v(t),\, 0 \,\right] = \max\left[\; k_n \cdot \Delta + k_n \cdot v^{(0)} + k_n \cdot \Re\left(v^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right),\, 0 \;\right], \tag{24}$$

being Δ the design value of interference (Δ>0) or gap (Δ<0) existing between the contact points before the external forces act on the system.

The static component of the contact normal load $N^{(0)}$, necessary to compute the periodic tangential load T(t), is directly linked to the static component of the normal relative displacement of the bodies in contact $v^{(0)}$ and to the interference (or gap) Δ. The influence of a static relative displacement $u^{(0)}$ over the contact plane on the determination of T(t) is now discussed. If during one period of oscillation the contact enters the slip-state or the lift-off state, it can be demonstrated that the static value $T^{(0)}$ of the tangential contact force T(t) is unique. This is also visible by looking at the hysteresis loop of Fig. 7 where, once the variable quantities u(t) and v(t) are determined as well as N(t) according to Equation (24), also the static value $T^{(0)}$ is directly determined. On the contrary, if the contact is fully stick, more than one solution exists. A clear example is visible in Fig. 13 f where the tangential contact force T(t) can vary being tangent either to μN or to –μN according to the limits given by the Coulomb's law. All the other solutions, where T(t) is in the middle of the two limits, are valid solutions as well. As a consequence, the static value $T^{(0)}$ is bounded within two values:

$$T^{(0)}_{\min} \le T^{(0)} \le T^{(0)}_{\max} \tag{25}$$

where $T^{(0)}_{min}$ and $T^{(0)}_{max}$ correspond to $T^{(0)}$ when T(t) is tangent respectively to –μN and μN (in the example of Fig. 13f it is $T^{(0)}= T^{(0)}_{min}=4N$).

In order to determine only one trend for T(t) even in case of a fully stick contact, the following predictor/corrector procedure is used: first T(t) is computed as

$$T = k_t \cdot u(t) = k_t \cdot u^{(0)} + k_t \cdot \Re\left(u^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right), \tag{26}$$

where $T^{(0)}$ is directly determined by the relative tangential displacement $u^{(0)}$. If T crosses the upper limit μN, then T is corrected as

$$T(t) = T_{\max}^{(0)} + k_t \cdot \Re\left(u^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right), \tag{27}$$

while if T crosses the lower limit -μN, T is corrected as

$$T(t) = T_{\min}^{(0)} + k_t \cdot \Re\left(u^{(n)} \cdot e^{i \cdot n \cdot \omega \cdot t}\right), \tag{28}$$

otherwise T is not modified.

A consequence of the refined method is that the static displacements becomes part of the set of unknowns which must be calculated within the algorithm (Figure 15 b) and the static equilibrium using the 0-th term of the Fourier series of the displacements and contact forces must be considered as well as the other n harmonics.

The same dynamic problem of Fig. 12, solved in the former section by means of the DTI, is now solved in the frequency domain using the HBM (see Section 2) according to two strategies:

1. Forced response calculation in the frequency domain (HBM), single harmonic (n=1), uncoupled approach (Algorithm of Fig. 15a).
2. Forced response calculation in the frequency domain (HBM), single harmonic (n=1), static/dynamic coupling approach (Algorithm of Fig. 15b).

Equation (8) is solved iteratively with the Newton-Raphson non-linear solver. In this case the sub-set of non-linear degrees of freedom is:

$$Q_{NL}^{(n)} = \left[q_{x1,l}^{(n)} \quad q_{y1,l}^{(n)} \quad q_{x1,r}^{(n)} \quad q_{y1,r}^{(n)} \quad q_{x,d}^{(n)} \quad q_{y,d}^{(n)} \quad \vartheta_d^{(n)} \right] \tag{29}$$

Similarly to Equation (22) the harmonic relative displacements $u_L^{(n)}$, $u_R^{(n)}$, $v_R^{(n)}$ and $v_L^{(n)}$ at the contacts are computed by means of equation (22), where the time-dependent quantities are substituted by the n harmonics approximating the real motion. In this case only the first term of the Fourier series is considered. If the uncoupled approach is adopted different strategies can be used to choose the normal pre-load N_0. In (Petrov & Ewins, 2007), for instance, the static equilibrium of the damper is solved assuming the same force distribution shown in Figure 14, i.e., two contacts in slip state. In this case, in fact, the four unknown contact forces ($N_{0,l}$, $N_{0,r}$, $T_{0,l}$, $T_{0,r}$) are reduced to only two since the friction forces are equal to the Coulomb's lower limit. Two equilibrium equations along the horizontal and vertical translation are sufficient to calculate the static normal loads easily. Another hypothesis is assumed in (Panning et al., 2004) where the contact is assumed frictionless, i.e., $T_{0,l}=T_{0,r}=0$. In other cases a quasi-static analysis is performed within in-house software (Cigeroglu et al., 2009) in order to find the normal load distribution when the contacts are more than one per side. This approach is closer to the industrial practice of performing the same simulation with FE commercial software (Szwedowicz et al., 2008). In this case, the static normal loads are introduced by considering the force distribution of Figure 14 that is also the final result of the DTI at the end of the first load step (N_0=25N). The variable normal load N(t) is then written as in Equation (17) according to the uncoupled approach.

Unfortunately, different hypotheses about the normal pre-load lead to different amount of damping generated by the contacts and to different forced responses. To this end, the

static/dynamic coupling of the non-linear equilibrium equations is used in order to obtain unique solution in the frequency domain. The two calculations are repeated for a range of frequencies including the first natural frequency of the blades. Fig. 16a shows the forced response in terms of vibration amplitude of the tip of the left blade (mass m_2). The dashed line is the free response of the blade without the underplatform damper (linear calculation), square marks are the responses of the blade computed with the DTI for different excitation frequencies (non-linear calculation).

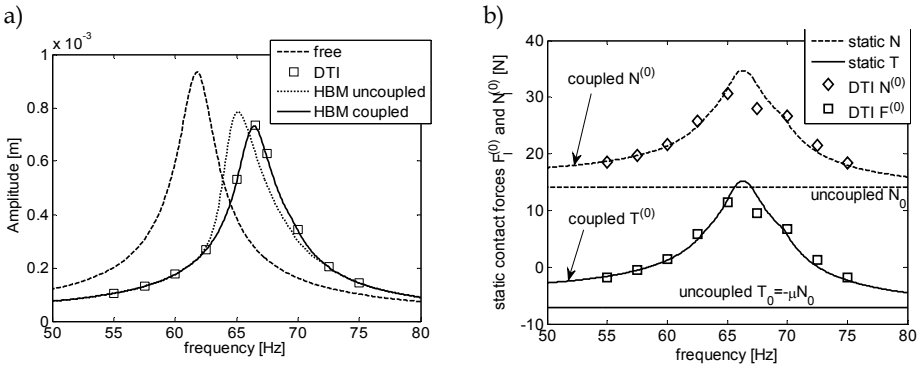


Fig. 16. (a) Comparison of the forced responses of the left blade calculated by means of different strategies (DTI vs. HBM) and (b) corresponding static contact forces on the damper

It is observed that the peak response shifts towards a higher frequency value since the contact with the damper is an additional constraint which stiffens the blade. At the same time, the peak amplitude decreases due to the damping introduced by friction. The dotted line is the non-linear response calculated by means of the uncoupled algorithm of Fig. 15a. Both the peak amplitude and the corresponding frequency are not correctly predicted. Finally, the solid line is the non-linear forced response of the left blade using the static/dynamic coupled approach (Figure 15 b). It is possible to see that the coupled solution perfectly matches the DTI of the equilibrium equations (20). The static normal loads of the left blade for the three different cases are reported in Fig. 16b. The markers are the static contact forces calculated by means of the DTI (diamond markers correspond to the static normal load while square markers are the static tangential force). The dashed and the solid lines refer respectively to the static normal and tangential loads when the HBM is used. The two horizontal trends are the static normal (N_0) and tangential (T_0) values that are chosen a priori within the uncoupled approach. Of course they are two constant values with respect to the excitation frequency since they are assumed regardless of the system dynamics. On the contrary, a good correspondence is found between the DTI method and the prediction of the static contact forces $N^{(0)}$ and $T^{(0)}$ when the static/dynamic coupled approach is used.

The linearization based on the first term of the Fourier series is not always sufficient to represent adequately the non-linear forced response of structures with friction damping. In particular, when the nonlinearities become larger and larger, more than one harmonic may be suitable to model the system dynamics. In the following example the stiffness k_x which links the mass m_1 to the mass m_2 is changed to 5.84e4 and the excitation forces on the blades have been increased to $F_{x,l}=F_{x,r}=30$N in order to generate more slipping at the contacts. The forced response of the modified blades is then calculated by means of the DTI and the static/dynamic approach using one and three harmonics ($N_H=1, 3$).

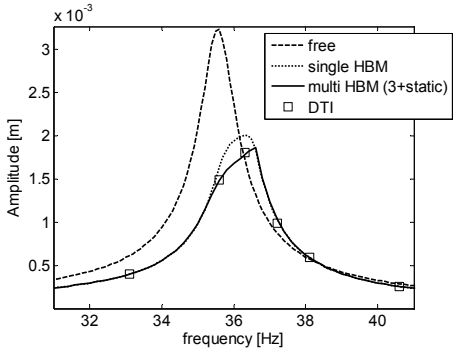


Fig. 17. Static/dynamic coupled approach, comparison of single and multi-HBM with DTI

The three results are shown in Fig. 17 for the sake of comparison, together with the free response of the blade. It is possible to see in this case that if only one harmonic of the Fourier series is retained the blade response is overestimated while three harmonics are sufficient to generate a more correct response.
Anyway, it was proved that in practical applications (Petrov & Ewins, 2005; Petrov, 2007) the additional contributions given by harmonics higher than the first one can be mostly neglected since their amplitude can be two or more orders of magnitudes below the amplitude of the fundamental harmonic. However, they may be important if the excitation is the sum of different loads vibrating at multiple values of frequencies.

5. Full scale applications

5.1 Reducing the size of the non-linear system

The solution of the balance equations (8) in the frequency domain can be a formidable task even with the newest computers in case of industrial applications, since the finite element models can be made of thousands, even millions of dofs. In order to reduce the long computation times, reduction techniques are applied to equations (8) and then the non-linear reduced model is solved with iterative methods, typical of non-linear algebraic equations. In the field of turbomachinery, in particular in the field of bladed disk dynamics, the first reduction technique consists in the application of the boundary conditions of cyclic symmetry (Petrov, 2004; Siewert et al., 2010).

5.1.1 Cyclic symmetry

When a rotating bladed disk is made of N_S identical sectors, it is not necessary to model the whole bladed disk to perform its dynamic analysis. It is possible to model only one of the sectors of the bladed disk, hence called fundamental sector, and then to apply the cyclic symmetry boundary conditions to the boundary nodes of the sector, where it would be connected with the adjacent sectors.
The dofs of the fundamental sectors can be divided into inner dofs Q_I, left dofs Q_L and right dofs Q_R. The first set includes all the dofs lying inside of the fundamental sector, while the second and the third includes dofs lying at the sector left and right boundary respectively, where the sector would be connected to adjacent sectors.

Then, since in cyclic symmetric structures under rotating excitation, like bladed disks, all the sectors vibrate with the same amplitude but with a phase delay depending on the angular speed of the rotor and on the harmonic order of the excitation forces, the following relationship holds for the arbitrary n^{th} harmonic component

$$Q_L^{(n)} = Q_R^{(n)} \cdot e^{in\phi} \tag{31}$$

where n is the generic harmonic index of the periodic rotating excitation and of the response and $\phi=2\pi/N_S$ is the physical angle between two consecutive sectors of the bladed disk.
The use of the cyclic symmetry, when possible, reduces the size of the non-linear system of N_S times, with respect to the full system.
After the application of the cyclic symmetry, the balance equations of the fundamental sector are

$$D^{(n)} \cdot Q^{(n)} = F_E^{(n)} + F_{NL}^{(n)} \text{ with n = 0..N}_H \tag{32}$$

where $D^{(n)} = -(n\omega)^2 M^{(n)} + in\omega C^{(n)} + K^{(n)}$ is the n-th dynamic stiffness matrix of the system and $M^{(n)}$, $C^{(n)}$ and $K^{(n)}$ the mass, damping and stiffness matrices of the fundamental sector obtained applying the nth boundary conditions of equation (31).

5.1.2 Modal superposition

Even after the application of the cyclic symmetry boundary conditions of equation (31), the size of the nonlinear system can be too large for a solution in a reasonable amount of time.
One of the possible strategies to further reduce the number of the equations, is to write the balance equations (6) in modal coordinates (Panning et al. 2003; Cigeroglu et al., 2009). The first step is a modal analysis of the linear system without contact elements, to generate the eigenvalues and the eigenvectors of the system. It is a rather simple step which can be easily performed with any commercial finite element solver available.
As a result of this step, the harmonic components $Q^{(n)}$ of the physical displacements are defined as

$$Q^{(n)} = \Psi^{(n)} \cdot \eta^{(n)} \tag{33}$$

where $\eta^{(n)}$ are the modal amplitudes and $\Psi^{(n)}$ are the mass-normalized mode shapes obtained solving the following eigen-problem

$$\det(K^{(n)} - \lambda M^{(n)}) = 0 \tag{34}$$

Being $K^{(n)}$ and $M^{(n)}$ the stiffness and the mass matrices of the fundamental sector with the cyclic symmetry boundary conditions of equation (31).
The physical balance equations (6) are then replaced by the modal balance equations

$$\eta^{(n)} = \eta_E^{(n)} + \alpha^{(n)} \cdot \varphi_{NL}^{(n)} \tag{35}$$

where $\eta_E^{(n)}$ is the linear response due to the external forces, $\alpha^{(n)}$ the modal receptance matrix and $\varphi_{NL}^{(n)}$ the modal non-linear forces due to friction contacts.
In order to include in the solution process the contact models described in Section 3, based on the procedure depicted in Fig. 2, transitions from the modal to the physical domain and

vice versa are necessary. Physical displacements of contact nodes $Q_{NL}^{(n)}$ are computed by equation (33), while physical contact forces $F_{NL}^{(n)}$ generated at the contact nodes are transformed in the modal forces $\varphi_{NL}^{(n)}$ of equation (35) by

$$\varphi_{NL}^{(n)} = \Psi^{(n)H} \cdot F_{NL}^{(n)} \tag{36}$$

where the superscript H denotes the Hermitian, or complex conjugate, operator.
The main advantage of modal superposition is that the size of the reduced model of equation (35) does not depend on the number of contact points but on the number of mode shapes retained to describe the dynamics of the system. On the other hand, if friction contacts modify strongly the dynamics of the bladed disk, the number of mode shapes necessary to represent accurately its dynamics could be relatively large. It is the case, for instance, of shrouded blades, where the friction contacts are located at the blade tip. In this case, the dynamics of the shrouded blade is completely different from the dynamics of the cantilevered free blade, used to extract the modal parameters in equation (34).

5.1.3 Component mode synthesis

When modal superposition is an unfeasible approach due to the large number of mode shapes necessary to represent accurately the dynamics of the system, a model reduction based on component mode synthesis (CMS) (Craig & Bampton, 1968) is possible.
The dofs $Q^{(n)}$ of the system can be partitioned into master dofs $Q_M^{(n)}$, and slave dofs $Q_S^{(n)}$. As a result, the homogeneous undamped linear system associated to the n-th harmonic order becomes

$$\left(\begin{bmatrix} K_{MM}^{(n)} & K_{MS}^{(n)} \\ K_{SM}^{(n)} & K_{SS}^{(n)} \end{bmatrix} - \omega^2 \begin{bmatrix} M_{MM}^{(n)} & M_{MS}^{(n)} \\ M_{SM}^{(n)} & M_{SS}^{(n)} \end{bmatrix} \cdot \right) \begin{Bmatrix} Q_M^{(n)} \\ Q_S^{(n)} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{37}$$

The physical vector of dofs $Q^{(n)}$ is then approximated by a linear superposition of static modes and of slave mass-normalized mode shapes.
The first step of the procedure is based on the static reduction of the homogenous system

$$\begin{bmatrix} K_{MM}^{(n)} & K_{MS}^{(n)} \\ K_{SM}^{(n)} & K_{SS}^{(n)} \end{bmatrix} \begin{Bmatrix} Q_M^{(n)} \\ Q_S^{(n)} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{38}$$

which leads to the following expression

$$\begin{Bmatrix} Q_M^{(n)} \\ Q_S^{(n)} \end{Bmatrix} = \begin{Bmatrix} I \\ K_{MM}^{(n)} - K_{MS}^{(n)} \cdot K_{SS}^{(n)-1} \cdot K_{SM}^{(n)} \end{Bmatrix} Q_M^{(n)} \tag{39}$$

where I is the identity matrix.
The second step consists in the calculation of the slave eigenvectors $\Psi_S^{(n)}$ computed solving the eigenproblem

$$\det(K_{SS}^{(n)} - \lambda M_{SS}^{(n)}) = 0 \tag{40}$$

and giving the relationship

$$Q_S^{(n)} = \Psi_S^{(n)} \cdot \eta_S^{(n)} \,. \tag{41}$$

By linear combination of equations (39) and (41), the physical dofs are expressed as

$$Q^{(n)} = \begin{Bmatrix} Q_M^{(n)} \\ Q_S^{(n)} \end{Bmatrix} = \begin{bmatrix} I & 0 \\ K_{MM}^{(n)} - K_{MS}^{(n)} \cdot K_{SS}^{(n)-1} \cdot K_{SM}^{(n)} & \Psi_S^{(n)} \end{bmatrix} \cdot \begin{Bmatrix} Q_M^{(n)} \\ \eta_S^{(n)} \end{Bmatrix} = \Phi_{CMS}^{(n)} \cdot \begin{Bmatrix} Q_M^{(n)} \\ \eta_S^{(n)} \end{Bmatrix} = \Phi_{CMS}^{(n)} \cdot q^{(n)} \tag{42}$$

The resulting reduced model has the following balance equations

$$d^{(n)} \cdot q^{(n)} = f_E^{(n)} + f_{NL}^{(n)} \text{ with n = 0..N}_\text{H} \tag{43}$$

where $d^{(n)} = -(n\omega)^2 m^{(n)} + in\omega c^{(n)} + k^{(n)}$ is the n-th dynamic stiffness matrix of the system and $m^{(n)}$, $c^{(n)}$ and $k^{(n)}$ are the reduced mass, damping and stiffness matrices defined as

$$k^{(n)} = {\Phi_{CMS}^{(n)}}^H \cdot K^{(n)} \cdot \Phi_{CMS}^{(n)}; \quad c^{(n)} = {\Phi_{CMS}^{(n)}}^H \cdot C^{(n)} \cdot \Phi_{CMS}^{(n)}; \quad m^{(n)} = {\Phi_{CMS}^{(n)}}^H \cdot M^{(n)} \cdot \Phi_{CMS}^{(n)}; \tag{44}$$

while the reduced forces are defined as $f_E^{(n)} = {\Phi_{CMS}^{(n)}}^H \cdot F_E^{(n)}; \quad f_{NL}^{(n)} = {\Phi_{CMS}^{(n)}}^H \cdot F_{NL}^{(n)}$

The correct strategy to implement the CMS technique to non-linear systems with friction contacts is to include the non-linear contact dofs $Q_{NL}^{(n)}$ in the set of master nodes. In this way, the physical displacements necessary to compute the contact forces $F_{NL}^{(n)}$ are explicitly included in the set of the unknowns of the reduced model and transition from the generalized to the physical coordinates and vice versa is not necessary during the calculation. Furthermore, as for the case of modal superposition, the CMS is a rather simple step which can be easily performed with any commercial finite element solver.

5.2 Modeling microslip

Contact elements described in Section 3 and used to model periodical contact forces at friction contacts assume an ideal gross-slip behavior of the contact. Once the modulus of tangential force T equals the Coulomb limit value μN, the contact enters the slip state.

Real contacts in industrial application are mostly conforming contact and the contact extends on a finite contact area larger than a single point. In these conditions, the transition between full sticking and slipping conditions is not so abrupt as modeled in Section 3. First of all, the outer parts of the contact surfaces enter the slipping conditions while the inner parts are still sticking. Then as the tangential contact load grows larger, the slipping area extends progressively to the whole contact area and gross slip occurs. The transition state between the sticking and the gross slipping conditions is called microslip (Filippi et al., 2004; Cigeroglu et al., 2007; Allara, 2009). It is an important contact state, since most of the friction contacts in industrial application works in microslip conditions.

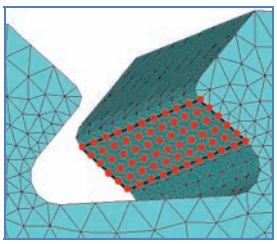


Fig. 18. Contact points of a blade root joint on a bladed disk

In order to model the microslip behavior of friction contacts in numerical simulations of frictionally damped structures, the contact area is divided into a grid made of several contact points, as shown in Fig. 18, where red points are used to identify the contact points on a bladed disk at the blade root joint.
Then the contact parameters (tangential and normal contact stiffness) are evaluated for the whole contact (Allara, 2009) and their value is evenly distributed among the contact nodes.

5.3 Case studies

The static/dynamic coupled approach described in Section 4.2, is now applied to two case studies, representing two of the most typical configurations of friction contacts in bladed disks: blade root joints and underplatform dampers.

5.3.1 Blade root joint

This application is very common in turbine bladed disks because the blades are connected to the disk rim by means of firtree or dovetail roots, inserted in the corresponding disk slots. In this case, the friction damping is generated at the contact between the blade root and the disk slot during the vibration of the assembly. The material of the mock bladed disk (Fig. 19) used for the analysis is steel and the number of blades is 12. The disk and the blades are modeled with the finite element method. Cyclic symmetry boundary conditions are applied to the disk, in order to model only its fundamental sector, while the blade is modeled with free-free boundary conditions.
Both the blade and the disk sector models are then reduced by means of the modal superposition principle. Each pair of contact nodes lying on the blade/disk interface are connected by means of two orthogonal 1D contact elements, taking into account variable normal load (See section 4.2).

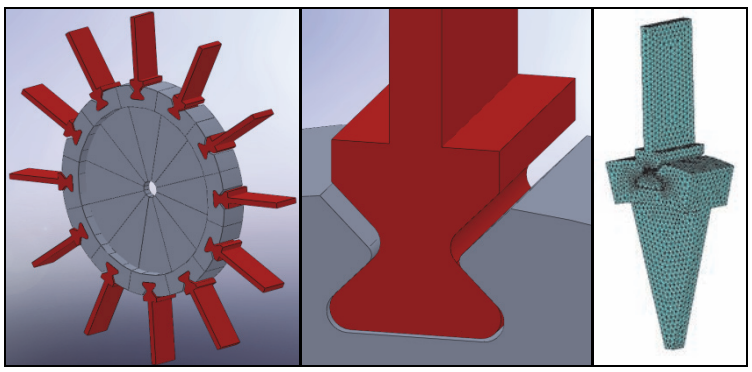


Fig. 19. Mock bladed disk: full view, detail of the root joint, finite element model.

The forced response of the system around the 1st bending mode of the blades with 6 nodal diameters is studied, corresponding to the out-of-phase vibration of two consecutive blades, using the 0th and the fundamental order of the Fourier terms.
The contact between the blade and the disk rim occurs when the bladed disks rotates and the centrifugal force presses the blade root onto the disk slots, the value of the design interference Δ is set equal to zero. The friction coefficient μ is set equal to 0.5 and the values

of normal and tangential contact stiffness are computed assuming a flat punch contact (Allara, 2009) and are evenly distributed among the contact nodes.

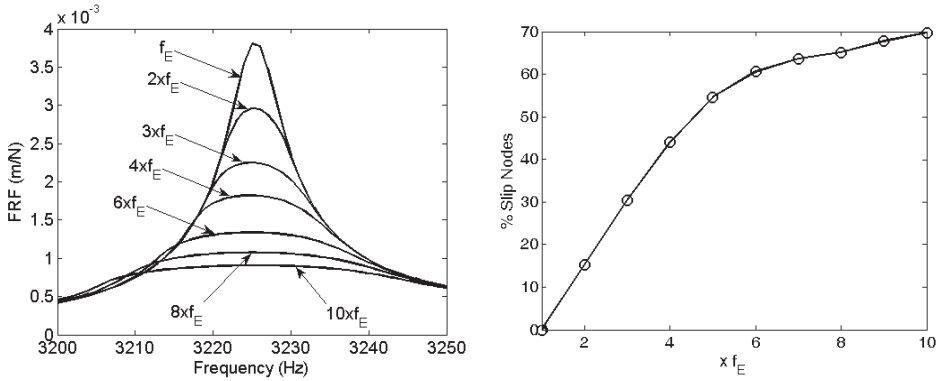


Fig. 20. Blade root joint: Forced response of the blade and slip growth.

The frequency response function (FRF) of the system, computed for different amplitudes of the engine order excitation, is shown in Fig. 20. The non-linear behavior of the blade root joint is evident, as the FRF of the system decreases as the amplitude of the excitation grows larger. At the lowest amplitude, the contact surfaces are fully stuck, while as the excitation force increases, slip begins at the boundary of the contact area and extends inward, as shown in Fig. 20, where the growth of the slipping area in percentage with respect to the total contact area is plotted.

5.3.2 Underplatform damper

The second application is an integral bladed disk (a.k.a. blisk) with underplatform dampers (Fig. 21). The dampers are wedge shaped devices located under the blade platforms and held in contact with them by the centrifugal force acting on the dampers themselves during rotation. Friction damping is generated at the contact surface between the left (L) and the right (R) damper surfaces and the walls of the blade cavities. Also in this case, the value of the design interference Δ is set equal to zero, since contact occurs only when the centrifugal force acts on the damper.

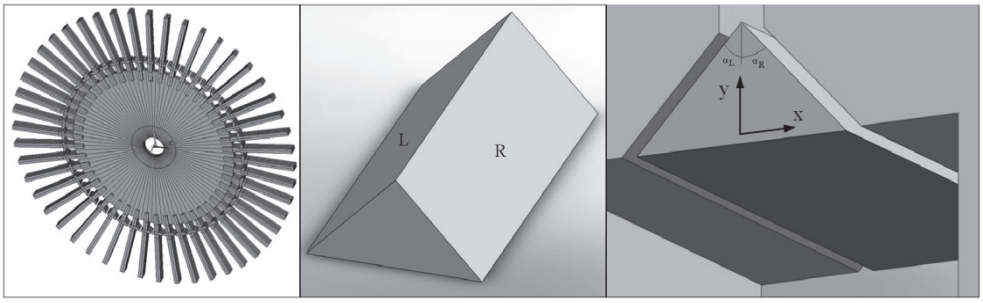


Fig. 21. Blisk geometry, damper geometry and damper location under the blade platforms.

Underplatform dampers are included in the design of turbine bladed disks as a system of passive control of vibration, when the detuning of the system is not possible, due to the high modal density of the bladed disk and/or to the wide spectrum of the excitation forces.
The material of the blisk and of the dampers is steel, the number of blades is 48, the damper is symmetrical with $\alpha_L=\alpha_R=45°$, the friction coefficient is set to $\mu=0.5$; the contact parameters k_t and k_n are computed assuming a flat punch contact (Allara, 2009) and then uniformly distributed over the contact nodes.
The blisk and the dampers are modeled with the finite element method. Cyclic symmetry boundary conditions are applied to the blisk, in order to model only one fundamental sector, while the damper is modeled with free-free boundary conditions. The modal superposition principle is used to reduce the size of the models and the 0^{th} and 1^{st} order harmonics are used in the analysis.
The effect of the main damper design parameter, the damper mass, on the dynamics of the assembly is investigated around the 1^{st} bending mode shape of the blades, in case of 12^{th} engine order excitation.
The FRF curves of the system are plotted in Fig. 22, where the effect of the damper mass on the system dynamics, both in terms of vibration amplitude and resonance frequency can be observed. An optimum damper mass exists and corresponds to the minimum vibration amplitude, while as the damper mass grows larger than the optimum value, the system response increases again and stabilizes when full sticking conditions are established. A full resume of the effect of the damper mass on the system response is shown in terms of the damper optimization curve, in Fig. 22.
As described in Section 5.2, in order to model microslip, the contact area is meshed with a regular grid of contact nodes. In Fig. 23, the evolution of the contact status versus the value of the damper mass is shown.

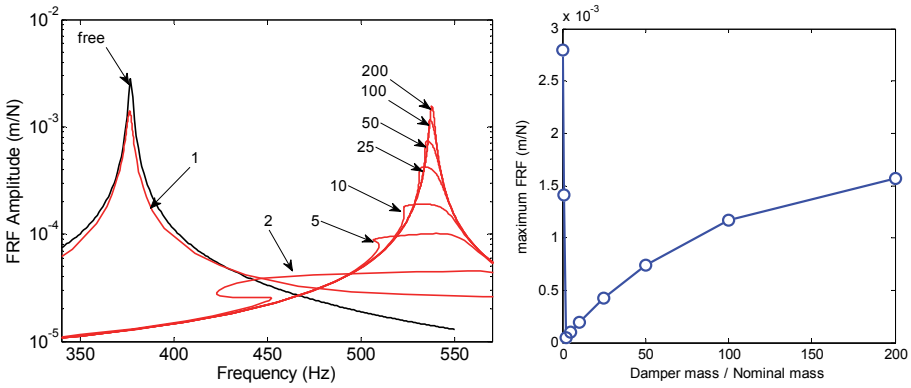


Fig. 22. Effect of the damper mass (m_d = n x m_{nom} with n = [1:200].): FRF curves (left) and optimization plot (right).

The FRF curves refer to a damper mass 25, 50, 100 and 200 times the nominal mass (m_{nom}). The status of the contact area is shown by means of the following colors (white: sticking; grey: alternate stick-slip; black: partial lift-off). The plots show that at 200x the contact is almost fully stuck, while the microslip increases as the damper mass becomes lower. At 25x no sticking areas are found and only alternating stick-slip and partial lift-off occur. The plots

also show that the upper part of the contact is subjected to partial lift-off in the whole range of explored damper masses, due to the blade platform kinematics, characterized by larger displacements at the higher radius.

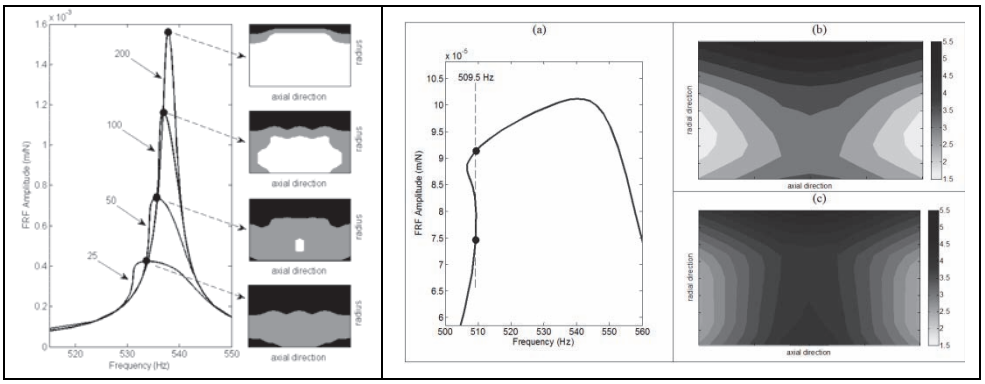


Fig. 23. Effect of the damper mass on the contact status (left). Different distributions of static normal load in case of coupled static/dynamic calculation (right).

On the right-hand side of Fig. 23, the FRF curve corresponding to 5x damper mass is shown. At 509.5 Hz multiple solution exist due to the softening effect produced by the partial lift-off of the contact surfaces. Two solutions are highlighted by the black points over the FRF curve. The distributions of static normal loads predicted at the upper point (b) and at the lower point (c) are also plotted. They differ to each other, despite being computed at the same frequency, i.e. at the same rotational speed and at the same centrifugal force. This result is possible only if the static and the dynamic contact forces acting on the damper are computed simultaneously as proposed with the coupled approach described in Section 4.2. In case of preliminary static analysis, the same distributions of static normal loads would be predicted.

6. Conclusions

One of the emerging directions in the design of complex structures is the modeling of joints. Today, friction contacts are one of the key phenomena to understand in order to predict correctly the dynamics of assemblies and the corresponding stress distribution. In the last two decades companies were more and more interested in including the nonlinear action of sliding surfaces within the design procedure by means of time efficient numerical tools.
In this paper the consolidated procedure of solving the nonlinear equilibrium equations of the system by means of the Harmonic Balance Method (HBM) is therefore described. HBM has the advantage of reducing the calculation time in spite of an acceptable approximation of the forced response with respect to the integration of the differential equations in the time domain (DTI).
A review of the classical modeling of friction contacts is presented and the main parameters which i) correctly describe the slip/stick/separation phenomenon and ii) mainly affect the response of the bodies in contact are defined and commented. Particular attention is paid to show why one contact model shall be preferred in spite of others in practical applications where the fast computation of the forced response is an asset.

The strong influence of the dynamic response of the structure on the determination of the static loads acting at the contact (and vice versa) is proved through a simple lumped parameter model whose equilibrium is solved by DTI. The classical contact models are not sufficient to simulate this dependence since the normal load is a parameter chosen a priori which does not depend on the dynamics of the structure.
In order to take into account the mutual influence, the authors proposed an improved contact model which couples the static and the dynamic equilibrium equations of the system.
Simulations obtained with the classical and refined contact models both implemented in a HBM algorithm are compared to the DTI response and it is demonstrated that the coupling of the static and dynamic equilibrium equations of the structure is of primary importance to predict the maximum oscillation amplitude of the system (therefore the maximum stress concentration) and the stiffening of the structure due to the contacts.
Finally, two practical examples of the application of the refined contact model to turbomachinery components are presented where friction dampers are optimized for two bladed disks in order to reduce the blade vibrations.

7. Appendix - numerical methods and solver optimization

In order to solve the non-linear balance equations of vibrating structures with friction contact, iterative numerical method are necessary. In the technical literature, the most exploited method is the continuation method in two of its possible versions: natural continuation and arc-length continuation.
In this appendix, we consider the non-linear balance equations of the structure

$$Q_{NL}^{(n)} = Q_{E,NL}^{(n)} + R_{NL,NL}^{(n)} \cdot F_{NL}^{(n)} \tag{45}$$

as described in Section 2 of this chapter, but the contents of this appendix are still valid also in case of reduced order models. In order to solve the algebraic system (45), the complex equations are first of all turned into a set of real equations with real unknowns. So the following real quantities are defined

$$X = \left\{ Q_{NL}^{(0)};\ \ \Re\left(Q_{NL}^{(1)}\right);\ \ \Im\left(Q_{NL}^{(1)}\right);\ \ \cdots\ \ \Re\left(Q_{NL}^{(N_H)}\right);\ \ \Im\left(Q_{NL}^{(N_H)}\right) \right\} \tag{46}$$

$$X_E = \left\{ Q_{E,NL}^{(0)};\ \ \Re\left(Q_{E,NL}^{(1)}\right);\ \ \Im\left(Q_{E,NL}^{(1)}\right);\ \ \cdots\ \ \Re\left(Q_{E,NL}^{(N_H)}\right);\ \ \Im\left(Q_{E,NL}^{(N_H)}\right) \right\} \tag{47}$$

$$F_X = \left\{ F_{NL}^{(0)};\ \ \Re\left(F_{NL}^{(1)}\right);\ \ \Im\left(F_{NL}^{(1)}\right);\ \ \cdots\ \ \Re\left(F_{NL}^{(N_H)}\right);\ \ \Im\left(F_{NL}^{(N_H)}\right) \right\} \tag{48}$$

$$R_X = \begin{bmatrix} R_{NL,NL}^{(0)} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \Re\left(R_{NL,NL}^{(1)}\right) & -\Im\left(R_{NL,NL}^{(1)}\right) & \cdots & 0 & 0 \\ 0 & \Im\left(R_{NL,NL}^{(1)}\right) & \Re\left(R_{NL,NL}^{(1)}\right) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \Re\left(R_{NL,NL}^{(N_H)}\right) & -\Im\left(R_{NL,NL}^{(N_H)}\right) \\ 0 & 0 & 0 & \cdots & \Im\left(R_{NL,NL}^{(N_H)}\right) & \Re\left(R_{NL,NL}^{(N_H)}\right) \end{bmatrix} \tag{49}$$

and the balance equations to solve become

$$X = X_E + R_X \cdot F_X(X) \tag{50}$$

7.1 Natural continuation

In this case, we want to compute the response of the system in a given range of frequencies $[\omega_i, \omega_f]$ of the fundamental harmonics of the external excitation with a frequency resolution $\Delta\omega$.
The balance equation is written in the form f(X)=0 as

$$f(X) = X - X_E - R_X \cdot F_X(X) = 0 \tag{51}$$

The analysis starts at $\omega = \omega_i$ with a guess value X_0, for instance the value of the response of the linear system without fiction contacts, and a Newton-Raphson procedure is then initialized. At the generic step, the $(k+1)^{th}$ value is computed as

$$X_{k+1} = X_k - J(X_k)^{-1} \cdot f(X_k) \tag{52}$$

with the value of the functions $f(X_k)$ defined as

$$f(X_k) = X_k - X_E - R_X \cdot F_X(X_k) = 0 \tag{53}$$

and the jacobian matrix J of the system defined as

$$J(X_k) = df(X)/dX\big|_{X=X_K} = I - R_X \cdot dF_X(X)/dX\big|_{X=X_K} . \tag{54}$$

When the norm of the residuals of f(X) is lower than a prescribed tolerance, the algorithm stops, the solution is saved and the frequency is increased to $\omega = \omega_i + \Delta\omega$. According to natural continuation, the guessed value at the generic frequency ω is the solution computed at $\omega - \Delta\omega$. In this way, the range of frequencies of interest is spanned with an ascending (or descending) order.
In order to improve the convergence of the algorithm and to shorten the calculation time, a feasible and effective option (Petrov & Ewins, 2003; Borrajo et al., 2006; Siewert et al, 2010) is the analytical calculation of the jacobian matrix J defined in equation (54). The key point is the analytical calculation of the derivatives of the Fourier coefficients of the non-linear contact forces dF_X/dX.

7.2 Arc-length continuation

When natural continuation is implemented, the range of frequencies of interest is spanned either with an ascending or with a descending order. In case of partial lift off of the contact points, the response curve may exhibit the so-called jump phenomenon, shown in Fig. 24. For a given frequency multiple solutions exist and turning points appear on the FRF curve where the resonance peaks shows either a hardening or a softening behavior. In these cases, the natural continuation approach is not able to follow the FRF curve and to compute the whole curve. A strategy based on two calculations, one with an ascending frequency order and the other with a descending frequency order, may partly help and compute both the lower and the upper branch of the solution but not the intermediate branch, shown in Fig. 24.

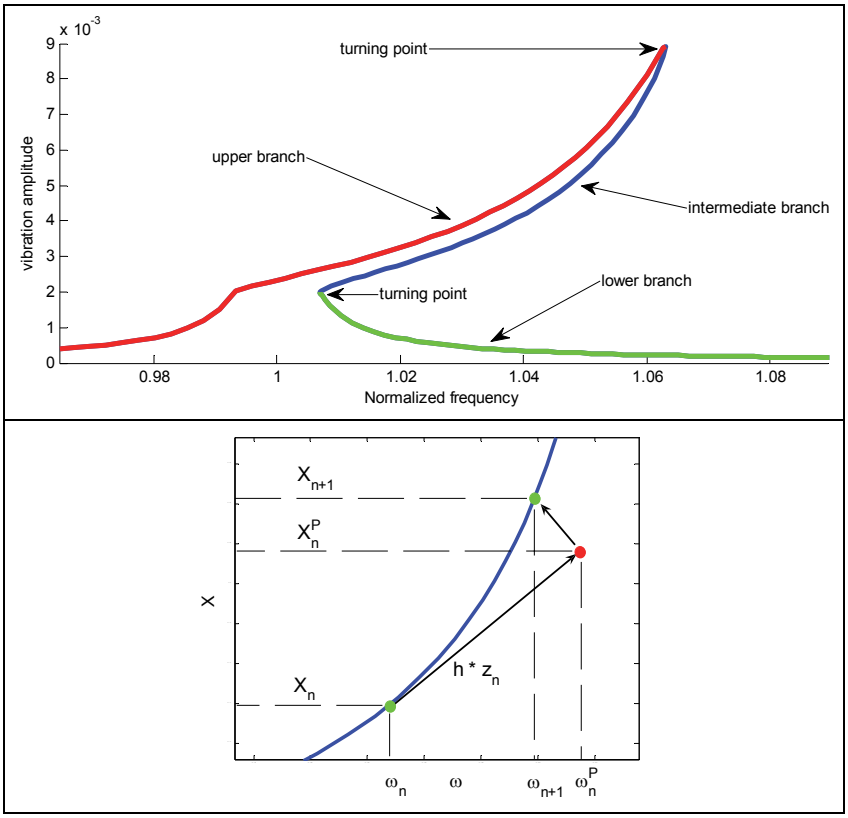


Fig. 24. Forced response curve with multiple solutions (top) and example of arc-length continuation (bottom).

In order to overcome this limitation of the natural continuation approach, an arc-length continuation method can be used (Chan & Keller, 1982), based on a predictor-corrector strategy. The first calculation at $\omega=\omega_i$ is performed with the classical Newton-Raphson method described in equation (52). Then (Fig. 24), the following arc-length continuation strategy can be used:

1. For a given n^{th} solution X_n at frequency ω_n, a predictor step is performed as

$$Y_n^P = \begin{Bmatrix} X_n^P \\ \omega_n^P \end{Bmatrix} = \begin{Bmatrix} X_n \\ \omega_n \end{Bmatrix} + h \cdot Z_n = Y_n + h \cdot Z_n \tag{55}$$

where Z_n is the unit vector tangent to the solution curve at the n^{th} solution and h a scalar controlling the length of the predictor step.

2. A corrector step is implemented searching for the solution in a direction orthogonal to the predictor step. The augmented system of equations solved in the corrector step consists of the following equations

$$\begin{cases} X - X_E - R_X \cdot F_X(X) = 0 \\ Z_n{}^T \cdot \left(Y_n^P - Y \right) = 0 \end{cases} \tag{56}$$

including the balance equations (51) and the equation used to define the corrector step direction. As a consequence, also the vectors of the unknowns is augmented including also the excitation frequency ω. Equation (56) is solved with the Newton-Raphson method, starting with a guessed value equal to Y_n^P .

8. References

Allara M. (2009). A model for the characterization of friction contacts in turbine blades, *Journal of Sound & Vibration*, Vol.320, No.3, (February 2009), pp. 527-544, ISSN 0022-460X

Allara, M; Zucca, S.; Gola, M.M (2007). Effect of crowning of dovetail joints on turbine blade root damping, *Key Engineering Materials*, Vol. 347, (September 2007), pp. 317-322.

Awrejcewicz J.; Supeł B.; Lamarque C.-H.; Kudra G.; Wasilewski G. & Olejnik P. (2008). Numerical and experimental study of regular and chaotic motion of triple physical pendulum, *International Journal of Bifurcation and Chaos*, Vol.18, No.10, (October 2008), pp.2883-2915, ISSN 0218-1274

Awrejcewicz J.; Grzelczyk D. & Pyryev Yu. (2008b). A novel dry friction modeling and its impact on differential equation and Lyapunov exponents estimation, *Journal of Vibroengineering*, Vol.10, No.4, (December 2008), pp. 475-482, ISSN 1392-8716

Awrejcewicz J. & Pyryev Yu. (2009). *Nonsmooth Dynamics of Contacting Thermoelastic Bodies*, Springer-Verlag, ISBN 978-0-387-09652-0, New York

Borrajo J.M.; Zucca S. & Gola M.M. (2006). Analytical Formulation of the Jacobian Matrix for Non-linear Calculation of the Forced Response of Turbine Blade Assemblies with Wedge Friction Dampers, *International Journal of Non-linear Mechanics*, Vol.41, No.10, (December 2006), pp. 1118-1127, ISSN 0020-7462

Cardona, A.; Lerusse, A. & Geradin, M. (1998). Fast Fourier Nonlinear Vibration Analysis, *Computational Mechanics*, Vol.22, No.2, (August 1998), pp. 128-142, ISSN 0178-7675

Chan, T.F.C.; Kelle,r H.B. (1982). Arc-Length Continuation and Multi-Grid Techniques for Nonlinear Elliptic Eigenvalue Problems, *SIAM Journal of Scientific and Statistical Computing*, Vol.3, No.2, pp. 173-194, ISSN 0196-5204

Charleux, D.; Gibert, C.; Thouverez, F.; Dupeux, J. (2006). Numerical and experimental study of friction damping blade attachments of rotating bladed disks, *International Journal of Rotating Machinery*, vol. 2006, Article ID 71302.

Cigeroglu, E.; An, N. & Menq, C. H. (2007). A Microslip Friction Model with Normal Load Variation induced by Normal Motion, *Nonlinear Dynamics*, Vol.50, No.3, (November 2007), pp. 609-626, ISSN 0924-090X

Cigeroglu E.; An N. & Menq C.H. (2009). Forced Response Prediction of Constrained and Unconstrained Structures Coupled Through Frictional Contacts, *Journal of*

Engineering for Gas Turbines and Power, Vol.131, No.2, (March 2009), ISSN 0742-4795

Craig R.R. & Bampton M.C.C. (1968). Coupling of Substructures for Dynamic Analyses, *AIAA Journal*, Vol.6, No.7, pp.1313-1319, ISSN 0001-1452

Csaba, G. (1998). Forced Response Analysis in Time and Frequency Domains of a Tuned Bladed Disk with Friction Dampers, *Journal of Sound and Vibration*, Vol.214, No.3, (July 1998), pp. 395-412, ISSN 0022-460X.

Filippi S.; Akay A. & Gola M.M. (2004). Measurement of tangential contact hysteresis during microslip, *ASME Journal of Tribology*, Vol.126, No.3, (July 2004), pp.482–489, ISSN 0742-4787

Firrone C.M.; Botto D. & Gola M.M. (2006) Modelling A Friction Damper: Analysis Of The Experimental Data And Comparison With Numerical Results, (ESDA2006-95605), *Proceedings of ESDA 2006*, ISBN 0-7918-4250-9, Torino, Italy, July, 2006

Firrone C.M.; Zucca S. & Gola M.M. (2011). The effect of underplatform dampers on the forced response of bladed disks by a coupled static/dynamic harmonic balance method, *International Journal of Non-Linear Mechanics*, Vol.46, No.2, (March 2011), pp.363-375, ISSN 0020-7462

Griffin J.H. (1980). Friction damping of resonant stresses in gas turbine engine airfoils, *Journal of Engineering for Power*, Vol.102, No.2, pp.329-333, ISSN 0022-0825

Griffin J.H. & Menq C.H. (1991). Friction Damping of Circular Motion and Its Implications to Vibration Control, *Journal of Vibration and Acoustics*, Vol.113, No.2, (April 1991), pp.225-229, ISSN 1048-9002

Menq C.H. & Yang B.D. (1998). Non-Linear Spring Resistance and Friction Damping of Frictional Constraint having Two-Dimensional Motion, *Journal of Sound and Vibration*, Vol. 217, No.1, (October 1998), pp.127-143, ISSN 0022-460X

Panning L.; Sextro W & Popp K. (2003). Spatial Dynamics of Tuned and Mistuned Bladed Disks with Cylindrical and Wedge-Shaped Friction Dampers, *International Journal of Rotating Machinery*, Vol.9, No.3, (July 2002), pp.219-228, ISSN (printed): 1023-621X. ISSN (electronic): 1542-3034.

Panning L.; Popp K.; Sextro W.; Goetting F.; Kayser A. & Wolter I. (2004). Asymmetrical underplatform dampers in gas turbine bladings: theory and application, (GT2004-53316), *Proceedings of ASME Turbo Expo*, ISBN 0-7918-4171-5, Vienna, Austria, June, 2004

Petrov E.P. & Ewins D.J. (2003). Analytical formulation of friction interface elements for analysis of nonlinear multiharmonic vibrations of bladed discs, *Transactions of ASME Journal of Turbomachinery*, Vol.125, No.2, (April 2003), pp.364-371, ISSN 0889-504X

Petrov E.P. (2004). A method for use of cyclic symmetry properties in analysis of nonlinear multiharmonic vibrations of bladed disks, *Journal of Turbomachinery*, Vol.126, No.1, (January 2004), pp.175-183, ISSN 0889-504X

Petrov E.P. & Ewins D.J. (2006). Effects of damping and varying contact area at blade-disc joints in forced response analysis of bladed disk assemblies, *Journal of Turbomachinery*, Vol.128, No.2, (April 2006), pp. 403-410, ISSN 0889-504X

Petrov E.P. & Ewins D.J. (2007). Advanced Modeling of Underplatform Friction Dampers for Analysis of Bladed Disk Vibration, *Journal of Turbomachinery*, Vol.129, No.1, (January 2007), pp.143-150, ISSN 0889-504X

Petrov E.P. (2007). Explicit Finite Element Models of Friction Dampers in Forced Response Analysis of Bladed Discs, (GT2007-27980), *Proceedings of ASME Turbo Expo 2007*, ISBN 0-7918-4794-2, Montreal, Canada, May, 2007.

Sanliturk K.Y. & Ewins D.J. (1996). Modelling Two-Dimensional Friction Contact and its Application using Harmonic Balance Method, *Journal of Sound and Vibration*, Vol.193, No.2, (June 1996), pp.511-523, ISSN 0022-460X

Sanliturk K.Y.; Ewins D.J. & Stanbridge A.B. (2001). Underplatform Dampers for Turbine Blades: Theoretical Modelling, Analysis and Comparison with Experimental Data, *Journal of Engineering for Gas Turbines and Power*, Vol.123, No.4, (October 2001), pp.919-929, ISSN 0742-4795

Siewert C.; Panning L.; Wallaschek J. & Richter C. (2010). MultiHarmonic Forced Response Analysis of a Turbine Blading Coupled by NonLinear Contact Forces, *Journal of Engineering for Gas Turbines and Power*, Vol.132, No.8, (August 2010), ISSN 0742-4795

Srinivasan A.V., (1997). Flutter and Resonant Vibration Characteristics of Engine Blades, *Journal of Engineering for Gas Turbines and Power*, Vol. 119, No.4 (October 1997), pp. 742-775.

Szwedowicz J. (2008). Bladed disks: non linear dynamics, *Structural design of aircraft engines: key objectives and techniques*, E. Seinturier & G. Paniagua, ISBN 978-2-930389-8-2-6, Belgium

Szwedowicz J.; Gibert C.; Sommer T.P. & Kellerer R. (2008). Numerical and Experimental Damping Assessment of a Thin-Walled Friction Damper in the Rotating Setup with High Pressure Turbine Blades, *Journal of Engineering for Gas, Turbines and Power*, Vol.130, No.1, (January 2008), ISSN 0742-4795

Yang B.D.; Chu M.L. & Menq C.H. (1998). Stick–Slip–Separation Analysis and Non-Linear Stiffness and Damping Characterization of Friction Contacts Having Variable Normal Load, *Journal of Sound and Vibrations*, Vol.210, No.4, (March 1998), pp.461-481, ISSN 0022-460X

Yang B.D. & Menq C.H. (1998a). Characterization of Contact Kinematics and Application to the design of Wedge Dampers in Turbomachinery Blading: part 1 - Stick-slip Contact Kinematics, *ASME Journal of Engineering for Gas Turbine and Power*, Vol.120, No.2, (April 1998), pp.410-417, ISSN 0742-4795

Yang B.D. & Menq C.H. (1998b). Characterization of Contact Kinematics and Application to the design of Wedge Dampers in Turbomachinery Blading: part 2 - Prediction of Forced Response and Experimental Verification, *ASME Journal of Engineering for Gas Turbine and Power*, Vol.120, No.2, (April 1998), pp.418-423, ISSN 0742-4795

Yang B.D. & Menq C.H. (1998c). Characterization of 3D contact kinematics and prediction of resonant response of structures having 3D frictional constraint, *Journal of Sound and Vibration*, Vol.217, No.5, (November 1998), pp.909-925, ISSN 0022-460X

Zucca S.; Botto D. & Gola M.M. (2008). Range of Variability in the Dynamics of Semi-cylindrical friction dampers for turbine blades, (GT2008-51058), *Proceedings of ASME Turbo Expo*, ISBN 978-0-7918-4315-4, Berlin, Germany, June, 2008

15

Modeling and Simulation of Biomechanical Systems - An Orbital Cavity, a Pelvic Bone and Coupled DNA Bases

J. Awrejcewicz[1], J. Mrozowski[1], S. Młynarska[1],
A. Dąbrowska-Wosiak, B. Zagrodny[1],
S. Banasiak and L.V. Yakushevich[2]
[1]*Technical University of Łódź, Department of Automation and Biomechanics, Łódź,*
[2]*Institute of Cell Biophysics of the Russian Academy of Sciences, Pushchino,*
[1]*Poland*
[2]*Russia*

1. Introduction

This chapter is organized in the following manner. Modelling of an orbital cavity using finite element method is presented in section 2. Finite element method (FEM) is one of the basic tools used for mechanical investigations of a skull, a pelvic bone, eye-socket and in reconstruction of a bony face deformed by congenital defects or injuries. For example, it was applied for modelling of a skull with gnathothisis (Boryor et al., 2008), for investigation of infant head injuries caused by impact (Roth et al, 2009) and for examination of facial skull dystosis of a child (Gautam et al, 2007). FEM served also for an radiological and mathematical analysis of facial deformation provoked by curved central axis of the skull (Iannetti et al., 2004). Furthermore, the method was utilized for modelling of orbit deformation being result of bunt injury (Al-Sukhun et al, 2006) and also for investigation of biomechanical properties of the orbit (Sander et al, 2006).
The aim of the study is to develop the numerical model of a bottom arch of an orbital cavity using a FEM. Based on the data obtained from computer tomography, the model of a healthy orbit was proposed. The results obtained from numerical analysis may serve as a basis for further investigations concerning stresses and deformations in orbital implants, including a direct implant application. Due to its geometrical complexity the whole skull (and especially its facial part) modelling presents a substancial engineering problem. To resolve it one subject the area of interest (the surface or the space) to a segmentation by means of finite number of elements averaging the physical state of the body. To generate the maps of stresses, deformations and displacements prior to calculations, it is indispensable to prepare the geometrical and material data of the model.
The main subject of section 3 is to present a model of a double layered pelvic bone and some phenomena during leg flexion, extension, adduction and abduction, using finite element method (FEM). In the musculoskeletal system, the pelvis is one of the most important bones. It is a support of whole body, transfer external and gravitational loads across sarco-illiac and

hip joints (Mrozowski & Awrejcewicz, 2004). According to reference (Wall, 2002), for the population over 65 years, more than 50% of all chronic diseases are joint diseases, most often spine and hip joint. Fracture rate is significantly higher for women over 65 years with ostheoporosis – even 4.6% (Cranney et al, 2007). A properly build model of this bone should bring satisfactory results, minimizing or just eliminating a difficulty of possible harmful medical activities on the examined subject. According to (Mazurkiewicz & Topoliński, 2009), the best method for today is BMD (Bone Mineral Density), painless and noninvasive, however it does not descript the shape of the bones. To make a FEM calculations, a model of a pelvic bone was introduced on the basis of high quality computer tomography (CT) images of an anonymous person pelvic bone. The authors propose some simplification of the model. The aim of the paper was to present the algorithm. To speed up the calculations, it was decided to use simple materials properties. Bone was modeled as homogenous, isotropic, elastic material. However, it is well known that bones have non-homogenous, anisotropic properties. To make model more accurate, this should be taken into account. Model of pelvis took into account a dual structure of the bone, consisting of cortical (compact) – outside part and trabecular (spongy) – inside part bone tissues. Stress in a bone comes from the body weight and from the attached muscles generating loads during movements of lower limb. Numerical computation results are presented in form of the stress and strain maps.
Section 4 deals with oscillations of coupled DNA bases. It is well known now, that the structure of the DNA molecule is not static but dynamic and that *in vivo* the double helix contains small amounts of open states in which bases are unpaired. We study nonlinear oscillations of the DNA base pairs (adenine – thymine and guanine-cytosine) which made substantial contribution to the process of opening DNA base pairs. We model the bases by pendulums and reduce the biophysical problem to the mechanical problem of nonlinear oscillations of base pairs consisting of two coupled unequal pendulums, oscillating in the horizontal planes. We analyze the dynamical behavior of the model system, investigate its stability and construct the diagram of bifurcations. Oscillations of complementary DNA bases forming the base pairs, in which Adenine (A) forms two hydrogen bonds with Thymine (T) and Guanine (G) forms three hydrogen bonds with Cytosine (C), are of special interest, because they made substantial contribution to the process of opening DNA base pairs that, in turn, is considered as one of important elements of the process of the DNA-protein recognition. To study these oscillations, different approaches including molecular dynamics methods (Norberg & Nilsson, 2002; Cheatham, 2004), quantum mechanics methods (Sponer, 1996), and empirical methods (Peyrard & Bishop, 1989; Yakushevich et al, 2002; Kovaleva et al, 2006) are used. In this work, we apply the approach that have been recently used to study oscillations of a single DNA bases (Yakushevich et al, 2007). The approach is based on the ideas of Englander and co-authors (Englander et al, 1980), who noticed an analogy between rotational oscillations of bases and rotational oscillations of pendulums.

2. Modelling of an orbital cavity using finite element method

A starting point for this part of investigations is to obtain a reliable geometry of the examined object, as close as possible to the real one. Nowadays, the most popular and effective way to perform it is the use of computer tomography (CT) scanning.
In the presented analysis a CT scan of an orbital cavity in the form of a DICOM (Digital Imaging and Communications in Medicine) file was utilized (Fig. 1).

Fig. 1. CT scan of an orbital cavity

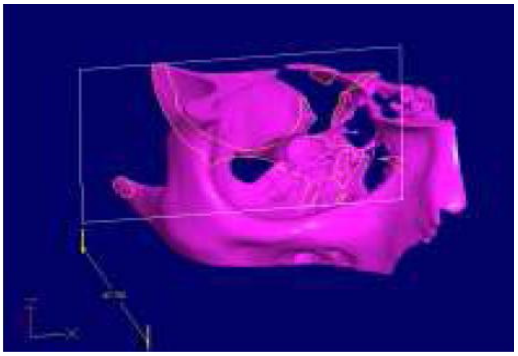


Fig. 2. Computer model of an eye-socket

Based on this file, a CAD eye-socket model was created by using a Solid Edge program (Fig. 2). The last step in preparation for calculations was the conversion of the CAD file to the format accepted by Ansys Workbench program.

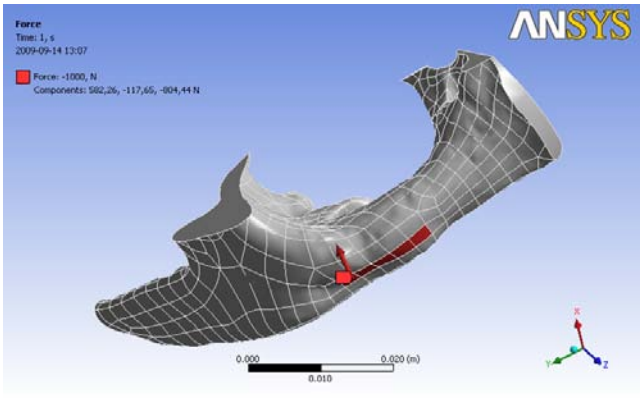


Fig. 3. Meshed orbital bone with applied force

Next, the grid of finite elements for the model subjected to numerical analysis has been generated. Covering the model's area by finite element network is based on the relationship between stress and strain (Hooke's law):

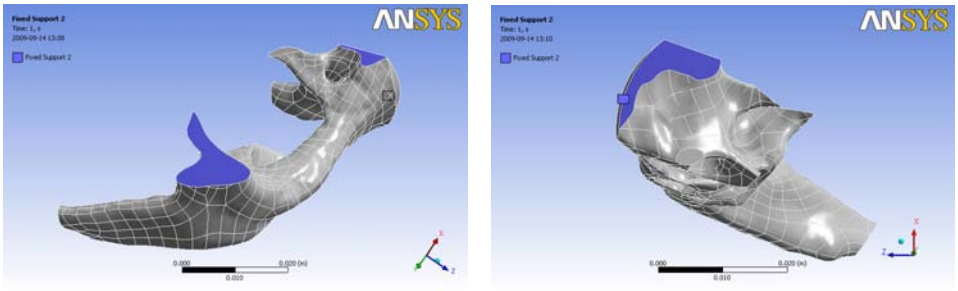
$$\sigma = \mathbf{D}\,\varepsilon\,, \tag{1}$$

where σ and ε denote stress and strain tensors, respectively, and $\boldsymbol{D}$ is the rigidity matrix (Rakowski & Kacprzyk, 2005).
Cortical bone material properties used in numerical calculation process was adopted from reference (Furusu et al., 2001), see Table 1. Loading, in the form of a concentrated force, was applied to the external part of a lower arch of zygomatic bone, in the place being statistically the most exposed to impact or pressure (Fig. 3).

Author	Cortical bone			Trabecular bone		
	E [MPa]	P [kg/m^3]	ν	E [MPa]	P [kg/m^3]	ν
Al-Bsharat [8]	12,200	2,120	0,22	1,300	990	0,22
Willinger [9]	15,000	1,800	0,21	4,500	1,500	0,01
Furusu [10]	11,500	2,000	0,3	40	862	0,45

Table 1. Material properties for cortical and trabecular bone.

For the numerical model of orbital bone the fixation scheme equivalent to the real eye-socket support was assumed (see Figures 4a and 4b).



(a) (b)

Fig. 4. Orbital bone fixation (two projections)

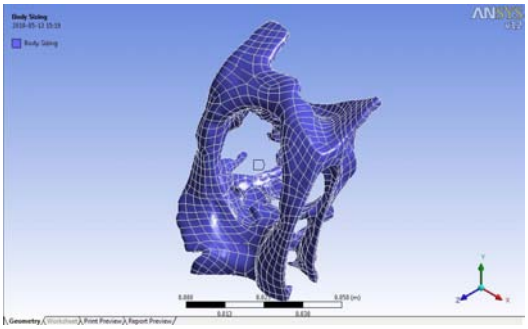


Fig. 5. Body sizing on whole right orbit

The same procedure as in the case of lower orbital arch was applied to the whole orbital cavity. A CAD model prepared in a Solid Edge program was exported to Ansys Workbench program in which geometry of the whole eye-socket was generated (Fig. 5).
Then, according to the aforementioned procedure, the whole model of eye-socket has been meshed using tetrahedron elements (Fig. 6).

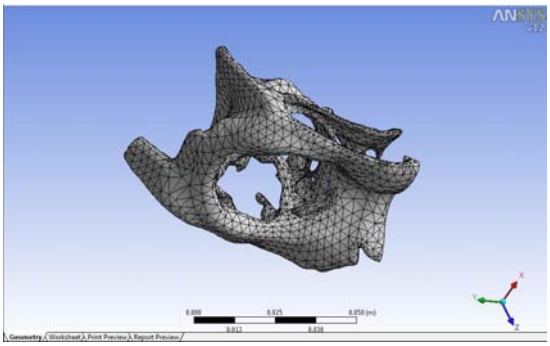


Fig. 6. Orbital cavity meshed by tetrahedron elements

As this was the case for a lower arch bottom, a force located in a lower arch of a zygomatic bone has been applied to the model. The force value as well as the bone material properties (Young modulus and Poisson ratio) was the same as in previous case.

2.1 Results

The results of numerical analysis were presented in the form of the map of stresses reduced according to the Huber-von Mises hypothesis, based on the strain energy of distortion. In conformity with this hypothesis the reduced stress can be calculated as follows:

$$\sigma_{red} = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2 - \sigma_x\sigma_y - \sigma_y\sigma_z - \sigma_z\sigma_x + 3(\tau_{xy} + \tau_{yz} + \tau_{zx})}\,, \tag{2}$$

where $\sigma_x, \sigma_y, \sigma_z$ are the normal stresses and $\tau_{xy}, \tau_{yz}, \tau_{zx}$ are the shear stresses.

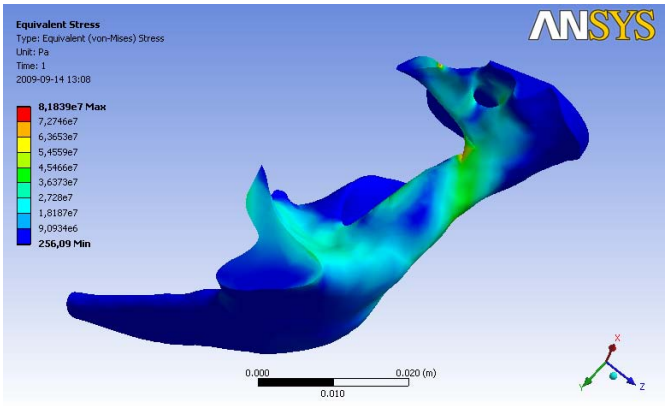


Fig. 7. Equivalent stress distribution in orbital bone

For the assumed boundary conditions and the place of force application the distribution of reduced stresses for a lower orbital arch was obtained (Fig. 7). It is readily to notice that the highest stress occur in its paranasal part. In case of an impact this is a potential region of bone fracture due to exceeding of allowable stress values.
The results of the analysis of the whole orbital cavity display a slightly different map of reduced stresses (Fig. 8). The stress concentration occurred here, in the external part of eye-socket. This is a confirmation of characteristics of real isolated fractures of orbital bottom.

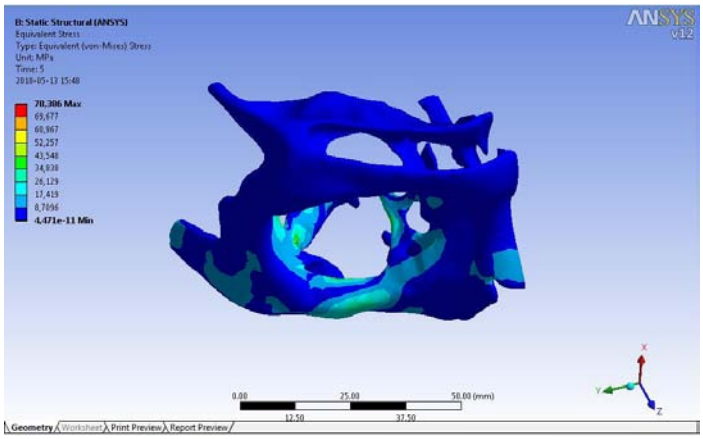


Fig. 8. Equivalent stress in orbital cavity

The last stage of a whole orbital cavity investigation was to calculate the total deformation. It may be concluded that the greatest deformations occur in the middle part of zygomatic bone (Fig. 9).

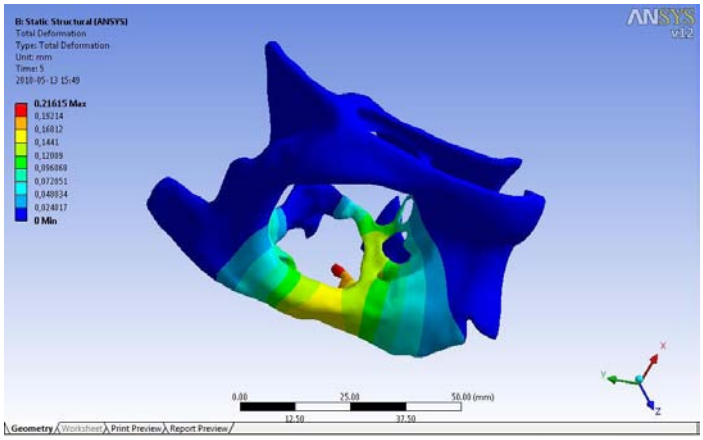


Fig. 9. Total deformation in orbital cavity

The results were obtained for material data presented in works (Al-Bsharat, et al., 1999) and (Willinger, 1999).

3. FEM pelvic bone modelling

The very first step was to prepare a model of a bone. For this purpose a wireframe model was created by using AutoCad. A set of 54 CT scan slices of heavy woman (100 kg, 50 years old), made with 5mm intervals along Z axis were used (see an example in Figure 10). Measuring of the bone layers thickness was made on the basis of CT scans and then compared with literature. Then, consequently slice images were imported to corresponding layers and one after another, closed boundary contours in XY plane, were sketched with straight lines. After that, all lines of adjacent contours were connected with straight lines in order to create triangles. The resultant wireframe was exported to ANSYS and meshed using triangular shell elements for cortical tissue (4 node triangular version of Shell 63 element type with real constant of 3mm for thickness and default element size of 5mm) and tetrahedral shape solid elements for trabecular tissue (10 node Solid 92 element type in tetrahedral shape, default element size was 5mm). In the end it gave 156738 of elements and 181980 nodes.
Because of strain concentration in acetabulum, it was decided to make there more precisely mesh, which gives 39286 surface elements and 194154 volume elements. Material properties of cortical bone and trabelacural bone were taken from (Kutz, 2003) and amounts: Young's modulus of 20000 MPa and Poisson's ratio of 0.3 for compact bone, and Young's modulus of 200 MPa and Poisson's ratio of 0.4 for spongy bone part. The thickness of the bone layers was taken from manual measurements compared with (Dalstra & Huiskes, 1995) and amounts 3 mm for the compact bone – the rest is the spongy one. The stress-strain state of the real pelvic bone is an effect of loads coming from the muscles attached to the bone through tendons and the weight of the upper part of the body. According to that, following boundary conditions were used (see Table 2 (forces) and Table 3 (movements restrictions)).

Muscle	**Component forces [N]**			**No. of nodes applied on**
	Fx	Fy	Fz	
RF - rectus femoris	15,2	0	834.8	110
S – Sartorius	-25.4	27	143.3	35
IP – iliacus	13.2	754.5	1681	1626
GRA – gracilis	35.6	-17.1	160.1	22
GMx - gluteus maximus	1050.7	490.2	2031.4	197
ST - semitendinosus	-11.9	5.9	225.6	25
SM – semimembranosus	239.2	35.3	1337.3	142
BCL - biceps femoris	131.2	0	733.4	72
ADM - adductor magnus	679.7	-117.8	1631.1	91
ADL - adductor longus	269.8	-257.2	461.2	30
ADB - adductor brevis	291	-273.7	211.4	24
PC – pectineus	107.6	-125.2	89.9	62
GMD - glutaeus medius	-97	-217.7	1198.8	741
GMU - glutaeus minimus	78.9	-67.8	684.4	1052
TFL - tensor fasciae-latae	-7.5	-23.8	284.9	43

Table 2. Muscle forces, forces direction and number of chosen nodes (based on (Wall, 2002; Cranney, 2007))

It should be noted that the used values were based on maximum forces determined by multiplication of crossectional area of each muscle in its physiological state and a value of maximum stress which can be induced in the muscle σ_{max}=0.5MPa.
In the assumed model we applied the following schemes of supports and movements restrictions (see Table 3):

Location	Restraint movements
Connnection with sacrum bone	X, Y, rotation in X, Y and Z axis
Acetabulum	X, Y, Z
Pubic psymphysis	X, Y, rotation in X and Z axis

Table 3. Movements restriction

Values of forces acting on the pelvic bone were taken from literature. Static forces values make calculations faster. During flexion following muscles take a part: rectus femoris, sartorius, iliacus, gracilis, pectineus, during extension: gluteus maximus, semitendinous, semimembranous, biceps femorisgluteus medius, during adduction: semitendinous, semimebranous, adductor magnus, adductor longus, adductor brevis, during abduction: tensor fasciale lateae, gluteus medius sartorius. An assumption was made that all the movements are made while standing on one leg only. From (Phillips, 2007) and (Huston, 2007) mass percentage of lower limb was estimated, equal to 19%. The resultant weight of 800 N coming from the upper body and single lower limb, was directed downwards along the Z axis and applied on nodes related with area of connection between pelvic and sacrum bones. For every motion a separate model was prepared with its boundary conditions. An example for flexion and extension can be seen in Figure 10.

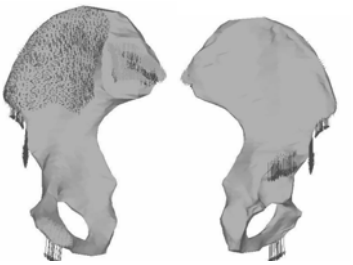


Fig. 10. Muscle forces configuration for flexion

3.1 Results and discussion

For all four load variants related to basic motions of flexion, extension, adduction and abduction the results were analyzed by plotting the three principal stresses, the Huber von

Mises equivalent stress, total displacement and the Huber von Mises equivalent strain. Plot in Figure 11 shows the Huber von Mises equivalent strain for flexion. The biggest value is placed in the middle part of acetabulum and amounts about 0.005.

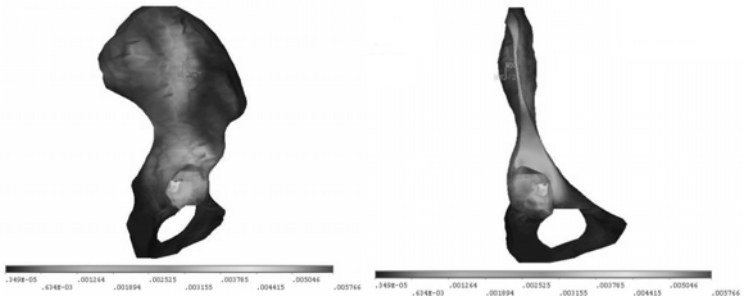


Fig. 11. Huber von Mises eq. strain for flexion and cross-sectional view

From 1st principal stresses it can be observed that the biggest tensile stresses occur during extension at acetabulum and are equal to 32.6 MPa (Fig. 12). Greatest compressive stress of -31 MPa comes from the 3rd principal stress and is found also for extension on the surface of acetabulum (Fig. 13).

Fig. 12. 1st principal stress for extension

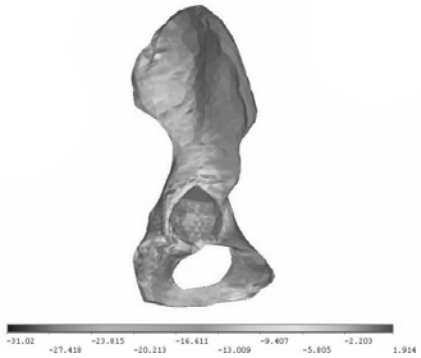


Fig. 13. 3rd principal stress for extension

The highest value of the Huber von Mises equivalent stress, as it was easy to expect, is also characteristic for acetabulum surface during extension movement and is equal to 28.6 MPa (Fig. 14). Here it can be seen from the crossectional view that the highest stresses are placed at the external parts of the cortical layer, where stresses for trabecular layer are much lower. Similar effect can be observed in models of other authors, even for full solid models (see (John, 2004) to compare).

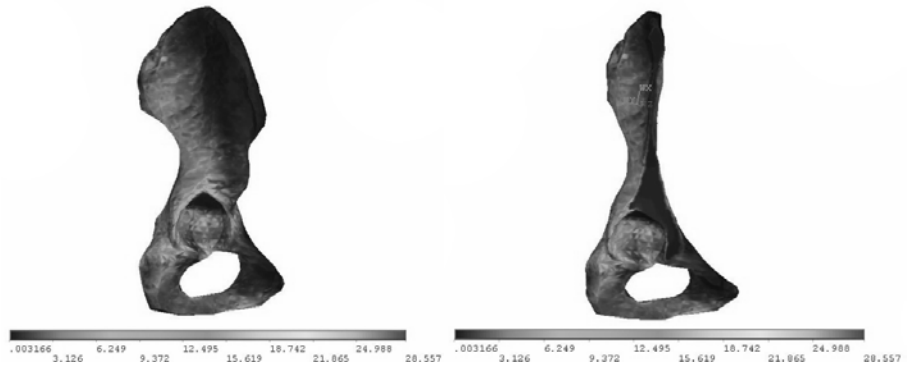


Fig. 14. Huber von Mises eq. stress for extension and cross-sectional view

Regarding the total displacements the biggest and equal to 0.28 mm are observed for extension (Fig. 15), located at ischial tuberosity and for abduction equal to 0.23 mm at front part of iliac crest (Fig. 19).
For all of the presented motion variants the greatest Huber von Mises strains occur in acetabulum. Similar conclusions can be found in the works (Kutz, 2003) and (Phillips, 2007). The highest strain is found for extension and amounts to 0.009 (Fig. 16) compared to adduction 0.006 (Fig. 17) and abduction 0.004 (Fig. 20).

Fig. 15. Total displacement for extension and cross-sectional view

Analysing the stress distribution for rest of the movements the maximum principal and equivalent stresses occur on acetabulum, except adduction for which greatest stresses are placed on the inside surface of superior ramus of ischium, where equivalent stress is equal to 14.8 MPa (Fig. 18).

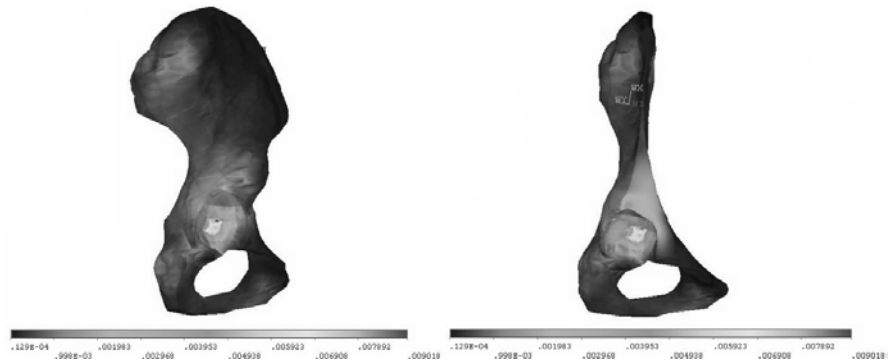


Fig. 16. Huber von Mises eq. strain for extension and cross-sectional view

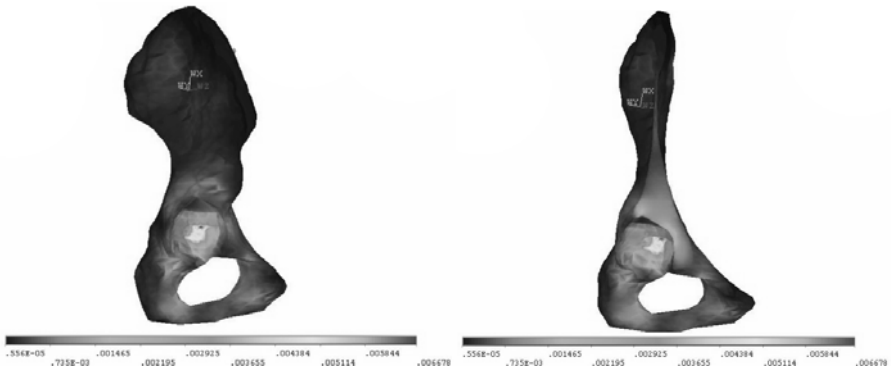


Fig. 17. Huber Von Mises eq. strain for adduction and cross-sectional view

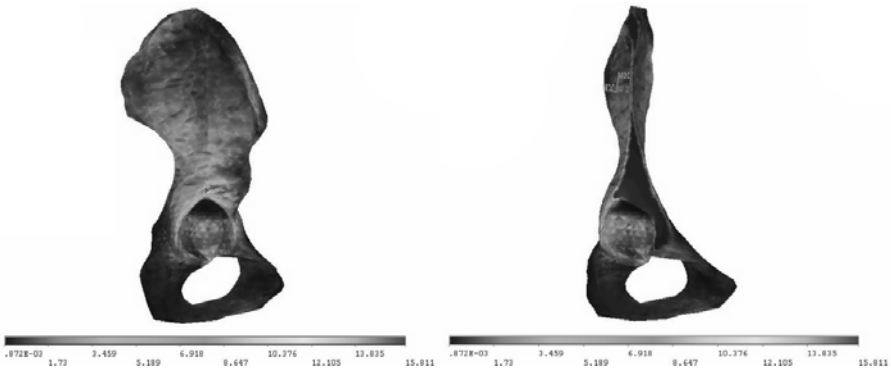


Fig. 18. Huber Von Mises eq. stress for abduction and cross-sectional view

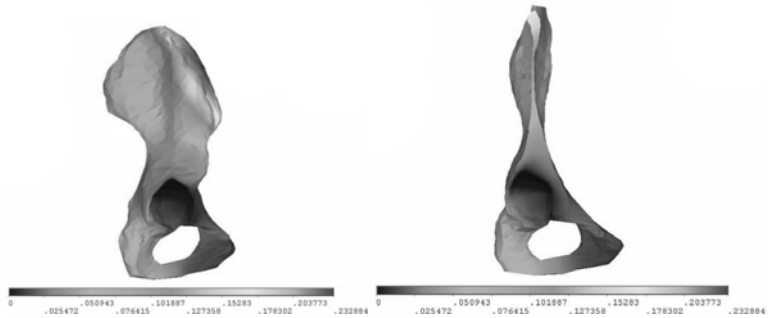


Fig. 19. Total displacement for abduction and cross-sectional view

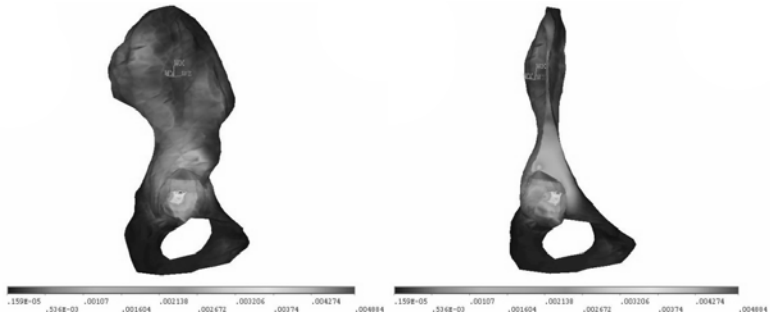


Fig. 20. Von Mises eq. strain for abduction and cross-sectional view

The smallest values of displacements are found on constraint surface of acetabulum for all motion variants.

The above observations prove that the acetabulum of pelvic bone is a main subject of pathology of the whole pelvis girdle. As a place of biggest stresses and strains concentration, acetabulum is exposed to a high risk of degeneration.

Earlier works on modeling pelvic bone, as an example (Dalstra & Huiskes, 1995) had not enough accurately mesh (inter alia, because of the possibility of computing) or models of the pelvic bone were simplified, for example authors were used only surface model (only trabercular bone), without the spongy one (Phillips, 2007; John, 2004). That is the reason why there is still a place for improvement of model and loads (Phillips, 2007) or models. This simulation was made for static situation – one leg standing. It is well known that the hip joint during normal walking is around 300% body weight (Bergmann et al, 2001) (we use 100% of body weight – mass of the leg + muscle forces). FEM method can give consistent results with medical observation even for more complicated situation as for example sideways fall (Majumder et al, 2007).

4. Modelling of DNA bases

The analogy permits to reduce the problem of oscillations of complementary DNA bases forming the base pairs (AT or GC) to the mechanical problem of oscillations of two non-identical coupled pendulums shown schematically in Fig. 21. In reference (Yakushevich et

al, 2009), linearized (particular) version of the problem has been investigated. Here we consider the nonlinear (general) version.

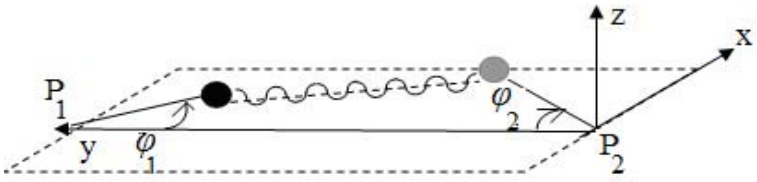


Fig. 21. Two coupled non-identical pendulums oscillating in the horizontal (xy) plane.

The function of Lagrange of the model imitating oscillations of two non-identical coupled pendulums shown in Fig. 21 has the form

$$L = T - V = \frac{I_1}{2}(\frac{d\phi_1}{dt})^2 + \frac{I_2}{2}(\frac{d\phi_2}{dt})^2 - \frac{1}{2}K_{12}(l - a^*)^2, \tag{3}$$

where φ_1 and φ_2 are the angles of inclination of the 1-st and 2-nd pendulums; I_1 and I_2 are the moments of inertia of the 1-st and 2-nd pendulums; K_{12} and l are the rigidity and the length of the spring connecting the pendulums; a^* is the length of the spring in its relax state being less than the distance a between masses of pendulums in the equilibrium state ($\varphi_1 = \varphi_2 = 0$). Therefore, $l > a > a^*$. After taking into account approximation $a^* << l$, damping and general external force, the corresponding equations of Lagrange take the form

$$I_1\frac{d^2\phi_1}{dt^2} + \frac{1}{2}K_{12}[A_1\sin\phi_1 - B\sin(\phi_1 + \phi_2)] = -\beta\frac{d\phi_1}{dt} + F, \tag{4}$$

$$I_2\frac{d^2\phi_2}{dt^2} + \frac{1}{2}K_{12}[A_2\sin\phi_2 - B\sin(\phi_1 + \phi_2)] = -\beta\frac{d\phi_2}{dt} + F, \tag{5}$$

where $A_1 = 2r_1(r_1 + r_2 + a), A_2 = 2r_2(r_1 + r_2 + a), B = 2r_1r_2$, and r_1, r_2 are the lengths of the first and of the second pendulums, respectively; β is the coefficient of dissipation, and F is an external generalized force. For simplicity only the case when F = const is considered.

4.1 Solutions of model equations and trajectories in the configuration space

Assuming that $t = \lambda\tau$, $\lambda = \sqrt{\frac{2I_1}{K_{12}B}}$ we have transformed equations (4)-(5) to non-dimensional system:

$$\frac{d^2\phi_1}{d\tau^2} + \frac{A_1}{B}\sin\phi_1 - \sin(\phi_1 + \phi_2) = -b_{01}\frac{d\phi_1}{d\tau} + k_{01}, \tag{6}$$

$$\frac{d^2\phi_2}{d\tau^2} + \frac{A_2}{B}I_{12}\sin\phi_2 - I_{12}\sin(\phi_1 + \phi_2) = -b_{01}I_{12}\frac{d\phi_2}{d\tau} + k_{01}I_{12}, \tag{7}$$

where $b_{01} = \beta\sqrt{\frac{2}{I_1 K_{12} B}}$, $k_{01} = \frac{2F}{K_{12}B}$, $I_{12} = \frac{I_1}{I_2}$.

We solved numerically equations (6)-(7) with the help of the MAPLE. To construct corresponding graphs, we used the values of the coefficients I, β, F, K_{12} and parameters A_1, A_2, B, a that are presented in Tables 4 - 5. Estimated values of parameters b_{01}, k_{01}, I_{12} are in Table 6.

Base	I [m^2kg] x10^{-44}
A	7.60703
T	4.86228
G	8.21744
C	4.10693

Table 4. Coefficient I (Yakushevich et al, 2005)

Base pair	A_1 [10^{-20} m^2]	A_2 [10^{-20} m^2]	B [10^{-20} m^2]	a [10^{-10} m]	β [10^{-34} Js]	F [10^{-22} J]	K_{12} [10^{-3} N/ m]
AT	208,8	172,8	55,68	7,4	4.25	3.12	2,51
GC	205,2	169,2	53,58	7,6	4.25	3.12	3,80

Table 5. Parameters A_1, A_2, B and coefficients β, F and K_{12} (Yakushevich et al, 2005, 2009)

Base pair	b_{01}	k_{01}	I_{12}
AT	0,0583	0,4465	1,5645
GC	0,0465	0,3065	2,0008

Table 6. Parameters b_{01}, k_{01}, I_{12}

Graphs of the solutions of equations (6)–(7) obtained for AT and for GC base pairs and for three cases: (a) ideal case when effects of dissipation and external fields are neglected; (b) non-ideal case when effects of dissipation are taken into account but effects of external fields are neglected; (c) non-ideal case when both effects of dissipation and effects of external fields are taken into account, are presented in Fig. 22. The trajectories of the considered dynamical system are presented in Fig. 23.

When constructing the graphs of the solutions and trajectories, we used initial conditions:

$$\phi_1(0) = 0.3753, \quad \phi_2(0) = 0.2347, \quad \frac{d\phi_1}{d\tau}(0) = 0.1, \quad \frac{d\phi_2}{d\tau}(0) = 0.1$$

which are closed to the point (0.2753, 0.3347, 0.0, 0.0) in the phase space, and we shall show in the next section that this point is stable in the case of AT base pair.

Summarizing the results presented in Fig. 22-23, we can state that effects of dissipation lead to decreasing the amplitudes of base oscillations. Effects of external field lead to displacement of the equilibrium positions, displacement in the case of GC base being less than displacement in the case of AT base pair. So, we can state that polynucleotide chains saturated by AT base pairs are more sensitive to the action of external field than the chains saturated by GC base pairs.

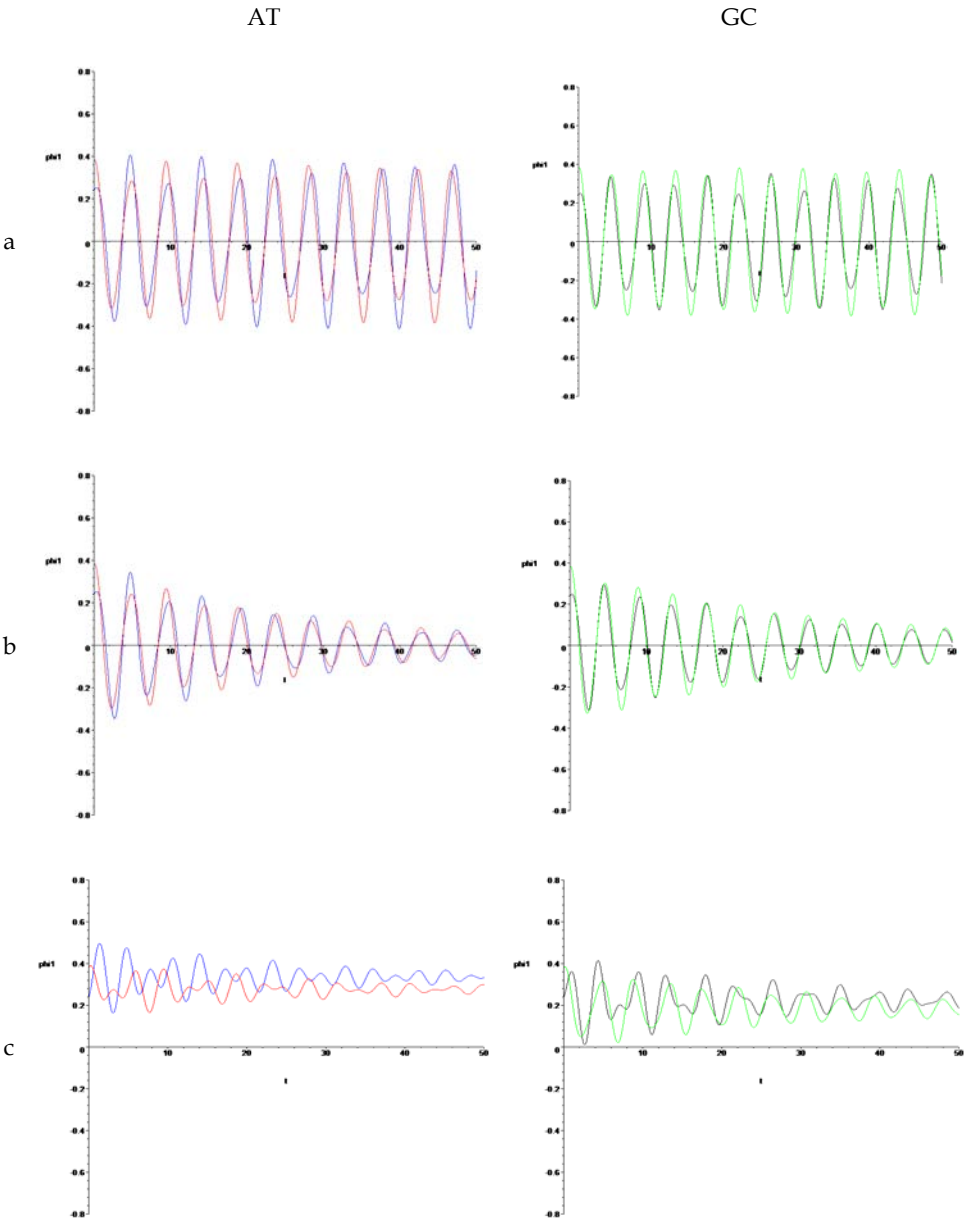


Fig. 22. Solutions of equations (4) – (5) obtained for AT and GC oscillating base pairs and for three different cases: (a) $\beta = 0$, $F = 0$; (b) $\beta \neq 0$, $F = 0$; (c) $\beta \neq 0$, $F \neq 0$.

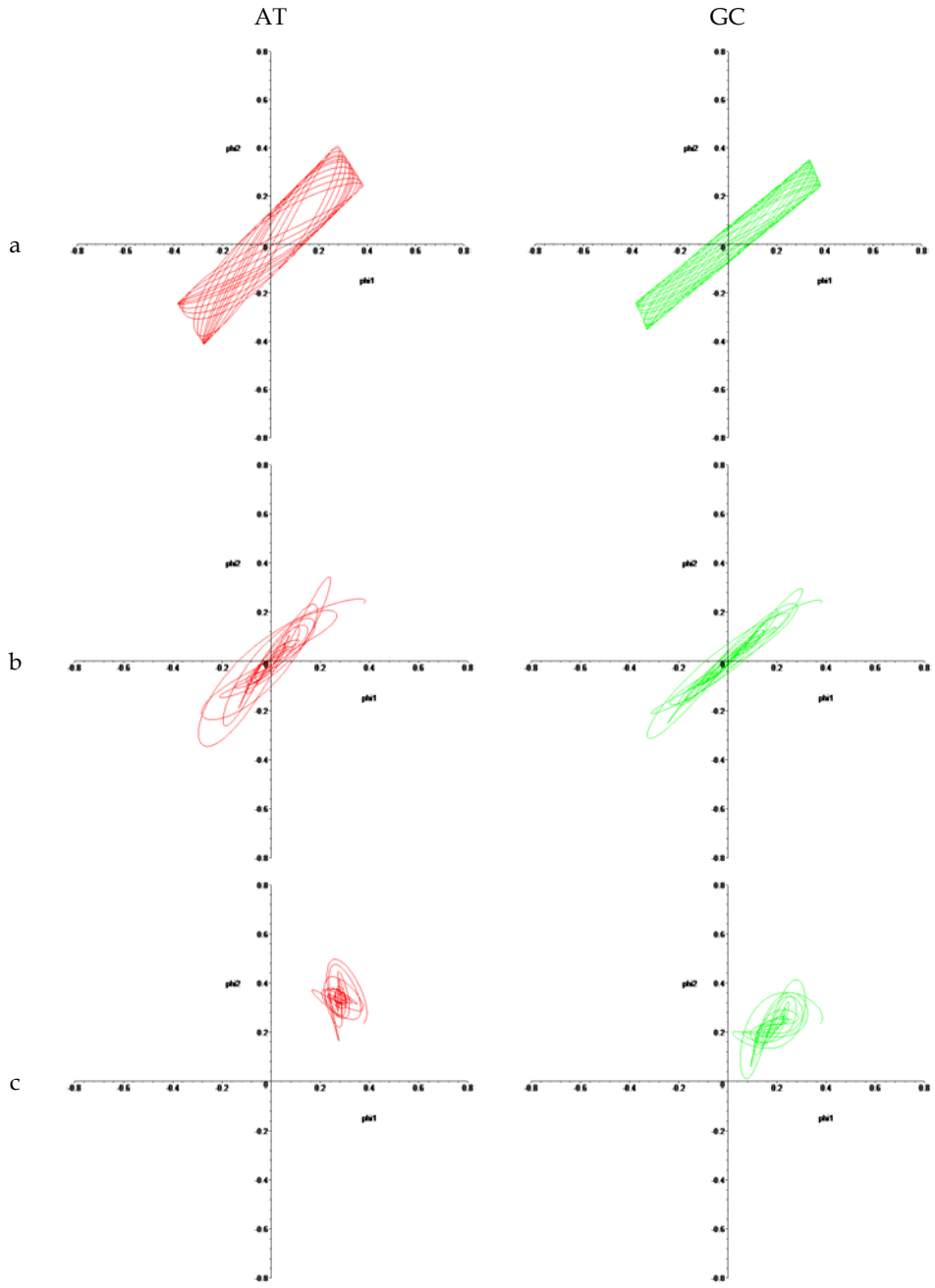


Fig. 23. Trajectories in the configuration space $\{\varphi_1; \varphi_2\}$ calculated for AT and GC oscillating base pairs and for three different cases: (a) $\beta = 0$, $F = 0$; (b) $\beta \neq 0$, $F = 0$; (c) $\beta \neq 0$, $F \neq 0$.

4.2 Analysis of stability and diagram of bifurcation

To analyze the stability of the system (6)-(7), we firstly transformed the system of two differential equations of the second order (6)-(7) to the system of four differential equations of the first order

$$\begin{aligned}
&\dot{\phi}_1 = \gamma_1,\\
&\dot{\phi}_2 = \gamma_2,\\
&\dot{\gamma}_1 = -\frac{A_1}{B}\sin\phi_1 + \sin(\phi_1+\phi_2) - b_{01}\gamma_1 + k_{01},\\
&\dot{\gamma}_2 = -\frac{A_2}{B}I_{12}\sin\phi_2 + I_{12}\sin(\phi_1+\phi_2) - b_{01}I_{12}\gamma_2 + k_{01}I_{12}.
\end{aligned} \tag{8-11}$$

We have noticed that (6)-(9) is invariant under the transformation $\phi_i \to \phi_i + 2\pi$. As a result, every solution $|\phi_i| > \pi$ we can identify with solution $\phi_i^* = \phi_i (\mathrm{mod}\, 2\pi)$.
The equilibriums of (6)-(9) are given by $(\phi_{10}, \phi_{20}, 0.0, 0.0)$, where ϕ_{10} and ϕ_{20} are solutions of the following algebraic equations

$$\begin{aligned}
&-\frac{A_1}{B}\sin\phi_1 + \sin(\phi_1+\phi_2) + k_{01} = 0,\\
&I_{12}\left(-\frac{A_2}{B}\sin\phi_2 + \sin(\phi_1+\phi_2) + k_{01}\right) = 0.
\end{aligned} \tag{12}$$

Since system of equations (12) is nonlinear, we cannot find solution analytically but we can do it numerically. We found that for each base pair there are four equilibrium points. Only one of them being stable, as all eigenvalues of the Jacobian of the right side of (8) – (11) have negative real parts. It is P^4_{AT} = (0.2753, 0.3347, 0.0, 0.0) point in the case of AT base pair and P^4_{GC} = (0.1658, 0.1658, 0.0, 0.0) point in the case of GC base pair.
Below we present the diagrams of bifurcation of the analyzed system. In general, diagrams of that type show possible equilibrium points or periodic orbits of a system as a function of one of the parameters of the system. For every value of that parameter we can choose many initial conditions and find the trajectories. If trajectory is attracted to the stable equilibrium, we obtain only one point in the diagram, if trajectory is approaching to periodic (or quasi-periodic) orbit we obtain two (or more) points in the diagram. If each trajectory is attracted by different point, we obtain a lot of points in the diagram which mean a chaotic behavior of the system. Bifurcation occurs when the change of the value of parameter causes the changes in the behavior of the analyzed system.
In Fig. 24 (a) - (b) we present the bifurcation diagrams as a function of k_{01} parameter, describing the external force. Because we assumed that the external force is constant, the increase of value of k_{01} parameter does not change the dynamical properties of the system (in both, AT and GC, cases) and we do not observe the bifurcation. It is worth to notice, that bigger value of k_{01} corresponds to the bigger displacement of the bases from the (0.0,0.0) equilibrium position
In Fig. 24 (c) - (d) we present the bifurcation diagrams as a function of b_{01} parameter, describing the parameter of dissipation. For very small values of b_{01} it seems that there exists some periodic or quasi-periodic trajectory (see Fig. 25.).

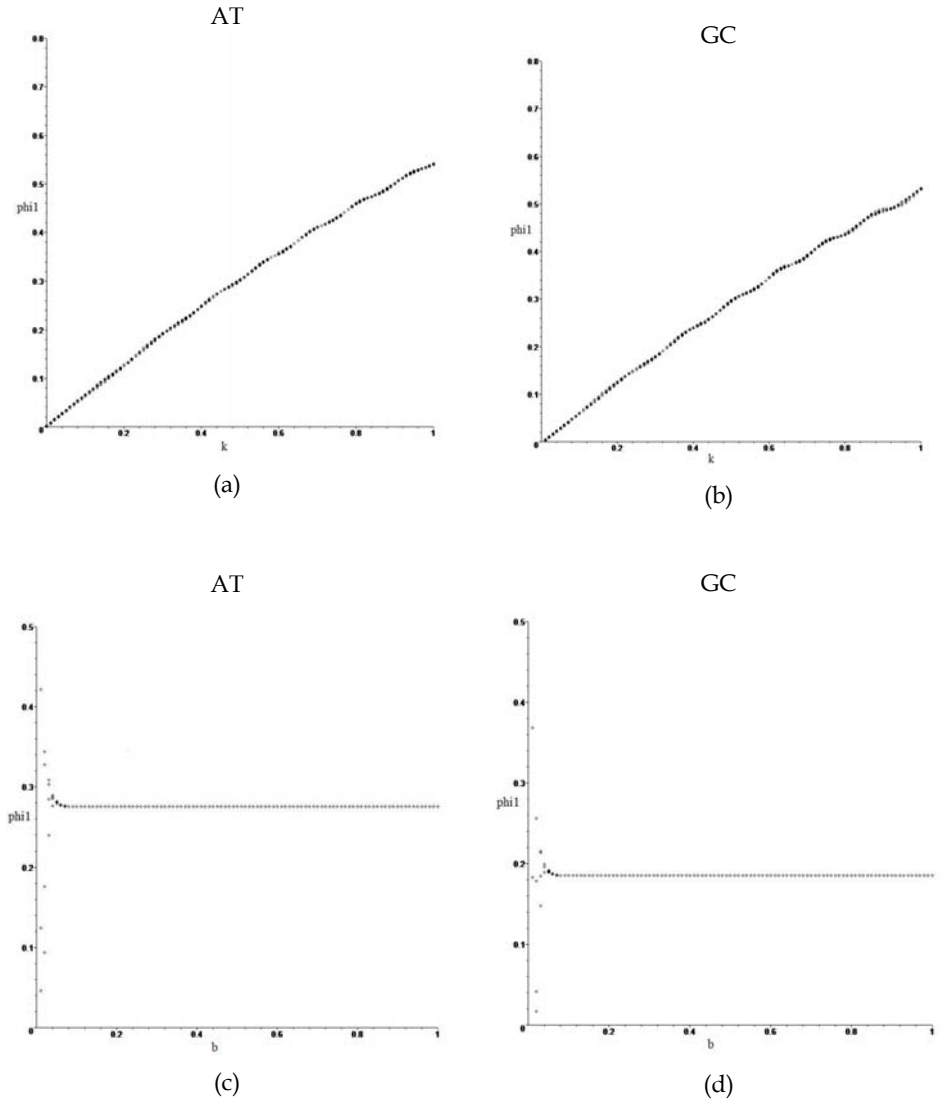


Fig. 24. Diagrams of bifurcation

In fact, trajectories are very slowly approaching to the stable point, and in the bifurcation diagrams we observe transient state of the system for small values of b_{01} parameter. For bigger values of b_{01} parameter we observe one solution for each value of the parameter. It means that trajectories more rapidly approach to the stable point. In Fig. 24 (c)-(d) we can notice also, that for each base pair there is a different equilibrium point, which is in agreement with our calculations.

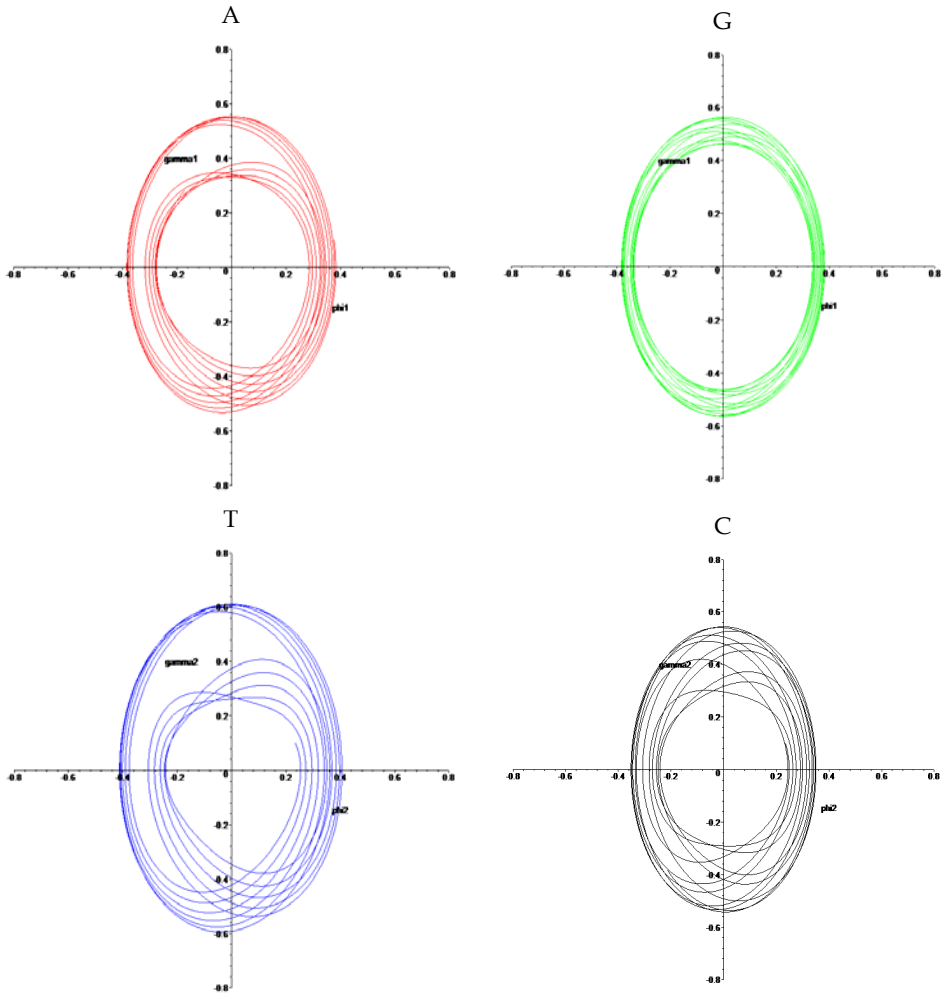


Fig. 25. Trajectories in the phase space $\{\varphi; \gamma\}$ calculated for A, T, G and C bases and for the case $\beta = 0.05$ and $F \neq 0$.

5. Conclusions

One of the most dangerous consequences of the road accidents is the facial skull injury. Especially, the eye-socket damage can cause the serious vision problems and brain dysfunctions. Hence we have presented the results of stress and strain analysis of an orbital cavity. Thanks to the numerical analysis in Ansys Workbench the concentration of the reduced stresses in eye-socket had been confirmed. In case of isolated fractures of an orbital bottom, the lower edge of an orbital cavity due to its elasticity does not break but only bends transmitting the force to a brittle bottom of an orbital which is prone to fracture (Smolarczyk-Wanyura, 2005). Further work is required in case of numerical model

optimization but also in the fact that simulation must include various loading cases. The study may provide an assistance for surgeons performing bony face operations.
In the case of a pelvic bone our results show, that most of the highest stresses and strains occurred in acetabulum. The comparison with medical data cases (Schmalzried et al, 2004), indicates that degeneration and wear of acetabular regions is the main case of pathology in the area of pelvic girdle. It is also consistent with the results presented in (Będziński, 1997). Results are comparable also with papers (Dalstra & Huiskes, 1995) and (Phillips, 2007), i.e. the same magnitude of results is reached. This method can be used for examination of human pelvic bone and can help, after some additional measurements regarding the bone density estimation, to predict when the failure (and if) could happen. It is also possible to test prototypes of pelvic bone implants.The presented method of creating a 3D model from images produced with Computer Tomography, although based on many simplifications, can be used for many other analyses, where the examined subject has an irregular shape. The presented model is not an ideal one. Each pelvic bone is different. The shape of it depends on sex and age. The presented algorithm serves as a proposal of a method of pelvic bone modeling. In future, if enough cases will be modeled, it will be possible to found a database of ready to use models. Minor changes will be needed for a data fitting. There is still a lot of space for model improvements like building a much more detailed 3D geometry model, creation of much finer mesh, considering anisotropic properties for material or modeling phenomena during walking, running, jumping (like in (Schmalzried et al, 2004)) or even skiing, when because of higher forces value, hip fracture is more likely. Creation of more accurate model and denser mesh will be possible, provided that more CT scans are available – depending on computational resources.This will allow to avoid mistakes such that bone contacts only in the fascia lunata and do not in the center where the ligamentum capitis femoris placed. Due to technical limitations adopted element are of size 5mm and constant thickness amounts 3mm. Authors of paper (John & Wysota, 2010) propose a method based on CT images analysis of tissue density and determination of Young modulus. They took into account the dual structure of the bones.
In section 4 we have studied rotational nonlinear oscillations of two coupled DNA bases that form a base pair: adenine-thymine (AT) or guanine-cytosine (GC). We have presented the graphs of solutions and trajectories in configuration space. We have also found stable equilibrium points for each base pair and constructed diagrams of bifurcations. This work part can be considered as a base for more complicated model of DNA, including not only one base pair, but a fragment of DNA consisting of three base pairs.

6. References

Al-Bsharat, A., Zhou, C., Yang, K.H., Khalil, T., King, A.I. (1999). *Intercranial Pressure in the Human Head Due to Frontal Impact Based on a Finite Element Model*. Detroit, Michigan, Wayne State University.

Będziński, R. (1997). *Biomechanical Engineering,* Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, in Polish.

Bergmann, G., Deuretzbacher, G, Heller, M., Greichen, F., Rohlmann, A., Strauss, J. et al., (2001). Hip forces and gait patterns from routine activities, *J. Biomech.*, 34, 859-871.

Boryor, A., Geiger, M., Homann, A., Wunderlich, A., Sander, Ch., Sander, F.M., Sander, F.G. (2008). Stress distribution and displacement analysis during an intermaxillary

disjunction-A three-dimensional FEM study of a human skull. *Journal of Biomechanics* 41, 376-382.

Cheatham, T.E III (2004). Simulation and modeling of nucleic acid structure, dynamics and interactions, *Curr. Opin. Struct. Biol*., 14, 360-367.

Cranney, A., Jamal, S.A., Tsang, J.F., Josse, R.G., Leslie, W.D. (2007). Low bone mineral density and fracture burden in postmenopausal women, *CMAJ: Canadian Medical Association journal,* 177(6), 575–80.

Dalstra, M., Huiskes, H.W.J. (1995). Load transfer across the pelvic bone, *J. Biomech.,* 28(6): 715-724.

Englander, S.W., Kallenbach, N.R., Heeger, A.J., Krumhansl, J.A., Litwin A. (1980). Nature of the open state in long polynucleotide double helices: possibility of soliton excitations, *Proc. Natl. Acad. Sci*., 77, 7222-7226.

Furusu, K. et al. (2001). Fundamental Study of Side Impact Analysis using the Finite Element Model of the Human Thorax. *Japan Society of Automobile Engineers Review,* 22, Elsevier Science BV 195-199.

Gautam, P., Valianthan, A., Adhikari, R. (2007). Stress and displacement patterns in the craniofacial skeleton with rapid maxillary expansion: A finite element method study. *Americvan Journal of Orthodontics and Dentofacial Orthopedics*.

Huston, R.L., (2009). *Principles of Biomechanics*, Taylor & Francis Group.

Iannetti, G., Fadda, M.T., Indrizzi, E., Gennaro, P., Spuntarelli, G. (2004). Scoliosis of the cranial base: radiological and mathematical analysis using finite elements system analysis (FESA) of a case. *Journal of Cranio-Maxillofacial Surgery,* 32, 220-227.

Jehad Al-Sukhun, Risto Kontio, Christian Lindqvist (2006). Orbital Stress Analysis—Part I: Simulation of Orbital Deformation Following Blunt Injury by Finite Element Analysis Method, available online 16 February. http://www.sciencedirect.com/.

John, A. (2004). *Identification and analysis of geometric parameters and mechanical properties of human pelvic bone*, ZN Politechniki Śląskiej, S. Mechanika, Gliwice; in Polish.

John, A., Wysota, P. (2010). The procedure to aid of diagnosis the osteoporosis in human pelvic bone, *Modelowanie Inżynierskie*, 35, 1-6.

Kovaleva, N.A., Savin, A.V., Manevitch, L.I., Kabanov, A.V., Komarov, V.M., Yakushevich, L.V. (2006). Topological solitons in inhomogeneous DNA molecule. *Polymer Science,* 48-A, 278-293.

Kutz, M. (2003). *Standard Handbook of Biomedical Engineering and Design,* McGraw-Hill.

Majumder, S., Roychowdhury, A., Pal, S. (2007). Simulation of hip fracture using a 3D finite element model of pelvis-femur-soft tissue complex with simplified representation of whole body, *Medical Engineering & Physics,* 29, 1167-1178.

Mazurkiewicz, A., Topoliński, T. (2009). Relationships between structure, density and strength of human trabecular bone, *Acta of Bioengineering and Biomechanics*, 11(4), 55-61.

Mrozowski, J., Awrejcewicz, J. (2004). *Fundamentals of Biomechanics*, The Technical University of Lodz Press, Łódź, in Polish.

Nagano, A., Komura, T., Fukashiro, S., Hinemo, R. (2005). Force, work and Power output of Lower limb muscles Turing human maximal-effort countermovement jumping, *Journal of Electromyography and Kinesiology*, 15, 367-376.

Norberg, J., Nilsson, J. (2002). Molecular dynamics applied to nucleic acids, *Acc. Chem. Res.*, 35, 465-472.

Peyrard, M., Bishop, A.R. (1989). Statistical mechanics of a nonlinear model for DNA denaturation, *Phys. Rev. Lett.,* 62, 2755-2758.

Phillips, A.T.M., Pamkaj, P., Howie, C.R., Usmani, A.S., Simpson, A.H.W.R. (2007). Finite element modeling of the pelvis: Inclusion of muscular and ligamentous boundary conditions*, Med. Eng. & Phys*., 29, 739-748.

Rakowski, G., Kacprzyk, Z. (2005). *Finite element method in structural mechanics* Oficyna Wydawnicza Politechniki Warszawskiej, (in Polish).

Roth, S., Raul, J-S., Willinger, R. (2009). Finite element modelling of paediatric head impact: Global validation against experimental data. *Computer methods and programs in biomedicine* , www.intl.elsevierhealth.com/journals/cmpb.

Sander, S., van den Bedem, S.P.W., van Keulen, F., van Helm, F.C.T., Simonsz, H.J. (2006). A finite-element analysis model of orbital biomechanics. *Vision Research,* 46, 1724-1731.

Schmalzried, T., Shepherd, E.F., Dorey, F.J., Jackson, W.J., Rosa, M., Fa'vae, F., McKellop, H., McClung, C.D., Martell, J., Moreland, J.R., Amstutz, H.C. (2000). Wear is a function of use, not time, *Clin Ort. Relat Res*., 381, 36-46.

Smolarczyk-Wanyura, D. (2005). Management of patients with isolated fractures of the floorof the orbit – the author's own observation, *Czas. Stoma.,* LVIII, 7. (in Polish).

Sponer, J., Leszczynski, J., Honza, P. (1996). Hydrogen bonding and stacking of DNA bases. A review of ab initio quantum-chemical studies. *J. Biomol. Struct. Dyn*., 14, 117-135.

Wall, A. (2002). Clinical aspects of total hip arthroplasty, *Acta of Bioengineering and Biomechanics*, 4(1), 16-19.

Willinger, R., Sung, H., Diaw, B. (1999). Development and Validation of a Human Head Mechanical Model. *Biomechanics,* 327[IIb], 125-131.

Yakushevich, L.V., Savin, A.V., Manevitch, L.I. (2002). On the Internal dynamics of topological solitons in DNA, *Phys. Rev*., E-66, 016614.

Yakushevich, L.V., Gapa, S., Awrejcewicz, J. (2009). *Mathematical analysis of equation imitating DNA base oscillations.* In: Mathematics. Computer. Education, 16(2), Edited by G.Yu. Riznichenko, Moscow – Izhevsk, RCD, pp. 135 – 144.

Yakushevich, L.V., Gapa, S., Awrejcewicz, J. (2009). *Mechanical analog of the DNA base pair oscillations*. In: Dynamical Systems – Theory and Applications, Edited by J. Awrejcewicz, Lodz, 879-886.

Yakushevich, L.V., Krasnobaeva, L.A., Shapovalov, A.V., Quintero, N.R. (2005). One- and two-soliton solutions of the sine Gordon equation applied to DNA, *Biophysics*, 50, 450-455.

16

Study Regarding Numerical Simulation of Counter Flow Plate Heat Exchanger

Grigore Roxana, Popa Sorin, Hazi Aneta and Hazi Gheorghe
"Vasile Alecsandri" University of Bacau
Romania

1. Introduction

Heat exchangers are equipments commonly found in industrial applications. Virtually not exist almost industrial area which has not a heat exchanger. This is used to exchange heat between two fluids, cooling and heating processes, heat recovery. The most performant, in terms of heat transfer, are the plate heat exchangers. These types of heat exchangers have a lot of advantages, including a high heat exchange area per unit volume and good heat transfer performance.
An important number of numerical studies applying finite element method have been made to research fluid flow and heat transfer into heat exchangers (Gut&Pinto,2003), (Saber&Mazaher Ashtiani,2010), (Awrejcewicz et al., 2007).
The paper presents a theoretical and experimental study on plate heat exchanger. It is performed a numerical simulation of a counter flow plate heat exchanger using finite element method. A 3D model was developed to analyze thermal transfer and fluid flow along the plate heat exchanger, using COSMOS/Flow program. The results are presented graphically and numerically. In parallel, starting from the same input data, it makes thermal calculations for the studied plate heat exchanger. The basic equations are the equation of heat balance for thermal agents and plate heat transfer equation. The calculation is iterative and has certain features related to channel geometry.
Validation of the models presented is made by comparing the measured values obtained on experimental study.

2. Presentation of studied plate heat exchangers - experimental results.

The studied heat exchanger is a pack of 8 stainless steel thermal plates with gaskets. These plates are assembled together in cast iron frames and there are chevron type plates. The hot water flows are in one direction in alternating chambers while the cold water flows are in counter flow in the other alternating chambers like in figure1. The number of passes is 1 and the thermal agents are directed into their proper chambers either by a suitable gasket made from ethylene propylene rubber (EPDM). The width of channel between plates is H_o= 0,003m and the number of channels is N_c =4. Overall heat transfer surface is S=0,218m^2.The geometric dimensions of the thermal plate are represented in table 1.
Such a heat exchanger can be used to warm the cold water, considered a secondary thermal agent, with hot water, a primary thermal agent. Figure 1 shows the simplified presentation of a plate heat exchanger with eight plates in counter flow arrangements.

Dimension	Notation	Value	Unit
Diameter of the inlet tube	d	0,003	m
Effective plate length, measured between ports	L	0,386	m
Effective plate width, measured between ports	l	0,088	m
Stainless steel plate thickness	δ_p	0,0006	m

Table 1. Geometric dimensions of the plate

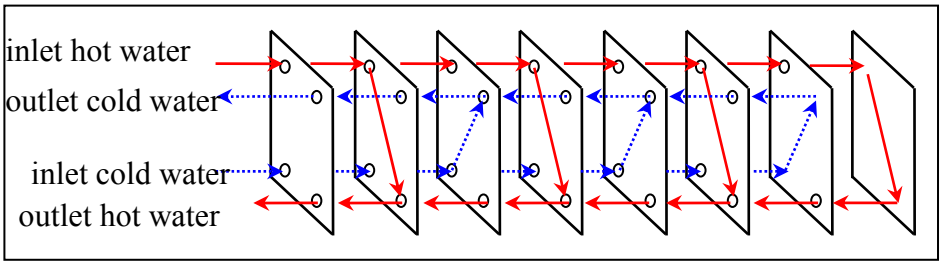


Fig. 1. Schematic presentation of counter flow plate heat exchanger with eight thermal plates

The plate heat exchanger is equipped with instrumentation for measuring pressure and temperature at the entrance and exit of thermal agents. Also it is measured the volume of hot water and cold water volume in adequate testing stand. The measured values are presented in table 2.

Description	Notation	M.U.	Value
Inlet temperature of hot water	$T_{1,in}$	°C	55
Outlet temperature of hot water	$T_{1,out}$	°C	21
Inlet temperature of cold water	$T_{2,in}$	°C	11
Outlet temperature of cold water	$T_{2,out}$	°C	19
Hot water volume	ΔV_1	m³	0,0044
Cold water volume	ΔV_2	m³	0,017
Measured time	τ	s	60

Table 2. Measured values

The volume flow rate of hot water and the volume flow rate of cold water are calculated with the next formula:

$$\dot{V}_i = \frac{\Delta V_i}{\tau}, \text{ [m}^3\text{/s]} \tag{1}$$

Where $i=1$ for hot water and $i=2$ for cold water.
The heat transfer rate from hot water is calculated with the next equation, for steady state conditions:

$$Q_1 = \dot{V}_1 \times \rho_1 \times c_{p1} \times (T_{1,in} - T_{1,out}) \text{ , [W]} \quad (2)$$

Where ρ_1 ,[kg/m^3]- density, c_{p1}, [J/kg*^{o}C] – heat capacity at constant pressure, determinate at average temperature of hot water $T_{1,m} = (T_{1,in}+T_{1,out})/2$.
The thermal flux received by cold water is given by equation (3), for steady state conditions:

$$Q_2 = \dot{V}_2 \times \rho_2 \times c_{p2} \times (T_{2,in} - T_{2,out}) \text{ ,[W]} \quad (3)$$

Where ρ_2 ,[kg/m^3]- density, c_{p2}, [J/kg*^{o}C] – heat capacity at constant pressure, determinate at average temperature of hot water $T_{2,m} = (T_{2,in}+T_{2,out})/2$.
The coefficient of heat retention for studied counter flow plate heat exchanger is defined with next relation:

$$\eta_t = \frac{Q_2}{Q_1} \text{ ,} \quad (4)$$

The determinate values are presented in table 3.

Description	M.U.	Value
$\dot{V}_1$	m^3/s	7,33*10^{-5}
$\dot{V}_2$	m^3/s	2,833*10^{-4}
Q_1	W	10347
Q_2	W	9486
η_t	-	0,917

Table 3. Determinate values from experimental results

3. Numerical simulation of the counter flow plate heat exchanger using finite element method

3D geometric model of the heat exchangers is created using SolidWorks program. Figure 2 shows the model of plate heat exchanger with eight thermal plates.
Mathematical modelling includes assignation of governing equations. The partial differential equations (pdes) governing fluid flow and heat transfer include the continuity equation, the Navier-Stokes equations and the energy equation. These equations are intimately coupled and non-linear making a general analytic solution almost impossible.
The governing equations for fluid flow and heat can be written as (Grigore&Popa, 2009):
Continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} + \frac{\partial \rho v}{\partial y} + \frac{\partial \rho w}{\partial z} = 0 \text{ ,} \quad (5)$$

x-, y-, z- momentum equations:

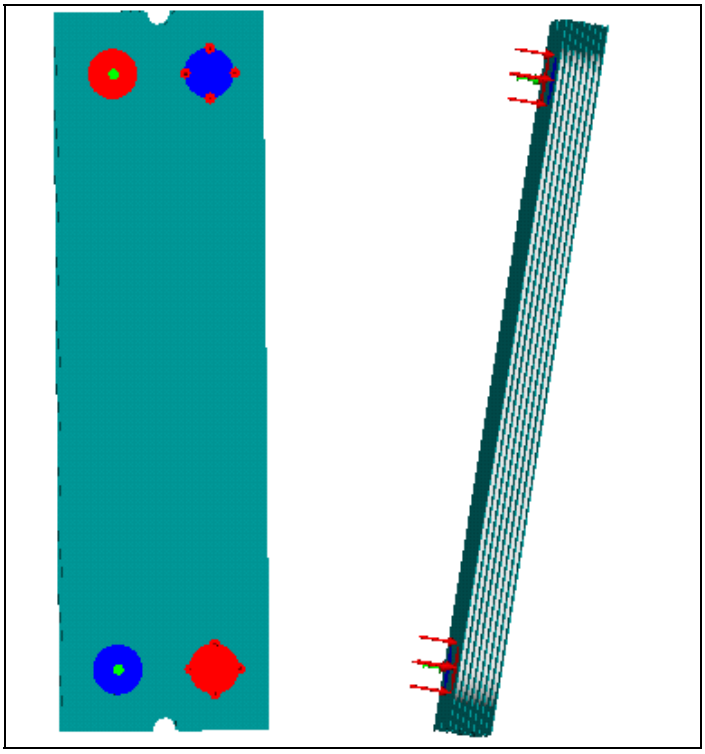


Fig. 2. 3D model of counterflow plate heat exchanger with 8 thermal plates

$$\rho\frac{\partial u}{\partial t}+\rho u\frac{\partial u}{\partial x}+\rho v\frac{\partial u}{\partial y}+\rho w\frac{\partial u}{\partial z}=$$
$$=\rho g_x-\frac{\partial p}{\partial x}+\frac{\partial}{\partial x}\left[2\eta\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[\eta\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)\right]+\frac{\partial}{\partial z}\left[\eta\left(\frac{\partial u}{\partial z}+\frac{\partial w}{\partial x}\right)\right]+S_\omega+S_{DR}, \tag{6}$$

$$\rho\frac{\partial v}{\partial t}+\rho u\frac{\partial v}{\partial x}+\rho v\frac{\partial v}{\partial y}+\rho w\frac{\partial v}{\partial z}=$$
$$=\rho g_y-\frac{\partial p}{\partial y}+\frac{\partial}{\partial y}\left[2\eta\frac{\partial v}{\partial y}\right]+\frac{\partial}{\partial x}\left[\eta\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)\right]+\frac{\partial}{\partial z}\left[\eta\left(\frac{\partial v}{\partial z}+\frac{\partial w}{\partial y}\right)\right]+S_\omega+S_{DR}, \tag{7}$$

$$\rho\frac{\partial w}{\partial t}+\rho u\frac{\partial w}{\partial x}+\rho v\frac{\partial w}{\partial y}+\rho w\frac{\partial w}{\partial z}=$$
$$=\rho g_z-\frac{\partial p}{\partial z}+\frac{\partial}{\partial z}\left[2\eta\frac{\partial w}{\partial z}\right]+\frac{\partial}{\partial x}\left[\eta\left(\frac{\partial u}{\partial z}+\frac{\partial w}{\partial x}\right)\right]+\frac{\partial}{\partial y}\left[\eta\left(\frac{\partial v}{\partial z}+\frac{\partial w}{\partial y}\right)\right]+S_\omega+S_{DR}. \tag{8}$$

The two source terms in the momentum equations, S_ω and S_{DR}, are for rotating coordinates and distributed resistances, respectively.

The distributed resistance term can be written in general as:

$$S_{DR} = -\left(K_i + \frac{f}{d} \right) \frac{\rho V_i^2}{2} - C\eta V_i \,, \tag{9}$$

Where *i* refer to the global coordinate direction (*u, v, w* momentum equation), *f*- friction factor, *d*- hydraulic diameter, *C* – permeability and the other factors are descript in table 4. Note that the *K*-factor term can operate on a single momentum equation at a time because each direction has its own unique *K*-factor. The other two resistance types operate equally on each momentum equation (Grigore&Popa, 2009), (Cosmos/Flow,2001).
The other source term is for rotating flow. This term can be written in general as:

$$S_{\omega} = -2\rho\omega_i \times V_i - \rho\omega_i \times \omega_i \times r_i \,, \tag{10}$$

Where *i* refer to the global coordinate direction, ω is the rotational speed and r is the distance from the axis of rotation.
For incompressible and subsonic compressible flow, the energy equation is written in terms of static temperature (Grigore et al., 2010):

$$\rho c_p \frac{\partial T}{\partial t} + \rho c_p u \frac{\partial T}{\partial x} + \rho c_p v \frac{\partial T}{\partial y} + \rho c_p w \frac{\partial T}{\partial z} = \frac{\partial}{\partial x}\left[U \frac{\partial T}{\partial x} \right] + \frac{\partial}{\partial y}\left[U \frac{\partial T}{\partial y} \right] + \frac{\partial}{\partial z}\left[U \frac{\partial T}{\partial z} \right] + q_V \quad ,(11)$$

The volumetric heat source term from equation (10) is considered zero for this model.
Table 4 presents the variable of the equations:

Variable	Description
c_p	specific heat at constant pressure
k	thermal conductivity
p	pressure
q_V	volumetric heat source
T	temperature
t	time
u	velocity component in x-direction
v	velocity component in y-direction
w	velocity component in z-direction
ρ	density
η	dynamic viscosity

Table 4. Variables of the governing equations

The equations describe the fluid flow and heat transfer under steady-state conditions for Cartesian geometries. For the turbulent flow, the solution of these equations would require a great deal of finite elements (on the order of $10^6 \div 10^8$) even for a simple geometry as well as near infinitesimal time steps. In this paper is used COSMOS/Flow program which solves the *time-averaged* governing equations.

The time-averaged equations are obtained by assuming that the dependent variables can be represented as a superposition of a mean value and a fluctuating value, where the fluctuation is about the mean value and a fluctuating value, where the fluctuation is about the mean. For example, the velocity component in y-direction can be written (Cosmos/Flow, 2001), (Grigore et al., 2010):

$$V = V + v', \text{ [m/s]} \tag{12}$$

where V, [m/s] – the mean velocity, v', [m/s] – the fluctuation about the mean. This representation is introduced into the governing equations and the equations themselves are averaged over time. If it uses the notation that the uppercase letters represent the mean values and lowercase letters represents fluctuating values, it can be written the governing equations (Cosmos/Flow, 2001), (Grigore et al., 2010):
Continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} + \frac{\partial \rho v}{\partial y} + \frac{\partial \rho w}{\partial z} = 0\,, \tag{13}$$

Momentum equations:

$$\rho\frac{\partial U}{\partial t} + \rho U\frac{\partial U}{\partial x} + \rho V\frac{\partial U}{\partial y} + \rho W\frac{\partial U}{\partial z} = \rho g_x - \frac{\partial P}{\partial x} + \frac{\partial}{\partial x}\left[2\eta\frac{\partial U}{\partial x} - \rho\overline{uu}\right] + \frac{\partial}{\partial y}\left[\eta\left(\frac{\partial U}{\partial y} + \frac{\partial V}{\partial x}\right) - \rho\overline{uv}\right] + \frac{\partial}{\partial z}\left[\eta\left(\frac{\partial U}{\partial z} + \frac{\partial W}{\partial x} - \rho\overline{uv}\right)\right] + S_\omega + S_{DR}\,, \tag{14}$$

$$\rho\frac{\partial V}{\partial t} + \rho U\frac{\partial V}{\partial x} + \rho V\frac{\partial V}{\partial y} + \rho W\frac{\partial V}{\partial z} = \rho g_y - \frac{\partial P}{\partial y} + \frac{\partial}{\partial y}\left[2\eta\frac{\partial V}{\partial y} - \rho\overline{vv}\right] + \frac{\partial}{\partial x}\left[\eta\left(\frac{\partial U}{\partial y} + \frac{\partial V}{\partial x} - \rho\overline{uv}\right)\right] + \frac{\partial}{\partial z}\left[\eta\left(\frac{\partial V}{\partial z} + \frac{\partial W}{\partial y} - \rho\overline{vw}\right)\right] + S_\omega + S_{DR} \tag{15}$$

$$\rho\frac{\partial W}{\partial t} + \rho U\frac{\partial W}{\partial x} + \rho V\frac{\partial W}{\partial y} + \rho W\frac{\partial W}{\partial z} = \rho g_z - \frac{\partial P}{\partial z} + \frac{\partial}{\partial z}\left[2\eta\frac{\partial W}{\partial z} - \rho\overline{ww}\right] + \frac{\partial}{\partial x}\left[\eta\left(\frac{\partial U}{\partial z} + \frac{\partial W}{\partial x} - \rho\overline{uw}\right)\right] + \frac{\partial}{\partial y}\left[\eta\left(\frac{\partial V}{\partial z} + \frac{\partial W}{\partial y} - \rho\overline{vw}\right)\right] + S_\omega + S_{DR} \tag{16}$$

Energy equation:

$$\rho c_p\frac{\partial T}{\partial t} + \rho c_p U\frac{\partial T}{\partial x} + \rho c_p V\frac{\partial T}{\partial y} + \rho c_p W\frac{\partial T}{\partial z} = \frac{\partial}{\partial x}\left[k\frac{\partial T}{\partial x} - \rho c_p\overline{uT'}\right] + \frac{\partial}{\partial y}\left[k\frac{\partial T}{\partial y} - \rho c_p\overline{uT'}\right] + \frac{\partial}{\partial z}\left[k\frac{\partial T}{\partial z} - \rho c_p\overline{wT'}\right] + q_V \tag{17}$$

The averaging process has produced extra terms in the momentum and energy equations. For turbulent flow, equations of continuity, of momentum and energy is a system of 5 equations with 14 unknowns (Cosmos/Flow,2001). To solve, it is used Boussinesq approximation which defines an eddy viscosity and eddy conductivity:

$$\eta_t = \frac{-\rho\overline{uu}}{2\dfrac{\partial U}{\partial x}} = \frac{-\rho\overline{uv}}{\dfrac{\partial U}{\partial y} + \dfrac{\partial V}{\partial x}} = \frac{-\rho\overline{vw}}{\dfrac{\partial V}{\partial z} + \dfrac{\partial W}{\partial y}} = \ldots\,, \tag{18}$$

$$k_t = \frac{-\rho c_p\overline{ut}}{\dfrac{\partial T}{\partial x}} = \frac{-\rho c_p\overline{vt}}{\dfrac{\partial T}{\partial y}} = \frac{-\rho c_p\overline{wt}}{\dfrac{\partial T}{\partial z}}\,. \tag{19}$$

These terms imply that the effect of turbulence is isotropic. With these approximations the governing equations become:

$$\frac{\partial \rho}{\partial t}+\frac{\partial \rho u}{\partial x}+\frac{\partial \rho v}{\partial y}+\frac{\partial \rho w}{\partial z}=0 \tag{20}$$

$$\rho\frac{\partial U}{\partial t}+\rho U\frac{\partial U}{\partial x}+\rho V\frac{\partial U}{\partial y}+\rho W\frac{\partial U}{\partial z}=\rho g_x-\frac{\partial P}{\partial x}+\frac{\partial}{\partial x}\left[2(\eta+\eta_t)\frac{\partial U}{\partial x}\right]+\frac{\partial}{\partial y}\left[(\eta+\eta_t)\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)\right]+\frac{\partial}{\partial z}\left[(\eta+\eta_t)\left(\frac{\partial U}{\partial z}+\frac{\partial W}{\partial x}\right)\right]+S_\omega+S_{DR} \tag{21}$$

$$\rho\frac{\partial V}{\partial t}+\rho U\frac{\partial V}{\partial x}+\rho V\frac{\partial V}{\partial y}+\rho W\frac{\partial V}{\partial z}=\rho g_y-\frac{\partial P}{\partial y}+\frac{\partial}{\partial y}\left[2(\eta+\eta_t)\frac{\partial V}{\partial y}\right]+\frac{\partial}{\partial x}\left[(\eta+\eta_t)\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)\right]+\frac{\partial}{\partial z}\left[(\eta+\eta_t)\left(\frac{\partial V}{\partial z}+\frac{\partial W}{\partial y}\right)\right]+S_\omega+S_{DR} \tag{22}$$

$$\rho\frac{\partial W}{\partial t}+\rho U\frac{\partial W}{\partial x}+\rho V\frac{\partial W}{\partial y}+\rho W\frac{\partial W}{\partial z}=\rho g_z-\frac{\partial P}{\partial z}+\frac{\partial}{\partial z}\left[2(\eta+\eta_t)\frac{\partial W}{\partial z}\right]+\frac{\partial}{\partial x}\left[(\eta+\eta_t)\left(\frac{\partial U}{\partial z}+\frac{\partial W}{\partial x}\right)\right]+\frac{\partial}{\partial y}\left[(\eta+\eta_t)\left(\frac{\partial V}{\partial z}+\frac{\partial W}{\partial y}\right)\right]+S_\omega+S_{DR} \tag{23}$$

$$\rho c_p\frac{\partial T}{\partial t}+\rho c_p U\frac{\partial T}{\partial x}+\rho c_p V\frac{\partial T}{\partial y}+\rho c_p W\frac{\partial T}{\partial z}=\frac{\partial}{\partial x}\left[(k+k_t)\frac{\partial T}{\partial x}\right]+\frac{\partial}{\partial y}\left[(k+k_t)\frac{\partial T}{\partial y}\right]+\frac{\partial}{\partial z}\left[(k+k_t)\frac{\partial T}{\partial z}\right]+q_V \tag{24}$$

In these conditions should be determined in addition to the 5 unknowns only k_t and η_t. The program COSMOS Flow uses a model with two equations for their determination (Cosmos/Flow,2001).

$$\eta_t=\frac{C_\eta \rho K^2}{\varepsilon}, \tag{25}$$

$$k_t=\frac{\eta_t c_p}{\mathrm{Pr}_t}, \tag{26}$$

Where Pr_t - turbulent Prandtl number and C_η – empirical constant. The transport equations for K and ε are derived using momentum equations:

$$\begin{aligned}&\rho\frac{\partial K}{\partial t}+\rho U\frac{\partial K}{\partial x}+\rho V\frac{\partial K}{\partial y}+\rho W\frac{\partial K}{\partial z}=\frac{\partial}{\partial z}\left[\left(\frac{\eta_T}{\sigma_k}+\eta\right)\frac{\partial K}{\partial z}\right]+\frac{\partial}{\partial x}\left[(\eta+\frac{\eta_T}{\sigma_k})\left(\frac{\partial K}{\partial x}\right)\right]+\\&\frac{\partial}{\partial y}\left[(\eta+\frac{\eta_T}{\sigma_k})\left(\frac{\partial K}{\partial y}\right)\right]-\rho\varepsilon+\eta_T\left[2\left(\frac{\partial U}{\partial x}\right)^2+2\left(\frac{\partial V}{\partial y}\right)^2+2\left(\frac{\partial W}{\partial z}\right)^2+\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)^2+\left(\frac{\partial U}{\partial z}+\frac{\partial W}{\partial x}\right)^2+\left(\frac{\partial V}{\partial z}+\frac{\partial W}{\partial y}\right)^2\right]\end{aligned} \tag{27}$$

$$\begin{aligned}&\rho\frac{\partial \varepsilon}{\partial t}+\rho U\frac{\partial \varepsilon}{\partial x}+\rho V\frac{\partial \varepsilon}{\partial y}+\rho W\frac{\partial \varepsilon}{\partial z}=\frac{\partial}{\partial z}\left[\left(\frac{\eta_T}{\sigma_\varepsilon}+\eta\right)\frac{\partial \varepsilon}{\partial z}\right]+\frac{\partial}{\partial x}\left[(\eta+\frac{\eta_T}{\sigma_\varepsilon})\left(\frac{\partial \varepsilon}{\partial x}\right)\right]+\\&\frac{\partial}{\partial y}\left[(\eta+\frac{\eta_T}{\sigma_\varepsilon})\left(\frac{\partial \varepsilon}{\partial y}\right)\right]-\rho C_2\frac{\varepsilon^2}{K}+C_1\eta_t\frac{\varepsilon}{K}\left[\begin{array}{l}2\left(\frac{\partial U}{\partial x}\right)^2+2\left(\frac{\partial V}{\partial y}\right)^2+2\left(\frac{\partial W}{\partial z}\right)^2+\\+\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)^2+\left(\frac{\partial U}{\partial z}+\frac{\partial W}{\partial x}\right)^2+\left(\frac{\partial V}{\partial z}+\frac{\partial W}{\partial y}\right)^2\end{array}\right]\end{aligned} \tag{28}$$

Where σ_K and σ_ε – turbulent Schmidt numbers, C_1, C_2 – empirical constants. With these 2 equations, there are 9 equations in 9 unknowns: U,V,W, P,T, μ_t, k_t, K,ε. Table 5 presents the constants associated to model (Cosmos/Flow,2001).

Constant	Value
C_η	0.09
C_1	1.44
C_2	1.92
σ_k	1.0
σ_ε	1.3

Table 5. Used constants in model

Finite element method is used to *discretize* the flow domain, thereby transforming the governing partial differential equations into a set of algebraic equations whose solution represents an approximation to the exact analytical solution (Awrejcewicz &Krysko, 2003), (Grigore et al., 2010), (Andrianov et al., 2004). A set of simplified hypothesis are introduced:

- Hot water and cold water are Newtonian fluids;
- No phase change occurs, the fluids are unmixed;
- Turbulent flow is fully developed;
- Working fluids are incompressible,
- Steady state conditions;
- Coefficient of heat retention equal with 1.

It is applied the Streamline Upwind Petrov Galerkin (SUPG) method. The method is used directly on the diffusion and source terms and for the advection terms, the streamline upwind method is used with the weighted integral method. These terms are transformed to stream-wise coordinates, like in next expression:

$$\rho U \frac{\partial \phi}{\partial x} + \rho V \frac{\partial \phi}{\partial y} + \rho W \frac{\partial \phi}{\partial z} = \rho U_S \frac{\partial \phi}{\partial S}, \tag{29}$$

Where s – stream-wise coordinate, U_s – the velocity component in the stream-wise coordinate direction, Φ- transported quantity.

In figure 3 is shown the used analysis scheme.

The disparagement mode (mesh) , is very important for the final results. The models are divided in a multitude of little parts with simple geometrical forms, defined as finite parts, and connected in common points called nods, like in figure 4.

The quality of mesh is high (10-node tetrahedral), mesh type is solid mesh, element size – 5,3588 mm, 365377 nodes and 244363 elements.

Boundary conditions from table 6 are proposed.

For incompressible flows, the most robust condition for the pressure equation is to specify a value at the outlet. Since only relative pressures are calculated by COSMOS/Flow, a value of 0 is recommended (Cosmos/Flow,2001). The numerical simulation of turbulent flow is modeled by k- ε turbulence model. ε represents the turbulent energy dissipation. To

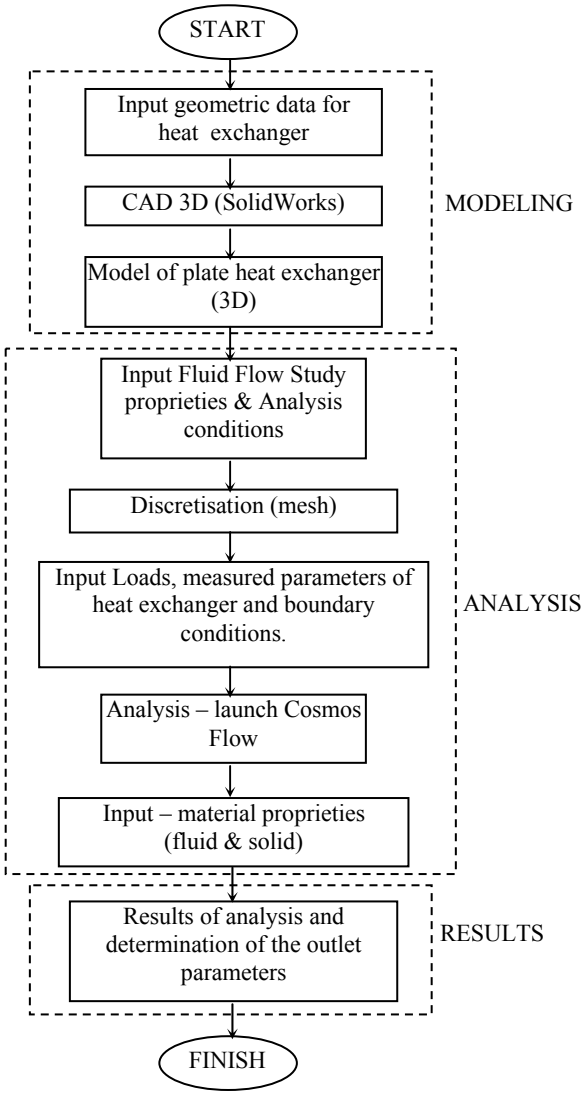


Fig. 3. Analysis scheme

calculate the boundary layer, either "wall functions" are used, overriding the calculation of k and ε in the wall adjacent nodes, or integration is performed to the surface, using a "low turbulent Reynolds (*low-Re*) k-ε" model (Grigore et. al., 2010).

After the analysis was processed it can be visualized the results, under graphical form or numerical value. Because the governing equations are non-linear, they must be solved iteratively. A Picard or successive substitution is used. In this method estimates of the solution variables are substituted in the governing equations. The equations are solved for new values which are the used as the estimates for the next pass. The convergence criterion

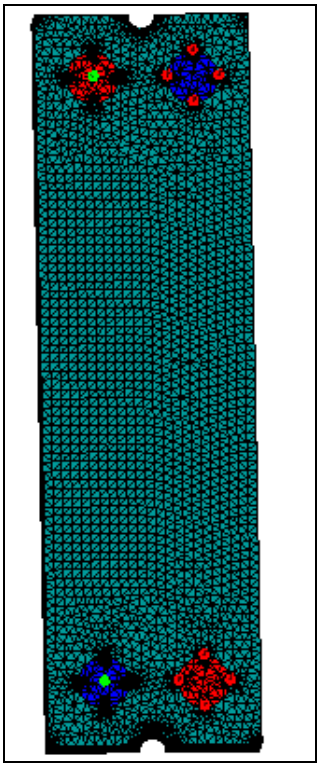


Fig. 4. Mesh

No.	Fluid	Description	unit	Value
1	Hot Water inlet	Temperature – $T_{1,in}$	^{o}C	55
2		Volume flow rate - $\dot{V}_1$	m^3/s	7,33*10^{-5}
4	Cold Water inlet	Temperature – $T_{2,in}$	^{o}C	11
5		Volume flow rate – $\dot{V}_2$	m^3/s	2,833*10^{-4}
6	Hot water outlet	Static pressure	N/m^2	0
7	Cold water outlet	Static pressure	N/m^2	0

Table 6. Boundary conditions

is the level at which the specified variable's residual norm must reach. With each pass, the residuals should become smaller if the solution is converging. The global iterations is shown below:

1. Solve x momentum equation;
2. Solve y momentum equation;

3. Solve z momentum equation;
4. Solve pressure equation and velocities;
5. Solve energy equation;
6. Solve turbulent kinetic energy equation;
7. Solve turbulent energy dissipation equation;
8. Check convergenge (GOTO1)

Analysis runs for 100 iterations, in turbulence conditions, for all eight cases. Profiles are obtained for the following parameters: u, v, w, T, k, ε. In figure 5 is shown the distribution of the nodal temperature, after 50 iterations.

The hot water temperature and the cold water temperature vary along their flow path, even in the case of constant thermal resistance, because of the flow distribution and temperature gradient variations across the plates.

Convergence control of a solution variable is accomplished by reduction the solution progression rate so that the change of divergence is minimized. COSMOS/Flow has the Graphical Convergence Monitor, where are presented the numerical data like the average, the average, minimum, maximum values for each degree of freedom over the completed range of iterations(Grigore et al.,2010).

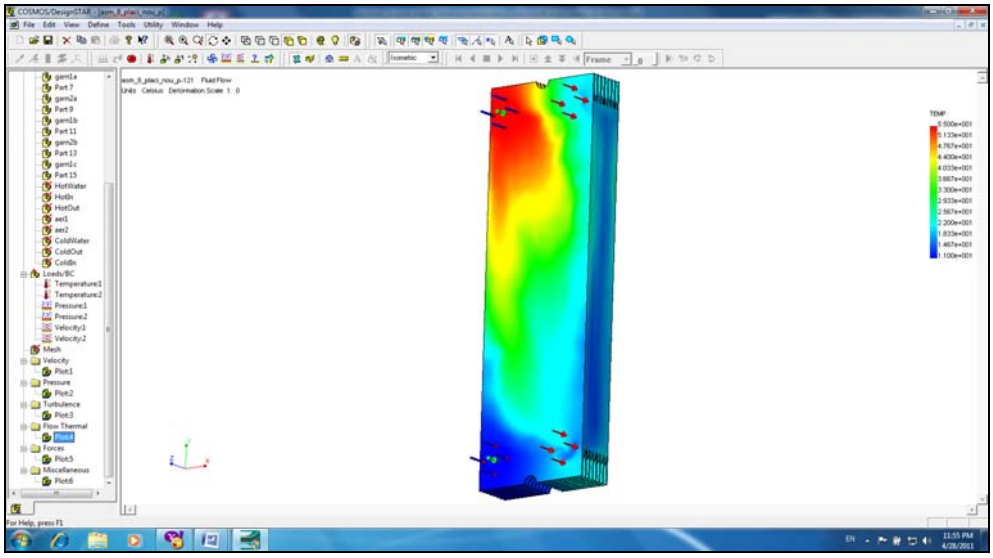


Fig. 5. Distribution of nodal temperature

No.	Fluid	Description	M.U.	Value
1	Hot water outlet	Temperature $T_{1,OUT}$	^{o}C	22,32
2	Cold water outlet	Temperature $T_{2,out}$	^{o}C	17,90

Table 7. Obtained values for average temperatures at the outlet

The convergence of the nodal temperature is shown in figure 6.
The values obtained for average temperature at the outlet of hot water and outlet of cold water are presented in table 7.

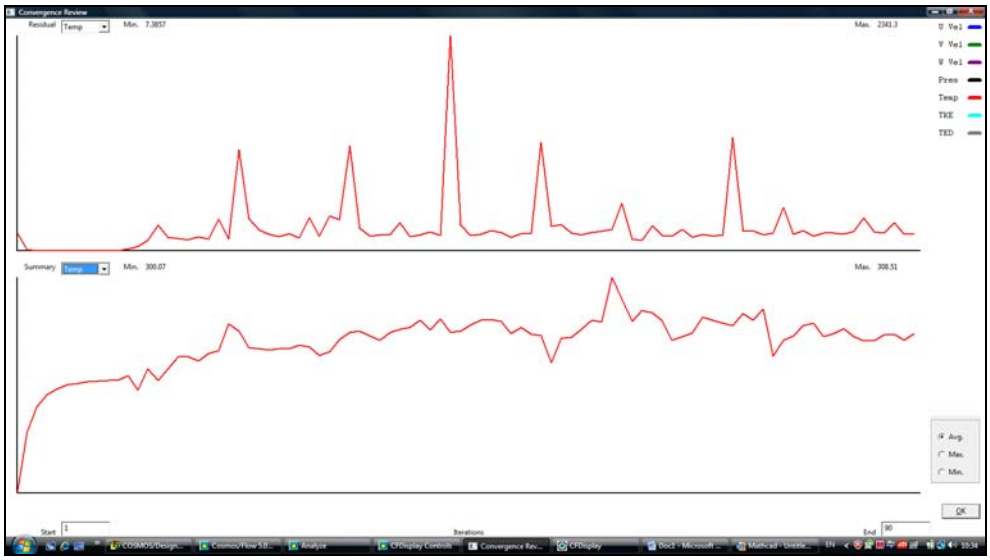


Fig. 6. Convergence of the nodal temperature after 90 iterations

4. Theoretical analysis of studied plate heat exchanger

The theoretic analysis is based on the equation of heat balance for thermal agents and plate heat transfer equation. The total rate of heat transfer between the hot and cold fluids passing through a plate heat exchanger may be expressed as:

$$Q = U \times S \times LMTD \text{ , [W]} \tag{30}$$

Where U, [W/m²K] - the overall heat transfer coefficient, $LMTD$, [K] - the log mean temperature difference in K. U is dependent upon the heat transfer coefficients in the hot and cold streams. $LMTD$ is computed under assumption of counter flow condition with next relation (Badea et al., 2003):

$$LMTD = \frac{\Delta T_{\max} - \Delta T_{\min}}{\ln\left(\frac{\Delta T_{\max}}{\Delta T_{\min}}\right)} \text{ ,[K]} \tag{31}$$

Where ΔT_{max} = max(ΔT1, ΔT2), ΔT_{min} =min (ΔT1, ΔT2); ΔT1, ΔT2 from figure 7. Figure 7 shows the hot and cold fluid temperature distributions in the counter flow heat exchanger. The heat transfer surface aria S is represented along the x-axis and the fluid stream temperature along the y-axis.
The boundary conditions are the same like in experimental case and analyze with finite element method. The following simplifications were considered:

- steady-state conditions;
- no leakage flow;
- no phase change;
- physical proprieties are constant in the plate heat exchanger;
- uniform temperature and uniform fluid distribution;
- no heat losses to the surrounding;
- efficiency of counter flow plate heat exchanger is considered 1.

The calculation is iterative and has certain features related to channel geometry and in the same time depends on the flow regimes and criterial relations for convection heat transfer coefficients. Next are presented the principal steps:

1. Approximation of the average temperatures of the thermal agents Tm_1 and Tm_2 and approximation of plates temperatures $Tp = (Tm_1 + Tm_2/2$.

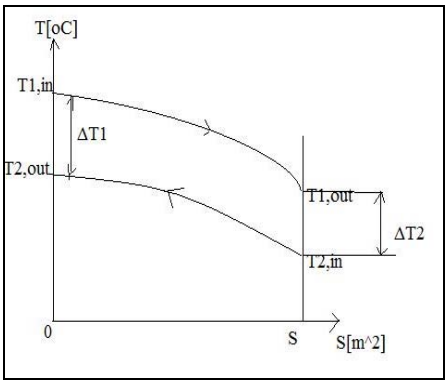


Fig. 7. The hot and cold fluid temperature distributions in the counter flow heat exchanger

2. Determination of hydraulic diameter characteristic for studied plate heat exchanger configuration, with next formula:

$$d_h = \frac{4 \times l \times H_o}{2 \times (l + H_o)}, \tag{32}$$

3. Determination of average velocity under channels for both thermal agents:

$$w_i = \frac{\dot{V}_i}{N_c \times H_o \times l}, [\mathrm{m/s}] \tag{33}$$

Where $i=1$ for hot water and $i=2$ for cold water.

4. Calculation of Reynolds numbers.

$$\mathrm{Re}_i = \frac{w_i \times d_h}{\nu_i}, \tag{34}$$

Where $i=1$ for hot water and $i=2$ for cold water, υ_i [$\mathrm{m/s^2}$] – cinematic viscosity.

5. For turbulent regimes, with next formula is calculated convection heat transfer coefficient, for each thermal agents(hot water and cold water)(Facultatea de Energetica, 2010):

$$\alpha_i = 63 \times \mathrm{Re}_i^{0,61} \times \mathrm{Pr}_i^{0,4} \times \lambda_i \times \left(\frac{\eta_i}{\eta_p} \right)^{0,14} \tag{35}$$

Where *i=1* for hot water and *i=2* for cold water, η_i [m/s²] – dynamic viscosity, λ_i [W/mK] – thermal conductivity, Pr- Prandtl number.

6. It calculates temperature of the plates:

$$Tp_i = \frac{Tm_1 \times \alpha_1 + Tm_2 \times \alpha_2}{\alpha_1 + \alpha_2}, \text{ [°C]} \tag{36}$$

And the error $\varepsilon = \left| \frac{Tp - Tp_i}{Tp} \times 100 \right|$.

If ε >2%, it goes to point 1 and the calculus begin again. If ε <2% is still calculating.

7. The overall heat transfer coefficient is determinated:

$$U = \frac{1}{\frac{1}{\alpha_1} + \frac{\delta_p}{\lambda_p} + \frac{1}{\alpha_2}}, \text{ [W/m²K]} \tag{37}$$

8. The maximum number of transfer units:

$$NTU_{\max} = \frac{U \times S}{W \min}, \tag{38}$$

Where $W_{min} = \min(W_1, W_2)$,[W/K] – minimum heat capacity rate.

$$W_1 = \dot{V}_1 \times c_{p,1} \times \rho_1, \text{ [W/K]} \tag{39}$$

$$W_2 = \dot{V}_2 \times c_{p,2} \times \rho_2, \text{ [W/K]} \tag{40}$$

9. Heat transfer effectiveness is definite like actual heat transfer divided by the maximum possible heat transfer.

$$\varepsilon = \frac{1 - e^{-NTU_{\max} \times \left(1 - \frac{W_{\min}}{W_{\max}}\right)}}{1 - \frac{W_{\min}}{W_{\max}} \times e^{-NTU_{\max} \times \left(1 - \frac{W_{\min}}{W_{\max}}\right)}}, \tag{41}$$

Where $W_{max} = \max(W_{min}, W_{min})$.

10. The heat exchanger duty is:

$$Q = \varepsilon \times W_{\min} \times (T_{1,in} - T_{2,in}) \tag{42}$$

11. The outlet temperature of the hot water and the outlet temperature of the cold water:

$$T_1, out = T_{1,in} - \frac{Q}{W_1}, [^oC] \tag{43}$$

$$T_2, out = T_{2,in} + \frac{Q}{W_2}, [^oC] \tag{44}$$

Obtained results are: $T_{1,out}$ = 22,36 ^{o}C and $T_{2,out}$ = 19,38 ^{o}C.

4. Conclusions

The paper presents a simplified model for a plate heat exchanger in a counter flow arrangement. It is realized a numerical simulation and it is observed that the model is in concordance with the experimental results and with the results from theoretical analysis.
Numerical simulation of plate heat exchanger using finite element method is very representative, although it is very laborious and consume more resources from a computer (the geometrical model is much complex, the simulation is more difficult), the results are well presented visual, graphic and numerical.

Value	M.U.	Experimental results	Theoretical analysis	Numerical simulation
$T_{1,out}$	^{o}C	21	22,36	22,32
$T_{2,out}$	^{o}C	19	19,38	17,9

Table 7. Results for outlet temperatures

There are small differences between results. The differences appear due to the simplifying assumptions considered and due to presence of fouling on the surface of the plates. Also a relative degree of uncertainty is introduced by the criterial relations used to calculate convection heat transfer coefficients. In the same time, the plate heat exchanger has corrugated plates patern. Numerical simulation cannot reflect the influence of the corrugation angle and corrugation height, but it offers a good understanding of the temperature distribution and fluid flow under turbulent motion.

5. References

Andrianov, I.V.; Awrejcewicz, J.& Manevitch, L.I.(2004). *Asymptotical Mechanics of Thin Walled Structures.A Handbook*, Springer-Verlag, ISBN: 3-54087602,Berlin,Germany

Awrejcewicz, J. & Krisko V.A.(2004). *Nonclassical Thermoelastic Problems in Nonlinear Dynamics of Shells*, Springer_Verlag, ISBN: 3-540-43880-7, Berlin, Germany

Awrejcewicz, J.; Krysko, V.A. & Krisko A.V. (2007). *Thermodynamics of Plates and Shells,* Springer-Verlag, ISBN: 9783540342618,Berlin,Germany

Badea, A; Necula, H; Stan, M.; Ionescu, L; Blaga, P. & Darie, G.(2003).*Echipamente şi instalaţii termice,* Editura Tehnică, ISBN: 973-31-2183-5, Bucureşti, Romania

COSMOS/Flow(2001) - Technical Reference

Grigore, R.& Popa, S.(2009). Modeling a Counter-flow Plate Heat Exchanger, *Proceedings of 4TH International Conference on Energy and Environment*, CIEM 2009, ISSN: 1454-23xx,Bucureşti, Romania

Grigore,R; Popa ,S; Hazi,A &Hazi, G.(2010). Study Regarding the Influence of the Plate Heat Exchanger Configuration on Its Performance, *WSEAS Transactions on heat and Mass Transfer,* Issue 3, vol.5, pp.133-142, ISSN: 1790-5079

Gut,J.A.W.& Pinto, J. M.(2003). Modeling of plate heat exchangers with generalized configurations, *International Journal of Heat and Mass Transfer*, , pp. 2571-2585, ISSN:0017-9310.

Facultatea de Energetica, Indrumar de schimbatoare de caldura, Bucuresti, 28.03.2010, Available from http://insttermind.3x.ro/Index_files/Iti_indrumar.pdf

Saber, M.H.&Mazaher Attain,H.(2010) Simulation and CFD Analysis of heat pipe exchanger using Fluent to increase of the thermal efficiency, *Proceedings of 5th IASME/WSEAS International Conference on Continum Mechanics*, ISBN: 978-960-474-158-8, University of Cambridge, UK,February 23-25, pp.184-189

17

Numerical Modelling and Simulation of Radial-Axial Ring Rolling Process

Lianggang Guo and He Yang
State Key Laboratory of Solidification Processing, School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an, PR China

1. Introduction

Ring rolling has been a kind of irreplaceable near-net-shape metal forming technology for the manufacture of various ring-shaped parts with high performance and high precision, such as various bearing races, ring gears, aero-engine casing, nuclear reactors parts and various connecting flanges, due to the most important advantages of the favourable grain flow and good surface quality of the rolled rings (Eruc & Shivpuri, 1992). Radial-axial ring rolling, the forming principle of which is shown in Fig.1, is a classic form of ring rolling process and is usually adopted to manufacture various high-quality large rings widely served in many important industry areas such as aerospace and wind power. During the process, the main roll rotates at a rotational speed n_1; the mandrel squeezes the ring wall at a feed rate v_f and runs idle because of the friction on the contact surface; the axial rolls, including the upper and lower conical rolls, are driven to rotate at an inverse speed n_a around their axes and to withdraw at a speed v_w with the increasing of the ring diameter to maintain a minimum relative slip between the axial rolls and the end faces of the ring; at the same time, the upper conical roll slides toward the lower conical roll at a feed rate v_a to cause axial height reduction of the ring, while the lower conical roll is held in a fixed position above the table plate of the radial-axial mill; the guide rolls contact the ring outer diameter to ensure the circularity of the ring, and any force imbalance and instability during the rolling process are removed by the actions of the guide rolls. Under the cooperative actions of all the rolls, the ring blank rotates and produces plastic deformation of reduction in cross-section and growth in diameter.

From the forming principle of radial-axial ring rolling, it can be known that the process is characterized by extremely complicated dynamic forming and high flexibility due to multiple independent control system for the three sets of rolls, namely, the radial main roll and mandrel, two axial conical rolls and two guide rolls, as illustrated in Fig. 1. What's more, in consideration of microstructure evolution of ring materials and the final performance of the rolled ring, the thermal-coupled plastic deformation behavior of the process and response of ring materials properties to the process are necessary and significant concerns for the design, operation and optimization of the process. In practice, the radial-axial ring rolling process usually presents uncooperative motions of the rolls, severe instabilities thus usually rolls rings with various macro and micro defects due to unreasonable design of process parameters (mainly including the sizes of ring blank,

forming temperature and various motion parameters such as v_f, n_1, v_a, n_a, and v_w, as show in Fig. 1) and severe dynamic contacts and collisions between the ring and the rolls. Thus, the optimal design and precise control of the actual process and the quality control of the rolled rings are faced with huge challenges and difficulties.

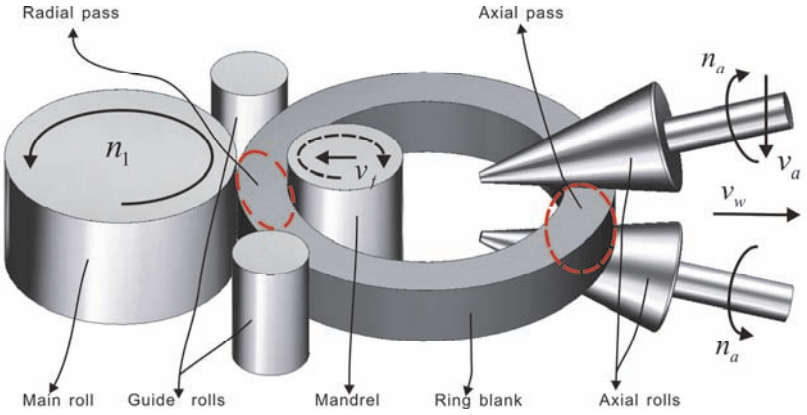


Fig. 1. Forming principle of radial-axial ring rolling

Due to the above concerns, the key science problem, which must be solved for the R&D of radial-axial ring rolling technology, is the evolution mechanism of geometry and microstructure of ring blank during the process. This is because the final precision of shape and size of the rolled ring depends on the evolution mechanism of the geometry of the ring blank, and the final performance of the rolled ring depends on the evolution mechanism of the microstructure of the ring blank. Solving the above science problem is the theoretical basis for the optimal design and steady control of the process. However, it is difficult to solve the above science problem only by experiment or theory analysis or numerical simulation because of the complexity of the process.

FE numerical modeling and simulation has been proven to be a powerful and accurate method to study plastic deformation behavior of various complicated metal forming processes (Yang et al. 2004), and can conduct more comprehensive, profound, and detailed investigations compared with the analytical and experimental methods. So, FE numerical modeling and simulation combined with theoretical analysis and experimental observation has been a powerful means for many metal forming processes such as forging, extrusion, spinning, tube bending and various rolling technologies including flat rolling, ring rolling, and so on.

So far, the ring rolling technology has evolved over 150 years. Allwood et al. (2005) reviewed the contributions of 174 papers by a thorough survey of work on ring rolling published in the English and Germany languages by 2004. Many studies on the ring rolling technology have been carried out by many researchers through experiment, theoretical analysis or numerical simulation. Johnson et al. (1968) carried out earlier experimental works on ring rolling. Hawkyard et al. (1973) reported a theoretical prediction for the roll force and torque during ring rolling between plain cylindrical rolls, and examined the prediction accuracy by experimental measurements of roll force and torque. Mamalis et al. (1975) investigated the cavity formation by rolling profiled (T-shaped) rings experimentally.

Lugora et al. (1987) analyzed the spread in plain ring rolling using Hill's general method of analysis. Hahn et al. (1994) reported the UBET analysis of the closed-pass ring rolling of rings having arbitrarily shaped profiles.

With the rapid development of numerical calculation and computer technologies, FE numerical simulation technology was widely employed to investigate deformation mechanics of ring rolling. Kim et al. (1990) reported a finite element code 'RING' which was developed for the three-dimensional deformation analysis of ring rolling. Yang et al. (1991) simulated the T-section profile ring rolling processes by rigid-plastic finite element method. Kang & Kobayashi (1991) and Joun et al. (1998) carried out the studies on preform design in ring rolling using the backward tracing scheme and an axisymmetric forging approach, respectively. Davey & Ward (2003) presented an ALE approach for finite element ring-rolling simulation to save computational cost. Based on the finite element method, Forouzan et al. (2003) proposed a new method (thermal spokes) to simulate the guide roll effect in FE analysis of the ring rolling process.

In recent years, many advances in ring rolling were obtained mainly by FE numerical simulations. For example, Wang et al. (2007) developed a virtual radial-axial ring rolling process for guiding process design and optimization using the LS-DYNA FE code; Jong et al. (2007) investigated the radial-axial ring-rolling design of a large-scale ring product of Ti-6Al-4V alloy using a calculation method and FEM analysis; Moon et al. (2008) predicted the polygonal-shaped defects during hot ring rolling using a rigid-viscoplastic finite element method; Hua et al. (2009) established a ring stiffness condition in radial-axial ring rolling whose validity was evaluated by numerical simulation; Zhou et al. (2010, 2011) numerically analyzed the coupled thermo-mechanical behaviors in radial-axial rolling of alloy steel large ring and revealed the effects of roll size on the process, and so on.

Facing with national needs in aerospace area, our research team, namely lab of precision plastic forming (LPPF) led by Professor He Yang in Northwestern Polytechnical University of China, has continuously engaged in investigating various complicated and special metal forming processes, such as NC tube bending (Yang et al., 2010; Zhan et al., 2006; Li et al., 2010), precision die forging and local loading forming (Yang et al. 2002; Liu et al., 2002; Fan et al., 2010), spinning (Yang et al., 2010; Zhan et al., 2007) and ring rolling (Guo et al., 2005; Yang et al., 2008; Guo & Yang, 2011). In the aspect of ring rolling, we achieved many important findings in past several years. For example, Yang et al. (2008) investigated the effects of blank size on strain and temperature distribution during hot rolling of titanium alloy large rings by 3D coupled thermo-mechanical FE simulations; Guo et al. (2005) revealed the plastic deformation behavior in cold ring rolling by FE simulation and proposed three kinds of plastic deformation behaviors of pure radial ring rolling process. Guo & Yang (2011) developed mathematical model of a steady forming condition for radial-axial ring rolling and demonstrated its validation by FE simulations; Li et al. (2008) reported a control method of guide rolls in 3D-FE simulation of ring rolling; Wang et al. (2009) revealed the coupled mechanical and thermal behaviours in hot rolling of large rings of titanium alloy using 3D dynamic explicit FEM, and so on. The relevant studies on the above metal forming technologies have attained supports of many national major research programs such as the Natural Science Foundation of China, National Basic Research Program of China (“973” Program), National Major Science and Technology Special Projects of China and National High Technology Research and Development Program of China (“ 863“ Program).

In this book chapter, we first propose a high-end research route for aerospace plasticity technology in terms of our understanding and research experiences on various metal

forming processes and give an application example of it for the investigation of radial-axial ring rolling technology, then discuss the involved key FE modelling technologies and reliability of the developed thermo-mechanical coupled 3D-FE model for the entire radial-axial ring rolling process, next report some simulation results including ring geometry evolution, stress field, strain field, temperature field, rolling forces and torques in the radial and axial directions during the process, afterwards summarize the conclusion and future work, and finally express our acknowledgement for the support given to this work.

2. High-end research route for aerospace plasticity technology

With the rapid development of aerospace industry, various key aerospace components need to be manufactured by plasticity technology for the fabrication of high-end aerospace equipments. But due to the severe service environment in aerospace area, manufacturing various key components, which have features of light weight, high precision, high performance, high reliability and high efficiency, has been the eternal goal for the R&D of advanced technology of plasticity. Therefore, the R&D of the aerospace plasticity technology is facing with huge challenges as discussed below.

2.1 Challenges for aerospace plasticity technology

In consideration of the requirements of light weight, high precision, high performance, high reliability and high efficiency for the key aerospace components, the main challenges for the R&D of aerospace plasticity technology are concisely summarized as follows.

1. Various difficult-to-deform and expensive materials such as titanium alloy, which have high strength, poor ductility, low elastic modulus and complicated microstructure but possess low density, good corrosion resistance and high temperature resistance, are exmployed to manufacture the key aerospace components, thus leading to difficulties for the R&D of aerospace plasticity technology.
2. The geometry structures of the key aerospace components become larger, thinner, more integral and more complex, which result in complicated preform design, ultrahigh forming load, inhomogeneous microstructure and performance, loose geometry tolerance and complicated dies for their plastic forming thus leads to huge challenges to the aerospace plasticity technology.
3. The plastic forming process of the key aerospace components is influenced interactively by multi-factors including material parameters (initial microstructure and various physical or chemical properties), process parameters (temperature, strain rate and degree of deformation, etc), geometry parameters (blank or preform size, product size and die size), so is highly nonlinear due to its material, boundary and geometry nonlinearity. The coupled effects of multi-factors and high nonlinearity lead to huge challenges to the aerospace plasticity technology.

Therefore, it is essential and urgent to find certain advanced methodology for the R&D of the aerospace plasticity technology. In terms of our understanding and research experiences on various metal forming technologies, we propose the following high-end research route for the aerospace plasticity technology.

2.2 Proposing of high-end research route for aerospace plasticity technology

In consideration of challenges for plasticity technology in aerospace area, a high-end research route for aerospace plasticity technology is proposed as illustrated in Fig. 2.

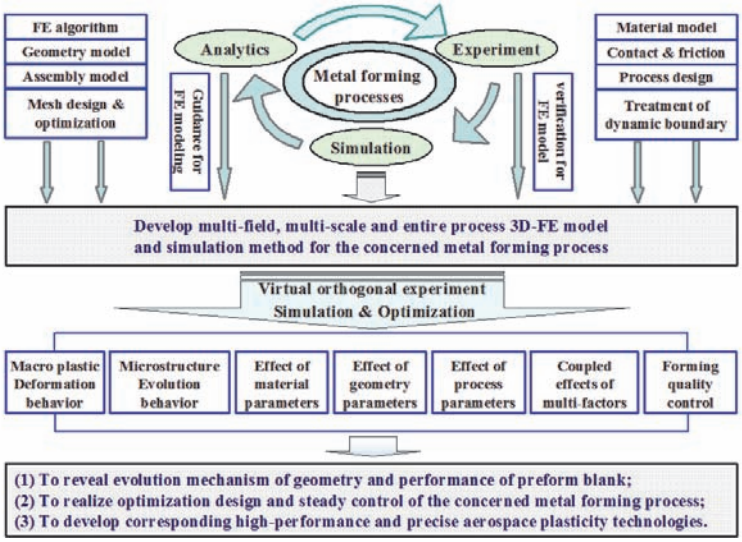


Fig. 2. High-end research route for aerospace plasticity technology

Due to the complexity of metal forming processes, FE numerical simulation organically combined with analytics and experiment is employed to investigate the forming mechanism of various metal forming processes. The relationships among the analytics, experiment and simulation are concisely discussed as follows.

The analytics is used to initially design the process parameters and provide basic understanding of the concerned metal forming process thus provides guidance and basis for FE modelling of the process. Of course, the analytics also can provide guidance for experiment. The experiment is used to verify the accuracy and reliability of the developed FE model. Also, the experiment can be used to verify the results obtained by analytics. And the prediction results obtained by simulation can provide detailed data information and important guidance for the analytics, experiment and actual production process.

Therefore, synthetically using analytics, experiment and FE numerical modeling technologies such as FE algorithm selection, mesh design and optimization, materials modelling, definition of contact and friction, and treatment of dynamic boundary, developing multi-field, multi-scale and entire process 3D-FE model and carrying out comprehensive simulation & optimization have been an advanced and unique methodology for investigating the concerned metal forming process. The multi-field, multi-scale and entire process 3D-FE model can be used as a virtual experimental platform to rapidly and inexpensively carry out various investigations about the concerned metal forming technology. The research contents, which can be carried out by the virtual experimental platform, include macro plastic deformation behavior, microstructure evolution behavior, effects and coupled effects of various factors (material, process and geometry parameters), and the prediction and control of various macro and micro forming defects, etc. The research aims for the concerned metal forming process are to reveal evolution mechanism of geometry and performance (microstructure) of the preform blank, to realize optimal design and steady control of the process and to develop high-performance and precise aerospace

plasticity technologies. And in Fig.2, the virtual orthogonal experiment is employed to design simulation experimental schedule for the purpose of saving calculation cost as far as possible. The simulation and optimization methods are combined to realize optimal design of the concerned metal forming process so as to obtain defect-free and high-performance deformed products.

2.3 High-end research route for radial-axial ring rolling technology

As an application example of the above proposed methodology for the investigation of aerospace plasticity technology, the high-end research route for radial-axial ring rolling technology has been given in Fig. 3 in consideration of the features of the process.

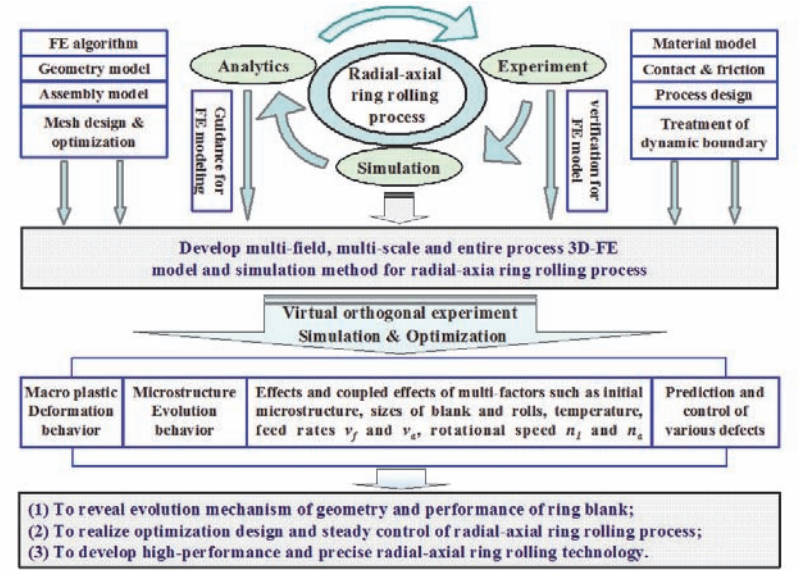


Fig. 3. High-end research route for radial-axial ring rolling technology

The high-end research route for radial-axial ring rolling technology actually outlines the key technologies, methods, contents and aims for the investigation of the process, which can be regarded as an overall planning for the implementation of various research tasks.

3. Numerical modelling of radial-axial ring rolling

Under ABAQUS/Explicit software environment, we developed a coupled thermo-mechanical 3D-FE model for the entire radial-axial ring rolling process, as shown in Fig.4 (Guo & Yang, 2011). The dynamic explicit FEM and mass scaling are used in the model to speed up the computation.

However, the model can only be used to investigate the coupled thermo-mechanical deformation behaviour but can not predict the microstructure evolution of ring blank. But the current model will be an important basis for the simulation of microstructure evolution. The involved key FE modelling technologies are discussed as follows in detail.

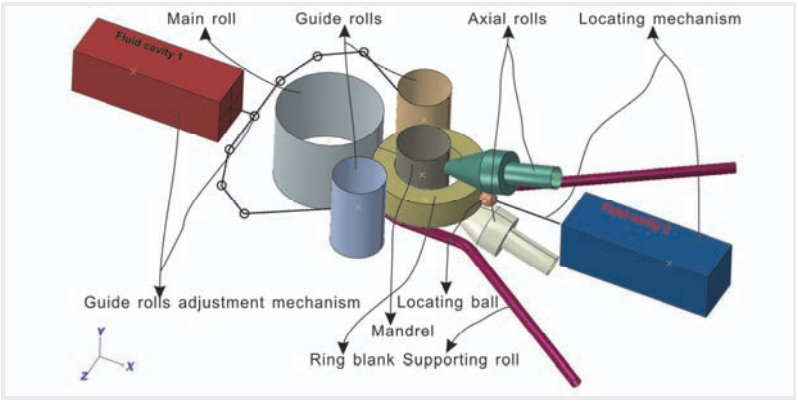


Fig. 4. The coupled thermo-mechanical 3D-FE model for the entire radial-axial ring rolling process

3.1 Key FE modeling technologies

The involved key technologies for the FE modelling of radial-axial ring rolling process mainly include geometry and assembly model, mesh design and optimization, material model, model of guide rolls control mechanism, contact and friction, and determination of the paths of the rolls.

3.1.1 Geometry and assembly model

Geometry model describes the shape and size of every geometry part in the FE model. In Fig.4, all the geometry parts include the ring blank, main roll, mandrel, two guide rolls, two axial conical rolls, supporting rolls, locating ball, guide rolls adjustment mechanism and locating mechanism used to capture the position of the locating ball so as to define the translation motion of the axial rolls in the X direction. The ring blank is modelled by deformable solid body which can be assigned certain material properties. All the rolls and locating ball are modelled by non-deformable analytic rigid surface bodies. For the guide rolls adjustment mechanism and locating mechanism, the fluid cavities are modelled by surface elements based on surface-based fluid cavities technology and the linkages used to control the guide rolls and locating ball are modelled by various connector elements such as beam, weld and hinge. The locating mechanism is designed to capture the growth behavior of the ring throughout the simulation of the radial-axial ring rolling process. Then based on the captured growth behavior of the ring, the translation motion of the axial rolls in the X direction can be conveniently controlled. A small force acting on the end-face of the fluid cavity 2 (as shown in Fig. 4) is preset to avoid appreciable effects on the radial-axial ring rolling process caused by the locating ball. The supporting roll is designed to support the gravity of the ring.

The assembly model is used to describe the relative position relationships among all the geometry parts. In consideration of the spread deformation of the ring in the axial direction, the supporting roll is located below the lower end-face of the ring about 3mm for the FE model shown in Fig.4. It is noted that the linkages for the control of the guide rolls and location ball should be designed to have enough motion space when the ring grows to the maximum diameter. Fig. 4 fully demonstrates the assembly model describing the relative position relationships among all the geometry parts.

3.1.2 Mesh design and optimization

Mesh design and optimization involves selection of element type, check of mesh quality, determination of element size and mesh convergence study, etc.

In the FE model shown in Fig.4, all the rolls and locating ball need not be meshed because they are modelled by rigid surface bodies; the fluid cavities are meshed by surface elements; the linkages are meshed by various connector elements such as beam, hinge and weld; and the ring is meshed by the coupled thermo-mechanical hexahedron element with eight nodes (C3D8RT). An adaptive mesh domain is created for the entire ring to maintain a high-quality mesh throughout an analysis, and reduction integration and hourglass control are employed to save computational time and avoid the zero-energy mode caused by the bending mode of deformation, respectively.

The mesh size is optimized by performing a mesh convergence study, where the same problem is simulated with a different level of refinement of the mesh, and then the simulation results are compared. If further mesh refinement produces a negligible change in the solution, the mesh is said to be converged.

3.1.3 Material model

Material model is used to describe the response of material properties to the coupled thermo-mechanical plastic deformation in metal forming processes and is the key concern for the prediction accuracy of the FE model. The material data, which should be given for the development of FE model, include stress-strain curves under different temperature and strain rate, temperature-dependent physical properties (including linear expansion coefficient, thermal conductivity, specific heat and Young's modulus), density and Poisson ratio, etc.

In Fig.4, the ring material used in the FE model is GH4169 alloy (Chinese grades), equivalent to IN718 alloy of American grades. All the material data stated above are given in the literature (Guo & Yang, 2011). For other geometry parts except the ring, material data need not be assigned to them because they are modeled by either rigid bodies or surface elements for fluid cavities and connector elements for linkages in the guide adjustment mechanism and locating mechanism.

3.1.4 Model of guide rolls control mechanism

The guide rolls play a significant role in both the circularity of the rolled ring and the stability of the radial-axial ring rolling process. Any force imbalance and instability during the rolling process can be removed by the actions of the guide rolls if well-controlled.

In the FE model shown in Fig.4, the guide rolls can be controlled adaptively by an adjustment mechanism which is modeled using the surface-based fluid cavities and connector element technologies in ABAQUS based on the method proposed by Li et al. (2008).

3.1.5 Contact and friction

Contacts are defined to transfer acting force, friction force and heat between the preform blank and various dies for the simulation of metal forming process by setting contact pairs, friction model and contact properties.

For the FE model of radial-axial ring rolling process shown in Fig.4, eight contact pairs are defined between the ring and the rolls. There exists friction and contact heat conduction at the interface of each contact pair. The modified Coulomb friction model is employed, and

the friction coefficients are assumed to be constant during an FE analysis. The guide rolls, supporting roll and locating ball are assumed as smooth surfaces so the friction coefficients on them are zero.

3.1.6 Determination of the paths of the rolls

Determination of the paths of the rolls is the basis for the definition of dynamic boundary conditions in the simulation of radial-axial ring rolling process. In ABAQUS, the amplitude curves with time of the rolls' motion parameters, including the feed rate of the mandrel v_f, rotational speed of the main roll n_1, feed rate of the upper axial conical roll v_a, and rotational speed of the axial conical rolls as shown in Fig.1, can be used to define the paths of the rolls. However, as previously stated, the radial-axial ring rolling process usually exhibits uncooperative motions of the rolls and severe instabilities. The uncooperative motions of the rolls often lead to unexpected distortion of the ring. So it is a great challenge to successfully establish a radial-axial ring rolling process and maintain its stability. How to properly design the motion parameters of the rolls is the key to this problem. Only when the motion parameters are well-designed can the radial-axial ring rolling process be established successfully thus can the FE model of the process be developed successfully to investigate its deformation behavior.

Guo & Yang (2011) proposed a steady forming condition describing both the mathematic correlations and the reasonable ranges of the process parameters of radial-axial ring rolling. And through the developed mathematic model of the steady forming condition, the variation curves with rolling time of n_a, v_f, n_1, v_a and n_a can be determined. These curves can be used as the motion amplitude curves of the rolls to define the dynamic boundary condition for the FE modeling of the radial-axial ring rolling process. The specific details for the determination of the rolls' paths can be referenced in the above literature. The determined paths of the rolls can ensure an approximately constant growth rate of the ring for the consideration of process stability. It is thus clear that analytics is a significant aid for the FE modeling of the metal forming process.

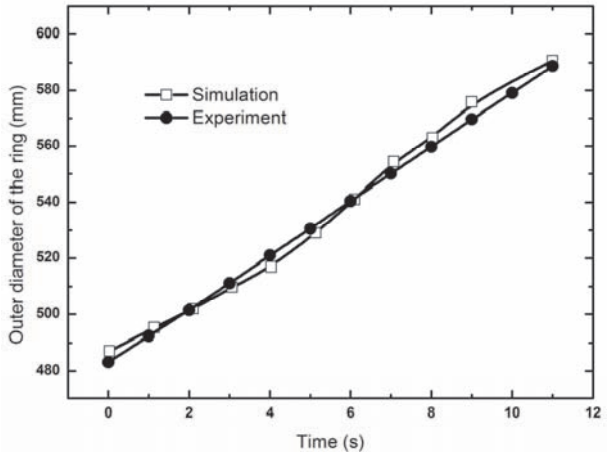


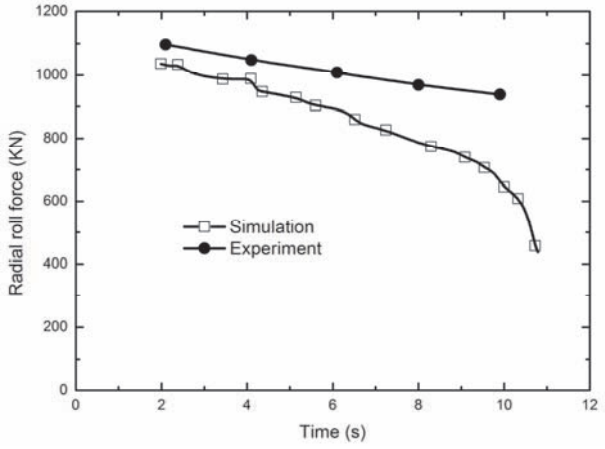
Fig. 5. Measured and predicted variation of the outer diameter of the ring with time

3.2 Model verification

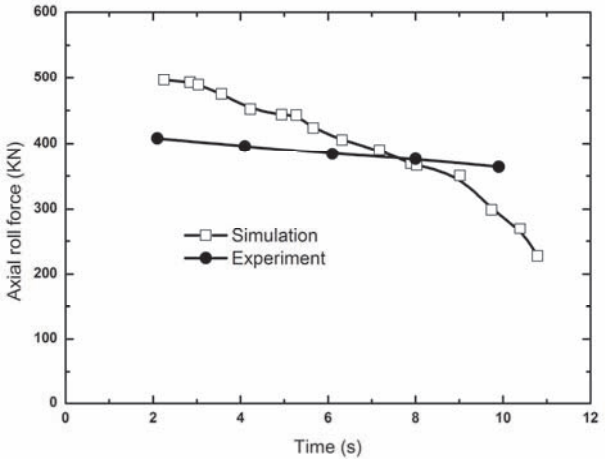
The developed FE model for the radial-axial ring rolling process was verified by experiment in terms of the variations of the outer diameter of the ring and the roll forces in the radial and axial directions, as discussed in the literature (Guo & Yang, 2011).

Fig. 5 shows the measured and predicted variation of the outer diameter of the ring with time. It is seen that the predicted results are in good agreement with the measured ones. This indicates that the accuracy of the developed FE model is sufficient for the prediction of plastic deformation.

Fig. 6 shows the measured and predicted variations of the roll forces in the radial and axial directions with time at the stable rolling stage. It is observed that the predicted roll forces are the same order of the measured ones but with errors to some extent. Through a



(a) radial roll force



(b) axial roll force

Fig. 6. Measured and predicted variations of the roll forces at the stable rolling stage

comparison between measured and predicted results, the maximum relative errors of the radial and axial roll forces are about 10.3% and 21.6% at the stable rolling stage, respectively. The discrepancy between them could be caused by the errors arising from material properties, temperature, thermal boundary conditions, measurement operations and rolling schedules, etc. Therefore, the developed FE model for the radial-axial ring rolling process can be deemed to have enough accuracy and reliability for investigating the plastic deformation behavior of the process.

4. Numerical simulation of radial-axial ring rolling

A simulation case study for radial-axial ring rolling process has been carried out carefully based on the established virtual experimental platform, namely the coupled thermo-mechanical 3D-FE model of the process. For the simulation, the needed various data and parameters include material data, sizes of ring blank, rolled ring and all the rolls, various motion parameters describing the paths of the rolls, etc. These data and parameters all can be selected or determined by the above stated key FE modelling technologies in consideration of both the equipment condition and steady forming condition of the process. The detailed calculation conditions are given as follows.

4.1 Calculation condition

Table 1 gives the used material, sizes of ring blank, rolled ring and all the rolls, friction coefficient, initial temperature of the ring blank, and the rotational speed n_1 and n_a. n_1 is selected according to the equipment condition and n_a is determined by the relation between n_a and n_1, i.e., $n_a=R_1n_1/R_{a1}$ given in the mathematic model of the steady forming condition (Guo & Yang, 2011). All the geometry size parameters listed in Table 1 are labelled in Fig. 7. And the relevant properties data of the used material can be found in the above literature.

Parameters	Value
The used material of ring blank	GH4169 alloy
Outer diameter of the ring blank D_0 (mm)	212.4
Inner diameter of the ring blank d_0 (mm)	130
Wall thickness of the ring blank b_0 (mm)	41.2
Height of the ring blank h_0 (mm)	57.4
Outer diameter of the rolled ring D_f (mm)	300
Inner diameter of the rolled ring d_f (mm)	240
Wall thickness of the rolled ring b_f (mm)	30
Height of the rolled ring h_f (mm)	50
Radius of the main roll R_1 (mm)	100
Radius of the mandrel R_2 (mm)	50
Half of cone angle of the axial rolls θ (°)	17.5
Rolling radius of the axial rolls R_{a1} (mm)	31.9
Friction coefficient μ	0.3
Initial temperature of the ring blank (℃)	1000
Rotational speed of the main roll n_1 (r/min)	60
Rotational speed of the axial rolls n_a (r/min)	188.1

Table 1. The needed various data and parameters in the simulation case study

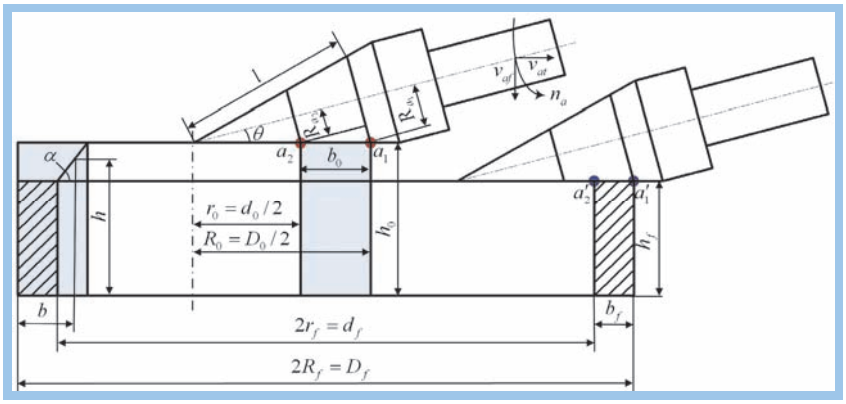


Fig. 7. Position variation of the upper conical roll before and after rolling

Another important work is to design the feed curves (paths) of the mandrel and the upper conical roll for the final determination of the calculation condition. Fig. 8 illustrates the feed curves versus rolling time of the mandrel and upper conical roll determined by the mathematical model of the developed steady forming condition as previously stated.

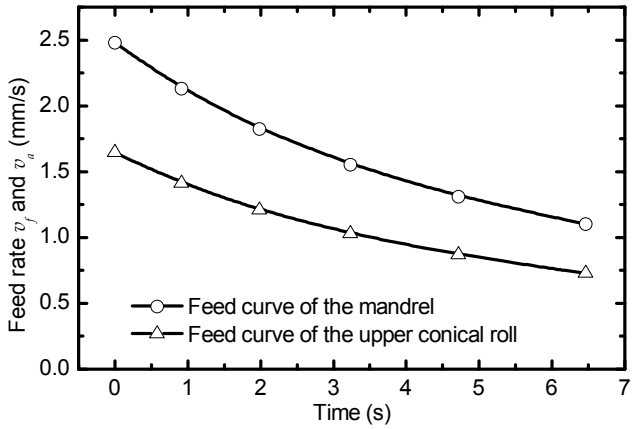
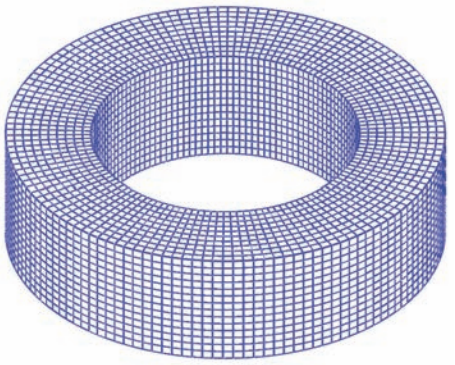


Fig. 8. Feed curves versus rolling time of the mandrel and upper conical roll

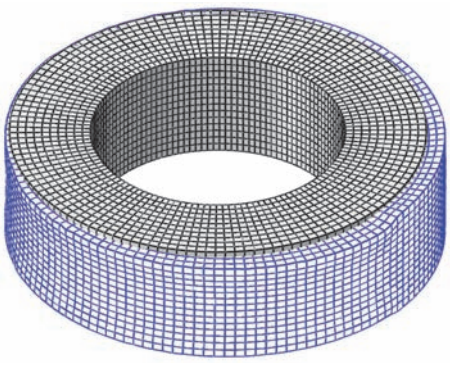
Based on the above calculation condition, a 3D-FE model of radial-axial ring rolling process is developed to predict the ring geometry evolution, stress field, strain field, temperature field, roll force and roll torque. The simulation results are discussed below.

4.2 Ring geometry evolution

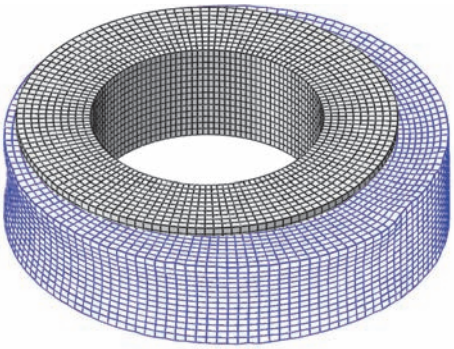
Fig. 9 gives the deformed meshes of the ring blank with the rolling process progressing. It can be observed that the ring produces deformation of reduction in thickness, reduction in height, and extension in diamter during the radial-axial ring rolling process and the circularity of the rolled ring is good. And the simulation indicates that the overall rolling process has good stability.



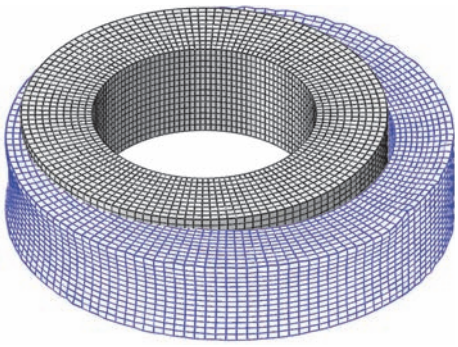
(a) start (t=0 s), iso view



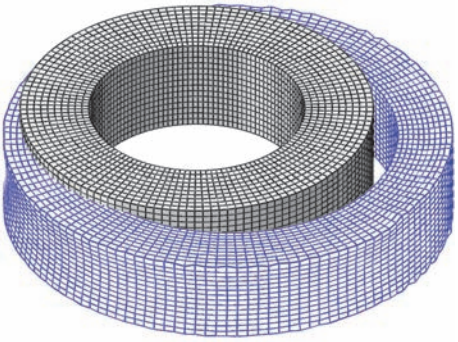
(b) t=2.262 s, iso view



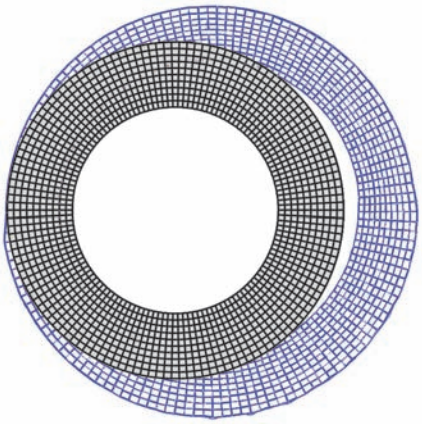
(c) t=4.460 s, iso view



(d) t=5.494 s, iso view



(e) end (t=6.464 s), iso view



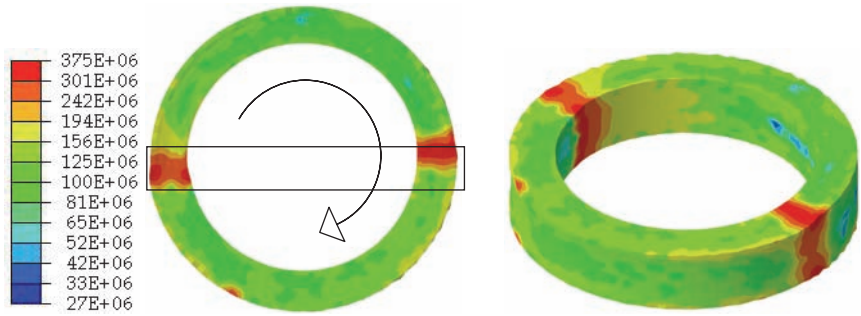
(f) end (t=6.464 s), top view

Fig. 9. Deformed meshes of the ring blank with the rolling process progressing

Observing the changes of deformed meshes of the ring, we can learn that the axial spread produced by the radial rolling is removed by the axial rolling of the axial conical rolls and the radial spread produced by the axial rolling is removed by the radial rolling between the main roll and mandrel. Just under the alternately multi-pass rolling in the radial and axial directions, the ring produces reductions of thickness and heigh and extension of diamter during the radial-axial ring rolling process.

4.3 Stress field

Fig. 10 gives the stress distribution contour of the rolled ring. It is seen that the maximum stress locates in the radial and axial deformation zones. And from the top view of the stress distribution contour, we can find that the radial and axial deformation zones are not on a diameter of the ring, although the radial and axial rolls are configured on a diameter of the ring. The arrow direction indicates the rotational direction of the ring. Just the deformation accumulation of ring materials in the radial and axial deformation zones leads to the reduction of cross-section and expansion of diameter of the ring blank.



(a) top view (b) iso view

Fig. 10. Stress distribution contour of the rolled ring

4.4 Strain field

Fig. 11 shows equivalent plastic strain (PEEQ) distribution contour under different time during the radial-axial ring rolling process. It can be seen that: (1) at the early stage of the process, the plastic deformation basically only produces on the contact surfaces, i.e., the inner and outer surfaces and the upper and lower faces of the ring and then gradually extends to the centre of the ring with the process progressing; and (2) at the end of the process, the maximum plastic deformation locates on the corner close to the outer surface of the ring, while at the centre of the ring, there is an approximately circular ring zone in which the plastic strain is relatively small and much smaller than that on the maximum strain zone.

4.5 Temperature field

Fig. 12 shows temperature distribution contour under different time during the radial-axial ring rolling process. We can see that: (1) the minimum temperature zone locates in the centre area of the inner surface of the ring and the maximum temperature zone locates on the corner close to the outer surface of the ring; (2) in the minimum temperature zone, the

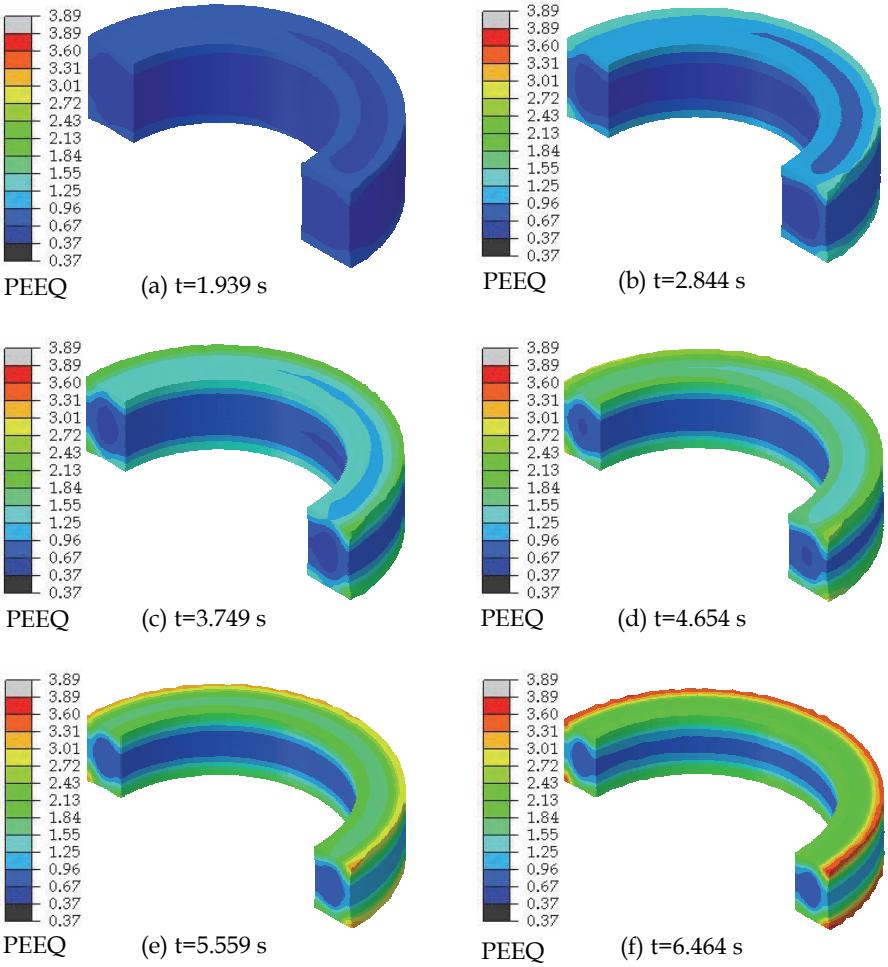


Fig. 11. Equivalent plastic strain (PEEQ) distribution contour under different time during the radial-axial ring rolling process

temperature decreases compared with the initial temperature of the ring due to contact heat dissipation; and (3) in the maximum temperature zone, the temperature increases compared with the initial temperature of the ring due to the heat generation of maximum plastic deformation.

4.6 Roll force and torque

Fig. 13 gives the variations of roll forces, which are measured by reaction forces on the mandrel, main roll, upper conical roll and lower conical roll, during radial-axial ring rolling process. From the figure it can be observed that: (1) the roll forces rapidly increase to maximum value at the early rolling stage and then gradually decrease at the steady rolling stage of the process; (2) the reaction force on the mandrel is greater than on the main roll, so

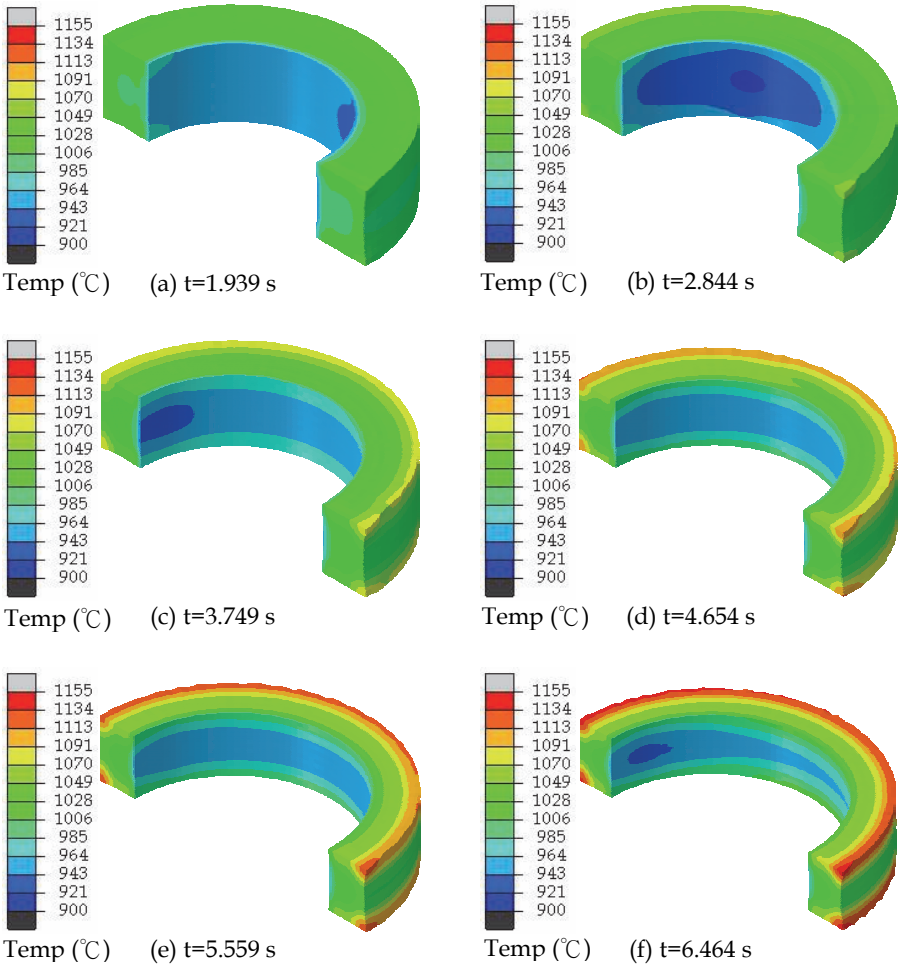


Fig. 12. Temperature distribution contour under different time during the radial-axial ring rolling process

the radial roll force should be determined by the reaction force on the mandrel for the selection of roll mill in consideration of safety for this simulation case; and (3) the reaction forces on the upper and lower conical rolls are basically equivalent, so the axial roll force can be determined by any one of them.

Fig. 14 shows the variation of the contact area between the ring and the rolls during the radial-axial ring rolling process. It is seen that the variation laws of the contact area during the process is similar to the ones of the roll forces shown in Fig.13. The size of the contact area between the ring and the rolls reflects the size of the deformation zone. The bigger is the contact area, the bigger is the deformation zone, and vice versa. So we can conclude from Fig.14 that the deformation zone in the radial pass is bigger than that in the axial pass for this simulation case.

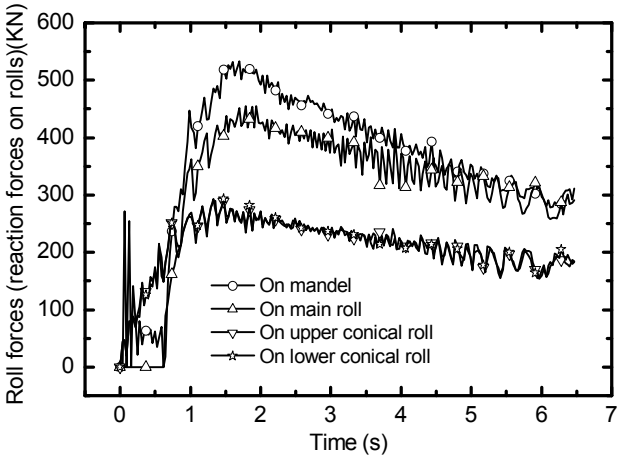


Fig. 13. Variation curves of roll forces during the radial-axial ring rolling process

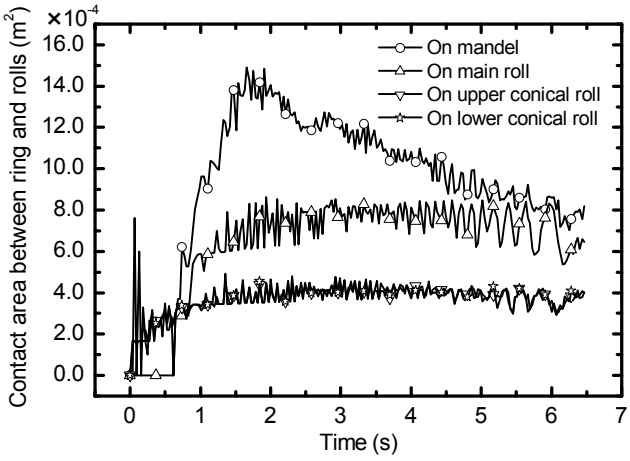


Fig. 14. The variation curves of the contact area between the ring and the rolls during the radial-axial ring rolling process

Fig. 15 gives the variations of the radial and axial roll torques, which are measured by reaction torques on the main roll and the axial upper and lower conical rolls, during radial-axial ring rolling process.

From Fig. 15 it can be observed that at the early stage, the variations of the radial and axial roll torques are relatively drastic and then tend towards stability. This demonstrates that the early rolling stage should be well-controlled for performing a successful radial-axial ring rolling operation. And the reaction torques of the upper and lower conical rolls have the same variation laws but reverse values because the axial conical rolls rotate in reverse direction during the rolling process.

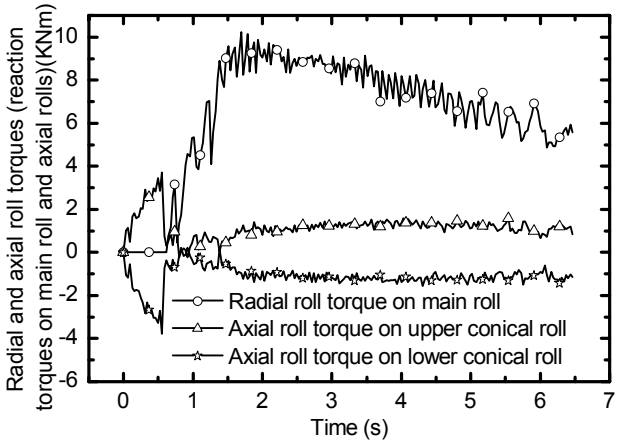


Fig. 15. Variation curves of the radial and axial roll forces during the radial-axial ring rolling process

5. Conclusion and future work

In consideration of the unique requirements of light weight, high precision, high performance, high reliability and high efficiency for the plasticity forming manufacture of various key aerospace components, the main challenges for the R&D of aerospace plasticity technology are summarized concisely. Facing the challenges, we have proposed a high-end research route for systematically and deeply investigating the aerospace plasticity technology, and pointed out that multi-field, multi-scale and entire process 3D-FE modeling, simulation and optimization has been an advanced and unique methodology for the rapid and inexpensive investigation of various aerospace plasticity technologies, especially for the R&D of new aerospace products and plasticity technologies for new materials.

As an application example of the proposed methodology for the investigation of aerospace plasticity technology, the high-end research route for radial-axial ring rolling technology has been given in consideration of the features of the process, which can be regarded as an overall planning for the implementation of various research tasks. Guided by the high-end research route, we first discussed the key FE modelling technologies such as geometry and assembly model, mesh design and optimization, material model, model of guide rolls control mechanism, contact and friction and determination of the paths of the rolls in detail, and then demonstrated the validation of the developed thermo-mechanical coupled 3D-FE model of radial-axial ring rolling process in terms of comparison of ring geometry and radial-axial roll forces with experiment.

Taking the coupled thermo-mechanical 3D-FE model as a virtual experimental platform, we carried out a simulation case study of radial-axial ring rolling process carefully. Some key simulation results, including ring geometry evolution, stress field, strain field, temperature field, roll forces and roll torques, are reported. The obtained simulation results are beneficial to gaining insight into the forming mechanism and laws of radial-axial ring rolling process and will provide important basis for the optimal design and steady control of the process.

However, the current studies carried out in this book chapter only basically realized multi-field (coupled thermo-mechanical) and entire process FE modelling and simulation. So,

according to the proposed high-end research route (as shown in Fig.3) and the key science problem proposed in introduction section of radial-axial ring rolling technology, the future work for the process is summarized as follows.

1. To develop microstructure evolution model of ring blank materials during radial-axial ring rolling process so as to numerically reveal the response of materials properties to the process;
2. To establish macro-micro coupled multi-scale FE model of radial-axial ring rolling process by embedding the microstructure evolution model of ring blank materials into the current coupled thermo-mechanical 3D-FE model of the process;
3. To reveal evolution mechanism of the geometry and microstructure of the ring blank by investigating macro plastic deformation and microstructure evolution behaviors, coupled effects of multi-factors, and macro and micro forming defects during radial-axial ring rolling process;
4. To establish optimization design and steady control method of radial-axial ring rolling process and to develop high-performance and precise radial-axial ring rolling technology.

6. Acknowledgment

The authors would like to thank the National Natural Science Foundation of China (50805120, 50935007), the National Basic Research Program of China ("973" Program) (2010CB731701), the Major National Science and Technology Special Project of China (2009ZX04014-074-03, 2010ZX04004-131-07), and the Research Fund of State Key Laboratory of Solidification Processing of China (KP200911) for the support given to this research.

7. References

Allwood JM, Tekkaya AE & Stanistreet TF. (2005). The development of ring rolling technology. *Steel Research International*, Vol.76, No.2-3, (September 2005), pp. 111-120, ISSN 0177-4832

Allwood JM, Tekkaya AE & Stanistreet TF. (2005). The development of ring rolling technology—Part 2: Investigation of process behaviour and production equipment. *Steel Research International*, Vol.76, No.7, (September 2005), pp. 491-507, ISSN 1611-3683

Davey K & Ward MJ. (2003). An ALE approach for finite element ring-rolling simulation of profiled rings. *Journal of Materials Processing Technology*, Vol.139, No.1-3, (August 2003), pp. 559-566, ISSN 0924-0136

Eruc E & Shivpuri R. (1992). A summary of ring rolling technology—I. Recent trends in machines, processes and production lines. *International Journal of Machine Tools and Manufacture*, Vol.32, No.3, (June 1992), pp. 379-98, ISSN 0890-6955

Eruc E & Shivpuri R. (1992). A summary of ring rolling technology—II. Recent trends in process modeling, simulation, planning and control. *International Journal of Machine Tools and Manufacture*, Vol. 32, No.3, (June 1992), pp. 399-413, ISSN 0890-6955

Forouzan MR, Salimi M, Gadala MS & Aljawi AA. (2003). Guide roll simulation in FE analysis of ring rolling. *Journal of Materials Processing Technology*, Vol.142, No.1, (November 2003), pp. 213-223, ISSN 0924-0136

Forouzan MR, Salimi M & Gadala MS. (2003). Three-dimensional FE analysis of ring rolling by employing thermal spokes method. *International Journal of Mechanical Sciences*, Vol.45, No.12, (December 2003), pp. 1975-1998, ISSN 0020-7403

Fan XG, Yang H, Sun ZC, & Zhang DW. (2010). Effect of deformation inhomogeneity on the microstructure and mechanical properties of large-scale rib – web component of titanium alloy under local loading forming. *Materials Science and Engineering A*, Vol.527, No.21-22, (August 2010), pp. 5391-5399, ISSN 0921-5093

Guo LG, Yang H & Zhan M. (2005). Research on plastic deformation behaviour in cold ring rolling by FEM numerical simulation. *Modelling and Simulation in Materials Science and Engineering*, Vol.13, No.7, (October 2005), pp. 1029–1046, ISSN 0965-0393

Guo LG & Yang H. (2011). Towards a steady forming condition for radial–axial ring rolling. *International Journal of Mechanical Sciences*, Vol.53, No.4, (April 2011), pp. 286-299, ISSN 0020-7403

Hawkyard J B, Johnson W, Kirkland J & Appleton E. (1973). Analyses for roll force and torque in ring rolling, with some supporting experiments. *International Journal of Mechanical Sciences*, Vol. 15, No.11, (November 1973), pp. 873-893, ISSN 0020-7403

Hahn YH & Yang DY. (1994). UBET analysis of the closed-pass ring rolling of rings having arbitrarily shaped profiles. *Journal of Materials Processing Technology*, Vol. 40, No.3-4, (January 1994), pp. 451-463, ISSN 0924-0136

Hua L, Pan LB & Lan J. (2009). Research on the ring stiffness condition in radial-axial ring rolling. *Journal of Materials Processing Technology*, Vol.209, No.5, (March 2009), pp. 2570-2575, ISSN 0924-0136

Johnson W & Needham G. (1968). Experiments on ring rolling. *International Journal of Mechanical Sciences*, Vol.10, No.2, (February 1968), pp. 95-113, ISSN 0020-7403

Joun MS, Chung JH & Shivpuri R. (1998). An axisymmetric forging approach to preform design in ring rolling using a rigid-viscoplastic finite element method. *International Journal of Machine Tools and Manufacture*, Vol.38, No.10-11, (October 1998), pp. 1183-1191, ISSN 0890-6955

Jong TY, Jeoung HK, Nho KP, Seung SC & Chong SL. (2007). Ring-rolling design for a large-scale ring product of Ti-6Al-4V alloy. *Journal of Materials Processing Technology*, Vol. 187-188, No. 12, (June 2007), pp. 747-751, ISSN 0924-0136

Kim N, Machida S & Kobayashi S. (1990). Ring rolling process simulation by the three dimensional finite element method. *International Journal of Machine Tools and Manufacture*, Vol.30, No.4, (1990), pp. 569-577, ISSN 0890-6955

Kang BS & Kobayashi S. (1991). Preform design in ring rolling processes by the three-dimensional finite element method. *International Journal of Machine Tools and Manufacture*, Vol.31, No.1, (1991), pp. 139-151, ISSN 0890-6955

Lugora CF & Bramley AN. (1987). Analysis of spread in ring rolling. *International Journal of Mechanical Sciences*, Vol.29, No.2, (1987), pp. 149-157, ISSN 0020-7403

Liu YL, Yang H, Zhan M & Fu ZX. (2002). A study of the influence of the friction conditions on the forging process of a blade with a tenon. *Journal of Materials Processing Technology*, Vol.123, No.1, (April 2002), pp. 42-46, ISSN 0924-0136

Li LY, Yang H, Guo LG & Sun ZC. (2008). A control method of guide rolls in 3D-FE simulation of ring rolling. *Journal of Materials Processing Technology*, Vol.205, No.1-3, (August 2008), pp. 99-110, ISSN 0924-0136

Li H, Yang H, Zhan M & Kou YL. (2010). Deformation behaviors of thin-walled tube in rotary draw bending under push assistant loading conditions. *Journal of Materials Processing Technology*, Vol.210, No.1, (January 2010), pp. 143-158, ISSN 0924-0136

Mamalis AG, Hawkyard JB & Johnson W. (1975). Cavity formation in rolling profiled rings. *International Journal of Mechanical Sciences*, Vol. 17, No.11-12, (1975), pp. 669-672, ISSN 0020-7403

Moon HK, Lee MC & Joun MS. (2008). Predicting polygonal-shaped defects during hot ring rolling using a rigid-viscoplastic finite element method. *International Journal of Mechanical Sciences*, Vol.50, No.2, (February 2008), pp. 306-314, ISSN 0020-7403

Wang ZW, Zeng SQ, Yang XH & Cheng C. (2007). The key technology and realization of virtual ring rolling. *Journal of Materials Processing Technology*, Vol.182, No.1-3, (February 2007), pp. 374-381, ISSN 0924-0136

Wang M, Yang H, Sun ZC & Guo LG. (2009). Analysis of coupled mechanical and thermal behaviors in hot rolling of large rings of titanium alloy using 3D dynamic explicit FEM. *Journal of Materials Processing Technology*, Vol.209, No.7, (April 2009), pp. 3384-3395, ISSN 0924-0136

Yang DY, Kim KH & Hawkyard JB. (1991). Simulation of T-section profile ring rolling by the 3-D rigid-plastic finite element method. *International Journal of Mechanical Sciences*, Vol.33, No.7, (1991), pp. 541-550, ISSN 0020-7403

Yang H, Zhan M & Liu YL. (2002). A 3D rigid–viscoplastic FEM simulation of the isothermal precision forging of a blade with a damper platform. *Journal of Materials Processing Technology*, Vol.122, No.1, (March 2002), pp. 45–50, ISSN 0924-0136

Yang H, Zhan M, Liu YL, Xian FJ, Sun ZC, Lin Y & Zhang XG. (2004). Some advanced plastic processing technologies and their numerical simulation. *Journal of Materials Processing Technology*, Vol. 151, No.1-3, (April 2004), pp. 151:63-69, ISSN 0924-0136

Yang H, Wang M, Guo LG, & Sun ZC. (2008). 3D coupled thermo-mechanical FE modeling of blank size effects on the uniformity of strain and temperature distributions during hot rolling of titanium alloy large rings. *Computational Materials Science*, Vol.44, No.2, (December 2008), pp. 611–621, ISSN 0927-0256

Yang H, Huang L, & Zhan M. (2010). Coupled thermo-mechanical FE simulation of the hot splitting spinning process of magnesium alloy AZ31. *Computational Materials Science*, Vol47, No.3, (January 2010), pp. 857-866, ISSN 0927-0256

Yang H, Li H, & Zhan M. (2010). Friction role in bending behaviors of thin-walled tube in rotary-draw-bending under small bending radii. *Journal of Materials Processing Technology*, Vol.210, No.15, (November 2010), pp. 2273-2284, ISSN 0924-0136

Zhan M, Yang H, Huang L & Gu RJ. (2006). Springback analysis of numerical control bending of thin-walled tube using numerical-analytic method. *Journal of Materials Processing Technology*, Vol.177, No.1-3, (July 2006), pp. 197-201, ISSN 0924-0136

Zhan M, Yang H, Zhang JH, Xu YL & Ma F. (2007). 3D FEM analysis of influence of roller feed rate on forming force and quality of cone spinning. *Journal of Materials Processing Technology*, Vol.187–188, (June 2007), pp. 486–491, ISSN 0924-0136

Zhou G, Hua L, Lan J & Qian DS. (2010). FE analysis of coupled thermo-mechanical behaviors in radial-axial rolling of alloy steel large ring. *Computational Materials Science*, Vol.50, No.1, (November 2010), pp. 65-76, ISSN 0927-0256

Zhou G, Hua L & Qian DS. (2011). 3D coupled thermo-mechanical FE analysis of roll size effects on the radial-axial ring rolling process. *Computational Materials Science*, Vol. 50, No.3, (January 2011), pp. 911-924, ISSN 0927-0256

18

Kinetostatics and Dynamics of Redundantly Actuated Planar Parallel Link Mechanisms

Takashi Harada
Kinki University
Japan

1. Introduction

Robotic systems with parallel link mechanisms (PLMs) have mechanical characteristics such as rigidity of the mechanism and precise positioning (Stewart, 1966), (Merlet 2006), (Wang & Liu, 2008). These characteristics enable them to stably perform contact tasks with sensitive force, e.g. mold grindings and rehabilitation robotics. On the other hand, mechanical interference and the singularity of the mechanisms (Merlet, 1989) restrict the robot's movable range. PLMs have therefore been conventionally applied not to general-purpose industrial robots, but to special-purpose machines (Weck, 2002), (Oiwa, 1997).
In order to expand this limited application of PLMs, we have proposed a new parallel link mechanism with multi drive linear motors (MDLMs) (Harada & Nagase, 2009, 2010). The multi drive is a control method for linear motors in which a number of moving parts are individually driven on one stator part. We have proposed various configurations of PLMs which have been constructed for MDLMs. These PLMs expand the robot's movable range while retaining the advantageous rigid mechanism and precise positioning that PLMs offer. Moreover, the proposed PLMs are suitable for force control, because the linear motors are directory driven without friction full gearings.
Several studies related to expanding the movable range of the PLM have previously been published (Honegger et al., 1997), (Kim et al., 2003), (Liu et al., 2004), (In et al., 2008), (Milutinovis et al., 2005), (Zhang, 2008). Notably, redundantly actuated 3-DOF $xy\theta$ planar PLMs on linear actuators (Zhang, 2008), (Wang et al., 2008), (Marquet et al., 2001) have been proposed as mechanisms that are similar to our PLM. A two 2-DOF PRRRP (P denotes prismatic joint and R denotes rotational) parallel manipulator (Liu et al., 2007) has been employed as a mechanical element of these planar PLMs, including ours. However, these planar PLMs, excluding ours, aim at position control, not at force control. Conventionally, gearings or ball screws are used for the actuator transmission of PLMs. However, it is difficult to compensate for the undesired internal force among the redundant link mechanisms with the position controlled actuators. This undesired internal force results in mechanical deformation around the transmission parts (Leong et al., 2004).
Our PLM is suitable for force control, because it employs directly driven linear motors. It can control the internal force and compensate for the mechanical deformation because of the favorable effect of force control and the back-drivability of the directly driven linear motors.

Moreover, it is designed to have constant inertia and decoupled dynamics in the x (horizontal) direction. Gravitational force only affects it in the y (vertical) direction. The simple characteristics of the dynamics make it easy to install a fast acceleration control or impedance control (Hogan, 1985) to the PLM, where the nonlinear dynamics of the PLM must be compensated for.

In this paper, we investigate the kinetostatics (kinametics and static force), and dynamics characteristics of the 3D4M PLM using symbolic mathematical analysis and numerical simulations. This paper is organized as follows: first, configurations of the 3D4M PLM on multi drive linear motors are introduced. Second, kinematic equations, forward kinematics and derivative kinematics of the 3D4M PLM are derived. The derived equations are symbolically programmed using Mathematica. Next, singularity and static forces of the 3D4M PLM are analyzed using Mathematica. Then, the decoupled dynamical design of the 3D4M PLM is introduced. The equations of motion of the 3D4M PLM are derived by symbolic programming using Mathematica. Finally, some examples are introduced for solving the equations of motion numerically using Mathematica.

2. Configurations of link mechanism

2.1 Multi drive linear motor

A ball screw driven by a rotational motor, as shown in Fig.1 (a), is generally used as a linear actuator in conventional PLMs. A single driving part moves in a straight line on a linear stator; we will refer to this below as a single drive. The single drive disturbs the space in which movement takes place, and restricts the general-purpose application of PLMs. Moreover, it is difficult in principal for the load of the tip to be transmitted back to the actuator, which in turn renders the ball screw drive incapable of force control.

To cope with these problems, multi drive linear motors (MDLMs), as shown in Fig.1 (b), are employed in our research. MDLMs offer a way to arrange more than one moving parts on one stator of a linear motor, with each moving part individually controlled and driven.

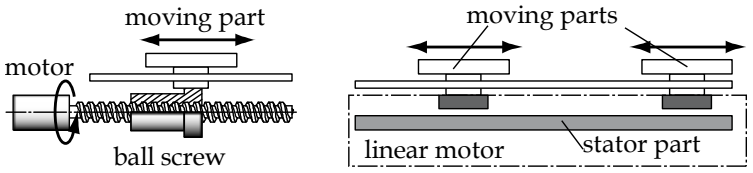


Fig. 1. Single and multi drive linear motors

2.2 Configuration of 3-DOF planar mechanisms

Configurations of a 3-DOF ($xy\theta$) planar PLM with 3 non-redundant moving parts (3D3M) and a 3-DOF ($xy\theta$) planar PLM with 4 redundant moving parts (4D4M), are shown in Figs. 2(a) and (b). The 3D4M PLM with 4 redundant moving parts is the centerpiece of our research. The redundancy of the PLMs is not used only for singularity avoidance as sought by conventional research, but is also used for forward kinematics computation (Merlet, 1996) and calibration of the mechanism (Zhuang & Liu, 1998), (Chiu & Perg, 2003), which have been standing problems with conventional PLMs.

We are planning to apply the PLM to a table mechanism of 5 axis machine tools. A schematic view and the motions of the 3D4M PLM are illustrated in Fig. 3 and Fig. 4.

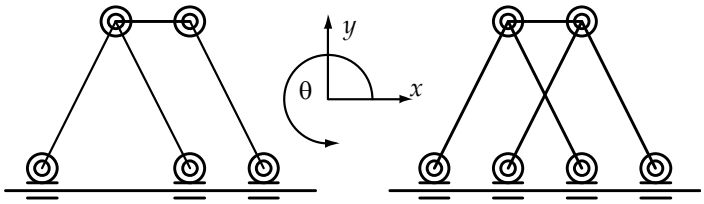


Fig. 2. 3-DOF planar parallel mechanisms

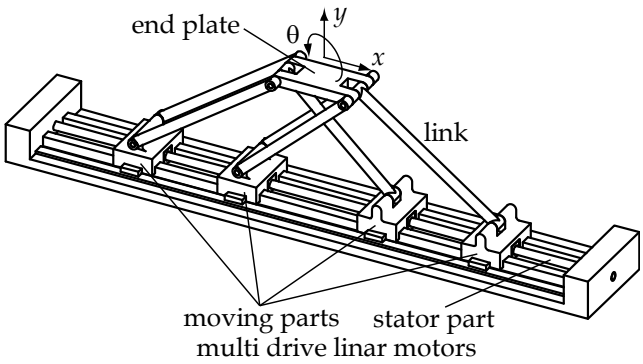


Fig. 3. Schematic view of the 3D4M mechanism

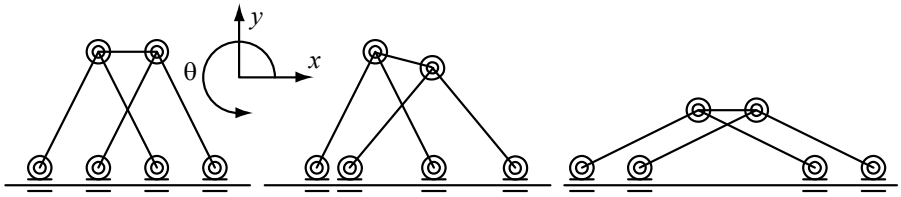


Fig. 4. Motions of the 3D4M mechanism

3. Kinematics of the parallel link mechanism

3.1 Kinematics of 3-DOF parallel link mechanisms

On the basis of the general kinematics formulation of parallel link mechanisms (Arai et al., 1991), kinematics equations for the proposed 3-DOF parallel link mechanism are derived. In turn, the parallel link mechanism of the particular configuration of our research can also be analyzed by the general method of kinematics.

The kinematic relationships of the 3D4M PLM, as shown in Fig. 5, are expressed as follows:

$$\begin{aligned}
&\mathbf{p}+\mathbf{R}\mathbf{s}_i=\mathbf{L}_i=c_i\mathbf{a}+l_i\mathbf{z}_i\\
&\mathbf{p}=[x,y]^T\\
&\mathbf{a}=[1,0]^T\\
&\mathbf{R}=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\\
&\mathbf{s}_1=\mathbf{s}_2=[-r_t,0]^T\\
&\mathbf{s}_3=\mathbf{s}_4=[r_t,0]^T
\end{aligned} \quad . \tag{1}$$

where c_i (i = 1,..,4) is the control variable of the ith actuator, i.e., the position of the ith moving part of the multi drive linear motor. The length of the ith rod (link) is expressed as l_i. The distance from the central point of the end plate to the ith pair of the end plate is expressed as r_t. Other symbols are indicated as in Fig. 5.

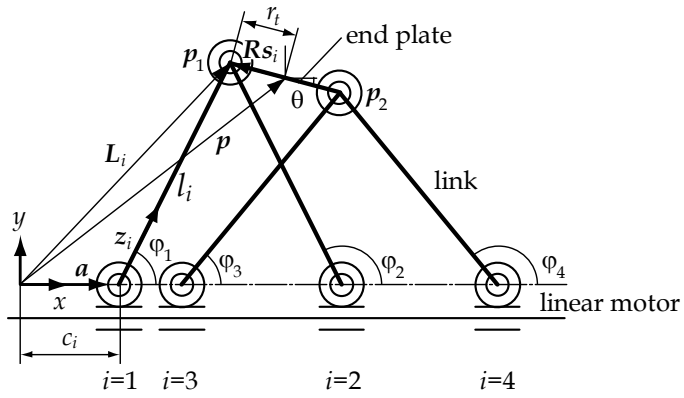


Fig. 5. Kinematic model of the 3-DOF PLM

Equation (1) expresses the relationship between the positions of the moving parts c_i (i=1,..,4) and the positions $\mathbf{p}$ and orientation θ of the end plate. By solving (1) as c_i, an inverse kinematics equation is derived as follows:

$$c_i=\mathbf{L}_i{}^T\mathbf{a}\pm\sqrt{(\mathbf{L}_i{}^T\mathbf{a})^2-(\mathbf{L}_i{}^T\mathbf{L}_i)^2+l_i{}^2}\ . \tag{2}$$

For the 1st and 3rd link, the plus-minus sign in (2) is given as positive, and for 2nd and 4th link is given as negative. The unit direction vector $\mathbf{z}_i$ and angle φ_i of the ith rod are given as

$$\begin{aligned}
&\mathbf{z}_i=[z_{xi},z_{yi}]^T=(\mathbf{L}_i-c_i\mathbf{a})/l_i\\
&\varphi_i=\tan^{-1}\left(\frac{z_{yi}}{z_{xi}}\right)
\end{aligned} \quad . \tag{3}$$

By applying derivatives to both sides of (1), the derivative relation of the 3D4M PLM is derived as follows:

$$\begin{aligned}
&\mathbf{J}_{e43}\Delta\mathbf{p}_3 = \mathbf{J}_{c4}\Delta\mathbf{c}_4 \\
&\Delta\mathbf{p}_3 = [\Delta x, \Delta y, \Delta\theta]^T \\
&\Delta\mathbf{c}_4 = [\Delta c_1, ..., \Delta c_4]^T \\
&\mathbf{J}_{e43} = \begin{bmatrix} \mathbf{z}_1^T & \mathbf{z}_1^T(\mathbf{R}_\theta\mathbf{R}\mathbf{s}_1) \\ \vdots & \vdots \\ \mathbf{z}_4^T & \mathbf{z}_4^T(\mathbf{R}_\theta\mathbf{R}\mathbf{s}_4) \end{bmatrix}, \mathbf{J}_{c4} = \begin{bmatrix} \mathbf{z}_1^T\mathbf{a} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{z}_4^T\mathbf{a} \end{bmatrix} \\
&\mathbf{R}_\theta = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}
\end{aligned} \quad . \tag{4}$$

Where $\mathbf{J}_{e43}$ and $\mathbf{J}_{c4}$ are Jacobian matrices of the system. Kinematic characteristics such as singular point and static force can be analyzed by using Eq. (4).
The kinematic equation of the 3D3M PLM is derived by removing the redundant part from Eqs. (1)-(4) as follows:

$$\begin{aligned}
&\mathbf{J}_{e33}\Delta\mathbf{p}_3 = \mathbf{J}_{c3}\Delta\mathbf{c}_3 \\
&\Delta\mathbf{p}_3 = [\Delta x, \Delta y, \Delta\theta]^T \\
&\Delta\mathbf{c}_3 = [\Delta c_1, \Delta c_2, \Delta c_4]^T
\end{aligned} \quad . \tag{5}$$

3.2 Symbolical programming using Mathematica

Equations (1)-(5) are symbolically programmed using Mathematica. Mathematica is a computational software program widely used in scientific, engineering, and mathematical fields and other areas of technical computing.
Positional vector **p** and matrix **R** in Eq. (1) are symbolically defined by Mathematica.

```
p = {x[t], y[t]};
R = {{Cos[θ[t]], -Sin[θ[t]]}, {Sin[θ[t]], Cos[θ[t]]}};
```
(M. 1)

Program code of Mathematica is indicated as (M. *i*). Valuables $x[t]$, $y[t]$ and $\theta[t]$ in (M. 1) are defined as function of time t.
Vectors $\mathbf{s}_i$, $\mathbf{L}_i$ (i=1,…4) and **a** in Eq. (1) are defined as

```
s1 = {-rt, 0};
s2 = {-rt, 0};
s3 = {rt, 0};
s4 = {rt, 0};
```
(M. 2)

```
L1 = p + R.s1;
L2 = p + R.s2;
L3 = p + R.s3;
L4 = p + R.s4;
```
(M. 3)

```
a = {1, 0};
```
(M. 4)

The position of each moving part c_i (i=1,...,4) in Eq. (2) is expressed as

```
c1 = L1.a - Sqrt[(L1.a)^2 - L1.L1 + l1^2];
c2 = L2.a + Sqrt[(L2.a)^2 - L2.L2 + l2^2];
c3 = L3.a - Sqrt[(L3.a)^2 - L3.L3 + l3^2];
c4 = L4.a + Sqrt[(L4.a)^2 - L4.L4 + l4^2];
```
(M. 5)

The unit direction vector $\mathbf{z}_i$ and angle φ_i (i=1,...,4) of the i th rod in Eq. (3) are defined as

```
z1 = (L1 - c1 a) / l1;
z2 = (L2 - c2 a) / l2;
z3 = (L3 - c3 a) / l3;
z4 = (L4 - c4 a) / l4;
```
(M. 6)

```
φ1 = ArcTan[z1[[1]], z1[[2]]];
φ2 = ArcTan[z2[[1]], z2[[2]]];
φ3 = ArcTan[z3[[1]], z3[[2]]];
φ4 = ArcTan[z4[[1]], z4[[2]]];
```
(M. 7)

Matrix $\mathbf{R}_\theta$ in Eq. (4) is defined as

```
Rθ = {{0, -1}, {1, 0}};
```
(M. 8)

Jacobian Matrices $\mathbf{J}_{e43}$ in Eq. (4) and $\mathbf{J}_{e33}$ in Eq. (5) are expressed as

```
Je43 = {Flatten[{z1, z1.Rθ.R.s1}],
   Flatten[{z2, z2.Rθ.R.s2}],
   Flatten[{z3, z3.Rθ.R.s3}],
   Flatten[{z4, z4.Rθ.R.s4}]};
```
(M. 9)

```
Je33 = {Flatten[{z1, z1.Rθ.R.s1}],
   Flatten[{z2, z2.Rθ.R.s2}],
   Flatten[{z4, z4.Rθ.R.s4}]};
```
(M. 10)

3.3 Forward kinematics

As shown in Fig. 5, two links are connected to the same pair of each end of the end plate. This makes it easy to solve the forward kinematics solution. The positions of the kinematic pairs p_1:$[x_1,y_1]^T$ and p_2:$[x_2,y_2]^T$ are expressed by the c_i, the position of the moving part of the linear motor, as follows:

$$\begin{aligned}
&[x_1, y_1]^T = [c_1 + l_1 \cos\varphi_1, l_1 \sin\varphi_1]^T \\
&\varphi_1 = \cos^{-1}\left(\frac{{l_1}^2 + (c_2 - c_1)^2 - {l_2}^2}{2l_1(c_2 - c_1)}\right) \\
&[x_2, y_2]^T = [c_3 + l_3 \cos\varphi_2, l_4 \sin\varphi_2]^T \\
&\varphi_2 = \cos^{-1}\left(\frac{{l_3}^2 + (c_4 - c_3)^2 - {l_4}^2}{2l_1(c_4 - c_3)}\right)
\end{aligned} . \tag{6}$$

If the length of each link l_i is designed with the same value as l, the forward kinematics of the PLM becomes a simple formula, as follows:

$$\begin{aligned}\left[x_1, y_1\right]^T &= \left[(c_2 + c_1)/2, \sqrt{l^2 - ((c_2 - c_1)/2)^2}\right]^T \\ \left[x_2, y_2\right]^T &= \left[(c_3 + c_4)/2, \sqrt{l^2 - ((c_4 - c_3)/2)^2}\right]^T\end{aligned} . \tag{7}$$

The orientation and the central position of the end plate are given from the positions of the actuators c_i, as follows:

$$\begin{aligned}x &= (x_1 + x_2)/2 \\ y &= (y_1 + y_2)/2 \\ \theta &= \tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right)\end{aligned} . \tag{8}$$

3.4 Singularity of the 3-DOF PLM

It is well known that parallel link mechanisms have two kinds of singularities (Gosselin & Angels, 1990). When PLM arrives at a position and orientation such that the Jacobian matrices $\mathbf{J}_{cm}$ in Eqs. (4)-(5) are singular, the output of the actuator does not transfer to the link mechanism. These situations are referred to as the 1st kind of singularity. When the PLM arrives at a position and orientation such that the Jacobian matrices $\mathbf{J}_{en}$ in Eqs. (4)-(5) are singular, the output of the end effecter does not transfer to the link mechanism. These situations are referred to as the 2nd kind of singularity. Variables m and n express the number of actuators and the degree of freedom of the end plate, respectively.

When each element of $\mathbf{J}_{cm}$ equals zero, the proposed PLM becomes the 1st kind of singularity. The condition of the 1st kind of singularity is expressed as follows:

$$\mathbf{z}_i^T \mathbf{a} = 0 . \tag{9}$$

Equation (9) implies that if the unit direction vector $\mathbf{a}$ of the actuator and the unit direction vector $\mathbf{z}_i$ of each link are orthogonal, the PLM becomes one of the 1st kind of singular points. The 2nd kind of singularity differs, depending on the configuration of the PLM. Here, the singular points of the 3D3M PLM are derived. The conditions of the 2nd kind of singularities are given as the determinant of the Jacobian matrix in Eq. (5), which equals zero as follows:

$$\det(\mathbf{J}_{e33}) = 0 . \tag{10}$$

Using symbolic mathematics software, equation (10) is solved. In case of each rod length l_i is identical as l, the following 4 types of singular points exist in the 3D3M PLM.

Type 1: $$\theta = \sin^{-1}(y / r_t) \tag{11. 1}$$

Type 2: $$\theta = \sin^{-1}((\pm l + y) / r_t) \tag{11. 2}$$

Type 3: $$\theta = \pm\cos^{-1}(\sqrt{((l - r_t)^2 - y^2)/(l - r_t)^2} \tag{11. 3}$$

Type 4: $$\theta = \pm\cos^{-1}(\sqrt{((l+r_t)^2 - y^2)/(l+r_t)^2} \quad (11.4)$$

3.5 Singularity analysis using Mathmatica

Here, we show how to derive the singularity conditions of Eqs. (11. 1) – (11. 4) using Mathematica.

When each rod length l_i is identical, Jacobian Matrix $\mathbf{J}_{e33}$ of (M. 10) is redefined $\mathbf{J}_{e33a}$ as follows:

```
Je33a = Je33 /. {l1 → 1, l2 → 1, l3 → 1, l4 → 1};
```
(M. 11)

At that time, det($\mathbf{J}_{e33}$) is simplified as follows:

```
Det[Je33a] // Simplify
```
(M. 12)

$$-\frac{1}{1^3}\,2\,rt\,(rt\,Sin[\theta[t]] - y[t])\,\sqrt{1^2 - (-rt\,Sin[\theta[t]] + y[t])^2}$$
$$\left(rt\,Sin[2\,\theta[t]] + 2\,Cos[\theta[t]]\,y[t] + 2\,Sin[\theta[t]]\,\sqrt{1^2 - (rt\,Sin[\theta[t]] + y[t])^2}\right). \quad (M.\ 13)$$

(M. 13) is the output of (M. 12) by Mathematica. Equation (10) is solved by Mathematica as follows:

```
Solve[Det[Je33a] == 0, θ[t]] // Simplify
```
(M. 14)

$$\left\{\left\{\theta[t] \to -ArcCos\left[-\frac{\sqrt{(1-rt)^2 - y[t]^2}}{\sqrt{(1-rt)^2}}\right]\right\}, \left\{\theta[t] \to ArcCos\left[-\frac{\sqrt{(1-rt)^2 - y[t]^2}}{\sqrt{(1-rt)^2}}\right]\right\},\right.$$
$$\left\{\theta[t] \to -ArcCos\left[\frac{\sqrt{(1-rt)^2 - y[t]^2}}{\sqrt{(1-rt)^2}}\right]\right\}, \left\{\theta[t] \to ArcCos\left[\frac{\sqrt{(1-rt)^2 - y[t]^2}}{\sqrt{(1-rt)^2}}\right]\right\},$$
$$\left\{\theta[t] \to -ArcCos\left[-\frac{\sqrt{(1+rt)^2 - y[t]^2}}{\sqrt{(1+rt)^2}}\right]\right\}, \left\{\theta[t] \to ArcCos\left[-\frac{\sqrt{(1+rt)^2 - y[t]^2}}{\sqrt{(1+rt)^2}}\right]\right\},$$
$$\left\{\theta[t] \to -ArcCos\left[\frac{\sqrt{(1+rt)^2 - y[t]^2}}{\sqrt{(1+rt)^2}}\right]\right\}, \left\{\theta[t] \to ArcCos\left[\frac{\sqrt{(1+rt)^2 - y[t]^2}}{\sqrt{(1+rt)^2}}\right]\right\},$$
$$\left.\left\{\theta[t] \to ArcSin\left[\frac{y[t]}{rt}\right]\right\}, \left\{\theta[t] \to ArcSin\left[\frac{-1+y[t]}{rt}\right]\right\}, \left\{\theta[t] \to ArcSin\left[\frac{1+y[t]}{rt}\right]\right\}\right\} \quad (M.\ 15)$$

(M. 15) is the output of (M. 14) by Mathematica. Singular points of Eqs. (11. 1) – (11. 4) are given by (M. 15).

3.6 Singularity avoidance by redundantly actuation

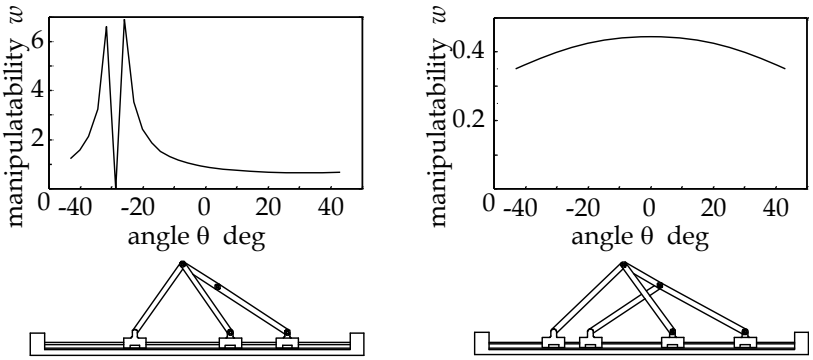
The singularity avoidance of the 3D4M PLM is shown through the definition of manipulability (Yoshikawa, 1985). The derivative kinematics of the PLM is transformed as

$$\Delta\mathbf{c}_3 = \mathbf{J}_{c3}^{-1}\mathbf{J}_{e33}\Delta\mathbf{p}_n = \mathbf{J}_{ce33}\Delta\mathbf{p}_n\,. \quad (12)$$

The manipulability of the PLM is defined as follows:

$$w = \sqrt{\det(\mathbf{J}_{ce33}{}^{T}\mathbf{J}_{ce33})}\ . \tag{13}$$

Numerical calculations of the manipulability of the 3D3M PLM and the 3D4M PLM in correspondence with the rotation angle of the end plate, are shown in Figs. 6(a) and (b). The ratio of each link length l_i and the length of the end effecter $2r_t$ is given as 2:1.
The 3D3M non-redundant PLM becomes the 4th type of singular point given in (11.4) when the angle of the end plate equals 28.7 degrees. Around this angle, the manipulability becomes to zero, as shown in Fig.6 (a). On the other hand, the singular point is avoided by the redundant 3D4M PLM, as shown in Fig.6 (b). This confirms that the redundancy of the 3D4M PLM greatly increases homogeneous manipulability.



(a) 3D3M PLM (b) 3D4M PLM

Fig. 6. Singularity analysis of the parallel link mechanisms

4. Static force analysis of the PLM

4.1 Static force analysis for the non-redundant PLM

Conventional static force analysis only derives the relationship between the generative force and torque of actuators and the **external forces** of the end effecter based on the principle of virtual force. Here, we expand this static force analysis in order to also calculate **internal forces** such as the constraint forces at the joints and the axial forces of the links.
First, the formula for static force is derived when the degree of freedom of the end effecter (n) and the actuators (m) is equivalent. For the sake of convenience, the external forces of the end plate $\mathbf{f}_e$ and the generative forces of the actuators $\mathbf{f}_c$ are expressed by vector forms as follows:

$$\begin{aligned} \mathbf{f}_e &= [f_x, f_y, \tau_\theta]^T \\ \mathbf{f}_c &= [f_{c1}, \ldots, f_{cm}]^T \end{aligned}\ . \tag{14}$$

The relationship between $\mathbf{f}_e$ and $\mathbf{f}_c$ is derived by the principle of virtual forces as

$$\begin{aligned} \mathbf{f}_e &= \mathbf{J}_{cenm}{}^{T}\mathbf{f}_c \\ \mathbf{f}_c &= (\mathbf{J}_{cenm}{}^{T})^{-1}\mathbf{f}_e \end{aligned}\ . \tag{15}$$

As shown in Fig. 7, f_{ni} and f_{li} are defined as quantities of the constraint force at the ith joint and axial force of the ith link, respectively. The values f_{ni} and f_{li} represent the internal forces of the PLM. The unit direction vector $\mathbf{n}$ along the constraint force f_{ni} is orthogonal to the unit direction vector $\mathbf{a}$ along the actuator force f_{ci}. The equilibrium of forces at the ith joint is given as:

$$f_{ci}\mathbf{a} + f_{ni}\mathbf{n} + f_{li}\mathbf{z}_i = \mathbf{0}\,. \tag{16}$$

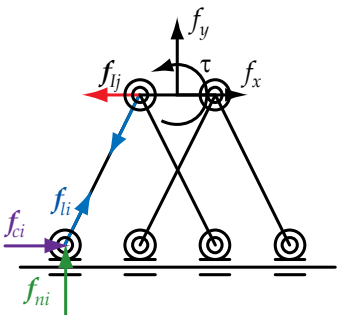


Fig. 7. Internal and external forces of the PLM

By applying the inner product to Eq. (16) with each vector $\mathbf{a}$ and $\mathbf{n}$, with the condition that vectors $\mathbf{a}$ and $\mathbf{n}$ are mutually orthogonal, the following formulas are derived.

$$\begin{aligned} f_{ci} + (\mathbf{z}_i{}^T\mathbf{a})f_{li} &= 0 \\ f_{ni} + (\mathbf{z}_i{}^T\mathbf{n})f_{li} &= 0 \end{aligned} \tag{17}$$

The constraint force f_{ni} and the axial force f_{li} of each link are combined as vector form $\mathbf{f}_n$ and $\mathbf{f}_l$ as follows:

$$\begin{aligned} \mathbf{f}_c &= -diag(\mathbf{z}_1{}^T\mathbf{a},\ldots,\mathbf{z}_m{}^T\mathbf{a})\cdot\mathbf{f}_l = -\mathbf{J}_{cm}\mathbf{f}_l \\ \mathbf{f}_n &= -diag(\mathbf{z}_1{}^T\mathbf{n},\ldots,\mathbf{z}_m{}^T\mathbf{n})\cdot\mathbf{f}_l = -\mathbf{J}_{nm}\mathbf{f}_l \end{aligned}\,. \tag{18}$$

From Eq. (15) to Eq. (18), the internal forces $\mathbf{f}_n$ and $\mathbf{f}_l$ are given by the following equations.

$$\begin{aligned} \mathbf{f}_n &= (\mathbf{J}_{nm}\mathbf{J}_{cm}{}^{-1})\mathbf{f}_c \\ \mathbf{f}_l &= -\mathbf{J}_{cm}{}^{-1}\mathbf{f}_c \end{aligned} \tag{19}$$

4.2 Static force analysis for the redundant PLM

In this section, the static forces for the redundant 3D4M PLM, as shown in Fig. 2 (b), are derived. Instead of the inverse matrix in Eq. (15), a generalized inverse matrix is applied to the calculation of the static force equation.

$$\begin{aligned} \mathbf{f}_c &= (\mathbf{J}_{ce43}{}^T)^+\mathbf{f}_e + (\mathbf{I} - (\mathbf{J}_{ce43}{}^T)^+\mathbf{J}_{ce43}{}^T)\mathbf{k} \\ \mathbf{J}_{ce43} &= \mathbf{J}_{c4}{}^{-1}\mathbf{J}_{e43} \end{aligned} \tag{20}$$

where $^+$ implies the pseudo inverse of a matrix, and $\mathbf{k}$ is a 1×4 arbitrary vector. In the case of the 3D4M PLM, the left side of (20) has 4 degrees of freedom. On the right side of (20), the 1st term has 3 degrees of freedom, which means the 2nd term will have 1 degree of freedom. The null space projection matrix $[\mathbf{I}-(\mathbf{J}_{ce43}{}^T)^+ \mathbf{J}_{ce43}{}^T]$ is 4×4, but its rank is just one. Therefore, the dimension of the 2nd term on the right side of (20) is reduced to

$$\begin{aligned}
&(\mathbf{I}-(\mathbf{J}_{ce43}{}^T)^+\mathbf{J}_{ce43}{}^T)\mathbf{k} = \mathbf{J}_{ce43}{}^T\mathbf{J}_i{}^{-1}\mathbf{f}_i \\
&\mathbf{J}_i = -\begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_3 & \mathbf{z}_4 \end{bmatrix} \\
&\mathbf{f}_i = \begin{bmatrix} \mathbf{f}_{iR} \\ \mathbf{f}_{iL} \end{bmatrix} = f_i[\cos\theta, \sin\theta, -\cos\theta, -\sin\theta]^T
\end{aligned} \tag{21}$$

The dimension is reduced to one by the independent variable f_i in (21). In its physical sense, f_i corresponds to the quantity of internal force acting on the end plate.

In Eq. (21), $\mathbf{f}_{iL}$ and $\mathbf{f}_{iR}$ are the internal forces which act upon the left and right ends of the end plate, respectively. For purposes of convenience, they are combined together as the vector $\mathbf{f}_i$. The direction of internal forces $\mathbf{f}_{iL}$ and $\mathbf{f}_{iR}$ coincides with the direction of the end plate, whose angle is defined by the rotation angle θ of the end effecter. The values $\mathbf{f}_{iL}$ and $\mathbf{f}_{iR}$ have the same quantity f_i, but have opposite directions. The one dimensional internal force that acts along the end plate is explicitly controlled by Eq. (21).

From Eq. (20) and Eq. (21), the generative forces of the actuators $\mathbf{f}_c$ are calculated from the external force of the end effecter $\mathbf{f}_e$ and the internal force of the end plate $\mathbf{f}_i$ as

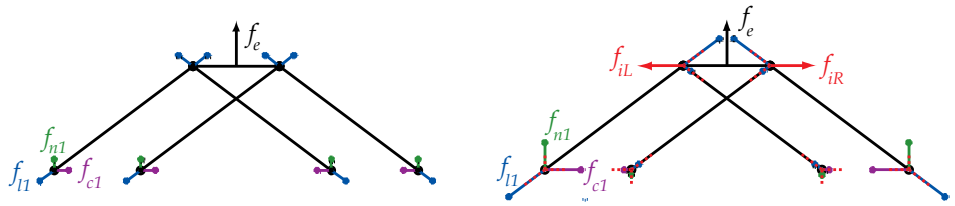
$$\mathbf{f}_c = (\mathbf{J}_{ce43}{}^T)^+\mathbf{f}_e + \mathbf{J}_{ce43}{}^T\mathbf{J}_i{}^{-1}\mathbf{f}_i\,. \tag{22}$$

The internal force of the end plate exerts tensile or compressive stress on the end plate, which helps to diminish joint backlash and increase mechanical rigidity (Adli et al., 1991).

4.3 Numerical simulation of the static force analysis

Numerical simulation software of the static force analysis for the 3D4M PLM has been developed. The external force $\mathbf{f}_e$, the internal force $\mathbf{f}_i$, the generative force of the actuator $\mathbf{f}_c$, the constraint force $\mathbf{f}_n$ and the tensile force of the link $\mathbf{f}_l$ are calculated for an arbitrary position and orientation of the PLM. Examples of static force simulations for the 3D4M PLM are shown in Figs. 8 (a) and (b).

Unit external force $\mathbf{f}_e$ is acting at the central position of the end plate. Figure 8 (a) shows the case in which the internal force of the end plate $\mathbf{f}_i$ is zero. Representations of the generative force of the actuator $\mathbf{f}_c$, the constraint force $\mathbf{f}_n$ and the tensile force of the link $\mathbf{f}_L$ are superimposed on the link mechanism as a solid line with the symbol * at the tip of the vector. Figure 8 (b) shows the case in which the unit internal force of the end plate $\mathbf{f}_i$ was applied. In Fig. 8 (b), the internal force $\mathbf{f}_i$ and additional forces caused by the $\mathbf{f}_i$ at each joint are shown by broken lines. Representations of the generative force of the actuator $\mathbf{f}_c$, the constraint force $\mathbf{f}_n$ and the tensile force of the link $\mathbf{f}_L$ are also superimposed on the link mechanism as a solid line with the symbol * at the tip of the vector. These forces include elements of the internal force $\mathbf{f}_i$,. The situation in Fig. 8 (b) indicates how compressive forces are exerted upon the end plate.



(a) without internal force (b) with internal force

Fig. 8. Numerical simulations about static forces

5. Dynamics of the parallel link mechanism

5.1 Equation of motion of the 3D4M PLM

The positions and orientation of the end plate $\mathbf{p}=[x,\ y,\ \theta]^T$ and the generative forces and torque at the end effecter $\mathbf{f}_e=[f_x,f_y,\tau_\theta]^T$ are considered as generalized positions and generalized forces, respectively. In this formulation, gravity is assumed to be affected by the negative direction of the y-axis. The equation of motion is derived by applying Lagrange's equation.

In particular, if the length, mass and moment of inertia of each link is designed with the same value, the equation of motion of the PLM becomes a simple formula, as follows:

$$\begin{bmatrix} f_x \\ f_y \\ \tau_\theta \end{bmatrix} = \begin{bmatrix} m_{xx} & 0 & 0 \\ 0 & m_{yy}(\mathbf{q}) & m_{y\theta}(\mathbf{q}) \\ 0 & m_{y\theta}(\mathbf{q}) & m_{\theta\theta}(\mathbf{q}) \end{bmatrix} \ddot{\mathbf{p}} + \begin{bmatrix} 0 \\ h_y(\mathbf{q},\dot{\mathbf{q}}) \\ h_\theta(\mathbf{q},\dot{\mathbf{q}}) \end{bmatrix} + \begin{bmatrix} 0 \\ g_y \\ 0 \end{bmatrix} \tag{23}$$

$$\mathbf{q} = [y,\theta]^T$$

Equation (23) represents that the dynamics of the **x** direction of the 3D4M PLM has both **decoupled and constant inertia** characteristics. The dynamics of the y and θ directions are also decoupled in the x direction. The constant of gravitational force affects only the y direction.

5.2 Deriving the equation of motion using Mathematica

Equation (23), the equation of motion of the 3D4M PLM, is derived using Mathematica. The equation of motion is derived by the Lagrange formulation as following steps.

Step 1. Positions (and orientations) of the center of gravity (c.o.g) of mechanical elements - rods, end plate and actuator, - are defined as function of the generalized positions, $x(t)$, $y(t)$ and $\theta(t)$.

Step 2. Velocities (and angular velocities) of the c.o.g. of the mechanical elements are derived as the time derivative of the positions (and orientations) of the c.o.g.

Step 3. Kinetic energies and potential energies of the mechanical elements are calculated.

Step 4. The Lagrangian L is derived as the difference between the total kinetic energy K and the total potential energy U of the mechanics.

$$L = K - U \tag{24}$$

Step 5. Lagurange's equation of motion is derived as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = F_i \,. \tag{25}$$

Where, q_i and F_i are generalized position and generalized force, respectively.

5.2.1 Step 1 - positions and orientations of the mechanical elements

The rod of the 3D4M PLM is designed as an even elongate bar. The c.o.g. of the rod is at the center of the bar. Positions of the c.o.g. of the rods are defined as follows:

```
pcm1 = c1 a + l1 / 2 z1;
pcm2 = c2 a + l2 / 2 z2;
pcm3 = c3 a + l3 / 2 z3;
pcm4 = c4 a + l4 / 2 z4;
```
(M. 16)

Orientations of the rods are alreday defined as (M. 7). Position and orientation of the end plate are given as (M. 1). Positions of the actuatos are already defined as (M. 5). Positions and orientations of the mechanical elements are defined as functions of the generalised positions, $x(t)$, $y(t)$ and $\theta(t)$.

5.2.2 Step 2 - velocities and angular velocities of the mechanical elements

Velocities of the rods are given as time derivative of (M. 16) as follows:

```
Dpcm1Dt = D[pcm1, t] // Simplify;
Dpcm2Dt = D[pcm2, t] // Simplify;
Dpcm3Dt = D[pcm3, t] // Simplify;
Dpcm4Dt = D[pcm4, t] // Simplify;
```
(M. 17)

Angular velocities of the rods are given as time derivative of (M. 7) as follows:

```
Dφ1Dt = D[φ1, t] // Simplify;
Dφ2Dt = D[φ2, t] // Simplify;
Dφ3Dt = D[φ3, t] // Simplify;
Dφ4Dt = D[φ4, t] // Simplify;
```
(M. 18)

Velocities of the actuators are given as time derivative of (M. 6) as follows:

```
Dc1Dt = D[c1, t];
Dc2Dt = D[c2, t];
Dc3Dt = D[c3, t];
Dc4Dt = D[c4, t];
```
(M. 19)

Velocity and angular velocity of the end plate are given as time derivative of (M. 1) as follows:

```
DpDt = D[p, t];
DθDt = D[θ[t], t];
```
(M. 20)

5.2.3 Step 3 - kinetic energies and potential energies of the mechanical elements

Kinetic energy *Ki* of each rod, and *Kt* of the end plate are defined as follows:

```
K1 = 1 / 2 m1 Dpcm1Dt.Dpcm1Dt + 1 / 2 I1 Dφ1Dt^2;
K2 = 1 / 2 m2 Dpcm2Dt.Dpcm2Dt + 1 / 2 I2 Dφ2Dt^2;
K3 = 1 / 2 m3 Dpcm3Dt.Dpcm3Dt + 1 / 2 I3 Dφ3Dt^2;
K4 = 1 / 2 m4 Dpcm4Dt.Dpcm4Dt + 1 / 2 I4 Dφ4Dt^2;
```
(M. 21)

```
Kt = 1 / 2 mt DpDt.DpDt + 1 / 2 It DθDt^2;
```
(M. 22)

Mass and moment of inertia of the *i* th rod are defined as *mi* and *Ii*, respectively. Mass and moment of inertia of the end plate are expressed as *mt* and *It*, respectively.
Sum of kinetic energies of the actuators are expressed as

```
Km = 1 / 2 mm1 Dc1Dt^2 + 1 / 2 mm2 Dc2Dt^2 + 1 / 2 mm3 Dc3Dt^2 +
    1 / 2 mm4 Dc4Dt^2;
```
(M. 23)

Mass of the *i* th actuator is expressed as *mmi*.
Potential energies of rods and end plate are expressed as follows:

```
U1 = m1 g pcm1[[2]];
U2 = m2 g pcm2[[2]];
U3 = m3 g pcm3[[2]];
U4 = m4 g pcm4[[2]];
Ut = mt g p[[2]];
```
(M. 24)

5.2.4 Step 4 - The Lagrangian of the mechanism

The Lagrangian *L* is derived as the difference between the total kinetic energy *K* of (M. 21) – (M. 23) and the total potential energy *U* of (M. 24) as follows:

```
L = K1 + K2 + K3 + K4 + Kt + Km - (U1 + U2 + U3 + U4 + Ut);
```
(M. 25)

5.2.5 Step 5 - Lagurange's equation of motion of the mechanism

Equation of motion of the mechanism is derived by applying Lagrange's equation of Eq. (25) to (M. 25) as follows.

```
S1 = D[D[L, x'[t]], t] - D[L, x[t]];
S2 = D[D[L, y'[t]], t] - D[L, y[t]];
S3 = D[D[L, θ'[t]], t] - D[L, θ[t]];
```
(M. 26)

S1, S2 and S3 are the left parts of Lagrange's equation of Eq. (25).

5.3 Simplify the dynamic characteristics of the 3D4M

As shown in Eq. (23), if the length, mass and moment of inertia of each link is designed with the same value, the equation of motion of the 3D4M PLM becomes a simple formula. Here we derive Eq. (23) using symbolic analysis by Mathematica.
Mass and moment of inertia of each mechanical part is descrived as follows.

```
lnklst = {mm1 → mm, mm2 → mm, mm3 → mm, mm4 → mm, l1 → l,
    l2 → l, l3 → l, l4 → l, m1 → ml, m2 → ml, m3 → ml, m4 → ml,
    I1 → Il, I2 → Il, I3 → Il, I4 → Il};
```
(M. 27)

By applying the list (M. 27) to (M. 26), the equation of motion of the 3D4M PLM becomes

```
S1b = S1 /. lnklst;
S2b = S2 /. lnklst;
S3b = S3 /. lnklst;
```
(M. 28)

Left side of equation of motion along the direction x becomes as follows:

```
S1b // Simplify
```

```
(4 ml + 4 mm + mt) x''[t]
```
(M. 29)

The equation of motion along the direction x is reduced to

$$(4m_l + 4m_m + m_t)\ddot{x}(t) = f_x \tag{26}$$

Equation (26) indicates that the dynamics of the direction **x** of the 3D4M PLM has both **decoupled and constant inertia** characteristics.
We obtain the gravity terms of the equation of motion as

```
Gr = {Coefficient[S1b, g], Coefficient[S2b, g],
  Coefficient[S3b, g]};
Gr // MatrixForm
```

$$\begin{pmatrix} 0 \\ 2\,ml + mt \\ 0 \end{pmatrix}$$
(M. 30)

(M. 30) indicates that the constant of gravitational force $g_y = (2m_l+m_t)g$ affects only to the direction y. Effective mass m_y and m_θ of the direction y in Eq. (23) are derived as

```
myy = Coefficient[S2b, y''[t]] // Simplify
```

```
((l^2 - rt^2 Sin[θ[t]]^2) (4 Il + l^2 (ml + mt) + (4 mm - mt) rt^2 Sin[θ[t]]^2) -
   (4 Il + l^2 (ml - 4 mm + 2 mt) - 2 (4 mm - mt) rt^2 Sin[θ[t]]^2) y[t]^2 +
   (-4 mm + mt) y[t]^4) /
 ((l^2 - rt^2 Sin[θ[t]]^2)^2 - 2 (l^2 + rt^2 Sin[θ[t]]^2) y[t]^2 + y[t]^4)
```
(M. 31)

```
myθ = Coefficient[S2b, θ''[t]] // Simplify
```

$$\frac{(4\,Il + l^2\,(ml + 4\,mm))\,rt^2\,\mathrm{Sin}[2\theta[t]]\,y[t]}{(l^2 - rt^2\,\mathrm{Sin}[\theta[t]]^2)^2 - 2\,(l^2 + rt^2\,\mathrm{Sin}[\theta[t]]^2)\,y[t]^2 + y[t]^4}$$
(M. 32)

Coeffcients of Coriolis and centrifigual forces of the durection y in Eq. (23) are reduced to

```
hy1 = Coefficient[S2b, y'[t] θ'[t]] // Simplify
```

```
((4 Il + l^2 (ml + 4 mm)) rt^2 Sin[2 θ[t]] ((2 l^2 - rt^2 + rt^2 Cos[2 θ[t]])^2 +
     4 (2 l^2 + rt^2 - rt^2 Cos[2 θ[t]]) y[t]^2 - 12 y[t]^4)) /
 (4 ((l^2 - rt^2 Sin[θ[t]]^2)^2 - 2 (l^2 + rt^2 Sin[θ[t]]^2) y[t]^2 + y[t]^4)^2)
```
(M. 33)

```
hy2 = Coefficient[S2b, y'[t]^2] // Simplify
```

$$\begin{aligned}&\left(2\left(4\,Il+l^2\,(ml+4\,mm)\right)\right)y[t]\\&\quad\left(8\,l^4+8\,l^2\,rt^2-9\,rt^4+\left(-8\,l^2\,rt^2+12\,rt^4\right)\text{Cos}[2\,\theta[t]]-3\,rt^4\right.\\&\qquad\left.\text{Cos}[4\,\theta[t]]-8\left(2\,l^2-rt^2+rt^2\,\text{Cos}[2\,\theta[t]]\right)y[t]^2+8\,y[t]^4\right)\Big)\Big/\\&\left(\left(2\,l^2-rt^2+rt^2\,\text{Cos}[2\,\theta[t]]\right)^2-4\left(2\,l^2+rt^2-rt^2\,\text{Cos}[2\,\theta[t]]\right)y[t]^2+\right.\\&\qquad\left.4\,y[t]^4\right)^2\end{aligned} \tag{M. 34}$$

```
hy3 = Coefficient[S2b, θ'[t]^2] // Simplify
```

$$\begin{aligned}&\left(2\left(4\,Il+l^2\,(ml+4\,mm)\right)\right)rt^2\,y[t]\\&\quad\left(-16\,l^4+56\,l^2\,rt^2-26\,rt^4+\left(48\,l^4-64\,l^2\,rt^2+33\,rt^4\right)\text{Cos}[2\,\theta[t]]+\right.\\&\qquad 8\,l^2\,rt^2\,\text{Cos}[4\,\theta[t]]-6\,rt^4\,\text{Cos}[4\,\theta[t]]-rt^4\,\text{Cos}[6\,\theta[t]]-\\&\qquad 8\left(-4\,l^2-7\,rt^2+4\left(3\,l^2+2\,rt^2\right)\text{Cos}[2\,\theta[t]]-rt^2\,\text{Cos}[4\,\theta[t]]\right)\\&\qquad\left.y[t]^2+16\,(-1+3\,\text{Cos}[2\,\theta[t]])\,y[t]^4\right)\Big)\Big/\\&\left(2\left(2\,l^2-rt^2+rt^2\,\text{Cos}[2\,\theta[t]]\right)^2-\right.\\&\qquad\left.8\left(2\,l^2+rt^2-rt^2\,\text{Cos}[2\,\theta[t]]\right)y[t]^2+8\,y[t]^4\right)^2\end{aligned} \tag{M. 35}$$

Then, $h_y(\mathbf{q},\dot{\mathbf{q}})$, Coriolis and centrifugal forces of the durection y in Eq. (23), is given as

$$h_y(\mathbf{q},\dot{\mathbf{q}}) = h_{y1}\dot{y}\dot{\theta} + h_{y2}\dot{y}^2 + h_{y3}\dot{\theta}^2 . \tag{27}$$

Effective moment of inertia $m_{\theta\theta}$ and Corioli and centrifugal moment h_θ of the direction θ are given as the same mannar of (M. 31) – (M. 35).

6. Numerical simulations of the dynamics

Numerical simulations of the dynamics of the 3D4M PLM has been tested using Mathematica. Values of the kinematic parameters of the 3D4M PLM are set as Table 1. These values are given from the prototype of the 3D4M PLM (Harada & Nagase, 2010) as show in Fig. 9.
Values of the kinematic parameters are given follows:

```
mmN = 0.375;(*mass of the actuator [kg]*)
lN = 0.140; (*length of the link [m]*)
rtN = 0.05;(*half length of the end plate [m]*)
mlN = 0.026; (*mass of the link [kg]*)
mtN = 0.132;(*mass of the end plate [kg]*)
(*moment of inertia of the link [kgm^2]*)
IlN = 9.84×10^4×10^-3×10^-6;
(*moment of inertia of the end plate [kgm^2]*)
ItN = 8.54×10^5×10^-3×10 ^-6;
```
(M. 36)

```
lnklstN = {mm1 → mmN, mm2 → mmN, mm3 → mmN, mm4 → mmN, l1 → lN,
   l2 → lN, l3 → lN, l4 → lN, m1 → mlN, m2 → mlN, m3 → mlN,
   m4 → mlN, I1 → IlN, I2 → IlN, I3 → IlN, I4 → IlN, rt → rtN,
   mt → mtN, It → ItN, g → 9.8};
```
(M. 37)

Link	Length [m]	0.140
	Mass [kg]	0.026
	Moment of inertia [kgm^2]	9.84×10^{-5}
End plate	Length [m]	0.100
	Mass [kg]	0.132
	Moment of inertia [kg m^2]	8.54×10^{-4}
Actuator	Total mass [kg]	0.375

Table 1. Specifications of the mechanical parts of the prototype

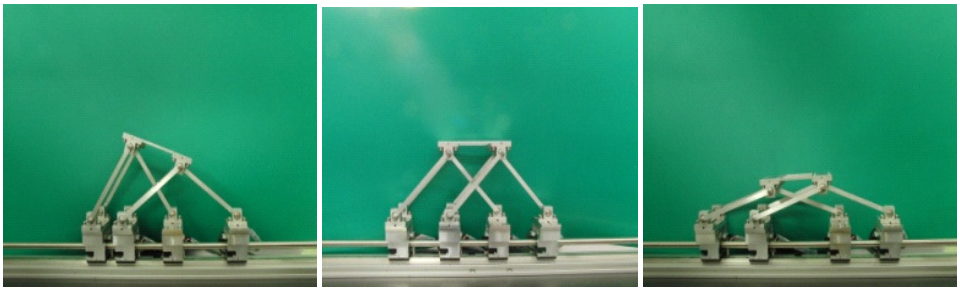


Fig. 9. Front views of the prototype

Equations of motion of Eq. (23) are nonlinear simultaneous ordinary differential equations about $x(t)$, $y(t)$ and $\theta(t)$ with respect to time t. They are numerically solved by the operator "NDSolve" in Mathematica.
We show an example of numerical simulation with initial conditions of the generalized positions and velocities are set as Table 2.

positions		velocities	
$x(0)$ [m]	0.0	$\dot{x}(0)$ [m/s]	0.0
$y(0)$ [m]	0.08	$\dot{y}(0)$ [m/s]	0.0
$\theta(0)$ [deg]	20.0	$\dot{\theta}(0)$ [rad/s]	0.0

Table 2. Initial conditions of the numerical simulation

Constant generalized forces and torque are applied as Table 3.

forces and torque	
f_x [N]	0.2
f_y [N]	$(2m_l+m_t)g$
τ_θ [Nm]	0.0

Table 3. Generalized forces and torque of the numerical simulation 1

Constant force $f_y=(2m_l+m_t)g$ compensates the gravitational effect along the direction y as shown in (M. 30).
Responses of the generalized positions with the conditions of Table 2 & 3 from time t=0 [s] to t=2.0 [s] are numerically calculated by Mathematica as follows:

```
ts = 2.0;

ans = NDSolve[{(S1 == 0.2) /. lnklstN, (S2 == g (2 m1 + mt)) /. lnklstN,
    (S3 == 0.0) /. lnklstN, x[0] == 0.0, y[0] == 0.08,
    θ[0] == -20 Degree, x'[0] == 0, y'[0] == 0, θ'[0] == 0}, {x, y, θ},
   {t, 0, ts}]

{{x → InterpolatingFunction[{{0., 2.}}, <>],
  y → InterpolatingFunction[{{0., 2.}}, <>],
  θ → InterpolatingFunction[{{0., 2.}}, <>]}}
```
(M. 38)

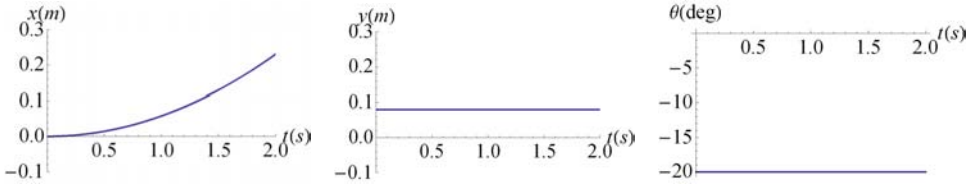
The results are graphically shown by using Mathematica with the following codes.

```
Plot[x[t] /. lnklstN /. ans, {t, 0, ts}, PlotStyle → Thick,
 LabelStyle → Large, PlotRange → {{0, 2}, {-0.1, 0.3}},
 AxesLabel → {t [s], x [m]}]
```
(M. 39)

```
Plot[y[t] /. lnklstN /. ans, {t, 0, ts}, PlotStyle → Thick,
 LabelStyle → Large, PlotRange → {{0, 2}, {-0.1, 0.3}},
 AxesLabel → {t [s], y [m]}]
```
(M. 40)

```
Plot[Evaluate[θ[t] 180 / π /. lnklstN /. ans],
 {t, 0, ts}, PlotStyle → Thick, PlotRange → All,
 LabelStyle → Large, AxesLabel → {t [s], θ [deg]}]
```
(M. 41)

Outputs of (M. 39) and (M. 41) are shown as Fig. 10 (a), (b) and (c), respectively.



(a) x (b) y (c) θ

Fig. 10. Results of the numerical simulation 1 (generalized positions)

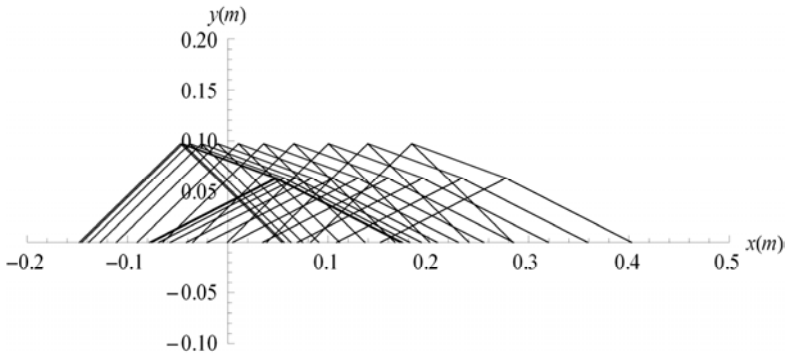


Fig. 11. Results of the numerical simulation 1 (skeletons)

Motions of the mechanism are graphically shown by the following code of Mathematica.

```
Graphics[
 Table[
  {Line[{{-rt Cos[θ[t]] + x[t] - Sqrt[l1^2 - (-rt Sin[θ[t]] + y[t])^2],
       0}, {-rt Cos[θ[t]] + x[t], -rt Sin[θ[t]] + y[t]}}],
   Line[{{-rt Cos[θ[t]] + x[t] + Sqrt[l2^2 - (-rt Sin[θ[t]] + y[t])^2],
       0}, {-rt Cos[θ[t]] + x[t], -rt Sin[θ[t]] + y[t]}}],
   Line[{{rt Cos[θ[t]] + x[t] - Sqrt[l3^2 - (rt Sin[θ[t]] + y[t])^2],
       0}, {rt Cos[θ[t]] + x[t], rt Sin[θ[t]] + y[t]}}],
   Line[{{rt Cos[θ[t]] + x[t] + Sqrt[l4^2 - (rt Sin[θ[t]] + y[t])^2],
       0}, {rt Cos[θ[t]] + x[t], rt Sin[θ[t]] + y[t]}}],
   Line[{{-rt Cos[θ[t]] + x[t], -rt Sin[θ[t]] + y[t]},
     {rt Cos[θ[t]] + x[t], rt Sin[θ[t]] + y[t]}}]} /. lnklstN /.
  ans, {t, 0, ts, 0.2}], AspectRatio → Automatic,
 PlotRange → {{-0.2, 0.5}, {-0.1, 0.2}}, Axes → True,
 AxesLabel → {x [m], y [m]}]
```

(M. 42)

Output of (M. 41) is shown as Fig. 11.
As shown in Fig. 10 and Fig. 11, the 3D4M PLM generated a constant motion of acceleration in the direction x while maintaining its initial configuration. The characteristiocs of the constant and decoupled dynamics of the 3D4M PLM along the direction x has been confirmed by this simulation.
We show another example of numerical simulation with initial conditions of the generalized positions and velocities are set same as Table 2, but other types of constant generalized forces and torque are applied as Table 4.

forces and torque	
f_x [N]	0.0
f_y [N]	$(2m_l+m_t)g$
τ_θ [Nm]	0.005

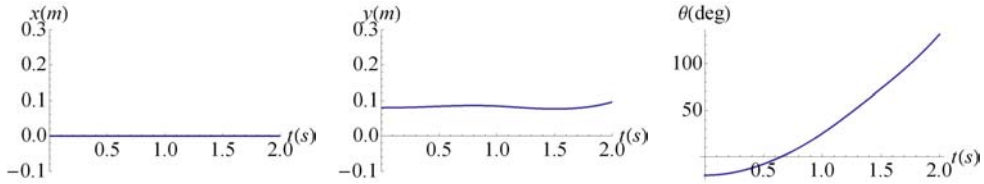
Table 4. Generalized forces and torque of the numerical simulation 2

Code for solving the nonlinear simultaneous ordinary differential equations is given as

```
ans =
 NDSolve[{(S1 == 0) /. lnklstN,
   (S2 == g (2 m1 + mt)) /. lnklstN, (S3 == 0.005) /. lnklstN,
   x[0] == 0.0, y[0] == 0.08, θ[0] == -20 Degree, x'[0] == 0,
   y'[0] == 0, θ'[0] == 0}, {x, y, θ}, {t, 0, ts}]
```

(M. 43)

Results of the simulation are shown in Fig. 12 and Fig. 13.



(a) x (b) y (c) θ

Fig. 12. Results of the numerical simulation 2 (generalized positions)

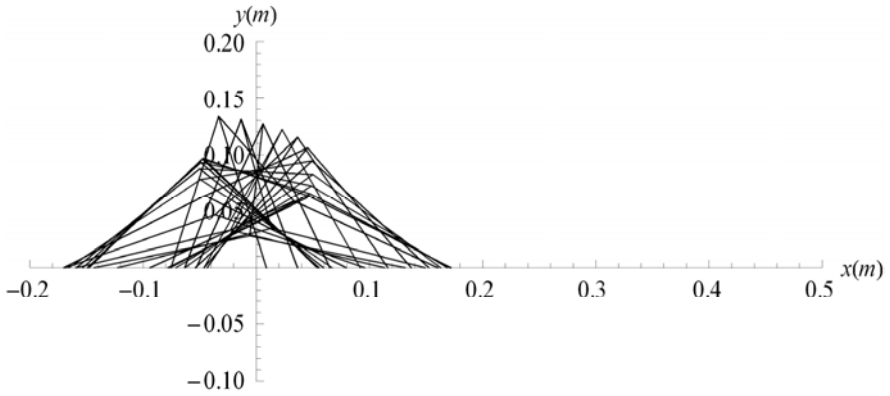


Fig. 13. Results of the numerical simulation 2 (skeletons)

As shown in Fig 12 and Fig. 13, the 3D4M PLM generated a constant motion of acceleration in the θ direction while maintaining its initial horisontal (x) position. As shown if Fig. 12 (b), position of the direction y slightly moves because of the dynamical coupling between the directions y and θ.

7. Conclusion

A novel redundantly actuated planar parallel link mechanism using multi drive linear motors has been proposed. It expands the range of motion, while retaining the advantages of rigid mechanism and precise positioning. The kinetostatics and dynamics characteristics of the mechanism have been analyzed using symbolic mathematical analysis and numerical simulations.

1. Kinematic equations, forward kinematics and derivative kinematics of the mechanism have been derived. The derived equations have been symbolically programmed using Mathematica.
2. Singularity and static forces of the mechanism have been analyzed using Mathematica. The conditions of the singularity have been symbolically solved using Mathematica.
3. The equations of motion of the mechanism have been symbolically derived using Mathematica by Lagrange's formulation.

4. If the length, mass and moment of inertia of each link were designed with the same value, the equations of motion of the mechanism have been showed the decoupled and constant inertia characteristics in the horizontal direction.
5. Examples have been introduced for solving the equations of motion numerically using Mathematica. The characteristics of the decoupled and constant inertia of the mechanism has been confirmed by the simulations.

8. Acknowledgment

This work was supported in part by the Electro-Mechanic Technology Advancing Foundation and Mazak Foundation, Japan.

9. References

Adli, M. A.; Nagai K.; Miyataka, K. & Hanafusa, H. (1991). Analysis of Internal Force Effect in Parallel Manipulators, *Transactions of the Society of Instrument and Control Engineers*, Vo. 27, No. 11, pp. 1266-1273, ISSN 0453-4654

Arai, T.; Stoughton, R.; Homma, K.; Adachi, H.; Nakamura, T. & Nakashima, K. (1991). Development of a parallel link manipulator," Proceedings of 5th International Conference on Advanced Robotics, Vol. 1, pp. 839-844, ISBN 0-7803-0078-5, Pisa , Italy, 19-22 Jun 1991

Chiu, Y. J. & Perg, M. H. (2003). Self-calibration of a general hexapod manipulator using cylindrical constraints, *International Journal of Machine Tools & Manufacture*, Vol. 43, Issue 10, pp. 1051-1066, ISSN 0890-6955

Gosselin, C. & Angeles, J. (1990). Singularity analysis of closed-loop kinematic chains," *IEEE Transactions on Robotics and Automation*, Vol. 6, No.3, pp. 281-290, ISSN 1042-296X

Harada, T. & Nagase M (2009). Configurations and Mathematical Models of Parallel Link Mechanisms Using Multi Drive Linear Motors, *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1974–1979, ISBN 978-1-4244-3803-7, St. Louis, MO, USA, 10-15 Oct. 2009

Harada, T. & Nagase M (2010). Impedance Control of a Redundantly Actuated 3-DOF Planar Parallel Link Mechanism Using Direct Drive Linear Motors, *Proceedings of the 2010 IEEE International Conference on Robotics and Biomimetics*, pp. 501–506, ISBN 978-1-4244-9317-3, Tianjin, China, 14-18 Dec. 2010

Hogan, N. (1985). Impedance Control: An Approach to Manipulation: Part I-III, *Transactions of the ASME, Journal of Dynamic Systems, Measurement, and Control*, Vol. 107, No. 1, pp. 1–24, ISSN 0022-0434

Honegger, M.; Codourey, A. & Burdet E. (1997). Adaptive Control of the Hexaglide, A 6 DOF Parallel Manipulator, *Proceedings of the 1997 IEEE International Conference on Robotics and Automations*, pp. 21–28, ISBN 0-7803-3612-7, Albuquerque, NM , USA, 20-25 Apr 1997

In, W.; Lee, S.; Jeong, J. I. & Kim, J. (2008). Design of a Planar-Type High Speed Parallel Mechanism Positioning Platform with the Capability of 180 Degrees Orientation, *CIRP Annals – Manufacturing Technology*, Vol. 58, Issue 1, pp. 421–424, ISSN 0007-8056

Kim, J.; Cho, Y. M.; Park, F. C. & Lee, J. M. (2003). Design of a Parallel Mechanism Platform for Simulating Six Degrees-of-freedom General Motion Including Continuous 360-

degree Spin, *CIRP Annals - Manufacturing Technology*, Vol. 52, Issue 1, pp. 347–350, ISSN 0007-8056

Leong, J. I.; Kang, D.; Cho, Y. M & Kim J. (2004), Kinematic Calibration for Redundantly Actuated Parallel Mechanisms, *Transactions of the ASME, Journal of Mechanical Design*, Vol. 126, Issue 2, pp. 307–319 , ISSN 1050-0472

Liu, X. J.; Wang, J. & Pritschow, G. (2004). A New Family of Spatial 3-DOF Fully Parallel Manipulators with High Rotational Capability, *Mechanism and Machine Theory*, pp. 475–494, ISSN 0094-114X

Liu, X. J.; Guan, L. & Wang J. (2007), Kinematics and Closed Optimal Design of a Kind of PRRRP Parallel Manipulator, *Transactions of the ASME, Journal of Mechanical Design*, Vol. 129, Issue 5, pp. 558–563, ISSN 1050-0472

Marquet, F.; Krut, S.; Company, O. & Pierrot, F. (2001). ARCHI: A New Redundant Parallel Mechanism - Modeling, Control and First Results, *Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 183-188, ISBN 0-7803-6612-3, Maui, HI, USA, 29 October - 03 November 2001

Merlet, J. P. (1989). Singular configurations of parallel manipulators and grassmann geometry, *The International Journal of Robotics Research*, Vol. 8, No. 5, pp. 45-56, ISSN 0278-3649

Merlet, J. P. (1996). Direct kinematics of planar parallel manipulators, *Proceedings of 1996 IEEE International Conference on Robotics and Automation,* Vol. 4, pp. 3744 - 3749, ISBN 0-7803-2988-0, Minneapolis, MN , USA, 22-28 Apr 1996

Merlet, J. P. (2006). *Parallel Robots* (Second Edition), Springer, ISBN 978-1-4020-4132-7, Netherland

Milutinovic, D. S.; Glavonjic, M.; Kvrgic, V. & Zivanovic S. (2005). A New 3-DOF Spatial Parallel Mechanism for Milling Machines with Long X Travel, *CIRP Annals – Manufacturing Technology,* Vol. 54, Issue 1, pp. 345–348. ISSN 0007-8056

Oiwa, T. (1997). New coordinate measuring machine featuring a parallel mechanism, *International journal of the Japan Society for Precision Engineering*, Vol. 31, No. 3, pp. 232-233, ISSN 0916-782X

Stewart, D. (1966). A platform with six degrees of freedom, *Proceedings of the Institution of Mechanical Engineers*, Vol. 180, No. 15, pp. 371-386, ISSN 0020-3483

Wang, J. & Liu, X. J. (2008). *Parallel Robots: Recent Advances in Research and Application,* Nova Science Pub Inc., ISBN 978-1-60456-859-2, New York

Wang, J.; Wu, J.; Li, T. & Liu, X. (2008). Workspace and Singularity Analysis of a 3-DOF Planar Parallel Manipulator with Actuation Redundancy, *Robotica*, Vol. 27, pp. 51–57, ISSN 0263-5747

Weck, M. (2002). Parallel kinematic machine tools-current state and future potentials, *CIRP Annals - Manufacturing Technology*, Vol. 51, Issue 2, pp. 671-683, ISSN 0007-8056

Yoshikawa, T. (1985). Manipulability of Robotic Mechanisms, *The International Journal of Robotic Research*, Vol. 4, No. 2, pp. 3-9, ISSN 0278-3649

Zhang, J. (2008), A Family of Parallel Manipulators with Mobile Platform Rotating Continuously, in *Parallel Robotics: Advances in Research and Application*, J. Wang, and X. J. Liu, Ed., New York: Nova Science Publishers, Inc., 2008, pp. 67–80.

Zhuang, H. & Liu, L. (1998). Determination of number of independent parameters for the self-calibration of parallel-link mechanisms, *Proceedings of ASME*, MED-vol.8, pp. 699-703, 1998.

19

Dynamics and Control for a Novel One-Legged Hopping Robot in Stance Phase

Guang-Ping He[1] and Zhi-Yong Geng[2]
[1]*School of Mechanical and Electrical Engineering, North China University of Technology,*
[2]*Sate Key Laboratory for Turbulence and Complex Systems, Peking University,*
Beijing,
P.R. China

1. Introduction

Legged locomotion systems are a class of biological robots by supporting on the ground with discrete points, and can traverse nature terrain in large range. This kind of robots is a class of important research objects in robotics community in a long time. In the early time, most of the legged robot systems are static balance locomotion systems, which show a slow moving speed and high requirement in energy dissipation. During the past twenty years, many scholars were interested in dynamic legged robots for improving the performance of the systems.

The one-legged hopping robots are a typical and the simplest systems of the dynamic legged robots, and can be found in many literatures, such as (Raibert,1986; Francois & Samson, 1998; Gregorio et al, 1997; Ahmadi & Buehler, 1997,2006; Ahmadi et al, 2007; Vakasi et al, 1991; Lapshin, 1992; Papadopoulos & Cherouvim, 2004; Hyon & Emura,2002; Zeglin,1999; Guang-Ping H. & Zhi-Yong G, 2008 etc.). As it pointed out that in (Papadopoulos & Cherouvim, 2004), most of researches about the one-legged hopping robots were limited to the systems with *Spring Loaded Inverted Pendulum* (SLIP) model, which is composed of a point mass attached on a telescopic spring that is free to rotate around its point of contact with the ground. The SLIP model hopping systems can be stabilized without much effort in control design since the simple decoupling dynamics (Raibert et al, 1984; Raibert,1986; Francois & Samson, 1998; Gregorio et al, 1997; Ahmadi & Buehler, 1997, 2006; Ahmadi et al, 2007; Vakasi et al, 1991; Lapshin, 1992; Papadopoulos & Cherouvim, 2004; Hyon & Emura,2002, 2004; Zeglin,1999; Hodgins & Raibert, 1990; Hyon et al, 2004). It was shown that this class of hopping robots not only could stably hop but also realize some acrobatic motion such as somersaults (Raibert et al, 1984; Hodgins & Raibert, 1990). Nevertheless, the main limitations of the SLIP model systems are that the mechanical systems lose the biological characteristics and the leg of the robot commonly has translational telescopic joint. The translational telescopic legs are generally actuated by pneumatic or hydraulic actuators, and show small motion range. The pumping station of the hydraulic systems is not easy to be embedded into the robots such that the robots could not move in large range. Though (Gregorio et al, 1997; Ahmadi & Buehler, 1997, 2006; Ahmadi et al, 2007) realized the hopping control on an electrically actuated experimental robot, the

theory analysis and experimental methods have little help for a legged hopping system with fully articulated joints.

For overcoming the limitations of the hopping systems with SLIP model, several scholars studied the hopping systems with non-SLIP model (Zeglin, 1991; Hyon & Mita, 2002; Takahashi et al, 2006; Guang-Ping H. & Zhi-Yong G, 2008, 2009a, 2009b, 2010), such as Uniroo (Zeglin, 1991) and Kenken (Hyon & Mita, 2002). The hopping systems with non-SLIP model generally have more biological characteristics. Thus the mechanism is more complex, and control design for this class system is more intractable since the highly nonlinear and strong coupling in dynamics. For instance, the Uniroo and Kenken are experimental robots, whereas the Uniroo employed the control method for SLIP systems and the best experimental results is 40 times jumps before falling down (Zeglin, 1991), the Kenken was controlled based on accurate simulations of the dynamics, and both of the two robots are actuated by hydraulic systems.

For the sake of reducing the energy dissipation caused by impact, the one-legged hopping systems commonly has small foot such that the robot only contacts with ground on a point. On the assumption that the foot of the robot has no slip, the point contacting with the ground can be regarded as a passive rotational joint. Thus the one-legged hopping robots are generally underactuated mechanical systems. In the field of nonlinear control, the underactuated mechanical systems are a class of interesting nonlinear systems that has been given many attentions in recent years. The benchmark systems of them include the Cart-pole (Graichen, 2007), Acrobot (Lai, 2008), Pendubot (Spong & Block, 1995), Plate-Ball (Oriolo & Vendittelli, 2005), underactuated planar manipulators (Arai et al, 1998), and underactuated surface vessel (Reyhanoglu, 1997) etc. It had been proved that the underactuated mechanical systems are second-order nonholonomic systems in gravitational circumstance if the passive generalized coordinates are not cyclic (Oriolo & Nakamura, 1991), and the nonholonomic systems cannot be stabilized by smooth time-invariant state feedback (Kolmanivsky & McClamroch, 1995). Thus the control methods presented in the literatures introduced non-continuous feedback, time-varying feedback, or the combination of the two classes method, and the control plants are mainly limited to the nonholonomic systems with special differential geometric or differential algebraic properties, such as differentially flat (Nieuwstadt & Murray, 1995; Guang-Ping H. & Zhi-Yong G, 2008, 2009a) or nilpotent (Murray, 1994; Guang-Ping H. & Zhi-Yong G, 2009b) systems. By nonlinear coordinates and inputs (control) changes, the differentially flat nonholonomic systems can be transformed into high order linear systems, the nilpotent systems can be transformed into chained form system under certain conditions (Murray et al, 1993). For the second-order nonholonomic underactuated mechanical systems without these properties, the control problem was not discussed adequately in the nonlinear control field.

The hopping robots with SLIP model are generally second-order nonholonomic underactuated systems because of the small foot, whereas they can be stablized to the periodical hopping orbits by smooth time-invariant state feedback (Raibert,1986; Gregorio et al，1997; Ahmadi & Buehler, 1997, 2006; Ahmadi et al, 2007; Hodgins & Raibert, 1990). The reason is that the special mechanical structure satisfies some conditions: ① the mass as well as the inertia of the leg are far less than them of the total system, swing the leg does not cause large orientation errors of the body; ② the position of the mass center(MC) of the robot is coincident to the hip join, then most of the nonlinear force of the dynamics disappears; ③ the robot has linearly telescopic leg, the telescopic motion of the leg is approximately decoupled from the rotational motion of the systems. These dynamic

properties ensure that the SLIP model robot can be stabilized by controlling partial variables without considering the nonholonomic constraints. Nevertheless, a one-legged hopping system with non-SLIP model shows complex nonlinear dynamics, the nonholonomic constraints of the system cannot be ignored. So far there does not have systemic method for designing the control for this class of hopping robots.

Since the obstacle for controlling a general underactuated mechanical system, it is important to investigate the dynamic synthesis problem such that the dynamics of the underactuated systems can be effectively simplified while the underactuated systems holds the controllability, high energy-efficiency, and dexterous mobility. This is helpful for inventing a new underactuated mechanical system with feasible control method, investigating the new applications of underactuated mechanical systems, or simplifying the control problem of the existing underactuated mechanical systems. For instance, Franch investigated the design method of differentially flat planar space robots (Franch et al, 2003), Agrawal & Fattch studied the dynamics synthesis for planar biped robots (Agrawal & Fattch, 2006). In this paper, we propose a novel biological mechanism for the one-legged planar hopping robots on the basis of the dynamics synthesizing. The mechanism is similar to the skeleton of kangaroos, and the dynamics of the mechanism possesses kinetic symmetry with respect to the passive joint variable, then the nonlinear dynamics of the novel mechanism can be transformed into the so-called *strict feedback normal form*, which can be potentially stabilized by backstepping technique. Thus the novel mechanism can be used to compare with the SLIP model robot for more adequately understanding the dynamic balance principle, high energy-efficiency, and dexterous mobility of kangaroos.

In this chapter, section 2 introduces the novel mechanism, and the dynamics of it is presented in section 3. In section 4, the proposition that confirms the nonlinear dynamics can be transformed into the strict feedback normal form is proved. Then a sliding mode backstepping control is introduced in section 5 and the exponential stability is also proved. The motion planning method for the hopping system in stance phase is presented in section 6. The feasibility of the mechanism and the stability of the control are verified by some numerical simulations in section 7.

2. The new mechanism for one-legged planar hopping robots

Fig.1 shows the new mechanism for designing a biological one-legged hopping robot. The robot mechanism has four rigid bodies, of which the shank, thigh, body and tail have length, mass and inertia with respect to their MC are l_i, m_i and I_i, $i=1,2,3,4$ respectively. Suppose the MC of every link deviates from the joint that nears to the ground, and the distance deviating from the joint is l_{ci}, $i=1,2,3,4$. The angle of shank is denoted as θ_1, the keen joint, hip joint, and tail joint are θ_2, θ_3, and θ_4 respectively. To simplify the dynamics of the system, let $l_{c3}=0$, and the mechanism is designed to joint the body, the thigh, and the tail at the same axis. The turning of the tail joint and the keen joint is synchronous (by parallel four bars mechanism or synchronous belt) with constant phase angle α_0. For improving the energy efficiency, a torsional linear spring with stiffness k is paralleled in the keen joint. The generalized coordinates of the mechanism in stance phase can be defined as $\mathbf{q}=[x_0, y_0, \theta_1, \theta_2, \theta_3]^{\mathrm{T}}$, of which θ_2, θ_3 are actuated joints, θ_1 is a passive joint, and (x_0, y_0) are constants in stance phase.

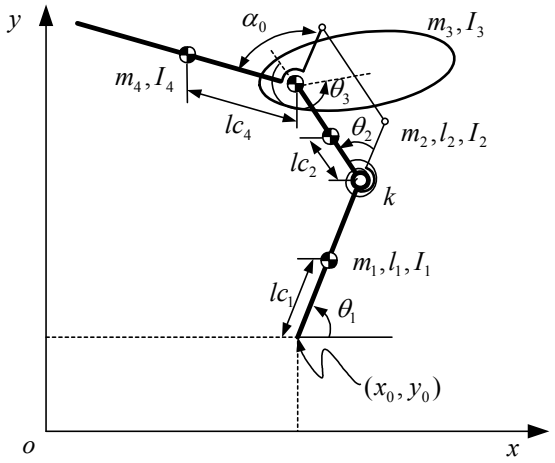


Fig. 1. A novel mechanism for one-legged hopping robot

The characteristics of the hopping robot mechanism can be summarized as follows:

a. The mechanism is underactuated because of the passive joint θ_1. In the following section, it is shown that the single passive generalized coordinate θ_1 does not appear in the kinetic energy but appears in the potential energy of the robot system. Thus the robot system is similar to an Acrobot system (Lai et al, 2008; Spong, 1995) in dynamics. This property in dynamics makes the nonlinear dynamics of the hopping robot can be transformed into the strict feedback normal form, which belongs to a special class of nonlinear system that can be stabilized by backstepping control.

b. The turning of the keen joint and the tail joint is synchronous, so the actuator of the keen joint can be installed on the tail. This special design can reduce the mass as well as the inertia of the leg. As proved by (Ahmadi & Buehler, 2006), a leg with less mass and inertia is helpful for improving the energy-efficiency of the hopping robot systems.

c. The MC of the body is coincident with the hip joint, i.e. $l_{c3} = 0$, such that the Coriolis and centrifugal forces about l_{c3} disappear, then the dynamics of the robot system is considerably simplified. In section 4, it will be shown that this property in dynamics makes the novel hopping system has the analytical coordinate transformations, which is necessary for designing the nonlinear controller.

d. The robot mechanism has articulate keen joint, so the leg can provide a larger clearance from the ground than it provided by a linearly telescopic leg in continuous hopping. This is beneficial for leaping over different size obstacles with the same energy cost.

3. The dynamics of the mechanism

In stance phase, the foot of the robot is contacting with the ground, thus the coordinates (x_0, y_0) are constants, and then the generalized coordinates of the hopping system shown in Fig.1 is reduced to $\mathbf{q}_1 = [\theta_1 \quad \theta_2 \quad \theta_3]^T$. If $L = T - U$ denotes the Lagrangian of the hopping system, where T is kinetic energy, U is potential energy of the system, then they can be given as form

$$T = \frac{1}{2}\sum_{i=1}^{i=4} m_i(\dot{x}_{ci}^2 + \dot{y}_{ci}^2) + \frac{1}{2}\Big[I_1\dot{\theta}_1^2 + I_2(\dot{\theta}_1 + \dot{\theta}_2)^2 + I_3(\dot{\theta}_1 + \dot{\theta}_2 + \dot{\theta}_3)^2 + I_4\dot{\theta}_1{}^2 \Big],$$

$$U = g\sum_{i=1}^{4}(m_i y_{ci}) + \frac{1}{2}k(\theta_2 - \theta_{20})^2,$$

where $(x_{ci}, y_{ci}), i = 1,2,3,4$ is the position of MC of every bodies of the mechanism, and can be written as

$$x_{c1} = x_0 + l_{c1}\cos\theta_1,$$

$$y_{c1} = y_0 + l_{c1}\sin\theta_1,$$

$$x_{c2} = x_0 + l_1\cos\theta_1 + l_{c2}\cos(\theta_1 + \theta_2),$$

$$y_{c2} = y_0 + l_1\sin\theta_1 + l_{c2}\sin(\theta_1 + \theta_2),$$

$$x_{c3} = x_0 + l_1\cos\theta_1 + l_2\cos(\theta_1 + \theta_2),$$

$$y_{c3} = y_0 + l_1\sin\theta_1 + l_2\sin(\theta_1 + \theta_2),$$

$$x_{c4} = x_0 + l_1\cos\theta_1 + l_2\cos(\theta_1 + \theta_2) + l_{c4}\cos(\theta_1 + \alpha_0),$$

$$y_{c4} = y_0 + l_1\sin\theta_1 + l_2\sin(\theta_1 + \theta_2) + l_{c4}\sin(\theta_1 + \alpha_0).$$

g is the gravitational acceleration, k is the stiffness of the spring in keen joint, θ_{20} is the position of keen joint with spring free.
The Euler-Lagrange dynamics of the hopping system in stance phase can be written as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = \tau_i, i = 1,2,3 \tag{1}$$

For more clearly, the dynamics (1) has form

$$\begin{aligned} &\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta}_1}\right) + \frac{\partial U}{\partial \theta_1} = 0 \\ &\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta}_2}\right) - \frac{\partial T}{\partial \theta_2} + \frac{\partial U}{\partial \theta_2} = \tau_2 \\ &\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta}_3}\right) = \tau_3 \end{aligned} \tag{2}$$

where $\tau_1 = 0$, $\partial T/\partial\theta_1 = 0$ and $\partial L/\partial\theta_3 = 0$ are considered. In other words, the kinetic energy is not depended on the generalized coordinates θ_1 and θ_3, and potential energy is not depended on θ_3. Olfati-Saber (Olfati-Saber, 2002) defines a coordinate to be *kinetic symmetry* if the coordinate does not appear in the kinetic energy of a mechanical system. The kinetic

symmetry is different from the well-known Lagrangian symmetry in classical mechanics. The existence of kinetic symmetries does not lead to the existence of conserved quantities in potential field. This is important to preserve the controllability of an underactuated mechanism (2). Since the relationship $\partial U/\partial\theta_1 \neq 0$ is satisfied outside the unstable balance point, the first equation of (2) cannot be integrated to a first-order differential equation, then it can be regarded as a second-order differential constraints (or second–order nonholonomic constraints (Oriolo & Nakamura, 1991)) of the actuated subsystem given by the last two equations of (2).
Further more, the dynamics (2) can be written as matrix form

$$\mathbf{M}_1(\mathbf{q}_1)\ddot{\mathbf{q}}_1 + \mathbf{C}_1(\mathbf{q}_1,\dot{\mathbf{q}}_1) + \mathbf{H}_1(\mathbf{q}_1) = \mathbf{Q}_1 \tag{3}$$

where $\mathbf{Q}_1 = \begin{bmatrix} 0 & \tau_2 & \tau_3 \end{bmatrix}^{\mathrm{T}}$ is the generalized force vector, $\mathbf{M}_1$ is the inertia matrix of the robot in stance phase and can be given by

$$\mathbf{M}_1 = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}$$

where

$$m_{11} = (m_2 + m_3 + m_4)l_1^2 + (m_3 + m_4)l_2^2 + \sum_{i=1}^{4}\left(m_i l_{ci}^2 + I_i\right) + 2m_2 l_1 l_{c2}\cos\theta_2 + 2m_3 l_1 l_2 \cos\theta_2 ,$$
$$+2m_4\left[l_1 l_2\cos\theta_2 + l_1 l_{c4}\cos(\alpha_0) + l_2 l_{c4}\cos(\theta_2 - \alpha_0)\right]$$

$$m_{12} = m_2 l_{c2}^2 + (m_3 + m_4)l_2^2 + \sum_{i=2}^{4} I_i + m_2 l_1 l_{c2}\cos\theta_2 + m_3 l_1 l_2\cos\theta_2 ,$$
$$+m_4\left[l_1 l_2\cos\theta_2 + l_1 l_{c4}\cos(\alpha_0) + l_2 l_{c4}\cos(\theta_2 - \alpha_0)\right]$$

$$m_{13} = I_3 ,$$

$$m_{21} = m_{12} ,$$

$$m_{22} = m_2 l_{c2}^2 + (m_3 + m_4)l_2^2 + \sum_{i=2}^{4} I_i ,$$

$$m_{23} = I_3 ,$$

$m_{31} = m_{13}$, $m_{32} = m_{23}$, and $m_{33} = I_3$.

$\mathbf{C}_1(\mathbf{q}_1,\dot{\mathbf{q}}_1)$ denotes the Coriolis and centrifugal forces, and has form

$$\mathbf{C}_1(\mathbf{q}_1,\dot{\mathbf{q}}_1) = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}^{\mathrm{T}}$$

where

$c_1 = \dot{m}_{11}\dot{\theta}_1 + \dot{m}_{12}\dot{\theta}_2 + \dot{m}_{13}\dot{\theta}_3$,

$c_2 = \dot{m}_{21}\dot{\theta}_1 - \frac{1}{2}\dot{\mathbf{q}}_1^{\mathrm{T}}\left(\frac{\partial}{\partial\theta_2}M_1\right)\dot{\mathbf{q}}_1$ and

$c_3 = 0$.

$\mathbf{H}_1(\mathbf{q}_1)$ is the potential forces including both gravity and elastic force of the coil spring in keen joint, and can be written as

$$\mathbf{H}_1(\mathbf{q}_1) = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix}^{\mathrm{T}}$$

where

$$h_1 = m_1 g l_{c1}\cos\theta_1 + (m_2 + m_3 + m_4) g l_1 \cos\theta_1 + (m_2 l_{c2} + m_3 l_2 + m_4 l_2) g\cos(\theta_1 + \theta_2) + m_4 l_{c4} g\cos(\theta_1 + \alpha_0)$$,

$$h_2 = (m_2 l_{c2} + m_3 l_2 + m_4 l_2) g\cos(\theta_1 + \theta_2) + m_4 l_{c4}\cos(\theta_1 + \alpha_0) g + k(\theta_2 - \theta_{20})$$,

and $h_3 = 0$.

The dynamics (3) can also be partitioned as following form according to the passive and actuated coordinates

$$\begin{aligned} &\mathbf{M}_{\mathrm{pp}}(\theta_2)\ddot{\mathbf{q}}_{\mathrm{p}} + \mathbf{M}_{\mathrm{pa}}(\theta_2)\ddot{\mathbf{q}}_{\mathrm{a}} + \mathbf{C}_{\mathrm{p}} + \mathbf{H}_{\mathrm{p}}(\theta_1, \theta_2) = 0 \\ &\mathbf{M}_{\mathrm{pa}}^{\mathrm{T}}(\theta_2)\ddot{\mathbf{q}}_{\mathrm{p}} + \mathbf{M}_{\mathrm{aa}}(\theta_2)\ddot{\mathbf{q}}_{\mathrm{a}} + \mathbf{C}_{\mathrm{a}} + \mathbf{H}_{\mathrm{a}}(\theta_1, \theta_2) = \boldsymbol{\tau} \end{aligned} \tag{4}$$

where $\mathbf{q}_{\mathrm{p}} = \theta_1$, $\mathbf{q}_{\mathrm{a}} = \begin{bmatrix} \theta_2 & \theta_3 \end{bmatrix}^{\mathrm{T}}$, and the subscript p and a denote "passive" and "actuated" respectively. Define $\mathbf{u} = \ddot{\mathbf{q}}_{\mathrm{a}}$ to be a new input, and by a control change

$$\boldsymbol{\tau} = -\left(\mathbf{M}_{\mathrm{aa}} - \mathbf{M}_{\mathrm{pa}}^{\mathrm{T}}\mathbf{M}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{pa}}\right)\mathbf{u} + \left[\mathbf{C}_{\mathrm{a}} + \mathbf{H}_{\mathrm{a}} - \mathbf{M}_{\mathrm{pa}}^{\mathrm{T}}\mathbf{M}_{\mathrm{pp}}^{-1}\left(\mathbf{C}_{\mathrm{p}} + \mathbf{H}_{\mathrm{p}}\right)\right]$$

the dynamics Eq.(4) can be transformed into the *partial feedback linearization form* that is due to (Spong, 1995).

$$\begin{aligned} \ddot{\mathbf{q}}_{\mathrm{p}} &= -\mathbf{M}_{\mathrm{pp}}^{-1}\left(\mathbf{C}_{\mathrm{p}} + \mathbf{H}_{\mathrm{p}}\right) - \mathbf{M}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{pa}}\mathbf{u} \\ \ddot{\mathbf{q}}_{\mathrm{a}} &= \mathbf{u} \end{aligned} \tag{5}$$

The main property of the underactuated system in (5) is the new control $\mathbf{u}$ appears in both the subsystems $\mathbf{q}_{\mathrm{p}}$ and $\mathbf{q}_{\mathrm{a}}$ of dynamics (5). This leads the control design for the system (5) is very difficult. In following section, we prove that the dynamics (5) can be further transformed into a special cascade nonlinear system, which simplifies the problem of designing a feasible control for the underactuated mechanical system (5) under certain additional conditions.

4. The strict feedback normal form of the dynamics

Olfati-Saber had been studied the normal form of underactuated mechanical systems in his excellent paper (Olfati-Saber, 2002). He presented three classes of cascade nonlinear systems in strict feedback form, feedforward form, and nontriangular quadratic form. As to the robot system considered in this paper, the following proposition ensures that the dynamics (5) can be transformed into the strict feedback normal form.

Proposition 1: (**strict feedback form of (5)**) The following global change of coordinates:

$$\begin{aligned} \mathbf{q}_{\mathrm{r}} &= \mathbf{q}_{\mathrm{p}} + \boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}) \\ \mathbf{p}_{\mathrm{r}} &= \partial T / \partial \dot{\mathbf{q}}_{\mathrm{p}} \end{aligned} \tag{6}$$

transforms the dynamics of (5) into a cascade nonlinear system in strict feedback form

$$\begin{aligned} \dot{\mathbf{q}}_{\mathrm{r}} &= \mathbf{M}_{\mathrm{pp}}^{-1}(\mathbf{q}_{\mathrm{a}})\mathbf{p}_{\mathrm{r}} \\ \dot{\mathbf{p}}_{\mathrm{r}} &= -\mathbf{H}_{\mathrm{p}}(\mathbf{q}_{\mathrm{r}} - \boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}), \mathbf{q}_{\mathrm{a}}) \\ \dot{\mathbf{q}}_{\mathrm{a}} &= \mathbf{p}_{\mathrm{a}} \\ \dot{\mathbf{p}}_{\mathrm{a}} &= \mathbf{u} \end{aligned} \tag{7}$$

where $\boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}) = \int_0^{\mathbf{q}_{\mathrm{a}}} \mathbf{M}_{\mathrm{pp}}^{-1}(\boldsymbol{\sigma})\mathbf{M}_{\mathrm{pa}}(\boldsymbol{\sigma})\mathrm{d}\boldsymbol{\sigma}$.

Proof: Considering the second equation of (6), it follows that

$$\mathbf{p}_{\mathrm{r}} = \partial T / \partial \dot{\mathbf{q}}_{\mathrm{p}} = \mathbf{M}_{\mathrm{pp}}(\mathbf{q}_{\mathrm{a}})\dot{\mathbf{q}}_{\mathrm{p}} + \mathbf{M}_{\mathrm{pa}}(\mathbf{q}_{\mathrm{a}})\dot{\mathbf{q}}_{\mathrm{a}}$$

thus

$$\mathbf{M}_{\mathrm{pp}}^{-1}(\theta_2)\mathbf{p}_{\mathrm{r}} = \dot{\mathbf{q}}_{\mathrm{p}} + \mathbf{M}_{\mathrm{pp}}^{-1}(\theta_2)\mathbf{M}_{\mathrm{pa}}(\theta_2)\dot{\mathbf{q}}_{\mathrm{a}} .$$

With considering the first equation of (6), then it can be obtained $\dot{\mathbf{q}}_{\mathrm{r}} = \mathbf{M}_{\mathrm{pp}}^{-1}(\theta_2)\mathbf{p}_{\mathrm{r}}$. The first equation of (7) is proved.

Once more, by the second equation of (6), we have

$$\dot{\mathbf{p}}_{\mathrm{r}} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{\mathbf{q}}_{\mathrm{p}}}\right) = \frac{\partial T}{\partial \mathbf{q}_{\mathrm{p}}} - \frac{\partial U}{\partial \mathbf{q}_{\mathrm{p}}} \tag{8}$$

since $\mathbf{q}_{\mathrm{p}} = \theta_1$ and $\partial T / \partial \mathbf{q}_{\mathrm{p}} = 0$, then

$$\dot{\mathbf{p}}_{\mathrm{r}} = -\frac{\partial U}{\partial \mathbf{q}_{\mathrm{p}}} = -\mathbf{H}_{\mathrm{p}}(\mathbf{q}_{\mathrm{p}}, \mathbf{q}_{\mathrm{a}})$$

and

$$\dot{\mathbf{p}}_{\mathrm{r}} = -\mathbf{H}_{\mathrm{p}}(\mathbf{q}_{\mathrm{r}} - \boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}), \mathbf{q}_{\mathrm{a}})$$

is following, with considering the first equation of (6). The second equation of (7) is proved.

Let $\mathbf{p}_{\mathrm{a}} = \dot{\mathbf{q}}_{\mathrm{a}}$, then the last two equation of (7) follows. This completes the proof.

Remark 1: An obstacle of using the Proposition 1 is that the integral $\boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}) = \int_0^{\mathbf{q}_{\mathrm{a}}} \mathbf{M}_{\mathrm{pp}}^{-1}(\boldsymbol{\sigma})\mathbf{M}_{\mathrm{pa}}(\boldsymbol{\sigma})\mathrm{d}\boldsymbol{\sigma}$ should be given in an explicit form. This is not always available for general underactuated mechanical system with satisfying the condition $\partial T / \partial \mathbf{q}_p = 0$. As to the robot system shown by Fig.1, thanks to the special design $l_{c3} = 0$, the integral can be explicitly obtained with slight errors. If let

$$A_1 = (m_2 + m_3 + m_4)l_1^2 + (m_3 + m_4)l_2^2 + \sum_{i=1}^{4}(m_i l_{ci}^2 + I_i) + 2m_4 l_1 l_{c4}\cos\alpha_0 ,$$

$$A_2 = m_2 l_{c2}^2 + (m_3 + m_4)l_2^2 + \sum_{i=2}^{4} I_i + m_4 l_1 l_{c4}\cos\alpha_0 ,$$

$$B_2 = m_2 l_1 l_{c2} + (m_3 + m_4)l_1 l_2 + m_4 l_2 l_{c4}\cos\alpha_0 ,$$

$$C_2 = m_4 l_2 l_{c4}\sin\alpha_0 ,$$

$$a = A_1 - 2B_2 ,$$

$$b = 4C_2 , \text{ and}$$

$$c = A_1 + 2B_2 ,$$

then the integral has form

$$\begin{aligned}\boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}) &= \int_0^{\mathbf{q}_{\mathrm{a}}} \mathbf{M}_{\mathrm{pp}}^{-1}(\boldsymbol{\sigma})\mathbf{M}_{\mathrm{pa}}(\boldsymbol{\sigma})\mathrm{d}\boldsymbol{\sigma} \\ &= \int_0^{\theta_2} \frac{m_{12}(\sigma_1)}{m_{11}(\sigma_1)} d\sigma_1 + \frac{m_{13}}{m_{11}(\theta_2)}\int_0^{\theta_3} \mathrm{d}\sigma_2 \\ &\approx \int \frac{A_2 + B_2\cos\theta_2 + C_2\sin\theta_2}{A_1 + 2B_2\cos\theta_2 + 2C_2\sin\theta_2} d\theta_2 + \frac{m_{13}}{m_{11}(\theta_{20})}\theta_3\end{aligned} .$$

Since $b^2 < 4ac$, $m_{13} = I_3$ is constant, and $m_{11}(\theta_2) - m_{11}(\theta_{20}) > 0$, then the integral can be written as

$$\begin{aligned}\boldsymbol{\psi}(\mathbf{q}_{\mathrm{a}}) \approx\ & (2A_2 - A_1)\frac{2}{\sqrt{4ac - b^2}}\arctan\left(\frac{2\mathrm{a}\cdot\tan\dfrac{\theta_2}{2} + b}{\sqrt{4\mathrm{ac}\text{-}\mathrm{b}^2}}\right) + \frac{\theta_2}{2} \\ & -(2A_2 - A_1)\frac{2}{\sqrt{4ac - b^2}}\arctan\left(\frac{b}{\sqrt{4\mathrm{ac}\text{-}\mathrm{b}^2}}\right) + \frac{m_{13}}{m_{11}(\theta_{20})}\theta_3\end{aligned} .$$

Remark 2: Since $\partial U / \partial \mathbf{q}_{\mathrm{p}} \neq 0$ is satisfied outside the unstable balance point, then $\partial U / \partial \mathbf{q}_{\mathrm{p}} = \mathbf{H}_{\mathrm{p}}(\mathbf{q}_{\mathrm{p}}, \mathbf{q}_a)$ can be regarded as the control input of the subsystem $(\mathbf{q}_{\mathrm{r}}, \mathbf{p}_{\mathrm{r}})$ of (7). Otherwise, the subsystem $(\mathbf{q}_{\mathrm{r}}, \mathbf{p}_{\mathrm{r}})$ is not controllable.

5. Design the control for the robot in stance phase

As analyzed in the introduction, there are feasible controls for underactuated mechanical systems to date only concentrated on a few classes of systems with special mathematical property in dynamics, such as the differentially flat or nilpotent. Whereas, there is still no a sufficient and necessary condition that makes certain a nonlinear systems to be differentially flat (Sira-Ramirez & Agrawal, 2004, Ch.8), and a general nilpotent nonlinear system is also difficult in control, with the major exception of the systems that can be transformed into the *chained form* (Kolmanovsky & McClamroch, 1995; Murray & Sastry, 1993; Murray, 1994). Despite that a systemic control method was proposed by (De Luca et al, 2001), for the general nilpotent systems, the convergent speed of control is slow and is hard to be applied to multi-inputs systems. The three classes of cascade nonlinear systems presented by Olfati-Saber (Olfati-Saber, 2002) are additional underactuated systems that exist feasible nonlinear control approaches. Olfati-Saber also proposed a globally stable control for two degrees of freedom (DOF) underactuated mechanical systems in (Olfati-Saber, 2000), nevertheless the suggested control is only applicable for stabilizing the system to it's origin but a trajectory tracking task is not. (Qaiser et al, 2007) also investigated the control problem for a class of underactuated mechanical systems with two DOF based on the result of (Olfati-Saber, 2000), and realized the globally exponential stabilization by Dynamic Surface Control, whereas he also didn't consider the trajectory tracking. For a nonholonomic nonlinear system, it is well known that the stabilization of the origin is not equal to it of a trajectory, since the former is a control problem for driftless nonlinear system under certain conditions, while the later is always a control problem with drift term for a nonlinear system.

To design a control for strict feedback form system (7), we define $\mathbf{z}_1 = \mathbf{q}_\mathrm{r} - \mathbf{q}_\mathrm{r}^\mathrm{d}$, $\mathbf{z}_2 = \mathbf{p}_\mathrm{r} - \mathbf{p}_\mathrm{r}^\mathrm{d}$, $\boldsymbol{\xi}_1 = \mathbf{q}_\mathrm{a} - \mathbf{q}_\mathrm{a}^\mathrm{d}$, and $\boldsymbol{\xi}_2 = \mathbf{p}_\mathrm{a} - \mathbf{p}_\mathrm{a}^\mathrm{d}$, where the superscript "d" denotes the desired or planned trajectory. Then the error system of (7) can be written as

$$\begin{aligned} \dot{\mathbf{z}}_1 &= \mathbf{f}_1 + \mathbf{g}_1 \mathbf{z}_2 \\ \dot{\mathbf{z}}_2 &= \mathbf{f}_2 + \mathbf{g}_2 \boldsymbol{\xi}_1 + \varepsilon \\ \dot{\boldsymbol{\xi}}_1 &= \mathbf{f}_3 + \mathbf{g}_3 \boldsymbol{\xi}_2 \\ \dot{\boldsymbol{\xi}}_2 &= \mathbf{f}_4 + \mathbf{g}_4 \mathbf{u} \end{aligned} \tag{9}$$

where $\mathbf{f}_1 = -\dot{\mathbf{q}}_\mathrm{r}^\mathrm{d} + \mathbf{M}_{\mathrm{pp}}^{-1}\mathbf{p}_\mathrm{r}^\mathrm{d}$, $\mathbf{g}_1 = \mathbf{M}_{\mathrm{pp}}^{-1}$,

$$\mathbf{f}_2 = -\dot{\mathbf{p}}_\mathrm{r}^\mathrm{d} - \left[\mathbf{H}_\mathrm{p}(\mathbf{z}_1 - \boldsymbol{\psi}(\boldsymbol{\xi}_1), \boldsymbol{\xi}_1)\right]_{\xi_1 = 0} = 0 ,$$

$$\mathbf{g}_2(\mathbf{z}_1, \boldsymbol{\xi}_1) = -\left(\frac{\partial}{\partial \boldsymbol{\xi}_1}\mathbf{H}_\mathrm{p}(\mathbf{z}_1 - \boldsymbol{\psi}(\boldsymbol{\xi}_1), \boldsymbol{\xi}_1)\right)_{\xi_1 = 0} ,$$

$$\mathbf{f}_3 = -\dot{\mathbf{q}}_\mathrm{a}^\mathrm{d} + \mathbf{p}_\mathrm{a}^\mathrm{d} = 0 , \ \mathbf{g}_3 = \mathbf{I} ,$$

$$\mathbf{f}_4 = -\dot{\mathbf{p}}_\mathrm{a}^\mathrm{d} , \ \mathbf{g}_4 = \mathbf{I} ,$$

$$\boldsymbol{\psi}(\xi_1)=\int_0^{\left(q_a^d+\xi_1\right)}\mathbf{M}_{pp}^{-1}(\theta_2)\mathbf{M}_{pa}(\theta_2)\mathrm{d}\boldsymbol{\sigma}\ ,$$

$\varepsilon\sim\xi_1^2$ is the errors item because of the affine approximation, and $\mathbf{I}$ is the identity matrix. To design a control for the affine nonlinear system (9) with strict feedback normal form, the following proposition can be proved.

Proposition 2: (Sliding mode backstepping control)

Consider the system (9), given $\|\varepsilon\|<\Gamma$, $\Gamma>0$ is a constant, and $\forall\eta>0$, then there exist a set of positive real numbers $k_i>0, i=1,2,\cdots,5$, and control

$$\begin{aligned}\mathbf{u}=&-\mathbf{g}_4^{-1}\left[k_4(\boldsymbol{\xi}_2-\boldsymbol{\alpha}_3)+k_5\mathrm{sign}(\boldsymbol{\xi}_2-\boldsymbol{\alpha}_3)\right]\\&-\mathbf{g}_4^{-1}\left[\mathbf{g}_3^{\mathrm{T}}(\boldsymbol{\xi}_1-\boldsymbol{\alpha}_2)+\mathbf{f}_4-\frac{\partial\boldsymbol{\alpha}_3}{\partial\boldsymbol{\xi}_1}(\mathbf{f}_3+\mathbf{g}_3\boldsymbol{\xi}_2)-\frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_2}(\mathbf{f}_2+\mathbf{g}_2\boldsymbol{\xi}_1)-\frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_1}(\mathbf{f}_1+\mathbf{g}_1\mathbf{z}_2)\right]\end{aligned}\tag{10}$$

where

$$\boldsymbol{\alpha}_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1)=-\mathbf{g}_3^{-1}\left[k_3(\boldsymbol{\xi}_1-\boldsymbol{\alpha}_2)+\mathbf{g}_2{}^{\mathrm{T}}(\mathbf{z}_2-\boldsymbol{\alpha}_1)+\mathbf{f}_3-\frac{\partial\boldsymbol{\alpha}_2}{\partial\mathbf{z}_2}(\mathbf{f}_2+\mathbf{g}_2\boldsymbol{\xi}_1)-\frac{\partial\boldsymbol{\alpha}_2}{\partial\mathbf{z}_1}(\mathbf{f}_1+\mathbf{g}_1\mathbf{z}_2)\right],$$

$$\boldsymbol{\alpha}_2(\mathbf{z}_1,\mathbf{z}_2)=-\mathbf{g}_2^{+}(k_2+\eta\Gamma^{-2})(\mathbf{z}_2-\boldsymbol{\alpha}_1)-\mathbf{g}_2^{+}\left[\mathbf{g}_1{}^{\mathrm{T}}\mathbf{z}_1+\mathbf{f}_2-\frac{\partial\boldsymbol{\alpha}_1}{\partial\mathbf{z}_1}(\mathbf{f}_1+\mathbf{g}_1\mathbf{z}_2)\right],$$

$\boldsymbol{\alpha}_1(\mathbf{z}_1)=-\mathbf{g}_1^{-1}(\mathbf{f}_1+k_1\mathbf{z}_1)$, and

$$\mathrm{sign}(x)=\begin{cases}1 & x>0\\0 & x=0\\-1 & x<0\end{cases},$$

renders the system (9) exponentially stabilize to the origin $(\mathbf{z},\boldsymbol{\xi})=(0,0)$.

Proof: Consider the subsystem $\mathbf{z}_1$, and select $V_1(z_1)=\frac{1}{2}z_1^{\mathrm{T}}z_1$ to be the candidate Lyapunov function, then one has

$$\dot{V}_1(\mathbf{z}_1)=\mathbf{z}_1^{\mathrm{T}}(\mathbf{f}_1+\mathbf{g}_1\mathbf{z}_2)\ .$$

Let $\mathbf{z}_2=\boldsymbol{\alpha}_1(\mathbf{z}_1)=-\mathbf{g}_1^{-1}(\mathbf{f}_1+k_1\mathbf{z}_1)$, then it follows that

$$\dot{V}_1(\mathbf{z}_1)=-k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1\ .$$

Further consider the subsystem $(\mathbf{z}_1,\mathbf{z}_2)$, select $V_2(\mathbf{z}_1,\mathbf{z}_2)=V_1(\mathbf{z}_1)+\frac{1}{2}(\mathbf{z}_2-\boldsymbol{\alpha}_1)^{\mathrm{T}}(\mathbf{z}_2-\boldsymbol{\alpha}_1)$ to be a new candidate Lyapunov function, and let $\mathbf{e}_{z_2}=\mathbf{z}_2-\boldsymbol{\alpha}_1$, $\forall\eta>0$. By the Young's inequality $2ab\le a^2+b^2$ and Cauchy-Schwarz inequality $\left|\mathbf{x}^{\mathrm{T}}\mathbf{y}\right|\le\|\mathbf{x}\|\cdot\|\mathbf{y}\|$, we have

$$\begin{aligned}\dot{V}_2(\mathbf{z}_1,\mathbf{z}_2) &= \dot{V}_1(\mathbf{z}_1) + \mathbf{e}_{z_2}^{\mathrm{T}}(\dot{\mathbf{z}}_2 - \dot{\boldsymbol{\alpha}}_1) \\ &= \mathbf{z}_1^{\mathrm{T}}\left[\mathbf{f}_1 + \mathbf{g}_1(\mathbf{e}_{z_2} + \boldsymbol{\alpha}_1)\right] + \mathbf{e}_{z_2}^{\mathrm{T}}\left[\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1 + \varepsilon - \frac{\partial \boldsymbol{\alpha}_1}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right] \\ &\le \mathbf{z}_1^{\mathrm{T}}\left[\mathbf{f}_1 + \mathbf{g}_1\boldsymbol{\alpha}_1\right] \\ &\quad + \mathbf{e}_{z_2}^{\mathrm{T}}\left[\mathbf{g}_1{}^{\mathrm{T}}\mathbf{z}_1 + \mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1 - \frac{\partial \boldsymbol{\alpha}_1}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right] + \eta\left\|\mathbf{e}_{z_2}^{\mathrm{T}}\right\|^2\|\boldsymbol{\varepsilon}\|^2 + \frac{1}{4\eta}\end{aligned}$$

$$\begin{aligned}&\le -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 \\ &+\mathbf{e}_{z_2}^{\mathrm{T}}\left[\mathbf{g}_1{}^{\mathrm{T}}\mathbf{z}_1 + \mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\alpha}_2 + \mathbf{g}_2(\boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2) - \frac{\partial \boldsymbol{\alpha}_1}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2) + \eta\Gamma^2\mathbf{e}_{z_2}\right] + \frac{1}{4\eta}\end{aligned}$$

Let $\boldsymbol{\alpha}_2(\mathbf{z}_1,\mathbf{z}_2) = -\mathbf{g}_2^{+}(k_2 + \eta\Gamma^2)\mathbf{e}_{\mathbf{z}_2} - \mathbf{g}_2^{+}\left[\mathbf{g}_1{}^{\mathrm{T}}\mathbf{z}_1 + \mathbf{f}_2 - \frac{\partial \boldsymbol{\alpha}_1}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right]$, then the last inequality can be written as

$$\dot{V}_2(\mathbf{z}_1,\mathbf{z}_2) \le -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} + \mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{g}_2(\boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2) + \frac{1}{4\eta}.$$

For the subsystem $(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1)$, select $V_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) = V_2(\mathbf{z}_1,\mathbf{z}_2) + \frac{1}{2}(\boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2)^{\mathrm{T}}(\boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2)$ as the candidate Lyapunov function, and let $\mathbf{e}_{\xi_1} = \boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2$, then

$$\begin{aligned}\dot{V}_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) &= \dot{V}_2(\mathbf{z}_1,\mathbf{z}_2) + (\boldsymbol{\xi}_1 - \boldsymbol{\alpha}_2)^{\mathrm{T}}(\dot{\boldsymbol{\xi}}_1 - \dot{\boldsymbol{\alpha}}_2) \\ &= -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} + \mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{g}_2\mathbf{e}_{\xi_1} \\ &\quad + \mathbf{e}_{\xi_1}^{\mathrm{T}}\left[\mathbf{f}_3 + \mathbf{g}_3\boldsymbol{\alpha}_3 + \mathbf{g}_3(\boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3) - \frac{\partial \boldsymbol{\alpha}_2}{\partial \mathbf{z}_2}(\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1) - \frac{\partial \boldsymbol{\alpha}_2}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right] + \frac{1}{4\eta}\end{aligned}.$$

Select

$$\boldsymbol{\alpha}_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) = -\mathbf{g}_3^{-1}\left[k_3\mathbf{e}_{\xi_1} + \mathbf{g}_2{}^{\mathrm{T}}\mathbf{e}_{z_2} + \mathbf{f}_3 - \frac{\partial \boldsymbol{\alpha}_2}{\partial \mathbf{z}_2}(\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1) - \frac{\partial \boldsymbol{\alpha}_2}{\partial \mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right],$$

then it follows that

$$\dot{V}_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) \le -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} - k_3\mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{e}_{\xi_1} + \mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{g}_3(\boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3) + \frac{1}{4\eta}.$$

For the system (9), let $V_4(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) = V_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) + \frac{1}{2}(\boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3)^{\mathrm{T}}(\boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3)$ be the candidate Lyapunov function and let $e_{\xi_2} = \boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3$, we have

$$\begin{aligned}\dot{V}_4(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) &= \dot{V}_3(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1) + \mathbf{e}_{\xi_2}^{\mathrm{T}}(\dot{\boldsymbol{\xi}}_2 - \dot{\boldsymbol{\alpha}}_3) \\ &\le -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} - k_3\mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{e}_{\xi_1} + \mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{g}_3(\boldsymbol{\xi}_2 - \boldsymbol{\alpha}_3) + \\ &\mathbf{e}_{\xi_2}^{\mathrm{T}}\left[\mathbf{f}_4 + \mathbf{g}_4\mathbf{u} - \frac{\partial\boldsymbol{\alpha}_3}{\partial\boldsymbol{\xi}_1}(\mathbf{f}_3 + \mathbf{g}_3\boldsymbol{\xi}_2) - \frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_2}(\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1) - \frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right] + \frac{1}{4\eta}\end{aligned}.$$

If we select the control to be

$$\begin{aligned}\mathbf{u} = &-\mathbf{g}_4^{-1}\left[k_4\mathbf{e}_{\xi_2} + k_5\mathrm{sign}(\mathbf{e}_{\xi_2})\right] \\ &-\mathbf{g}_4^{-1}\left[\mathbf{g}_3^{\mathrm{T}}\mathbf{e}_{\xi_1} + \mathbf{f}_4 - \frac{\partial\boldsymbol{\alpha}_3}{\partial\boldsymbol{\xi}_1}(\mathbf{f}_3 + \mathbf{g}_3\boldsymbol{\xi}_2) - \frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_2}(\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\xi}_1) - \frac{\partial\boldsymbol{\alpha}_3}{\partial\mathbf{z}_1}(\mathbf{f}_1 + \mathbf{g}_1\mathbf{z}_2)\right].\end{aligned}$$

Obviously, it follows that

$$\dot{V}_4(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \le -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} - k_3\mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{e}_{\xi_1} - k_4\mathbf{e}_{\xi_2}^{\mathrm{T}}\mathbf{e}_{\xi_2} - k_5\left\|\mathbf{e}_{\xi_2}\right\| + \frac{1}{4\eta},$$

then if one selects $k_5 > 0$ and $k_5\left\|\mathbf{e}_{\xi_2}\right\| > \frac{1}{4\eta}$, the following inequality is satisfied

$$\dot{V}_4(\mathbf{z}_1,\mathbf{z}_2,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) < -k_1\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - k_2\mathbf{e}_{z_2}^{\mathrm{T}}\mathbf{e}_{z_2} - k_3\mathbf{e}_{\xi_1}^{\mathrm{T}}\mathbf{e}_{\xi_1} - k_4\mathbf{e}_{\xi_2}^{\mathrm{T}}\mathbf{e}_{\xi_2} < 0,$$

thus the origin of the affine nonlinear system with drift term is asymptotically stable. If we select $k_i = \frac{\lambda}{2} > 0, i = 1,2,3,4$, then the last inequality can be rewritten as

$$\dot{V}_4 < -\lambda V_4.$$

The solution of the differential inequality is given by

$$V_4(t) < V_4(t_0)e^{-\lambda(t-t_0)}.$$

Thus the control (10) will render the system (9) exponentially stabilize to the origin $(\mathbf{z},\boldsymbol{\xi}) = (0,0)$. This completes the proof.

Remark 3: As $\mathbf{g}_2 \in R^{1\times 2}$ is not a square matrix for the robot system considered in this paper, the inverse matrix of $\mathbf{g}_2$ is calculated by Moore-Penrose pseudo-inverse $\mathbf{g}_2^{+} = \mathbf{g}_2^{\mathrm{T}}(\mathbf{g}_2\mathbf{g}_2^{\mathrm{T}})^{-1}$.

Remark 4: Different from the most of control plants of nonholonomic systems in literatures, the affine nonlinear system (9) with drift terms $\mathbf{f}_i \neq 0$ is considered in this paper, thus the control presented by (10) can both stabilize the unstable balance point and track a feasible trajectory of the hopping robot system (2). The nonlinear control methods suggested by (Olfati-Saber, 2000, 2002) and (Qaiser, 2007) can not be utilized to the trajectory tracking problem.

6. Motion planning for the hopping robot in stance phase

Motion planning for a hopping robot with non-SLIP model is not intuitional. Fig.2 shows a sketch of motion of the hopping robot in stance phase. In the figure, MC denotes the mass

center of the robot, x_s indicates the moving distance of MC of the robot in stance phase, β is the angle of MC deviating from the vertical line that traverses the point of foot contacting with the ground. For an underactuated mechanical system, not arbitrary motions are feasible, the planned motion must satisfy the second-order nonholonomic constraints of the system. As to the robot system in Fig.1, the second-order nonholonomic constraints are given by the first equation of (4).

Denote $\mathbf{X}_c = [x_c \quad y_c]^T$ to be the position of MC of the robot in stance phase, the kinematics of the MC of the robot can be formulated as

$$\mathbf{X}_c = \mathbf{F}(\mathbf{q}_1). \tag{11}$$

The acceleration equation of (11) has form

$$\ddot{\mathbf{X}}_c = \mathbf{J}\ddot{\mathbf{q}}_1 + \dot{\mathbf{J}}\dot{\mathbf{q}}_1, \tag{12}$$

where $\mathbf{J} = \partial\mathbf{F}/\partial\mathbf{q}_1$. With considering the second-order nonholonomic constraints, the feasible motion in joint space can be calculated by

$$\ddot{\mathbf{q}}_1^d = \begin{bmatrix} \mathbf{J} \\ \mathbf{M}_{pp} \quad \mathbf{M}_{pa} \end{bmatrix}^{-1} \left(\begin{bmatrix} \ddot{\mathbf{X}}_c^d \\ 0 \end{bmatrix} - \begin{bmatrix} \dot{\mathbf{J}}\dot{\mathbf{q}}_1 \\ \mathbf{C}_p + \mathbf{H}_p \end{bmatrix} \right), \tag{13}$$

where $\mathbf{X}_c^d$ is the desired trajectory of MC that is planned in Cartesian space.

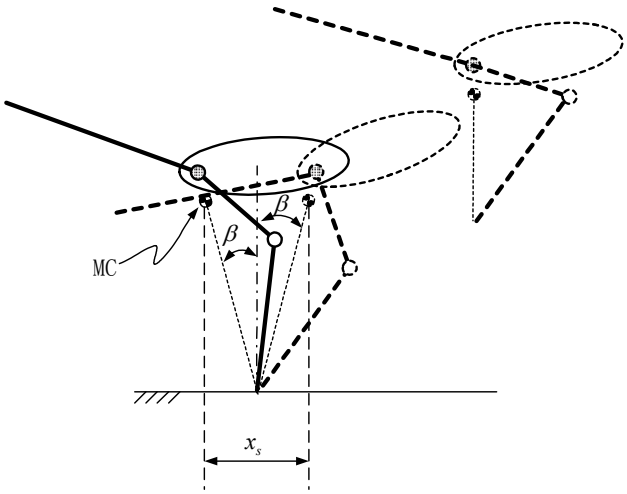


Fig. 2. Motion sketch of the robot in stance phase

The configuration with balance potential forces is an important point in joint space of the robot in stance phase. This point is both the control target of the robot for static balance and the reference point of planning a feasible trajectory of MC of the robot in every stance phase for continuous hopping. For searching the static balance configuration, one has to employ a numerical method since the complexity of the robot dynamics. For instance, define an optimization measure as follow

$$\mu(\mathbf{q}_1) = \mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1 + x_{\mathrm{c}}^2, \tag{14}$$

where $\mathbf{H}_1$ is the potential force term in equation (3), and x_{c} is the coordinate of MC of the robot along horizontal direction. If the measure $\mu(\mathbf{q}_1) > 0$ is minimized such that it has $\mu(\mathbf{q}_1^*) \approx 0$, then the corresponding optimized variables $\mathbf{q}_1^*$ is the searching configuration that satisfies the static balance condition.

Many algorithms can be employed to search for the optimal solution. One simple algorithm can be given as

$$\begin{cases} \mathbf{q}_1(i+1) = \mathbf{q}_1(i) - \lambda \dfrac{\nabla_{\mathbf{q}_1}\mu}{\left(\nabla_{\mathbf{q}_1}\mu\right)^{\mathrm{T}}\nabla_{\mathbf{q}_1}\mu}, \\ \mathbf{q}_1(0) = \mathbf{q}_1^0 \end{cases} \tag{15}$$

where $\lambda > 0$ is the iterative step length, $\nabla_{\mathbf{q}_1}\mu = \dfrac{\partial\mu}{\partial\mathbf{q}_1}$ is the grads of the measure $\mu(\mathbf{q}_1)$ along the smooth vector field $\mathbf{q}_1$, and i is the iterative times.

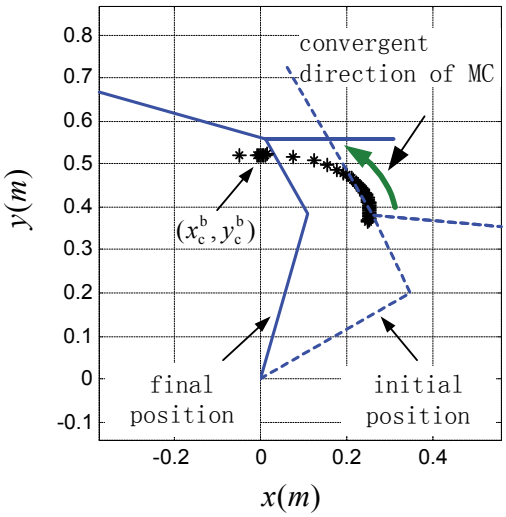


Fig. 3. Simulation for searching the balance configuration of the robot in stance phase

The static balance configuration of the robot can be calculated to be $\mathbf{q}_1^* = (\theta_1, \theta_2, \theta_3) \approx (72.763^{\mathrm{o}}, 45.980^{\mathrm{o}}, -120^{\mathrm{o}})$ when the physical parameters of the robot is selected as

a. The gravitational acceleration is $g = 9.8\,\mathrm{N/kg}$;
b. The stiffness of spring in keen joint is $k = 46.19\,\mathrm{Nm/rad}$;
c. The phase angle between the tail and thigh is $\alpha_0 = 90^{\mathrm{o}}$;
d. The mass, length and location of MC, and the inertia of every body are given by

$$m_1 = 0.6\mathrm{kg}\ ,\ l_1 = 0.4\mathrm{m}\ ,\ l_{c1} = \frac{1}{2}l_1\ ,\ I_1 = \frac{1}{12}m_1 l_{c1}^2\ ,$$

$$m_2 = 1.5\text{kg}\ ,\ l_2 = 0.2\text{m}\ ,\ l_{c2} = \frac{1}{2}l_2\ ,\ I_2 = \frac{1}{12}m_2 l_{c2}^2\ ,$$

$$m_3 = 5.0\text{kg}\ ,\ l_3 = 0.3\text{m}\ ,\ l_{c3} = 0\ ,\ I_3 = \frac{1}{2}m_1 l_3^2\ ,$$

$$m_4 = 1.5\text{kg}\ ,\ l_4 = 0.4\text{m}\ ,\ l_{c4} = \frac{1}{2}l_4\ ,\ I_4 = \frac{1}{12}m_4 l_{c4}^2\ .$$

Fig.3 shows the convergent procedure for finding the balance configuration, and $(x_c^b, y_c^b) \approx (0\text{m}, 0.5219\text{m})$ denotes the position of MC when the robot is static balance.

7. Numerical simulations

In this section, some numerical simulations are provided for verifying the feasibility of the robot mechanism and the nonlinear controller proposed in the former sections. The physical parameters of the robot mechanism are listed in section 6. Fig.4 shows the simulation results for stance balance control. In this simulation, the initial configuration errors of the robot is given by $\mathbf{e}_{q_1} = \begin{bmatrix} 10^\text{o} & -15^\text{o} & -50^\text{o} \end{bmatrix}^\text{T}$ and the target configuration is $(\theta_1, \theta_2, \theta_3) \approx (72.763^\text{o}, 45.980^\text{o}, -120^\text{o})$, which is obtained in section 6. One can find the configuration errors converge to zero (Fig.4 (b)) and finally stable at the balance configuration (Fig. 4 (a)). Fig. 4 (c) shows the trajectory of MC of the robots during the stabilizing procedure. Fig.4 (d) shows the corresponding torques of the actuators. For more intuitively, Fig.5 also shows the configurations snapshots of the robot during the stance balance control.

The general stance phase motion for the hopping robot is commonly a continuous trajectory that nears to the stance balance configuration. Given the desired motion of MC of the robot is

$$\ddot{\mathbf{X}}_\text{c}^\text{d}(t) = \begin{cases} \begin{bmatrix} -0.07t\omega_n^2 \sin(\omega_n t) & 0 \end{bmatrix}^\text{T} & t < 1s \\ \begin{bmatrix} 0.07\omega_n^2 \sin(\omega_n t) & 0 \end{bmatrix}^\text{T} & t \geq 1s \end{cases},$$

where ω_n is the nature angular frequency of the robot. The desired motion is a periodical motion of MC moving along direction x. Fig.6 depicts the simulation results and one can find that the desired motion is approximately realized (see Fig.6 (b)). In Fig.6 (a)-(c), the curve in dashed indicates the desired motion, and the solid curve is the controlled motion of the corresponding variable. The large fluctuation of MC along direction y is induced by the special synchronous transmission system of the keen joint and tail joint.

Fig.7 shows another simulation result for trajectory control of MC moving along the direction y. In this simulation, the planned motion of MC is given by

$$\mathbf{X}_\text{c}^\text{d}(t) = \begin{cases} \begin{bmatrix} 0 & 0.03t\omega_n^2 \sin(\omega_n t) \end{bmatrix}^\text{T} & t < 1s \\ \begin{bmatrix} 0 & 0.03\omega_n^2 \sin(\omega_n t) \end{bmatrix}^\text{T} & t \geq 1s \end{cases}.$$

In Fig.7 (a)-(c), the curve in dashed also indicates the desired motion, and the solid curve indicates the controlled motion of the corresponding variable. There is no capsizal torque

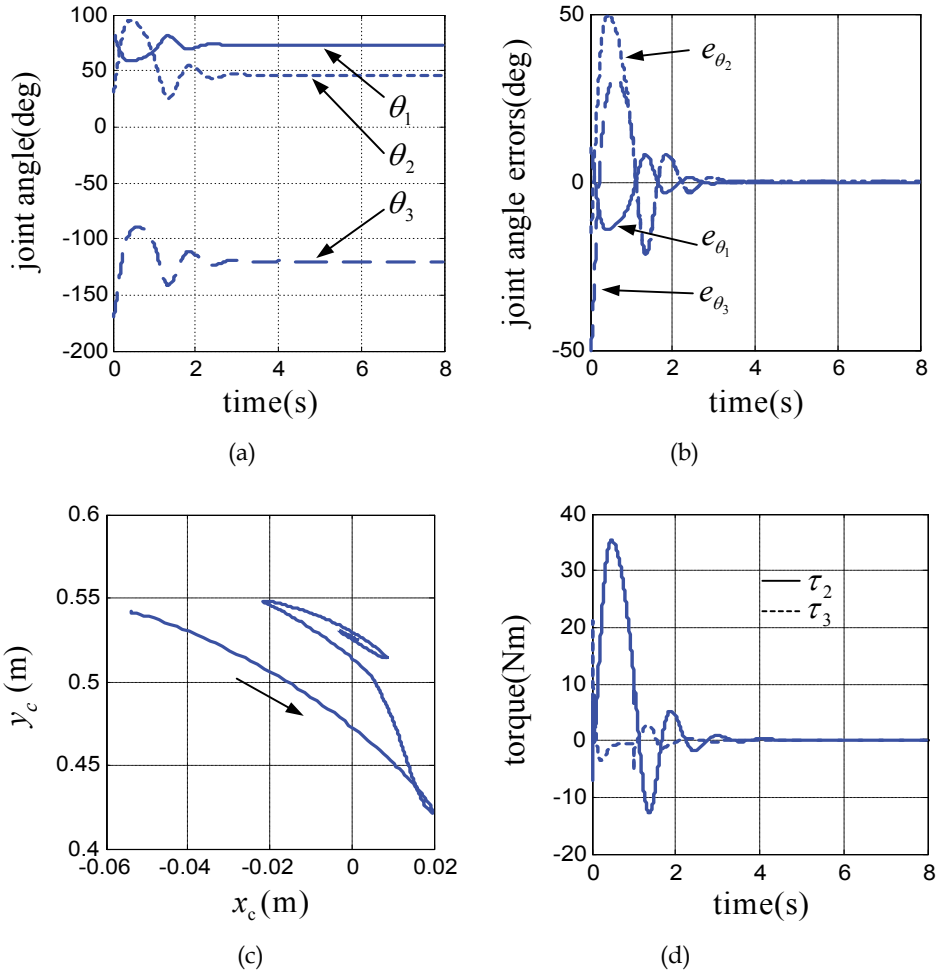


Fig. 4. Simulation for the stance balance control

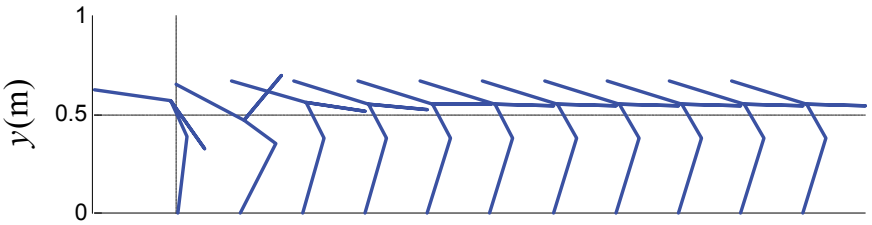
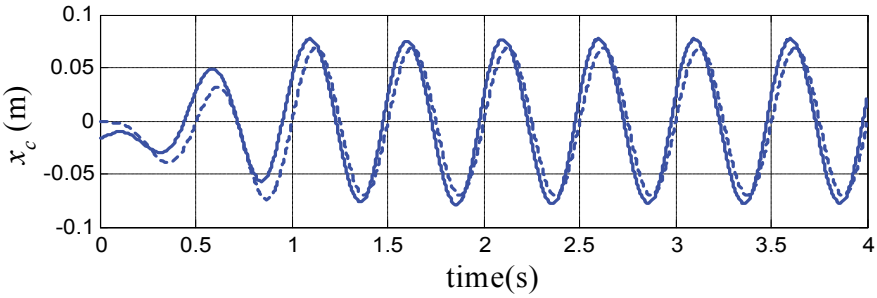
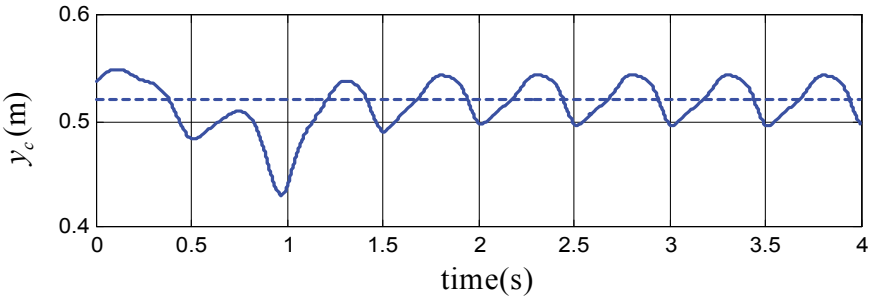


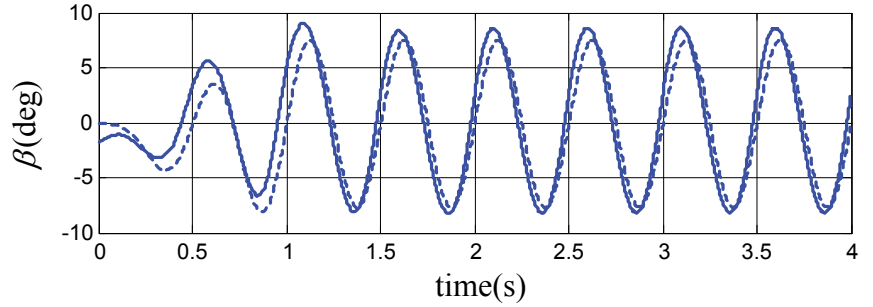
Fig. 5. Some configuration snapshots of the hopping robot during stance balance control shown by Fig.4



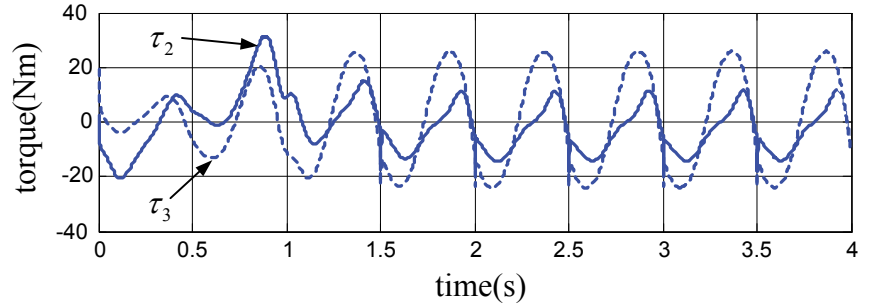
(a)



(b)



(c)



(d)

Fig. 6. Simulation for a periodical motion of MC along direction x .

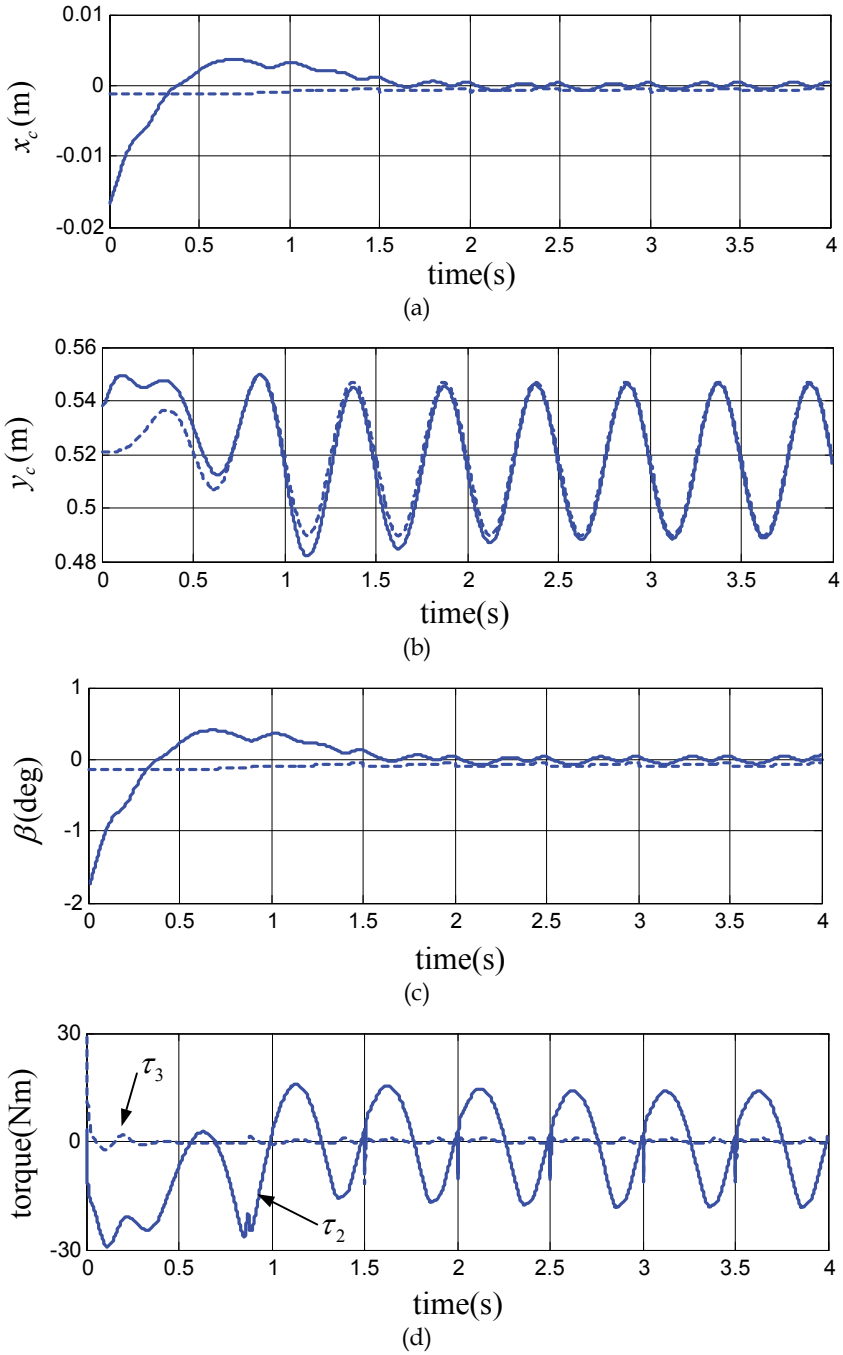


Fig. 7. Simulation for a periodical motion of MC along direction y .

need to be balanced since that the desired motion follows the vertical direction through the supporting point, one can find that the motion is realized with less error than it in Fig.6. Both Fig.6 (d) and Fig. 7(d) illustrate the torque of actuators during the corresponding controlled motion in two different directions of MC.

It is well known that the angular momentum of a hopping robot system in flight phase is conserved, and the angular momentum is generally unintegrable, thus the hopping robot systems in flight phase are first order nonholonomic system. For the first order nonholonomic systems with two inputs, there is a theorem that confirms the system can be transformed into the so-called chained form (Murray & Sastry, 1993; Murray, 1994). As to the chained from systems, there are many feasible control method in literatures. Therefore, one can expect that the novel hopping system presented in this paper will stable hopping under appropriate motion planning. The optimal motion plan and control problems for the novel hopping system with considering the energy-efficiency, and comparing the presented mechanism with the SLIP model system from the point of view of energy-efficiency and mobility, are interesting works in the future.

8. Conclusions

On the basis of dynamic synthesis, a novel mechanism for one-legged hopping robot is proposed. Different from the most of relative researches in literature, the proposed hopping robot mechanism is a non-SLIP model system, which generally shows more biological characteristics while the control problem of it is intractable, due to the complex nonlinear dynamics and the second-order nonholonomic constraints. Thanks to the special design, it is proved that the dynamics of the presented hopping robot mechanism can be transformed into the non-affine strict feedback normal form. Further more, it is shown that the normal form can also be rewritten as affine system with slightly approximation. Then a sliding model backstepping control is proposed for stabilizing the nonlinear dynamic system to its origin as well as a given trajectory around the balance configuration of the robot in stance phase. The stability of the presented control is proved, and verified by some numerical simulations.

9. Acknowledgments

This work is supported by Nature Science Foundation of China (No.50975004) and PHR(IHLB)(No. PHR200906107).

10. References

Agrawal, S. K. & Fattah A. (2006). Motion Control of a Novel Planar Biped With Nearly Linear Dynamics, *IEEE/ASME Transactions on Mechatronics*, Vol.11, No.2, pp.162-168, ISSN: 1083-4435.

Ahmadi, M. & Buehler, M. (1997). Stable Control of a Simulated One-Legged Running Robot with Hip and Leg Compliance, *IEEE Transactions on Robotics and Automation*, Vol.13, No.1, pp. 96-104, ISSN: 1042-296X.

Ahmadi, M. & Buehler, M. (2006). Controlled Passive Dynamic Running Experiments With the ARL-Monopod II, *IEEE Transactions on Robotics*, Vol.22, No.5, pp. 974-986, ISSN: 1552-3098.

Ahmadi, M.; Michalska, H. & Buehler, M.(2007). Control and Stability Analysis of Limit Cycles in a Hopping Robot, *IEEE Transactions on Robotics*, Vol.23, No.3, pp.553-563, ISSN: 1552-3098.

Arai,H.; Tanie,K. & Shiroma,N. (1998). Time-scaling Control of an Underactuated Manipulator, *Journal of Robotic Systems*, Vol.15, No.9, pp. 525-536, ISSN: 1097-4563.

De Luca, S. I. & Oriolo, G. (2001). Stabilization of a PR Planar Underactuated Robot, *IEEE International Conference on Robotics and Automation*, pp.2090-2095, ISBN: 0-7803-6576-3, Seoul.

Franch, J.; Agrawal, S.K. & Fattah, A. (2003). Design of Differential Flat Planar Space Robots: A Step Forward in their Planning and Control, *Proceedings of the 2003 IEEE/RSJ Intentional Conference on Intelligent Robots and Systems*, pp.3053-3058, ISBN: 0-7803-7860-1.

Francois, C.; Samson, C. (1998). A New Approach to the Control of the Planar One-Legged Hopper, *The International Journal of Robotics Research*, Vol.17, No.11, pp.1150-1166, ISSN: 0278-3649.

Graichen, K.; Treuer, M. & Zeitz,M. (2007). Swing-up of the double pendulum on a cart by feedforward and feedback control with experimental validation, *Automatica*, Vol.43, pp.63-71, ISSN: 0005-1098.

Gregorio,P.; Ahmadi, M. & Buehler, M. (1997). Design, Control, and Energetics of an Electrically Actuated Legged Robot, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol.27, No.4, pp. 626-634, ISSN: 1083-4419.

Guang-Ping H. & Zhi-Yong G. (2008). Optimal Motion Planning for Differential flat Underactuated Mechanical Systems, *Proceedings of the IEEE International Conference on Automation and Logistics*, pp.1567-1572, ISBN: 978-1-4244-2502-0, Qingdao, China.

Guang-Ping H. & Zhi-Yong G. (2009a). Optimal Motion Planning for Differential flat Underactuated Mechanical Systems, *Chinese Journal of Mechanical Engineering*, Vol.22, No.3, pp. 347-354, ISSN: 0257-9731.

Guang-Ping H. & Zhi-Yong G. (2009b). Exponentially Stabilizing An One-Legged Hopping Robot With Non-SLIP Model In Flight Phase, *Mechatronics*, Vol.19, No.3, pp. 364-374, ISSN: 0957-4158.

Guang-Ping H. & Zhi-Yong G. (2010). Robust backstepping control of an underactuated one-legged hopping robot in stance phase, *Robotica*, Vol.28, No. pp.583-596, ISSN: 0263-5747.

Hodgins, J. K. & Raibert, M. H. (1990). Biped Gymnastics, *The International Journal of Robotics Research*, Vol.9, No.2, pp. 115-132, ISSN: 0278-3649.

Hyon, S. & Emura, T. (2002). Quasi_periodic Gaits of Passive One-Legged Hopper, *IEEE/RSJ International Conference on Intelligent Robotics and Systems*, pp. 2625-2630, ISBN: 0-7803-7398-7, Switzerland.

Hyon, S. H. & Emura, T. (2004). Energy-preserving Control of a Passive One-Legged Running Robot, *Advanced Robotics*, Vol.18, No.4, pp.357-381, ISSN: 0169-1864.

Hyon, S. H. & Mita, T. (2002). Development of a Biologically Inspired Hopping Robot-"Kenken", *IEEE International Conference on Robotics and Automation*, pp.3984-3991, ISBN: 0-7803-7272-7, Washington DC.

Hyon, S. H.; Jiang, X.; Emura, T. & Ueta, T. (2004). Passive Running of Planar 1/2/4-Legged Robots, *IEEE/RSJ International Conference on Intelligent Robots*, pp. 3532-3539, ISBN: 0-7803-8463-6, Sendai.

Kolmanovsky, I. & McClamroch, H. (1995). Developments in Nonholonomic Control Problems, *IEEE Control Systems magazine*, Vol. 15, No.6, pp.20-36, ISSN: 0272-1708.

Lai,X.Z.; She,J.H.; Yang, S.X. & Wu, M. (2008). Control of acrobot based on non-smooth Lyapunov function and comprehensive stability analysis, *IET Control Theory*, Vol.2, No.3, pp. 181-191, ISSN: 1751-8644.

Lapshin, V.V. (1992). Vertical and Horizontal Motion Control of a One-Legged Hopping Machine, *The International Journal of Robotics Research*, Vol.11, No.5, pp.491-498, ISSN: 0278-3649.

Murray, R. M. & Sastry, S. (1993). Nonholonomic motion planning: Steering using sinusoids, *IEEE Transactions on Automation Control*, Vol. 38, No.5, pp. 700-716, ISSN: 0018-9286.

Murray, R. M. (1994). Nilpotent bases for a class of non-integrable distributions with applications to trajectory generation for nonholonomic systems, *Math. Controls, Signals, and Systems*, Vol.7, No.1, pp.58-75, ISSN: 0932-4194.

Nieuwstadt,M. J. & Murray, R. M. (1995). Approximate Trajectory Generation for Differentially Flat Systems with Zero Dynamics, *Proceedings of the 34th Conference on Decision and Control*, pp. 4224-4230, ISBN: 0-7803-2685-7, New Orleans.

Olfati-Saber,R. (2000). Control of Underactuated Mechanical Systems with two Degrees of Freedom and Symmetry, *Proceedings of American Control Conference*, pp.4092-4096, ISBN: 0-7803-5519-9.

Olfati-Saber,R. (2002). Normal Forms for Underactuated Mechanical Systems with Symmetry, *IEEE Transactions on Automatic Control*, Vol.47, No.2, pp.305-308, ISSN: 0018-9286.

Oriolo, G. & Nakamura, Y. (1991). Free-Joint Manipulators: Motion Control under Second-Order Nonholonomic Constraints, *IEEE/RSJ International Workshop on Intelligent Robots and Systems*, pp. 1248-1253, ISBN: 0-7803-0067-X, Osaka.

Oriolo, G. & Vendittelli, M. (2005). A Framework for the Stabilization of General Nonholonomic Systems With an Application to the Plate-Ball Mechanism, *IEEE Transactions on Robotics*, Vol.21, No.2, pp. 162-175, ISSN: 1552-3098.

Papadopoulos, E. & Cherouvim, N. (2004). On Increasing Energy Autonomy for a One-Legged Hopping Robot, *IEEE International Conference on Robotics and Automation*, pp. 4645-4650, ISBN: 0-7803-8232-3, New Orleans.

Qaiser,N.; Iqbal,N.; Hussain,A. & Qaiser, N. (2007). Exponential Stabilization of a Class of Underactuated Mechanical Systems using Dynamic Surface Control, *International Journal of Control, Automation, and Systems*, Vol. 5, No.5, pp. 547-558, ISSN: 1598-6446.

Raibert, M .H. (1986). *Legged Robots That Balance*, MIT Press, ISBN-10: 0-262-18117-7, Cambridge, MA.

Raibert, M. Benjamin Brown, Jr. H. H. & Chepponis, M. (1984). Experiments in Balance with a 3D One-Legged Hopping Machine, *The International Journal of Robotics Research*,Vol.3, No.2, pp. 75-92, ISSN: 0278-3649.

Reyhanoglu,M. (1997). Exponential Stabilization of an Underactuated Autonomous Surface Vessel, *Automatica*,Vol.33, No.12, pp. 2249-2254, ISSN: 0005-1098.

Sira-Ramirez, H. & Agrawal, S. K. (2004). *Differentially Flat Systems*, Marcel Dekker, Inc., ISBN-10: 0824754700, New York.

Spong, M. W. (1995). The Swing Up Control Problem For The Acrobot, *IEEE Control Systems Magazine*, Vol.15, No.1, pp. 49-55, ISSN: 0272-1708.

Spong,M. & W. Block, D. J. (1995). The Pendubot: A Mechatronic System for Control Research and Education, *Proceedings of the 34th Conference on Decision and Control*, pp.555-556, ISBN: 0-7803-2685-7, New Orleans.

Takahashi, T.; Yamakita, M. & Hyon, S.H. (2006). An Optimization Approach for Underactuated Running Robot, *SICE-ICASE International Joint Conference*, pp.3505-3510, ISBN: 89-950038-4-7.

Vakasi, A. F.; Burdick, J. W. & Caughey, T.K. (1991). An "Interesting" Strange Attractor in the Dynamics of a Hopping Robot, *The International Journal of Robotics Research*, Vol.10, No.6, pp. 606-618, ISSN: 0278-3649.

Zeglin, G. J. (1991). *Uniroo: A One Legged Dynamic Hopping Robot*, Massachusetts Institute of Technology, Bachelor Dissertation.

Zeglin, G. J. (1999). *The Bow Leg Hopping Robot*, Carnegie Mellon University, Doctor Dissertation.

20

Mechanics of Cold Rolling of Thin Strip

Z. Y. Jiang
School of Mechanical, Materials and Mechatronic Engineering,
University of Wollongong, Wollongong,
Australia

1. Introduction

Cold rolled thin strip has a wide application in electronic and instrument industries, and its production has always been of major interest to the manufacturers and researchers in the area of metal plasticity. Thin strip rolling involves significant metal plasticity to produce a desired product. Iwamoto (2004), Stoughton & Yoon (2004) and Huh et al. (2004) were interested in dealing with the plastic deformation and plastic yielding of steel, and its micro-mechanics. With the need for higher quality and productivity in cold strip mill, mathematical models of cold rolling of a strip with a desired shape and dimension, both for mill set-up and for on-line control, have become a key issue in the steel rolling process. One major part of these models concerns the strip and roll deformation, plastically deformed strip shape and profile. The development of the roll deformation model can be divided into three groups, which includes simple beam model, slit beam model and finite element analysis model (Ginzburg, 1989). Stone & Gray (1965) modelled the roll deformation as the deflection of a simple beam on an elastic foundation. Shohet & Townsend (1968) proposed a slit beam deflection model, and then Edwards & Spooner (1973), Wang (1986) improved this theory and introduced a matrix method to solve the beam deflection considering strip plastic deformation. It has now been widely used in analysis of the roll deformation and strip shape and profile. Timoshenko & Goodier (1970), Jiang et al. (2003a, b, c), Komori (1998) and Lin & Lee (1997) used finite element model and numerical methods to analyse the strip rolling and to improve the simulation accuracy of the strip shape and profile. In order to improve the quality of the produced products, Kim & Oh (2003) used finite element method to analyse grain-by-grain deformation by crystal plasticity with couple stress, Simth et al. (2003) conducted a study of the effect of the transverse normal stress on sheet metal formability and Ho et al. (2004) developed integrated numerical techniques to predict springback in creep forming thick aluminum sheet components. Buchheit et al. (2005) performed simulations of realistic looking 3-D polycrystalline microstructures generated. The simulation on precipitate induced hardening in crystal plasticity was conducted (Han et al., 2004). Martin & Smith (2005) investigated the influence of the compressive through-thickness normal stress on sheet metal formability and tried to explore the ways to improve the sheet metal formability. However, the finite element analysis is rather complicated and may have a convergence problem, which is difficult to be used for on-line control of the thin strip rolling. An influence function method analysis considering the strip plastic deformation and roll deformation can be directly used in the control of strip rolling, especially in the control of the shape and profile of strip.

In practical rolling of thin strip, there is a phenomenon that the upper and lower work rolls may contact each other beyond the edges of strip if the strip is very thin and there is no work roll bending applied as shown in Fig. 1. m and n are the number of the slab elements along the half roll barrel and half strip width respectively. Δx is the width of each element; D_w the diameter of work roll; D_b the diameter of backup roll; d_w the diameter of work roll neck; d_b the diameter of backup roll neck; L_w the half-width of work roll barrel; L_b the half-width of backup roll barrel; F the bending force; P the rolling force; q_j the intermediate force between the work roll and backup roll at element j; p_j the rolling force at element j; q_{em} the edge contact force between the upper and down work rolls acting on slab element m; B the strip width; L_1 the central distance between the work roll bending cylinders, L_2 the central distance between the housing screws and L_e the roll edge contact length. This case often occurs during the thin strip rolling, and the rolled strip shape and profile will be affected significantly if the control model is not applicable. Roll edge contact force between the upper and down work rolls will change with different rolling conditions. The delivered thickness distribution of strip depends on the material properties, the reduction of plastic deformation, roll thermal and mechanical crown, roll wear profile, the roll deformations due to the deflection of the rolls, the local contact effect which includes the flattening between the work roll and backup roll, the flattening between the work roll and strip and the edge contact of the work rolls. The edge contacts of the work rolls affect the deformation of the rolls and the strip shape, thus forming a new deformation feature including strip plastic deformation and roll deformation in the cold rolling process. In this case, the models of deformation and plasticity are different from the traditional analysis in plasticity in cold strip rolling. Not only will the distribution of the roll pressure change when the work rolls contact beyond the edges of the strip, but also the plastic deformation of strip and the deformation model of work rolls (Edwards & Spooner, 1973, Kuhn & Weinstein, 1970), friction variation at the interface of the rolls and the strip, and work roll wear (Lenard, 1992, 1998, Liu et al., 2001, Jiang & Tieu, 2001). Sutcliffe et al. (1998, 2001) developed a robust model for rolling of thin strip and foil and carried out experimental measurements of load and strip profile. A comparison of roll torque and lateral spread was also conducted for thin strip rolling (Shi et al., 2001). The real contact area is relevant to the contact friction coefficient. Stupkiewicz & Mroz (2003) developed a phenomenological model to calculate the real contact area accounting for bulk plastic deformation in metal forming. How to determine the distribution of rolling force and the strip shape and to find a method to improve its shape and profile when the work rolls contact beyond the strip edges are the main features of this study. The effect of the strip width and transverse friction on the roll edges contact length, the rolling force and strip shape will be quantified and discussed in this study.

In this chapter, a modified semi-infinite body model was introduced to calculate the flattening of work roll/backup roll, work roll/strip, and the Foppl model (Ginzburg & Azzam, 1997) was employed to simulate the edge contact between the upper and down work rolls. Based on the theory of the slit beam, this special cold rolling of thin strip was calculated using an influence function method. The rolling force determined from the plasticity of metal forming was iterated in the simulation, and the analysis of the mechanics of the rolls is for dealing with the plasticity of this special rolling through factors such as the

rolling force and torque, and the strip plastic deformation such as the strip shape and profile. A comparison of the forces and the strip shape with or without the work roll edge contact was carried out. The effects of the different rolling parameters, such as the reduction, strip width, friction coefficient and transverse friction distribution, on the mechanics and deformation of the cold rolling of thin strip were analyzed. The developed model is useful in improving the shape and flatness quality of thin strip in cold rolling process. Based on the developed method and research results, a comprehensive model which is suitable for considering the work roll edge contact or not can be obtained.

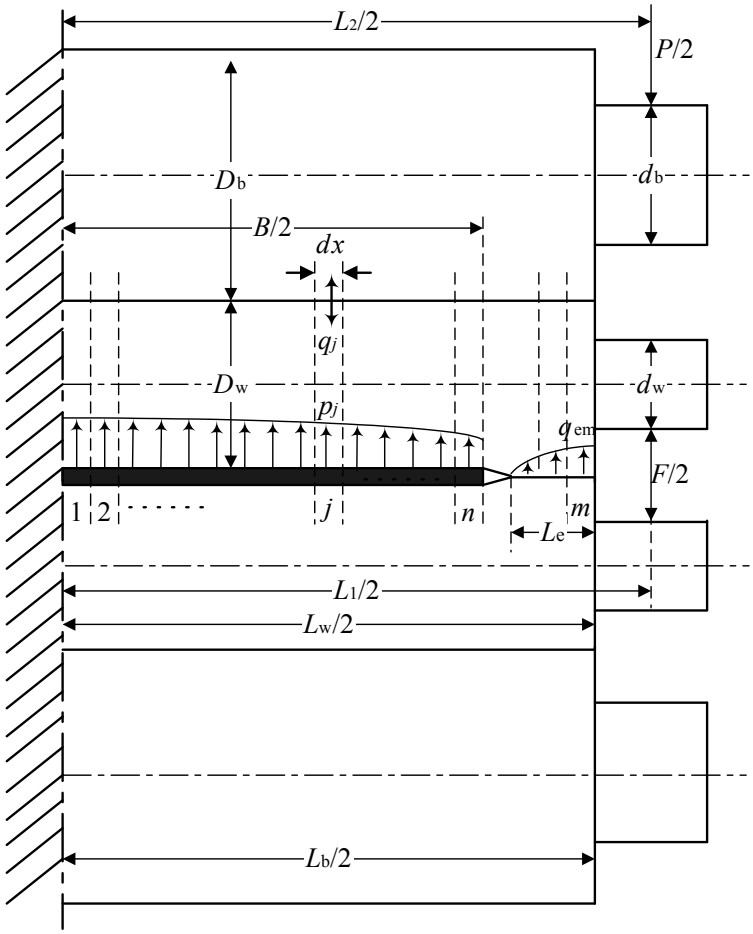


Fig. 1. Edge contact of work rolls and Slit beam deflection mode for a 4-high rolling mill

2. Roll deformation of 4-high rolling mill

As symmetry of the mill about the mid-span of the rolls, the calculation process involves one half of a roll system. Given in Fig. 1, one half of the roll barrel and the strip are divided into

m and n slab elements respectively. The rolling pressure, the pressure between the work roll and backup roll, and that between the upper and down work rolls are uniform in each element, which are replaced by a concentrated load applied to the middle of each element. The profile of the deformed work roll and backup roll are obtained by calculating the roll deflections due to bending and shear forces. Local deformations due to the flattening in the contact region between the work roll and backup roll, between the work roll and strip, and between the upper and down work rolls are added to the roll deflections.

As shown in Fig. 2, if $G(x,x')$ is the deformation of the beam at position x caused by a unit load which is applied to the beam at position x', the deformation of the beam at position x caused by an arbitrary load distribution along the beam can be calculated by the following equation,

$$y(x) = \int G(x,x') \cdot p'(x')dx' \tag{1}$$

where $G(x,x')$ is the influence function in the linear mechanical field. If the load distribution is handled as a number of concentrated loads at the middle of each element, Eq. (1) can be expressed as

$$y(i) = \sum_{j}^{m} g(i,j)p_j \tag{2}$$

where the influence function, $g(i,j)$, is defined as the deflection in the middle of the element i due to a unit load applied to the middle of the element j. The deformation, $y(i)$, in Eq. (2) not only indicates the deflection of the roll, but also represents the flattening of the contact zone.

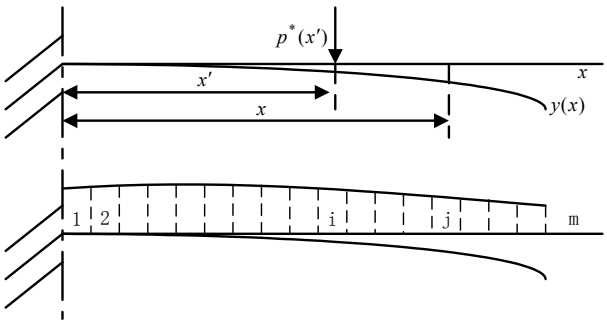


Fig. 2. Deformation of the roll from an arbitrary load distribution

2.1 Deflection of the work rolls

The deflection of the work rolls due to bending and shear forces can be described as the vertical displacement of the work roll at element i by the following equation,

$$y_w(i) = \sum_{j=1}^{m} g_w(i,j)q_j - \sum_{j=1}^{n} g_w(i,j)p_j - \sum_{j=k}^{m} g_w(i,j)q_{ej} - g_{wf}(i)F \tag{3}$$

where $y_w(i)$ is the vertical deflection of the work roll at element i, $g_w(i,j)$ is the influence function for the work roll deflection due to the combined bending and shear forces generated by rolling load, $g_{wf}(i)$ is the influence function for the work roll deflection due to the force generated by the roll bending mechanism, k represents the minimum element number of the edge contact length caused by the flattening between the upper and down work rolls.

The exact forms of the influence functions $g_w(i,j)$ and $g_{wf}(i)$ are described in Reference (Edwards & Spooner, 1973) according to the theorem of Castigliano (Timoshenko & Goodier, 1970), which can be defined as follows

$$\begin{aligned} g_w(i,j) &= \frac{1}{6E_wI_w}(x_j^2(3x_i - x_j) + (1+\nu_w)D_w^2 x_j) \qquad x_i \geq x_j \\ &= \frac{1}{6E_wI_w}(x_i^2(3x_j - x_i) + (1+\nu_w)D_w^2 x_i) \qquad x_i < x_j \end{aligned} \tag{4}$$

$$g_{wf}(i) = \frac{1}{6E_wI_w}(x_i^2(3L_1 - x_i) + (1+\nu_w)x_i D_w^2) \tag{5}$$

where E_w is the work roll modulus of elasticity, ν_w is Poisson's ratio of the work roll and I_w is the moment of inertia of the work roll section.

2.2 Deflection of the backup rolls

The deformation of the backup roll can be expressed as the vertical displacement of the backup roll at the i th element,

$$y_b(i) = \sum_{j=1}^{m} g_b(i,j)q_j \tag{6}$$

where $y_b(i)$ is the vertical deflection of the backup roll at element i, $g_b(i,j)$ is the influence function for the backup roll deflection, which has been derived by Edwards & Spooner (1973)

$$\begin{aligned} g_b(i,j) &= \frac{1}{6E_bI_b}(3x_i^2(L_b - x_j) - (x_i - x_j)^3 + (1+\nu_b)(x_i - x_j)D_b^2) \qquad x_i \geq x_j \\ &= \frac{1}{6E_bI_b}(3x_i^2(L_2 - x_j)) \qquad x_i < x_j \end{aligned} \tag{7}$$

where E_b is the backup roll modulus of elasticity, ν_b is Poisson's ratio of the backup roll and I_b is the moment of inertia of the backup roll section.

2.3 Flattening between the work roll and strip

As shown in Fig. 3, the infinite plane π is the boundary of a semi-infinite body, the upper side of the plane π is the semi-infinite space and the down side is the semi-infinite solid. When a force P acts on this plane at position O, the vertical displacement at point B produced by the force P is as follows

$$w = \frac{P}{2\pi E}\left[(1+\nu)z^2(r^2+z^2)^{-3/2} + 2(1-\nu^2)(r^2+z^2)^{-1/2}\right] \tag{8}$$

where E and v are the modulus of elasticity and Poisson's ratio of the semi-infinite body respectively.

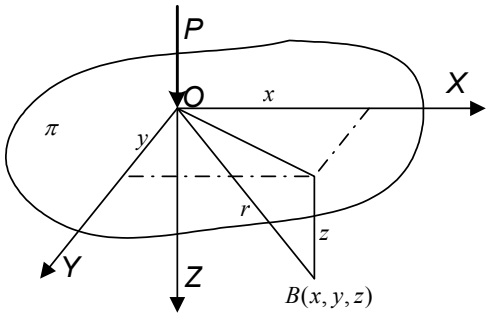


Fig. 3. Semi-infinite body model

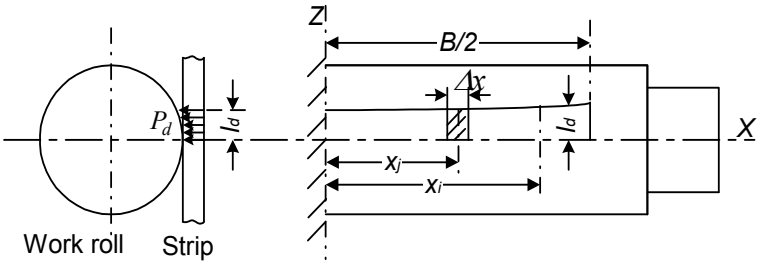


Fig. 4. Flattening between the work roll and strip

Given in Fig. 4, the projected arc of contact between the work roll and strip, l_d, is not a constant and changes along the strip width that can be deduced by Hitchcock model (Wang, 1986)

$$l_d = \sqrt{R_w(\Delta h_i + \frac{16p_i}{\Delta x_i}(\frac{1-\nu_w^2}{\pi E_w})} \tag{9}$$

where Δh_i is the absolute reduction at element i, R_w is the radius of the work roll. Due to the maximum value of l_d is far less than the work roll diameter, so the work roll can be approximatively treated as a semi-infinite body. The influence function for the flattening between the work roll and the strip was derived according to a semi-infinite body model (Wang, 1986). However, the roll is not a real semi-infinite body, a modified model and the flattening between the work roll and strip can be calculated by

$$y_{ws}(i) = \sum_{j}^{n} g_{ws}(i,j)p_j \tag{10}$$

where y_{ws} is the flattening between the work roll and strip, the influence function for the flattening between the work roll and strip can be written as

$$g_{ws}(i,j)=\Phi\left(\left|x_i-x_j\right|\right)+\Phi\left(\left|x_i+x_j\right|\right) \tag{11}$$

where $\Phi(x)$ can be written as Eq. (12).

$$\begin{aligned}\Phi(x)=\frac{1-\nu_w{}^2}{E_w\pi\Delta x}\Bigg\{&\ln\frac{\sqrt{l_d{}^2+(x+\Delta x/2)^2}+x+\Delta x/2}{\sqrt{l_d{}^2+(x-\Delta x/2)^2}+x-\Delta x/2}\\&+\frac{x+\Delta x/2}{l_d}\ln\frac{\sqrt{l_d{}^2+(x+\Delta x/2)^2}+l_d{}^2}{|x+\Delta x/2|}\\&-\frac{x-\Delta x/2}{l_d}\ln\frac{\sqrt{l_d{}^2+(x-\Delta x/2)^2}+l_d{}^2}{|x-\Delta x/2|}\\&-\frac{1}{2(1-\nu_w)}\left[\frac{x+\Delta x/2}{\sqrt{R_w{}^2+(x+\Delta x/2)^2}}-\frac{x-\Delta x/2}{\sqrt{R_w{}^2+(x-\Delta x/2)^2}}\right]\\&-\ln\frac{\sqrt{R_w{}^2+(x-\Delta x/2)^2}-(x-\Delta x/2)}{\sqrt{R_w{}^2+(x+\Delta x/2)^2}-(x+\Delta x/2)}\Bigg\}\end{aligned} \tag{12}$$

2.4 Flattening between the backup roll and work roll

Fig. 5 shows the flattening between the backup roll and work roll. It can also be found from Foppl model (Ginzburg & Azzam, 1997) that the flattening contact width between the backup roll and work roll is far less than the diameters of the work roll and backup roll, and it is suitable to calculate the flattening according to a semi infinite body model (Wang, 1986).

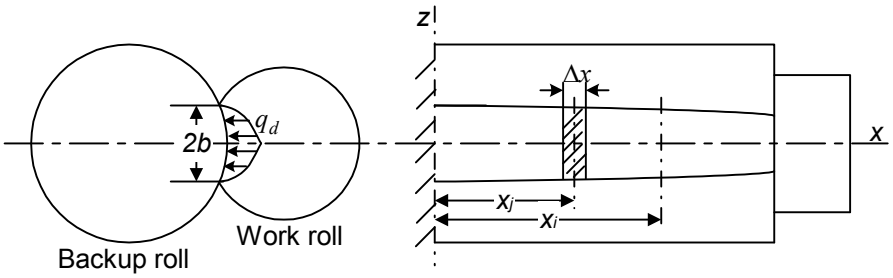


Fig. 5. Flattening between the backup roll and work roll

Assuming the contact pressure between the backup roll and work roll, q_d, is a parabolic distribution along the flattening contact width, $2b$ (see Fig. 6). As the backup roll and work roll are flattened at the same time, the flattening between the backup roll and work roll can be expressed as

$$y_{wb}(i)=\sum_{j}^{m} g_{wb}(i,j)q_j \tag{13}$$

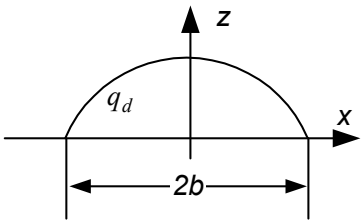


Fig. 6. Distribution of contact pressure along the contact width

where y_{wb} is the flattening between the work roll and backup roll, the influence function is given

$$g_{wb}(i,j)=F_w\left(\left|x_i-x_j\right|\right)+F_w\left(\left|x_i+x_j\right|\right)+F_b\left(\left|x_i-x_j\right|\right)+F_b\left(\left|x_i+x_j\right|\right) \tag{14}$$

where $F(x)$ is a complex function derived by Wang (1986). $F_w(x)$ and $F_b(x)$ can be obtained when the corresponding values for the work roll and backup roll parameters respectively are introduced in Eq. (15).

$$\begin{aligned}F(x)=\frac{1-\nu^2}{E\pi\Delta x}\Bigg\{&\frac{3}{4b}\Bigg[2b\ln\frac{\sqrt{b^2+(x+\Delta x/2)^2}+x+\Delta x/2}{\sqrt{b^2+(x-\Delta x/2)^2}+x-\Delta x/2}+2(x+\Delta x/2)\ln\frac{\sqrt{b^2+(x+\Delta x/2)^2}+b}{|x+\Delta x/2|}\\&-2(x-\Delta x/2)\ln\frac{\sqrt{b^2+(x-\Delta x/2)^2}+b}{|x-\Delta x/2|}+\frac{1}{3b}(x-\Delta x/2)\sqrt{b^2+(x-\Delta x/2)^2}\\&-\frac{1}{3b}(x+\Delta x/2)\sqrt{b^2+(x+\Delta x/2)^2}-\frac{2}{3}b\ln\frac{\sqrt{b^2+(x-\Delta x/2)^2}-(x-\Delta x/2)}{\sqrt{b^2+(x+\Delta x/2)^2}-(x+\Delta x/2)}\\&+\frac{1}{6b^2}(x+\Delta x/2)^3\ln\frac{\sqrt{b^2+(x+\Delta x/2)^2}+b}{\sqrt{b^2+(x+\Delta x/2)^2}-b}-\frac{1}{6b^2}(x-\Delta x/2)^3\ln\frac{\sqrt{b^2+(x-\Delta x/2)^2}+b}{\sqrt{b^2+(x-\Delta x/2)^2}-b}\Bigg]\\&-\frac{1}{2(1-\nu)}\Bigg[\frac{x+\Delta x/2}{\sqrt{R^2+(x+\Delta x/2)^2}}-\frac{x-\Delta x/2}{\sqrt{R^2+(x-\Delta x/2)^2}}\Bigg]-\ln\frac{\sqrt{R^2+(x-\Delta x/2)^2}-(x-\Delta x/2)}{\sqrt{R^2+(x+\Delta x/2)^2}-(x+\Delta x/2)}\Bigg\}\end{aligned} \tag{15}$$

2.5 Flattening between the upper and down work rolls

The work roll contacts at the edges outside the strip width are considered here as a new analysis of the cold rolling of thin strip. As shown in Fig. 7, the contact length between the upper and down work rolls due to strip plastic deformation, L_e, is far less than the length of roll barrel, it is not suitable for calculation by a semi-infinite body model. In the mean time, it is difficult to satisfy the accuracy during the iterative loop by using a semi-infinite body model due to few elements touching at the edges of the work rolls. Based on the study of

Wang (1983), when the modulus of the elasticity of the upper work roll is equal to that of the down work roll, i. e. $E_{wu} = E_{wl} = E$, the flattening between a pair of work rolls can be given directly by

$$y_{ww}(k) = \frac{2q_{ek}(1-v_w{}^2)}{\pi E_w}\left(\frac{2}{3} + \ln\frac{2R_{wu}}{b} + \ln\frac{2R_{wl}}{b}\right) \tag{16}$$

where R_{wu} and R_{wl} are the radii of the upper and down work rolls respectively, b is a half of flattened contact width between the upper and down work rolls, which is given by

$$b(k) = \sqrt{\frac{8(1-v_w^2)q_{ek}}{\pi E_w}\frac{R_{wu}R_{wl}}{(R_{wu}+R_{wl})}} \tag{17}$$

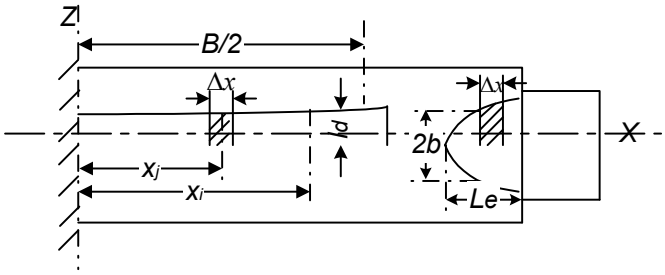


Fig. 7. Flattening between the upper and down work rolls

2.6 Contour of compatibility

Under a rolling load, the deformation of the work roll, backup roll and the strip are described in Fig 8. Compatibility for the contact of the work roll and backup roll varies with the sum of the contour of the deformed work roll and backup roll, and the local flattening of the rolls. It can be calculated by the relationship

$$y_{wb}(i) = y_{wb}(0) + y_b(i) - y_w(i) - m_b(i) - m_w(i) \tag{18}$$

where $y_{wb}(0)$ is the centreline value of flattening between the work roll and backup roll, $m_b(i)$ and $m_w(i)$ are the combined machined and thermal cambers of the backup roll and work roll at element i respectively, m_b and m_w are the combined machined and thermal cambers of the backup roll and work roll at the centre of the roll barrel respectively.

The contour of the work roll surface in contact with the strip is determined by the combined influence of the rolling load, machined and thermal crown and the local flattening between the work roll and strip. The exit thickness of the strip at any point is the same as the loaded gap height at that point. Thus, the compatibility for contact of the work roll and strip can be expressed as

$$h(i) = h(0) + 2\cdot(y_{ws}(i) - y_{ws}(0)) + 2\cdot(m_w(i) - y_w(i)) \tag{19}$$

where $y_{ws}(0)$ is the centreline value of flattening between the work roll and strip.

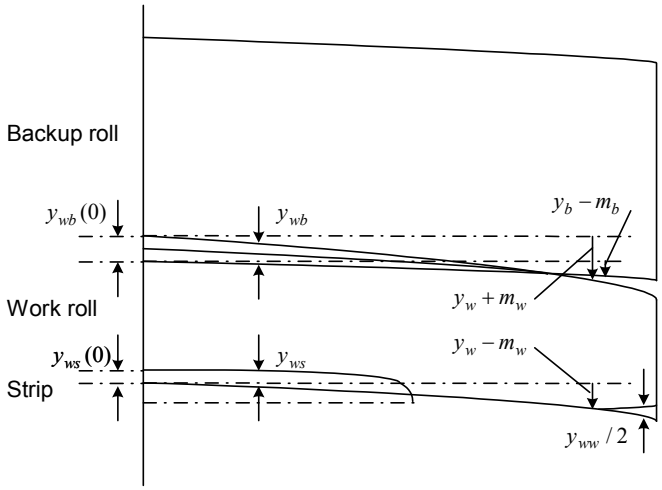


Fig. 8. Compatibility for contact of the work roll and backup roll, the work roll and strip, the upper and down work rolls

In the rolling of thin strip, the sides of work rolls beyond the strip width may touch and deform. The edge contacts between the upper and down work rolls affect the deformation of the roll and the strip. The compatibility for edge contact of the upper and down work rolls is calculated from the deformed work roll profile and the centreline value of the flattening between the work roll and strip, which is written as Eq. (20).

$$y_{ww}(i) = 2 \cdot y_{ws}(0) - h(0) + 2 \cdot (y_w(i) - m_w(i)) \tag{20}$$

2.7 Static equilibrium of work roll

Static equilibrium of the work roll is obtained by summing vertically the load between the work roll and backup roll, the load between the work roll and strip, the load between the upper and down work rolls, and the load applied to the work roll by the work roll bending mechanism. It can be expressed as Eq. (21).

$$\sum_{i=1}^{m} q_i - \sum_{i=1}^{n} p_i - \sum_{i=k}^{m} q_{ei} - F/2 = 0 \tag{21}$$

2.8 Solution of equations

The iterative method was used to calculate the roll and strip deformation and the strip shape, as shown in Fig. 9.

3. Simulation conditions

The parameters used in the simulation for cold rolling of thin strip are as follows.

Work roll diameter: 63 mm	Work roll barrel: 249 mm
Work roll crown: 0 μm	Poisson's ratio of work roll: 0.3
Young's modulus of work roll: 22000 Kg/mm^2	Distance between housing screw: 340 mm

Backup roll diameter: 228 mm
Backup roll barrel: 249 mm
Backup roll crown: 0 μm
Poisson's ratio of backup roll: 0.3
Young's modulus of backup roll: 22000 Kg/mm^2
Slab thickness of strip: 0.5 mm
Central distance between bending cylinder: 340 mm
Entry thickness of strip: 0.30 mm
Exit thickness of strip: 0.10 mm
Width of strip: 140 mm
Back tension: 0 kN
Front tension: 0 kN
Rolling speed: 1 m/s
Initial crown of strip at entry: 0.0 mm
Friction coefficient: 0.1
Work roll bending force: 0 kN/chock
Defining point of strip crown from edge: 10 mm

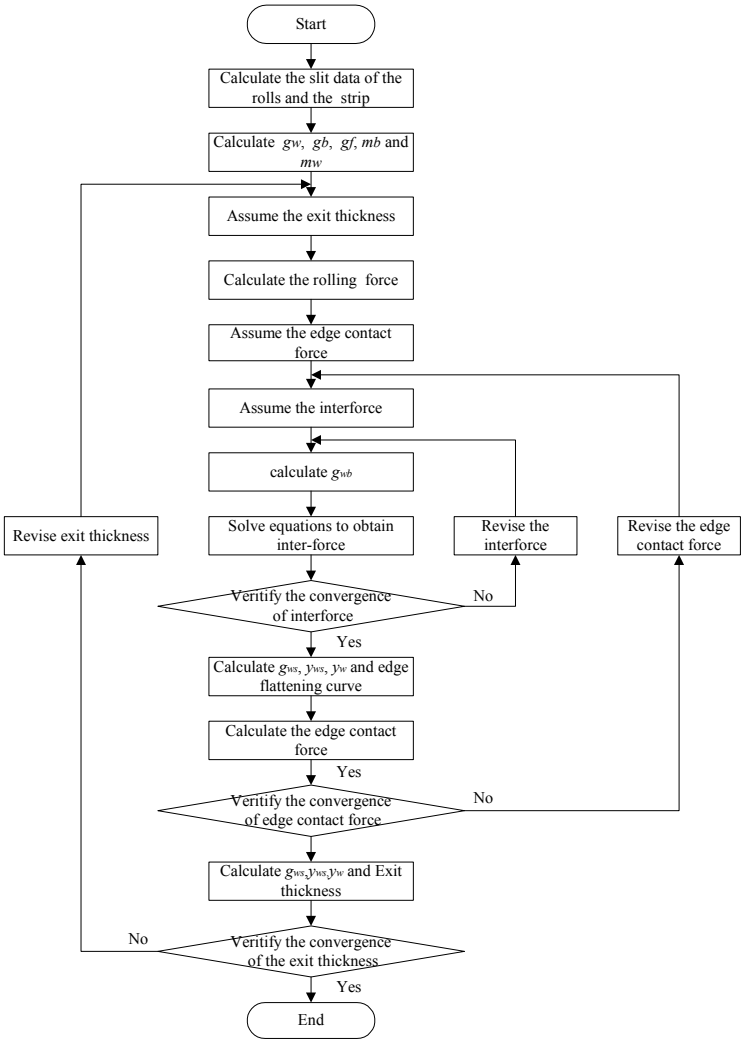


Fig. 9. Flow chart of the roll and strip deformation and strip shape calculation

Deformation resistance equation for strip is written as Eq. (22).

$$k_s = 740(\varepsilon_m + 0.01)^{0.23} \quad \text{(MPa)} \tag{22}$$

The slab thickness is 0.5 mm and the exit thickness of strip is 0.12 mm. The front and back tension is zero. The effects of the different rolling parameters such as the reduction, strip width, friction coefficient and the friction distribution along the strip width, on the mechanics and deformation of the cold rolling of thin strip are analyzed when an influence of edge contact force of the work rolls is considered. In the calculation, a significant concern is the rolling force, which is determined from the plasticity of the metal forming. The calculated rolling force is employed first, and then the further iterations are carried out.

For carbon steel rolling, the rolling force is calculated by using Bland-Ford-Hill model (Wang, 1983) considering the strip plastic deformation.

$$F = B \cdot k_p \cdot \sqrt{R'\Delta h} \cdot D_P \cdot \kappa \tag{23}$$

where κ is the tension factor, k_P is the dynamic deformation resistance which can be described by Eq. (24)

$$k_p = k_s \cdot (1000 \cdot \dot{\varepsilon})^{\alpha} \tag{24}$$

where α is a constant, $\dot{\varepsilon}$ is the stain rate and

$$k_s = k_0 \cdot (\varepsilon_m + m_1)^{n_1} \tag{25}$$

where k_0 is a constant, in this simulation $k_0 = 740$ MPa, m_1 and n_1 are constant, $m_1 = 0.01$ and $n_1 = 0.23$, ε_m is an average integral reduction which can be described as

$$\varepsilon_m = \ln \frac{H_1}{h_m} \tag{26}$$

where H_1 is slab thickness and

$$h_m = (1-\beta) \cdot H + \beta \cdot h \tag{27}$$

where β is a constant (0.75). R' is the flatten radius of work roll which can be deduced by Hitchcock model

$$R' = R \cdot \left\{ 1 + \frac{C_H \cdot P}{B \cdot (H-h)} \right\} \tag{28}$$

where H and h are the entry and exit thickness of strip, respectively, C_H is Hitchcock coefficient (Wang, 1983). D_P can be described as

$$D_P = 1.08 + 1.79\varepsilon f \sqrt{R'/H} - 1.02\varepsilon \tag{29}$$

where ε is the reduction and f the friction coefficient.

4. Results and discussion

4.1 Effect of edge contact on specific forces and strip profile

When the entry and exit thickness of strip is 0.2 mm and 0.12 mm respectively, friction coefficient 0.1, strip width 160 mm and the work roll bending force is zero. The calculated results such as the exit thickness distribution of strip along the strip width and the specific force distribution between the work roll and backup roll, between the upper and down work rolls, between the work roll and strip along the roll barrel with or without edge contact of the work rolls are shown in Fig. 10 and Table 1. It can be seen that the intermediate force closer to the edge of the roll barrel increases and the rolling force close to the side of the strip reduces due to the work roll edge contact. The maximum edge contact force at the edge of the roll barrel is larger than the backup work roll intermediate force, which will result in further wear of work rolls at this zone. The edge contact force between the upper and down work rolls is nearly 11 % of the rolling force. Due to the effect of the edge contact force of the work rolls, the crown of the strip reduces from 45.56 to 36.54 μm, and the edge contact of the work rolls can improve the strip shape when there is no work roll bending force applied. When the cold thin strip is rolled, the edge contact effect may occur and its effect must be introduced for calculating the roll and strip deformation, strip shape, thus forming a new analysis feature of the rolling process.

Status	Rolling force (kN)	Intermediate force (kN)	Edge contact force (kN)	Crown of exit strip (μm)	Edge contact length (mm)
Edge contact	595.99	661.29	65.30	36.54	36.0
No edge contact	620.25	620.25	0	45.56	0

Table 1. Comparison of specific forces and strip crown with or without edge contact effect

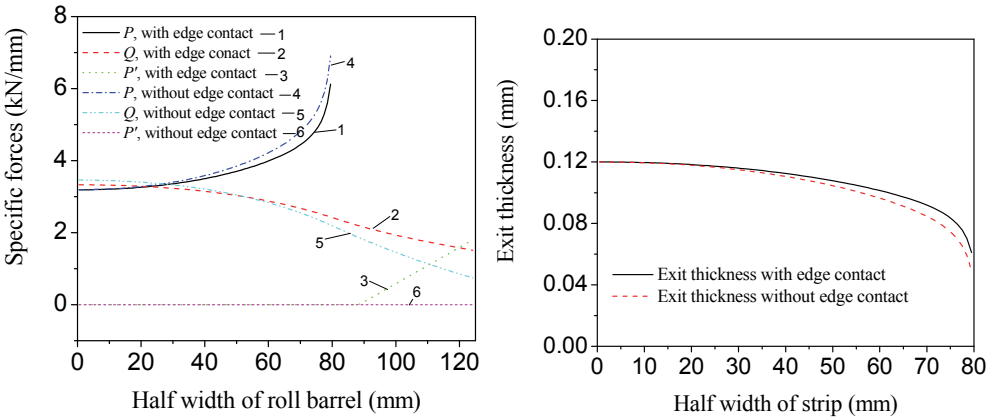


Fig. 10. Comparison of specific forces and exit thickness distribution with or without edge contact effect

4.2 Effect of reduction on forces and strip profiles

When the exit thickness of strip is 0.12 mm, friction coefficient 0.1, strip width 160 mm and no work roll bending force is applied, the entry thickness of the strip are 0.15, 0.17 and 0.2 mm respectively. The effect of the reduction on the specific forces (rolling force, intermediate force and edge contact force) and the exit thickness distribution is shown in Fig. 11 and Table 2. It can be seen that with increasing entry thickness of strip, the rolling force and intermediate force increase significantly, and at the same time the edge contact force and edge contact length of the work rolls have a tendency to increase, which are caused by an increase of reduction. It can also be seen that the strip profile (strip crown) reduces significantly when the reduction decreases (see Table 2). In the simulation, it is found that when the entry thickness is less than 0.1425 mm, which indicates that the reduction is less than 15.8 %, the upper and down work rolls do not touch and the edge contact force is zero. Therefore, under a certain exit thickness of strip, the strip shape and profile become poor with an increase of the reduction although there is a tendency of an increase of the edge contact forces.

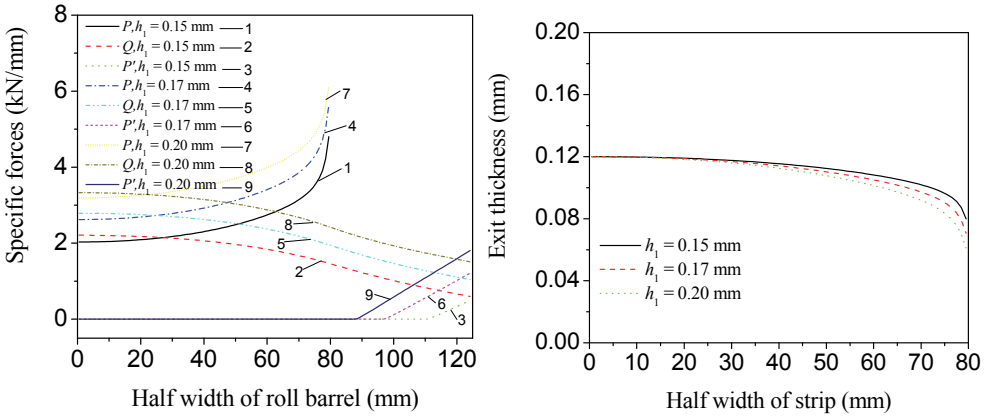


Fig. 11. Effect of entry thickness on specific force and exit thickness distribution

Entry thickness (mm)	Rolling force (kN)	Intermediate force (kN)	Crown of exit strip (μm)	Edge contact force (kN)	Edge contact length (mm)
0.15	401.30	408.09	23.62	6.79	13.0
0.17	504.11	536.81	29.67	32.69	27.0
0.20	595.99	661.29	36.54	65.30	36.0

Table 2. Comparison of specific forces and strip profiles for different entry thickness

4.3 Effect of strip widths on simulation results

The effect of the strip widths on the work roll edge contact force is more complex. In the calculation, the strip entry and exit thickness are 0.17 mm and 0.12 mm respectively, the

friction coefficient is 0.1 and the work roll bending force is zero. The strip widths of 80, 100, 120, 140, 160 and 180 mm were introduced in the analysis. The effect of the strip widths on the exit strip thickness distribution is shown in Fig. 12. It can be seen that with a narrower strip, the strip shape and profile are improved and the rolling force, intermediate force and edge contact force also reduce significantly (see Table 3). If the strip width is larger and more close to the edge of the rolls, the deflection of the work roll increases by a combined effect of the change of the distribution of the rolling force and intermediate force when the strip widths vary (see Fig. 13a and b), so the edge contact force and contact length of the work rolls increase accordingly, as shown in Fig. 13c. If the strip widths are less than 100 mm, the deflection of the work roll at the edge and the edge contact force will reduce with a narrower strip. It can be concluded that the strip width has a significant influence on the edge contact force and edge contact length of work rolls, which can result in an unstable work roll edge wear.

4.4 Effect of friction on calculation results

The lubrication and friction of the strip is a key issue in cold rolling process. The values of the friction coefficient may change significantly in different cold rolling mills and different operating conditions. The effects of the different friction coefficients on the edge contact of the work rolls are shown in Table 4 and Fig. 14. The entry and exit thickness of strip are 0.17 and 0.12 mm respectively, the strip width is 160 mm, and no work roll bending force is applied. With an increase of friction coefficient, the rolling force, intermediate force and edge contact force of the work rolls increase significantly. On the other hand, the strip shape reduces to 25.91 μm from 31.61 μm when the friction coefficient increases from 0.07 to 0.13. Although a higher rolling force has a tendency to make the strip shape poorer, the increase of edge contact forces with friction coefficient has a major effect on the improvement of the strip profile. Therefore, when the friction coefficient increases, the edge contact force of the work rolls increases, which is helpful in improving the strip profile.

The friction coefficient along the strip width is not a constant due to the change along the strip width of operating parameters, i.e. the rolling force, reduction and rolling speed etc.

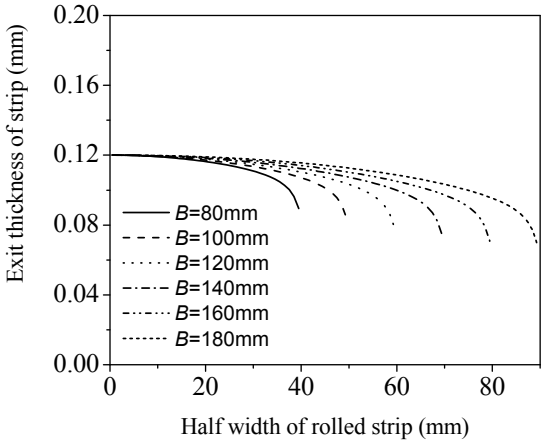
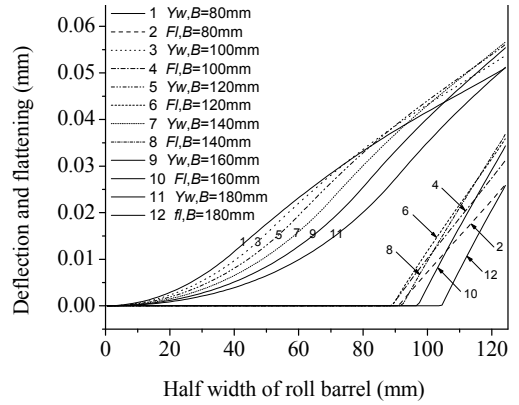
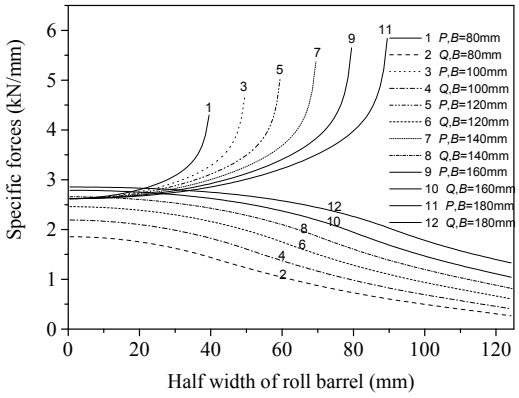


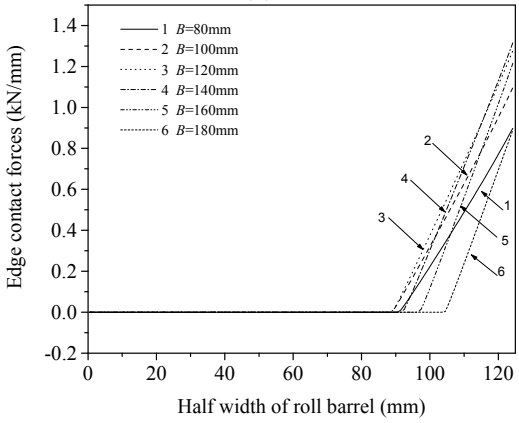
Fig. 12. Effect of the strip widths on the distribution of exit strip thickness



(a)



(b)



(c)

Fig. 13. Effect of the strip widths on the roll deflection and flattening (a), specific forces (b) and edge contact forces (c)

However, there are no reports on the effect of the friction variation along the strip width due to the complexity of this problem. In this section, the entry thickness of strip is 0.30 mm, exit thickness of strip 0.10 mm, strip width 140 mm, back tension 0 kN, front tension 0 kN, metal flow resistance is described by Eq. (22), and no work roll bending force is applied, and the different transverse friction coefficients were assumed to be constant f, parabolic increasing f_i and parabolic decreasing f_d along the strip width as shown in Fig. 15.

Width of strip (mm)	Rolling force (kN)	Intermediate Force (kN)	Crown of exit strip (μm)	Edge contact force (kN)	Edge contact length (mm)
80	235.39	264.05	14.48	28.66	33.0
100	300.27	337.74	18.78	37.48	35.0
120	367.83	411.52	23.05	43.68	35.0
140	435.91	478.38	26.70	42.47	32.0
160	504.11	536.81	29.67	32.69	27.0
180	569.96	587.76	31.65	17.80	20.0

Table 3. Comparison of specific forces and strip crowns with different strip widths

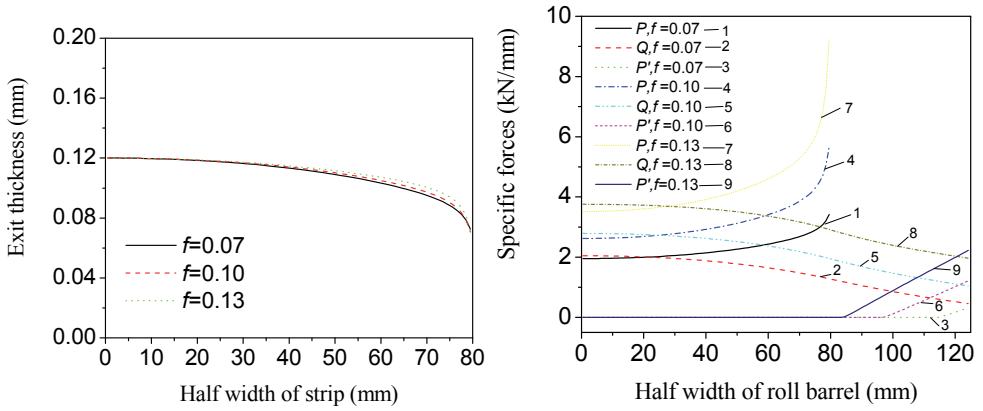


Fig. 14. Effect of friction coefficient on exit thickness distribution and specific forces

Friction coefficient	Rolling force (kN)	Intermediate force (kN)	Crown of exit strip (μm)	Edge contact force (kN)	Edge contact length (mm)
0.07	362.04	365.20	31.61	3.17	9.0
0.10	504.11	536.81	29.67	32.69	27.0
0.13	686.28	776.22	25.91	89.94	40.0

Table 4. Comparison of specific forces and strip crowns with different friction coefficients

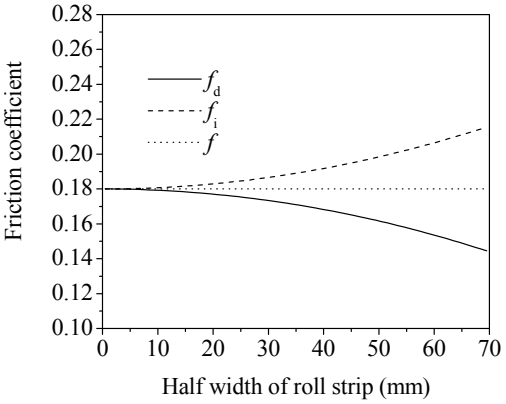


Fig. 15. Distribution of friction coefficient along the strip width

The effects of the transverse friction on the strip profile and specific forces i.e. the rolling force, intermediate force and edge contact force are shown in Figs. 16 and 17. It can be seen that the transverse friction has a significant effect on the strip profile. As the friction coefficient at the edge of strip increases, the exit crown of strip reduces, which indicates that the strip shape becomes better. Thus the strip shape and profile of thin strip can be improved by increasing the edge friction along the strip width. It can also be seen that the rolling force P increases significantly with the friction coefficient at the edges of strip. The intermediate force Q and edge contact force P' increase substantially with a higher friction coefficient at the edge of strip, and the length of edge contact also increases. Therefore, the length of edge contact can be determined from this developed model, which is helpful in understanding the feature of the thin strip rolling with work roll edge contact.

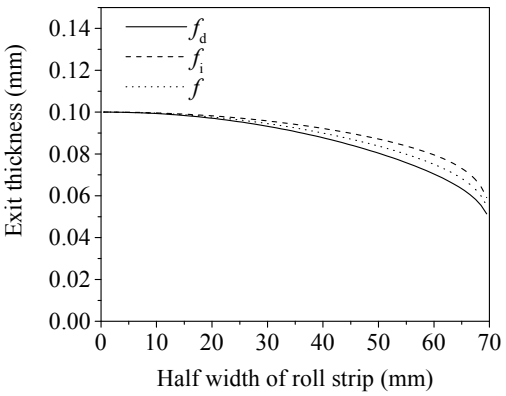


Fig. 16. Effect of transverse friction on strip shape

In order to verify the simulation results, the clod rolling of thin strip was conducted in lab. When the rolling speed is 0.27 m/s, entry thickness of strip is 0.55 mm, exit thickness 0.12 - 0.17 mm, strip widths 100 - 160 mm, a low carbon steel was rolled on Hille 100 rolling mill, friction coefficient is 0.1, its deformation resistance as described in Eq. (30) replaces Eq. (22) in the simulation.

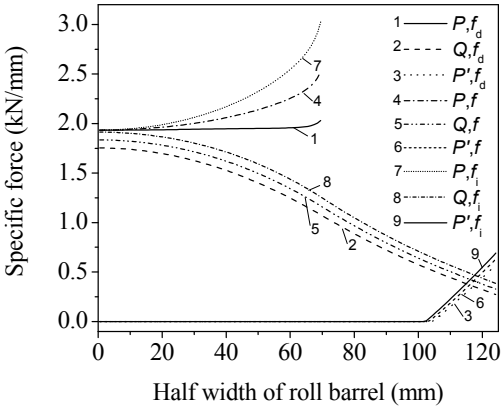


Fig. 17. Effect of transverse friction on specific forces

$$k = 403.53 \cdot (\varepsilon_m - 0.0067)^{0.1271} \times (1000\dot{\varepsilon})^{0.0123} \tag{30}$$

Other rolling mill parameters are the same as above. Comparison of the calculated rolling forces with the measured values for various strip widths is shown in Fig. 18. It can be seen that the calculated rolling force increases with the strip width, and it is in good agreement with the measured value. Fig. 19 shows a comparison of the measured rolling force with the calculated rolling force under various strip widths and rolling speeds. It can be seen that the rolling force decreases significantly with an increase of the rolling speed. The variation of interference friction features between the roll and strip under various rolling speeds is the main reason for this result. The calculated rolling force is in good agreement with the measured value, which verifies the plastic deformation model we have developed for this thin strip rolling. At lower rolling speeds, the work roll edge contact force becomes higher.

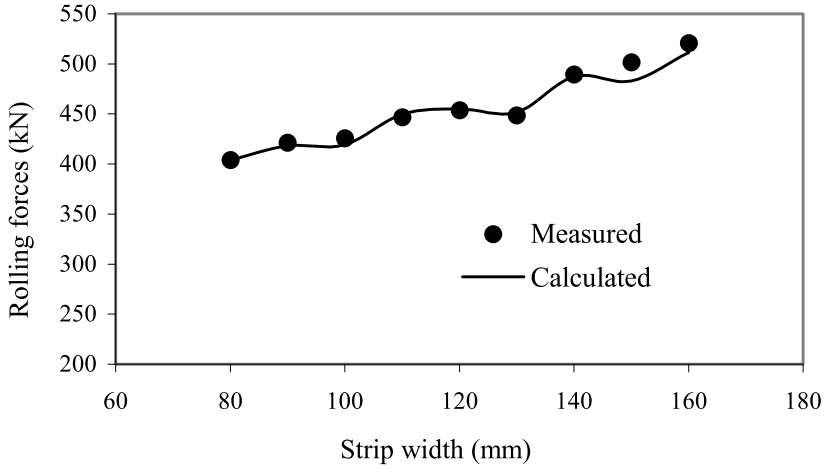


Fig. 18. Comparison of the calculated rolling forces with the measured values

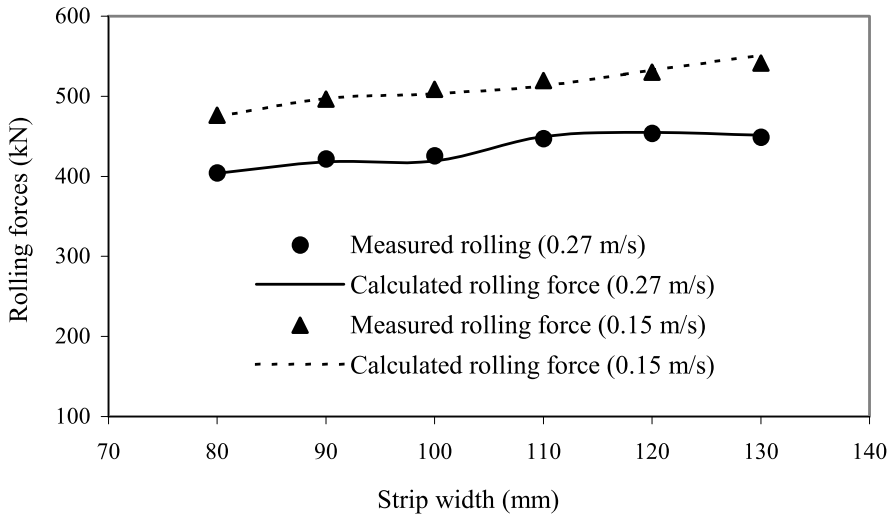


Fig. 19. Effects of the rolling speeds and strip widths on rolling forces

The average percentages of work roll edge contact force with respect to the total rolling force for various strip widths are 15.0 and 15.6 % for rolling speeds of 0.27 and 0.15 m/s, respectively. Therefore, the work roll edge contact force plays an important role in this rolling process.

5. Conclusions

A new model for rolling mechanics of thin strip in cold rolling has been developed successfully when the work rolls edge contacts. A strip plastic deformation-based model of the rolling force was employed in the calculation, and a modified semi-infinite body model was introduced to calculate the flattening between the work roll and backup roll, and the flattening between the work roll and strip, as well as a Foppl model was employed to calculate the edge contact between the upper and down work rolls. Based on the theory of the slit beam, the special rolling and strip deformation was simulated using a modified influence function method.

The calculated results show that the specific forces such as the rolling force, intermediate force and the shape and profile of the strip for this special rolling process are significantly different from the forces in the traditional cold rolling process, and those form a new theory of metal plasticity in metal rolling. The edge contact of the work rolls can improve the strip shape when no work roll bending force is applied. With an increase of reduction, the rolling force, intermediate force and edge contact force increase significantly, however the strip shape becomes poor. Strip width has a significant influence on the edge contact force and edge contact length of the work rolls, which can result in an unstable work roll edge wear. When the friction coefficient increases, the edge contact force between the two work rolls increases, this can improve the strip profile. The transverse friction has a significant effect on the rolling force, edge contact force and the length of edge contact. It affects the strip shape

and profile significantly, which is helpful in improving the strip shape and profile by modifying transverse friction. The calculated rolling force increases when the_strip width increases and the rolling speed decreases, and it is in good agreement with the measured value. At lower rolling speeds, the work roll edge contact force becomes higher as a percentage of the total rolling force.

6. Nomenclature

b A half of flattened contact width between the upper and down work rolls
B Width of strip
D_w Diameter of the work roll
D_b Diameter of the backup roll
d_w Diameter of the work roll neck
d_b Diameter of the backup roll neck
E_w , E_b Young's modulus of the work roll and backup roll respectively
fl Roll flattening
F Bending force
f Friction coefficient
H_1 Slab thickness
H Entry thickness of strip
h Exit thickness of strip
I_b Moment of inertia of the backup roll section
I_w Moment of inertia of the work roll section
l_d Projected arc of contact between the work roll and strip
k_0 Constant
k_p Dynamic deformation resistance
k_s Static deformation resistance
L_w Width of the work roll barrel
L_b Width of the backup roll barrel
L_1 Central distance between the work roll bending cylinders
L_2 Central distance between the housing screws
L_e Edge contact length between the upper and down work rolls
m Slab number of half of the roll barrel
m_1 Constant
m_w , m_b Combined machined and thermal cambers of the work roll and backup roll at the centre of the roll barrel respectively
n Slab number of half of the strip
n_1 Constant
P Rolling force
q_j Intermediate force between the work roll and backup roll at element j
p_j Rolling force at element j
q_{em} Edge contact force between the upper and lower works at element m
R' Flatten radius of work roll
R_{wu} Radius of the upper work roll

R_{wl}	Radius of the down work roll
y_w	Vertical deflection of the work roll
y_b	Vertical deflection of the backup roll
y_{ws}	Flattening between the work roll and strip
y_{wb}	Flattening between the work roll and backup roll
y_{ww}	Flattening between the upper and down work rolls
$y_{wb}(0)$	Centreline value of flattening between the work roll and backup roll
$y_{ws}(0)$	Centreline value of flattening between the work roll and strip
Yw	Roll deflection
Δh_i	Absolute reduction at element i
Δx	Width of each element
v_w , v_b	Poisson's ratio of the work roll and backup roll respectively
ε	Reduction
ε_m	Average integral reduction
$\dot{\varepsilon}$	Strain rate
β	Constant

7. Acknowledgements

The work was supported by Australian Research Council (ARC) Discovery-Project grant including an Australian Research Fellowship.

8. References

Buchheit, T.E., Wellman, G.W., Battaile, C.C., 2005. Investigating the limits of polycrystal plasticity modeling. Int. J. Plas. 21(2), 221-249.

Edwards, W.J., Spooner, P.D., 1973. Analysis of strip shape. In: Bryand, G.F. (ed), Automation of Tandem Mills. Iron and Steel Institute, London, p. 177-212.

Ginzburg, V.B., 1989. Steel-rolling technology: theory and practice. Marcel Dekker Inc., New York, pp. 730-748.

Ginzburg, V.B., Azzam, M., 1997. Selection of optimum strip profile and flatness technology for rolling mills. Iron & Steel Engineer 74(7), 30-38.

Han, C.-S., Wagner, R.H., Barlat, F., 2004. On precipitate induced hardening in crystal plasticity: algorithms and simulations. Int. J. Plas. 20(8-9), 1441-1461.

Ho, K.C., Lin, J., Dean, T.A., 2004. Modelling of springback in creep forming thick aluminum sheets. Int. J. Plas. 20(4-5), 733-751.

Huh, M.Y., Lee, K.R., Engler, O., 2004. Evolution of texture and strain states in AA 3004 sheet during rolling with a dead block. Int. J. Plas. 20(7), 1183-1197.

Iwamoto, T., 2004. Multiscale computational simulation of deformation behavior of TRIP steel with growth of martensitic particles in unit cell by asymptotic homogenization method. Int. J. Plas. 20(4-5), 841-869.

Jiang, Z.Y., Tieu, A.K., 2001. A method to analyse the rolling of strip with ribs by 3-D rigid visco-plastic finite element method. J Mater. Process. Technol. 117(1-2), 146-152.

Jiang, Z.Y., Tieu, A.K., 2003a. Modelling of thin strip rolling with friction variation by a 3-D finite element method. JSME Int. 46(A3), 218-223.

Jiang, Z.Y., Tieu, A.K., Zhang, X.M., Lu, C., Sun, W.H., 2003b. Finite element simulation of cold rolling of thin strip. J Mater. Proc. Technol. 140(1-3), 544-549.

Jiang, Z.Y., Zhu, H.T., Tieu, A.K., 2003c. Effect of rolling parameters on cold rolling of thin strip during work roll edge contact. J Mater. Proc. Technol. 140(1-3), 537-543.

Kim, H.-K., Oh, S.-I., 2003. Finite element analysis of grain-by-grain deformation by crystal plasticity with couple stress. Int. J. Plas. 19(8), 1245-1270.

Komori, K., 1998. Analysis of cross and vertical buckling in sheet metal rolling. Int J Mech. Sci. 40(12), 1235-1246.

Kuhn, H.A., Weinstein, A.S. 1970, Lateral distribution of pressure in thin strip rolling. J Eng. for Industry, 453-460.

Le, H.R., Sutcliffe, M.P.F., 2001. A robust model for rolling of thin strip and foil. Int. J. Mech. Sci. 43, 1405-1419

Lenard, J.G., 1992. Friction and forward slip in cold strip rolling. Tribol. Trans. 35(3), 423-428.

Lenard, J.G., 1998. The effect of lubricant additives on the coefficient of friction in cold rolling. J Mater. Process. Technol. 80-81, 232-238.

Lin, Z-C., Lee, S-Y., 1997. An investigation of contact problem between strip and work roll with a smooth straight surface during cold rolling. Int J Mech. Sci. 39(12), 1385-1404.

Liu, Y.J., Tieu, A.K., Wang, D.D., Yuen, W.Y.D., 2001. Friction measurement in cold rolling. J Mater. Process. Technol. 111, 142-145.

Martin, P.H., Smith, L.M., 2005. Practical limitations to the influence of through-thickness normal stress on sheet metal formability. Int. J. Plas. 21(4), 671-690.

Shi, J., McElwain, D.L.S., Langlands, T.A.M., 2001. A comparison of methods to estimate the roll torque in thin strip rolling. Int. J. Mech. Sci. 43, 611-630.

Shohet, K.N., Townsend, N.A., 1968. Roll bending methods of crown control in four-high plate mill. J Iron and Steel Institute 11, 1088-1098.

Smith, L.M., Averill, R.C., Lucas, J.P., Stoughton, T.B., Matin, P.H., 2003. Influence of transverse stress normal stress on sheet metal formability. Int. J. Plas. 19(10), 1567-1583.

Stone, M.D., Gray, R., 1965. Theory and practice aspects in crown control. Iron Steel Eng. XLII(8), 73-83.

Stoughton, T.B., Yoon, J.-W., 2004. A pressure-sensitive yield criterion under a non-associated flow rule for sheet metal forming. Int. J. Plas. 20(4-5), 705-731.

Stupkiewicz, S., Mroz, Z., 2003. Phenomenological model of real contact area evolution with account for bulk plastic deformation in metal forming. Int. J. Plas. 19(3), 323-344.

Sutcliffe, M.P.F., Rayner, P.J., 1998. Experimental measurements of load and strip profile in thin strip rolling. Int. J. Mech. Sci. 40, 887-899.

Timoshenko, S.P., Goodier, J.N., 1970. Theory of elasticity. McGraw-Hill, Third edition, New York.

Wang, G.D., 1986. The shape control and theory. Metallurgical Industry Press, Beijing, pp. 225-379 (in Chinese).
Wang, T., 1983. Rolling technology. Metallurgical Industry Press, Beijing (in Chinese).

21

Performance Evaluation of Single-Channel Receivers for Wireless Optical Communications by Numerical Simulations

M. Castillo-Vázquez, A. Jurado-Navas, J.M. Garrido-Balsells
and A. Puerta-Notario
Communications Engineering Department, University of Málaga
Campus de Teatinos, Málaga
Spain

1. Introduction

In the last decade, the growing demand of high-speed portable computer terminals in office environment is promoting the development of broadband wireless local area networks (WLAN). Within this field, wireless infrared communications represent a viable alternative to radio frequency for short-range indoor transmission, with many significant advantages such as enormous potential bandwidth, security and low component price (Kahn & Barry, 1997). However, system design for indoor WLAN is conditioned by the especially harmful characteristics of the optical channel, i.e. high attenuation, elevated ambient light noise and multipath distortion (Kahn et al., 1994). These inconveniences make infrared systems demand high levels of optical power to achieve a sufficient signal-to-noise ratio (SNR), and limit high-speed performance.

To mitigate such inconveniences, a wide choice of transmitters and receivers designs have been proposed so far (Alves et al., 2000; Bellon et al., 1999; Carruther & Kahn, 2000; Jivkova & Kavehrad, 2001; Jungnickel et al., 2003; Yun & Kavehrad, 1992). Among all, the combined use of multibeam transmitters and angle diversity receivers has become the preferred solution in the literature (Carruther & Kahn, 2000; Djahani & Kahn, 2000; Jivkova & Kavehrad, 2001). In this configuration, the multibeam transmitter (MBT) is used to distribute the optical power within the room, creating a regular lattice of spots on the ceiling where the light is concentrated. Signals from these spots are collected by the multiple optical elements of conventional angle diversity receivers (ADR) and properly combined in order to increase the overall SNR.

There are two main ways to implement an angle diversity receiver. In the conventional approach, the receiver consists of various branches which are oriented in different directions. Each branch comprises a separate nonimaging concentrator and a small photodetector. Since each concentrator only receives signals from a small region of the room, the conventional angle diversity receiver can reduce ambient light noise and multipath distortion by tuning the weights of each branch. The second structure, referred to as imaging angle diversity receiver, consists of a single imaging optical concentrator (e.g., a lens) that forms an image of

the received light on a segmented array of photodetectors (pixels) placed at its focal plane. This receiver has two major advantages over nonimaging angle-diversity receivers. First, all photodetectors share a common concentrator, thus reducing size and cost. Second, all photodetectors can be laid out in a single planar array, enabling the use of a large number of receiving elements or pixels. The use of multibeam transmitters in conjunction with angle diversity receivers (imaging and non-imaging) improves channel bandwidth and received SNR (Carruther & Kahn, 2000; Kahn et al., 1998; Tang et al., 1996).

However, in spite of the advantages of the above mentioned structures, the performance of wireless optical links can be further improved. In this sense, two novel receiver designs, alternative to angle diversity receivers, were proposed by the authors in (Castillo-Vázquez et al., 2004) and (Castillo-Vázquez & Puerta-Notario, 2005), respectively. These new receivers, known as single-channel receivers (SCR), share the same design idea, which is to provide their single optical front-end with self-orienting capability. Thus, they can automatically aim at the optimum ceiling area in terms of illumination. The self-orienting capability of both SCR, together with the very narrow field of view (FOV) employed, drastically reduces path loss, background noise and multipath distortion. Moreover, its single-channel structure minimizes hardware complexity to a minimum and, consequently, also reduces power consumption, compared to ADR which use multiple channels.

In this chapter, the work presented in (Castillo-Vázquez et al., 2004) and (Castillo-Vázquez & Puerta-Notario, 2005) is extended to investigate the impact of both SCR on channel characteristics. By numerical simulations, the main performance indicators of two link configurations, formed by a MBT and the proposed SCR, are obtained. These link simulations are used to analyze the effect of diverse design aspects in the system performance. In particular, two points are investigated (a) the effect of transmitter spots size and ambient light sources (natural and artificial) on SNR and channel bandwidth (BW), and (b) the impact of the receivers total FOV and blockage on the transmitter power requirements. The results obtained in these simulations show the robustness and weaknesses of each receiver structure and prove the great potential of both SCR when operating in a multispot diffusing configuration. The chapter is organized as follows. The second section is focused on the characteristics and structure of single-channel receivers. In the third section, the transmitter and ambient light models employed in the numerical analysis are presented. In section four, the performance evaluation of receivers is carried out. Conclusions are presented in the last section.

2. Single-channel receivers

Fig. 1 depicts the structure of the two single-channel receivers under study. Both receivers use an optical concentrator to increase the received optical signal power, and an optical bandpass filter to reject ambient light. For the sake of clarity, this filter is not shown in the figure.

As shown in Fig. 1 (a), a conventional single-channel receiver (CSCR) consists of an objective lens-photodetector set mounted on an electromechanical orienting system that moves the optical front-end in azimuth ϕ_r and elevation θ_r angles. In this structure, a SNR estimator and a maximum search algorithm are used to automatically aim the receiver at the room area with higher SNR. A CSCR, thus, is able to point to the best ceiling spot, avoiding the areas with strong ambient light. This is based on the fact that this light is normally received from a different direction than the desired signal, as shown in Fig. 1 (a). Thus, this receiver can completely reject the light from artificial lamps if they are placed far from the selected

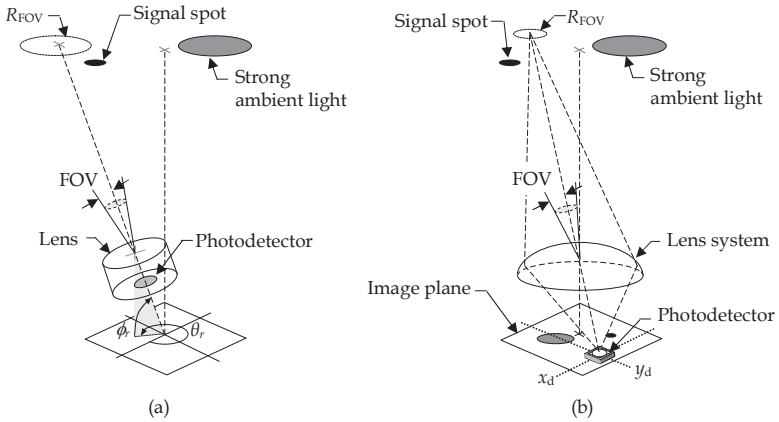


Fig. 1. (a) Conventional single-channel receiver (CSCR). (b) Single-channel imaging receiver (SCIR). The CSCR employs a lens in conjunction with a single photodetector. The SCIR employs an imaging lens system combined with a small photodetector placed on a miniaturized positioning mechanism.

spot. In addition, the receiver can work with an extremely narrow FOV, since it is pointed to the optimum direction, which allows the use of a positive lens as optical concentrator with high gain and, consequently, reducing the detector dimensions. Note that this reduction is in accordance with the purpose of attaining high bit rates. The FOV value optimizing signal and noise levels in the receiver can be obtained geometrically as

$$\mathrm{FOV} = \tan^{-1}(D_{\mathrm{spot}}/2h), \tag{1}$$

where D_{spot} is the diameter of the ceiling spots and h is the distance between receiver and ceiling. Since in many office rooms the receivers are placed at a desktop level (about 90 cm above the floor), it can be assumed that the distance h is approximately 2 m. Moreover, as will be explained later, a spot diameter D_{spot}=10 cm is considered as a reasonable size for our system. Thus, Eq. (1) yields to a receiver FOV of approximately 1.5°. To obtain this FOV, the CSCR uses a positive lens with a 2-cm diameter and an f-number ($f/\#$) of 2 (i.e., $f/2$), in conjunction with a circle photodetector of diameter

$$D_{\mathrm{det}} = 2D\mathrm{f}/\#\tan(\mathrm{FOV}). \tag{2}$$

Hence, a $D_{\mathrm{det}} = 2$ mm is needed to achieve the required FOV. Another important parameter of a CSCR is the total field of view ($\mathrm{FOV}_{\mathrm{total}}$). Such a parameter defines the permitted range of pointing directions for the receiver front-end, depending on the constraints of the electromechanical system. Although, in practice, any value for this parameter is possible, it has been chosen a $\mathrm{FOV}_{\mathrm{total}} = 45°$ in order to keep the blockage probability under reasonable limits. Despite all, if a blockage is produced, receiver must find an alternative spot. Blockage in MBT configurations is further investigated in (Jivkova & Kavehrad, 2003). Note that, with a $\mathrm{FOV}_{\mathrm{total}} = 45°$, the CSCR covers a total ceiling area of 12.5 m^2.

The second receiver structure is depicted in Fig. 1 (b). This structure, referred to as single-channel imaging receiver (SCIR), comprises an imaging lens system combined with a

small photodetector placed on a miniaturized positioning mechanism. Here, the positioning system allows the photodetector to move along the X-Y axes defining the focal plane of the lens system. With such a structure, the receiver can modify its orientation, determined by the elevation and azimuthal angles (θ_r, ϕ_r), changing the photodetector coordinates (x_d, y_d). In a SCIR, the photodetector size defines the receiver FOV and, consequently, the dimensions of the ceiling region R_{FOV}, as shown in Fig. 1 (b). Unlike CSCR, the optical concentrator here is not a simple lens, but a system consisting of several lenses acting as an imaging concentrator. Thus, it images part of the room ceiling onto the receiver image plane. Such an image contains areas with signal light (spot images) and areas with ambient light, as shown in Fig. 1 (b). In our design, a SNR estimator and a maximum search algorithm are used to move the photodetector to the coordinates with better SNR. Note that these coordinates usually coincide with the position of the spot image closest to the receiver. Moreover, since spot images are minute, the receiver can used a very small photodetector to detect them, thus ensuring an optimal ratio between collected signal and noise levels. In a SCIR, the size of the photodetector is chosen equal to the size of spot images. Therefore, its diameter is given by

$$D_{\text{det}} = |M_T|\, D_{\text{spot}}, \tag{3}$$

where M_T is the magnification of the lens system, obtained from the distance to the image plane d_{img} and h as $M_T = -d_{img}/h$. Note that, in this design, d_{img} is approximately the focal length.

To achieve a large total FOV at a low level of optical aberrations, while maintaining a reasonable size of the image plane, a lens system similar to the designs proposed in (Djahani & Kahn, 2000) and (Jivkova et al., 2004), is used. In particular, the lens system used has an aperture diameter of 3 cm and $f/1$. This large aperture is needed to ensure an acceptable signal power in the photodetector. Likewise, to obtain a good image quality and avoid shifting problems in the optical interference filter (Kahn & Barry, 1997), the receiver total FOV is limited to 30°. In these conditions, the distance from the lens system to the image plane is d_{img}=3.05 cm, and the image plane diameter is 35 mm. Assuming, again, a spot diameter D_{spot}=10 cm, Eq. (3) yields to D_{det}= 1.52 mm. With a $\text{FOV}_{\text{total}} = 30°$, the resulting SCIR covers a total ceiling area of approximately 4.1 m^2.

3. Signal and ambient light modeling

In this section, we describe the model of the multibeam transmitter and the model of the light noise sources used in our numerical analysis. Together with this description, the expressions for the signal and light power received on both single-channel receivers, are also deduced. These expressions will be used in section 4 to analyze system performance.

3.1 Multibeam transmitter

As mentioned in the introduction, a multibeam transmitter is formed by multiple narrow beams pointing in different directions towards the ceiling room. These narrow beams project onto the ceiling a spot pattern that can adopt very diverse geometries. In the literature, a great variety of geometries have been proposed so far. These geometries include uniform patterns, in which the spots are organized in regular lattices as those in (Jivkova & Kavehrad, 2005; 2000) and non-uniform patterns, in which the spots are arranged in a more elaborated

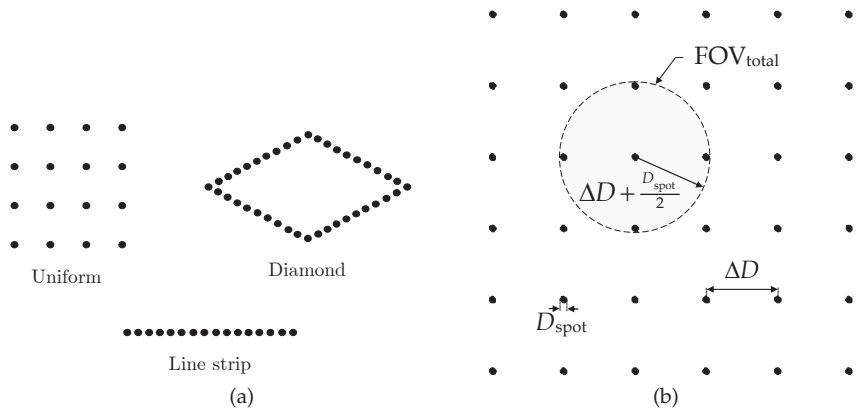


Fig. 2. (a) Transmitter pattern geometries proposed in the literature. (b) Transmitter pattern used in our analysis.

way, forming a strip line as in (Al-Ghamdi & Elmirghani, 2004a) or a diamond shape as in (Al-Ghamdi & Elmirghani, 2004b). Some examples of these geometries are shown in Fig. 2 (a). Although line strip and diamond transmitters offer some advantages when used in combination with pyramidal receivers (Al-Ghamdi & Elmirghani, 2003a;b; 2004a), in our case, uniform patterns are best suited to the operation of single channel receivers. Therefore, for this analysis, we have chosen a uniform transmitter that creates a square mesh pattern on the ceiling as the one depicted in Fig. 2 (b). This pattern is formed by M circular spots of diameter D_{spot} with a spot grid spacing ΔD. In practice, this transmitter can be implemented by using a single laser source together with a holographic or diffractive optical element, as described in (Eardley et al., 1996; Kavehrad & Jivkova, 2003; Pakravan et al., 1996). The choice of ΔD, M and D_{spot} greatly affects system performance. Here, some criteria to choose these parameters are described.

In order to provide immunity against blockage, the transmitter pattern is designed so that at least two signal spots lie within the receiver total FOV. Thus, if the selected spot is obstructed, receivers can point to a different one to avoid a link failure. To fulfill this condition, the grid spacing ΔD must satisfy that $h \tan(\text{FOV}_{\text{total}}) \geq \Delta D + D_{\text{spot}}/2$, as deduced from the geometry of Fig. 2 (b). Thus, from this relation, the condition becomes

$$\Delta D \leq h \tan(\text{FOV}_{\text{total}}) - \frac{D_{\text{spot}}}{2}. \tag{4}$$

According to this expression, a CSCR (with a $\text{FOV}_{\text{total}} = 45°$) would require a transmitter with higher grid spacing ΔD than a SCIR (with a $\text{FOV}_{\text{total}} = 30°$), and, as a consequence, for the same room dimensions, the MBT employed with CSCR would need fewer signal spots than the MBT of a SCIR. For instance, if we assume a rectangular room of X m $\times$ Y m, the minimum number of spots required to provide immunity against blockage is given by

$$M = (\lfloor X/\Delta D \rfloor + 1)(\lfloor Y/\Delta D \rfloor + 1) \tag{5}$$

where $\lfloor\rfloor$ represents the floor function, and ΔD is the maximum grid spacing satisfying Eq. (4).

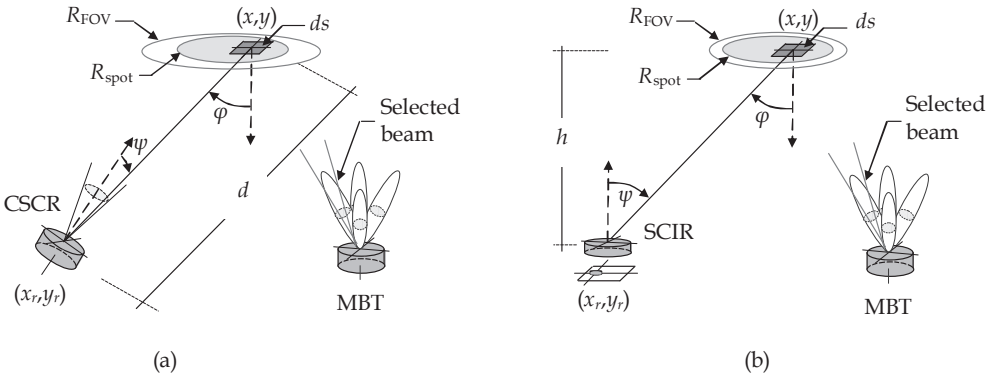


Fig. 3. (a) Link geometry used to compute the received power signal in a conventional single-channel receiver. (b) Link geometry used to compute the received power signal in a single-channel imaging receiver.

After choosing ΔD and M, the last parameter of the considered MBT is the spot diameter D_{spot}. This is the most important parameter, since it affects simultaneously several aspects of system performance. Note that, whenever the spot size is reduced, the SNR and the channel BW are significantly improved. Our results show that channel BW doubles and SNR increases approximately by 6 dB every time the spot diameter D_{spot} is halved. However, such an improvement in SNR and BW is achieved at the expense of complicating the search process in the pointing algorithm, which needs approximately five times the time to find the optimum direction. Thus, a trade-off solution must be found. For this purpose, different simulations in a typical room office are carried out, varying the spot diameter from 5 to 20 cm in steps of 5 cm. Based on the results of these simulations, it is concluded that a spot diameter of 10 cm is needed to achieve a channel BW higher that 1 GHz with a reasonable search time.
In our analysis, the considered MBT emits a total average power P_{tx}, which is equally divided among M beams. The value of M depends on the type of receiver employed (a CSCR or a SCIR), as deduced from Eq. (4) and Eq. (5). Due to the highly collimated nature of beams emerging from the transmitter, the path loss between the transmitter and the ceiling is neglected in our simulations. Thus, all transmitter power is assumed to be concentrated within M small ceiling spots with a diameter $D_{\text{spot}} = 10$ cm. Among these spots, only one (the best) is selected by single-channel receivers and, hence, only one is used in the calculations. To compute the received power from this spot, the covered region of the ceiling, R_{spot}, is divided into differential surface elements ds, as shown in Fig. 3. Then, as in (Carruther & Kahn, 2000), theses elements are modeled as first order Lambertian emitters with a transmit power per steradian given by

$$I_{\text{spot}}(\varphi) = \frac{4\rho P_{tx} ds}{\pi^2 M D_{\text{spot}}} \cos(\varphi), \tag{6}$$

where ρ represents the ceiling reflectivity and φ is the emission angle relative to the normal. It is worth noting here that the assumption that the ceiling behaves as a diffuse reflector (i.e., a Lambertian emitter) is only valid for angles of incidence below 65°, since for higher angles, especially above 70°, the ceiling reflections exhibit a strong specular component, as described in (Gfeller & Bapst, 1979). This condition limits the maximum size of a room illuminated by a single transmitter. For instance, assuming a distance between the transmitter and the ceiling

$h_t = 2$ m, the maximum room diagonal is given by $2h_t \tan(65^\circ) = 8.6$ m (Jivkova & Kavehrad, 2001). Therefore, a square room of 6 m $\times$ 6 m would fulfill this condition.
According to the geometry of Fig. 3, the average power reflected by a ceiling element ds and detected by a SCR can be written as

$$dP_r = \frac{4\rho P_{tx} ds}{\pi^2 M D_{\text{spot}}^2} \cos(\varphi) \cdot \frac{A \cos(\psi)}{d^2}, \tag{7}$$

where A is the receiver entrance area, ψ is the angle between the surface normal to the receiver and the incident ray, and d is the distance between the transmitter and the surface element ds. Here, the value of $\cos(\varphi)$ is h/d, and the distance d can be written as a function of h, the reflector coordinates (x, y) and the receiver position (x_r, y_r) as

$$d = \sqrt{(x - x_r)^2 + (y - y_r)^2 + h^2}. \tag{8}$$

Note that the only difference between the two single-channel receivers when calculating dP_r is the value of the angle ψ. For the CSCR, this angle depends on the receiver orientation (θ_r, ϕ_r), and can be geometrically obtained from Fig. 3 (a) as

$$\psi = \cos^{-1}\left(\frac{(x - x_r)\cos(\varphi_r)\cos(\theta_r) + (y - y_r)\sin(\varphi_r)\cos(\theta_r) + h\sin(\theta_r)}{d}\right). \tag{9}$$

On the other hand, in the case of the SCIR, ψ equals ϕ and, therefore, $\psi = \cos^{-1}(h/d)$, as deduced from Fig. 3 (b).
The total reflected signal incident on the receivers is then computed by integrating the contributions of all surface elements ds within the region illuminated by the selected beam R_{spot}. In a CSCR, this total power is given by

$$P_{r,\text{CSCR}} = \frac{4\rho P_{tx} h A}{\pi^2 M D_{\text{spot}}^2} \int\limits_{(x,y)\in R_{\text{spot}}} \frac{\cos(\psi)\, ds}{\left[(x - x_r)^2 + (y - y_r)^2 + h^2\right]^{3/2}}, \tag{10}$$

where the value of the angle ψ is obtained from Eq. (9). Similarly, the total power for a SCIR is given by

$$P_{r,\text{SCIR}} = \frac{4\rho P_{tx} h^2 A}{\pi^2 M D_{\text{spot}}^2} \int\limits_{(x,y)\in R_{\text{spot}}} \frac{ds}{\left[(x - x_r)^2 + (y - y_r)^2 + h^2\right]^2}. \tag{11}$$

3.2 Ambient light sources

In an indoor environment, the mayor sources of ambient light include sunlight, incandescent light and fluorescent light sources (Barry, 1994; Boucouvalas, 1996; Gfeller & Bapst, 1979; Otte et al., 1999). These sources radiate a substantial amount of power within the wavelengths of silicon photodetectors, inducing shot noise in receivers (Moreira et al., 1997; Tavares et al., 1995). Shot noise is the main degrading factor in wireless optical communications. Here, the case of sunlight and incandescent illumination is considered.
Sunlight is, certainly, the strongest source of ambient light. Normally, this light illuminates rooms after being reflected on ceiling and walls. For this reason, to model sunlight, it is assumed that ceiling acts as a secondary source of radiation. The effect of walls is neglected

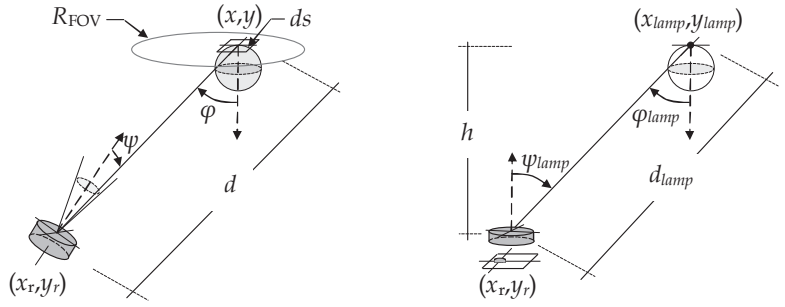


Fig. 4. (a) Link geometry used to compute the sunlight power received by a conventional single-channel receiver. (b) Link geometry employed to compute the artificial light power received by single-channel imaging receiver.

in the analysis because it does not affect single-channel receivers. As in (Carruther & Kahn, 2000), to compute the received ambient light power, the ceiling area within the receivers FOV, R_{FOV}, is divided into differential surface elements ds, acting as first order Lambertian emitters with a spectral radiant emittance given by $S_n(x,y)$. Then, according to the geometry of Fig. 4, the sunlight power detected by a single-channel receiver is obtained as

$$P_{sun} = \int\limits_{R_{\text{FOV}}} S_n(x,y)\Delta\lambda \frac{\cos(\varphi)}{\pi} \cdot \frac{A\cos(\psi)}{d^2} \cdot ds, \tag{12}$$

where, $\Delta\lambda$ is the bandwidth of the receiver optical filter, φ is the angle between the reflected light and the normal of ds given by $\varphi = \cos^{-1}(h/d)$, ψ is the angle between the surface normal of detector and the incident ray, and d is the distance between the receiver and each differential surface element ds. Here, the differences in computing Eq. (12) between both types of receivers are two: the value of the angle ψ, and the size and geometry of the region R_{FOV}. Referring to Fig. 4 (a), the total sunlight power received in a CSCR is obtained from Eq. (12) as

$$P_{sun,\text{CSCR}} = \int\limits_{(x,y)\in R_{\text{FOV}}} \frac{S_n\Delta\lambda(x,y)\,hA\cos(\psi)}{\pi\left[(x-x_r)^2+(y-y_r)^2+h^2\right]^{3/2}} ds, \tag{13}$$

where R_{FOV} is an ellipse whose size depends on the receiver orientation, and the angle ψ is obtained from Eq. (9). Likewise, the total natural light power for the SCIR is given by

$$P_{sun,\text{SCIR}} = \int\limits_{(x,y)\in R_{\text{FOV}}} \frac{S_n\Delta\lambda(x,y)\,h^2A}{\pi\left[(x-x_r)^2+(y-y_r)^2+h^2\right]^{2}} ds, \tag{14}$$

where now R_{FOV} is a circle whose area depends on the receiver photodetector diameter and the lens magnification M_T. For this receiver, the angle ψ equals φ.

As previously mentioned, artificial light from incandescent lamps is much weaker than sunlight and greatly depends on the relative position of the receivers with respect to these lamps. Different measurements from ceiling lamps carried out by (Tavares et al., 1995) show that an excellent model for their radiant intensities is a generalized Lambertian pattern of

order m. Thus, in our analysis incandescent lamps are modeled as point sources with a radiant intensity given by

$$I_{\text{lamp}}\left(\varphi_{lamp}\right) = P_l \frac{m+1}{2\pi} \cos^m\left(\varphi_{lamp}\right), \tag{15}$$

where P_l represents the lamp power within the receiver optical bandwidth $\Delta\lambda$, and φ_{lamp} is the emission angle respect to the normal, as shown in Fig. 4 (b).
Referring to Fig. 4 (b) and employing the Lambertian model of Eq. (15), the received ambient light power from an incandescent lamp within the receiver FOV (of a CSCR or a SCIR) is given by

$$P_{lamp} = \frac{P_l\,(m+1)\,h^m A \cos\left(\psi_{lamp}\right)}{2\pi\left[(x_{lamp}-x_r)^2 + (y_{lamp}-y_r)^2 + h^2\right]^{\frac{m+2}{2}}} \tag{16}$$

where h is the vertical separation between the receiver and the lamp, ψ_{lamp} is the reception angle respect to the normal, and (x_{lamp}, y_{lamp}) are the ceiling lamp coordinates. Again, the only difference in computing Eq. (16) between the two types of receivers is the value of the angle ψ_{lamp}. For the CSCR, ψ_{lamp} is calculated using Eq. (9) replacing (x, y) by (x_{lamp}, y_{lamp}) and changing d by the distance between the lamp and the receiver d_{lamp}. For the SCIR, ψ_{lamp} is given by $\cos^{-1}(h/d_{lamp})$, as deduced from Fig. 4 (b).

4. Performance evaluation

In order to analyze the effect of SCR structures on system performance, two link configurations formed by a MBT and a SCR have been simulated. Both the MBT and the SCR are designed according to the criteria described in Sections 2 and 3, respectively. In these simulations, the main system performance indicators: the SNR and the channel BW, have been computed in a typical indoor environment under natural and artificial light sources. To compute the SNR, it is has been assumed that shot noise induced by ambient light is the dominant noise in the receivers. Under this assumption, the average electrical SNR in our direct-detection receivers is given by

$$\text{SNR} = \frac{(rP_r)^2}{2qrB_nP_{amb}}, \tag{17}$$

where r is the photodetector resposivity, P_r is the signal power received in the photodetector, which is obtained using Eq. (10) or Eq. (11) depending on the type of receiver considered (the CSCR or the SCIR), q is the electron charge, B_n is the receiver equivalent noise bandwidth, and P_{amb} is the received ambient light power. The value of P_{amb} is obtained adding natural and artificial light powers

$$P_{amb} = P_{sun} + P_{lamp}. \tag{18}$$

Here, P_{sun} is the ambient light power received from the sunlight, which is computed using Eq. (13) and Eq. (14) for CSCR and SCIR, respectively, and P_{lamp} is the ambient light power received from a ceiling lamp given by Eq. (16). Note that, due to the narrow FOV of SCR, these receivers only collect light from a single ceiling lamp.
On the other hand, to compute the channel BW, the impulse responses $h(t)$ of all the links are firstly obtained. To this end, it has been used a simulation tool similar to the one developed by (Barry et al., 1993), but considering only the first bounce. This simplification is possible

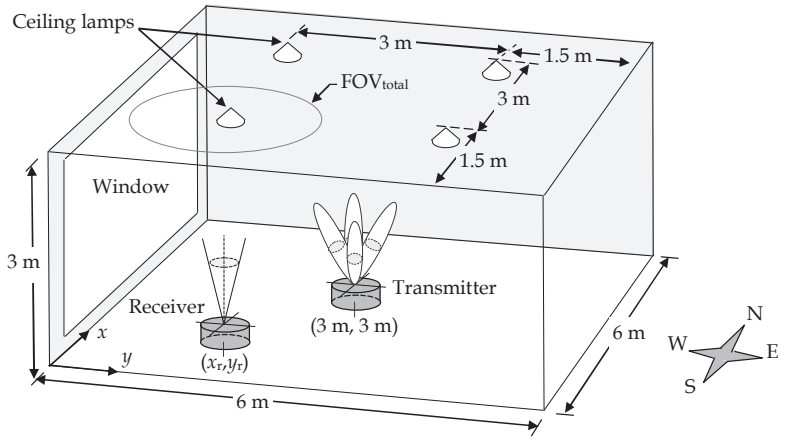


Fig. 5. Room model considered in our numerical analysis. Ambient light includes sunlight (through the window) and four incandescent lamps.

because when a MBT and a narrow FOV receiver are employed, the first bounce provides enough information on the achievable channel bandwidth, as described in (Akavan et al., 2002). Then, from $h(t)$ the frequency response is computed using $H(f) = \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}dt$, and, finally, $H(f)$ is used to determine the 3-dB channel BW (B).

4.1 Simulation set-up

Fig. 5 shows the room considered in the simulations. As can be observed, it is an empty room with dimensions 6 m × 6 m × 3 m, in which a big window occupies the west wall. The effect of the sunlight which illuminates the room through the window, is taken into account considering that the ceiling spectral emittance $S_n(x, y)$ varies linearly from 0.03 W/m^2/nm at the west edge ($y = 0$), to 0.01 W/m^2/nm at the east edge ($y = 6$). The values of S_n are taken from measurements made by Carruthers, published in (Carruther & Kahn, 2000). Four ceiling lamps have been also included in the room placed at the coordinates (x_{lamp}, y_{lamp})=(1.5, 1.5), (1.5, 4.5), (4.5, 1.5) and (4.5, 4.5), respectively, as depicted in Fig. 5. According to the explanation given in Section 3, these lamps have been modeled as Lambertian emitters of order m =2, with a power spectral density $p_{lamp} = 0.037$ W/nm within the receiver filter passband. The values of m and p_{lamp} have been taken from (Carruther & Kahn, 2000) and (Djahani & Kahn, 2000). Note that, the lamp power $P_l = p_{lamp}\Delta\lambda$. A ceiling reflectivity $\rho = 0.7$ is assumed.

In the simulations, two link configurations have been analyzed. The first configuration (CSCR links) employs a MBT and a CSCR. The second (SCIR links) uses a different MBT combined with a SCIR. In CSCR links, the receiver is composed of a lens with an aperture area $A = 3.14$ cm^2 and a photodetector with $D_{\text{det}} = 2$ mm. The resulting CSCR achieves a FOV=1.5° within a FOV$_{\text{total}} = 45°$. Together with this receiver, the considered MBT is formed by $M = 16$ circular spots with $D_{\text{spot}} = 10$ cm and a pattern grid spacing $\Delta D = 1.95$ m. On the other hand, in SCIR links, the receiver uses an imaging lens system with an aperture area $A = 7$ cm^2 and a FOV$_{\text{total}} = 30°$. In this SCIR, the circular photodetector has a diameter D_{det}= 1.52 mm. Combined with this receiver, now, the considered MBT is composed of $M = 36$ spots with a $\Delta D = 1.1$ m. The values of ΔD and M have been computed using

Eq. (4) and Eq. (5), respectively. We have also assumed that both receivers have an equivalent noise BW $B_n = 500$ MHz, an optical bandpass filter having a half-power BW $\Delta\lambda = 50$ nm and photodetectors with $r = 0.6$ A/W of resposivity. It should be pointed out here that, although particular values for r, $\Delta\lambda$, B_n and ρ are assumed in the simulations, the conclusions obtained based on the comparison of the two receivers are valid, regardless of the values employed.
In all simulated links, it is assumed that MBT are kept fixed in the center of the room at a height of 1 m, while receivers are placed in different locations. Therefore, MBT and receivers positions are given by $(3, 3)$ and (x_r, y_r), respectively, as show in Fig. 5. The distance between receivers and ceiling is $h = 2$ m.

4.2 Simulation results

Fig. 6 presents SNR and channel BW spatial distributions for the two link configurations, as function of the the receiver position within the room. To obtain these results, receiver coordinates (x_r, y_r) has been changed along the room surface with a grid spacing of 10 cm. These coordinates are given with respect to the southwest corner, as shown in Fig. 5. A transmitter power $P_{tx} = 200$ mW is assumed in all cases. In this figure, it is observed that the two SNR spatial distributions present M local maximum corresponding to the coordinates of the ceiling spots: 16 maximum for CSCR links (Fig. 6 (a)) and 36 for SCIR links (Fig. 6 (c)). It is also observed that both SNR distributions show a clear west-east asymmetry caused by the sunlight. Therefore, CSCR and SCIR placed close to the window (in the west edge) obtain approximately 4.6 dB and 4 dB less SNR than receivers located in the opposite edge of the room (in the east edge). However, apart from this variation, both SNR figures exhibit very smooth spatial changes. In fact, if the inclination caused by sunlight is neglected, the maximum SNR variation for the CSCR links is of only 5 dB, while the variation for SCIR links is of hardly 0.6 dB. Thus, SCIR links produce more uniform channels in terms of SNR than CSCR links. This intuitive result is due to the higher number of spots employed by SCIR links, as a consequence of the receiver total FOV constraints. Note that, absolute values of SNR are excellent in both figures, higher than 24.6 dB in all receiver positions. On the other hand, and as expected, artificial light from the lamps does not affect the SNR. This is because both receivers are able to avoid these light sources, rejecting all shot noise they induce.
On the contrary, the spatial distribution of the channel BW presents very abrupt variations, as depicted in Fig. 6 (b) and (d). These abrupt variations, of tens of GHz, are due to the changes in the receiver orientation when it moves along the room. However, despite these variations, the obtained BW is always higher than 1 GHz and 1.2 GHz for CSCR links and SCIR links, respectively. Note that, now, the positions with higher BW do not coincide with the coordinates of the transmitter spots. Again, Fig. 6 (b) and (b) show a clear west-east asymmetry of the channel BW due to the higher level of sunlight in the west part of the room caused by the window. However, here the channel BW behavior is different from the SNR behavior and, now, receivers located in the west half of the room obtain better results than receivers located in the east half.
Since SNR and channel BW depend on the receiver position, to compare the performance of the two link configurations, a statistical approach has been applied. Thus, a large number (10^3) of receiver positions have been randomly chosen and the corresponding communications links have been simulated for each in order to obtain statistical results for the system performance. In these new simulations, the power requirements to achieve a bit error rate (BER) not exceeding 10^{-9} and the channel BW have been computed. The BER is given by

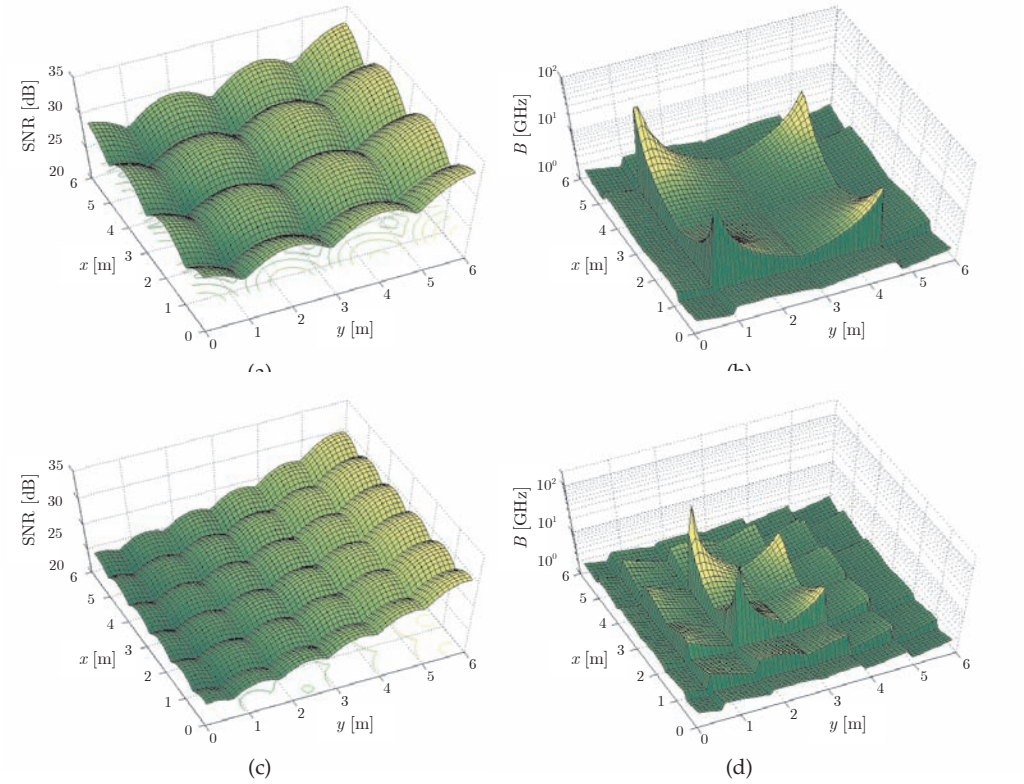


Fig. 6. Signal-to-noise ratio and channel bandwidth spatial distributions for CSCR links and SCIR links. (a) SNR for CSCR links. (b) Channel BW for CSCR links. (c) SNR for SCIR links. (b) Channel BW for SCIR links. CSCR links employ a MBT with 16 beams and a receiver with a $\text{FOV}_{\text{total}} = 45°$. SCIR links uses a MBT with 36 beams and a receiver with a $\text{FOV}_{\text{total}} = 30°$. A transmitter power $P_{tx} = 200$ mW is assumed in all cases.

BER$= Q(\text{SNR})$, so a SNR of 15.6 dB is required to achieve 10^{-9} BER, neglecting the effects of multipath distortion and assuming on-off keying (OOK) modulation (Kahn & Barry, 1997). Fig. 7 (a) presents the average transmitter power to achieve 10^{-9} BER for CSCR links and SCIR links, as a function of the receiver location percentage. In this figure, it is clearly observed that both types of links require very similar transmitter powers. In fact, for 95% location percentage, CSCR links require 17.5 dBm while SCIR links need 17.8 dBm. Thus, CSCR links are lightly superior to SCIR links in terms of power efficiency. This is due to CSCR links divide the same power between fewer beams and, hence, they waste less power illuminating spots different from the selected one. However, when receivers performance is compared in terms of channel BW, the result is just the opposite and, in this case, SCIR links achieve higher BW than CSCR links, as shown in Fig. 7 (b). In particular, for 95% location percentage, the BW of CSCR links is approximately 1.1 GHz, while in SCIR links the BW is better than 1.3 GHz.
The effect of blockage and optical aberrations in system performance has also been analyzed in these links. To emulate signal blockage, the best spot (i.e., the spot which provides the highest

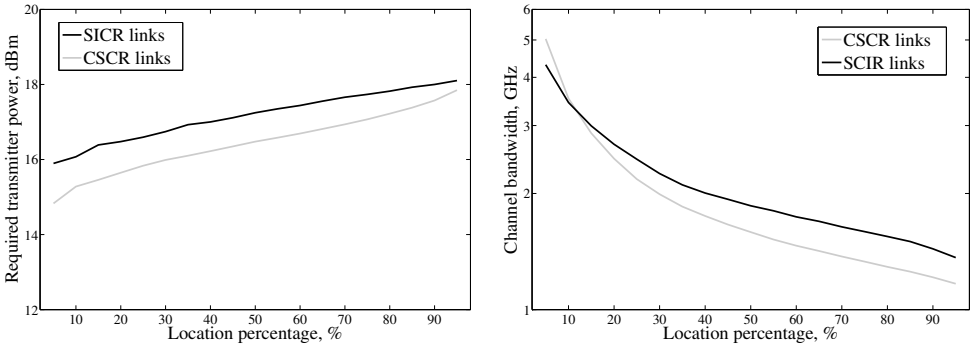


Fig. 7. (a) Required transmitter power to achieve a BER$< 10^{-9}$ and (b) channel bandwidth as a function of the receiver location percentage.

SNR) within the receivers $\text{FOV}_{\text{total}}$ has been discarded in all links. Thus, since receivers are forced to point at an alternative spot, the transmitter power requirements increase. However, due to the proximity between spots, for 95% location percentage, such an increase is only by 1.6 dB for CSCR links and 0.5 dB for SCIR links. Finally, to include optical aberrations in the simulations, an expansion of the projected spots has been assumed so that for a 10-cm diffusing spot 2 m away from the receiver, the diameter of the image spots varies between 1.52 mm an 2.28 mm, for arrival angles $\psi = 0^\circ$ and $\psi = 30^\circ$, respectively. Thus, since the area of image spots is greater than the photodetector area, part of the received power is lost. In particular, our simulations show that, for 95% location percentage, optical aberrations cause an increase of 0.7 dB in transmitter power requirements.

5. Conclusion

In this chapter, the impact of single-channel receivers on channel properties has been investigated. By numerical simulations, the main performance indicators of two link configurations consisting of a MBT and the proposed CSCR and SCIR have been calculated in a typical indoor environment. The obtained results show the advantages and weaknesses of each receiver structure and have proved the great potential of both SCR when operating in a multispot diffusing configuration. From the analysis of these results, it can be concluded that: (a) both receiver structures obtain excellent results in terms of SNR and channel BW, (b) when reducing transmitter spot size, both SNR and BW are significantly improved at the expense of increasing the search time of the receivers algorithm, (c) as receivers move towards the window sunlight degrades SNR, but channel BW is improved, (d) artificial light does not affect system performance, (e) the joint selection of the receiver total FOV and the number of spots makes CSCR links better than SCIR links in terms of power efficiency and worse in terms of channel BW, and (f) blockage and optical aberrations have a low impact on system performance.

6. Acknowledgment

This work was supported by the Spanish Ministerio de Ciencia e Innovación, Project TEC2008-06598.

7. References

Akavan, K., Kavehrad, M. & Jivkova, S. T. (2002). High-speed power-efficient indoor wireless infrared communication using code combining-Part I, *IEEE Transactions on Communications* 50(7): 1098–1109.

Al-Ghamdi, A. & Elmirghani, J. (2003a). Optimization of a triangular PFDR antenna in a fully diffuse OW system influenced by background noise and multipath propagation, *IEEE Transactions on Communications* 12: 2103–2113.

Al-Ghamdi, A. & Elmirghani, J. (2003b). Performance evaluation of a pyramidal fly-eye diversity antena in an indoor optical wireless multipath propagation environment under very directive noise sources, *IEE Proc-Optoelectron.* 150: 482–489.

Al-Ghamdi, A. & Elmirghani, J. (2004a). Line strip spot-diffusing transmitter configuration for optical wireless systems influenced by background noise and multipath dispersion, *IEEE Transactions on Communications* 52(01): 37–45.

Al-Ghamdi, A. & Elmirghani, J. (2004b). Spot diffusing technique and angle diversity performance for high speed indoor diffuse infra-red wireless transmission, *IEE Proc.-Optoelectron.* 151: 46–52.

Alves, L., Aguiar, R., de Vasconcelos, E. & Cura, J. (2000). A sectored receiver for infrared wireless networks, *IEEE International Symposium on Circuits and Systems*, Geneva, pp. 429–432.

Barry, J. (1994). *Wireless Infrared Communications*, Kluwer Academic Publishers.

Barry, J., Kahn, J., Krause, W. J., Lee, E. A. & Messerschmitt, D. G. (1993). Simulation of multipath impulse response for wireless optical channels, *IEEE Journal on Selected Areas in Communications* 11(3): 367–379.

Bellon, J., Sibley, M., Wisely, D. & Greaves, S. (1999). Hub architecture for infra-red wireless networks in office enviroments, *IEE Proc.-Optoelectron.* 146(2): 78–82.

Boucouvalas, A. (1996). Indoor ambient light noise and its effect on wireless optical links, *IEE Proc.-Optoelectron.* 143: 334–338.

Carruther, J. & Kahn, J. (2000). Angle diversity for nondirected wireless infrared communication, *IEEE Transactions on Communications.* 48(6): 960–969.

Castillo-Vázquez, M., García-Zambrana, A. & Puerta-Notario, A. (2004). Angle diversity with rate-adaptive transmission using repetition coding and variable silence periods for wireless infrared communications, *IEEE 59th Vehicular Technology Conference, VTC '04-Spring*, Vol. 1, pp. 415–419.

Castillo-Vázquez, M. & Puerta-Notario, A. (2005). Single-channel imaging receiver for optical wireless communications, *IEEE Communications Letters* 9(10): 897–899.

Djahani, P. & Kahn, J. (2000). Analysis of infrared wireless links employing multibeam transmitters and imaging diversity receivers, *IEEE Transactions on Communications* 48(12): 2077–2088.

Eardley, P. L., Wisely, D. R., Wood, D. & McKee, P. (1996). Holograms for optical wireless LANs, *IEE Proc.-Optoelectron.* 143(6): 365–369.

Gfeller, F. & Bapst, U. (1979). Wireless in-house data communications via diffuse infrared radiation, *Proceedings of the IEEE* 67: 1474–1486.

Jivkova, S., Hristov, B. & Kavehrad, M. (2004). Power-efficient multispot-diffuse multiple-input-multiple-output approach to broad-band optical wireless communications, *IEEE Transactions on Vehicular Technology* 53(3): 882–889.

Jivkova, S. & Kavehrad, M. (2001). Receiver designs and channel characterization for multi-spot high-bit-rate wireless infrared communications, *IEEE Transactions on Communications* 49(12): 2145–2153.

Jivkova, S. & Kavehrad, M. (2003). Shadowing and blockage in indoor optical wireless communications, *IEEE Global Telecommunications Conference, GLOBECOM '03.*, San Francisco, USA, pp. 3269–3273.

Jivkova, S. & Kavehrad, M. (2005). Transceiver design concept for cellular and multispot diffusing regimes of transmission, *EURASIP Journal on Applied Signal Processing* 1: 30–38.

Jivkova, S. T. & Kavehrad, M. (2000). Multispot diffusing configuration for wireless infrared access, *IEEE Transactions on Communications* 48(6): 970–978.

Jungnickel, V., Forck, A., Haustein, T., Kruger, U., Pohl, V. & von Helmolt, C. (2003). Electronic tracking for wireless infrared communications, *IEEE Transactions on Wireless Communications* 2(5): 989–999.

Kahn, J. & Barry, J. (1997). Wireless infrared communications, *Proceedings of the IEEE* 85(2): 265–298.

Kahn, J. M., Barry, J. R., Audeh, M. D., Carruthers, J. B., Krause, W. J. & Marsh, G. W. (1994). Non-directed infrared links for high-capacity wireless lans, *IEEE Personal Communications* 1(2).

Kahn, J. M., You, R., Djahani, P., A.G.Weisbin, Teik, B. K. & Tang, A. (1998). Imaging diversity receivers for high-speed infrared wireless communications, *IEEE Communications Magazine* 36: 88–94.

Kavehrad, M. & Jivkova, S. (2003). Indoor broadband optical wireless communications: optical subsystems design and their impact on channel characteristics, *Optical Wireless Communications* pp. 30–35.

Moreira, A., Valadas, R. & Duarte, A. O. (1997). Optical interference produced by artificial light, *Wireless Networks* 3(2): 131–140.

Otte, R., de Jong, L. & van Roermund, A. (1999). *Low-Power Wireless Infrared Communications*, Kluwer Academic Publishers.

Pakravan, M. R., Simova, E. & Kavehrad, M. (1996). Holographic diffusers for indoor infrared communication systems, *Proc. 'Communications: The Key to Global Prosperity Global Telecommunications Conf. GLOBECOM '96*, Vol. 3, pp. 1608–1612.

Tang, A., Kahn, J. & Ho, K.-P. (1996). Wireless infrared communication links using multi-beam transmitters and imaging receivers, *IEEE International Conference on Communications, ICC 96, Conference Record, Converging Technologies for Tomorrow's Applications.*, Vol. 1, pp. 180–186vol.1.

Tavares, A., Valadas, R. & de Oliveira Duarte, A. (1995). Performance of an optical sectored receiver for indoor wireless communications systems in presence of artificial and natural noise sources, *SPIE's Photonics East '95 International Symposium*, Philadelphia.

Yun, G. & Kavehrad, M. (1992). Spot-diffusing and fly-eye receivers for indoor infrared wireless communications, *IEEE Conf. on Sel. Topics in Wireless Communications*, Vancouver, Canada, pp. 286–292.

22

Estimation of Rotational Axis and Attitude Variation of Satellite by Integrated Image Processing

Hirohisa Kojima
Tokyo Metropolitan University
Japan

1. Introduction

As a result of the increased number of missions in space, the number of satellites that have completed their missions or have broken down has increased, leaving a great deal of space debris in orbit. Most space debris is found in GEO or low-altitude polar orbits and more than 9,600 pieces of debris having a diameter of over 10 cm are currently in orbit. The number of pieces of debris may increase further due to break up, which increases the chance of debris colliding with other spacecraft. To solve this problem, the development of space robots to capture and eliminate space debris from orbit has been explored extensively, and areas such as attitude estimation, formation flying or rendezvous(Kojima , 2005), manipulator control(Inaba & Oda , 2000), de-orbiting of space debris using electro-dynamic tether systems(Forward et al. , 2000; Ishige et al. , 2004) have been investigated. If the target satellite is incorporative, that is, for an example, if radar communication between the target satellite and the debris eliminator satellite is not possible, then image processing will be required in order to monitor the attitude of the satellite and to capture it by means of a robot manipulator. Image processing algorithms with lower computational costs are desired, because the computer resources installed on satellites are usually fewer than those on the ground. A great number of image processing algorithms have been developed for various purposes, such as edge extraction(Harris & Stephens , 1988; Kitchen & Rosenfeld , 1982) and silhouette extraction(Tomasi & Kaneda , 1991). In the EST-VII mission(Inaba & Oda , 2000), the target markers were installed on the daughter satellite so that the mother satellite can easily monitor the attitude of the daughter satellite.

However, normally it is not easy to recognize the attitude of satellites in orbit, because commercial satellites that are not equipped with target markers are usually covered by multi-layer insulator (MLI) with numerous wrinkles that randomly reflect the Sun's light, and such random reflection makes silhouette extraction more difficult. Furthermore, satellites often overlap the Earth, and the direction of the Sun's light relative to the satellite varies with time.

To estimate the attitude of a satellite, the iterative closest point (ICP) algorithm(Besl & McKay , 1992) has been studied by JAXA(Terui et al. , 2002), but this algorithm needs a computational cost. In order to avoid the computational cost of ICP, the grid closest point (GCP) algorithm (Yamany et al. , 1998) was developed using a hash-table technique. The GCP algorithm has

been studied by JAXA and MD-Robotics (Cropp & Palmer , 2002) in Canada. Both ICP and GCP algorithms, however, require a stereovision system to obtain the three-dimensional position of the feature points on the satellite and require a priori reference data in comparing the data set of the feature points on the satellite. Therefore, if a priori reference data is not recorded on the ground in advance, or a stereo camera system is not installed on the debris monitoring satellite, then neither ICP nor GCP can be employed to estimate the attitude of a satellite from the images.

To overcome this problem, the development of an image processing method that can treat wrinkles on the surface of satellites and handle images of satellites overlapping the Earth albedo without using a stereo-vision system is desired. In this paper, such an image-processing scheme is proposed. The proposed scheme consists of six steps: (1) searching the position of a target satellite in the image by color information, (2) extraction of feature points on the satellite using a Harris corner detector (Harris & Stephens , 1988) (3) optical flow estimation by template matching and random sample consensus (RANSAC)(Lacey et al. , 2000), (4) deleting the incorrect optical flow using the epipolar condition(Zhang et al. , 1995), (5) initial guess of the rotational axis and attitude variation from the extracted optical flow by a heuristic approach, and (6)an iterative algorithm to obtain the precise rotational axis and the attitude variation from the initial guess. The image processing method described herein is commonly used for attitude estimation. The contribution of the present paper is the modification of this method by using RGB color information to effectively extract the feature points on a satellite from images, by deleting noises in the images with respect to the consistency of the epipolar condition, and by the heuristic technique to obtain the initial guess of the rotational axis and attitude variation, and shorten the computational cost for the estimation. A space light simulator, which consists of a dark room, is constructed to emulate the light environment in space and is used to verify the validity of the proposed method. The results of experiments using the images of a satellite model taken in the simulation room reveal that the proposed method can estimate the rotational axis and attitude variation of the satellite model under a good lighting condition within approximately 15 deg relative angle, and with an error range of 1.7 deg, which has not yet been achieved under the condition that neither a stereo system nor the priori geometry information of the satellite can be used.

2. In-space lighting simulation room

In order to simulate in-space light conditions, a dark room called the “in-space lighting simulator.” was constructed. The dimensions of the in-space light simulator are as follows: height; 2.5 m, width; 2.5 m, and length; 3 m. Figure 1 shows the schematic representation of this simulator which consists of a halogen lamp for imitating the light from the Sun, a reflector for imitating the Earth albedo, and a spacecraft model representing a malfunctioning satellite, specifically, the MDS-1, launched by NASDA. Although the spectral distribution of the halogen lamp is not the same as that of the Sun, the halogen lamp is used in the present study for the sake of simplicity. The albedo of the reflector is set to approximately 0.3 so as to imitate that of the Earth. The reflector is made of high polymeric paper, onto which an image of the Earth, including clouds, is printed. Although MDS-1 was covered with a black multi-layer insulator (MLI), the satellite model in the present study is covered with a gold MLI, because most satellites currently on orbit are covered with gold MLI. The rotational motion of the satellite model is simulated by a three-gimbal system that represents the Euler angles in the form z-y-z, as shown in Fig 2. A commercial digital CCD camera is used to

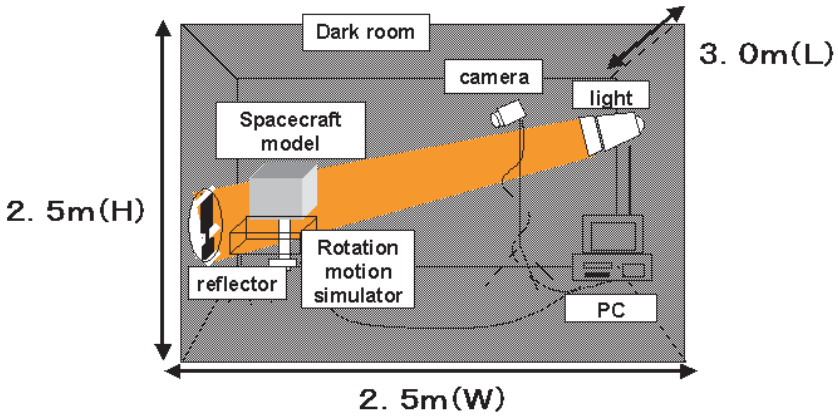


Fig. 1. Schematic representation of the in-space lighting simulator.

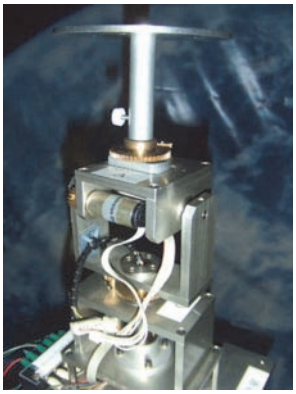


Fig. 2. Rotational motion simulator using z-y-z Euler angles.

take images of the satellite model, based on the assumption that such a camera will likely be installed on a debris monitoring satellite in order to reduce the development costs of the satellite.

The sizes of the reflector and the satellite model are determined so as to imitate the relative image, considering the geometric relationship, as shown in Fig. 3, where R is the radius of the Earth, H is the altitude of the satellite with a camera for monitoring debris, and θ is the angle of the Earth as seen from the orbiting satellite with altitude of H. The angle is obtained as a function of R and H as

$$\theta = \arcsin\left(\frac{R}{R+H}\right) \tag{1}$$

Using the distance between the reflector and the digital camera, Z, the diameter of the reflector d is obtained as follows:

$$d = 2Z\tan\theta \tag{2}$$

Table 1 shows the parameters of the satellite and distance between the satellite model and the

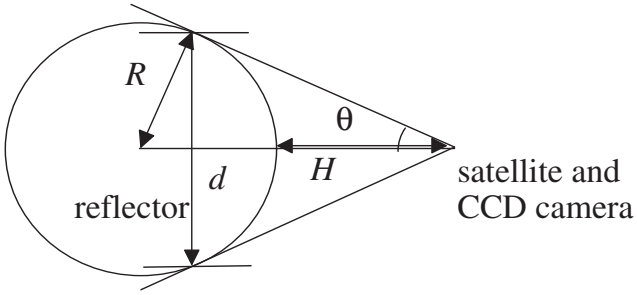


Fig. 3. Geometric relationships between the satellite and the size of the reflector.

Size of Tsubasa	1.2 x 1.2 x 1.5 m
Size of the satellite model	0.21 x 0.21 x 0.21 m
Distance between the satellite model and the reflector	1.5 m
Diameter of the reflector	3.0 m
Emulated altitude	2,500 km

Table 1. Parameters of the In-space lighting simulation.

reflector. The imitated altitude determined from these parameters is also indicated in Table 1. The altitude of the spacecraft in this study is approximately 2,500 km, which is much higher or lower than that of existing debris orbits. The altitude of the satellite model can be imitated by adjusting either the distance between the reflector and the camera, or the distance between the camera and the spacecraft model.

3. Integrated image processing method to estimate attitude variation

The image processing method considered in this study for detecting the rotational axis and attitude variation consists of six steps: (step1) target searching based on color information, (step2) feature points extraction by Harris corner detector(Harris & Stephens , 1988), (step3) optical flow estimation using template matching and random sample consensus (RANSAC)(Lacey et al. , 2000), (step4) deleting incorrect paired points using the epipolar condition(Zhang et al. , 1995), (step5) initial guess of the rotational axis and attitude variation from the extracted optical flow by a heuristic approach, and (step6) an iterative algorithm for obtaining the precise rotational axis and the attitude variation from the initial guess. In this section, each step will be briefly explained.

3.1 Target searching by color information

Before explaining the method used to search the center of the target in the image, the color characteristics of images of a satellite are addressed.

Color images can be typically shown in RGB (red, green, and blue) form. First, single spectra representing these three colors are assumed, respectively, to be:

$$\bar{r}_\lambda = 700.0[nm], \quad \bar{g}_\lambda = 546.1[nm], \quad \bar{b}_\lambda = 435.8[nm] \tag{3}$$

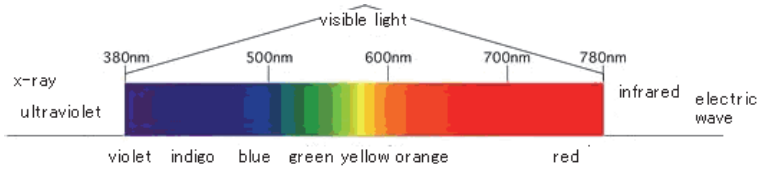


Fig. 4. Color and wavelength within the rage of the visible light spectrum.

Then, comparing each single spectrum of 1 W with the wave in the range of from 380 nm to 780 nm, which corresponds to the visible wavelength(Fig. 4), yields the distribution of each spectrum. When the color spectrum is given by $C(\lambda)$ W, the RGB value is calculated as

$$R = \int_{380}^{780} C(\lambda)\,\bar{r}_\lambda d\lambda \tag{4}$$

$$G = \int_{380}^{780} C(\lambda)\,\bar{g}_\lambda d\lambda \tag{5}$$

$$B = \int_{380}^{780} C(\lambda)\,\bar{b}_\lambda d\lambda \tag{6}$$

If colors are indicated by eight bits per layer (blue, green, and red), then the number of indicated colors is 16,777,216.
Satellites usually reflect either gold, orange, or yellow, because they are covered by gold multi-layer insulator (MLI), and have solar array panels. On the other hand, the surface of the Earth is blue and white, unless viewed from a high angle. To extract feature points on the satellite efficiently, let us consider the reflection characteristics of the materials of a satellite. Figure 4 shows the relationship between color and wavelength, and Figure 5 shows the distribution of the spectral response of a CCD camera to light of the visible wavelengths. Figures 5(a), 5(b), and 5(c) show the distributions of spectral response of the panel, space, and albedo, respectively, of the upper image. It is found in Fig.5(a) that the RGB brightness of solar array panels has the relation R $>$ G $>$ B.
Taking the above color characteristics into account, the center of the target in the image can be recognized as the center of the areas in which the color distribution is R $>$ G $>$ B as shown in Fig.6. After determining the center of the target, the feature point extraction can be concentrated around the recognized center.

3.2 Extraction of feature points by Harris corner detector

A number of corner detection methods, including Harris(Harris & Stephens , 1988), Kitchen-Rosenfeld, KLT, and SUSAN, have been proposed. A photo image taken in space may include undesired signals for attitude detection due to the effects of camera halation, for example. However, since the Harris corner detector (Harris & Stephens , 1988) is known to be insensitive to noises, in this study, the Harris corner detector is employed to extract the feature points on the satellite model. In this detector, the corner detective function given R_h by

$$R_h = \frac{\det \mathbf{M}}{(tr\mathbf{M})^2} \tag{7}$$

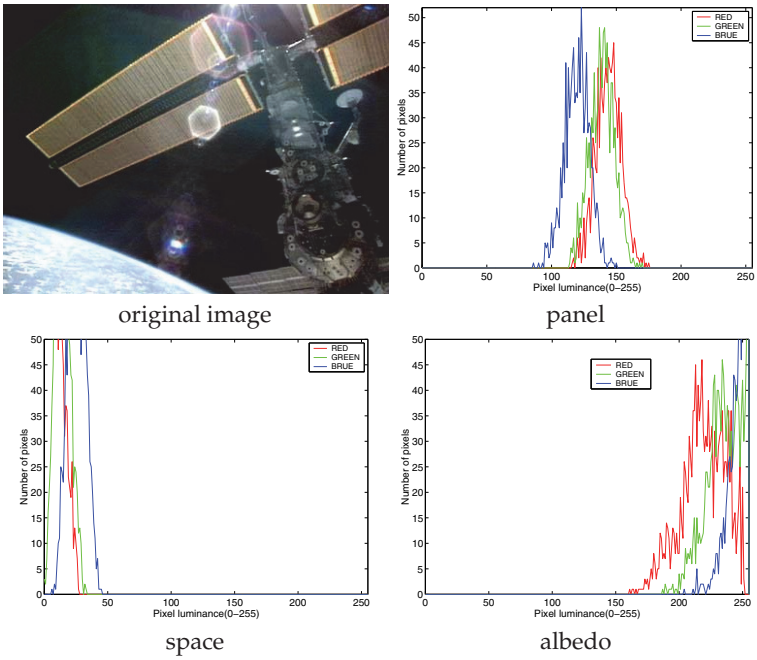


Fig. 5. Spectral response curve.

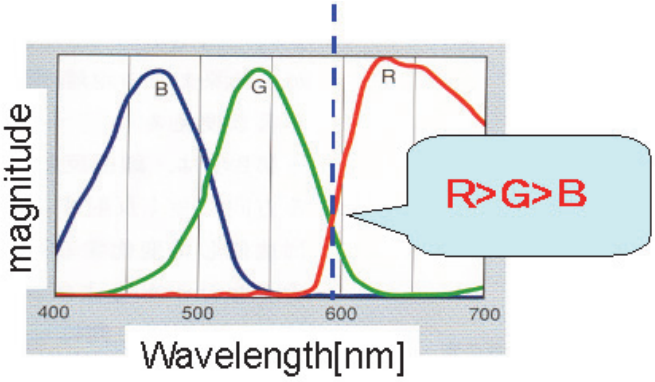


Fig. 6. Schematic representation of spectral response curve of the panel.

is used, where the matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{pmatrix} \left(\frac{\partial I}{\partial j}\right)^2 & \left(\frac{\partial I}{\partial j}\right)\left(\frac{\partial I}{\partial i}\right) \\ \left(\frac{\partial I}{\partial j}\right)\left(\frac{\partial I}{\partial i}\right) & \left(\frac{\partial I}{\partial i}\right)^2 \end{pmatrix} \tag{8}$$

and $I(i,j)$ is the pixel value at an image point (i,j). If the two eigenvalues of $\mathbf{M}$ exceed a specified threshold value, then the corresponding point may be a corner. All of the candidate

corners are checked using the value of R_h, and points with values greater than those of the neighboring points are then extracted as the feature points.
When a color image is directly used in the Harris corner detector, incorrect feature points that are not on the target may be also extracted. In order to eliminate incorrect feature points to the extent possible, the gray scale of the input image was smoothed by an averaged filter, and then candidate feature points were extracted. Feature points were then selected based on RGB color information. As a result of this process, the incorrect were eliminated.

3.3 Paring of the extracted feature points

Two sets of feature points detected by the Harris corner detector are compared, and the points corresponding to the points in the other set are detected by the template matching technique. A template is a pixel data set of size $m \times n$, the center of which is the feature point. The detection process is as follows.
At first, a template of fixed size is generated at each feature point in the first image. The similarity of the template to the feature points in the second image is then calculated. The point with the highest similarity is then accumulated. This process is repeated until all of the candidate corresponding feature points are recorded.
The similarity rate, which is often called the normalized cross correlation function, is given by

$$R_{NCC} = \frac{\sum_{j=0}^{n-1}\sum_{i=0}^{m-1} I(i,j)\,T(i,j)}{\sqrt{\sum_{j=0}^{n-1}\sum_{i=0}^{m-1} I(i,j)^2 \times \sum_{j=0}^{n-1}\sum_{i=0}^{m-1} T(i,j)^2}} \tag{9}$$

where $m \times n$ is the template size, $T(i,j)$ is the pixel value at the point (i,j) in the template image, and $I(i,j)$ is the pixel value at point (i,j) in the second image.
The correlation range is between 0 and 1. If the value is near 1, then the point has high similarity. Template matching has a limitation with respect to the angle between two groups of the feature points and the rotational velocity of the target satellite. In addition, when the images include the same patterns, the algorithm fails to detect the correct corresponding feature points.
To overcome the above-described template-matching problem, the RANSAC algorithm(Lacey et al. , 2000), which is an optimization method, is used in this study. An initial estimate of the optimal solution is usually required for the optimization methods, but the reasonable selection of an initial estimate is difficult. On the other hand, the RANSAC algorithm does not require an initial estimate because an initial estimate of the solution is randomly set in the algorithm. In addition, the objective function in this method does not need to consider constraints, such as the requirement that certain points be near other points or the requirement that certain points be in the positive range.
In this study, RANSAC is first used to select the candidate corresponding feature points. The template matching method is then employed to detect the optimal correspondence of the feature points.

3.4 Error delete by the epipolar condition

At least eight corresponding points are needed to make an essential matrix that enables detection of the relative translational unit vector, rotational axis vector, and rotational angle of the target. The essential matrix may be not precisely obtained when simply using the

eight-point algorithm (Zhang et al. , 1995) to images of a satellite, because the extracted pairs include incorrect pairs. As such, the following two techniques are employed sequentially to delete probably incorrect paired feature points in this study.
First, an optical flow having a length that exceeds 1.5 times the average of the total optical flow length is deleted because such an optical flow may contain incorrectly paired feature points.
Second, an essential matrix is calculated using all of the pairs after deleting the paired points that exceed 1.5 times the average of the optical flow length. If the extracted feature points of images 1 and 2 are represented as $\mathbf{p}_i = [\, u_i \; v_i \; 1 \,]^T$ and $\mathbf{p}'_i = [\, u'_i \; v'_i \; 1 \,]^T$, respectively, then the epipolar condition can be represented as

$$\mathbf{p}_i'^T \mathbf{E} \mathbf{p}_i = 0 \tag{10}$$

where $\mathbf{E}$ is the essential matrix. This condition can be converted into a problem to find vector $\mathbf{h}$

$$\min |\mathbf{A}\mathbf{h}| \tag{11}$$

where

$$\mathbf{A} = \begin{bmatrix} u_1u'_1 & u_1v'_1 & u_1 & v_1u'_1 & v_1v'_1 & v_1 & u'_1 & 1 \\ u_2u'_2 & u_2v'_2 & u_2 & v_2u'_2 & v_2v'_2 & v_2 & u'_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_nu'_n & u_nv'_n & u_n & v_nu'_n & v_nv'_n & v_n & u'_n & 1 \end{bmatrix} \tag{12}$$

$$\mathbf{h} = \begin{bmatrix} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 \end{bmatrix} \tag{13}$$

Using vector $\mathbf{h}$, which is the eigenvector of $\mathbf{A}^T\mathbf{A}$ with the minimum eigenvalues, the essential matrix is given by

$$\mathbf{E} = \sqrt{2} \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \tag{14}$$

Each pair of points is confirmed to satisfy the epipolar condition, which is represented by

$$det(\mathbf{E}\mathbf{p}_i) = 0 \tag{15}$$

From the practical point view, it is impossible to satisfy the above epipolar condition. Thus, instead of the above condition, more practical condition is used.

$$det(\mathbf{E}\mathbf{p}_i) \leq \epsilon \tag{16}$$

If the epipolar condition is not satisfied within a specified tolerance for a pair of points, then those paired points are deleted to purify the data of the pairs. This verification is repeated several times to obtain a more reliable essential matrix for the paired images.

3.5 Initial guess of rotational axis and attitude variation by a heuristic approach

The final process is the estimation of the rotational axis vector and the attitude variation from the obtained essential matrix. As mentioned earlier, it is difficult to obtain the precise attitude variation using the eight-point algorithm (Zhang et al. , 1995), because the algorithm is originally used on the ground to estimate the motion of the measure camera relative to the static scene. In other words, the motion of the target satellite relative to the chaser satellite's camera is not always equivalent to the motion of the chaser satellite's camera relative to the target without motion. In addition, in a case that the image is taken from the

direction perpendicular to the rotational axis, the optical flow might be likely recognized as the translational motion by the eight-point algorithm.

In order to overcome the above problem, in this study, provided that the translational motion can be cancelled in advance by tracking the center of the target as recognized by means of the color information, and provided that the shape of the target satellite is approximately cylindrical, that is, provided that the optical flow lengths are approximately the same in three dimensions, firstly the following heuristic approach is employed to obtain the initial guess for the rotational axis vector and attitude variation. The iteration based on the least square method is then repeated several time using the initial guess resulting from the heuristic approach to obtain reasonable precision. The objective of the heuristic approach is to reduce the computational cost of the iterative scheme. A schematic diagram of the heuristic approach is shown in Fig.7, and the process is as follows:

(1) Record the maximum length of the optical flow ($F_{\max 1}$) and its direction ($\theta_{\max 1}$).

(2) Rotate the image and the optical flow so that the optical flow with the maximum length follows the y-axis of the camera screen coordinate. Record the length between the center of the target and the y-edge of the target (R_{target}) after rotating the image.

(3) Search the optical flow with maximum x length after rotating the image and the optical flow and recording the optical flow ($F_{\max 2}$).

(4) Determine whether the camera views from the top or bottom of the target by taking the optical flow around the y-edge of the target into account. If the optical flow around the y-edge moves upward, then the camera views the target from the top, otherwise the camera views the target from the bottom.

(5-1) Calculate the inclination of the rotational axis of the target toward the camera view direction, which may be estimated as the ratio between the maximum x length of the optical flow ($F_{\max 1}$) and the maximum y length of the optical flow ($F_{\max 2}$) as follows:

$$v_z = \pm F_{\max 2} / F_{\max 1} \tag{17}$$

The sign of the above equation, plus or minus, is determined in accordance with the direction of the optical flow with maximum x length $F_{\max 2}$ with respect to the y-edge of the target. If the direction is upward and outward, or upward and inward, then the sign is minus. Otherwise the sign is plus.

(5-2) Calculate the inclination of the rotational axis of the target with respect to the y-axis of the camera screen, which may be perpendicular to the optical flow with the maximum length θ_{Fmax1}.

$$v_x = \cos\theta_{F\max 1} \cos\left(\sin^{-1}(v_z)\right) \tag{18}$$

$$v_y = -\sin\theta_{F\max 1} \cos\left(\sin^{-1}(v_z)\right) \tag{19}$$

(6) Calculate the attitude variation (θ_{vary}) from the ratio between the maximum y length of the optical flow ($F_{\max 2}$) and the radius of the target (R_{target}), that is,

$$\theta_{\text{vary}} = \frac{F_{\max 1}}{R_{\text{target}}} \tag{20}$$

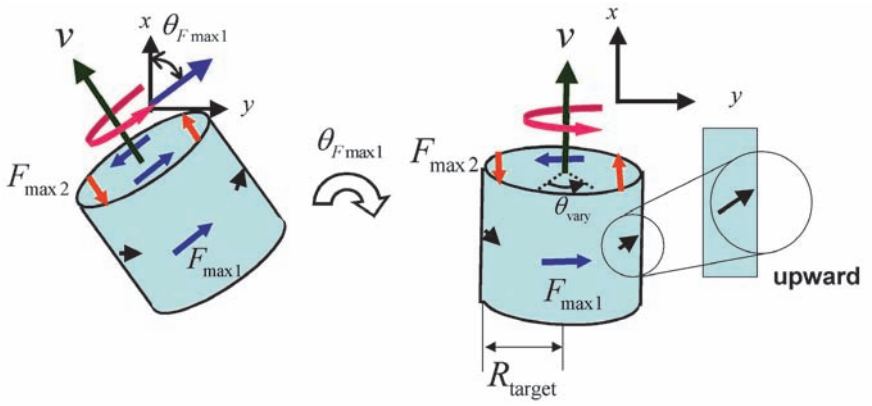


Fig. 7. Schematic diagrams for the estimation of the rotational axis.

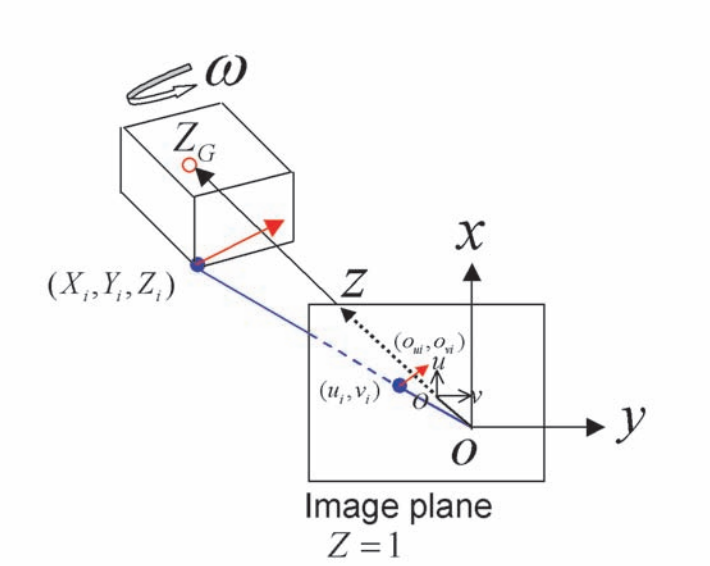


Fig. 8. Relationships between the rotational motion of the target and optical flow.

3.6 Iterative method for estimating rotational axis and attitude variation, from the initial guess

Figure 8 shows a representation of the relationship between the camera coordinate, the position and optical flow of the feature point. Note that in this study, without loss of generality, the camera screen is assumed to be located at $Z = 1$.

Under the assumption that the translational motion of the target relative to the chaser satellite has been already cancelled by tracking the center of the target image and the center is maintained on the origin of the screen coordinate, each optical flow of the paired feature points

$[\, o_{ui}\ o_{vi}\,]^T (= \mathbf{o}_i)$ is represented by

$$\begin{bmatrix} o_{ui} \\ o_{vi} \end{bmatrix} = \left(\begin{bmatrix} 1 & (1+u_i^2)\ 0 \\ -(1+v_i^2) & 1 \quad 0 \end{bmatrix} + \frac{Z_G}{Z_i} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \tag{21}$$

where (u_i, v_i) is the position vector of the i-th feature point in the first image, $\boldsymbol{\omega} (= [\omega_x, \omega_y, \omega_z]^T)$ is the rotational angular velocity of the target represented in the camera coordinate, Z_G is the z position of the center of the rotating target, and Z_i is the z position of the i-th feature point in the camera 3D coordinate.
Introducing

$$\mathbf{B}_\mathrm{i} = \begin{bmatrix} 1 & (1+u_i^2)\ 0 \\ -(1+v_i^2) & 1 \quad 0 \end{bmatrix} \tag{22}$$

$$\eta_i = \frac{Z_G}{Z_i} \tag{23}$$

$$\mathbf{C} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{24}$$

then, Eq.(21) is rewritten as

$$\mathbf{o_i} = (\mathbf{B}_i + \eta_i \mathbf{C})\, \boldsymbol{\omega} \tag{25}$$

Gathering all the optical flows (21), that is, denoting

$$\mathbf{o} = \begin{bmatrix} \mathbf{o}_1 \\ \mathbf{o}_2 \\ \vdots \\ \mathbf{o}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_n \end{bmatrix}, \quad \mathbf{C}' = \begin{bmatrix} \eta_1 \mathbf{C} \\ \eta_2 \mathbf{C} \\ \vdots \\ \eta_n \mathbf{C} \end{bmatrix} \tag{26}$$

yields

$$\mathbf{o} = (\mathbf{B} + \mathbf{C}')\, \boldsymbol{\omega} \tag{27}$$

By solving Eqs.(25) and (27) with respect to the relative distance η_i, and the angular velocity of the target ω, one has

$$\eta_i = \mathbf{C}^{\#} (\mathbf{o}_i - \mathbf{B}_i \omega) \tag{28}$$

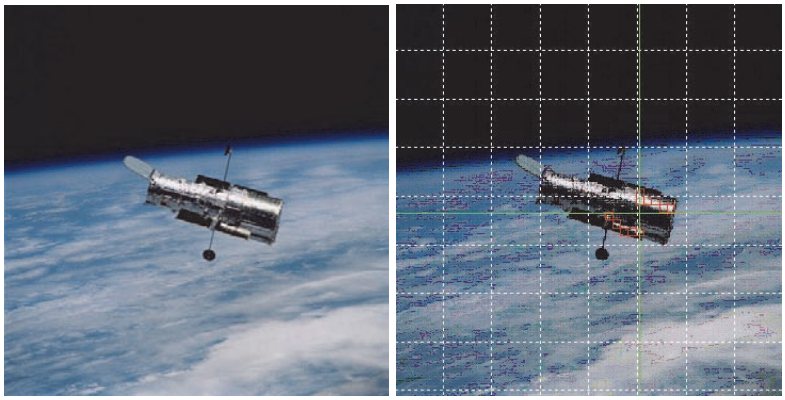
$$\boldsymbol{\omega} = (\mathbf{B} + \mathbf{C}')^{\#}\, \mathbf{o} \tag{29}$$

where superscript # denotes pseudo inverse matrix.
Starting from the initial guess resulting from the heuristic approach described above, Eqs.(28) and (29) are alternately repeated several times to obtain more precise estimation of the angular velocity (attitude variation) of the target.

4. Experimental results

4.1 Results of searching the center of the target

The Hubble Space Telescope (HST) is chosen as an example of a target. Figure 9(a) shows the input photo image, in which the HST is overlapping the Earth albedo, along with the searched center of the target indicated in green shown in Fig. 9(b). It is seen in Fig. 9(b) that the center of a satellite is almost correctly extracted by the color-information-based method.



(a) (b)

Fig. 9. The original image(a) and the detection center of the target.

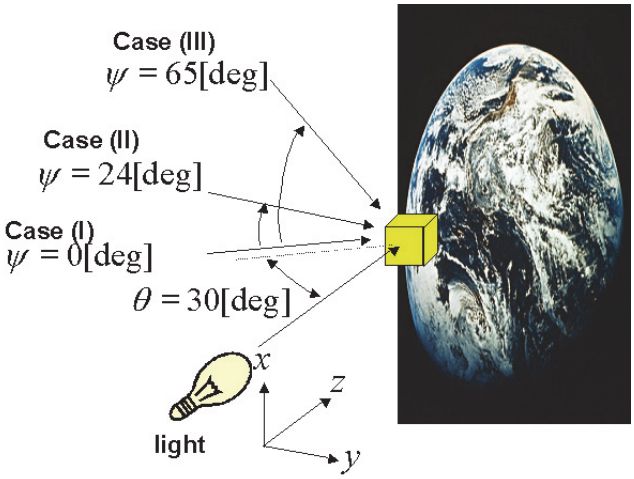
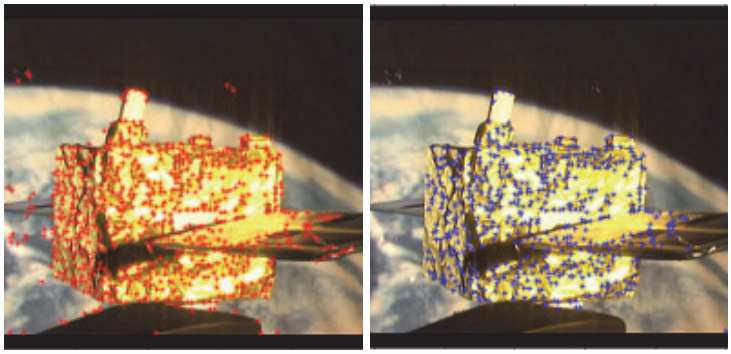


Fig. 10. Experimental conditions for the CCD camera location

4.2 Results of optical flow detection

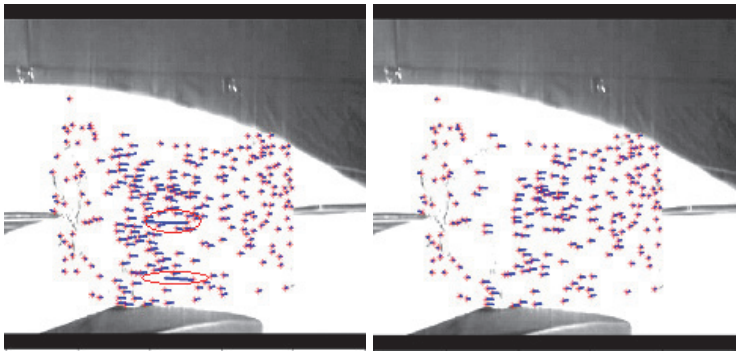
A satellite model covered by golden MLI was produced and set on a rotator to emulate rotational motion. For the sake of simplicity, the emulated rotational motion is a single spin at 3 deg/frame around the z-axis. The lighting direction is set to be perpendicular to the reflector. The camera direction with respect to the reflector is set to be inclined at 30 deg. In order to study the effect of the camera view direction on the estimation of the rotational axis and attitude variation, three cases are studied for the camera view direction with respect to the target (as shown in Fig. 10): case (I) parallel to the horizon, case (II) 24 deg from horizontal above the target, and case (III) 65 deg from horizontal above the target.

Two photographs are taken at an angle interval of 3 deg in the in-space lighting simulation room using the rotational motion produced by the rotator. Figure 11 shows the feature points



(a) (b)

Fig. 11. Feature points extracted by the Harris corner detector without (a) and with employing color information (b).



(a) (b)

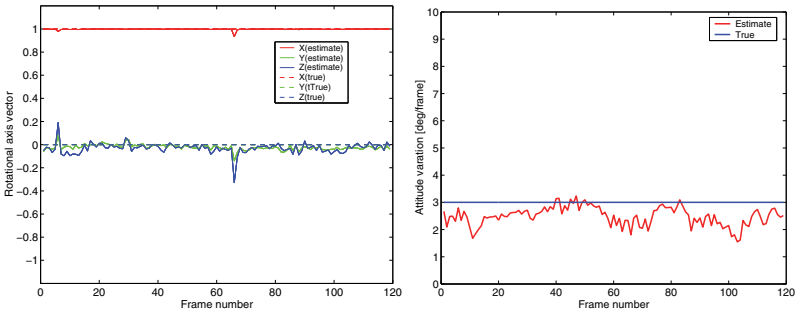
Fig. 12. Optical flow extraction before (a) and after (b) eliminating undesired paired points.

extracted by the Harris corner detector, where the threshold for the corner detective function is set to $R_h > 500$.

Figures 12(a) and 12(b) show results of the optical flow before and after deleting the undesired optical flow, respectively. Some undesired flows, which may be caused by incorrectly paired points, are contained in the optical flows, as shown in Fig. 12(a). On the other hand, as shown in Fig. 12(b), all such undesired paired points are successfully deleted, where $\epsilon = 10^{-2}$ is used as the tolerance for epipolar condition. Consequently, the use of the epipolar condition is effective in obtaining the desired optical flow. However, this matching process requires a long time. Thus, a method for reducing the process time of template matching and RANSAC should be investigated to obtain the optical flow within a reasonable process time.

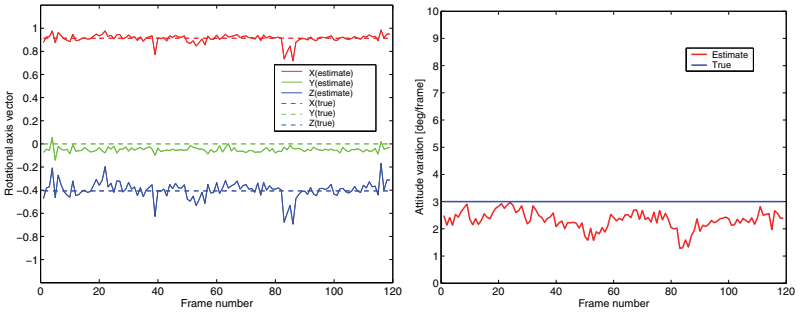
4.3 Results of estimation of rotational axis vector and attitude variation

The results of rotational axis vector and attitude variation estimated by the proposed method according to the frame number are shown in Figs. 13, 14, and 15, respectively, for cases (I), (II), and (III). The correct rotational axis vector and attitude variation are also shown in these figures. The figures show that rough agreement was obtained between the correct direction



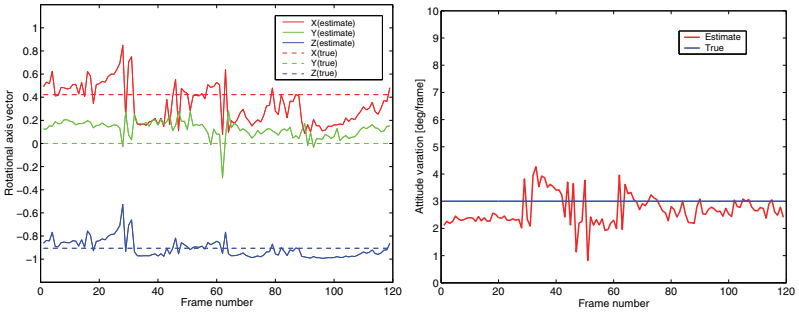
(a) (b)

Fig. 13. Estimated rotational axis vector (a) and the attitude variation (b) for case I.



(a) (b)

Fig. 14. Estimated rotational axis vector (a) and the attitude variation (b) for case II.



(a) (b)

Fig. 15. Estimated rotational axis vector (a) and the attitude variation (b) for case III.

and the estimated direction. The maximum difference between the estimated axis vector and the correct one is approximately 15 deg for cases (I) and (II). The estimated attitude variation per one frame for cases (I) and (II) does not change dramatically, compared to that of case (III), but its mean is less than the correct one by approximately 0.75 deg. This may be because the relative depth ratio, η_i, was estimated to be greater than the correct value due to shade on the target. On the other hand, the estimated attitude variation for case (III) changes dramatically,

compared to that of cases (I) and (II), but the mean of the estimated attitude variation roughly agrees to the correct one. This may be because the reflection of the target was not sufficiently received by the monitoring camera for case (III), in which the mostly observable surface of the target was perpendicular to the light, and this situation resulted in a sensitive estimation to the small number of the extracted optical flow, while the relative depth ratio could be appropriately estimated, compared to cases (I) and (II), because a most observable surface for the camera is a top flat surface of the target, which is almost perpendicular to the camera view direction. The differences of the estimated axis vector and the attitude variation from the correct ones have not yet been achieved under the condition that neither a stereo system nor the priori geometry information of the satellite can be used.

The estimated rotational axis direction oscillates randomly because the MLI wrinkles reflect the light randomly and the heuristic approach described herein estimates the axis based on representative optical flows that depend on random reflections. Furthermore, the estimated attitude variation contains slow oscillations in accordance with the rotational motion of the target. This is because the reflection depends on the angle of the target surface toward the camera view direction and the lighting direction. If taking into account the direction of the reflecting surface to the light direction and the camera view direction, in other words, if the monitoring direction of a camera to the target is appropriately set in order to receive the sufficient reflection, a more precise attitude variation will be obtained. This is a future study.

5. Conclusions

In this study, it was confirmed that feature points and optical flow of a rotating target can be extracted from images taken by only one camera. This was achieved by using the Harris corner detector, template matching, and RANSAC, and by deleting the undesired points in accordance with the RGB color information and the length of the optical flows, even if the optical markers are not equipped on the target. After the optical flow was obtained, the eight-point algorithm was used to obtain a more reliable essential matrix subject to the optical flow. A heuristic approach was introduced to estimate the rotational axis vector and attitude variation by selecting a representative optical flows, provided that the translational motion can be eliminated by tracking the center of the target obtained from the color information. In addition, to improve the rotational axis and attitude variation estimated by the heuristic approach, an iterative square least method was used.

The experimental results showed that the estimated rotational axis vector and attitude variation agree roughly with the correct values under a good lighting condition, but the accuracy is not still yet within an acceptable range for easily approaching or capture in a case that the monitoring direction is approximately perpendicular to the lighting direction, and that they rapidly change due to the random refection of wrinkles of multi-layer insulator. As the next stage, we will attempt to improve the estimation accuracy by taking into account the shade on the target or the lighting direction with respect to the target, and applying some filtering techniques.

6. References

Besl, J. P. & McKay, D. N.(1992). A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.14, No.2, (February 1992), pp.(239-256), 0162-8828

Cropp, A. & Palmer, P.(2002). Pose Estimation and Relative Orbit Determination of a Nearby Target Microsatellite using Passive Imagery, 5th Cranfield Space Dynamics Conference, July 2002

Forward, R. L., Hoyt, R. P. & Uphoff, C. W. (2000). Terminator Tether: A Spacecraft Deorbit Device. *Journal of Spacecraft and Rockets*, Vol.37, No.2, (March-April 2000), pp.(187-196), 00224650

Harris, C. & Stephens, M. (1988). A Combined Corner and Edge Detector, *Proceedings of the Fourth Alvey Vision Conference*, Manchester, August-September 1988, pp.(147-151).

Inaba, N. & Oda, M. (2000). Autonomous satellite capture by a space robot world first on-orbit experiment on a Japanese robot satellite ETS-VII, *Proceedings of IEEE International Conference on Robotics and Automation*, San Francisco, April 2000, pp.(1169-1174)

Ishige, Y., Kawamoto, S. & Kibe, S.(2004). Study on Electrodynamic Tether System for Space Debris Removal. *Acta Astronautica*, Vol.55, No.11, (December 2004), pp.(917-929), 00945765

Kitchen, L. & Rosenfeld, A. (1982). Gray Level Corner Detection. *Pattern Recognition Letters*, Vol.1, No.2, (December 1982), pp.(95-102) 01678655

Kojima, H.(2005). Fly-around Motion Control Based on Exact Linearization with Adaptive Law. *Journal of Guidance, Control, and Dynamics*, Vol.28, No.1, (January-February 2005), pp.(167-169) 07315090

Lacey, A. J., Pinitkarn, N. & Thacker, N. A. (2000). An Evaluation of the Performance of RANSAC Algorithms for Stereo Camera Calibration, 11th British Machine Vision Conference, Bristol, September 2000

Terui, F., Kamimura, H., Nishida, S., Takaya, K. & Kawamura, E.(2002). A Stereo Image Processing for the Attitude Estimation of Large Space Debris Objects, *Proceedings of 23rd International Symposium on Space Technology and Science*, Matsue, May-June 2002, pp.(851-856)

Tomasi, C. & Kaneda, T.(1991). Detection and Tracking of Point Features, Carnegie Mellon University, Tech. Report, Pittsburgh, April 1991

Yamany, S. M., Ahmed, M. N., Hemayed, E. E. & Farag, A. A.(1998). Novel Surface Registration using the Grid Closest Point (GCP) Transform, *Proceedings of International Conference Image Processing*, 0-8186-8821-1, Chicago, October 1998, pp.(809-813).

Zhang, Z., Deriche, R., Faugeras, O. & Luong, Q. T. (1995). Robust Technique for Matching Two Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry, *Artificial Intelligence*, Vol.78, Nos.1-2, (October 1995), pp.(87-119) 00043702

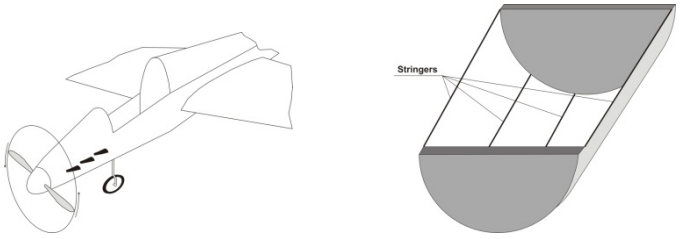
23

Coupling Experiment and Nonlinear Numerical Analysis in the Study of Post-Buckling Response of Thin-Walled Airframe Structures

Tomasz Kopecki
Rzeszów University of Technology, Poland

1. Introduction

Rational approach to design of load-carrying structures seems to suggest the necessity of focusing special attention on crucial areas of these structures, which are decisive for durability and reliability of the structure. The presence of such crucial areas in a designed solution, resulting usually from its practical functions, should be given careful consideration in view of opportunity to introduce appropriate changes in the design solutions possibly before costly and time-consuming workshop realization of a prototype.
The intent of the author is to draw attention to gravity of the factor integrating nonlinear numerical analysis with an experiment – in a broad sense of this word. Then, this chapter presents a methodology that can be used for assessment and current improvement of numerical models thus ensuring correct interpretation of results obtained from nonlinear numerical analyses of a structure.
The proposed methodology is based on carrying out experimental examination of selected crucial elements of load-carrying structures parallel with their nonlinear numerical analysis. Special attention is paid to factors determining proper realization of adequate experiments with emphasis placed on the role which the model tests can play as a fast and economically justified research tool that can be used in the course of design work on thin-walled load-carrying structures.
The presented considerations are illustrated by an example structures, whose degree of complexity and deformation range is characteristic for modern solutions in the design of airframe load-carrying structures. As a representative part of the construction design, a fragment of a load-carrying structure containing an extensive cutout, in which the highest stress levels and gradients occur in the conditions of torsion resulting in post-buckling deformation states within the range of permissible loads, was selected. Stress distributions observed within that range of deformation constitute a basis for determination of the structure's fatigue life.
Research described in this chapter were carried out in two stages:
In first of them, the objects of research constituted open section cylindrical shells with relatively small diameter in comparison to the length. In this case, the point of analysis was the determination of critical loads values and forms of deformations of free edges of examined structures, which was devoid of reinforcements or reinforced with the stringers.



a) b)

Fig. 1. (a) an example of extensive cutout in the airframe structure;
(b) isolated, critical fragment of the structure

In the second stage of research, objects of research were the cylindrical shells reinforced with the stringers protected from the buckling thanks to their considerable geometrical moments of inertia. Special attention was focused on the problem of the loss of stability of the shells.

2. Post-buckling study of open section cylindrical steel shells subjected to constrained torsion

The subject to be considered concerned numerical-experimental analysis of open cross section thin-walled cylindrical shells, characterized by the relatively big ratio between length and diameter. The range of numerical calculations included post-buckling stress analysis and determination of critical loads. Two geometrically varied structures were considered. The first is an open section cylindrical shell without stiffening. The second structure represented the shell reinforced by three stringers in the closed section. Special attention was focused on the problem of buckling of the edge of the shell, which is tantamount to the collapse of the structure.

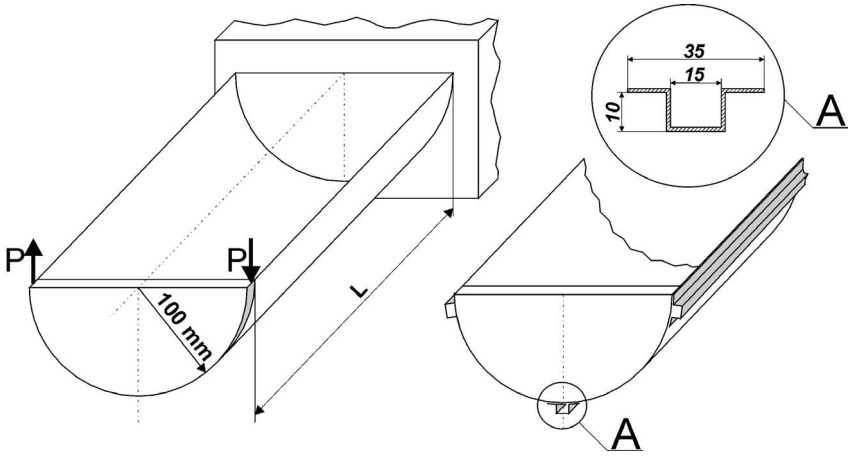


Fig. 2. Installation, loading device, and dimensions of the cross section.

Parallel to the numerical computations, for both two variants mentioned above, actual experiments were performed. A ratio of L/D=5 was assumed in both cases.

Experimental models were made of steel. The installation, loading device and structure dimensions with reinforcing stringers cross section is shown in Fig. 2.

2.1 Actual experiments

Parallel to the numerical calculations, qualitative experimental testing was performed on both variants of the structure. This provided a method for comparison of post-buckling shape and critical load values obtained in the numerical manner.

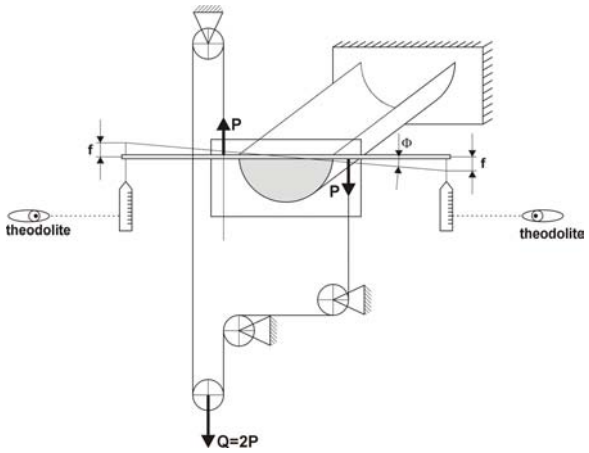


Fig. 3. The experiment device

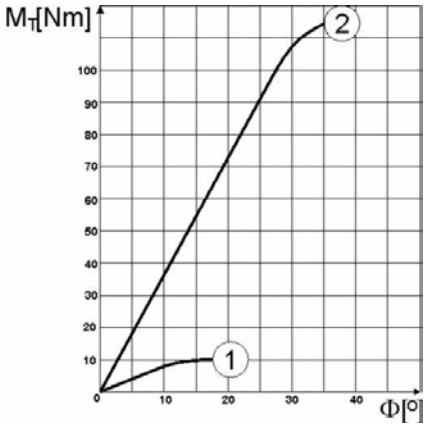


Fig. 4. Representative equilibrium path: torsion angle versus torque moment: 1 - shell without reinforcement, 2 - structure with stiffeners

The experiment was carried out on a special station able to provide the assumed loading and boundary conditions. Two cases were considered: the shell without reinforcement, and a structure with three stringers of closed section. In both cases the specimen length was L=1[m]. Figure 3 shows the design of this device.

The plot of torsion angle versus torsion moment, accepted as the representative equilibrium path, is shown by Fig. 4.
Fig. 4 shows, in the range of pre-buckling, that the load relationship between torsion moment and torsion angle is linear, exactly up to the moment of abrupt variation. Post-buckling forms of deformations are shown on fig.5 and fig.6.

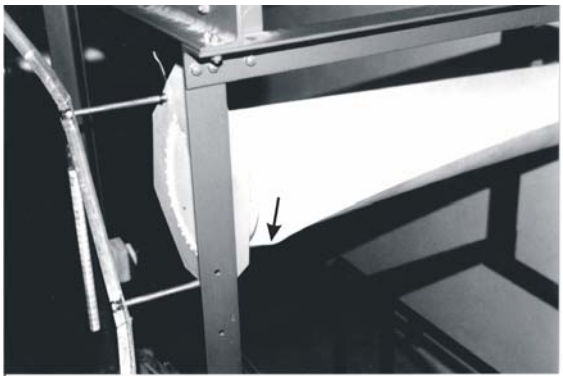


Fig. 5. Local buckling of the structure without reinforcement; L=1[m].

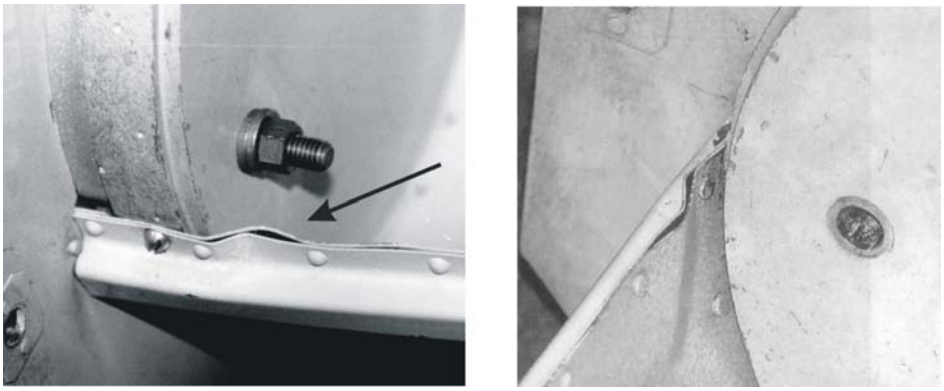


Fig. 6. Post-buckling plastic deformation. Structure reinforced by closed section stiffeners.

These experiments show that the structures considered are characterised not only by low torsional rigidity, but also by large deformations. Therefore application of the linear FEM analysis can only refer to the under-critical deformation range. It provides a way to identify stress concentration zones, possibly local buckling areas.
In order to determine the stress distribution in the post-buckling state, nonlinear static analysis was done. The stress-strain relation of uniaxial tension for actual material was simplified by the model of the ideal elastic - plastic body with a yield point of 240[MPa] (Fig.7).

2.2 Nonlinear FEM analysis

The results of the experiments showed that even a small load increment over the critical value leads to local plastic deformation. Numerical simulation of post-buckling deformation

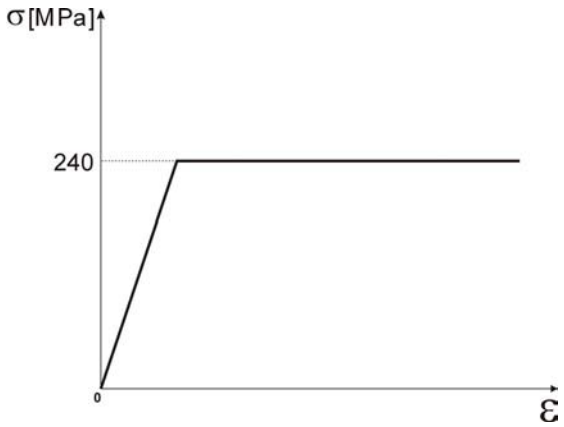


Fig. 7. Stress-strain relation of idealised material.

requires a nonlinear application. Large deformations and the change of the structure's rigidity have to be taken into consideration.
Nonlinear formulation of the problem is managed by the discrete equilibrium equations encountered in nonlinear static structural analysis, formulated by the displacement method presented in the compact force residual form

$$\mathbf{r}(\mathbf{u},\boldsymbol{\Lambda}) = \mathbf{0}. \tag{1}$$

Here **u** is the state vector, containing the degrees of freedom that characterize the configuration of the structure; **Λ** is an array of control parameters, containing the components of external loading, whereas **r** is the residual vector containing out-of-balance forces conjugated to **u.** Varying the vector **r** with respect to the components of **u** in the assumption - **Λ=**constant, a tangent stiffness matrix **K** in a structural mechanics application can be written as:

$$\mathrm{K}=\frac{\partial r}{\partial u}\,. \tag{2}$$

An alternative version of equation (1) is the force-balance form:

$$\mathrm{p(u)=f(u,\Lambda)}\,. \tag{4}$$

The **p** vector contains components of internal forces, resulting from deformation of the structure; however **f** are the control-dependent external forces, composing the set introduced respectively during the next stages of the analysis, which may also be dependent on the current geometry of the structure.
The philosophy of the nonlinear analysis in FEM is based on the gradual application of control parameters, completed in further stages. It corresponds to the stage for every reliable state of the structure in which a static balance is specific for a corresponding solution of equation (1). Control parameters connected to external force components are generally expressed as functions of reliable quantity **λ,** called the stage control parameter. The result of the nonlinear analysis composes a set of solutions, corresponding to each value of the **λ**

parameter. They create the equilibrium path of the system. The unambiguous graphical interpretation of the equilibrium path is possible for at most two degrees of freedom. However with the knowledge of external loading, the value of stage control parameters and related geometric structural configurations, it is possible to obtain an approximated dependence between selected values describing deformation of the structure versus external loading. For the numerical models considered, equilibrium paths were determined in the method: torsional moment versus total torsional angle.

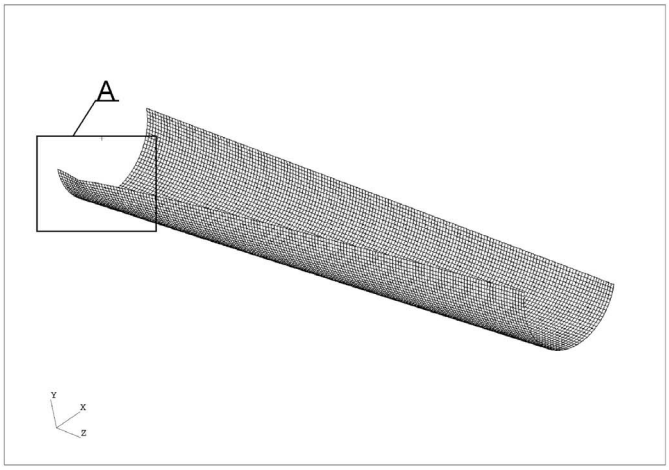
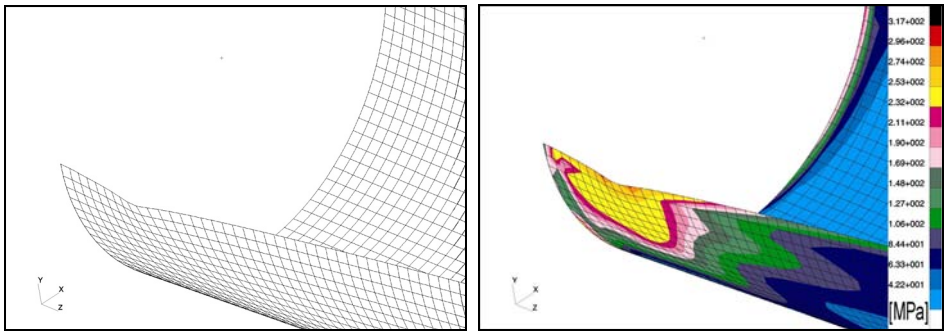


Fig. 8. Post buckling deformation. Structure without stringers

Algorithms of nonlinear analysis are mainly based on iterative and incremental - iterative procedures. The stiffness matrix **K** is treated in every equation stage as a constant and it is increased as far as the **λ** stage control parameter is increaded. The Newton-Raphson algorithm constitutes the basic iterative method. Its' drawback is that it cannot obtain the solution convergence. This method is bound up with the appearance of the limit of bifurcation points on the equilibrium path. In such situations the arc length method is applied, which makes it possible to determine the balance of the system.
Nonlinear numerical analyses of this problem were done applying the MSC MARC 2007 programme. This programme allows the user to intervene in the iteration parameter selection.
Two diversified numerical models were analyzed, with the structure stiffened by the stringers of "omega" in the cross section of each. The first has a surface - stringer rivet joint simulated by beam elements. Contact was also reflected between the surfaces of stringers and the surface itself. A simplification was applied in the second model, relying on the continuous connection of stringers with surface.
After several numerical tests, the boundary conditions of all models were changed, due to their excessive stiffness. The establishment of the back edges of the shells, as shown in Fig.1 was replaced by ribs with additional supports.
The effects illustrating the character of deformation in numerical calculations and effective stress distribution on external structural surfaces are shown by Figs. 8-13.



a) b)

Fig. 9. Form of post buckling deformation; a) area of A detail in actual scale; b) effective stress distribution on external surface according to von Mises' criterion.

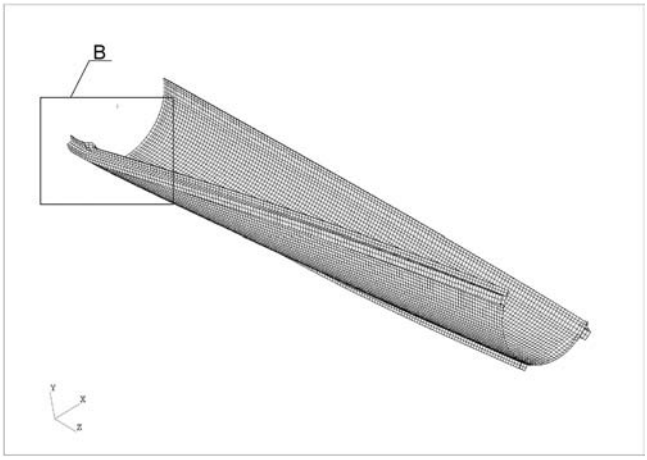


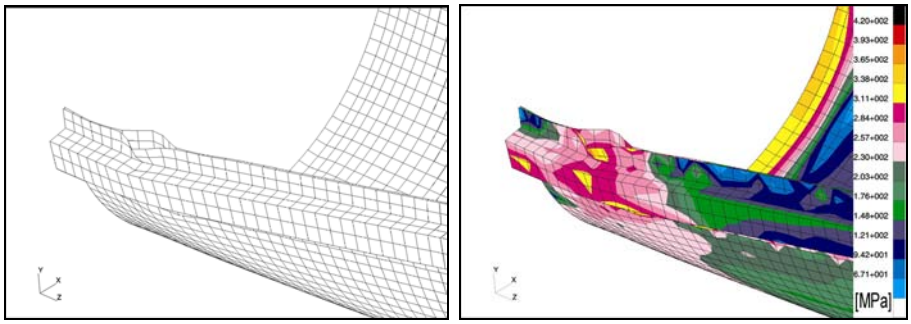
Fig. 10. Reinforced structure -version 1- riveted joint of the surface with the stringer. Local post-buckling deformation.

The results of the calculations presented in Figs. 8 and 9 prove the existence of local plastic deformation areas in the vicinity of the boundary fixing of the structure. These effects show satisfactory compatibility with the experiment (compare Fig. 5), both in the location and the character of the plastic deformation range.

Figures 10 and 11 show the numerical results for the model, where the connection was reflected by inner rivets. The result obtained, describing the state of local plastic deformation in a post-buckling state, differs qualitatively from the effect noted in the experiment.

Several attempts were made (not presented here) to identify reasons for this divergence. It is possible, on this basis, to make an attempt to explain this phenomena.

Looking at Fig. 11, we can notice that the elements of the surface and stringer in the zone between rivets were subjected to transverse dislocations moving in opposite directions, while the direction of dislocations were the same. This divergence could be the result of a loss of stability bifurcation. A reasonable suggestion arises from the fact that in both



a) b)

Fig. 11. Form of post-buckling deformation; a) area of C detail in actual scale; b) effective stress distribution on external surface according to von Mises' criterion.

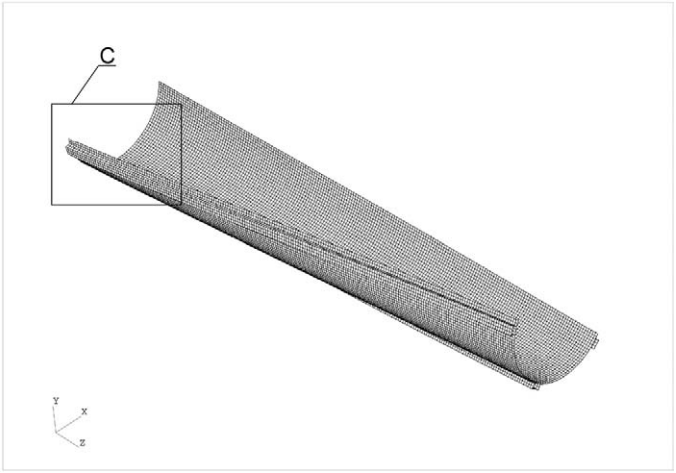
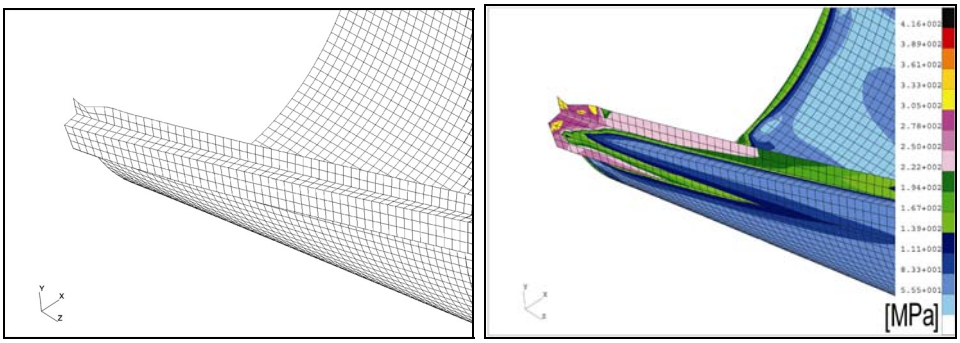


Fig. 12. Reinforced structure - version 2 - continuous connection of surface and stringers. Local post-buckling deformation.

elements the bifurcation had a stable-symmetrical character. In the actual structure, geometric imperfections could determine the identical direction of the dislocation of both surfaces already initiated during the riveting process.

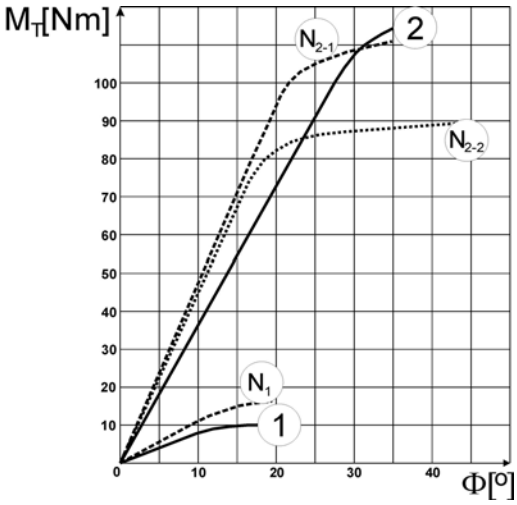
The second model (Figs. 12-13) was of considerable interest. The results of numerical calculations correspond exactly to the results of the experiment (Fig. 6). Applied simplifications adjust conditions of the iteration parameters selection in the actual structure transformation. They rely on the continuous connection between stringers and surfaces which eliminates the possibility of local stress concentration in the proximity of the rivets. Taking into consideration the character of advanced plastic deformation as noted, and the responding stress distribution, it is possible to regard the results obtained as satisfactory.

In Fig. 14, the relationship between the torsion moment and the torsion angle is shown as obtained both in the experiment and in the calculation.



a) b)

Fig. 13. Form of post-buckling deformation; a) area of B detail in actual scale; b) effective stress distribution on the external surface according to von Mises' criterion.



1 - Structure without stringers – experimental result
2 - Structure reinforced by closed section stiffeners – experimental result
N_1 - Structure without stringers – numerical result
$N_{2\text{-}1}$ - Version with continuous connection of surface and stringers – numerical results
$N_{2\text{-}2}$ - Version with riveted joint of the surface with the stringers – numerical results

Fig. 14. Relationship between the torsion moment versus the torsion angle.

It is necessary to emphasize that the results of the numerical calculations present approximate relations between the loading and the accepted parameter determining structure deformation. In fact, the obtained characteristics express the relationship between the torsion angle of the structure and the product: $\mathbf{M_{max} \cdot p_t}$, where $\mathbf{p_t}$ denotes a pseudo-time coefficient, as the step of load advantage application in the particular step of counts, whereas $\mathbf{M_{max}}$ is the maximum value of the structure loading. In the case considered, it is a maximum value of the torsional moment. The relationship is dependent on the accepted

method of the solution, the parameters of the iteration, and the shape of the equilibrium path between the products mentioned above and the actual loading of the numerical model. It should be noted that the loading of the numerical model is the maximum accepted value of the torsional moment for $\mathbf{p_t = 1}$.

2.3 Concluding remarks

On the basis of the numerical and experimental results several statements could be formulated, essential for engineering practice.

- The results obtained numerically show higher critical load values in all considered cases. It is possible that this can be explained by a rather imprecise rigidity reflecting the actual design in the numerical model as whole. It is related to the plate boundary conditions in particular. Additionally, the structure stiffness execution process should be considered.
- Establishing the back edge of the shell by limiting its degrees of freedom causes excessive stiffness in the numerical model of the structure. It is necessary to apply boundary conditions reproducing actual mount flexibility.
- The obtained divergence between the nonlinear numerical analysis and the results of the experiments suggests that appropriate imperfections of structure geometry in the numerical model should be taken into consideration. The effect would be the ability to propose reliable inferences, if not requirements, in relation to the technological process, particularly for neuralgic zones determining the load capacity of the structure.
- The presented study denotes experimental revision, information about structure behaviour under loading, and a verification function for the numerical FEM model in particular, where the solution of the problem requires nonlinear formulation.

3. Numerical and experimental analysis of torsined cylindrical shells reinforced with strenghten stringers

The subject of the second stage of research was a thin-walled open-section cylindrical shell stiffened by means of longitudinal stringers (Fig. 15), which model a zone with an extensive cutout (e.g. cockpit in an airframe structure). Such zones are usually joined with neighboring closed-section structural elements of cylindrical or slightly convergent shape. The planes of the joints determine boundary conditions eliminating the possibility of free deplanation of the outward sections.

A dimensioning form of load applied to the structure that, in the range of permissible load levels, can lead to loss of local stability of the shell, is the torsion retaining the character of a constrained one because of boundary conditions.

In order to determine the effect of longitudinal stiffening members on torsional rigidity within the full range of the analyzed deformations, structures with three, five and seven stringers were examined.

In the course of experimental work, photographic registration of subsequent deformation phases was carried out with simultaneous recording of vales of the torsion angle as the parameter enabling development of a representative equilibrium path. The tests were carried out both in sub-buckling range and in advanced post-buckling deformation states.

In order to determine the stress field in post-buckling deformation states that could constitute a basis for numerical analysis of the structure's fatigue life, a numerical model based on the finite elements method, for one of design solutions, was developed. The final

shape of the model was developed on the grounds of comparative analysis of deformation patterns and the nature of representative equilibrium paths that were obtained experimentally and numerically.

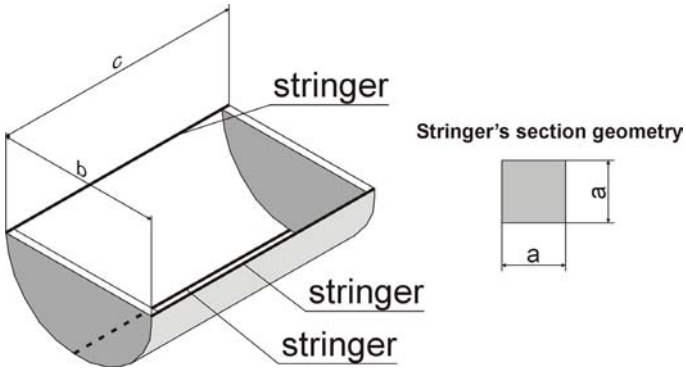


Fig. 15. Schematic geometry of the structure (a=10mm)

3.1 Experimental research

Three variants of the examined structure, differed by the number of stringers were used. In the first variant, the structure was reinforced with three stringers, in the second one — with five, and in the third variant — with seven stringers. To reproduce boundary conditions ensuring torsional rigidity of the structure corresponding to the constrained torsion, both end sections of the shell were provided with 20 mm plates. Schematic diagram of the experimental setup is shown in Fig. 16.

The structure was made of polycarbonate, for which the tensile strength test was carried out and material constants determined, i.e. Young's modulus E = 3000 MPa and Poisson ratio ν = 0.36.

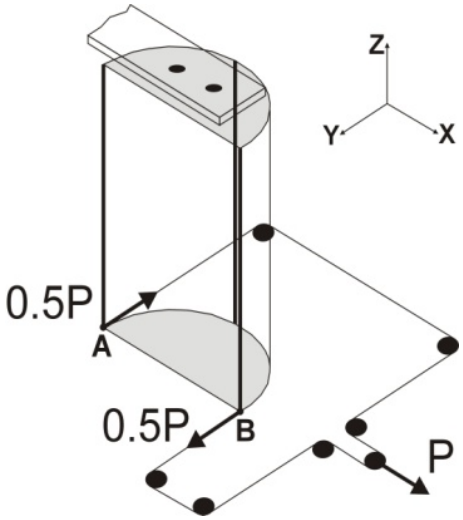


Fig. 16. Schematic diagram of the experimental setup

Fig. 17 shows the characteristic of the above-mentioned material corresponding to one-dimensional tensile stress. Clearly visible elastic and inelastic deformation zones suggest the possibility that the actual material characteristic can be approximated by an ideal elastic-plastic model. Moreover, because of low value of its elasticity modulus (by two orders of magnitude lower than that of steel) it was possible to carry out experiments at low values of external loads.

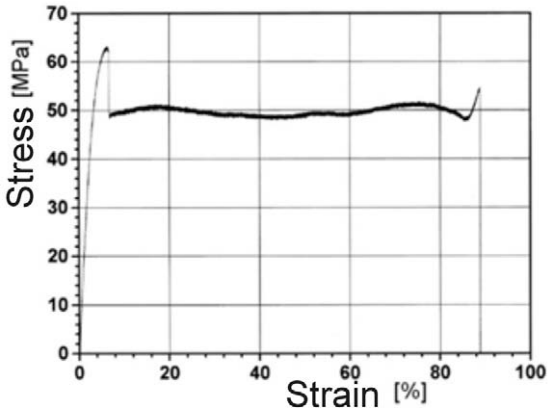


Fig. 17. Tensile stress plot for polycarbonate sample

The choice of the material, apart from the above-mentioned physical characteristics, was also justified by its high optical activity thanks to which it became possible to obtain qualitative information about optical effect distribution in circular polarization conditions. Joints between the shell and the stringers were realized by means of steel bolts spaced 20 mm apart. In order to avoid possible assembly stress at bolt joints, continuous observation of isochromatic fringe pattern fields in the vicinity of each bolt was performed throughout the whole assembly work. A view of the experimental stand with the model mounted on it is presented in Fig. 18a.

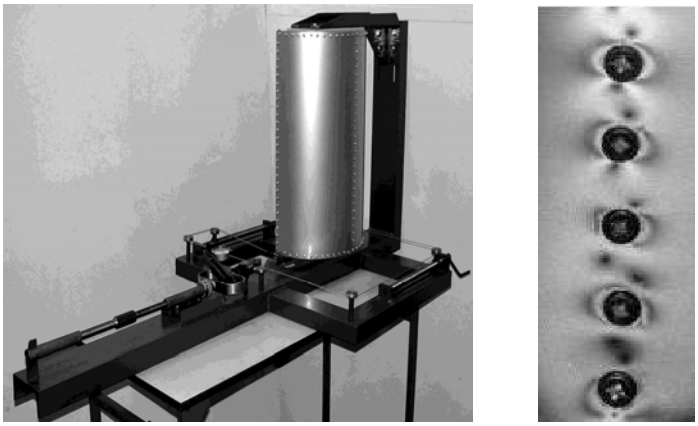


Fig. 18a. Experimental setup with the model fixed and ready for tests. 18b. Distribution of isochromatic fringe patterns in vicinity of bolts subject to control during assembly.

3.2 Structure reinforced with three stringers

The experimental work started with the structure reinforced by means of three stringers. The experiment was carried out with the load controlled by means of gravitational method ensuring good stability of load values. At the same time, one performed measurements of maximum values of the torsion angle. Based on these results, one determined a functional relationship between the twisting moment values and the model torsion angles representing an equilibrium path for a selected representative degree of freedom (cf. Fig. 22).

The first perceptible indications of loss of stability were observed in the vicinity of frames and external stringers, at the twisting moment value of $M_t \approx 20$ Nm, corresponding to torsion angle $\Theta \approx 2°$. While the load was increasing, the post-buckling equilibrium pattern covered larger and larger portion of the shell gradually reaching a global character.

It must be emphasized that the bifurcation-free character of the equilibrium path determined in the experiment was the result of the way in which the model was loaded and the possibility to perform the measurements only in steady-state conditions. Actually, the occurrence of large shell deformations and changes in their forms are connected with the jumps, in the course of which deformation increases although the load does not change. In fact, one deals here with a number of bifurcation processes, while the obtained equilibrium path represents the general character of dependence of the total torsion angle vs. twisting moment.

Fig. 19 presents an advanced form of post-buckling elastic deformation of the structure. The outward stringers were subject to significant deflection, while the central stringer did not change its original form. The obtained deformations confirm low torsional rigidity of the structure in advanced states, therefore it can be concluded that the use of solutions based on small number of stringers would be of little practical interest.

View from the inside. View from the outside.

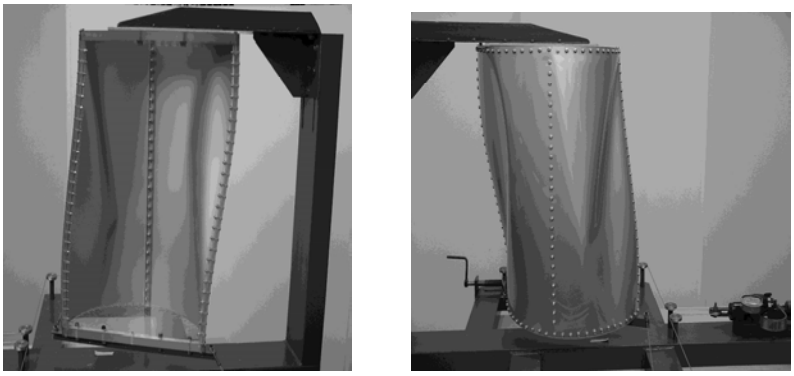


Fig. 19. Advanced post-buckling deformation, M_t = 60 Nm. M_t=60 Nm. (structure with three stringers)

3.3 Structure reinforced with five stringers

Another variant subject to examination represented a shell reinforced by means of five equally-separated stringers. The increase of the number of stringers was aimed at examining the expected increase of torsional rigidity, especially in advanced deformation states. As in

the previous variant, the result of the experiment consisted in developing a plot showing the relationship between the twisting moment and the structure's torsion angle (Fig. 22).

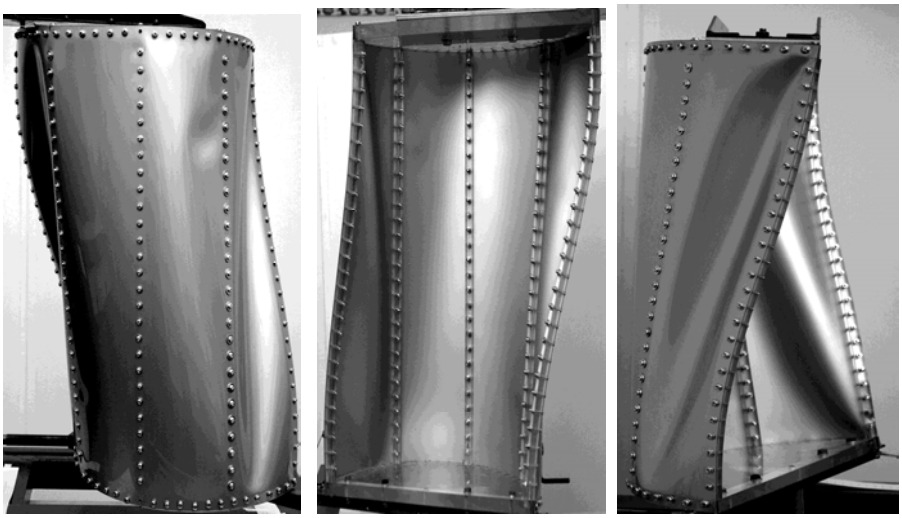


Fig. 20. Advanced post-buckling deformation phase (structure with 5 stringers), M_t = 75 Nm.

The representative equilibrium path shows that loss of stability occurred here in a way similar to that observed in the case of structure with three stringers. The presence of additional stringers placed between the outward ones and the central member (Fig.20) did not have any significant effect on rigidity of the examined structure.

3.4 Structure reinforced with seven stringers

The increase of the number of stringers to seven resulted in a significant increase of torsional rigidity of the structure. A considerable change was observed in the form in which the shell was loosing its stability, consisting in distinct increase of load on the stringers. The relationship of the structure's torsion angle vs. the twisting moment obtained in the course of experiment is shown in Fig. 22. Post-buckling deformation patterns are presented in Figs. 21. The process of local buckling was initiated in the vincinity of boundary stringers at twisting moment value of M_t = 33 Nm. Deformation of the shell proceeded smoothly. The experiment ended after the twisting moment reached M_t = 110 Nm. It was found that largest deformations occurred in two outermost segments adjacent to the shell edges. The presence of additional stringers resulted in a decrease of the depth of the folds which, in fact, ment an increase of the system's torsional rigidity. Deformations of the shell and stringers in the central part of the structure remained small compared to the deformations observed in the first two variants.

Fig. 22 presents a comparison of three characteristics corresponding to the three analyzed solutions of the structure design. In comparison with the first two variants, the structure reinforced with seven stringers revealed a significant increase of torsional rigidity. For instance, at twisting moment value M_t = 55 Nm, the torsion angle of the structure with three stringers amounted to 20° compared to only about 6° in the case of the third variant. The applied reinforcement shows therefore significant effect on the increase of torsional rigidity as well as on the values and patterns of stress distribution in the structure.

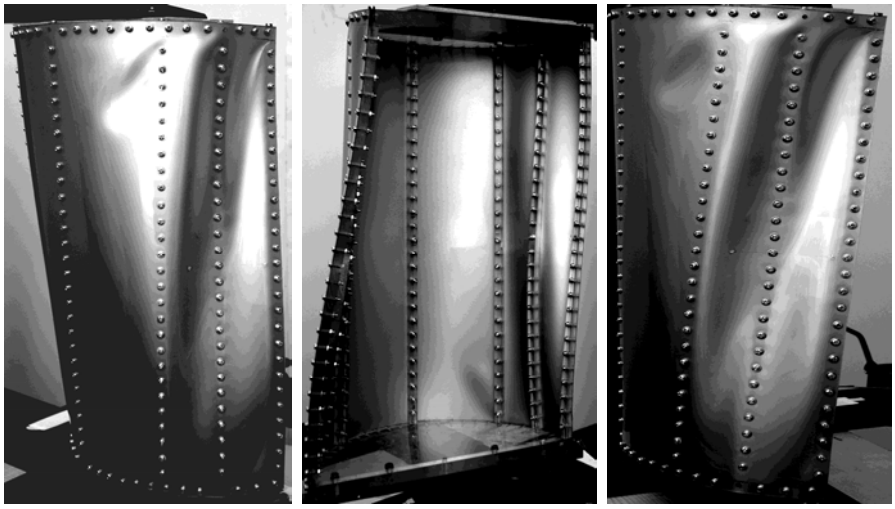


Fig. 21. Advanced post-buckling deformation pattern (structure with 7 stringers), M_t=110 Nm

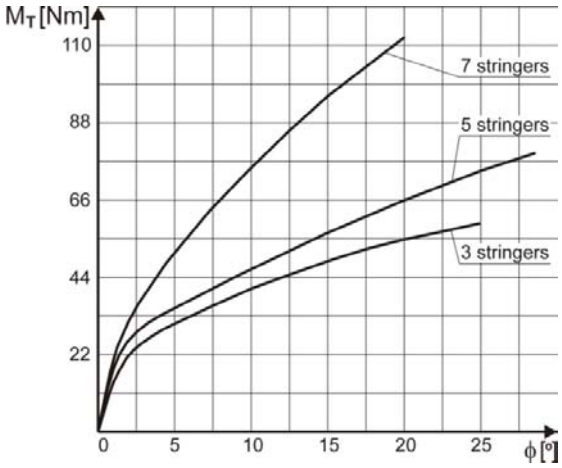


Fig. 22. Comparison of representative equilibrium paths.

3.5 Examination by means of optical polarization methods

Optical sensivity of the material of which the test structure was constructed allowed for observation of distributions of optical effects in polarized light. In the case of bending/membrane state occurring in post-buckling deformation conditions, the optical effect observed in circularly-polarized light can not be identified with isochromatic fringe patterns as the principal stress axes can vary along the shell thickness.

These effects are overestimated compared to isochromatic fringe pattern values, and the degree of overestimation increases with the increasing angle between the directions of principal stressesof bending and membrane states.

Fig. 23. Optical effect distributions: a model with 3 stringers (left), a model with 5 stringers (right)

The quantitative analyses of the observed patterns can be biased with a significant error. Nevertheless, the observed optical effects represent a source of vital information useful in determination of high stress gradients zones that can then be particularly helpful when developing a numerical model aimed at determination of stress fields characteristic for advanced post-buckling deformation states.
Figs. 23–24 present an example distributions of optical effects observed in the examined models are.

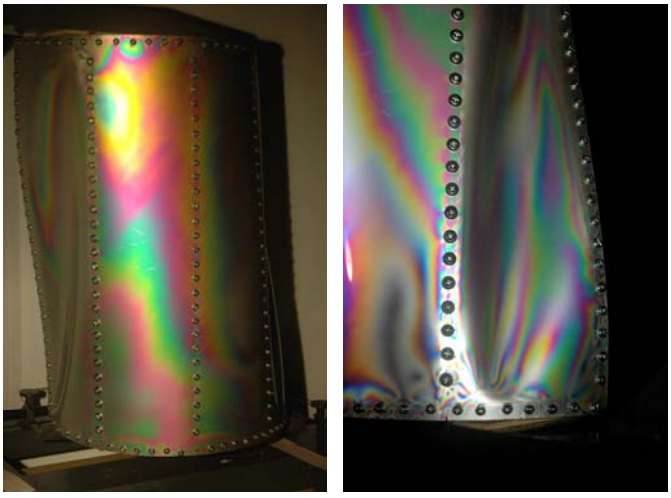


Fig. 24. Optical effect distributions — a model with 5 stringers

3.6 Nonlinear numerical analysis

In the considered issue, one accepted the maximum values of torsion angle of the structure and the twisting moment of a couple of the forces, acting on the frame closing the boundary

section as representative parameters for determining the equilibrium path. As a base of the analysis process, one applied the Newton – Raphson method and the corrective strategy based on arc-length control concept formulated by Riks – Wempner. The attempts to apply the Newton – Raphson method with the basic correction phase, did not allow for obtaining a convergent solution. Reliability of the obtained results was assessed by comparing the equilibrium path to the result of experiment, as well as assessing the compatibility of deformations obtained in the numerical way and those obtained experimentally. Information from both of these comparisons suggesting the necessity of making multiple corrections in the numerical models. Creating those models, one took into consideration a number of factors resulting from the necessity to retain some features of the real structure. Most important was to use a correct set of finite elements to make an appropriate simplifications in the geometry of the modeled structure and to appropriate reproduce the boundary conditions. In order to meet the above requirements, it was necessary to perform a number of numerical tests. Fig 25 illustrates the concept of boundary conditions reproduction, according to which the nodes fixing the structure are represented by two constraint points with blocked translational and rotational degrees of freedom. This corresponds to the real manner of fixing the structure by means of two bolts fastening one of the outward frames to the test stand.

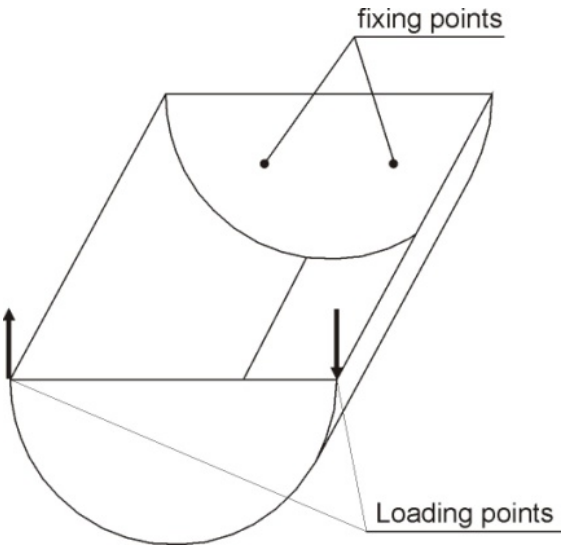


Fig. 25. Schematic diagram of model fix and load

Nonlinear numerical analyses were performed using the MSC MARC-2007 software package. Several model versions of the structure, with three stringers, and with various ways of reproduction of stringers and their joints with the shell were created.
In all cases, the shell was modeled using a bilinear, thin-shell element. The frames, and in some cases the stringers, were modeled using bilinear, four-node thick-shell element with global displacement and rotations as the degree of freedom. Bilinear interpolation was used for the coordinates, displacement and the rotations. The six degrees of freedom per node are as follows:

u, v, w - displacement components defined in global Cartesian x,y,z – coordinate system, Φ_x, Φ_y, Φ_z - rotation components about global: **x, y, z** – axis respectively.
In order to reproduce the structure stiffness ensuring compatibility with experiment results, three kinds of elements were applied for the modeling stringers: bilinear thick-shell element, elastic beam element and three-dimensional arbitrary-distorted brick elements.
Elastic beam with transverse shear is a straight beam in space which includes transverse shear effects with elastic material response. Linear interpolation is used for the axial and the transverse displacements as well as for the rotations. Section forces are output as: the axial force, local $\mathbf{T_x}$, $\mathbf{T_y}$ - shear forces, bending moments about the **x, y** – axes of the cross section respectively, torque moment about the beam axis.
The three-dimensional arbitrary-distorted brick element is an eight-node, isoparametric, arbitrary hexahedral. As this element uses trilinear interpolation functions, the strains tend to be constant throughout the element. The element can be used for all the constitutive relations. There are three global degrees of freedom: u, v, w – at each node.
In the one of the models, joints between stringers and the shell were reproduced in a discrete way, by means of beam-type elements, with simultaneous use of contact preventing elements interpenetrate in advanced deformation states. In other cases, joints of continuous kind were used. The results of calculation show that the way in which the joints between shell and stringers were modeled, had little influence on the character of effective stress distribution in the structure shell. Moreover, the model with discrete joints is characterized by high complexity, which – as it was proved in numerical tests – makes the convergence of nonlinear analyses much more difficult to obtain and leads to deformation patterns incompatible with those obtained in the experimental way in the post-buckling range.
Fig. 26 presents comparison of effective stress distribution according to Huber-Mises-Hencky criterion in the pre-buckling range, for the models with stringers constructed of 3D elements and joints realized by means of two different methods.

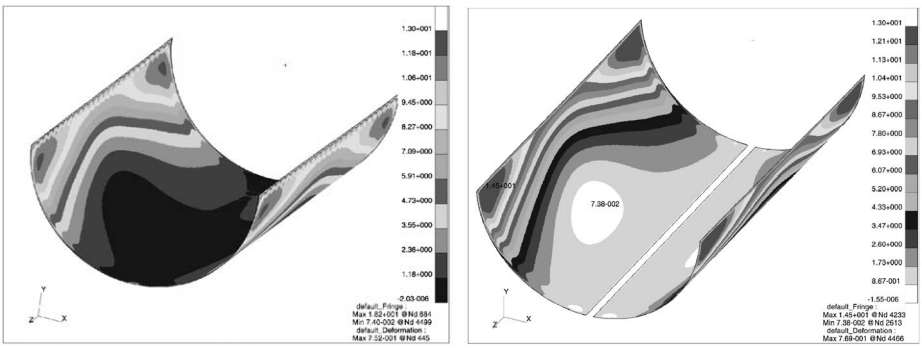
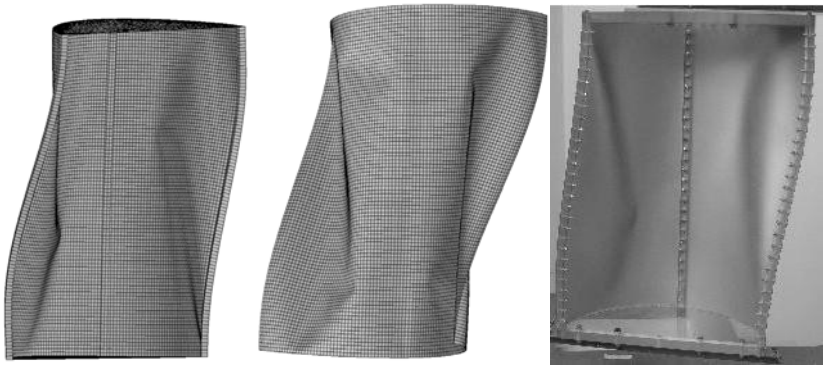


Fig. 26. Comparison of stress distributions in the middle layer according to Mises' hypothesis obtained for the discrete joints model (left) and the continuous joints model (right)

Conformity of deformations was obtained for two version of the models, with continuous joints. In the first one, the stringers made of three-dimensional elements were used. In the second case, one applied one-dimensional stringers, made of beam-type elements.

Comparison of the calculated deformation patterns corresponding to maximum load with the deformation observed in the course of experimental investigation is presented in Fig. 27. For both models, the equilibrium paths representing the relationship of twisting moment vs. torsion angle, complied with the similar ones, measured experimentally. The comparison of equilibrium paths obtained by means of numerical calculations with the results of experiment is presented in Fig. 28.
By comparing the values of the representative state parameter (torsion angle) pertaining to specific values of the control parameter (twisting moment) one can conclude that the maximum divergence between results obtained numerically with those measured in the experiment for the highest applied loads amounts to about 30%.



a) b) c)

Fig. 27. Post-buckling deformation patterns under the same load:
(a) a model with stringers made of 3D elements;
(b) a model with stringers made of beam elements;
(c) actual structure under load

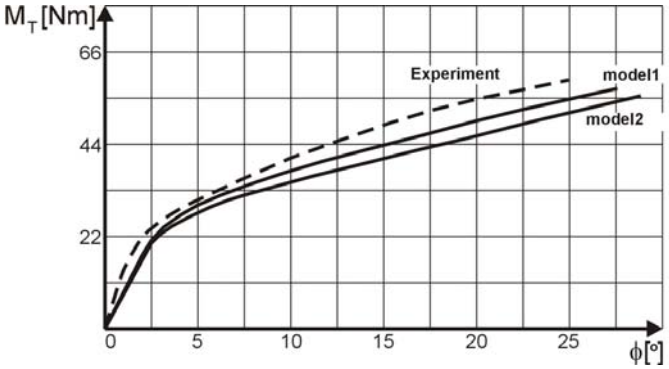
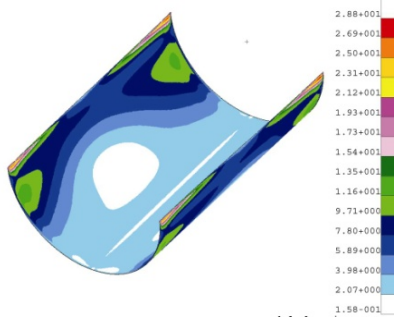


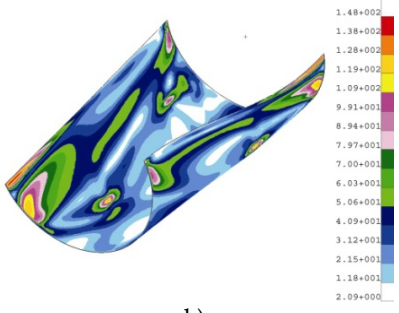
Fig. 28. Comparison of representative equilibrium paths: model1 — stringers of 3D elements; model2 — stringers of beam elements

As it follows from the above comparison of the characteristics, the model reproducing stiffness of stringers based on 3D elements shows better conformity with experiment than

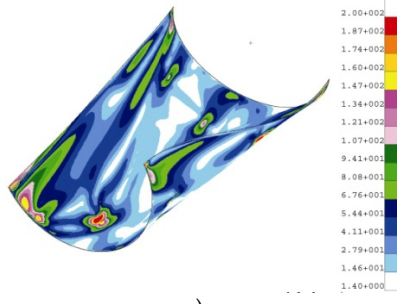
the model based on beam elements. The obtained convergence of equilibrium paths and deformation patterns allows to conclude that stress distributions obtained as the result of numerical analyses are close to the actual ones, and therefore can be adopted as a basis for estimates concerning durability of the structure. Fig. 29 and 30 present comparison of stress distributions according to Huber-Mises-Hencky hypothesis for different deformation phases. The analysis of that juxtaposition leads to the conclusion that in the post-buckling deformations phase some significant redistribution of stresses occur, and therefore the use of simplified calculation algorithms based on linearized analysis of stability has no grounds in the case of structures of that type.



a)

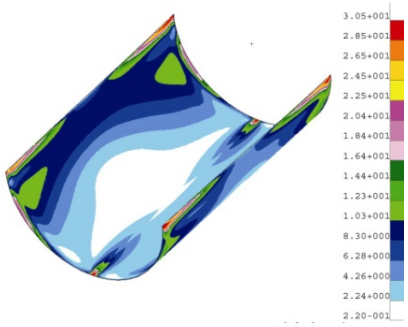


b)

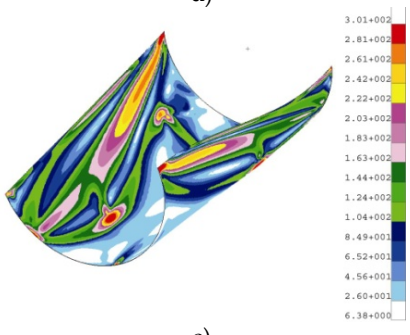


c)

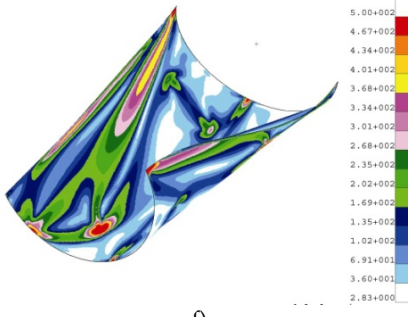
Fig. 29. Stress distributions according to Mises' hypothesis: (a model with stringers represented by beam elements) a) 25% load — middle layer; b) 70% load — middle layer; c) 100% load — middle layer



d)



e)



f)

Fig. 30. Stress distributions according to Mises' hypothesis: (a model with stringers represented by beam elements): d) 25% load — external layer; e) 70% load — external layer; f) 100% load — external layer.

After suitable comparison, the satisfactory similarity between isochromatic fringe pattern distribution in the experimental model and the stress distribution according to Mises' hypothesis in numerical model was found (Fig. 31). It allowed to accept numerically obtained results as reliable.

At the next step, nonlinear numerical calculations of models reinforced with five and seven stringers were carried out. Creating FEM models, one retained identical boundary conditions as described above. All analyses based on the same prediction methods and correction strategies as in case of the structure with three stringers.

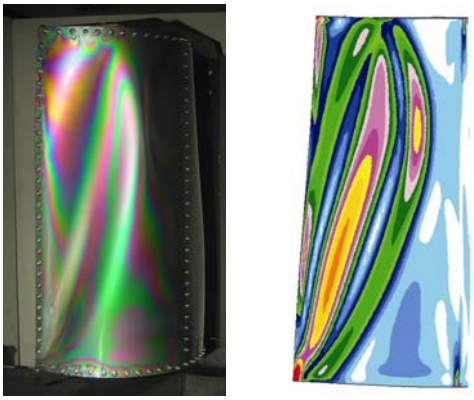


Fig. 31. Comparison of isochromatic fringe pattern distribution in the experimental model with stress distribution according to Mises' hypothesis — model with three stringers

Fig. 32 and 33 present comparisons of deformations obtained during experiments with suitable results of nonlinear numerical calculations, for the same value of twisting moment (80% of the maximum). Representative equilibrium paths are presented on Fig. 28.

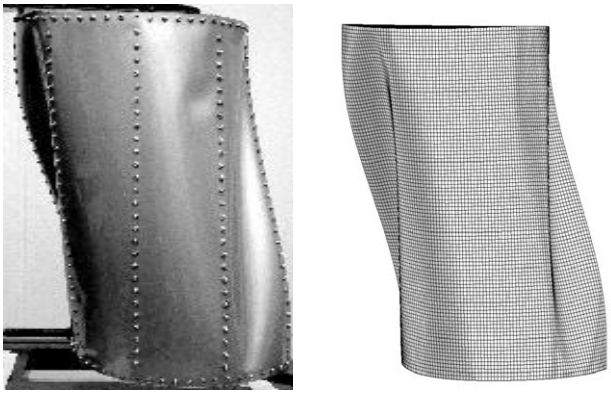


Fig. 32. Structure reinforced by 5 stringers – comparison of deformation forms obtained as the result of experiment (left) and nonlinear numerical analysis (right) – stringers reproduced with beam elements

3.7 Concluding remarks

The performed numerical analyses and experimental investigation lead to several conclusions of, how it seems, cognitive and utilitarian significance.

- The concept of performing the numerical analyses associated with experiment refers to the observed tendency, according to which the contemporary design of load-carrying structures, should be supported by permanent improvement of numerical formulations. The common feature of the two approaches to structure design is idealization of structures. We have in mind that, in engineering applications, the FEM is an approximate method and the received results refers not to the real structure, but to an

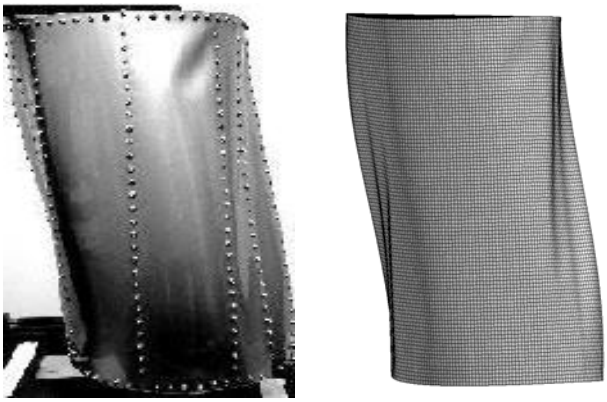


Fig. 33. Structure reinforced by 7 stringers – comparison of deformation forms obtained as the result of experiment (left) and nonlinear numerical analysis (right) – stringers reproduced with brick elements

idealized model. So that, without denying the importance of non-linear numerical analyses as tools of unquestionable effectiveness, we must keep in mind the problem of unreliability of results. It seems then that experimental verifications of FEM analyses is justified and sometimes indispensable. Such verification significantly increases reliability of results obtained on the numerical way.

- Despite the limitations connected with the possibilities of proper interpretation of optical effects, analyzed in terms of quality considerations, the results of photo-elastic examinations may provide the significant information about the existence of stress concentration areas in pre-buckling and post-buckling states, even before numerical models are created.
- The results of experimental investigation used for improving numerical models and for correcting computing procedures, pointed the necessity of taking into consideration the corrective phase of nonlinear analysis. Neglecting this phase leads to incompatibilities of numerical and experimental results or may resort in the lack of convergence of the solution.
- It is necessary to emphasize that compatibility of deformations in experimental and numerical models validated the credibility of stress fields received as the result of numerical computation.
- There are some other, more detailed remarks and conclusions that follow on the suggested procedure. For example, the results of numerical analyses and experimental examinations proved, that the method of modeling of connections between the stringers and the shell did not have any significant effect on the character of global deformation. It is also important to notice that the model based on stringers fixed with discrete connections to the shell, which has great complexity, makes it more difficult to obtain a convergent solution in the non-linear analysis.

4. References

Andrianov J., Awrejcewicz J., Manewitch L.I. (2004) *Asymptotical Mechanics of thin-walled structures.* Springer, Berlin, Germany

Awrejcewicz J., Krysko A. (2003) *Nonclassical thermoelastic problems in nonlinear dynamics of shells.* Springer, Berlin, Germany

Awrejcewicz J., Krysko V.A., Vakakis A.F. (2004) N*onlinear dynamics of continuous elastic systems.* Springer, Berlin, Germany

Arborcz J. (1985). *Post-buckling behavior of structures. Numerical techniques for more complicated structures.* Lecture Notes In Physics, 228, USA

Bathe K.J. (1996). *Finite element procedures*, Prentice Hall, USA

Doyle J.F. (2001). *Nonlinear analysis of thin-walled structures.* Springer-Verlag, Berlin, Germany

Felippa C. A. (1976): *Procedures for computer analysis of large nonlinear structural system in large engineering systems.* ed. by A. Wexler, Pergamon Press, London, UK

Kopecki T. (2010). *Advanced deformation states in thin-walled load-bearing structure design work.* Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów, Poland

Lynch C., Murphy A., Price M., Gibson A. (2004). *The computational post buckling analysis of fuselage stiffened panels loaded in compression.* Thin-Walled Structures, 42:1445-1464, USA

Marcinowski J. (1999). *Nonlinear stability of elastic shells.* Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland

Mohri F., Azrar L., Potier-Ferry M. (2002). *Lateral post buckling analysis of thin-walled open section beams.* Thin-Walled Structures, 40:1013-1036, USA

Niu M. C. (1988). *Airframe structural design.* Conmilit Press Ltd., Hong Kong

Rakowski G., Kacprzyk Z. (2005). *Finite elements method in structure mechanics.* Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, Poland

Ramm E. (1987). *The Riks/Wempner Approach – An extension of the displacement control method in nonlinear analysis*. Pineridge Press, Swensea, UK

Aben H. (1979). *Integrated photoelasticity.* Mc Graw-Hill Book Co., London, UK

24

Numerical Simulation for Vehicle Powertrain Development

Federico Millo, Luciano Rolando and Maurizio Andreata
Politecnico di Torino,
Italy

1. Introduction

Increasing concerns about environmental issues, such as global warming and pollutant emissions, as well as the predicted scarcity of oil supplies have made energy efficiency and reduced pollutant emissions a primary selling point for automobiles. As a result, the design of vehicle powertrain becomes more challenging since it has to achieve these additional targets, without compromising other performance such as power, torque or fun to drive. In this context, thanks to the exponential grow of computational power, ground transportation industry has accepted the reality that fast, efficient, and cost effective engine and vehicle development necessitates the use of numerical simulation at every stage of the design process. Within the vehicle powertrain design and development process, three different levels of modelling can generally be found:

- *Detailed modelling:* it is performed during the research and early development stages. It is mainly focused on single powertrain components, such as for instance the internal combustion engine or the electric motor of a Hybrid Electric Vehicle (HEV), providing detailed information about their behaviour, while it cannot study the whole system.
- *Software in the Loop (SiL) modelling:* it is carried out later in the development cycle but often before production hardware is available. It has a global perspective and it can be used to evaluate vehicle performance. Moreover today SiL becomes popular in control system development, such as for instance in the development of HEVs energy management systems.
- *Hardware in the Loop (HiL) modelling*, which is conducted once production controllers are available.

Therefore, this chapter will provide a description of different methodologies which can support engineers in each phase of the vehicle powertrain design process. Starting from the description of different modelling approaches, the chapter will then go deeper into the analysis of numerical models for the main powertrain subsystems (such as for instance the internal combustion engine, the electric motors for HEVs, etc.).

Finally, two case studies of numerical simulation applied to powertrain development will be presented; the first focused on the evaluation of vehicle performance, paying particular attention to the engine behaviour under transient conditions, the second aiming instead to the assessment of the fuel economy potential of different Hybrid Electric Vehicle architectures.

2. Modelling methodology

Since one of the main objectives of vehicle models is to estimate the engine fuel consumption and pollutant emissions in different driving conditions, the next section will provide a brief description of the three most common modelling approaches suitable for this application.
A conventional powertrain, with a unique power source represented by an internal combustion engine will be considered for the sake of simplicity.

2.1 Kinematic approach

The kinematic approach is based on a backward methodology (see Fig. 1) where the input variables are the speed of the vehicle and the grade angle of the road (Genta, 1997). The engine speed can therefore be easily determined from simple kinematic relationships, starting from the wheel revolution speed and the total transmission ratio of the driveline, while the traction force that should be provided to the wheels to drive the vehicle according to the chosen speed profile can be calculated from the main vehicle characteristics (i.e. vehicle mass, aerodynamic drag and rolling resistance).

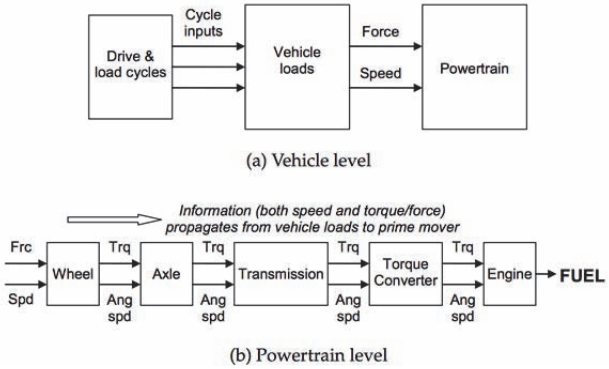


Fig. 1. Information flow in a kinematic or backward simulator (from Guzzella & Sciarretta, 2007).

Once both engine torque (or the Brake Mean Effective Pressure, BMEP) and speed have been determined, a 0D black box model of the engine (see following sections) can be used to find the instantaneous fuel consumption or emission rate, as shown in Fig.2. Finally, instantaneous fuel consumption and emission rate are integrated over the driving cycle to obtain cumulative data.
Obviously, this approach neglects all the dynamic phenomena considering transient conditions as a sequence of stationary states; therefore it is often used only for a first preliminary estimation of the fuel consumption or engine emissions of a motor vehicle, although the simulation results can differ significantly from the experimental data due to these simplifying assumptions. Moreover, the backward approach ensures that the driving profile will be exactly followed, but, on the other hand, there are no guarantees that a given vehicle will actually be able to meet the desired speed trace, since the power request is directly computed from the speed and it is not checked vs. actual powertrain capabilities (Guzzella & Sciarretta, 2007). Finally such an approach also neglects the thermal transient occurring after an engine cold start, which is, on the contrary, usually taken into account by most of the type approval driving cycles (such as NEDC, FTP, etc.).

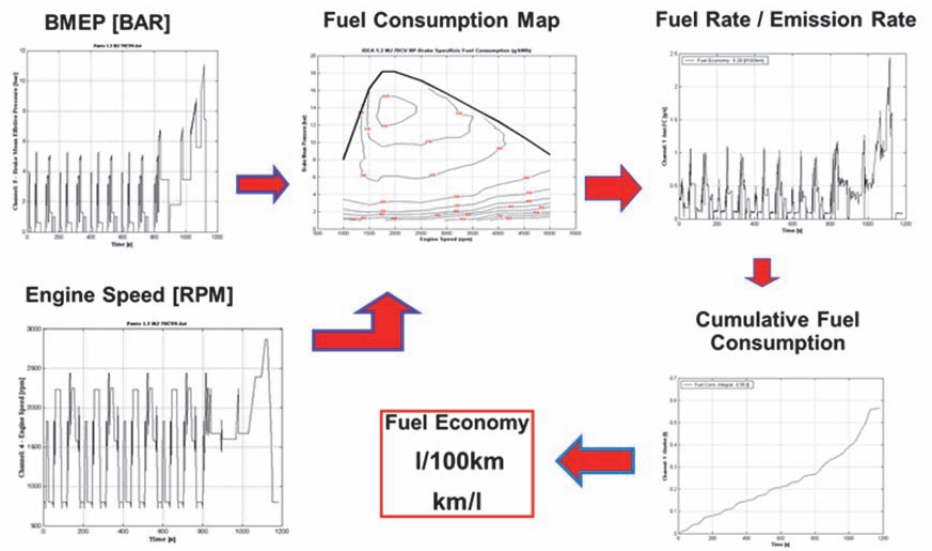


Fig. 2. Information flow in a backward model for motor vehicles fuel consumption calculation.

2.2 Quasi static approach

In the quasi static approach, a driver model (typically a PID) compares the target vehicle speed with the actual speed and generates a power demand profile in order to follow the target vehicle speed profile, by solving the longitudinal vehicle dynamics equation as shown in Fig.3 (Guzzella & Sciarretta, 2007; GAMMA TECHNOLOGY, 2009).

Once the torque and speed of the engine have been determined, fuel consumption or pollutant emissions can then be calculated by means of interpolation of engine maps, as for the kinematic methodology which was described in the previous section.

The simulation model can therefore be regarded as a "quasi-static" model, since, although system dynamics are taken into account, the behaviour of the main devices (ICE, EM, batteries) is described by means of steady state performance maps.

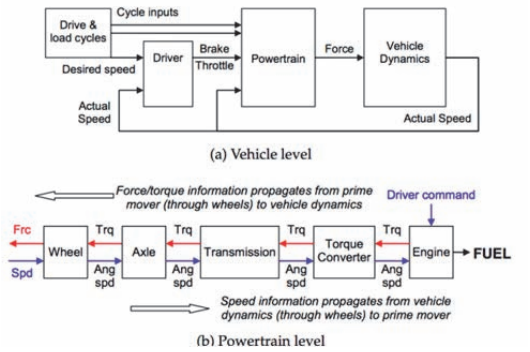


Fig. 3. Information flow in a quasi-static powertrain model (from Guzzella & Sciarretta, 2007).

Depending on the simulation targets, on the transient to be simulated and on the powertrain characteristics, a quasi-static approach can be suitable or not. For instance, for the evaluation of the fuel consumption of a vehicle equipped with a conventional powertrain on the NEDC, it can provide a reasonable accuracy, as shown in Fig. 4 (Vassallo et. al., 2007): as a matter of fact, since load and speed transients are relatively smooth for a conventional powertrain on the NEDC the simplifying assumptions of this method do not deteriorate significantly its prediction capabilities. Similar remarks could also be made as far as NOx emissions are concerned, also shown in Fig.4.

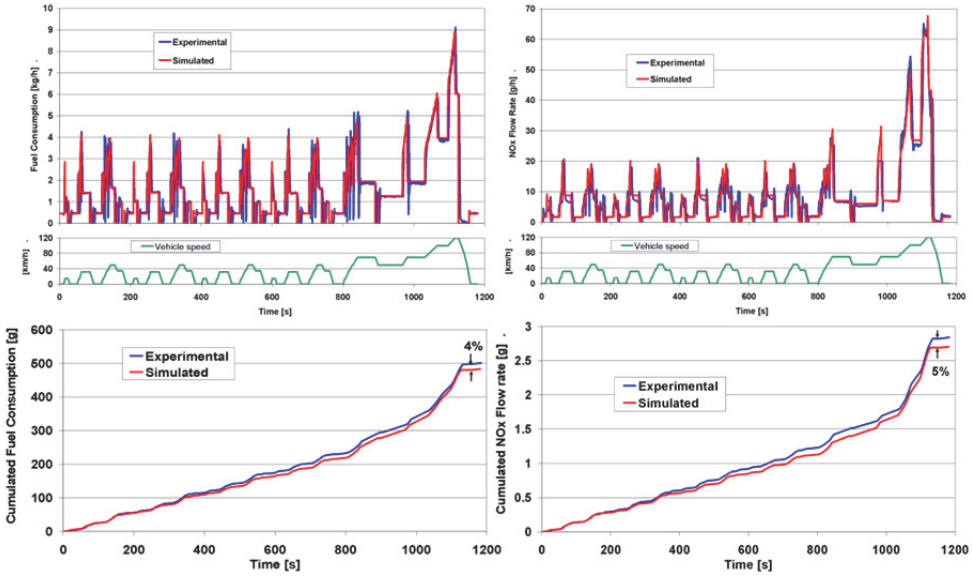


Fig. 4. Top: fuel consumption and NOx emission rate over warm NEDC cycle for a conventional powertrain. Bottom: cumulated fuel consumption and NOx emissions on the same driving cycle (from Vassallo et. al., 2007).

On the other hand, for the same driving cycle and for the same powertrain, the same approach does not provide satisfactory results when used to predict soot emissions, since for this kind of pollutant the acceleration transients and the related "turbo-lag" phenomena significantly contribute to the cycle cumulative emissions, thus requiring a more detailed engine simulation model, capable to properly capture also the engine transient behaviour. More details can be found in the case studies described in section 4.

2.3 Dynamic approach

Finally, in the fully dynamic approach, not only the longitudinal vehicle dynamics equation is solved to determine the engine speed and the torque demand, but also the internal combustion engine behaviour during transients is modelled by means of detailed OD or 1D fluid-dynamic models. For instance, for an internal combustion engine, the intake and exhaust systems can be represented as a network of ducts connected by junctions that represent either physical joints between the ducts, such as area changes or volumes, or subsystems such as the engine cylinder.

The solution of the equations governing the conservation of mass, momentum and energy of the flow for each element of the network can then obtained using a finite difference technique.
In this way, even highly dynamic events, such as abrupt vehicle accelerations during tip-in manoeuvres can be properly and reliably simulated with a reasonable accuracy, as shown for instance in Pettiti et al., 2007.

3. Subsystems analysis

Regardless of the simulation approach chosen, the performance of a vehicle powertrain model strongly depends on the methodologies which are applied to describe each powertrain component. Therefore the next sections will provide a description of the most common modelling techniques for the following powertrain sub-systems:

- Tyre
- Transmission
- Internal Combustion Engine
- Electric Motor
- Energy Storage System

3.1 Tyre

In most of the studies concerning vehicle powertrain, a detailed model of the vehicle dynamic is not required and its mathematical description can be limited to longitudinal motion neglecting the equations dealing with the lateral behaviour, (Andrzejewski & Awrejcewicz, 2005).
Obviously it is necessary to provide a model of the tyre which represents the link between the powertrain and the external environment, allowing the calculation of the forces at the interface between the wheel and the road surface. The simplest tyre model is a perfect rolling model in which the deformation of the tyre is neglected and the torque applied to the wheel shaft is transformed into a traction force considering a pure rolling motion between the tyre and the soil (Genta, 1997). According to this approach the dynamic response of the tyre can be approximated by a first order delay and the maximum force generated at the road interface can be assumed to be proportional to the vertical load on the wheel. The first order delay is useful to avoid numerical issues at very low vehicle speed, and to simulate (although in a quite approximate way) the tyre damping. The brakes are modelled as an additional torque that reduces the net torque acting on the tyre. Therefore the net torque acting on the wheel is:

$$T_{wh} = T_{shaft} - T_{brake} \tag{1}$$

Where T_{wh} is the wheel torque, T_{shaft} is the torque at the driveshaft, and T_{brake} is the braking torque. The effective traction force is then:

$$F_x = \frac{1}{1 + \tau_x s} \frac{T_{wh}}{R_e} \tag{2}$$

Where F_x is the longitudinal force at the ground, R_e the effective rolling radius, τ_x a time constant that introduces a delay between the torque and the force and s is the Laplace variable. The wheel speed ω is then:

$$\omega = \frac{V_{veh}}{R_e} \tag{3}$$

Being V_{veh} the longitudinal vehicle speed. The value of longitudinal force is determined by the vertical load acting on the wheel:

$$-F_z\mu_{x,max} \leq F_x \leq F_z\mu_{x,max} \tag{4}$$

Where F_z is the vertical force on the wheel, and $\mu_{x,max}$ is the peak value of the road/tyre friction coefficient (usually around 0.8-0.9 for dry asphalt).
If the study purposes also require the dynamic response of the tyres a *semi non-linear* model developed by Pacejka (Genta, 1997; Andrzejewski & Awrejcewicz, 2005) can be taken into account.
Once the force generated at the road/wheel interface by each individual tyre is available, it is possible to introduce the vehicle equilibrium equation:

$$M_{eqv}\dot{V}_{veh} = \sum F_x - F_{roll} - F_{aero} - F_{grade} \tag{5}$$

where M_{eqv} is the vehicle equivalent mass, $\dot{V}_{veh}$ is the vehicle longitudinal acceleration, F_x represents the traction and braking force generated by each tyre, F_{roll} is the rolling resistance, F_{aero} is the aerodynamic resistance, F_{grade} is the grade resistance. The terms in Eq. 5 can be determined as follows.

- **Rolling resistance:**

$$F_{roll} = c_{roll}(V_{veh}, p_{tyre}, \dots)M_{veh}g\cos\alpha \tag{6}$$

Where g is the gravity acceleration, α is the road slope (so that $M_{veh}g\ cos\ \alpha$ is the vertical component of the vehicle weight) and c_{roll} is the rolling resistance coefficient which, in principle, is a function of vehicle speed, tyre pressure p_{tyre}, external temperature etc. (Genta, 1997). In most cases however, c_{roll} is assumed to be constant, or to be an affine function of the vehicle speed. In this case the approximation commonly used is:

$$c_{roll} = c_0 + c_1V_{veh} + c_2V_{veh}^2 + c_0V_{veh}^3 \tag{7}$$

- **Aerodynamic resistance:**

$$F_{aero} = \frac{1}{2}\rho_{air}A_fC_dV_{veh}^2 \tag{8}$$

Where ρ_{air} is the air density, A_f the vehicle frontal area, C_d the aerodynamic drag coefficient.

- **Road slope:**

$$F_{grade} = M_{veh}g\sin\alpha \tag{9}$$

- **Inertia:**

$$F_{inertia} = M_{eqv} \cdot \dot{V}_{veh} \tag{10}$$

Where M_{eqv} is the vehicle equivalent mass that takes into account all the inertia of the driveline components and can be expressed as (if the transmission inertia can be neglected):

$$M_{eqv} = M_{veh} + J_{wheel}\frac{1}{R_e^2} + J_{eng}\frac{\tau_{gear}^2\tau_{fd}^2}{R_e^2} \tag{11}$$

Where M_{veh} is the vehicle mass, J_{wheel} is the wheel inertia, J_{eng} is the engine inertia, τ_{gear} is the transmission gear ratio and τ_{fd} is the final drive gear ratio.
The aerodynamic and rolling resistances can be experimentally determined by means of a so called *coast-down test*, consisting in a free vehicle deceleration. Since in these conditions the deceleration is due only to the rolling and aerodynamic resistances, by measuring the instantaneous vehicle speed the total drag force acting on the vehicle can be determined. The resulting relationship is then usually fitted by a quadratic function of the vehicle speed (BOSCH, 2011):

$$F_{roll+aero} = C_0 + C_1V_{veh} + C_2V_{veh}^2 \tag{12}$$

Where C_0, C_1 and C_2 are called "Coast-Down Coefficients".

3.2 Transmission

In a motor vehicle the term transmission or driveline refers to all the mechanical devices used to connect the engine to wheels. It may include clutches, torque converters, differential and, more generally, gears.
Since nowadays most of the European passenger cars are equipped with manual gearboxes, this section will focus on the description of these transmissions. They can be represented as gear couples the transmission ratio of which can change dynamically: gears are mechanical components, whose external interfaces are two mechanical connections representing input and output shafts; they can be modelled through their transmission ratios and through an efficiency function η (depending on gear ratio, revolution speed, and input torque) which takes into account all power losses due to friction. Since the speed ratio is fixed, being determined by kinematic constraints, the power losses imply the reduction of the torque at the output shaft:

$$\begin{cases} \omega_{out} = \dfrac{\omega_{in}}{\tau} \\ T_{out} = \eta\,\tau\,T_{in} \end{cases} \tag{13}$$

Where ω is the revolution speed, T the torque, τ the transmission ratio, η the gear efficiency and the subscripts *in* and *out* refer to the input and output shafts according to the power flow. The corresponding power loss is then calculated as:

$$P_{loss} = \varpi_{in}T_{in}(1-\eta) \tag{14}$$

The variable gear ratio signal deriving from the gear selection index (which is determined on the basis of the driving cycle schedule) is filtered with a 1st order transfer function that simulates the delay involved in the actual procedure of gear shifting, that usually takes a few tenths of second to be completed (Serrao, 2009).
In vehicles equipped with manual transmissions, during gear shift phases in which the vehicle and the engine speeds have to be kinematically decoupled a clutch is needed.
Dry or wet friction clutches have no torque amplification capability and produce substantial losses especially during the first acceleration phase when the vehicle starts at zero velocity.

If the engine speed ω_e is assumed to be constant during this drive away phase, the clutch dissipates the following amount of mechanical energy:

$$E_{clutch} = \frac{1}{2} J_v \varpi_{c,0}^2 \tag{15}$$

Where $\omega_{c,0}$ is the wheel velocity at which the clutch input speed ω_e and the clutch output speed ω_{gb} coincide for the first time. The inertia J_v includes the vehicle inertia and all inertia due to the rotating parts located downstream of the clutch. The amount of energy dissipated does not depend on the clutch torque profile during the clutch-closing process.
Finally, it should be pointed out that during all phases in which the clutch is slipping, the torque $T_{gb}(t)$ at the gearbox input is not limited by the engine, but is defined by the clutch characteristics and by its actuation system. The clutch torque $T_1(t)$ depends on the speed difference and the actuation input $u(t)$:

$$T_1(t) = T_{1,max}\big(\Delta\omega(t)\big) \cdot u(t), \qquad 0 < u(t) < 1 \tag{16}$$

The maximum clutch torque T_{1max} can then be approximated by:

$$T_{1,max}(t) = sign(\Delta\omega(t))\left[T_b - (T_b - T_a)e^{-|\Delta\omega(t)|/\Delta\omega_0}\right] \tag{17}$$

The parameters $\Delta\omega_0$, T_a and T_b must be determined experimentally, and they generally depend on the clutch temperature and wear (Guzzella & Sciarretta, 2007).

3.3 Engine

Depending on the application and on the required level of detail different simulation approaches can be used for the internal combustion engine modelling, as depicted in Fig. 5.

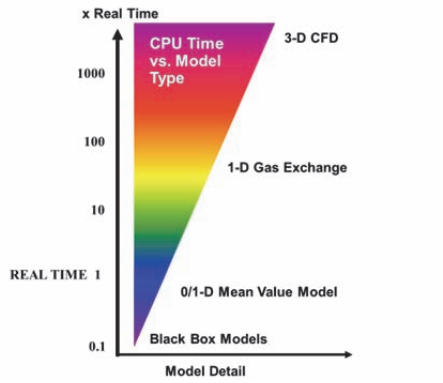


Fig. 5. Flowchart of the main engine modelling methodologies, showing model detail vs. computational time.

The most complete and detailed methodology is the 3D CFD: it is widely used for modelling flow dynamics in intake and exhaust systems of internal combustion engines, as well as for modelling in-cylinder phenomena such as gas flow through the intake and exhaust valves, direct fuel injection, mixture formation, combustion process etc. It can usually provide only component level details (such as for instance, pressure drops, flow distributions, fuel and air

mixing, etc.), but cannot usually provide a system level perspective, since, because the computational time increases with the system volume to be discretized, this approach is usually applied only to a specific engine component, such as for instance a single cylinder or a manifold, although some examples of application of 3D CFD modelling to a whole internal combustion engine, including intake and exhaust systems, have been reported in literature (Chiodi, 2010). In order to gain a system level perspective, 1D fluid-dynamic simulation tools are generally used (Keribar et. al., 2000; Morel et al., 2006), in which the engine intake and exhaust systems are usually modelled as a network of ducts connected by junctions that may represent either physical joints between the ducts or subsystems such as the engine cylinder, and the solution of the equations governing the conservation of mass, momentum and energy of the flow for each element of the network is obtained using a finite difference technique. Typical applications of 1D fluid-dynamic models may include volumetric efficiency, torque and fuel consumption predictions under steady state operating conditions, as well as predictions of engine transient response to throttle tip-in and tip-out manoeuvres, turbocharging system response, etc.

Beyond these traditional uses, 1D simulations are today expanding in the area of control system modelling towards Software in the Loop (SiL), which is nowadays a popular activity in control system development prior to prototype hardware availability. At this level, 0D black-box models which follow a quasi-static approach based on experimental steady-state maps are currently the preferred options, due to their superior real-time capabilities. This approach is generally suitable for fuel consumption and emission calculations on type approval driving cycles, where transients are quite smooth and can be simulated by means of a sequence of stationary states.

However, 0D black box models, which neglect most (if not all) of the engine dynamic phenomena, are definitely unsuitable for the simulation of fast transients, such as for instance for the predictions of engine response to throttle tip-in and tip-out manoeuvres, or of turbocharging system response, for which 1D fluid-dynamic models would certainly be more appropriate, but are usually inapplicable because of their high computational time requirements.

For this reason the development of the so called “mean value” model seems to be the most valuable solution to combine the low computational requirements of black-box models with the accuracy of 1D models: the purpose of these models is the reduction of the complexity of a detailed 1D approach, while maintaining, at the same time, a physical description of the main phenomena in order to achieve the best compromise between detail level and computational requirements. Basically mean value models simplify both intake and exhaust systems in single equivalent volume where, as in a filling emptying model, the equation of conservation of energy and mass are applied. The separation of the thermodynamic properties between adjacent volumes is then possible thanks to restricting elements which imposes to the adjacent volumes a mass flow rate, which is calculated according to its characteristics and to the boundary conditions. Therefore the overall system behaviour, including turbocharger, is still represented although there is no detail about pressure wave dynamics (see also Fig. 6). Thanks to these features, the computational time is significantly reduced, with only limited impoverishment of the model detail, thus achieving an intermediate level between system black-box models and detailed engine models, as shown in Fig. 7.

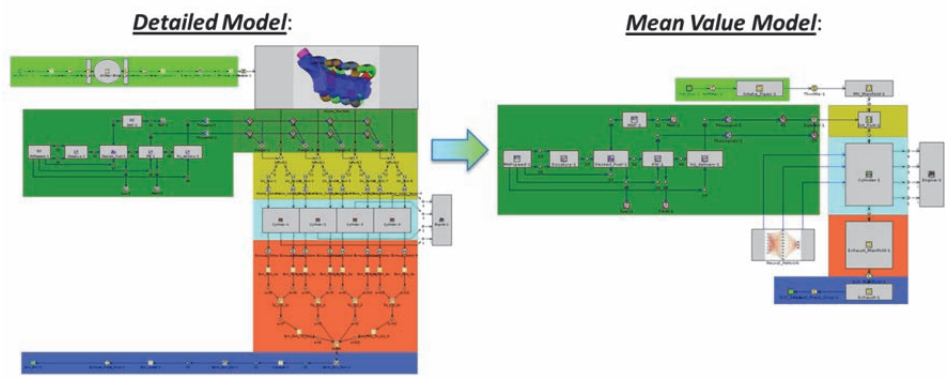


Fig. 6. Comparison between the structure of a detailed 1-D engine model (on the left) and a mean value model (on the right) – Example from Gamma Technology, 2009.

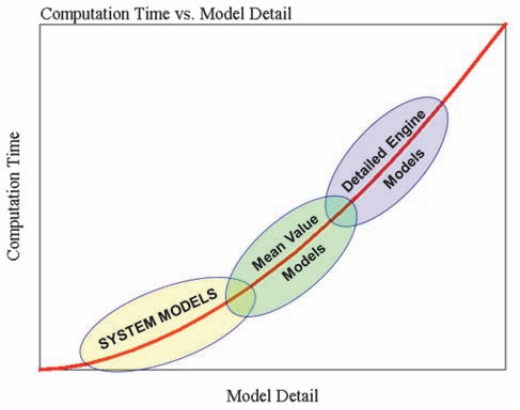


Fig. 7. Computational requirement vs. model detail for different engine modelling methodologies.

3.4 Electric motor

Nowadays the increasing share of HEVs and EVs in the automotive market makes electric machines (and in particular Electric Motors, EM) key components in vehicle powertrain development (Szumanowski, 2000).

Although the wide diffusion of electric machines in several areas of application has led to the development of different technologies (such as for instance AC and DC machines, synchronous vs. asynchronous motors, etc.) for powertrain development purposes electric machines can usually be modelled through a system-level approach similar to the 0D methodology previously described for the internal combustion engines, by means of torque and efficiency maps in which desired values of electrical power or torque are used as control inputs (see Fig. 8). Rotor inertia is the only dynamic element modelled, as the electrical dynamics are generally much faster. Thus, in most of the cases this quasi static modelling approach shows very good agreement with experimental data, and, as consequence, it is the

most widely used technique in HEVs powertrain modelling. The relation between the input and output power can be simply obtained as shown here below:

$$P_{electric} = \frac{P_{mech}}{\eta(\omega, T)} = \frac{\omega \cdot T}{\eta(\omega, T)} \tag{18}$$

Where $P_{electric}$ is the EM electrical power, P_{mech} the EM mechanical power, T the EM torque, ω the EM speed and $\eta(\omega,T)$ the EM efficiency depending on torque and speed, as shown in Fig. 8.

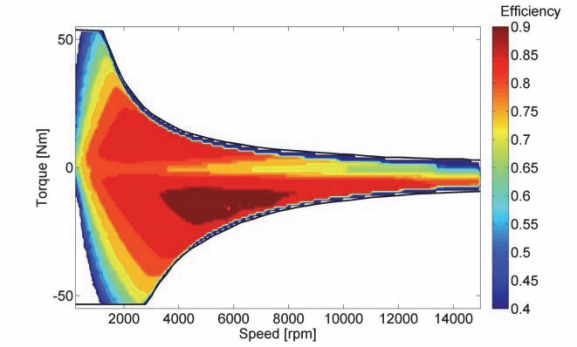


Fig. 8. Example of synchronous electric motor efficiency map.

3.5 Energy storage system: electrical battery

As for the electric machines, the relevance of electrical batteries in automotive applications has been rapidly increasing in the last years along with the rising interest for EVs and HEVs. Battery modelling is a complex task since all the main parameters that influence the battery properties (State Of Charge, voltage, current, temperature) are dynamically correlated to each other in a highly non-linear fashion.

Different modelling approaches can be adopted, depending on the detail level which is required and on the affordable computational requirement, but, regardless of the battery type (e.g. Ni-Mh, Li-Ion, etc.), the battery model purpose is generally to compute the battery State Of Charge or SOC, which is defined as the ratio between the actual charge stored in the battery and the maximum charge level, as shown in Eq. 19:

$$SOC(t) = \frac{Q(t)}{Q_{max}} = \frac{\int_0^t i(t)dt}{Q_{max}} \tag{19}$$

Where $i(t)$ is the electric current flowing into or from the battery, $Q(t)$ is the instantaneous charge stored in the battery and Q_{max} is the battery maximum charge.

The simplest battery model can then be obtained by means of the equivalent circuit shown in Fig. 9, consisting in an ideal voltage generator (U_{oc}) in series with a resistor (R_i). The Kirchhoff's law then yields the following equation:

$$U_2 = U_{OC} - R_i I_2 \tag{20}$$

The open circuit voltage U_{oc} depends on the State Of Charge SOC, and it can be computed either through performance maps experimentally measured, either through their linear approximation:

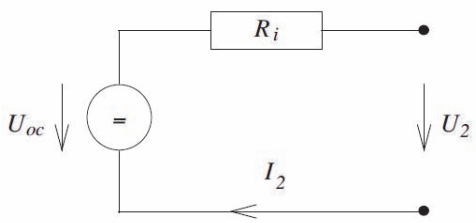


Fig. 9. Battery static model

$$U_{OC} = k_1 SOC(t) + k_2 \tag{21}$$

Where constants k_1 and k_2 depend on battery characteristics (i.e. number and connections of the cells, electrodes materials, etc.) but are generally independent from operating parameters, and are therefore constant with time.
The battery resistance R_i can then be split into the sum of three different contributions (Szumanowski, 2000):

- an ohmic resistance R_0 representing all the ohmic resistances in the electrolyte, in the electrodes and in the battery interconnections;
- a charge transfer resistance R_{ct}, which is associated to the chemical reactions at the electrodes;
- a diffusion resistance R_d which models the diffusion of reactants and products in the layer between the electrode and the electrolyte.

These resistances values depend on the battery temperature, on the SOC and on the electric current in a highly non-linear fashion. Due to the difficulties in modelling the effects of these parameters, experimental data from constant current battery discharging test are often used to define a black box model of the battery resistance R_i.
The simple battery model above described is often referred to as a *static model*, since it does not take into account the dynamic behaviour of the battery during transient operating conditions. Whenever more complex models are needed for HEVs and EVs modelling, detailed reviews can be found in literature, such as for instance in González-Longatt, 2006.

4. Case studies

The last section of this chapter will show two overviews on numerical studies concerning vehicle powertrain development. The former describes the build of a mean value engine model in order to investigate the dynamic performance of vehicles equipped with turbocharged diesel engines especially from the acceleration transient point of view. The latter shows an evaluation of different hybrid powertrains for an European mid-size passenger car, in order to obtain a preliminary assessment of their potentialities in terms of fuel consumption and pollutant emissions reductions.

4.1 Mean value model for the analysis of turbolag phenomena in automotive diesel engines (Pettiti et al., 2007)

A positive perception of the vehicle performance by the driver is currently a key factor for customer acceptance of any new car model. So, particular attention must be paid to optimize the engine and vehicle behaviour during sudden acceleration transients, such as during the

tip-in manoeuvres which are usually employed to evaluate vehicle performance. However turbocharged engines, which are becoming more and more popular to satisfy the demand of high specific power outputs, suffer from the well-known "turbo-lag" phenomenon, showing a delayed response to an abrupt torque demand, due to the time which is required to speed-up the turbocharger so to obtain an adequate boost level.
This drawback is even more pronounced in diesel engines, because due to the insufficient boost level during acceleration transients, the amount of injected fuel has to be markedly reduced to avoid excessive smoke levels, thus resulting in poor engine performance. Therefore a simulation tool allowing a quick but reliable estimation of the impact of different design and calibration choices on vehicle performance during tip-in could be extremely helpful (Gambarotta, 2001; Canova & Gambarotta, 2002), because the simplest quasi static engine models are completely unsuitable, since they do not consider any transient phenomenon, as described in the previous sections. As a matter of fact, if these models were used to simulate tip-in manoeuvres, they would lead to an unacceptable overestimation of vehicle performance for any turbocharged engine. The aim of this work is therefore the design and development of a quite simple and flexible model for turbocharged diesel engines, able to simulate a tip-in manoeuvre with satisfactory accuracy. The model can be classified as a mean value model because it does not take into account the variation of the thermodynamic properties of the working gases within the engine cycle, like a detailed 1D model does, but it deals only with cycle-averaged thermodynamic quantities.

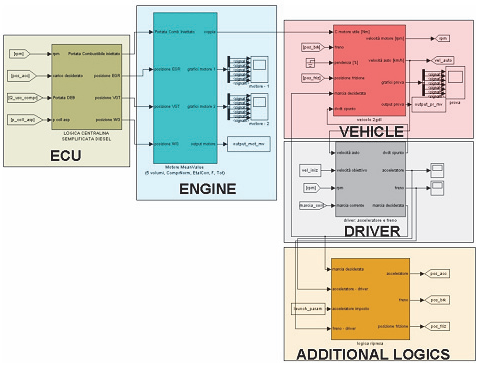


Fig. 10. Simulink vehicle mean value model: global layout

The global model layout is shown in Fig. 10: as one can see, it is made up of the following five main subsystems:

- ECU block
- Engine block
- Vehicle block
- Driver block
- Additional logics block

The ECU block receives as inputs the signals of the air mass flow meter, of the accelerator pedal position and of the current engine speed and it provides as outputs for the engine block the injected fuel mass, the VGT (Variable Geometry Turbine) and the EGR (Exhaust Gas Recirculation) actuators position. The engine model, which is the core block of the

system, takes into account main engine devices such as intake and exhaust systems, turbocharger and engine cylinder which are modelled through single control volumes.
As in filling and emptying models the equations of conservation of energy and mass are used to determine the thermodynamic state of each control volume, which is treated like an open thermodynamic system. Five control volumes were used in the model for the intake and exhaust systems, for the volume between compressor and intercooler and for intake and exhaust manifolds. The separation of the thermodynamic properties between adjacent volumes is possible thanks to restricting elements such as for instance valves, turbochargers, etc... The flow restricting element imposes to the adjacent volumes a mass flow rate, that is calculated according to its characteristics and to the boundary conditions.
The turbocharger was modelled through the efficiency and mass flow rate maps which are usually provided by the manufacturer. However, at the beginning of a tip-in manoeuvre, the initial turbocharger speed may easily fall to very low values, far from the range included in the original map of the compressor, requiring additional extrapolation routines. Finally, a simplified method was developed to quickly evaluate the turbocharger inertia, which is quite rarely supplied by the manufacturer together with the mass flow and efficiency maps.
The main output of the engine block is the delivered torque whose current value is passed to the vehicle block where the dynamic equilibrium equations for the driveline and the engine are solved in order to determine vehicle performance (i.e. longitudinal acceleration) and engine speed. This block also needs as inputs the brake and clutch positions, determined by the driver which is basically modelled through a PID controller.
The vehicle was simply modelled with two degree of freedom (one for the driveline and the other for the engine), i.e. a simple rigid driveline model was used. Despite its simplicity, this model showed to be suitable enough to simulate tip-in manoeuvres, mainly because in such conditions the turbocharged engine gradually delivers its torque, so that oscillating phenomena of the driveline are remarkably reduced if compared to naturally aspirated engines. Nevertheless at least a tyre model including slip would be necessary to improve simulation accuracy during "fast" tip-in manoeuvres, such as for instance second gear manoeuvres.
To assess the reliability and the accuracy of the model, an extensive comparison with experimental data was carried out. The test vehicle was an European compact car equipped with a 1.3 litre displacement VGT turbocharged diesel engine. Because a previous analysis carried out in Coltro, 2005 showed that the parameter which can be more closely related to the driver perception of vehicle performance is the vehicle acceleration pattern as a function of time, the comparisons were mainly focused on the abovementioned quantity. Moreover, this parameter can be particularly useful to highlight how the use of steady state models for evaluating the performance of turbocharged engines during tip-in transients may be inaccurate especially at the beginning of the manoeuvre. As one can see in Fig. 11 for a tip-in manoeuvre from 40 km/h in IV gear, the acceleration predicted with a steady model can be even double than the experimental value. On the contrary, the simulation carried out with the mean value model appears to be accurate, although a not negligible overestimation of the vehicle acceleration can still be noticed.
This can be ascribed to a corresponding overestimation of the boost pressure, as shown in Fig. 12. The possible reasons could be the absence of a fully detailed ECU control logic reproduction, as well as of a detailed model of the VGT actuator. Nevertheless, although an appreciable overestimation of the calculated vehicle acceleration can be observed also in this case, the accuracy of the prediction could be still considered as satisfactory.

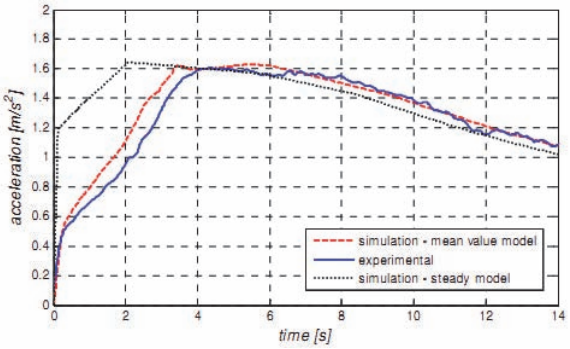


Fig. 11. Acceleration vs. time. Manoeuvre: tip-in from 40 km/h, IV gear.

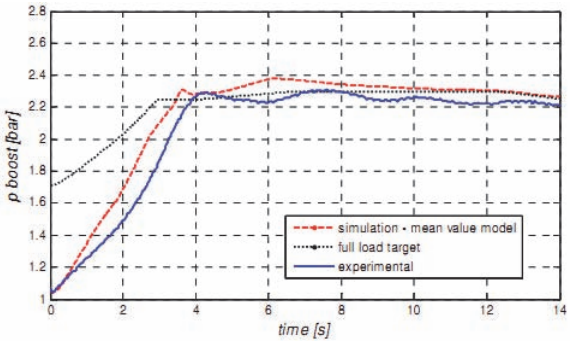


Fig. 12. Boost pressure vs. time. Manoeuvre: tip-in from 40 km/h, IV gear.

After assessing the model reliability and accuracy, the simulation tool could be used to evaluate the impact on vehicle performance of different design choices (such as, for instance, different gear ratios values) or calibration strategies (such as, for instance, the use of different "smoke maps" for fuel limitation during transients). As an example, the comparison of the acceleration patterns that could be obtained by means of three different smoke maps which limits the maximum injected mass is shown in Fig. 13.

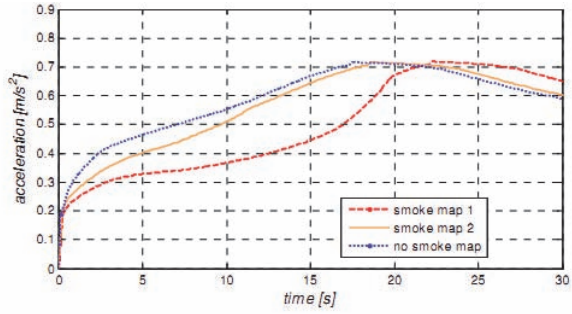


Fig. 13. (Effect of different smoke maps) Acceleration vs. time. Manoeuvre: tip-in from 60 km/h, VI gear.

In conclusion, the proposed model allows an easy and quick investigation on the impact on vehicle performance of several design parameters, such as, for instance, the inertia of the turbocharger or the gear ratios of the driveline. Moreover, the model is also suitable to evaluate the impact of different control strategies, concerning the smoke-related fuel injection limitation, the boost control through the VGT position and the EGR control.

4.2 Development of a control strategy for complex light-duty diesel-hybrid powertrains (Millo et al., 2010)

The new CO_2 emission targets set by the EC (130 g/km over NEDC to be reached within 2012, and 95 g/km as a long term goal for 2020) are extremely demanding, especially for the long term target.

For an "average" European passenger car (compact size, about 1300 kg mass) the energy requested over NEDC can be estimated approximately equal to 0.4 MJ/km, that means that engine average efficiencies values around 32-33% will be required to reach the 95 g/km target (assuming a "pure powertrain" approach to the CO_2 target achievement, that is, without considering possible benefits coming from reduction in rolling resistance due to "green" tyres, aerodynamics improvements, vehicle body lightening, etc.).

However, even if remarkable improvements have been achieved in the last decade, it should be noted that over most of the NEDC the ICE is operated at low load, and thus with poor efficiency, thus making the 95 g/km CO_2 target far beyond current system capabilities. Consequently, it seems that a breakthrough from current technologies will be mandatory and coupling the high efficiency levels of diesel engines with the fuel saving potentials offered by hybridization could undoubtedly represent a valuable option. Nevertheless, despite these impressing potentialities, although diesel propulsion systems could undoubtedly add further benefits to Hybrid Electric Vehicles the high values of efficiencies that can be already attained by conventional diesel vehicles may make more difficult the development of hybridization strategies enabling further improvements in a cost effective way (Cooper et al., 2009).

Furthermore, some hybridization strategies aiming only to reach low fuel consumption targets may lead to unbearable penalties in terms of emissions, especially for NOx, and the add-up of the costs of hybrid and aftertreatment technologies which could be needed to meet future emissions regulations could seriously limit Diesel HEVs growth.

A careful analysis is therefore required in order to properly evaluate the potentialities of different hybrid architectures, taking into account also possible peculiar issues that may hinder their widespread application. Different degrees of hybridizations, from micro to mild hybrids, and different energy management systems were therefore evaluated in this work through numerical simulation, in order to obtain a preliminary assessment of the potentialities of different hybrid systems for the European passenger car market.

Four different hybrid architectures were considered (see Fig. 14):

- **Architecture A.** It corresponds to a typical medium size European passenger car, equipped with a Common Rail DI diesel engine (1.6 litre displacement) and a Manual Transmission (MT) gearbox.
- **Architecture B.** It is a Micro Hybrid with a Belt Alternator Starter (BAS). A small (2kW) Motor Generator Unit (MGU) is coupled to the internal combustion engine by means of a belt thus allowing the replacement of all the functionalities of both the starter and the alternator of the conventional vehicle.

- **Architecture C.** It features a Flywheel Alternator Starter (FAS): the Motor Generator Unit is installed between the engine and the gearbox and the electric power installed is one order of magnitude greater (18 kW) than the BAS. Two different variants were analysed: while the first one (which will be referred to as C1) features one clutch only located between the FAS and the gearbox, the second one (which will be referred to as C2) features a further clutch between the ICE and the FAS.
- **Architecture D.** It features the same Motor Generator Unit as architecture C, but coupled with a fixed transmission ratio to the output shaft of the gearbox. In this case the electric machine must be designed with quite different features, because it cannot benefit of the torque multiplication factor provided by the transmission, but will anyway be connected to the driveline by means of a dedicated gear with a transmission ratio equal to 5. Therefore the MGU power was reduced to 15 kW, while its maximum speed was increased up to 18000 rpm.
- **Architecture E.** It features the same 18 kW electric machine as architecture C, but connected to one of the primary shafts of a Dual Clutch Transmission (DCT)**.** This provides an additional degree of freedom because the energy management system can choose the MGU gear ratio when the ICE torque flows through the other shaft in order to achieve further fuel consumption savings.

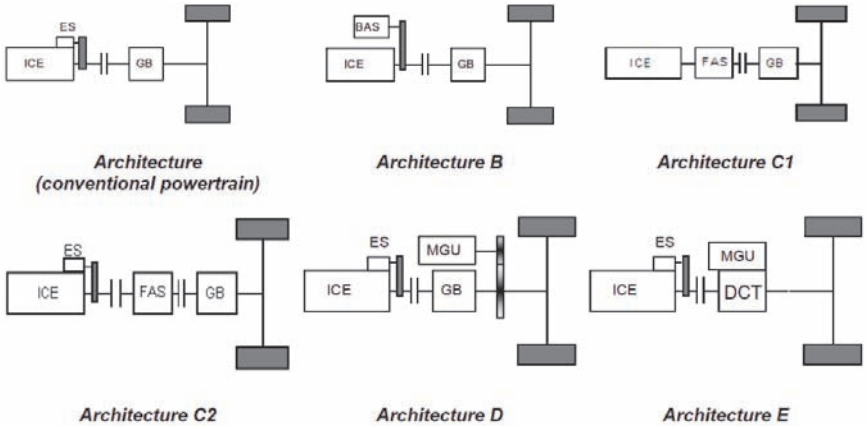


Fig. 14. Hybrid Electric Vehicles architectures (B, C, D and E) and reference conventional powertrain A

For all hybrid architectures, except the Micro Hybrid B, a Li-Ion battery constituted by 64 cells with 25 Wh energy capacity, and 730 W peak power each, for a total energy capacity of 1600 Wh was considered. A mean charge-discharge efficiency of 0.9 was considered.
HEVs simulations have been carried out with a vehicle model developed in the GT-Drive environment (GAMMA TECHNOLOGY, 2009; Morel et al., 2000), exploiting the quasi-static approach described in the previous sections. Fuel Consumption and engine out NOx emission were evaluated over the NEDC driving cycle on the basis of performance maps (see Fig. 15) neglecting warm-up phenomena due to lack of experimental data concerning the different engine calibration strategies adopted during a cold start. Moreover, since NOx emissions alone were evaluated (because they usually represent the major constraint for diesel hybridization strategies) and no NOx aftertreatment device was considered, engine-out emissions only were evaluated, and no aftertreatment model was used.

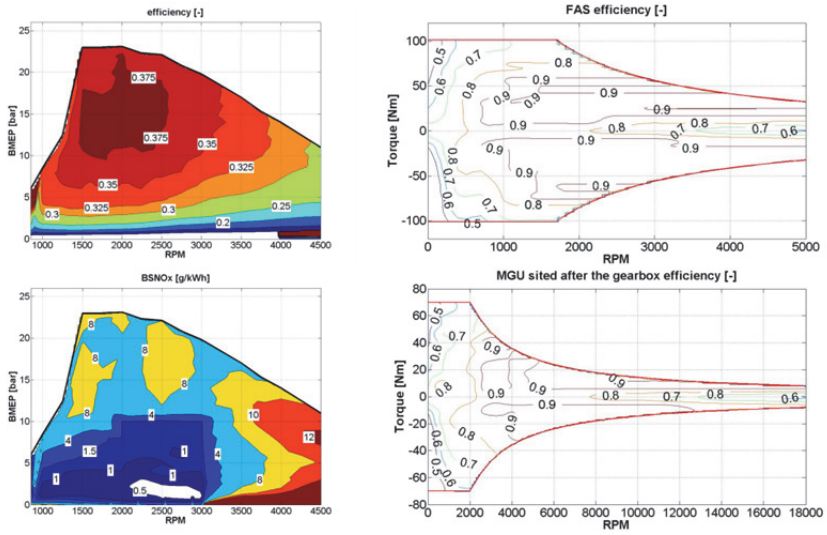


Fig. 15. Performance Maps – Left: Engine Efficiency and BSNOx maps. Right: MGUs efficiency maps.

Different power sources management strategies may be enabled depending on the hybrid architecture, as shown in Table 1. The base function enabled by the architecture B is the Stop&Start strategy. With the increase of electrical power available (architectures C, D and E), more options become available, from electrical power assistance during acceleration transients to improve vehicle performance (e-Boost), to regenerative braking. A larger battery allows also the "mild hybrid power management" strategy: whenever the State of Charge (SOC) of the battery is above a certain level, the pure electric mode is maintained until a proper power and/or vehicle speed threshold is exceeded, thus eliminating the ICE operations in the low load region which is characterized by poor efficiency figures.
Moreover, a further strategy specifically targeted to achieve a reduction of NOx emissions for the diesel powertrain can also be implemented, by using the MGU to assist the ICE even during the moderate vehicle acceleration prescribed by the NEDC driving cycle, in order to cut off the power peaks that are responsible of high NOx emissions rates. This strategy was specifically developed to solve a typical issue of diesel hybrids, that is the increase in NOx emissions produced by hybridization strategies aiming only to reach low fuel consumption targets by means of load point shift techniques.

	B	C	D	E
Start&Stop	√	√	√	√
e-Boost		√	√	√
Regenerative Braking		√	√	√
NOx cut		√	√	√
Mild Hybrid Power management		√	√	√

Table 1. Hybrid functionalities enabled depending on the hybrid architecture.

It should be pointed out that all the above mentioned strategies share a common target of a neutral battery energy balance at the end of the driving cycle, with a SOC variation compared to the initial value smaller than 1% of chemical energy consumed during the driving cycle.
The potential of the different hybrid architectures in terms of fuel consumption reduction over the NEDC has been firstly evaluated by applying to all architectures, except architecture B, a mild hybrid power management strategy with the purpose of achieving the lowest fuel consumption that will be referred to as "conventional", since it encompasses start&stop, regenerative braking and load point shift with a target of lowest fuel consumption, but without any specific attention for NOx emissions, which undoubtedly represent the Achille's heel of these solutions.
Afterwards, further investigations were carried out considering a "NOx cut" strategy. Fig. 16 and Fig. 17 show the main results obtained with the two different strategies.
The micro-hybrid architecture B, that only allows a simple Stop&Start, is capable to provide a 6% fuel consumption saving and a roughly equal NOx emissions reduction, thanks to the elimination of the idle phases.

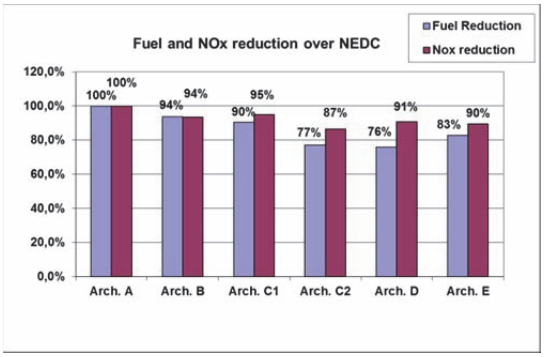


Fig. 16. Fuel consumption and NOx emissions reduction over NEDC for different diesel hybrid architectures with a "conventional" energy management strategy.

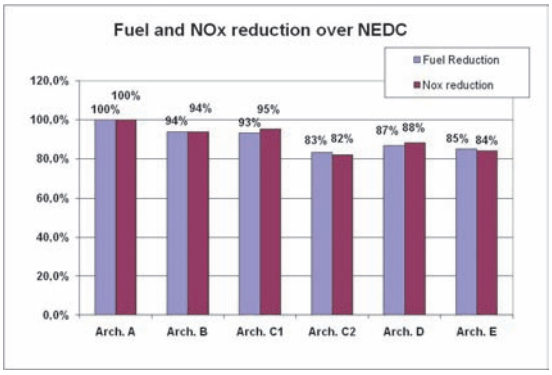


Fig. 17. Fuel consumption and NOx emissions reduction over NEDC for different diesel hybrid architectures with a "NOx cut" strategy.

On the other hand, the mild-hybrid architectures C, D and E, thanks to the higher electric power installed, increase significantly the fuel saving capabilities, reducing the fuel consumption up to 23-24% for architectures C2 and D, while allowing also an appreciable 10% average reduction of NOx emissions. The remarkable improvements in powertrain efficiency that can be achieved with architectures C2 and D are mainly due, in addition to the start&stop and regenerative braking savings, to the application of a load point shift strategy. The different performances of architectures C1 and C2 are due to the engine frictions, which in C1 diminish the regenerative braking efficiency and require more power in electric driving phases.
The results of the "NOx cut" strategy are shown in Fig. 17, where further improvements in NOx emissions in comparison with the conventional strategy can be clearly seen. On the other hand, fuel consumption penalties to be paid may range from affordable (2% for architecture E) to unacceptable (11% for architecture D). Architecture C2, which provided the highest fuel consumption and NOx emissions reductions with the conventional power management strategy, remains the best option also with the NOx cut strategy, although its results are now approached also by architecture E, the fuel consumption of which increases by only 3% while NOx emissions fall down of about 6%. Only architecture C1 (and architecture B, obviously, which does not have other capabilities beyond the start&stop) does not seem to get any emission advantage from the NOx cut strategy; again, the main reason could be the poor efficiency of the regenerative braking, due to the unavoidable ICE motoring, which is subtracting a remarkable amount of energy. For this reason, the ICE is therefore called, through load point shift, to store in the battery a huge amount of energy that will be then consumed in the EUDC segment in order to assist the ICE with the MGU. On the contrary, for all the other architectures the regenerative braking is very efficient and the ICE has therefore to generate a smaller energy quantity.
Finally, the analysis through numerical simulations of the potentialities of different hybrid powertrains highlighted that diesel hybrid powertrains, although being subject to NOx emissions constraints that could jeopardize their benefits, offered substantial advantages both in NOx emissions and in fuel consumption reduction. FAS (Flywheel Starter Alternator) with double clutch and the hybrid architecture with a MGU directly connected to the differential case by a dedicated gear were able to achieve the lowest fuel consumption levels.

5. Conclusions

In conclusion, ground transportation industry has nowadays accepted the reality that fast, efficient, and cost effective engine and vehicle development necessitates the use of numerical simulation at every stage of the design. Within the vehicle powertrain design and development process, different modelling approaches can be followed, depending on the simulation targets, ranging from fully dynamic, high fidelity models to black-box faster than real time models: it is therefore of crucial importance a clear target setting for the simulation activity, so to allow the selection of the most suitable modelling approach, in order to achieve the best compromise between detail level and computational requirements. The two case studies presented in the last section clearly support these statements: the first analysis, focusing on the evaluation of vehicle performance, requires a detailed description of the powertrain dynamics, such as turbo-lag phenomena, while the second, aiming to the assessment of the fuel economy potential of different Hybrid Electric Vehicle architectures

can exploit a simpler quasi-static approach since the analysis deals only with driving cycles with moderate dynamics.
Despite of these significant differences both models can provide valuable information to effectively support the powertrain development.

6. Nomenclature

AC	Alternative Current
BSFC	Brake Specific Fuel Consumption
BMEP	Brake Mean Effective Pressure
CFD	Computational Fluid Dynamics
DCT	Dual Clutch Transmission
EC	European Commission
ECU	Electronic Control Unit
EGR	Exhaust Gas Recirculation
EM	Electric Machine
EUDC	Extra Urban Driving Cycle
EV	Electric Vehicle
FAS	Flywheel Alternator Starter
FTP	Federal Test Procedure
GB	Gearbox
HEV	Hybrid Electric Vehicle
HIL	Hardware in the Loop
ICE	Internal Combustion Engine
MGU	Motor Generator Unit
MT	Manual Transmission
PID	Proportional Derivative Integrative (controller)
NEDC	New European Driving Cycle
NOx	Nitrogen Oxides
SIL	Software in the Loop
SOC	State of Charge
VGT	Variable Geometry Turbine

7. References

Andrzejewski, R. Awrejcewicz, J. (2005), *Nonlinear Dynamics of a Wheeled Vehicle*. Springer-Verlag, Berlin.

BOSCH R., (2011), *Bosch Automotive Handbook*, (8th Edition), Bentley Publishers, ISBN: 978-0-7680-4851-3.

Canova, M., Gambarotta, A., (2002) *Automotive engine modelling for real -time control using an object-oriented simulation library*, 2nd Int. Workshop on Modelling, Emissions and Control of Automotive Engines-MECA02, Salerno, Italy.

Chiodi, M., (2010) *An innovative 3D-CFD-Approach towards Virtual development of Internal Combustion Engines*, PhD Dissertation University of Stuttgart, Germany.

Coltro, R., (2005) *Experimental investigation on the impact of engine characteristics on the vehicle behaviour during tip-in manoeuvres*, Mechanical Engineering Graduation Thesis, Politecnico di Torino, Turin, Italy, (in Italian).

Cooper, B., Bar, F., Beasly, M., and Penny, I., (2009) *The Challenge of CO2 and Emissions Targets for Diesel Engines: Can Both Be Combined?* 9th Stuttgart International Symposium on Automotive and Engine Technology, Stuttgart, Germany.

Gambarotta, A., (2001) *A Control-Oriented Library for the Simulation of Automotive Diesel Engines*, 3rd Int. Conference on Control and Diagnostics in Automotive Applications, Sestri Levante, Italy.

GAMMA TECHNOLOGY, (2009) *GT-Suite User Manuals*, Chicago, United States.

Genta, G. (1997) *Motor Vehicle Dynamics: Modeling and Simulation,* World Scientific Pub Co Inc, ISBN: 9789810229115, Singapore.

González-Longatt, F. M. (2006) *Circuit Based Battery Models: A Review*, Proceedings of 2nd Congreso IberoAmericano De Estudiantes de Ingenieria Electrica, Puerto la Cruz, Venezuela.

Guzzella, G., Sciarretta, A., (2007) *Vehicle Propulsion Systems: Introduction to Modeling and Optimization,* Springer, ISBN: 9783642094156, Berlin.

Keribar, R., Ciesla C., and Morel, T., (2000) *Engine/Powertrain/Vehicle Modeling Tool Applicable to all Stages of the Design Process,* SAE Technical Paper 2000-01-0934.

Millo, F., Badami, M., Ferraro, C.V., Rolando, L., (2009) *Different Hybrid Powertrain Solutions for European Diesel passenger cars*, SAE INTERNATIONAL JOURNAL OF ENGINES, 2009, Vol. 2, pages 493 - 504, ISSN: 1946-3936.

Millo, F., M Badami, M., Ferraro, C.V., Lavarino, G., Rolando, L., (2010) *A Comparison Between Different Hybrid Powertrain Solutions for an European Mid-Size Passenger Car*, SAE Technical Paper 2010-01-0818.

Morel, T., Keribar, R. and Leonard, A., (2003) *Virtual Engine/Powertrain/Vehicle Simulation Tool Solves Complex Interacting System Issues,* SAE Technical Paper 2003-01-0372.

Pettiti, M., Pilo, L., Millo, F., (2007) *Development of a new mean value model for the analysis of turbolag phenomena in automotive diesel engines,* SAE Technical Paper 2007-01-1301.

Serrao, L., (2009) *A Comparative Analysis of Energy Management Strategies for Hybrid Electric Vehicle,* PhD thesis, The Ohio State University, Columbus, Ohio, United States.

Szumanowski, A., (2000), *Fundamentals of Hybrid Vehicle Drives*, Radom, ISBN: 372041148, Warsaw, Poland.

Vassallo, A., Cipolla, G., Mallamo, F., Paladini, V., Millo, F., Mafrici, G., (2007) *Transient Correction of Diesel Engine Steady-State Emissions and Fuel Consumption Maps for Vehicle Performance Simulation,* 16th Aachener Kolloquium Fahrzeug und Motorentechnik, Aachen, Germany.

25

Crash FE Simulation in the Design Process - Theory and Application

S. Roth[1], D. Chamoret[1], J. Badin[1,2], JR. Imbert[2] and S. Gomes[1]
[1]*Laboratoire M3M: Mécatronique, Méthodes, Modèles, Métiers, Université de Technologie de Belfort-Montbéliard, Cedex,*
[2]*DPS - Digital Product Simulation, Croissy sur Seine, France*

1. Introduction

With the evolution of computer science, numerical methods such as finite element methods are more and more used to understand physical problems. These tools are often used as an alternative to very costly experimental methods. Finite element analysis is already used in fields such as mechanical and civil engineering, crash analysis and biomechanics, allowing an interesting investigation of local strain and stress. In biomechanics of impacts for example, FE codes are used to simulate human body (soft tissues, bones etc.) which can be simulated with its environment: helmets, airbags, cars, barriers (Whitworth et al., 2004). A majority of mechanical design is based on static analysis but there are numerous applications however, in which it is appears necessary to take into account the highly non-linear, dynamic phenomena (Awrejcewicz et al. 2003, 2004).
For the design of a mechanical product, numerical methods are nowadays widely used at different step of the life cycle: at the beginning of the design process for design optimization (to investigate different solutions), for the comprehension of physical phenomenon which happened during a test (at a diagnostic point of view), in the development of standards etc…
For mechanical structures under impact, a lot of problems remain at different steps of the design, even if the lots of improvement have been made the last decades. A particular point concerns the way to transfer CAD models towards finite element model without loss of information. The problems of standard exchange and the data management can be raised.
The objective of the present chapter is to give theoretical foundations of crash analysis and to show how this simulation step can be integrated in the design process. Explicit Finite Element software as Radioss (Altair ©) can be used to the crash analysis. But many difficulties can arise during this analysis. Problems can come from the size of the model which can generate a time consuming simulation. So, for numerical models with lot of elements, how can this time step be reduced in order to optimize the simulation duration?
During the design process, how can the simulation, the data and the results can be managed in a context of collaborative design?
All these questions have to be raised in order to have a critical point of view and in order to use the numerical simulation for an optimized and a competitive design in a industrial framework.

2. The use of numerical simulation in the process of product design

2.1 Evolution of the product design

Within the current economical and industrial context, companies would like to obtain a better cost control and to streamline their product design in order to reach the famous "cost/quality/delay" objectives. It involves the development of new methods in design process with the enhancement of concurrent engineering contexts.

The engineering process is a set of interlinked activities and involving many actors in different areas of expertise but dependent on each other. The design process is an activity of the engineering process which is absolutely essential in the product lifecycle (AFNOR, 1994).

In the context of minimizing design time and parallelism of the activities of the design process, industrial practices have evolved from engineering process divided into sequences or phases to a concurrent engineering or integrated engineering process (Fig. 1). These concurrent design methods aim to enhance collaborative work in order to increase the responsiveness of the company to reduce costs. They are realized by a parallel design activities and the enhancement of collaborative sharing of data between resources and actors in the company.

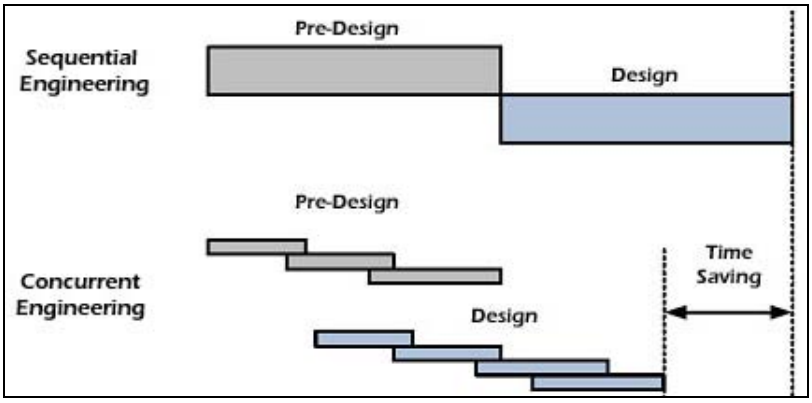


Fig. 1. Concurrent engineering approach for time saving in the design process

Design process evolutions were followed by new design method and nowadays, with the use of 3D geometrical product components in CAD files, engineers include parameters and expert rules (considering as knowledge) to drive the geometry in CAD models through parametric and variational approaches (Fig. 2). These models are termed associative because they allow engineers to easily modify the geometric of a component by changing parameters values and generate new product architecture very quickly.

The aim is to reduce routine design (80% of the estimated design process), test a large range of product architectures vey quickly, especially in the upstream phase of the design process and enhance the product quality with time and cost reduction. This is in accordance with DFX: the Design For X approach which emphasizes the importance of considering the overall constraints of several design activities, and especially in the upstream phase of design process, to avoid major conflicts and to limit the redesign cycle.

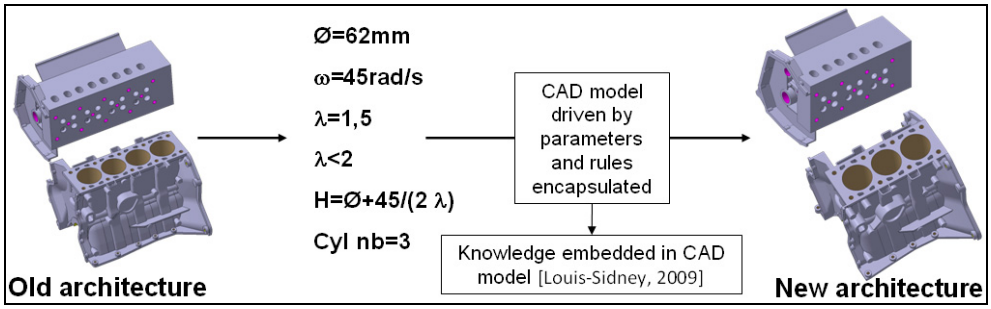


Fig. 2. Parametric design method on an automotive Power Unit

2.2 The role of numerical simulation in the design process

Although the design process has evolved, the numerical simulations have also evolved considerably to become a key area in product design. Initially used at the end of the design process as validation or presentation of activities, simulation is currently used in the overall design process and especially in the upstream phase (trade-off, pre-design) using CAD/CAE[1] integration and parametric models to drive the design and identify the better concepts of product's architecture earlier.

Thus, nowadays, it seems as necessary to use numerical simulation, especially the finite element simulation to lead the way to innovation. In the early design phases, numerical simulation allows for the management of a better design and quicker. This is particularly true in the area of mechanical systems more specifically in the automotive industry where the development speed has to be increased. That is the reason why the crash simulation techniques are gaining an increasing role in the product development instead of time-consuming validation testing.

These evolutions have led to a strong connexion between the design process and the numerical simulation and today we talk about *"simulation driven design method"* (Fig. 3). In this way it helps to streamline the design process, and to better take into account the constraints from the various expert domains in the product design and a better control. Indeed, the idea is to minimize physical prototypes which are very expensive to make and use the simulation even for the certification. Thus one of classical objective in automotive industry for future is to produce a car with just one prototype good at the first time.

3. CAD/CAE integration method

These evolutions have led to a strong connexion between the design process and the numerical simulation and today we talk about *"simulation driven design"*.

More than a connexion, it can be talk about a dependency which is not well handled yet. Indeed there exist a gap between designers and analysts. The large number of heterogeneous information handled in the design process combined with the low level of interconnection between CAD and crash simulation software tools often lead to data discrepancy and incoherence. Thus, the data and information are often scattered and duplicated, thus preventing data coherence, traceability, and reuse, and inhibiting the

[1] CAD/CAE : Computer Aided Design / Computer Aided Engineering

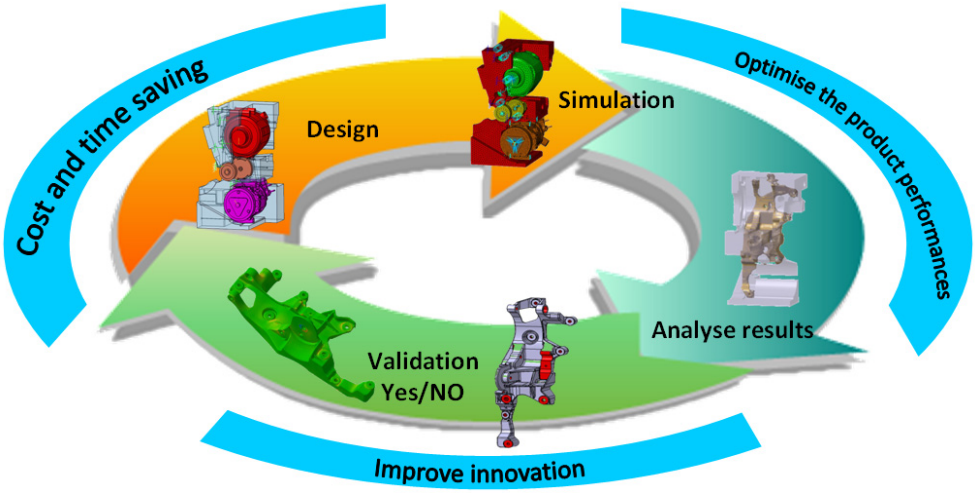


Fig. 3. Using numerical simulation overall the design process to better and quicker design (simulation driven design)

respect of design steps sequences. This situation prevents companies from turning the information and know-how embedded in their geometric and simulation models into a shared structured knowledge that can be capitalized.
Different kinds of approaches exist nowadays to facilitate this connexion between the design process and the use of numerical simulation. Overall, these approaches try to develop the integration design / simulation in a collaborative environment.

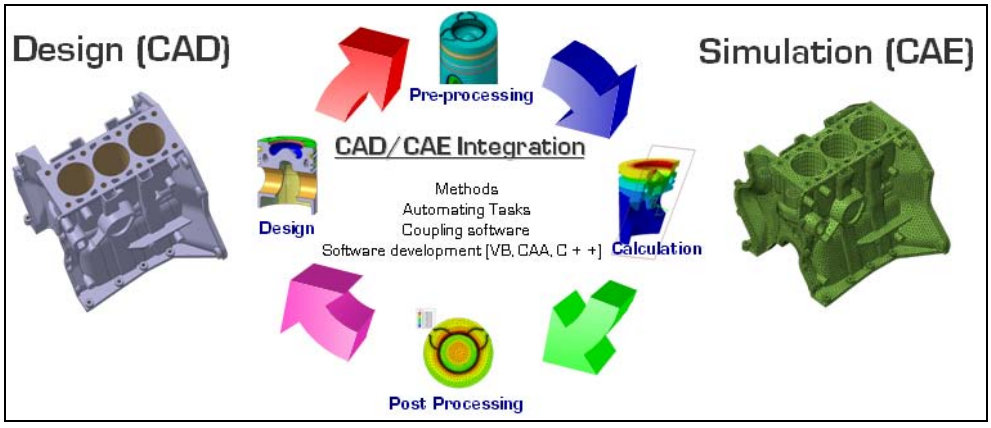


Fig. 4. CAD/CAE integration process

In this context, support tools or methods for simulation in design have been developed. The aim of this type of tools and methods is to bind design and simulation tools (coupling software in a unique environment), create a real link between the geometry of a component and the simulation context, and automate the simulation task for the designers (Fig. 4). It is

based on the methods of parametric models, idealization, meshing, and optimization. Some approaches are able to automate the transition from the geometric model (CAD File) to a numerical model by methods of automatic generation (from idealized models) or mesh discretization to generate a finite element model.
To give an industrial example, CAD/CAE integration models used in upstream design activity enable linking between the geometrical design and numerical simulation to construct "workbenches" dedicated to specific product components and physical domains. The workbench allows engineers to modify the geometry of a generic component by parametric driven design method. Then the model (idealized models used) is automatically or semi-automatically re-meshed and the calculation job is launched on a CPU. Finally, engineer retrieves the results for analyzing (Fig. 5).

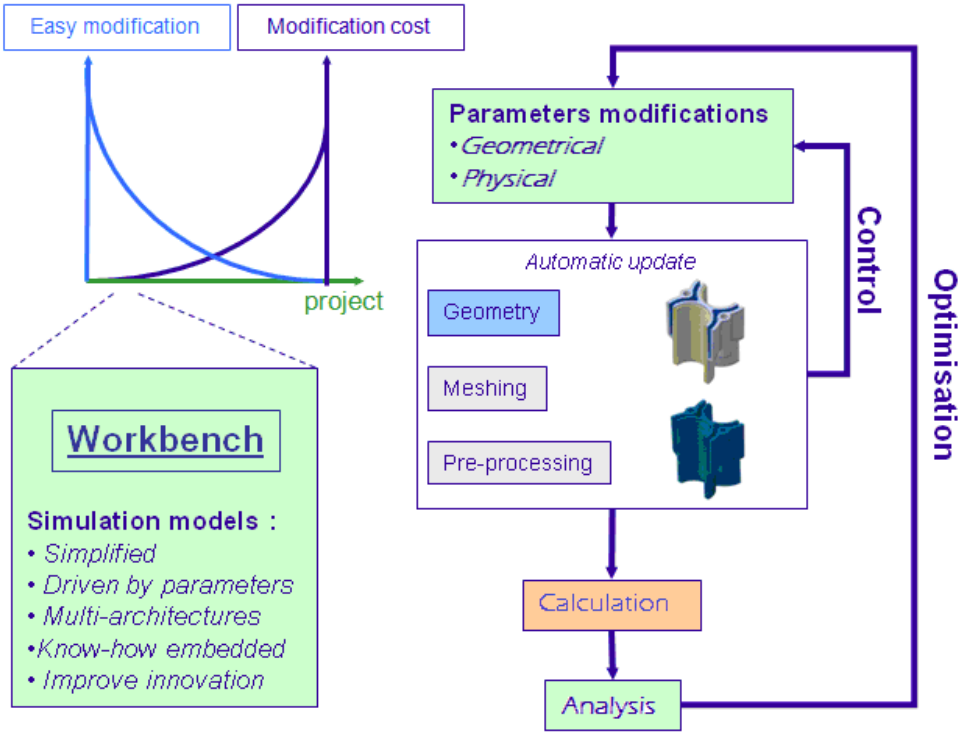


Fig. 5. Workbench process

The workbench used into several iteration loops allows for engineers to test several component architectures very quickly and identify the main design concepts with validation or not using simulation. The entire workbenches (also called expert models) are very different and heterogeneous because they are used in a large diversity of practice, with a diversity of tools, in a diversity of physical domains and moments in the design process. The expert models are based on various geometric representations with the advantage of product representations tailored to each individual situation.

Following this example, using design/simulation overall the design process allows for more flexibility and performance and show an important interest in the scientific and industrial domains.

3.1 Interest of design/simulation integration

The interests to bring closer together the design and the simulation are multiples and they can be grouped into three main parts:

- First, the collaborative work with tractability and coherence between design and simulation activities:
  Today engineers work on concurrent engineering context which mean they need to share an important volume of information in heterogeneous design and simulation activities. Each activity may takes place in different site using a large range of tools which are not able to communicate together. If design and simulation are totally independent and unsynchronised it is very difficult to take account of update in a model which impacts other models. The aim is to gather engineers on a collaborative model or a common tool which guaranteed the link between design and simulation and allowing better performances for traceability and coherence. Thus, design and simulation integration improve collaborative work in a project.
- Next, reduce routine design and better take into account of constraint from several area of expertises:
  Link design and simulation allow for better take into account of constraint from several areas of expertises. With parametric models (using the associativity), the constraints from geometric design are faster take into account in the simulation process, and reverses, which mean simulation results can impact the design and drive it. Well, it is easier to make loops between design and simulation and validate concepts by the simulation.
  For example, with classical method we use design tool for component modelling and then specifics simulation tools for meshing, pre-processing, calculation, post-processing for the first design concept tested, and then it is necessary to start again for the next design modification. It takes very long time and engineers cannot test numerous product architecture.
  With a design/simulation integration method it is possible to reduce the routine design several times (4 times and more) start for the second loop of modification (Fig. 6). This method carried out to earn in quality because engineers can test a large range of product design and identify the better, and limit the time consuming.
- Better control in the design and the simulation activities which allow to streamline the design process:
  Simulation groups several complexes activities which need experts to use specifics tools. With design/simulation integration, automatics processes used allow for designers to use the simulation with low level of knowledge in simulation. Also, it allows for capitalizing and secure know-how (simulation process, constraints, etc.) into models and thus streamlines the design process.

If design/simulation integration is now commonly used by industrials and offers significant gains in performance in the design process, some domain has particularity as crash. These particularities come from the size of the design and simulation model handled and the specific design context of crash activities.

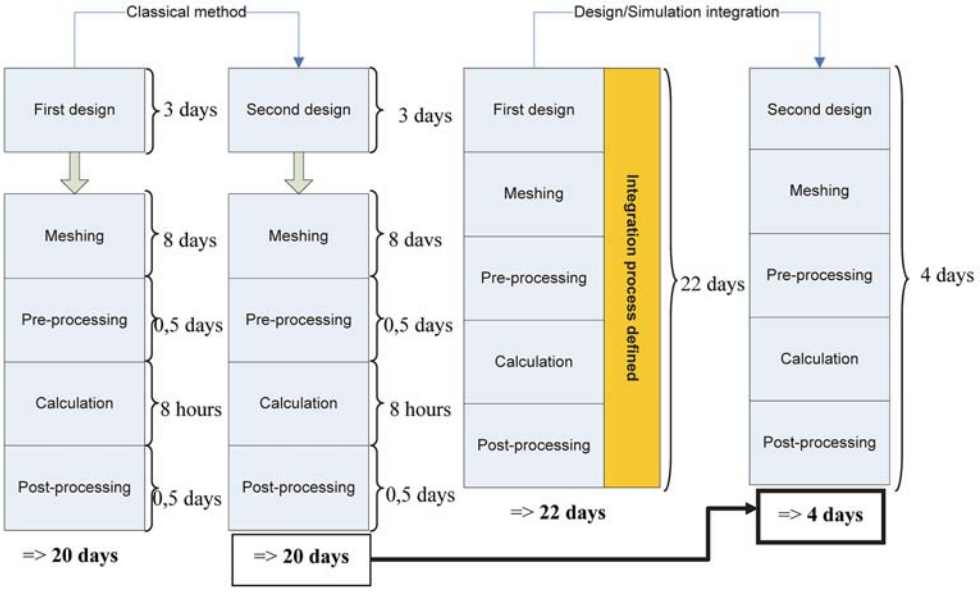


Fig. 6. Classical method vs design/simulation method

3.2 Characteristics of crashworthiness simulations

Compared to other kind of simulations, such as structural or vibration analysis, vehicle crash analysis has got some typical characteristics we may speak about.

We can start our discussion dealing with the fact that all the carmakers around the world decided some years ago to reduce their need for physical prototypes.

It's seems very difficult at that time to believe we can avoid real vehicle experiments regarding crash. But it's one of their objectives.

We'll not deal with pedestrian protection evaluation in this article; neither about biomechanical aspects in crashworthiness analysis.

All these topics have also very particular aspect.

Complex models

The first aspect that appears when reviewing FEA models for crash is the complexity of such models.

Most of the time crash models embed hundred of parts, from main body panels to small hinge. They embed visible parts, (wheels) and none visible ones, (outer CV joint). Models include heavy parts, (battery) but also light, (foam), etc.

Because crash analysis is mainly a problem of intrusion of one part in another, geometries are often modelled as close as possible of their reality.

Shapes are complex

The facts that models include a lot of parts imply that the connections between these parts must be defined.

In reality parts and components are assemble using welding, (seam welding and spot welding); using bolt and even (more and more) glue.

Thus, in addition to include hundred of parts, the FEA model for crash will also need thousand of connections definitions, thousand of connection properties definitions, (fracture limits, etc.).
Designing vehicle body does not really depend on routine design but there is a strong impact on several other parts of the cars as the power unit, the cockpit, the frame, etc.

Huge problems

Hundred of parts, thousand of connections, millions of nodes: huge problem in terms of DOF.
Because of the size of the problem and because of the transient aspect of the simulation, the computation phase of a crashworthiness analysis can last several days.
Each year, even if the power of computer increases the duration of a typical crash computation remains the same.
Engineers are not yet in the process of stabilizing their models. They enrich them with more and more detailed parts, with finer meshes, with more precise contact management, etc. This way the computer power is harnessed to serve the quality of results instead of reducing computing time.
The big size of crash simulation models unfortunately also deals with the difficulties to manipulate these complex models during the pre-processing and post processing phases. These phases are also very demanding in terms of PC power regarding crash simulations.

Marketing aspects

Whereas some years ago the car-style was so important for final customers, the crashworthiness aspect appears more and more as a key point in the choice for a new vehicle.
Nowadays it is not uncommon to have some information regarding the Euro NCAP results of a new car directly in the advertisements for this new car; nor to see crash test dummy "playing" in such advertising.
Priority between Style and Crashworthiness has changes these few last years.
Thus crash as a strong impact on product design cost and that is the reason why industrials show an important interest and make research or developments. We propose to see some of them in the next section.

4. Several industrial methods for CAD/ CEA integration

4.1 Approaches with strong links to CAD

It's now possible to deploy some approaches based on strong links between geometries and FEA components.
Unfortunately in these cases body shapes have to be simplified. Most of the time these kind of approaches are used at the very early stage of a new project. This lat point made these approaches very interesting.
Thanks to the strong link between CAD geometries and FEA, and because geometries are simplified, meshing operations can be performed using automatic mesher.
The link between CAD & FEA, the high level of automation makes short loops iteration possible.

4.1.1 AVP - an example based on skeleton and simplified geometry

AVP is a set of methodologies and CATIA V5 workbenches developed for a French car maker. It offers a team of engineers with expertise in CAD & Analysis the possibility to quickly model a new vehicle.

Within four weeks, (Fig. 7), a new body style can be defined. The whole structure product is divided into hollows parts, junctions and panels. Generally the same platform can be reused from a project to another.
A skeleton controls links between parts, consisting in strong parametric geometry.
For the most, parts are represented by multi-sections. Each sections consisting in a five segments polyline.
The high level of simplification guarantees the whole body automatic update process on design change. It made also possible the automatic quadrangle meshing of the body. The process complies the organisation's meshing standards.
Specific CATIA V5 workbenches have been developed in order to pre-process a crash analysis case within the software.
User can model the specific connexions that are validated within his organisation. He can define all type of features needed for crash analysis, (sections, accelerometers, contact interfaces, etc.)
Finally, the high level of automation and the complete integration of these methodologies and tools within a unique software interface allow short loop iterations.

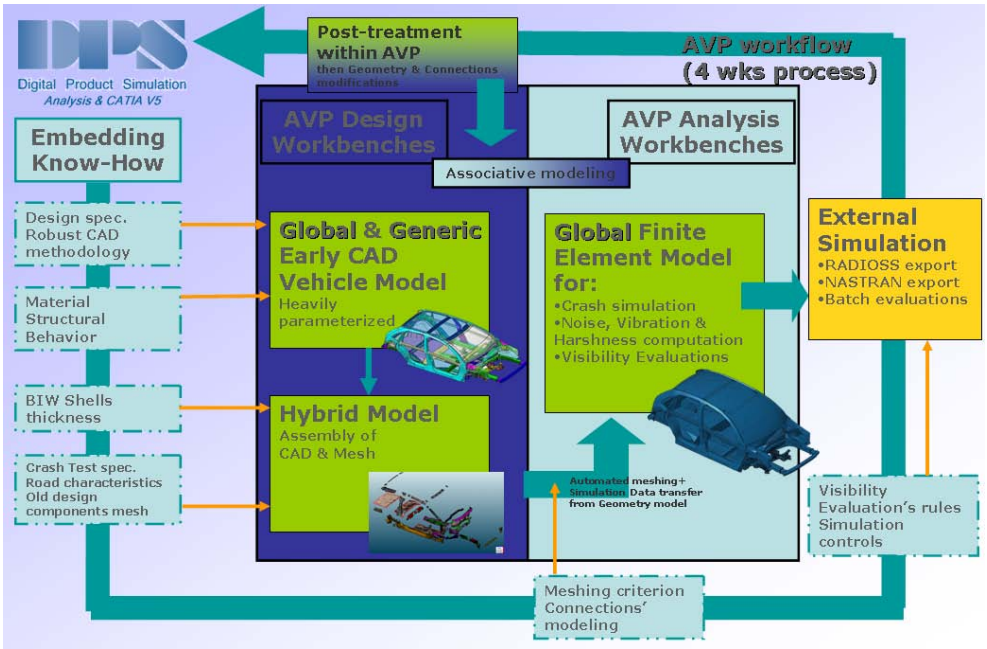


Fig. 7. AVP Workflow

This kind of approach is very efficient during early design phases. It proposes an agile geometry, able to represent several architectures.
But when the car concept becomes mature, engineers need to design more and more precisely. Although geometries are parametric, they cannot evolve to more detailed geometries.
That is the big limitation of the approach.

4.1.2 Fast Concept Modeller - an example based on productive Design tools

FCM is a set of tools aiming at helping designers to create very quickly and easily vehicle geometries.
Based on very productive tools, FCM allows users to model a vehicle manipulating geometry objects directly on the screen.
Fast Concept Modeller is a single CATIA V5 workbench. The user interface favours "free hand" actions.
The geometries results, as they were with the AVP approach, are parametric and very simplified.
During the new project vehicle geometries can become more and more detailed. They can evolve from a beam model to a more complex beam-shell model including fillets, multi-flanges, etc.
Regarding the FEM functions included in the software, shell can be mesh using batch meshing technique (ANSA). In that case properties and connexions attributes defined on geometry are directly transferred on finite element model.
For the early stage of the process Beams are used. In that case, the car geometry is automatically discretized using variable cross section beams. This process is very powerful if optimization loops are engage on the beam structure. FCM can pilot the vehicle geometry from the result of such an optimisation.

4.1.3 Approach using software of the shelf

It exists powerful pre-processing software fully integrated into CATIA V5 and allowing expert simulations set-up (Fig. **8**).
These software lies on the CATIA V5 philosophy, (all the model features have geometry support) but also extend the natural capability of CATIA V5 providing the user with direct access to nodes & elements.
Such as FCM, these kind of software offer batch meshing capabilities. This possibility bridges the gap in the CATIA mesher.
In this way geometry model can be much more detailed. In another hand, the possibility to deal directly with nodes & elements entities brings user the change to handle orphan mesh. Meshes perform with more dedicated software or meshes of a previous project can be easily used.

5. Limitations and opening

5.1 Detailed geometries

As we mentioned above, the approaches based on a strong link between FEM and geometry imply - most of the time - a poor level of detail in geometry: the simplest geometry is, the more automated the update on changes will be.
The gap between simple and detailed geometries is not easy to fill.
Even if during the early design phases geometries have to be very simple in order to be able to evaluate a lot of architectures and alternatives, while the project run engineers need to study the influence of small modifications. Teams quickly have to integrate manufacturing process parameters.
Unfortunately it not easy, (not possible) to use these very simplified model for the next steps.

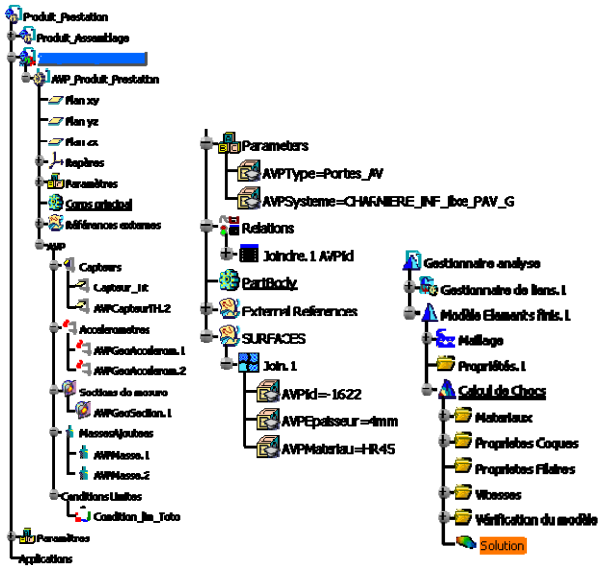


Fig. 8. Specific Crash Analysis Features within CATIA V5

Moreover, because CAD/ CAE integrated approaches embed geometries, meshes and analysis features, the associated numerical models are quite big and require the use of powerful workstation.

5.2 Integrated approaches = CAD + CAE

For an organisation, saying the same team will handle geometry & FEM models is a big challenge. It means FEA engineers have to be trained in CAD software, (more rarely, Designers are trained in simulation).
The FEA engineer job is changing slowly...
Double competency, Design + Simulation, will be on tomorrow a must have for young engineers.

5.3 Simulation life cycle management

With the natural trend bringing closer CAD and FEA, some techniques now enter in the simulation field.
Among them SLM, Simulation Life Cycle Management, is surely something which will become more and more important.
Designers and Simulation engineers are now working together. They need to share the same data. They will need some specific tools to do that more easily.
We already hear many testimonies speaking about the fact that integrated CAD/ CAE approaches urge team to ensure a better data traceability.

5.4 Optimisation and more

Probably the main advantage of a CAD/ CAE integration is the fact that organisation can perform short loop of iteration.

On each change the remaining work is automatically update then performed.
Optimisation is possible and even effective for more and more complex cases.
The next challenge will be the coupling between several types of simulations.
Being able to take into account the forging or stamping process of each part during the crash worthiness simulation is a big challenge, but will ensure an important level of accuracy.
Performing optimisation loops including crash and stamping simulation is yet unrealistic, but...

5.5 Knowledge management for design and simulation

We have seen design/simulation integrations method focused on models closer, but problems still exists about knowledge embedded in models. Indeed, each expert model manages parameters and rules independently from other model which uses the same knowledge. This knowledge is often duplicated and dependant of the models which using it. This situation favours knowledge inconsistency between models and it often happens that simulations are launched on different models sharing same parameter but on wrong values.
Indeed it is very difficult to make expert models communicate together because they are used with several tools which are not able to communicate together despite CAD/CAE integration method. It appears that is their no communication platform for this type of knowledge and today with the massive using of design and simulation models it is a real problematic.
Nowadays, researches are focused on this problematic in accordance with global PLM (Product Life Cycle) approach. The aim is to define a method and a model or meta-model (in UML[2] or MOF[3] which are modelling standards defined by OMG[4]) allowing to manage knowledge and share it through experts models with coherence. Some of research work proposes to capitalize parameters and rules extracted forms design and simulation models into generic information baseline and to built knowledge configuration synchronized with experts models. We propose to explain one of these researches called KCModel.

5.6 Perspectives with KCModel (Knowledge Configuration Model)

5.6.1 KCModel objectives

The aim of this research is to propose a new tool which helps users to ensure data, information, and knowledge consistency when shared in several and heterogeneous experts CAD and CAE models. This tool will focus on a new generic approach called KCModel: "Knowledge Configuration Model" based on knowledge configurations synchronized with expert models.
KCModel is formalized into meta-models in UML Language. In the context of KCModel, we consider as:

- technical data, the parameters and expert rules extracted from experts models,
- information, the data capitalized on, structured and organized into a specific entity to construct a technical and generic product information baseline,

[2] UML: Unified Modelling Language is a language used to formalise model object oriented. UML is defined by OMG.
[3] MOF: Meta Object facility is a language used to formalise meta-models object oriented. MOF is defined by OMG.
[4] OMG : Object Management Group – www.omg.org

- knowledge, a set of technical product information entities instantiated from the baseline in a configuration used in specific design or simulation activity. This configuration is synchronized with a specific CAD or CAE model.

The purpose of the KCModel is to **C**apitalize, **T**race, **R**e-use, and ensure the **C**onsistency **(CTRC)** of technical data shared by several experts model, especially in the upstream step of design process (Fig. 9):

1. Capitalize on parameter and rules as a generic and cross functional baseline.
2. Share and trace through several users.
3. Re-use parameters and rules in expert models.
4. Ensure the consistency and save the modifications.

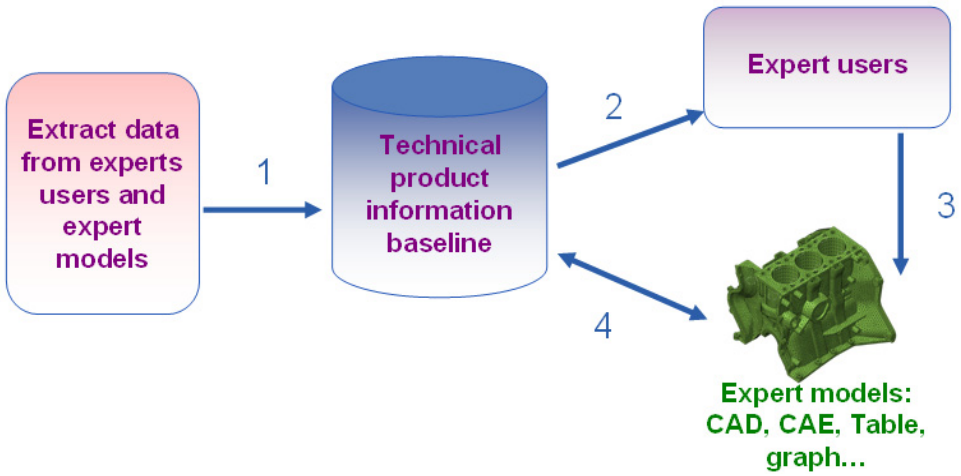


Fig. 9. Global KCModel method

We propose now to explain the KCModel method and then to focus on the knowledge configurations and how the consistency can ensured.

5.6.2 Global KCModel method

The KCModel (Fig. 10) allows for capitalization of technical data extracted from different expert models, into an abstract generic information entity called "Information Core Entity" (ICE is the smaller information entity used). Data capitalized on, structured, organized and documented in these entities is then considered as technical information and all ICE centralized in a single point in a generic and a cross-functional baseline. To be used in a specific context (e.g. thermal load case on a piston for a milestone X in a project), we create a "Configuration Entity" (CE) instantiating ICE corresponding with the context of use. The configurations are then synchronized with the different expert models and managed in a consistent way. Each configuration is a representation of knowledge embedded in expert models. Configurations are compared between them to warn conflict to engineers.

This approach lets to manage the technical product information and its instances (set of parameters of values) by using configurations and versions. Explicit knowledge is handled in these models. Indeed, data is capitalized on ICE to become information. Information is transformed into knowledge when an individual understands its necessity to an activity,

which means by creating knowledge configurations with ICE instantiate needed to a specific design or simulation activity.

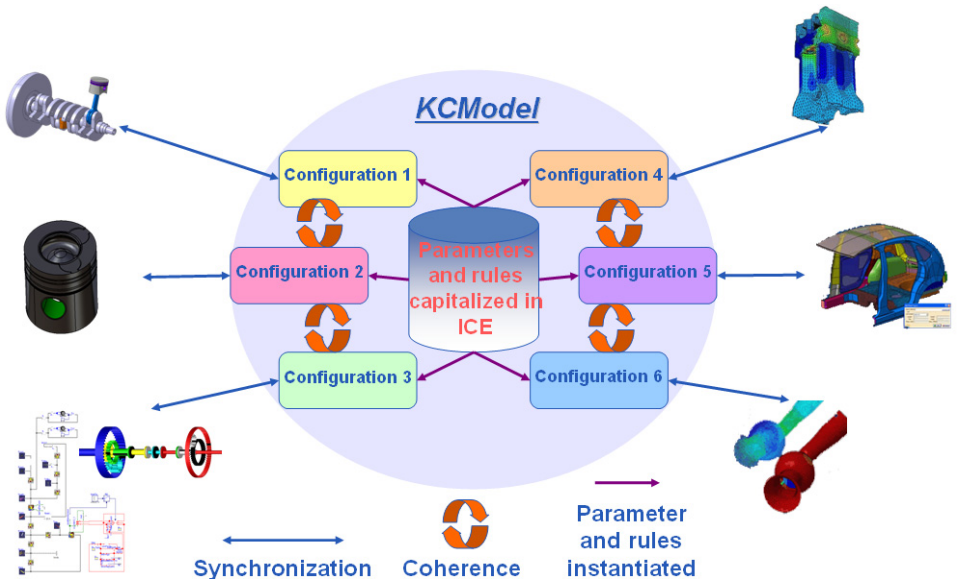


Fig. 10. Using knowledge configuration in KCModel to ensure the consistency between expert models

This type of approach is highly compliant with design/simulation method and allows for reaching a high level of collaboration and performance in the design process.

6. Specificities of FE codes for impact simulation

6.1 Introduction

Finite element codes are based on the spatial discretization of a continuous field, and consist in solving differential equations system. Engineers have to go through several steps, from the choice of the dimension of the model, to the constitutive laws or the choice of element formulations.

First of all, this complex task goes through the choice of the dimension of the model to be solved. Indeed, numerical engineers have to analyse the type of analysis to be solved in order to optimize the resolution of the physical phenomenon, and to face to the choice of the geometry and the physical model: can this phenomenon be modelled in a single dimension, or does it need 3 dimensions? The basic example of the flexion of a beam, which could be modelled with a beam, with shells or with solids elements can illustrate this first point.

This chapter will only deal with the finite element method, which is the most used method in the field of crash analysis. For dynamic simulations, such as crash simulations, the time dependence is added to the complex solver process. The problem to be raised is to have an optimized time discretization using either implicit or explicit scheme, with the introduction of a complex concept in dynamic numerical methods, the time step. Depending on the choice of the spatial discretization, on the choice of the material properties, this time step

plays an important role in the simulation. Finally, dynamic finite element codes are complex codes and its specificities will be explained in the next sections.

6.2 Equation of the motion - dynamic formulation

A nonlinear finite element equation of motion is usually obtained from the principle of virtual work. This is the weak form for equilibrium equations which includes internal force, contact/friction force, inertia force, damping force, external force, and boundary condition. Finite element method (FEM) discretization of the equations of motion leads to the following matrix form of coupled set of second-order nonlinear deferential equations:

$$[M].\left(\ddot{X}\right)+[K](X)=F_{ext}$$

where (X) is the vector of the nodal positions at current time and $\left(\ddot{X}\right)$ the vector of nodal accelerations. $[M]$ is the mass matrix, $[K]$ the stiffness matrix and F_{ext} the vector of external forces. This equation is non-linear (in (X) and $\left(\ddot{X}\right)$) due to the presence of contact, an possible material and geometrical nonlinearities. A time integration scheme must be chosen, it must be able to struggle with this strong non linear problem.

6.3 Basic contact notions

The consideration of contact boundary conditions in the finite element simulation of interacting components is nowadays established as state of the art. According to the principles of continuum mechanics, the contact conditions can be expressed as follows. Let us consider two deformable bodies Ω^1 and Ω^2 in potential contact and two potential contact surfaces are noted Γ^1 and Γ^2. Let $x=\varphi(\mathrm{X},t)$ be the current position vector at an instant $t\in \mathrm{I}_t$. The orthogonal projection of x on the body surface Γ^2 is defined by x^2. The contact distance vector (or gap vector) is defined by

$$g=x-x^2=g_n\,\nu$$

g_n is the oriented contact distance.

Let t be the contact stress vector exerted by Γ^2 on the body Ω^1. Next, the displacement vector u, the velocity vector $\dot{u}$ and the contact stress vector t can be uniquely decomposed into a normal part and a tangential part as follows:

$$\mathrm{u}=\mathrm{u}_t+u_n\nu,\quad u_n=u\cdot\nu,$$

$$\dot{\mathrm{u}}=\dot{\mathrm{u}}_t+\dot{u}_n\nu,\quad \dot{u}_n=u\cdot\nu,$$

$$\mathrm{t}=\mathrm{t}_t+p_n\nu,\quad p_n=\mathrm{t}\cdot\nu.$$

The unilateral contact law is characterized by a geometric condition of non-penetration, a static condition of no-adhesion and a mechanical complementary condition. These three conditions, known as the Signorini conditions, can be formulated as

$$g_n \geq 0, \quad p_n \geq 0, \quad g_n p_n = 0.$$

In the case of dynamic contact, the Signorini conditions can be formulated, on Γ^1, via the relative velocity

$$\dot{u}_n \geq 0, \quad p_n \geq 0, \quad \dot{u}_n p_n = 0 \quad .$$

The bodies are separating when $\dot{u}_n > 0$ and remain in contact for $\dot{u}_n = 0$. The formulation of the Signorini conditions can be combined with the sliding rule to derive the complete frictional contact law for the contacting part. Γ^1. This complete law specifies possible velocities of bodies that satisfy the unilateral contact conditions and the sliding rule.
A key issue in the treatment of contact constraints in explicit dynamics is the choice of contact constraints to enforce at contacting nodes. The contact constraints evaluation has significant effect on the accuracy and efficiency of the analysis. A variety of numerical methods have been proposed in the literature to deal with this problem: lagrange multiplier methods, penalty (Chamoret, et al., 2004), augmented Lagrangian approach (Alart, et al., 1991) and bipotential method (Feng, et al., 2006) are the most frequently used. A good evaluation of this force requires in a first step locating all the potential contacting nodes. In an industrial context, where the number of contact nodes is important, it appears essential to develop contact searching algorithms (Zhong, et al., 1994), (Weyler, et al., 2011). A good procedure should be accurate to detect all the potential contact nodes and efficient to avoid the unnecessary research (and so an increased of the computation time).

6.4 Time integration scheme

For the simulation of dynamic problems such as crash analysis, the time discretisation is one of the major point that can strongly influence the accuracy and efficiency of the algorithm. The two main solution procedures are the explicit and implicit algorithms. The implicit scheme is unconditionally stable. But it has two main drawbacks: the first one is that a linear set of equations must be solved repeatedly so the computation time increases with the size of the model when using a direct solver. The second one concerns convergence which is sometimes hard to reach. In general finite element code dedicated to the simulation of transient dynamic phenomena such as crash or impact (e.g. Radioss, Altair Hyperworks, Michigan, USA), the temporal explicit scheme is used. Explicit numerical time schemes such as the well-known central difference scheme have been widely used as they do not require numerical iterations at each time step, and also for their good properties in term of accuracy and robustness with possible nonlinearities.
The state of the system is evaluated at each time step. The state at a given time t, is used to calculate the state at the $t + \Delta t$, where Δt is representing the time step. Furthermore, the inertia and mass of the system is taken into account. The explicit scheme is a specific method where the equilibrium state is evaluated at a time where displacements are already known at each point of the mesh.

$$[M].\left(\ddot{X}_n\right) = \left(F_{ext}(t_n)\right) - [K]\left(X_n\right) = \left(F_{ext}(t_n) - F_{\text{int}}(t_n)\right)$$

In this process, displacements are known at the time where the dynamic equilibrium of the system is solved, and needs only the inversion of the mass matrix. Furthermore, if a lumped

mass matrix scheme is used, the mass matrix is diagonal and does not need inversion. The resolution of the system is very quick since each degree of freedom is calculated separately. Each stress are evaluated in each elements individually. At each time step, the state of equilibrium is updated, which corresponds to the propagation of a wave into the element. This important point lead to the conditional stability of the scheme, which means the existence of a critical time step for the stability of the resolution.

7. Specific problems to explicit scheme

For high speed simulations, temporal discretization can be performed by the central difference methods (CDM). In such explicit time integration method, specific conditions on the maximal time step for numerical stability are assumed. The maximum time step $\Delta t_{\max}$ to be used is determined by the Courant number C (Courant, et al., 1967):

$$C = \frac{c.\Delta t_{\max}}{\Delta x} \leq 1$$

This requirement means that, during one time step, the distance travelled by the fastest wave in the model ($c.\Delta t_{\max}$) should be smaller than the smallest characteristic element size (Δx) in the mesh, representing the shortest length for a wave arriving on a node to cross the element. c represents the wave velocity for a given material or the time step for the analysis should be smaller than the time a wave needs to cross the element.
With elements of 5 mm, and for a typical steel material law, this condition leads to an order of magnitude of 10^{-3} ms. Indeed with this order of magnitude of the time step, it appears that this specific scheme is an appropriate method to solve very rapid phenomenon, with high velocity leading to non highly non linear problems. For typical impact duration of 100-200 ms, it appears necessary to use this kind of integration scheme for an accuracy of the results.
This time step also depends on the number and the type of elements used to model the system. In automotive industry, FE models are developed using 4 nodes shell elements (Belytschko, et al., 1981). The following picture illustrates a FE model of a window, developed with shell elements.

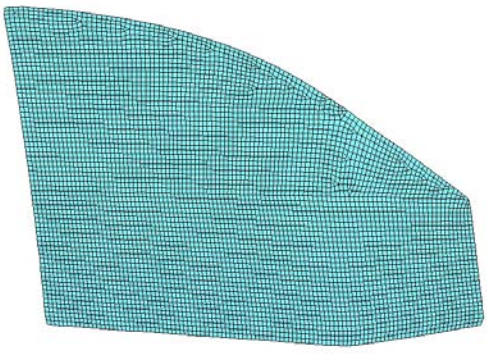


Fig. 11. Mesh of a window for FE simulation

With specific geometrical hypothesis (such as the ratio length / thickness), the use of these elements are powerful when modelling shell structures such as windshield, bonnet, doors in automotive engineering.
The use of such elements are used in the case study of the SIA vehicle of the University of Technology of Belfort-Montbeliard, which is submitted to impact.

8. A numerical study - UTBM's vehicle for SIA Trophy - Esphyra

All the theoretical concepts previously defined lead to a case study, which is explained in order to illustrate the feasibility of FE simulations in the crash field.
The SIA trophy is an automotive challenge for automotive designers, manufacturers, universities, whose aim is to build and design a vehicle able to face to today's new specifications in terms of innovations, respect of the environment. From a numerical point of view, standard procedures have been used in order to perform crash FE simulations.
The mesh have been performed based on the CAD geometry models (Figure 12 and 13).

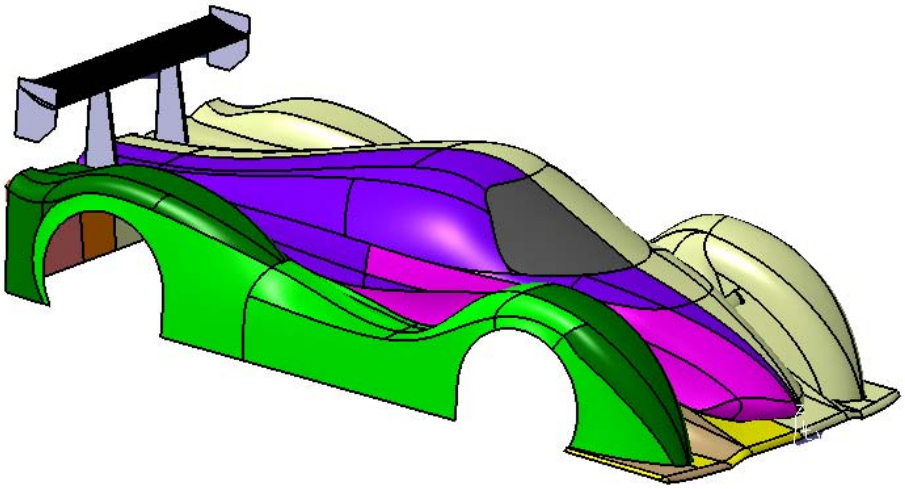


Fig. 12. CAD model of the E-Sphyra's bodywork

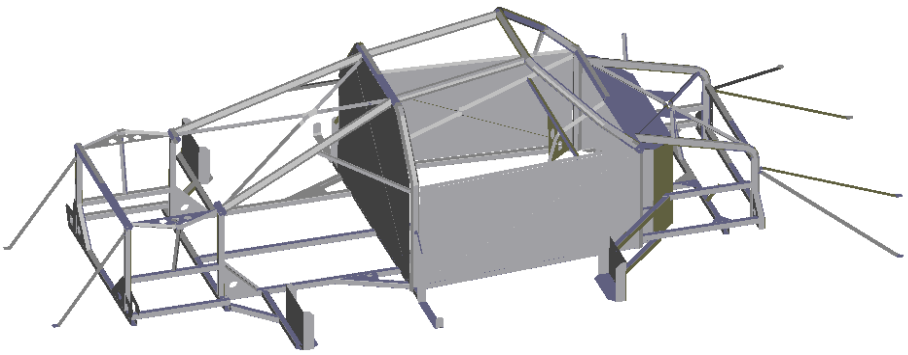


Fig. 13. CAD model of the E-Sphyra's frame

The bodywork and the frame of the vehicle have been modelled with solid parts into a CAD software, in taking into account their thickness. Mid-surfaces of the bodywork and of the frame have been extracted in order to mesh them with 4 nodes shell elements as recommended in section 7, and illustrated in figure 14.

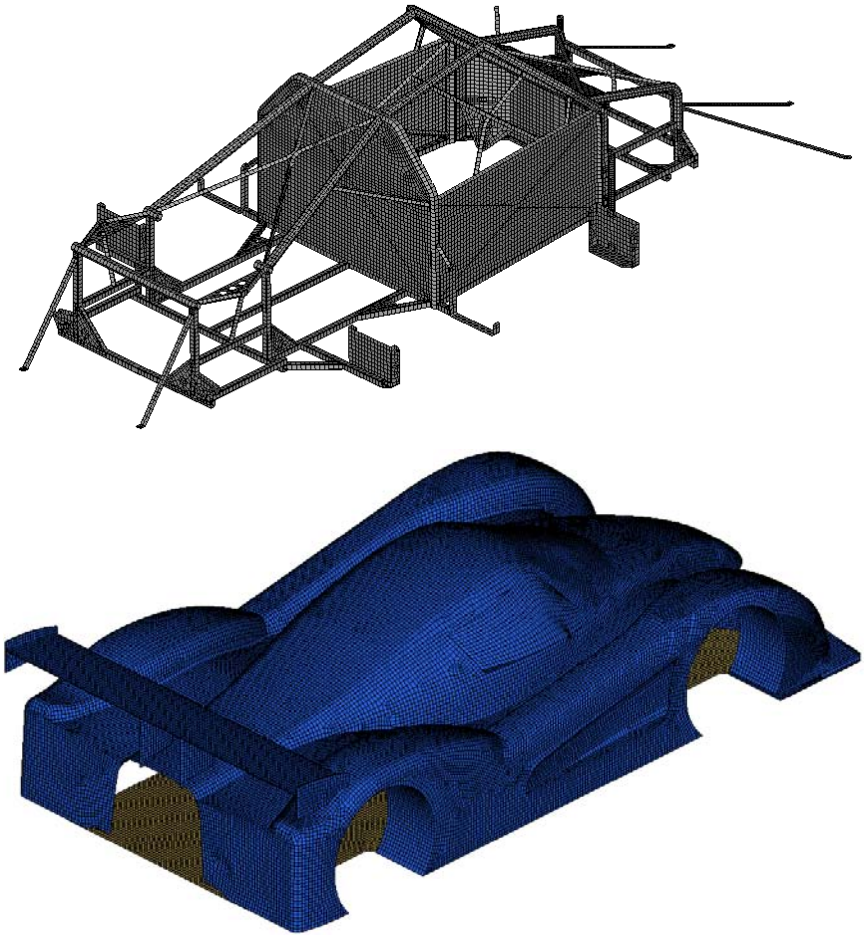


Fig. 14. Mesh of the E-Sphyra's frame and bodywork

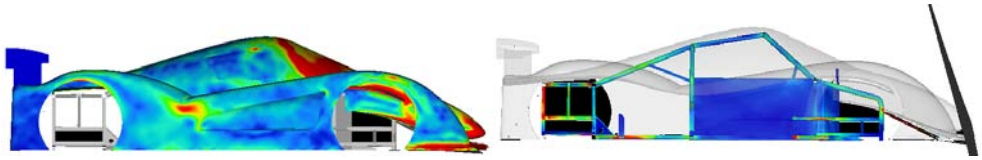


Fig. 15. Von Mises stress distribution inf the frame and in the bodywork

With steel material laws, whose parameters are listed in the following table, the structure was submitted to an impact with an initial velocity of 10 m/s against a deformable plane.
Results in terms of Von Mises stress distribution in the frame and in the bodywork are illustrated in figure 15.
The simulation of this impact used the theory described in previous section in terms of integration scheme used, in terms of time step and mesh, in finally in terms of contact.

Mechanical parameter	target
Density	$7.9.10^{-3}$ kg/m^3
Young's Modulus	210000 MPa
Poisson ratio	0.3
Sigma yield	210 MPa
Sigma max	240 MPa

8.1 Conclusion and opening of crash simulations in the field of biomechanical simulations

Specific approach for FE simulations in crash fields are necessary for optimal calculations. Indeed, several points are in favour of explicit scheme in such simulations.
Despite of the conditional stability of the sheme and the little time step, there are lots of advantages using explicite scheme in the crash field, like its precision, its easy capability of the mass matrix inversion (if lumped mass matrix is used), quite low CPU cost etc…The more the velocity is high, the more the explicite scheme is adapted for the simulations.
These kind of numerical algorithms has prove itself, in terms of robustness and accuracy of the results. Since the beginning of the 70's and the need of the investigation of "what happens" during a vehicle impact, these methods have not stop to improve. With the evolution of computer science, FE simulations are more and more used to investigate physical problems. It allowed creating complex models with a large number of elements. With 10 000 elements at the end of the 80's, typical FE models of vehicle can reach today to several hundreds of thousands of elements. Indeed, precise FE models of vehicles are developed allowing a precise investigations of what happens during an impact, with the aim of an optimization of the structures and an improvement of the safety of the vehicles. This concept of safety become more and more important, and recent finite element simulations couple the finite element model of the vehicle with a FE model of a human. That is the concept of numerical impact biomechanics.
Automotive engineering and biomechanical engineering can gather their knowledge to improve the security of the vehicle occupants. At a numerical level, lots of studies have dealt with the development of finite element models of human structures: head (Roth, et al., 2010), or shoulders (Duprey, et al., 2005) for example which can be coupled to human environment in order to improve its security. In developing "biofidelic models" research can be lead on injury mechanisms (Roth, et al., 2009) which can help to evaluate the dangerous behaviour of a structure (Meyer, et al., 2009).
Finally, simulations have helped engineers to develop powerful models leading to an improvement of the life cycle for the design of a mechanical system. However, in a context of "world development" and "collaborative engineering", it is necessary to have a specific design methodology in order to optimize the design process. Numerical simulation being a part of the design process, it is necessary to involve the simulation data in a PDM : the development of Simulation Data Management is today's compulsory.

9. Conclusion

Finally, these last decades have shown the development of numerical simulation which became essential in the design process, especially in automotive engineering. In crash field the requirements and the standards have increase compared to the last decades, and lead to a number of tests which are now compulsory for the probate and the industrialization of a vehicle. Furthermore, the coupling of vehicle FE models with human body FE models for an improvement of the safety of a vehicle allowed the development of numerical biomechanics. Human body is now investigated at a numerical level, allowing optimizing the design of vehicle or protecting devices.
Furthermore, the numerical simulation makes the lifecycle decrease with an optimized and an performing management of simulation data. However limits in terms of interaction between CAD system and FE platform still exist.

10. Acknowledgment

This chapter has been writing with the collaboration of DPS – Digital Product Simulation, which is an expert in design and simulation integration for product development.
Authors would like to thanks DPS (Digital Product Simulation) for their expertise, their help and their contribution to the writing of this chapter.

11. Reference

Alart, P & Curnier, A 1991, 'A mixed formulation for frictional contact problems prone to Newton like solution methods', *Computer Methods in Applied Mechanics and Engineering*, vol 92, no. 3, pp. 353-375.

Awrejcewicz J., Krysko V. A., Nonclassical Thermoelastic Problems in Nonlinear Dynamics of Shells, Springer, Berlin, 2003

Awrejcewicz J., Krysko V. A., Vakakis A. F, Nonlinear Dynamics of Continuous Elastic Systems, Springer, Berlin, 2004

Belytschko, T & Tsay, CS 1981, 'Explicit algorithms for nonlinear dynamics of shells', *AMD-Vol. 48, ASME*.

Chamoret, D, Saillard, P, Rassineux, A & Bergheau, JM 2004, 'New smoothing procedures in contact mechanics', *Journal of Computational and Applied Mathematics*, vol 168, no. 1-2, pp. 107-116, Selected Papers from the Second International Conference on Advanced Computational Methods in Engineering (ACOMEN 2002).

Courant, R, Friedrichs, K & Lewy, H 1967, 'On the partial difference equations of mathematical physics', *IBM J. Res. Dev.*, vol 11, pp. 215-234.

Duprey, S, Bruyere, K & Verriest, J-P 2005, 'Numerical Simulation of Shoulder Response to a Lateral Impact with the HUMOS Model', in MD Gilchrist (ed.), *IUTAM Symposium on Impact Biomechanics: From Fundamental Insights to Applications*, Springer Netherlands.

Feng, Z-Q, Peyraut, F & He, Q-C 2006, 'Finite deformations of Ogden's materials under impact loading', *International Journal of Non-Linear Mechanics*, vol 41, no. 4, pp. 575-585.

Hamri, O, L\', Giannini, F & Falcidieno, B 2010, 'Software environment for CAD/CAE integration', *Adv. Eng. Softw.*, vol 41, pp. 1211-1222.
http://e-sphyra.sup.fr/.

Meyer F, Willinger R, 2009, 'Three Years Old Child Head-Neck Finite Element Modeling. Simulation Of The Interaction With Airbag In Frontal And Side Impact' *International journal of vehicle safety,* vol 4 (4) pp 285-299.

Robinson, TT, Armstrong, CG & Fairey, R 2011, 'Automated mixed dimensional modelling from 2D and 3D CAD models', *Finite Element in Analysis and Design*, vol 47, pp. 151-165.

Roth, S, Raul, J-S & Willinger, R 2010, 'Finite elements modelling of paediatric head impact. Global validation against experimental data.', *Computer Methods and Programs in Biomedicine*, vol 99, pp. 25-33.

Roth, S, Vappou, J, Raul, J-S & Willinger, R 2009, 'Child head injury criteria investigation through numerical simulation of real world trauma', , *Computer Methods and Programs in Biomedicine*, vol 93, pp. 32-45.

Shephard, MS, Beall, MW, O'Bara, RM & Webster, BE 2004, 'Toward simulation-based design', *Finite Element in. Anaysis andDesign*, vol 40, pp. 1575-1598.

Weyler, R, Oliver, J, Sain, T & Cante, JC 2011, 'On the contact domain method: A comparison of penalty and Lagrange multiplier implementations', *Computer Methods in Applied Mechanics and Engineering*, In Press, doi:10.1016/j.cma.2011.01.011

Whitworth, HA, Bendidi, R, Marzougui, D & Reiss, R 2004, 'Finite element modeling of the crash performance of roadside barriers', *International Journal of Crashworthiness*, vol 9, pp. 35-43, 10.1533/ijcr.2004.0270.

Zhong, Z-H & Nilsson, L 1994, 'Automatic contact searching algorithm for dynamic finite element analysis', *Computers \& Structures*, vol 52, no. 2, pp. 187-197.

26

Translational and Rotational Motion Control Considering Width for Autonomous Mobile Robots Using Fuzzy Inference

Takafumi Suzuki and Masaki Takahashi
Keio University
Japan

1. Introduction

Obstacle avoidance methods for mobile robots have proposed in a broad range of studies and the availabilities have been discussed. Most of these studies regard the robots as points or circles and control methods of the translational movements are discussed. In these studies, it is pointed out that a non-circle robot can be transformed into a point robot by expanding the obstacles by the largest radius or maximum size of the robot. The effectiveness of avoiding obstacles by these approaches have been confirmed, however, according to the shape of the robot, these approaches reduce and waste the available free-space and can decrease the likelihood of getting to the goal. If wide-robots, which are horizontally long, are regarded as circles according as conventional approaches, they have possibilities not to go through between two divided objects due to the largest radius of the robot, even if they ought to be able to go through by using their shortest radius. This suggests necessity of suitable orientation angle at the moment of avoidance. Consequently, to enable wide-robots to avoid obstacles safely and efficiently, it is necessary to control not only the translational movement but also the rotational movement. In our current research, wide-robots with omni-directional platform have been employed, as shown in Fig.1. In situations like Fig.1, both wide-robots can go through only by changing the orientation angle in real time.

Some researches focus attention on the orientation angle of the robot (Kavraki, 1995)(Wang & Chirikjian, 2000). In these studies, by convolving the robot and the obstacle at every orientation and constructing the C-space, the suitable orientation angles of the robot for path planning are decided. However, these methods need environmental map and do not show the effectiveness for autonomous mobile robots about avoidance of unknown obstacles in these studies. Therefore, in order to avoid unknown obstacles reactively considering the orientation angle, the wide-robot needs an algorithm that can decide the orientation angle and rotational velocity command on the spot based on the current obstacle information.

Meanwhile, decision methods of the translational movement have been proposed in many studies (Wang et al., 2000) (Du et al. 2007) (Khatib, 1986) (Borenstein & Koren, 1991) (Dieter, 1997), we employ fuzzy potential method (FPM) (Tsuzaki & Yoshida, 2003). This method realizes some tasks in dynamic environment by fuzzy calculation about desire for each direction of the robot. In this research, it was shown that wheeled robots succeeded getting to the goal with conveying a soccer ball and avoiding obstacles.

Fig. 1. A situation where two robots move on narrow corridor

In this paper, a control method using a capsule-shaped case is described for both translational and rotational movement based on the FPM, and it takes into consideration the width of the robot. With this new approach, real-time control of the orientation angle is easily achieved. Conventional FPM has only been able to deal with translational velocity. The proposed method is able to control the rotational and translational velocity simultaneously within the framework of FPM.

2. Capsule case

2.1 Need to consider the width of mobile robot

In recent years, non-circle robots have been developed, of which vertically long robots, wide-robots, appear. In studies of humanoid robots, the robots have two arms mounted to the stationary torso with wheels because these robots can be used in terms of mobility, manipulation, whole-body activities, and human-robot interaction (Ambrose et al., 2004) (Du et al., 2007). During last two decades, vast number of algorithms of obstacle avoidance for mobile robot, and recently some researches and developments of the mobile robot in practical use have been reported. These robots have problems that conventional methods are inconvenient for applying for the wide-robot because most of conventional methods of obstacle avoidance regard the robot as points or circles. Or due to the postulate of the conventional methods, the robots have needed to be designed in a circle. We also have developed the wide -robot, which has torso with two arms and a head, for making the robot perform not only moving but also communication with human by means of gestures or speeches based on a perspective of human interaction. This robot is also horizontally long. In addition, when the robot opens an arm slightly, as shown in Fig. 1, or both arms, it becomes increasingly harder to apply conventional methods. If these wide-robots are regarded as circles according as conventional approaches, they have possibilities not to go through between two divided objects due to the largest radius of the robot, even if they ought to be able to go through by using their shortest radius. In this study, enable the wide-robots, which move automatically, to move smoothly and safely in the environment with obstacles, a capsule case is introduced.

2.2 Design of capsule case

The capsule shaped case is modeled by two circles and two lines tangent to the circles as shown in Fig.2.

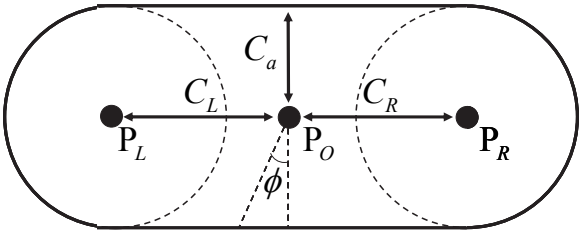


Fig. 2. A capsule case

This closed contour is defined as $l(\phi)$ with the origin at the point P_O and as follows equation:

$$l(\phi)=\begin{cases} C_a/\cos\phi & \text{if } 0\le\phi<\phi_1, \\ & \phi_4\le\phi<2\pi, \\ -C_a/\cos\phi & \text{if } \phi_2\le\phi<\phi_3, \\ \sqrt{X(\phi)^2+Y(\phi)^2} & \text{if } \phi_1\le\phi<\phi_2, \\ & \phi_3\le\phi<\phi_4, \end{cases} \tag{1}$$

where φ is clockwise from the back direction of the robot. ϕ_i are respectively $\phi_1=\arctan(C_L/C_a)$, $\phi_2=\pi-\arctan(C_L/C_a)$, $\phi_3=\pi+\arctan(C_R/C_a)$, $\phi_4=2\pi-\arctan(C_R/C_a)$. $X(\phi)$ and $Y(\phi)$ are calculated as following equations:

$$X(\phi)=\begin{cases} \dfrac{-C_L-\sqrt{C_L^2-(C_L^2-C_a^2)\{1+\tan^2(\pi/2-\phi)\}}}{1+\tan^2(\pi/2-\phi)} & \text{if } \phi_1\le\phi<\phi_2. \\ \dfrac{C_R+\sqrt{C_R^2-(C_R^2-C_a^2)\{1+\tan^2(\pi/2-\phi)\}}}{1+\tan^2(\pi/2-\phi)} & \text{if } \phi_3\le\phi<\phi_4. \end{cases} \tag{2}$$

$$Y(\phi)=X(\phi)\cdot\tan(\pi/2-\phi)\ . \tag{3}$$

In the proposed method, C_L, C_R, C_a are decided in a way that wide-robot shape falls within the capsule case.

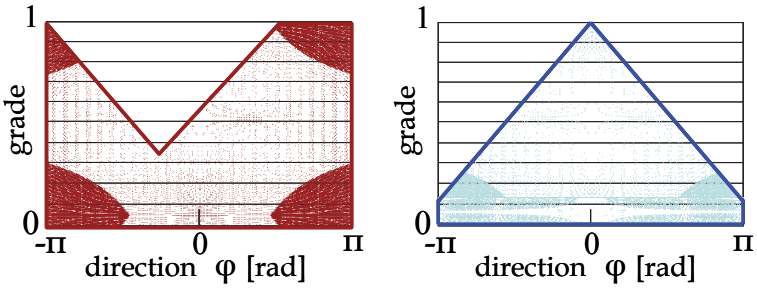


Fig. 3. Examples of PMFs

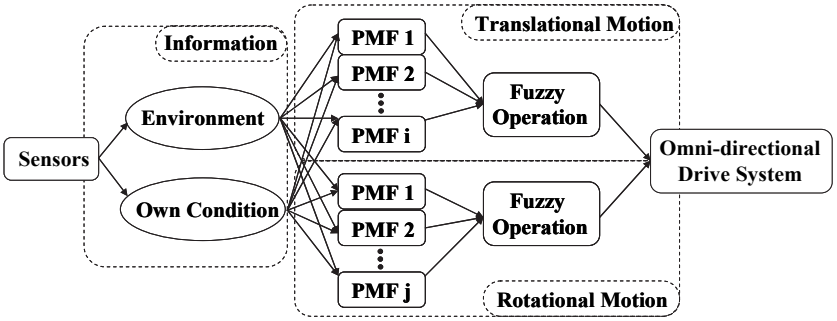


Fig. 4. A concept of fuzzy potential method considering translational and rotational motion with an omni-directional platform

3. Fuzzy potential method (FPM) using capsule case

3.1 Concept

In the fuzzy potential method (FPM), a current command velocity vector that takes into consideration element actions is decided by fuzzy inference. Element actions are represented as potential membership functions (PMFs), and then they are integrated by means of fuzzy inference. The directions on the horizontal axis in Fig. 3 correspond to the directions, which are from $-\pi$ to π radian measured clockwise from the front direction of the robot. Grades for the directions are represented on the vertical axis. The grades, direction, and configured maximum and minimum speeds, are used to calculate the current command velocity vector. Previously, the FPM dealt with the problem of the translational motion control in the same way as other control methods for autonomous mobile robots. In this paper, it is shown that modifying the FPM enables it to deal with rotational motion control, which is achieved concurrently with translational motions, within the FPM framework. In the modified framework as shown in Fig.4 ,PMFs for translational motions and rotational motions are designed respectively based not only on the environmental information but also on the robot's own condition. Environmental information and the robot's own condition are treated separately and divided into a translation problem and a rotational problem Then the PMFs of each problem are independently integrated using fuzzy inference. Finally, translational and rotational velocities, which are calculated by defuzzification of mixed PMFs, are realized by an omni-directional drive system.

3.2 PMF for translational motions

3.2.1 PMF for obstacles

In order to enable a wide-robot to avoid obstacles safely and efficiently in real time, a concave shaped PMF $\mu_{oj}^{t}(j=1,2,\cdots,n)$, which is considering the capsule case, is generated. This PMF is specified by depth and width, which are calculated based on geometrical relation between an obstacle and a robot as shown in Fig.6. By generating based on some variables, which are φ_L, φ_R, φ'_L, φ'_R, a and $\varphi_{r,o}$ in Fig. 5, the choice of safe direction becomes possible.

First, φ_L, φ_R, φ'_L, φ'_R are calculated as following equations:

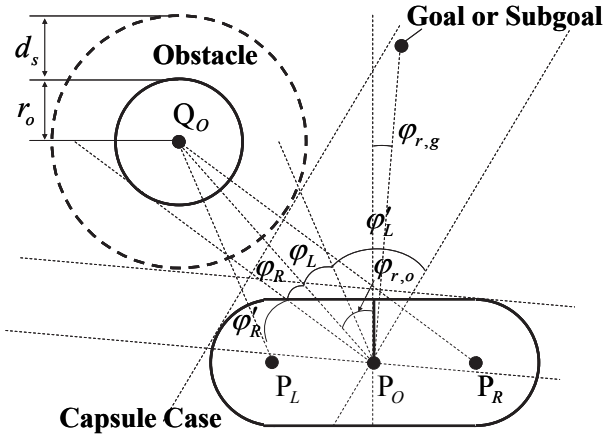
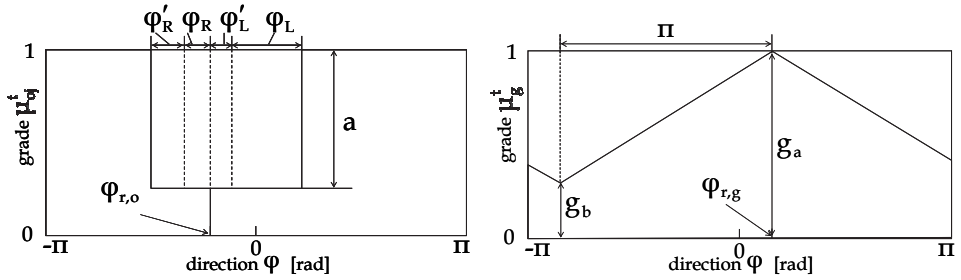


Fig. 5. A wide-robot and an obstacle



(a) (b)

Fig. 6. PMFs for translational motions: μ^t_{oj} is a PMF for an obstacle (a), μ^t_g is a PMF for a goal (b)

$$\varphi_L = \arccos\left(\frac{\left\|\overrightarrow{P_OQ_O}\right\|^2 + \left\|\overrightarrow{P_LQ_O}\right\|^2 - \left\|\overrightarrow{P_OP_L}\right\|^2}{2\left\|\overrightarrow{P_OQ_O}\right\| \bullet \left\|\overrightarrow{P_LQ_O}\right\|}\right). \tag{4}$$

$$\varphi_R = \arccos\left(\frac{\left\|\overrightarrow{P_OQ_O}\right\|^2 + \left\|\overrightarrow{P_RQ_O}\right\|^2 - \left\|\overrightarrow{P_OP_R}\right\|^2}{2\left\|\overrightarrow{P_OQ_O}\right\| \bullet \left\|\overrightarrow{P_RQ_O}\right\|}\right). \tag{5}$$

$$\varphi'_L = \begin{cases} \arcsin\left(\dfrac{D}{\left\|\overrightarrow{P_LQ_O}\right\|}\right) & \text{if } D < \left\|\overrightarrow{P_LQ_O}\right\|. \\ \Pi - \arcsin\left(\dfrac{\left\|\overrightarrow{P_LQ_O}\right\| - d_s}{D - d_s}\right) & \text{if } D \geq \left\|\overrightarrow{P_LQ_O}\right\|. \end{cases} \tag{6}$$

$$\varphi'_R = \begin{cases} \arcsin\left(\dfrac{D}{\left\|\overrightarrow{P_R Q_O}\right\|}\right) & \text{if } D < \left\|\overrightarrow{P_R Q_O}\right\|. \\ \pi - \arcsin\left(\dfrac{\left\|\overrightarrow{P_R Q_O}\right\| - d_s}{D - d_s}\right) & \text{if } D \geq \left\|\overrightarrow{P_R Q_O}\right\|. \end{cases} \tag{7}$$

Next, as a measure to decide how far the robot should depart from the obstacle, a is defined as the depth of the concave shaped PMF. a is described as following equation:

$$a = \frac{\alpha - \left\|\mathbf{r}_{r,o}\right\|}{\alpha - D} \qquad \text{if } \left\|\mathbf{r}_{r,o}\right\| < \alpha \ . \tag{8}$$

where $\mathbf{r}_{r,o} = (r_x, r_y)$ is current position vector of the obstacle relative to the robot
If the current obstacle position is ins
ide of a circle with radius α from the robot position, the PMF for obstacle avoidance is generated. In other words, if a relative distance $\left\|\mathbf{r}_{r,o}\right\|$ is below α, a is defined and the concave shaped PMF corresponding to the obstacle is generated. D is decided to ensure the safety distance as following equation:

$$D = C_a + r_o + d_s \ . \tag{9}$$

where C_a is the minimum length of capsule case as shown in Fig.2. r_o and d_s denote respectively the radius of the obstacle and safety distance. $\varphi_{r,o}$ is the angle of direction to the obstacle relative to the robot, which is calculated as following equation:

$$\varphi_{r,o} = \arctan\left(\frac{r_y}{r_x}\right) . \tag{10}$$

The PMF μ_{oj}^t is generated for all obstacles which the robot has detected. And then, they are all integrated by calculating logical product μ_o^t, as following equation:

$$\mu_o^t = \mu_{o1}^t \wedge \mu_{o2}^t \wedge \ \cdots \ \wedge \mu_{oj}^t \ . \tag{11}$$

As mentioned above, by deciding the depth and the base width of concave, PMF μ_o^t, which aims to early starting of avoidance behavior and prompt the direction of the velocity vector to be far from obstacle direction in response to the fast-moving obstacle, is generated.

3.2.2 PMF for a goal

To head to the goal, a PMF μ_g^t shaped like triangle as shown in Fig.6 (b). μ_g^t is specified by g_a, g_b, $\varphi_{r,g}$. As a measure to decide how much the robot want to head to the goal, g_a is defined as the height of the triangular PMF. As a measure to decide how much the robot is allowed to back away from obstacles, g_b is defined. μ_g^t gets the maximum value as g_a at an angle of the goal direction relative to the front direction of the robot, $\varphi_{r,g}$, and gets the minimum value as g_b at an angle of a direction opposite to the goal direction. g_a and g_b are described as following equations:

$$g_a = \begin{cases} \dfrac{\left\|\mathbf{r}_{r,g}\right\|}{\varepsilon} & \text{if } \left\|\mathbf{r}_{r,g}\right\| \le \varepsilon \\ 1.0 & \text{if } \left\|\mathbf{r}_{r,g}\right\| > \varepsilon \end{cases}, \tag{12}$$

$$g_b = \eta g_a \qquad (0 \le \eta < 1), \tag{13}$$

where $\left\|\mathbf{r}_{r,d}\right\|$ is an absolute value of the position vector of the goal relative to the robot. ε and η are constants. If $\left\|\mathbf{r}_{r,d}\right\|$ is below ε, g_a is defined. The shorter the distance between the obstacle and the robot is, the smaller g_a becomes. The robot decelerates and stops stably.

3.3 Calculation of translational command velocity

The proposed method employs fuzzy inference to calculate the current command velocity vector. Specifically, The PMF μ_o^t and the PMF μ_g^t are integrated by fuzzy operation into a mixed PMF μ_{mix}^t as shown in Fig. 7. μ_{mix}^t is an algebraic product of μ_o^t and μ_g^t as following equation:

$$\mu_{mix}^t = \mu_g^t \cdot \mu_o^t \ . \tag{14}$$

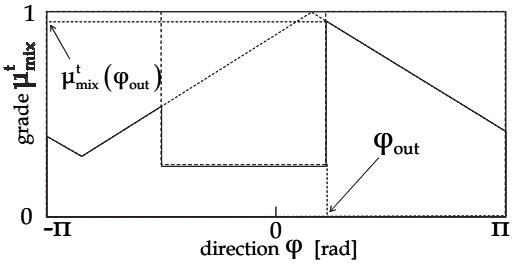


Fig. 7. A mixed PMF for translational motion

Finally, by defuzzifier, the command velocity vector is calculated as a traveling direction φ_{out} and an absolute value of the reference speed of the robot base on the mixed PMF μ_{mix}^t. φ_{out} is decided as the direction which makes the PMF $\mu_{mix}^t(\varphi)$ maximum.
Based on φ_{out}, v_{out} is calculated as following equation:

$$v_{out} = \mu_{mix}^t(\varphi_{out})(v_{max} - v_{min}) + v_{min} \ . \tag{15}$$

where $\mu_{mix}^t(\varphi_{out})$ is the mixed PMF for translational movement corresponding to the φ_{out}, v_{max} and v_{min} are configured in advance respectively as higher and lower limit of the robot speed.

3.4 PMF for rotational motions

3.4.1 PMF for obstacles

In order to enable a wide-robot to decide the appropriate angle of the direction for obstacle avoidance in real time, PMF μ_o^r is generated based on a PMF μ_e^r, which considers the environmental information, and another PMF μ_c^r as following equation:

$$\mu_o^r = \mu_e^r - \mu_c^r . \tag{16}$$

μ_e^r is generated based on the information of distances from the center of the robot to obstacles corresponding to all directions, as shown in Fig. 8. The relative information of the distances is obtained by use of range sensors such as the laser range finder, the ultra sonic sensors or the infrared sensors. μ_c^r is generated based on the capsule case which is introduced in 2. is calculated with Eq. (1) as following equation:

$$\mu_c^r(\varphi) = \frac{1(\varphi + \pi)}{\alpha}. \tag{17}$$

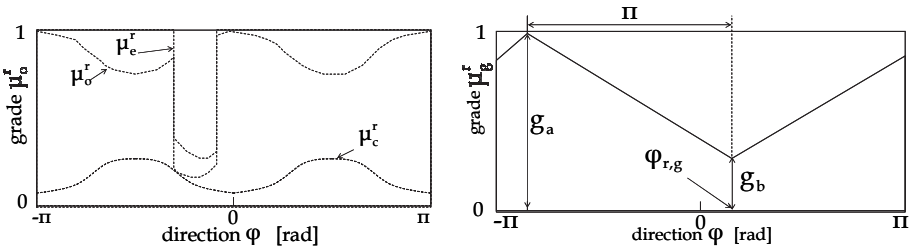


Fig. 8. PMFs for rotational motions: μ_o^r is a PMF for an obstacle (a), μ_g^r is a PMF for a goal (b)

The aim of the PMF μ_o^r in (16) is to search an orientation angle of the robot which enables the distance between a point on capsule case and each obstacle to maximize, by turning front or back side of the robot on the direction that there is a closest point to each obstacle. By considering the capsule case, a design of PMF can deal with the width of the robot for rotational motion.

3.4.2 PMF for a goal

In order to turn front on the goal direction or traveling direction if there is no obstacle to avoid, PMF for a goal for rotational motion is generated as μ_g^r. This shape is decided in same way with μ_g^t, by using (12), (13).

3.5 Calculation of rotational command velocity

As for the rotational movement, like the translational movement, the proposed method employs fuzzy inference to calculate the current rotational command velocity vector. Specifically, The PMF μ_e^r, which considers the own condition by using Eq. (1), and the PMF μ_g^r, which is to head to the goal, are integrated by fuzzy operation into a mixed PMF μ_{mix}^r as shown in Fig. 9. μ_{mix}^r is an algebraic product of μ_o^r and μ_g^r as following equation:

$$\mu_{mix}^r = \mu_g^r \cdot \mu_o^r . \tag{18}$$

Finally, by defuzzifier, the command velocity vector is calculated as a traveling direction φ_{ori} and an absolute value of the reference speed of the robot base on the mixed PMF μ_{mix}^r. φ_{ori} is decided as the direction φ_i which makes a following function $h(\varphi)$ minimum.

$$h(\varphi) = \int_{\varphi-\zeta}^{\varphi+\zeta} \mu_{mix}^{r}(\psi)\, d\psi \tag{19}$$

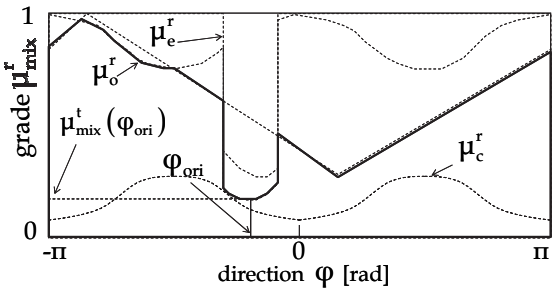


Fig. 9. A mixed PMF for rotational motion

where ζ is the parameter to avoid choosing undesirable φ_i caused by such as noises on the sensor data. Based on φ_{ori}, ω is calculated as following equation:

$$\omega = \mathrm{sgn}(\varphi_{ori}) \tag{20}$$

where $\mu^r_{mix}(\varphi_{ori})$ is the mixed PMF for translational movement corresponding to the φ_{ori}, ω_{max} and ω_{min} are configured in advance respectively as higher and lower limit of the rotational speed of the robot.

3.6 Calculation of wheel speeds

To realize the movement, in this study, an omni-directional platform is employed for a autonomous mobile robot. The command velocity vector is realized by four DC motors and omni wheels using following equations:

$$v_r^x = \|\mathbf{v}_{out}\| \cos\varphi_{out} \ . \tag{21}$$

$$v_r^y = \|\mathbf{v}_{out}\| \sin\varphi_{out} \ . \tag{22}$$

$$\begin{pmatrix} v_1^w \\ v_2^w \\ v_3^w \\ v_4^w \end{pmatrix} = \begin{pmatrix} \cos\delta & \sin\delta & R \\ -\cos\delta & \sin\delta & R \\ -\cos\delta & -\sin\delta & R \\ \cos\delta & -\sin\delta & R \end{pmatrix} \begin{pmatrix} v_r^x \\ v_r^y \\ \omega \end{pmatrix}. \tag{23}$$

where $\mathbf{v}_{out}$ and ω are respectively current command translational velocity vector and rotational speed. δ is an angle of gradient for each wheel. R is a half of a distance between two catawampus wheels. v_i^w is a command movement speed of each i-th wheel.

4. Simulation results

The effectiveness of the proposed method was verified by numerical simulations intended for omni-directional autonomous mobile robots. As postulates, the robot supposed to be

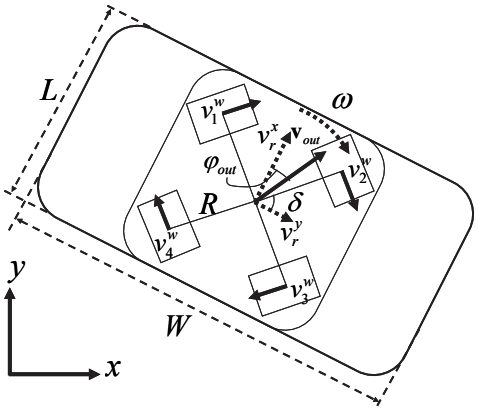
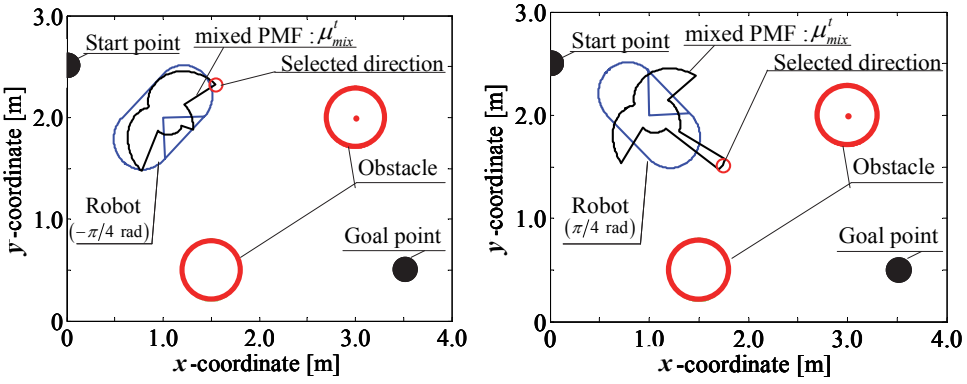


Fig. 10. An omni-directional platform with wide body

able to detect obstacles and has information about the relative position vector. The measuring range was assumed to be 4.0m at all directions. Each parameter was as follows: The wide-robot size was assumed as L=0.4m, W=1.0m, which are in Fig.10. Considering this L and W, C_a, C_L and C_R in Fig.2 were all set at 0.3m. r_o, d_s in Fig.5 were both set at 0.3m. Consequently D=0.9m in Eq.(9). α in Eq. (8) was 4.0m. η in Eq. (13) was 0.2. ε in Eq. (12) was 1.0m. v_{max}, and v_{min} in Eq. (15) were respectively 0.5m/s and 0.0m/s.

4.1 Performance of capsule case

In this section, the effectiveness of using capsule case and the design method of PMF based on the capsule case are verified, by comparing the results of chosen direction of movement at following two different situations about the orientation angle for a wide-robot. As common assumption, the positions of the robot and two obstacles were immobilized on each point respectively (1.0m, 2.0m), (1.5m, 0.5m) and (3.0m, 2.0m), as shown in Fig.11.

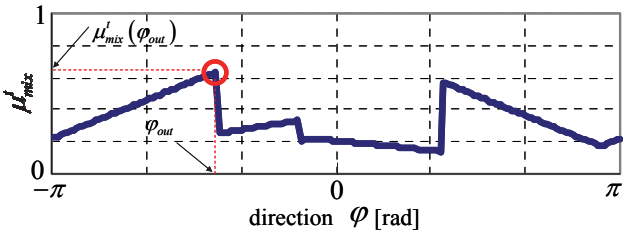


(a) Situation I (b) Situation II

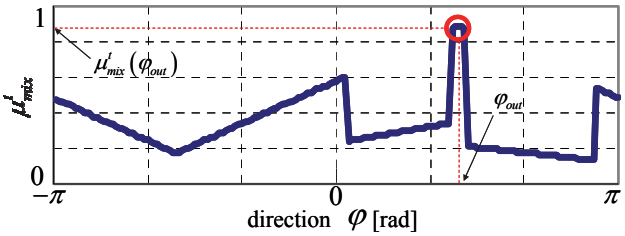
Fig. 11. Simulation results : in Situation I (a), the robot cannot find the direction between the two obstacles, in Situation II (b), the robot can find the direction between the two obstacles.

4.1.1 Situation I

The orientation angle of the robot was fixed to -π /4 radian clockwise from *x*-axis on the absolute coordinate. Therefore, the robot faced to a goal point, as shown in Fig.11(a), however, the chosen direction of the current movement of the robot was calculated as -1.35 radian, which was clockwise from the front direction of the robot, as shown in Fig.11(a). This value of the chosen direction was calculated based on the mixed PMF μ^{t}_{mix} in this situation as shown in Fig.12(a). As a result, the robot chose the direction, which the robot would go the roundabout route.



(a) Situation I : the robot cannot find the direction between the two obstacles



(b) Situation II: the robot can find the direction between the two obstacles

Fig. 12. Aspects of mixed PMFs for translational motion in two different situations: (a) is in relation to the situation of Fig.11(a), and (b) is in relation to Fig.11(b).

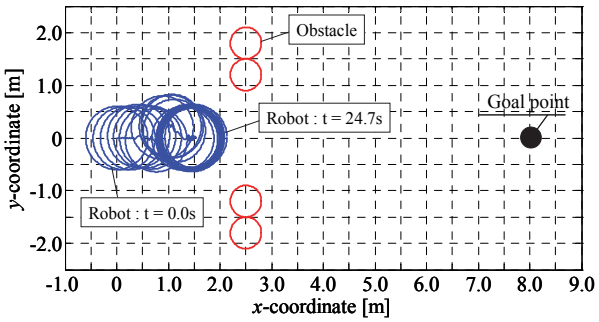


Fig. 13. A simulation result of Method I (conventional) : The robot not using capsule case didn't succeeded in going through between two divided objects

4.1.2 Situation II

The orientation angle of the robot is fixed to π /4 radian on the absolute coordinate. As contrasted to Situation I, the robot didn't faced to a goal point, as shown in Fig.11(b), however, the chosen direction of the current movement of the robot was calculated as 1.37 radian, which was clockwise from the front direction of the robot, as shown in Fig.11(b). This value of the chosen direction was calculated based on the mixed PMF μ^{t}_{mix} in this situation is shown in Fig.12(b). As a result, the robot chose the direction, which the robot would take a shorter route without collision.

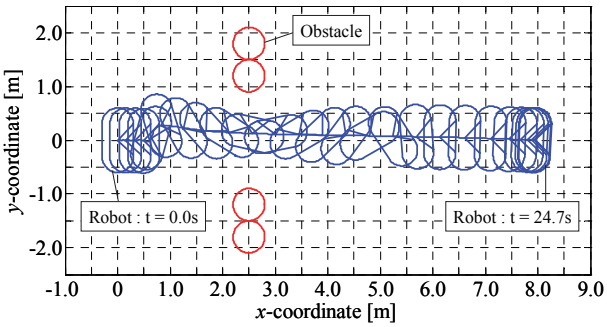


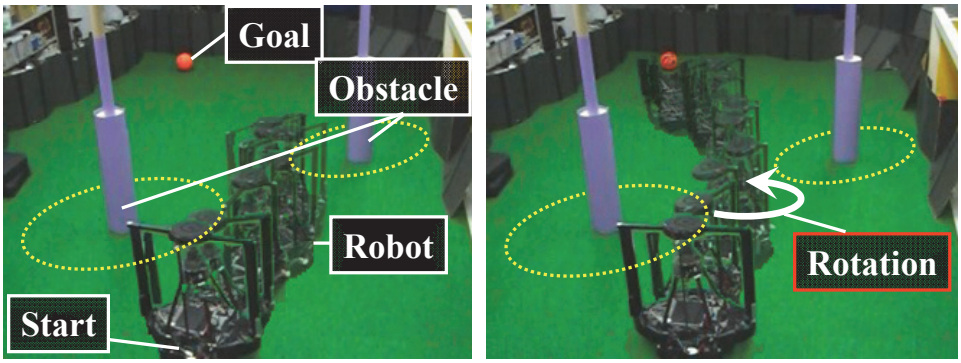
Fig. 14. A simulation result of Method II (proposed): The robot using capsule case succeeded in going through between two objects with translational and rotational motion in real time.

Through these two results, the effectiveness of the capsule case is confirmed. The wide-robot can decide the direction of translational motion with considering the own orientation, goal position and obstacle positions simultaneously in real time.

4.2 Capability of going through between objects

The effectiveness of the proposed method was tested by comparing two design methods, I and II, based on PMF, for obstacle avoidance problem. Start and goal point of the robot are respectively (0.0m,0.0m) and (8.0m,0.0m). The trajectory of the robot and the aspects considering the orientation angle on the position every 1 second are plotted in Fig. 13 and Fig. 14. Obstacles positions are respectively (2.5m,−1.8m), (2.5m,−1.2m), (2.5m,1.2m), (2.5m,1.8m). In Method I, as a conventional method, the robot was regarded as a circle with radius 0.6m. In Method II, as a proposed method, the capsule case was used and rotational motion of the robot was taken into consideration. When the Method I was used, the robot was based on the maximum radius and did not take into consideration the rotational motion. Therefore, the robot did not succeed in going between two objects as shown in Fig. 13. While the robot did not collide with the obstacles, the robot did not get to the goal.

On the other hand, in the Method II, the capsule case and real-time control based on FPM were used. As shown in Fig. 14, the robot performed translational and rotational motions simultaneously and succeeded in going between two objects. In addition, the robot succeeded in getting to the goal with the orientation angle 0 radian using PMF for rotational motion.



(a) a robot without capsule case (b) a robot with capsule case

Fig. 15. Experimental results: (a) shows a result of Method I. The robot not using capsule case didn't succeed in going through between two objects. (b) shows a result of Method II. The robot using capsule case succeeded in going through between two objects with translational and rotational motion in real time.

The effectiveness of the proposed method was verified also by simplified experiments using omni-directional autonomous mobile robots as shown in Fig.15. In each picture of the Fig.15, aspects of the robot every 1 second are plotted. The robot recognized environment by the omni-directional camera. A position of a goal and that of obstacles relative to the robot were calculated by extracting features in images based on objects' colours. A ball was assumed as the goal and columns were assumed as obstacles, as shown in Fig.15. A dashed circle enveloping a column in Fig.15 corresponds to a dashed circle in Fig.5. As shown in Fig. 15(a), the robot was not able to between two objects without the capsule case. However, as shown in Fig. 15(b), the robot with the capsule case performed translational and rotational motion simultaneously in real time and succeeded in going between two obstacles.
These results showed that motion control without a capsule case made it difficult for the robot to go between two objects due to the largest radius of the robot, even if it would be able to go through by using its shortest radius. Applying the capsule case to a wide robot enhances the possibility of going between two objects.

5. Conclusion

In this paper, the real time control method of simultaneously translational and rotational motions for an autonomous mobile robot, which is horizontally long, has been introduced. This method employs omni-directional platform for the drive system and is based on the fuzzy potential method (FPM). The novel design method of potential membership function (PMF), which is considered the width of the robot by using the capsule case, has been shown. According to this proposed method, the wide-robot can decide the current direction of translational motion to avoid obstacles safely by using capsule case. In addition, by controlling the rotational motion in parallel with the translational motion in real time, the wide-robot can go through narrow distance between two objects. The effectiveness has been verified by numerical simulations and simplified experiments. It has been shown that the proposed method enables simultaneous control of the translational and rotational velocity within the framework of FPM.

5. References

Kavraki, L., 1995. Computation of Configuration Space Obstacles Using the Fast Fourier Transform, *IEEE Trans. on Robotics and Automation*, Vol. 11, No. 3, pp. 408-413.

Wang, Y., Chirikjian, G. S., 2000. A New Potential Field Method for Robot Path Planning, *Proc. IEEE Int. Conf. on Robotics and Automation*, San Francisco, CA, pp. 977-982.

Ambrose, R. O., Savely, R. T., Goza, S. M., Strawser, P., Diftler, M. A., Spain, I., and Radford, N., 2004. Mobile manipulation using NASA's robonaut, *Proc. IEEE ICRA*, pp. 2104-2109.

Du, Z., Qu, D., Yu, F. and Xu, D, 2007. A Hybrid Approach for Mobile Robot Path Planning in Dynamic Environments, *Proc. IEEE Int. Conf. on Robotics and Biomimetics*, pp.1058-1063.

Khatib, O., 1986. Real-time Obstacle Avoidance for Manipulators and Mobile Robots, *Int. J. of Robotics Research*, Vol.5, No.1, pp.90-98.

Borenstein, J., Koren, Y. , 1989. Real-Time Obstacle Avoidance for Fast Mobile Robots, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol.19, No.5, pp.1179-1187.

Borenstein, J., Koren, Y., 1991. The Vector Field Histogram Fast Obstacle Avoidance for Mobile Robots, *IEEE Trans. on Robotics and Automation*, Vol.7, No.3, pp.278-288.

Lumelsky, V. J., Cheung, E. , 1993. Real Time Obstacle Collistion Avoidance in Teleoperated Whole Sensitive Robot Arm Manipulators, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.23, No.1, pp.194-203.

Dieter, F., Wolfram, B., Sebastian, T., 1997. The Dynamic Window Approach to Collision Avoidance, *IEEE Robotics and Automation*, Vol. 4, No. 1, pp.1-23.

Tsuzaki, R., Yoshida, K., 2003. Motion Control Based on Fuzzy Potential Method for Autonomous Mobile Robot with Omnidirectional Vision". *Journal of the Robotics Society of Japan*, Vol.21, No.6, pp.656-662. Takahashi, M., Suzuki, T., 2009. Multi Scale Moving Control Method for Autonomous Omni-directional Mobile Robot, *Proc. of the 6th Int. Conf. on Informatics in Control, Automation and Robotics.*

27

Obstacle Avoidance for Autonomous Mobile Robots Based on Position Prediction Using Fuzzy Inference

Takafumi Suzuki and Masaki Takahashi
Keio University
Japan

1. Introduction

In the future, it's not difficult to image that we will often come across many autonomous mobile robots traversing densely populated place we live in. In such situation, because the autonomous mobile robots need to carry out their tasks in a place with unknown obstacles, the obstacle avoidance is one of the important functions of the robots. With a view to implementation of autonomous mobile robot working in doors, we employ an omni-directional platform as shown in Fig. 1 (left). For experimental verification, an omni-directional mobile robot shown in Fig. 1 (right) is developed. The robot has an omni-directional camera for environmental recognition, and can move to all directions by four omni wheels.

While there are many studies about obstacle avoidance method focusing attention on possibility of avoidance, this paper presents the method focusing on not only possibility but also safer trajectory of avoidance. Even if there are the same situations that the robot needs to avoid a static obstacle, timing of beginning avoidance behaviour should vary according to the robot speed. If the obstacles are moving also, the timing should vary according to the velocities of the obstacles. To cite a case, in a situation that a robot and an obstacle go by each other as shown in Fig. 2, the robot should avoid along the curved line like (iii) according to the speeds of the obstacle and own speed. In order to get to the goal with efficient and safe avoidance behaviour in the unknown environment for the robots, predicting the future obstacles' positions by their current motions is needed.

This paper introduces a real-time obstacle avoidance method introducing the velocity of obstacle relative to the robot. By means of considering predicted positions of the robot and the obstacle calculated from the time and the relative velocity, the robot can begin the avoidance behaviour at an appropriate time according to the velocity of the obstacle and the robot.

Some researches focus attention on the velocity of obstacle (Ko & Lee, 1996) to avoid moving obstacles efficiently. In this research, virtual distance function is defined based on distance from the obstacle and speed of obstacle, however, only projection of the obstacle velocity on the unit vector from the obstacle to the robot is considered. In other words, the velocity of the robot is not considered.

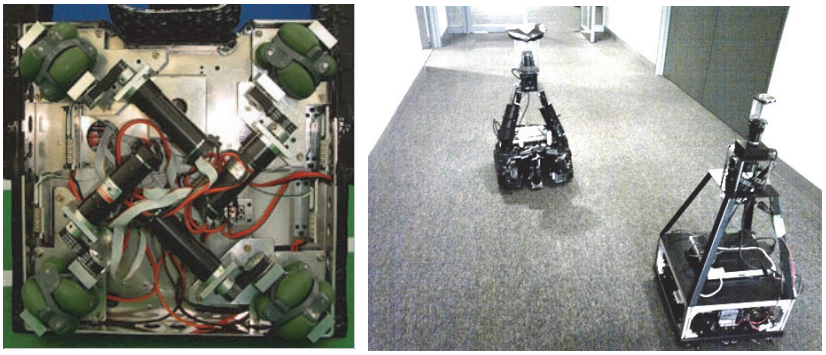


Fig. 1. An omni-directional platform of a prototype robot (left), a situation that the robot needs to avoid the other robot (right)

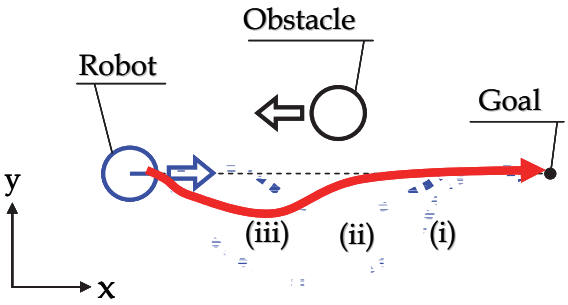


Fig. 2. A situation of obstacle avoidance

On the other hand, in (Ge & Cui, 2002), the velocity of the obstacle relative to the robot is considered. Our approach also employs the relative velocity. In addition to this approach, a position vector of the obstacle relative to the robot in the future is calculated by the relative position and the velocity.

In order to solve the real-time motion planning problem, there are effective methods (Khatib, 1986) (Arkin, 1989) (Borenstein & Koren, 1991) (Fox et al., 1997). Fuzzy potential method (FPM) (Tsuzaki & Yoshida, 2003) (Otsuka et al. 2005) is also one of the effective methods. In this research, the method was applied to an autonomous mobile robot which plays soccer. By adequate designing of potential membership function (PMF), it was realized that wheeled robots got to the goal with conveying a soccer ball and avoiding obstacles. This method is easy to understand at a glance. However, in dynamic environment, to avoid moving obstacles efficiently, more specific guideline of designing is desired. In this paper, we introduce design method of PMF considering the predicted positions and discuss the availability by comparing the design of PMF considering the relative velocity and that not considering.

In this paper, for the purpose of avoiding the moving obstacle safely and smoothly, design methods of the potential membership function (PMF), which is considering the velocity of the obstacle relative to the robot, are presented.

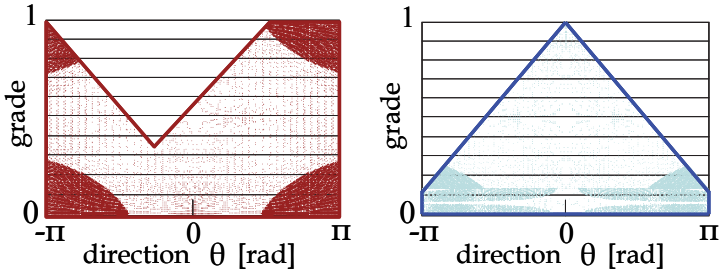


Fig. 3. Potential membership functions (PMFs): for an obstacle (left) and a goal (right).

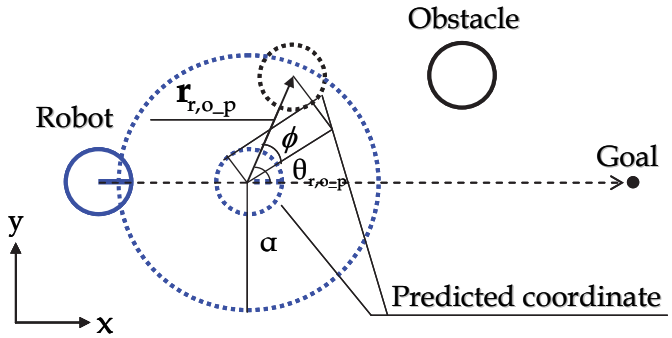


Fig. 4. Predicted coordinates of a robot and an obstacle

2. Fuzzy potential method (FPM)

In the fuzzy potential method (FPM), a current command velocity vector considering element actions is decided in real time. Element actions are represented as potential membership functions (PMFs), and then they are integrated by means of fuzzy inference. Furthermore, by using a state evaluator, the PMFs are modified adaptively according to the situation. The directions on the horizontal axis in Fig. 3 correspond to the directions which are from $-\pi$ to π radians and measured clockwise from the front direction of the robot. The grade for the direction is represented on the vertical axis. By use of the grade, direction and configured maximum and minimum speed, the current command velocity vector $\mathbf{v}_{out} = (v_r^x, v_r^y)$ is calculated. The command velocity vector is realized by an omni directional platform.

PMF idea allows us to represent our knowledge and experiences easily, and furthermore it gives us easy understanding. The grade can be seen as a desire for each direction of the robot. In this paper, to discuss an obstacle avoidance problem, methods for generating of PMF to head to a goal and to avoid moving obstacles are introduced. This method has two steps. First step is generating PMFs. Second step is deciding the command velocity vector by use of fuzzy inference to integrate the PMFs. Hereinafter, design method of PMF considering the obstacle velocity relative to the robot and way to decide the command velocity vector by fuzzy inference are described

3. FPM considering the relative velocity

To realize the obstacle avoidance in dynamic environment, the proposed method employs two different PMFs, one is considering vectors of obstacle's position and obstacle's velocity relative to the robot and the other is to head to the goal. PMF is denoted by μ which is function of θ. Note θ is the direction from $-\pi$ to π radians measured clockwise from front direction of the robot. To simplify the analysis, it is assumed that the autonomous mobile robot detects obstacles by equipped external sensors and is capable of calculating the positions and velocities of obstacles relative to the robot. The shapes of the robot and the obstacles are treated as circles on 2D surface.

3.1 Design of PMFs

3.1.1 PMF for an obstacle

To avoid moving obstacles safely and efficiently, an inverted triangular PMF by specifying a vertex, height and base width is generated. Because this PMF considers future positions of the robot and the obstacle, the robot can start avoiding the obstacle early and be prompted not to go on to the future collision position. For the purpose of safe avoidance, the PMF μ_o is generated.

First, in order to predict the future state of both obstacle and robot with the aim of efficient avoidance, a predicted relative position vector, in γT seconds, $\mathbf{r}_{r,o\_p} = (r_{x\_p}, r_{y\_p})$ is calculated as following equation:

$$\mathbf{r}_{r,o\_p} = \mathbf{r}_{r,o} + \gamma T \mathbf{v}_{r,o} \tag{1}$$

where $\mathbf{r}_{r,o} = (r_x, r_y)$ is current position vector of the obstacle relative to the robot, and $\mathbf{v}_{r,o} = (v_x, v_y)$ is the current velocity vector of obstacle relative to the robot. γ is an arbitrary parameter from 0 to 1. T ,which is the time until the distance between the obstacle and the robot is minimum, is defined as following equation:

$$T = \frac{\left\| \mathbf{r}_{r,o} - \mathbf{p} \right\|}{\left\| \mathbf{v}_{r,o} \right\|} \tag{2}$$

where $\mathbf{p} = (p_x, p_y)$ is a position vector of the obstacle relative to the robot when a distance in the future between the obstacles and the robot is minimum. $\mathbf{p}$ is calculated by means of relative position and velocity vector as following equation:

$$\begin{pmatrix} p_x \\ p_y \end{pmatrix} = \begin{pmatrix} \left\{ (v_y / v_x) r_y - r_x \right\} \Big/ (v_y / v_x + v_x / v_y) \\ -(v_y / v_x) p_x \end{pmatrix} \tag{3}$$

As described above, the predicted relative position vector, at the time γT seconds from now, $\mathbf{r}_{r,o\_p}$ is calculated as Fig. 4 shows. By use of this position vector, a predicted obstacle direction relative to the robot θ_{r,o_p} is calculated as following:

$$\theta_{r,o\_p} = \arctan\left(\frac{r_{y\_p}}{r_{x\_p}} \right) \tag{4}$$

where θ_{r,o_p} is decided to be the vertex of the inverted triangle.
Next, as a measure to decide how far the robot should depart from the obstacle, a is defined as the height of the inverted triangular PMF. a is described as following equation:

$$a = \frac{\alpha - \left\|\mathbf{r}_{r,o\_p}\right\|}{\alpha - R_{r,o}} \qquad \text{if } \left\|\mathbf{r}_{r,o\_p}\right\| < \alpha \tag{5}$$

$$R_{r,o} = R_r + R_o \tag{6}$$

where R_r and R_o denote respectively the radius of the robot and that of the obstacle treated as circles. If the calculated obstacle position at γT seconds later is inside of a circle with radius α from the robot position at γT seconds later, the PMF for obstacle avoidance considering the relative velocity is generated. In other words, if a predicted relative distance $\left\|\mathbf{r}_{r,o\_p}\right\|$ is below α, a is defined and the inverted triangular PMF corresponding to the obstacle is generated. Smaller the predicted relative distance is, larger the value of a is.
In addition, a base width of inverted triangular PMF is decided by following equation:

$$b = \eta\left\|\mathbf{v}_{r,o}\right\| + \phi \tag{7}$$

where ϕ is decided based on the sum of radiuses of the robot and the obstacle ($R_{r,o}$) , and predicted relative position vector ($\mathbf{r}_{r,o\_p}$) as Fig. 4 shows. ϕ is calculated by following equation:

$$\phi = \arcsin\left(\frac{R_{r,o}}{\left\|\mathbf{r}_{r,o\_p}\right\|}\right). \tag{8}$$

b increases up to π radian in proportion to an absolute value of the relative velocity and predicted relative distance. If the obstacle comes at rapidly, for instance, the value of b increases. Hence, the base width grows shown in Fig. 5, and the value of grade for the direction of the obstacle relative to the robot comes about to be reduced. η is a gain.
As mentioned above, by deciding the vertex, the height and the base width of inverted triangle considering the predicted relative position, PMF μ_o, which aims to early starting of avoidance behavior and prompt the direction of the velocity vector to be far from obstacle direction in response to the fast-moving obstacle, is generated.

3.1.2 PMF for a goal

To head to the goal, a PMF μ_d shaped like triangle is generated as shown in Fig. 6. As a measure to decide how much the robot want to head to the goal, c is defined as the height of the triangular PMF. c gets the maximum value at an angle of the goal direction θ_d, which is relative to the front direction of the robot, and is described as following equation:

$$c = \begin{cases} \dfrac{\left\|\mathbf{r}_{r,d}\right\|}{\varepsilon} & \text{if } \left\|\mathbf{r}_{r,d}\right\| \le \varepsilon \\ 1.0 & \text{if } \left\|\mathbf{r}_{r,d}\right\| > \varepsilon \end{cases} \tag{9}$$

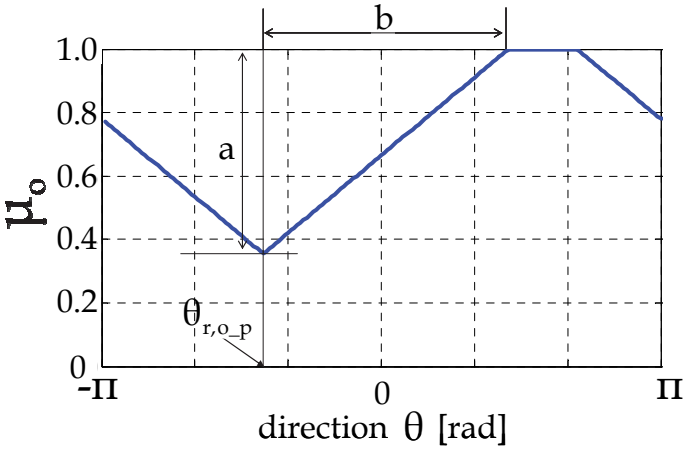


Fig. 5. A PMF for an obstacle considering relative velocity

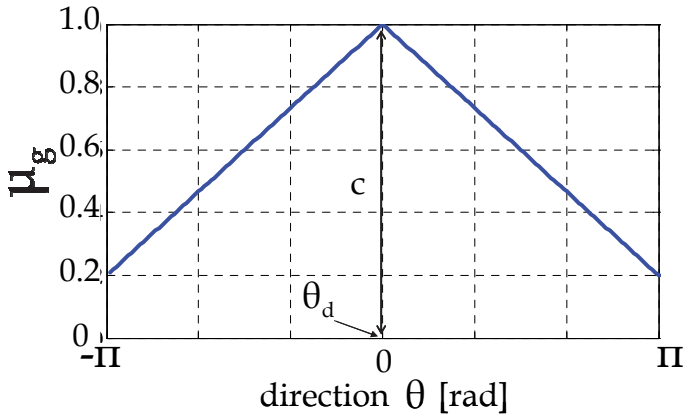


Fig. 6. A PMF for a goal

where $\|\mathbf{r}_{r,d}\|$ is an absolute value of the position vector of the goal relative to the robot. ε is constant. If $\|\mathbf{r}_{r,d}\|$ is below ε, c is defined. The shorter the distance between the obstacle and the robot is, the smaller c becomes. Therefore the robot can decelerate and stop stably.

3.2 Calculation of command velocity vector by fuzzy inference

The proposed method employs fuzzy inference to calculate the current command velocity vector. Specifically, The PMF μ_o, which considers the velocity of obstacle relative to the robot, and the PMF μ_d, which is to head to the goal, are integrated by fuzzy operation into a mixed PMF μ_{mix} as shown in Fig. 7. μ_{mix} is an algebraic product of μ_o and μ_d as following equation:

$$\mu_{mix} = \mu_d \cdot \mu_o \tag{10}$$

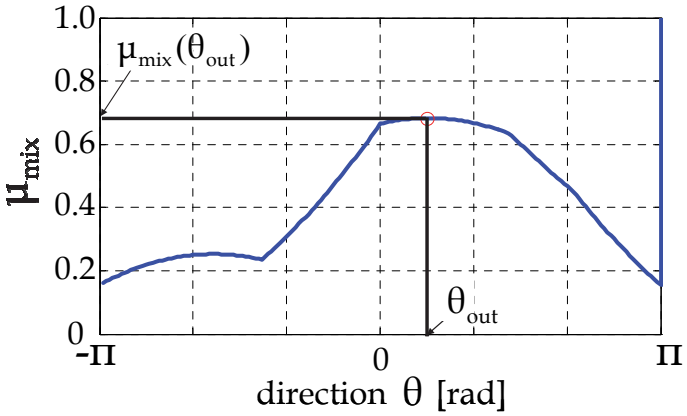


Fig. 7. A mixed PMF

Finally, by defuzzifier, the command velocity vector $\mathbf{v}_{out} = (v_r^x, v_r^y)$ is calculated as a traveling direction θ_{out} and an absolute value of the reference speed of the robot v_{out} base on the mixed PMF μ_{mix}. θ_{out} is decided as the direction θ_i which makes a following function $f(\theta)$ maximum.

$$f(\theta) = \sum_{i=j-n}^{j+n} \mu_{mix}(\theta_i) \tag{11}$$

where n is the parameter to avoid choosing undesirable θ_i caused by such as noises on the sensor data. Based on θ_{out}, v_{out} is calculated as following equation:

$$v_{out} = \mu_{mix}(\theta_{out})(v_{max} - v_{min}) + v_{min} \tag{12}$$

where $\mu_{mix}(\theta_{out})$ is the mixed PMF μ_{mix} corresponding to the θ_{out}, v_{max} and v_{min} are configured in advance respectively as higher and lower limit of the robot speed.
Based on θ_{out} and v_{out}, $\mathbf{v}_{out} = (v_r^x, v_r^y)$ is calculated as following equation:

$$v_r^x = v_{out}\cos\theta_{out} \ . \tag{13}$$

$$v_r^y = v_{out}\sin\theta_{out} \ . \tag{14}$$

3.3 Visualization for PMF on two-dimension surface

It would be convenient to have a visualizer that show us why the robot will go on to the direction. In the proposed method, we can see aspects of the PMFs on two dimension surface and understand easily the reason for choice of the direction. For example, a PMF described on polar coordinate shown in Fig. 9 (left) is comparable to the PMF described on x-y coordinate shown in Fig. 9 (right).

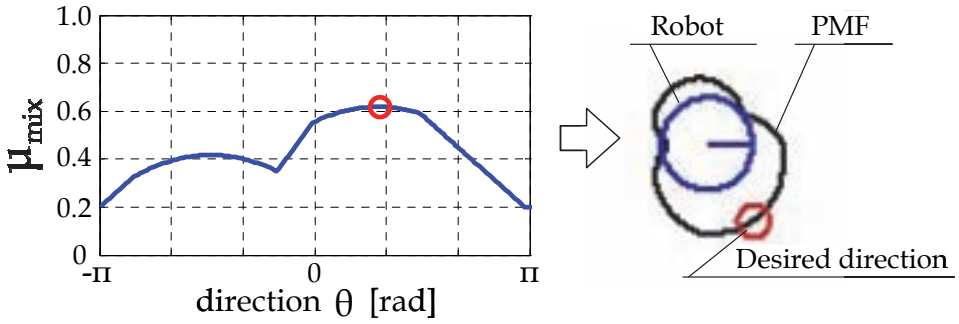
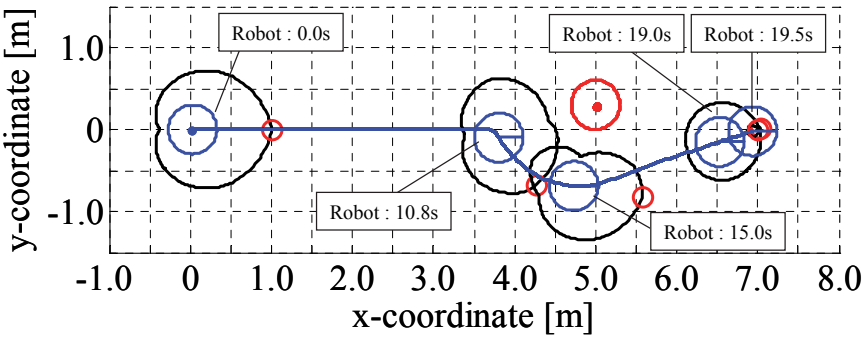
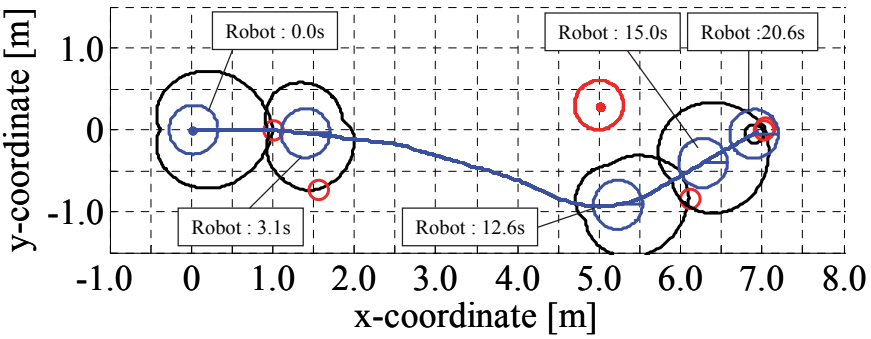


Fig. 9. Visualization of PMF



(a) not using PMF considering relative velocity



(b) using PMF considering relative velocity

Fig. 10. Simulation results of an obstacle avoidance going by each other when speed of obstacle was 0.0 m/s and of a robot was 0.5 m/s

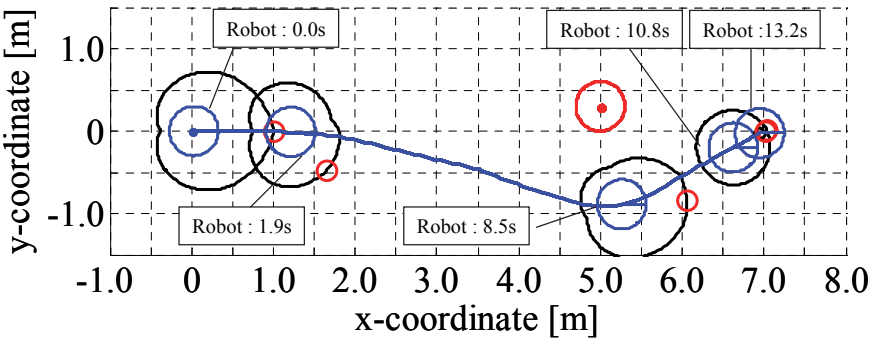


Fig. 11. Simulation results of an obstacle avoidance going by each other when speed of obstacle was 0.0m/s and of a robot was 0.8m/s

4. Simulation results

To verify the effectiveness of the proposed method that employs PMF considering the velocity of the obstacle of the robot, numerical simulations which assumed an obstacle avoidance of autonomous omni-directional mobile robot were carried out.

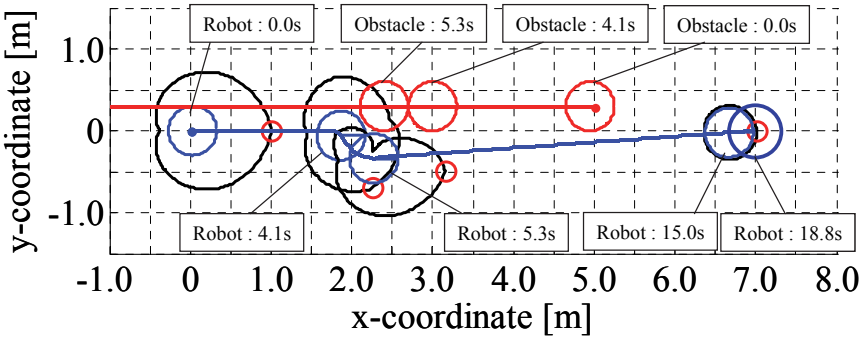
In this simulations as shown in Figs. 10-12, the radius of robot and obstacle were supposed to be both 0.3m, therefore, $R_{r,o} = 0.6\text{m}$. α in Eq. (5) was 1.6m. γ in Eq. (1) was 0.7. ε in Eq. (9) is 1.0m.

Figure 10, 11 and 12 show the simulation results when the robot passed the obstacle. Initial positions of the robot and the obstacle were respectively $(0.0\text{m}, 0.0\text{m})$ and $(5.0\text{m}, 0.3\text{m})$. The goal position of the robot was $(7.0\text{m}, 0.0\text{m})$. In the situation in Figure 10, the higher limit of robot speed was $v_{max} = 0.5\text{m/s}$, the lower one was $v_{min} = 0.0\text{m/s}$. The higher limit of acceleration of the robot was $a_r = 1.0\text{m/s}^2$. The simulations have done with two different obstacle speed $v_o = 0.0, 0.5\text{m/s}$, which the direction was negative on x -axis. Figure 10(a) and (b) show respectively the trajectory of the robot that the PMF for obstacle avoidance is generated without considering the relative velocity and that with considering the relative velocity, when $v_o = 0.0\text{m/s}$.

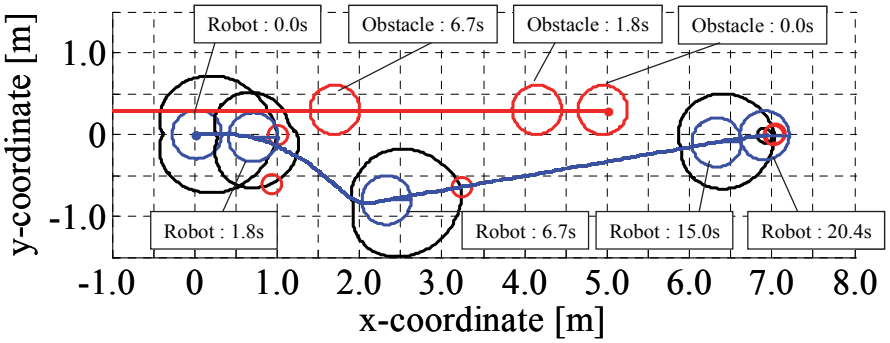
As shown in Figure 10(a), the robot got close to the obstacle because the relative velocity was not considered. On the other hand, in Figure 10(b), the robot succeeded at the early starting of avoidance behaviour due to generating PMF by use of predicted information based on the relative velocity. In addition to the situation as in Figure 10(b), in Figure 11, the higher limit of the robot speed has been changed: $v_{max} = 0.8\text{m/s}$. Even if the robot speed becomes higher, the robot succeeded in efficient avoidance.

In Figure 12(a) and (b), the trajectories of the robot, with PMF considering the relative velocity and not considering that, when the obstacle speed $v_o = 0.5\text{m/s}$. In (a), due to delay of starting avoidance behaviour, the robot collided with the obstacle. On the other hand, in (b), due to the early starting of the avoidance behaviour, the robot succeeded at the obstacle avoidance.

From these simulation results, it is confirmed that by an associating the PMF for avoidance with the relative velocity, higher the obstacle speed is, earlier the timing of the avoidance behaviour of the robot is, therefore the ability of avoiding obstacle can be enhanced.
In order to verify the effectiveness of the proposed method, simplified experiments were also carried out.
In this experiments as shown in Fig.13 A ball was supposed to be a moving obstacle and is rolled toward the robot. The robot recognized the environment by the omni-directional camera. A position of a goal and that of an obstacle relative to the robot were calculated by extracting features based on objects' colours. The robot size is L 0.4 × W 0.4 × H 0.8m and the ball diameter is 0.2m. The radius of robot and obstacle were supposed to be 0.3m and 0.1m respectively, therefore, $R_{r,o} = 0.4$m . α was set to 1.4m when the robot used the proposed PMF which was considering relative velocity. When the robot didn't use the proposed PMF, α was set to 2.4m. γ was 0.7. ε was 1.0m. v_{max} was 0.5m/s, v_{min} was 0.0m/s. a_r was $1.0\text{m}/\text{s}^2$.



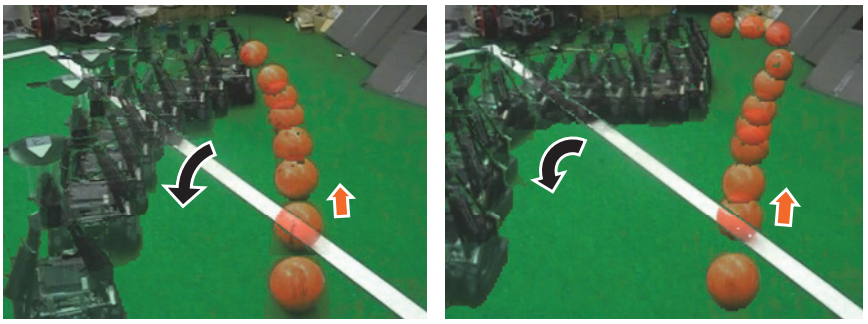
(a) not using PMF considering relative velocity



(b) using PMF considering relative velocity

Fig. 12. Simulation results of obstacle avoidance going by each other when speed of an obstacle was 0.5 m/s and of a robot was 0.5 m/s.

When the robot used the proposed PMF, which was considering relative velocity, as shown in Fig. 13 (a), it succeeded in avoiding the moving ball with smooth trajectory. On the other hand, the robot with the PMF, which was not considering relative velocity, diverged once as shown in Fig. 13 (b).



(a) (b)

Fig. 13. Trajectories of the obstacle (ball) and the robot with the PMF considering relative velocity (a) and not considering relative velocity (b)

5. Conclusion

In this paper, for the purpose of avoiding the moving obstacle safely and smoothly, design methods of the potential membership function (PMF), taking into consideration the velocity of the obstacle relative to the robot have been presented. The proposed PMF for an obstacle and PMF for a goal are unified by fuzzy inference. By defuzzification, the command velocity vector of the robot is calculated and the obstacle. Numerical simulations and simplified experiments, which assumed an obstacle avoidance of an autonomous omni-directional mobile robot, were done. As the result of the comparison between the design method of PMF using relative velocity and not using, it was confirmed that the PMF using relative velocity enhanced the ability of avoiding the moving obstacle.

6. References

Khatib,O., 1986. Real-time Obstacle Avoidance for Manipulators and Mobile Robots, In *Int. J. of Robotics Research*, vol.5, no.1, pp.90-98.

Arkin, R., C., 1989. Motor schema-based mobile robot navigation, In *Int. J. Robotics Research*, vol. 8, no. 4, pp. 92-96.

Ko, Y., N., and Lee, H., B., 1996. Avoidability Measure in Moving Obstacle Avoidance Problem and Its Use for Robot Motion Planning, In *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 96)*, pp.1296-1303.

Ge, S., S., and Cui, J., Y., 2002. Dynamic Motion Planning for Mobile Robots Using Potential Field Method, In *Autonomous Robots,* vol.13, pp.207-222.

Borenstein, J., and Koren, Y., 1991. The Vector Field Histogram Fast Obstacle Avoidance For Mobile Robots, In *IEEE Trans. on Robotics and Automation*, Vol.7, No.3, pp.278-288.

Fox, D., Burgard,W. and Thrun,S., 1997. The Dynamic Window Approach to Collision Avoidance, In *IEEE Robotics and Automation*, Vol. 4, No. 1, pp.1-23.

Tsuzaki, R., Yoshida , K., 2003. Motion Control Based on Fuzzy Potential Method for Autonomous Mobile Robot with Omnidirectional Vision. In *Journal of the Robotics Society of Japan* . vol.21, no.6, pp.656-662.

Otsuka, F., Fujii, H., Yoshida, K.,,2005. Action Control Based on Extended FPM for an Autonomous Mobile Robot. *The 26th Annual Conference of the Robotics Society of Japan*.

28

Numerical Simulation Research and Use of The Steel Sheet Pile Supporting Structure in Vertical Excavation

Qingzhi Yan and Xiangzhen Yan
The Institute of Storage-Transportation & Architectural Engineering,
China University of Petroleum,
China

1. Introduction

Deep foundation pit construction is frequently limited by construction site, which is usually through vertical excavation instead of slope excavation. The steel sheet pile supporting structure is a special supporting method for vertical excavation widely applied in the base of high-rise buildings, underground railway, municipal engineering and hydraulic engineering, with better economic benefits and environment effects. The supporting structure is used to protect the foundation pit from sliding in the process of excavation of the foundation pit.

2. Survey of deep foundation pit bracing

In the 1930s, Terzaghi and other scholars studied excavation engineering in geotechnical engineering problems.
Early in the 20th century, steel plate piles were first developed in Europe. In 1903, in Japan, the steel plate was used for the first time in Mitsui library of retaining piles construction and found with special performance. Hence steel sheet pile were used in Japan for large and repair works after 1923 Great Kanto earthquake. In 1931, steel sheet pile was rapidly developed and produced in Japan, and then it also witnessed positive development, application and dissemination in Europe, South Korea, the United States and other countries.

3. Steel sheet pile supporting overview

3.1 Concepts of steel sheet pile supporting

The retaining part refers to the excavation, in order to ensure pit wall not to collapse, protect the security of underground structures and surrounding environment. Steel sheet pile supporting of foundation pit supporting structure is used in a type of steel sheet pile (general concurrently waterproof curtain) into the soil, set up necessary support or pull anchor, to resist earth pressure and water pressure, to keep the stability of strata, to maintain the balance of deep excavation and guarantee the smooth construction of.

3.2 Steel foundation pit supporting structure stability analysis

The purpose of the foundation pit supporting of excavation is to ensure stability in construction. Pit instability and failure mainly involves two kinds of problems.

For the first kind of problem of foundation pit stability, the supporting structure (including supports wall body, interior support, anchor, etc.) of the internal force (mainly moment) and displacement as the research object, is to determine the foundation pit supporting structure satisfy the intensity and rigidity requirement the stability problem. The second kind of foundation pit instability issues is mainly for supporting structure static equilibrium conditions, the main problems in the research of the supporting structure static condition are to satisfy the static equilibrium condition stability problem.

4. Numerical analysis of steel sheet pile foundation pit supporting structure stability

4.1 Introduction of finite element method in the application of foundation excavation stability analysis

There are two kinds of steel sheet pile supporting finite element analysis methods, elastic foundation beam method and plane strain finite element method. The elastic foundation beam method is the foundation pit medial soil as soil spring, regardless of the pile and soil contact. The property of slippage of interface between soil and steel sheet pile can be calculated in the process of constructing surface subsidence and the of foundation pit bottom uplift. Therefore, in the stability analysis of steel sheet pile supporting, we should choose the plane strain finite element analysis.

4.2 Steel sheet pile plane strain finite element analysis assumptions

Steel sheet pile supporting structure relies on steel sheet pile, soil anchor (or support) and soil under the common function of passive earth pressure, water pressure, earthquakes and other load. The finite element method to analyze steel sheet pile supporting structure, involves stem cell, beam element, soil unit, and contact elements and so on many kinds of element types. Considering the nonlinear problem, the situation is more complicated.

The length of the deep foundation pit bracing is usually long, plane strain finite element method along the length direction is often applied to take unit length calculation.

4.3 Nonlinear finite element μ

There will be through elastic non-linear model Duncan - Zhang (D - C) hyperbolic model to calculate the displacement and stress field of sliding body, and then through the form of table definition unit Mohr -Column criterion, and for each unit of the yield of state judge which satisfy Mohr -Column criterion the slope plastic differentiate code layout.

E_t the tangent modulus of elasticity

Soil tri-axial test, σ_3 remain unchanged, exert partial stress, point draw $(\sigma_1-\sigma_3)/\varepsilon_\alpha$ curve (see figure 1), Connor, etc, found that $(\sigma_1-\sigma_3)$ hyperbola can be used to fit these curves. For one σ_3, the relations of $(\sigma_1-\sigma_3)/\varepsilon_\alpha$ can be expressed as

$$\sigma_1 - \sigma_3 = \frac{\varepsilon_\alpha}{a + b\varepsilon_\alpha} \tag{1}$$

The Eq. (1) also takes the form

$$a + b\varepsilon_\alpha = \frac{\varepsilon_\alpha}{\sigma_1 - \sigma_3} \tag{2}$$

In which $(\sigma_1-\sigma_3)$ refers to partial stress, ε_α axial strain, and a, b significance for test constants, the intercept and slope of curve in figure 2, are presented respectively as follows:

$$\begin{cases} a = \dfrac{1}{E_i} \\ b = \dfrac{1}{(\sigma_1 - \sigma_3)_u} \end{cases} \tag{3}$$

Here $(\sigma_1-\sigma_3)_u$ refers to the value of $(\sigma_1-\sigma_3)$ when $\varepsilon_\alpha \to \infty$, $(\sigma_1-\sigma_3)$ refers to progressive values (as shown in figure 3), E_i refers to the initial tangent modulus, obtains

$$E_i = kp_\alpha \left(\frac{\sigma_3}{p_\alpha} \right)^n \tag{4}$$

The P_a refers to atmospheric pressure, generally take 100kPa, K and n refers to trials have certain parameters, its significance see figure 1, available by the calculation

$$E_t = \frac{\partial(\sigma_1 - \sigma_3)}{\partial \varepsilon_\alpha} \tag{5}$$

Then $\varepsilon_\alpha = a(\sigma_1-\sigma_3)/[1-b(\sigma_1-\sigma_3)]$, Will ε_α generation into Eq.(5), obtains

$$E_t = \frac{1}{a}\left[1 - b(\sigma_1 - \sigma_3)\right]^2 = E_i\left[1 - R_f s\right]^2 \tag{6}$$

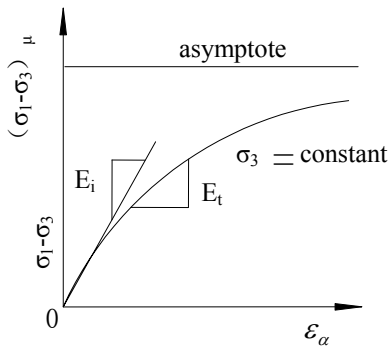


Fig. 1. The relation between $(\sigma_1-\sigma_3)$ and ε_α

In which S refers to stress level, $s=(\sigma_1-\sigma_3)/(\sigma_1-\sigma_3)_f$, and R_f refers to the destruction of the stress ratio, $R_f=(\sigma_1-\sigma_3)_f/(\sigma_1-\sigma_3)_u$, less than 1, generally in between (0.75-1), and failing stress $(\sigma_1-\sigma_3)_f$ .With relevant consolidation pressure σ_3, as shown in figure 4, obtains

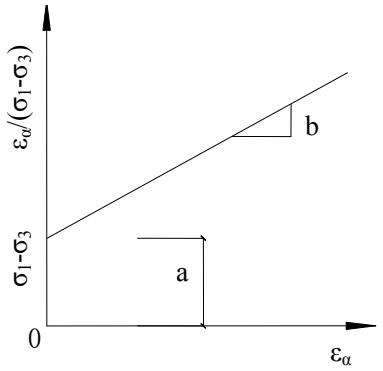


Fig. 2. The relation between $\varepsilon_\alpha/(\sigma_1-\sigma_3)$ and ε_α

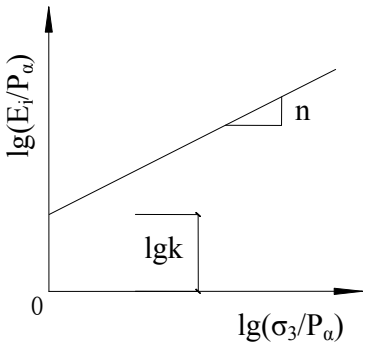


Fig. 3. The relation between $\lg(E_i/P_a)$ and $\lg(\sigma_3/P_a)$

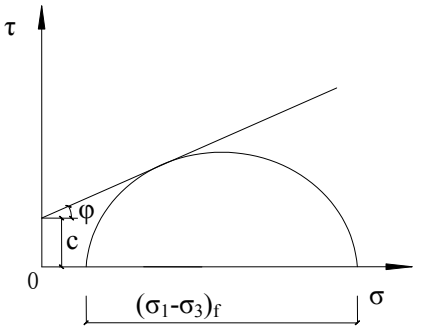


Fig. 4. Limit Mohr's circle

$$(\sigma_1-\sigma_2)_f=\frac{2(\cos\varphi+\sigma_3\sin\varphi)}{1-\sin\varphi} \tag{7}$$

Substituting Eq.(6)into Eq.(7), it follows that

$$E_t = \left[1 - R_f \frac{(1-\sin\varphi)(\sigma_1 - \sigma_3)}{2(c\cos\varphi + \sigma_3 \sin\varphi)}\right]^2 kp_\alpha \left(\frac{\sigma_3}{p_\alpha}\right)^n \tag{8}$$

Here E_t refers to the tangent modulus of elasticity.

(2) Tangent of Poisson ratio

Kulhway and Dunkcan think conventional tri-axial test measured with the curvilinear relationship between ε_α and $\varepsilon_\gamma$may adopt hyperbola to fitting (see figure 5), and the curvilinear relationship between $-\varepsilon_\gamma/\varepsilon_\alpha$ and $-\varepsilon_\gamma$for a straight line (see figure 6).Its interception is *f*, slope is *D*, then

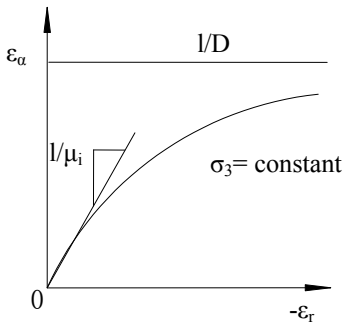


Fig. 5. The relation ε_α and $-\varepsilon_r$

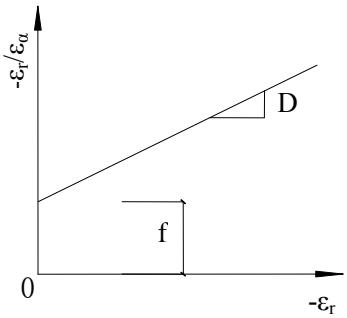


Fig. 6. The relation $-\varepsilon_r/\varepsilon_\alpha$ and $-\varepsilon_r$

$$\frac{-\varepsilon_r}{\varepsilon_\alpha} = f + (-\varepsilon_r) \tag{9}$$

Same is

$$-\varepsilon_r = \frac{f\varepsilon_\alpha}{1 - D\varepsilon_\alpha} \tag{10}$$

Because of the lateral pressure increment is zero, next Eq. is usable for Poisson ratio

$$\mu = \frac{-\Delta\varepsilon_r}{\Delta\varepsilon_\alpha} = \frac{\partial(-\varepsilon_r)}{(\varepsilon_\alpha)} \tag{11}$$

Substituting Eq. (10) into Eq. (11), and, will $\varepsilon_\alpha = a(\sigma_1-\sigma_3)/[1-b(\sigma_1-\sigma_3)]$ into Eq. (11), it follows that

$$\mu = \frac{f}{(1-A)^2} \tag{12}$$

In which A refers $A = \dfrac{D(\sigma_1-\sigma_3)}{kp_\alpha\left(\dfrac{\sigma_3}{p_\alpha}\right)^n\left[1-R_f\dfrac{(1-\sin\varphi)(\sigma_1-\sigma_3)}{2(c\cos\varphi+\sigma_3\sin\varphi)}\right]^2}$

By Eq.(10), when $-\varepsilon_r\to\infty$, ε_α is progressive values, then D the same of $D=1/\varepsilon_\alpha$, when the reciprocal of the,

$$f = \left(\frac{-\varepsilon_r}{\varepsilon_\alpha}\right)_{\varepsilon\to 0} = \mu_i \tag{13}$$

In which μ_i refers to the initial tangent Poisson ratio.

For different σ_3 have different μ_i values, in half logarithmic coordinate system μ_i in the relationship with σ_3/P_a is a linear curve approximation (see figure 4-7), its intercept is G_γ, slope is F, therefore

$$\mu_i = G - F\lg\left(\frac{\sigma_3}{p_\alpha}\right) \tag{14}$$

So the tangent Poisson's ratio obtain by Eq. (15) solution

$$\mu_t = \frac{G - F\lg\left(\dfrac{\sigma_3}{p_\alpha}\right)}{(1-A)^2} \tag{15}$$

(3) Resilience modulus

Eq. (8) is elastic modulus in loading cases. For unloading cases, elastic modulus E_{ur} obtained by unloading test. In figure 8, OA is stress-strain curve of loading status, the slope is E_t, AB stress-strain curve of unloading status, the slope is E_{ur}. Obviously, $E_{ur} > E_t$. Duncan etc have supposed E_{ur} not changes with $(\sigma_1-\sigma_3)$, only changes with σ_3, it is concluded that the relation curve of $\lg(E_{ur}/P_a)\sim\lg(\sigma_3/P_a)$ is a straight line (see figure 9), its intercept is $\lg E_{ur}$, slope is n. Generally speaking, n basically consistent (elastic modulus as tangent modulus) in loading cases, while $K_{ur}=(1.2\sim3.0)K$. For close-grained sand and hard clays $K_{ur}=1.2K$, for loose sand and soft soil $K_{ur}=3.0K$, general soil K_{ur} value is between them. Resilience modulus can be calculated by next Eq.

$$E_{ur} = k_{ur}p_\alpha\lg\left(\frac{\sigma_3}{p_\alpha}\right)^n \tag{16}$$

The finite element calculation to give a standard K_{ur} under specific circumstance, this is actually a rough yield criterion. Can use such standard: when $(\sigma_1-\sigma_3)<(\sigma_1-\sigma_3)_0$, and $S<S_0$ use K_{ur}, otherwise use K_t. Here $(\sigma_1-\sigma_3)_0$ is the biggest variable stress in history has achieved. S_0 is the maximum stress historical levels history once reached.

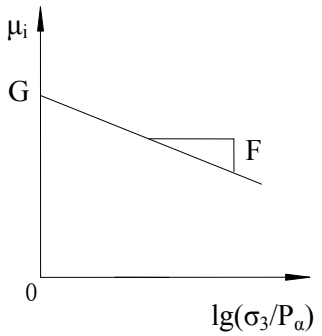


Fig. 7. $\mu_i \sim \lg(\sigma_3/P_a)$ relation curve

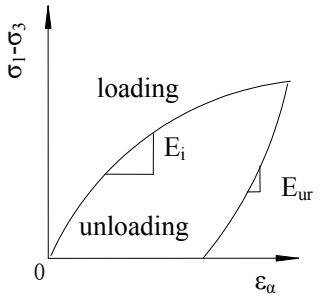


Fig. 8. Loading and unloading relation curve

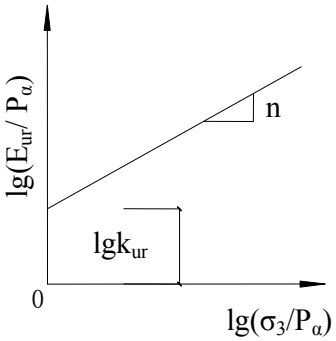


Fig. 9. $\lg(E_{ur}/P_a) \sim \lg(\sigma_3/P_a)$ relation curve

(4)Property indexes

Adopting Duncan - Zhang (D - C) elastic non-linear model of E-μ, the stress and strain of soil is of nonlinear properties. The concentrated reflection is summed up in Eq. (8) and (15) and (16), all parameters contains in Eq. are to be determined by conventional tri-axial test. Nonlinear plane strain finite element numerical analysis of steel sheet pile needs to determine that the physical and mechanical indexes such as γ, ϕ and c, and the elastic modulus index R_f, K and n, and Poisson's ratio index G, F and D, spring-back modulus index E_{ur} and m, all 11 parameters. Zhujun Gao, Zongze Yin put forward a optimal method of determining soil constitutive model parameters, put absolute error between the tri-axial test measurement of the stress-strain relationship curves and soil constitutive model of curve as objective function, to determine the soil constitutive model parameters.

2) Construction process simulation

In the normal soil excavation and backfilling calculation, it is often assumed that they will be instantaneously completed. But the construction process of excavation and backfilling could have influence on internal force and deformation produce of steel sheet pile structure should not be neglected, it is necessary to simulate the influence of foundation pit excavation and construction of backfill process. The excavation and backfilling is actually a hierarchical loading process, and backfill belongs to primary loading, the excavation is the unloading or an unloading reloading process. It has been proved in practice that incremental method can be used to simulate construction process. The key issue is that the proposed method can simulate the construction process adopted and load history, soil constitutive relationship should reflect the different stress path of soils, and the input parameters in calculation should be in accordance with the actual calculation, and the boundary conditions of should be as far as possible and reasonable.

Construction process simulation mainly includes a step-by-step excavation, exertion and removal of soil anchor and support, etc. The finite element equation the simulation of mechanical behavior of the different construction stage of can be written as

$$\left\{\left[K_0\right]+\left[\Delta K_i\right]\right\}\left\{\Delta\delta_i\right\}=\left\{\Delta F_{ir}\right\}+\left\{\Delta F_{in}\right\}\left(i=1,m\right) \tag{17}$$

In which m refers to total number of construction steps, K_0 the initial general strength degree matrix before excavation, ΔK_i incremental of stiffness of geotechnical body and the support structure in the construction process, the value is element stiffness of geotechnical body and supporting structure settled or dismantled, $\{\Delta F_{ir}\}$ refers released boundary incremental node force matrix of produced by excavation, determined in first excavation by geotechnical body self-weight, groundwater load, ground overloading, in followed excavation steps determined by the current stress state; $\{\Delta F_{in}\}$ refers the increased node force matrix in the construction process, $\{\Delta\delta_i\}$ refers the incremental displacement matrix of any construction stages. $\{\Delta\delta_i\}$ refers displacement, ε_i strain and stress σ_i matrix in ith step in the construction process

$$\left\{\delta_i\right\}=\sum_{k=1}^{i}\left\{\Delta\delta_i\right\},\ \left\{\varepsilon_i\right\}=\sum_{k=1}^{i}\left\{\Delta\varepsilon_i\right\},\ \left\{\sigma_i\right\}=\left\{\sigma_0\right\}+\sum_{k=1}^{i}\left\{\Delta\sigma_i\right\} \tag{18}$$

Generally speaking, load steps divided more, the analytical results is more close to reality. The more accurate the loading steps, the results are more close to reality. However, the fact that how many exact steps are divided is determined by the purpose of the relevant analysis. If the main purpose is to understand the deformation and stress of the foundation

under the backfilling, the load steps can be fewer; If you need to understand the displacement and stress of backfilling itself, there can be more the load steps be added. Typical division into 10-12 may obtain satisfactory results. The typical process can be divided into 10- 12 steps to obtain satisfactory results. The amount of increment of the displacement in backfilling itself is very sensitive needs to be judged according to the measured data and the experience of the engineering representative.

There is another thing to be noted and is of vital importance is the unit application in the construction process simulation. All units generated in the pre-treatment, and then put the soil anchor or support unit killed, then put soil anchor or support unit killed after every excavation step, brought each unit of support or anchor activation after a layer each construction completed. Attention should be paid to reactivate or kill units in every step.

4.4 Unit types and strength degree matrix.

Steel sheet pile of plane strain finite element analysis will involves four units, which is described as follows: the soil unit, contact elements, beam element and line unit. The following is a simple introduction of various units unit strength degree matrix.

1. Beam element strength degree matrix

The plane beam element is shown in figure 4-10. Classic beam bending theory generally corresponding to the smaller depth-span beam, it meets normal sections the assumption. The strength degree matrix under local coordinate is

$$[K]^e = \begin{bmatrix} \frac{AE}{l} & & & & & \\ 0 & \frac{12EI}{l^3} & & & & \\ 0 & \frac{-6EI}{l^2} & \frac{4EI}{l} & & & \\ -\frac{AE}{l} & 0 & 0 & \frac{AE}{l} & & \\ 0 & \frac{12EI}{l^3} & \frac{6EI}{l^2} & 0 & \frac{12EI}{l^3} & \\ 0 & \frac{-6EI}{l^2} & \frac{2EI}{l} & 0 & \frac{6EI}{l^2} & \frac{4EI}{l} \end{bmatrix} \tag{19}$$

In which A refers to cross-sectional area of beam element, E elastic modulus of beam element materials, I moment of inertia of beam element cross-sectional inertia, and L refers to the beam element length.

Eq. (19) is the unit strength degree matrix of beam element in local coordinate, if θ is the angle between local coordinate system and integral coordinate system, the unit strength degree matrix of beam element under integral coordinate system is:

$$\left[\bar{K}\right]^e = [T]^T [K]^e [T] \tag{20}$$

In which $[K]^e$ is the unit strength degree matrix of beam element in the local coordinate, $\left[\bar{K}\right]^e$ is the unit strength degree matrix of beam element in the global coordinate system of units, $[T]$ is the coordinate transformation matrix of strength degree matrix. The calculation Eq. is

$$[T]=\begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos\theta & \sin\theta & 0 \\ 0 & 0 & 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

2. Line element strength degree matrix

Planar line element as shown in figure 4-11, the line element strength degree matrix in local coordinate is

$$[K]^e=\frac{EA}{l}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{22}$$

In which A refers to cross-sectional area of line element, E elastic modulus of line element materials of and L refers to unit length.

Line element unit strength degree matrix in local coordinate and integral coordinate system conversion also can process according (20), calculation is

$$[T]=\begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{bmatrix} \tag{23}$$

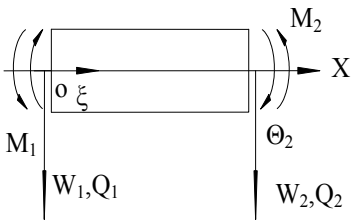


Fig. 10.

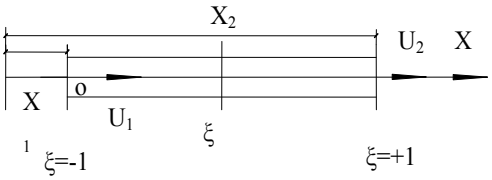


Fig. 11.

The support and soil anchor in three-d is discontinuous structures, how in the analysis of plane strain finite element simulation, which is an important issue of the steel sheet pile analysis. In situ and Clough study, once pointed out that exact a treatment is the board wall of strut (soil anchor) axial stiffness press unit length added. To support (soil anchors), it in the plane strain finite element analysis of axial stiffness by the support of the actual axial stiffness divided by supporting spacing.

3. Soil unit strength degree matrix

Soil element adopted the plane four nodes iso-parametric unit, the unit strength degree matrix calculation method is of the same as strength degree matrix of usual iso-parametric unit, just instead E and μ by E or E_{ur} and μ_t in the calculation, and the unit strength degree matrix of soil to be calculated by:

$$[K]^e = t\int_{-1}^{1}\int_{-1}^{1}[B]^T[D][B]|J|d\xi d\eta \tag{24}$$

In Eq. (24), $[B]=[B_1 \quad B_2 \quad B_3 \quad B_4]$ and $[B_i]=\begin{bmatrix} \frac{\partial N_i}{\partial x} & 0 \\ 0 & \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} \end{bmatrix}$;

$$[D]=\frac{E}{1-\mu^2}\begin{bmatrix} 1 & \mu & 0 \\ \mu & 0 & 0 \\ 0 & 0 & \frac{1-\mu}{2} \end{bmatrix},\ .\ |J|=\begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{vmatrix}$$

4. Contact surface elements strength degree matrix

In contact surface of steel sheet pile and soil, there is a great difference in material property. In some conditions, slippage or craze could be generated possibly on the contact interface, so it is suggested to set contact surface elements to simulate the interaction between steel sheet pile and soil.

Goodman and others propose the joints units, commonly used in contact elements. This unit is no thickness four nodes unit, as shown in figure 12, the idea is to have countless normal and tangential tiny spring associated between two contact interfaces, and stress and relative displacement relations is described as:

$$\begin{Bmatrix} \sigma \\ \tau \end{Bmatrix}=\begin{bmatrix} k_n & 0 \\ 0 & k_s \end{bmatrix}\begin{Bmatrix} w_s \\ w_n \end{Bmatrix} \tag{25}$$

In which w_s refers to tangential displacement, and w_n normal relative displacement, K_s tangential spring coefficient, and K_n normal spring coefficient.The unit strength degree matrix is available under local coordinate.

$$[k]=\frac{L}{6}\begin{bmatrix} 2k_s & & & & & & & \\ 0 & 2k_n & & & & & & \\ k_s & 0 & 2k_s & & & & & \\ 0 & k_n & 0 & 2k_n & & & & \\ -k_s & 0 & -2k_s & 0 & 2k_s & & & \\ 0 & -k_n & 0 & -2k_n & 0 & 2k_n & & \\ -2k_s & 0 & -k_s & 0 & k_s & 0 & 2k_s & \\ 0 & -2k_n & 0 & -k_n & 0 & k_n & 0 & 2k_n \end{bmatrix} \tag{26}$$

To determine K_s by direct shear tests, to determine point remit $\tau\sim w_s$ relation curves by hyperbola assumptions:

$$k_s=\left(1-\frac{R_f\tau}{\sigma_n tg\delta}\right)^2 k_1\gamma_w\left(\frac{\sigma_n}{p_\alpha}\right)^n \tag{27}$$

In which K_i, n and R_f, refers to as the test parameters, δ refers to the friction angle between soil and structural materials, δ refers to water volume weight. Elasticity coefficient is relevant with stress state, in response contact surface pulled open, give K_n a small value, or take a big value.

By coordinate transformation, the unit strength degree matrix [K] is obtianed in the global coordinate system

$$[K]=[Q]^{-1}[k][Q] \tag{28}$$

[Q] is coordinate transformation matrix

$$[Q]=\begin{bmatrix} \alpha & & & \\ & \alpha & & \\ & & \alpha & \\ & & & \alpha \end{bmatrix},\ [\alpha]=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

5 The numerical analysis of engineering examples

5.1 Project profile

The engineering site stratum structure and the causes of stratum structure are very simple, and the variation in thickness of the stratum is low. The average ground elevation is 5.50m, and the depth of the foundation pit is 10.00m. The area of engineering site is 3745 square meters, and the area of supporting structures is 4180 Square meters.

5.2 Geological conditions of the site

There are nine geological stratums in the depth range of 35m according to field exploration and comprehensive analysis of laboratory test, by the litho logy of the geological stratums are composed mainly of plain fill, silt, silt-clay and mealy sand.

5.3 Design decision of the supporting structure

The standard section of foundation pit supporting structure is shown in Figure 12. The model of steel sheet pile is H formed steel, whose width is 486mm, and the depth of section is 420mm, and elastic section modulus is 3.12×106mm3. The length of pile is 16.00m. The elevation of pile cap is +5.50m, and toe is -10.50m. The location's elevations of two layers of pre-stressed strands anchor are +1.50 m and -1.50m, and the angle between soil anchor and horizontal plane is 15. The first layer of soil anchor is made by 3 bunches of steel strands and each bunch is consisted of 7 lines. The freedom length is 5.00m, and the designed length of anchoring section is 10.00m; The second layer of soil anchor's component is similar to the first one. The freedom length is 5.00m, and the designed length of anchoring section is 10.00m. The length between soil anchors is 2.50m. Because it is very difficult to control the location of steel sheet pile exactly and steel sheet piles can not bond join each other very well, it needs to set a row of DJM piles acting as curtain for cutting off water, for the steel sheet pile has the effect of cutting off water. The length of DJM pile is 13.00m, diameter is 0.50m, cement-mixed ratio is 15%, and the length between piles is 0.40m.

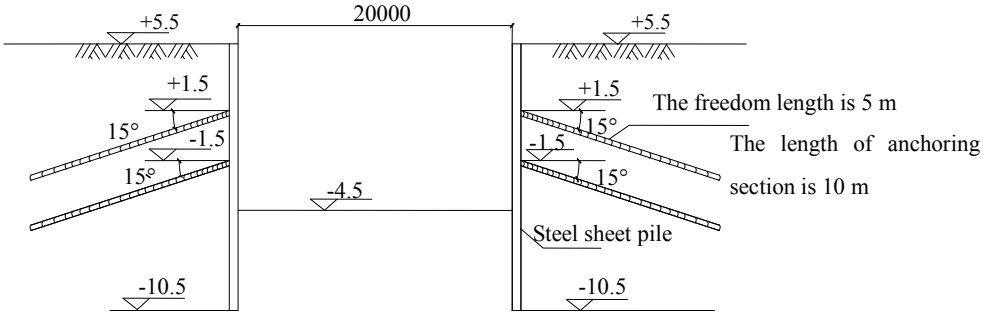


Fig. 12. Foundation pit support sectional drawing

5.4 Decision of computational model

It takes half of model to carry out analogue computation, because the structure is symmetrical. The width of that the foundation pit can effect is 3-4 times as the width of excavation and the influence depth which is 2-4 times as [a, b] according to the project experience. In this project, the width of excavation is 20.00m, the depth of excavation is 10.00m, the influence width is 30.00m, and the influence depth is 35.00m.

The finite element analysis method of steel sheet pile is applied to simulate construction of foundation pit supporting structure. It uses the 2—dimensional, 4—node iso-parametric element and D-C constitutive model to simulate soil mass. It uses 2—dimensional 2—node beam element to simulate steel sheet pile. It uses bar element to simulate the soil anchor, supporting structure and bottom brace. It takes equivalent elastic modulus as elastic modulus. It takes calculation under ten-stage loading, and there are ten increment steps in each stage (load sub-step). Some special construction stages of the first and the second layer of soil anchors are applied, including pre-construction, after construction and construction completion, to research horizontal displacement and settlement of steel plate pile, moment of steel sheet pile and axial force of soil anchor.

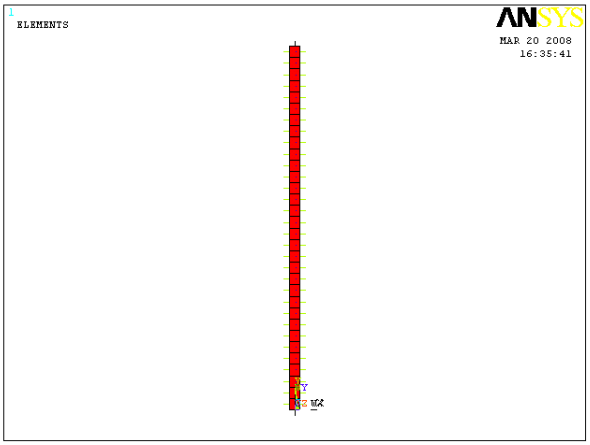


Fig. 13. Setting contact surface element between steel sheet pile and soil

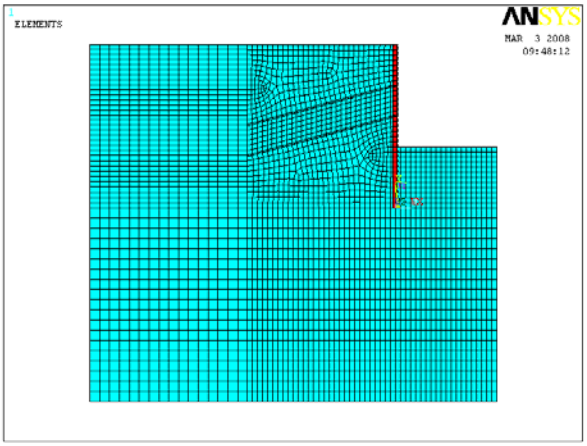


Fig. 14. Finite element analysis mesh of steel sheet pile foundation pit support schematic diagram

5.5 Results and analysis (horizontal displacement and settlement)

1. Horizontal displacement

From the Figure 15, the horizontal displacement curve calculated with the finite element method is similar to the measured displacement curve by and large. While the calculated displacement value is bigger than measured displacement value, the calculated maximum displacement value is 0.041m, and the depth is about 8.00m. The measured displacement value is 0.035m, and the depth is about 8.00m too.

2. Settlement

The settlement curves of ground after steel sheet pile in the different stages are shown in Figure16. The settlement will increase with the increase of excavation depth, and the location

in which the maximum settlement happened is as long as excavation depth away from piles. Pre-stress will increase and maximum settlement points will be far away from pit wall after we set the first row of anchors. With the excavation of the foundation pit, the settlement will go on increasing, and the maximum settlement points are still away from the pit wall owing to the effect of the last layer of anchors. After setting the second row of anchors, there are few effects on the settlement. The reason may be that the anchoring section of the second row of anchors is too long. With the increase of excavation depth, ground settlement will increase again.

3. Uplift of foundation pit

From comprehensive analysis of these three excavation stages, the vertical settlement values of foundation pit are 6.2cm, 10.1cm and 12.3cm, and the settlement value is in the range of safety. The measured values are 5.5cm, 11.1cm and 13.8cm.

5.6 Internal force

1. Moment of steel sheet pile

The moment curves of steel sheet pile in the different stages are shown in Figure 17. The moment of pile shaft above the first layer of soil anchors is caused by active earth pressure. The moment remains about the same in the course of excavation, because the anchor can be seen as pivot and the part above pivot can be seen as cantilever structure. The cantilever structure is almost not changed in the course of excavation, because the steel sheet pile has no negative displacement, and this part has always been bearing consistent active earth pressure. After the first anchorage starts being constructed and exerts pre-stress, negative moment decreases and positive moment increases, but the absolute value of moment decreases, so soil anchor not only controls the horizontal displacement of the steel sheet pile well, but also improves its condition of forces. After the second excavation, excavation face goes down, earth pressure and moment increase. After the second anchorage starts construction and exerts pre-stress, negative moment increases, positive moment decreases and the maximum absolute value of moment decreases. So the second row of soil anchors control the horizontal displacement of the steel sheet pile and improve the condition of

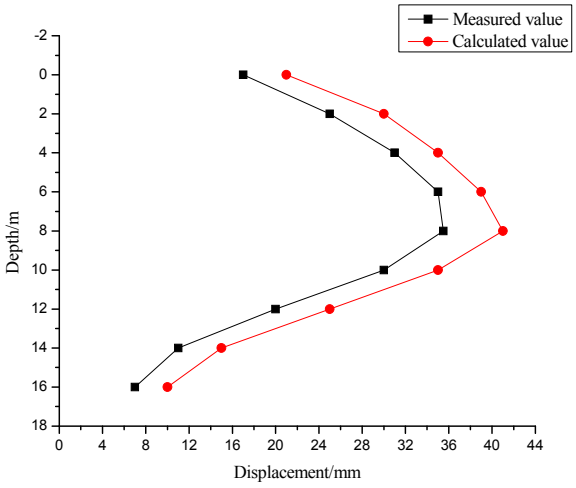


Fig. 15. Final calculated and measured values of horizontal displacement curve graph

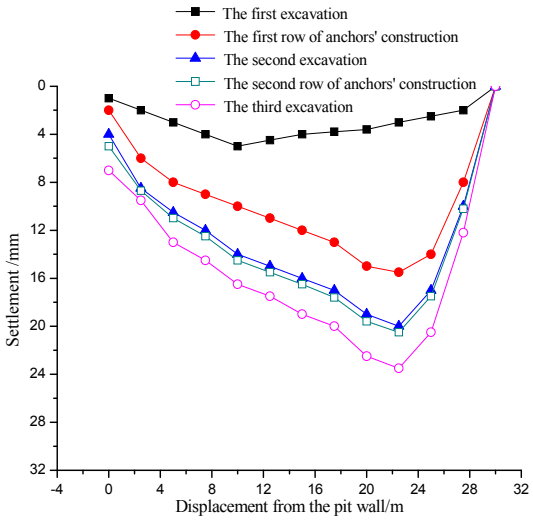


Fig. 16. Ground settlement behind pile curve graph

forces too. The maximum moment of the steel sheet pile happens in the stage of the last sub-step, and the maximum bending stress satisfies with the request of strength.

2. Anchor's shaft force

From chart 18, the shaft force of every layer has different levels of increase with the carrying out of excavation. After the consummation of the second excavation, the first layer of anchorage's shaft force has greatly increased. After the second anchorage starts being constructed and exerts pre-stress, the upper anchor will emerge stress relaxation, and the shaft force has a little decrease. After the third excavation, the shaft force of two layers of anchorage will bear new load, so it increases. The axial force of the first layer of anchors which is calculated by FEM increases from 151.21 KN to 249.63 KN with the process of loading. It is similar to the measured data. And because the shaft force of the second goes down to 180.34KN and then increases to 238.95KN, it is similar to the measured data 238.15KN. Looking at the second layer of anchors, the shaft force increases from 149.25KN to 264.87KN, and it's also similar to the measured data 263.22KN.

6. Conclusion

6.1 Conclusion

1. This paper mainly studies on the models and mechanism of steel sheet pile, and proposes two kinds of instability problems about the steel sheet pile: First, the supporting structure has not enough strength or stiffness to support the load and there are several destruction forms including support buckling, pull-anchor damage, excessive deformation of the supporting structure and bending failure. The second problem is the soil instability of the foundation pit.

   The forms of instability include sliding of foundation pit, subversion of the supporting structure, kick damage of the supporting structure, uplift instability of the foundation pit soil, leakage instability of foundation pit (piping or drifting sand) and heavy-piping instability of the foundation pit soil. The mechanics method is applied to obtain the

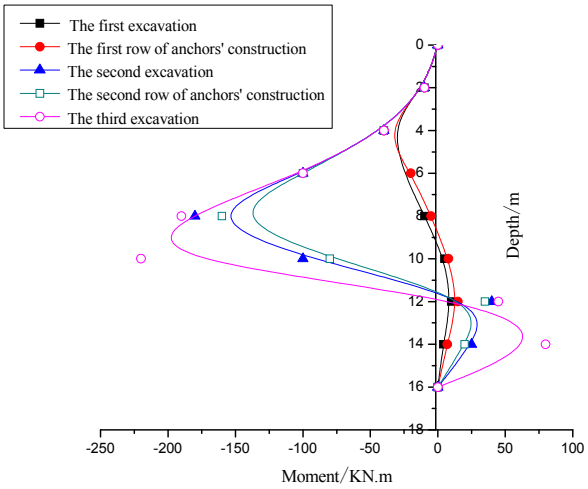


Fig. 17. Moment of steel sheet pile curve graph

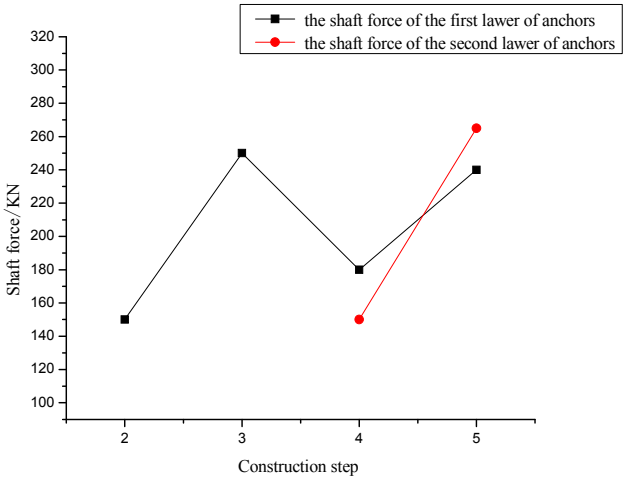


Fig. 18. Axial force of soil anchor curve

code formula from a reasonable discussion and systematical analysis. In the first instability problem, it uses the equivalent beam method and "m" method of elastic foundation beam methods to obtain the conclusion that the finite element method is a more ideal stability analysis method which we can use to deal with the strength problems and deformation problems, because the equivalent beam method does not involve the structure's deformation and "m" method does not involve the strength problems of soil.

2. In the second type of instability problems, according to the different steel sheet pile supporting basic form, and put forward different form steel sheet pile foundation pit supporting overall sliding stability analysis superposition methods, namely: Steel sheet pile supporting the stability analysis method, the influence of water that its safety coefficient calculation Eq. is:

$$K_S = \frac{\sum_{i=1}^{n}(W_i \cos\theta_i - \mu_i L_i) tg\varphi_i' + C_i' L_i}{\sum_{i=1}^{n} W_i \sin\theta_i} \tag{29}$$

3. Supported steel sheet pile retaining stability analysis method. Its safety coefficient calculation Eq. is:

$$K_S = \frac{\sum_{i=1}^{n}[(W_i \cos\theta_i + Q_H \alpha_k \sin\theta_i) tg\varphi_i + C_i L_i]}{\sum_{i=1}^{n} W_i \sin\theta_i - Q_H \alpha_k \cos\theta_i} \tag{30}$$

4. Have anchor steel sheet pile retaining stability analysis method. Its safety coefficient calculation Eq. is:

$$K_S = \frac{\sum_{i=1}^{n}(W_i \cos\theta_i tg\varphi_i + C_i L_i) + \sum_{j=1}^{m} T_j(\sin\beta_j tg\varphi_j + \cos\beta_j) / S_h}{\sum_{i=1}^{n} W_i \sin\theta_i} \tag{31}$$

The two types of instability problems combine, focus on the soil and steel sheet pile between interface slippage characteristics of plane strain finite element method and the methods of realization in software, in f foundation excavation engineering examples, this method sheet pile supporting stability analysis of practicality.

6.2 Prospects

In this paper, the stability of the steel sheet pile supporting has just made some preliminary studies in numerical simulation analysis, due to the limited conditions of the foundation pit supporting soil constitutive model parameters, the Duncan a (D - C) soil constitutive model parameters are given access to relevant information, resulting in the numerical calculation results there are certain differences, due to the complexity of soil properties of materials, the parameter value is often very difficult to grasp, should pass tests, and to determine the soil constitutive model parameters are optimized in the analysis; and this paper a symmetric model is used to take the half, this analogy is not quite reasonably, because even before the soil level of excavation isotropic, but often undergone asymmetry of foundation pit excavation, it will cause the whole system of

foundation pit of stress and deformation fields of asymmetry. Therefore, there is still large room for further improvement in using numerical simulation to analyze the stability of steel sheet pile supporting.

7. References

HuangQiang , 1995, Deep foundation pit bracing engineering design technology, *China building material industry publishing house*, Beijing, pp. 1-8

Zhonghan Chen, 1999,&Liping Cheng, deep foundation pit engineering, China Machine Press, house,Beijing, pp. 210-238 Xinxin Zhang, 2006, *Steel sheet pile rolling process of groove design and finite element simulation*, Northeastern University,Senyang ,pp. 3-4

Yuxu,2005, Steel sheet pile support in Xiamen soft soil area applied research [D].*Tongji University*,Shanghai, pp. 3-4

Juyun Yuan, Jingpei Li, Xiaoming Lou,2001, foundation engineering design principle [M]. *Tongji University Press*, Shanghai, pp. 128-141

Jianghua Fang,Yuanbing Chen, 2003, Deep foundation pit supporting technology overview[J], *Urumchi*, pp 28-30

Huangqiang, 1999, Building foundation pit supporting technical code for applications handbook[M]. *China Architecture and Building Press*, Beijing, pp. 196-264

Zhaoquan Zhu, 2004, Soil anchor steel sheet pile structure calculation analysis research [D], *Hohai University*,Nanjing, pp. 30-34

H. H. Vaziri. Theory and application of an eficient computer program for analysis of flexible earth-retaining structures [J].*Computers& structures*, 1995, V56 (l)pp.177-187

Nakakyama, M Beaudoin, B. B. A novel technique determining bond strength developed between cement paste and steel [J].*Cement and Concrete Research*, 1987, V 22(3)pp. 478-488

Mingyang Peng,Yanrong Peng, 2001, 02, Deep excavation pile wall supporting structure practical methods of numerical analysis, *Guangdong Civil Engineering & Construction,Guangzhou*, pp.7-12

Linde Yang,Xiangxia Zhang, 2005, 22(4),Geotechnical constitutive model research review and discuss [J]. *Journal of Hebei Institute of Architectural Scienc Technology*, Handan, pp.138-140

Jiahuan Qian, Zongze Yin, 2000, Geotechnical principle and calculation [M]. *China WaterPower Press, Beijing*, pp. 302-345

Baili Zhu,Zhujiang Shen, 1990, Computational Soil Mechanics [M]. *Shanghai Scientific and Technical Publishers*, Shanghai, pp.86-95

Jiahuan Qian, Zongze Yin, 1992, Geotechnical numerical analysis [M].*China railway publishing house*, Beijing, pp.230-241

Maocheng Wang, Shaoming, 1988, Finite element method and the basic principle and numerical analysi [M]. *Tsinghua University Press*, Beijing, pp.112-135

Xionghua Liao,Xikuei Li, 2002, 19(4),About soil –structure interaction interface mechanics behavior of the numerical simulation, [J]. *Chinese Journal of Computational Mechanics*,Dalian, pp.450-455

Zhanhua Qingxiang metal materials company 550 cubic meters blast furnace ore focal slot geotechnical investigation report [R]. 2007 pp.6-12

29

Collision Avoidance Law Using Information Amount

Seiya Ueno and Takehiro Higuchi
Yokohama National University
Japan

1. Introduction

The collision avoidance control has been one of the key technology for future transportation. Recently, many unmanned systems are developed in shapes of robots, cars, ships, aircraft, etc. In these environments, proper navigation and control systems including collision avoidance is needed. This paper is on collision avoidance control law for air vehicles under uncertain information. The control law uses information amount as one of the physical parameter for control system.

In the field of guidance, navigation, and control, collision avoidance of automated transportation system has been one of main interest of researchers. Many researches started from collision avoidance of ships (Ciletti et al., 1997) where collision avoidance has been one of the problems due to the increasing demand for the naval transportations. Wide varieties of studies on collision avoidance are treated in fields of robots (Fukuda & Kubota, 1999), cars (Hiaoka et al., 2009a, 2009b) and satellites. Some of these researches treat avoidance problems with the formation control which requires the cooperative information control (Slater et al.,2006; Stipanovic et al., 2007).

In the field of aeronautics, the Traffic alert and Collision Avoidance System (TCAS) has been one of the references for the collision avoidance. TCAS exchanges the information of aircrafts and advises the aircraft to avoid in vertical direction. For the conflicts in collision avoidance control, Frazzoli et al. (2001) have shown feasible strategy to treat the conflict problem. Gates (2009) has proposed rule-based collision avoidance control strategies for real-time online collision avoidance. Miele et al. (2010) has proposed collision avoidance control for case of abort landing with low computational load which can be calculated by on-board computer.

Conventional avoidance problems assume that all information about avoidance (intruders and environments) is certain. Therefore, control law is designed based on certain information. However in real cases, all information may not be correct and most of it is uncertain. These uncertainty of information differes by the relative position of the evader and intruder or the absolute position of intruder. There has been no research on control law to deal with uncertain information. This paper proposes control law that treats uncertain information. New parameters quantifying information amount are defined for this purpose. The proposed control law provides new performance by enabling the aircraft to obtain information and to check the certainty of the information.

Two different cases of numerical simulations are used to investigate the usage of the information amount. The first case defines the problem as the uncertainty of the information changes by the relative position of the evader and the target. The problem treats the case

where the amount of infomation changes by relative position, for example, the flight in fog or smoke. The information is clearer as the evader gets closer to the fog. These uncertainties are quantified and used as parameters for collision avoidance control law. The second case defines the problem as the uncertainty of the infomation is given as absolute position. The infomation can or cannot be obtained by the position itself, for example, the flight around urban buildings or moutains. In both cases, the information amount is obtained from focused area assigned by the user. Using the information amount, the control law is designed for safer flight of the air vehicles.

2. Information amount

In this study, amounts of information are treated as parameters for the control law. First, the focused area : S_E is treated as the region of the area that the user focuses. This area can be large if the vehicle is moving fast or very small if the vehicle is in urban area moving very slowly. The cleared area : S_C is the area where the information are certain. In the cleared area, all of the infomation is available, meaning if there is an intruder in that region, the evader can obtain all the infomation of the intruder. In the other hand, the blurred area : S_B is the area where infomation is uncertain.
From these parameters of the areas, the infomation amount is derived qunatatively as physical value to be used. One of the important factor used in this paper is information localization : I_L which is calculated from the amount of cleared area out of the focused area. The I_L is,

$$I_L \equiv S_C / S_E \tag{1}$$

and the schematic image of this areas are shown in Fig.1.
Another important factor for the information amount is information acquisition requirement : I_R which is requirement of the I_L for safe flight. The evader selects the proper amount for the vechicle to obtain. The higher the value is, more infomation have to be obtained. In the other hand, evader have to move more when the value is high. This amount can be changed by the requirement of the user.
The infomation amounts can be changed according to the users request and experience. If the evader is moving fast, the S_E might be large but the I_R could be small. If the evader is trying to turn the corner, S_E might be small but the I_R could be very large. These amounts are similar to that of human sence of avoiding the dangers, which makes this amount unique and useful for colision avoidance.

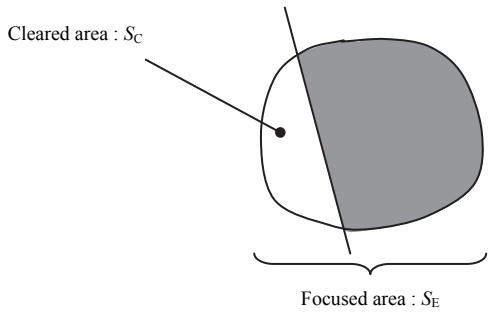


Fig. 1. Schematic image of information amounts.

3. Collision avoidance law

The total system of collision avoidance law in this paper consists of three types of control laws. They are actual collision avoidance, information gathering, and cource keeping. They are switched by the risk of collision and amount of infomation obtained. All of the simulation are in 2-dimensions and either acceralation or angular velocity of the vehicle is used as input variables. Fig.2 shows the basic definition of variables and constants used in this paper.

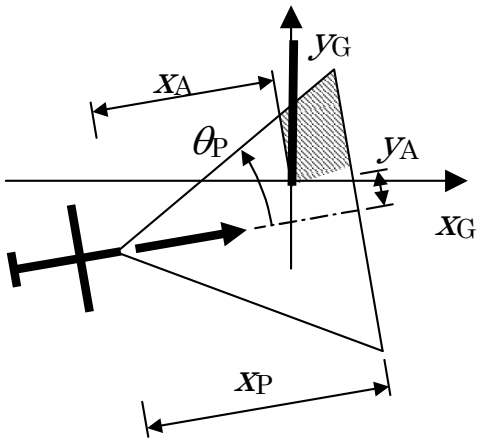


Fig. 2. Definitions of variables and constants.

The risk of collision is described numerically for collision avoidance control law. Two values are introduced in this paper. One is Range to Closest Point to Approach: R_{CPA}, which is the minimum range between two vehicles when their velocities and directional angles are kept at present value. The closest point is shown in Fig.3. This value indicates the future risk of collision. The other is Time to Closest Point to Approach: T_{CPA}, which is time to the range between two vehicles is R_{CPA}. Even if R_{CPA} is small, it is not necessary to avoid quickly

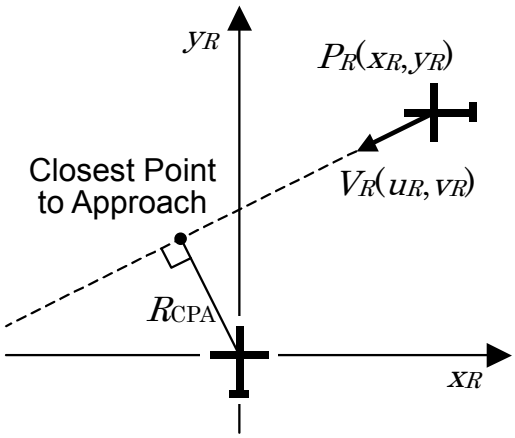


Fig. 3. Image of closest point to approach.

when T_{CPA} is large, in the othere hand, when the T_{CPA} is small, the evader have to start the motion very quickly. The collision avoidance control law is derived from the combination of these two risk functions.
R_{CPA} and T_{CPA} are driven using the relative position and velocity in the body fixed coordinate shown in Fig.3. The T_{CPA} is derived as,

$$T_{CPA} = -\left(x_R u_R + y_R v_R\right) \Big/ \left(u_R^2 + v_R^2\right) \tag{2}$$

Then, the relative position and range at closest point of approach is given as,

$$\begin{pmatrix} x_C \\ y_C \end{pmatrix} = \begin{pmatrix} x_R \\ y_R \end{pmatrix} + T_{CPA} \cdot \begin{pmatrix} u_R \\ v_R \end{pmatrix} \tag{3}$$

$$R_{CPA} = \sqrt{x_C^2 + y_C^2} \tag{4}$$

The risk function is defined as the following equation.

$$\varphi \equiv R_0^2 \Big/ \left(R_0^2 + R_{CPA}^2\right) \tag{5}$$

where R_0 represents the safety range. Collision will occur when φ=1. When the minimum range between two vehicle is equal to R_0, φ =0.5. The collision avoidance control law is designed to reduce φ less than 0.5 in a period of T_{CPA}. These parameters that show the risk of collisions are used for collision avoidance. When the risk is high, the direct collision avoidance control is activated to avoid the collision. One of the examples of collision avoidance control law is described in the following.
The collision avoidance control law satisfies the following equation.

$$\dot{\varphi} = -\varphi / T_C \tag{6}$$

T_C is time constant that is derived from the following requirement.

$$T_C = -T_{CPA} / \ln\left(2\varphi_0\right) \tag{7}$$

where φ_0 is the initial value of risk function. The left hand side of Eq. (6) can be derived from the derivative of Eq. (5) and eliminating the effect of the angular velocity of the evader. The effect of the angular velocity is momentary, where they return to their original values after the avoidance. Let the absolute velocity and angle of direction of the evader be, V_0 and y_0, and the intruder be, V_1 and y_1, the relative velocity of the vehicles can be expressed as,

$$\begin{pmatrix} u_R \\ v_R \end{pmatrix} = \begin{pmatrix} y_R \\ -x_R \end{pmatrix} \dot{\psi}_0 + \begin{pmatrix} \cos(\psi_1 - \psi_0) & 1 \\ \sin(\psi_1 - \psi_0) & 0 \end{pmatrix} \begin{pmatrix} V_1 \\ V_0 \end{pmatrix} \tag{9}$$

The first factor of the right hand side is the angular velocity of the evader, so by eliminating this factor, the following can be derived by taking the derivative of the relative velocity.

$$\frac{d}{dt}\begin{pmatrix} u_R \\ v_R \end{pmatrix} = \begin{pmatrix} v_R \\ -u_R \end{pmatrix} \dot{\psi}_0 \tag{10}$$

As total, the angular velocity for collision avoidance can be derived as,

$$(\omega_0)_A = \frac{1}{p_C v_R - q_C u_R} \cdot \frac{R_0^2}{2\varphi T_A} \tag{11}$$

where,

$$\begin{aligned}
p_C &= x_C p_x + y_C p_y \\
q_C &= x_C q_x + y_C q_y \\
p_x &= v_R\left(-u_R p_o - v_R q_i\right) \Big/ \left(u_R^2 + v_R^2\right) \\
q_x &= u_R\left(2v_R p_i - y_R\left(u_R^2 + v_R^2\right)\right) \Big/ \left(u_R^2 + v_R^2\right) \\
p_y &= v_R\left(2u_R p_i - x_R\left(u_R^2 + v_R^2\right)\right) \Big/ \left(u_R^2 + v_R^2\right) \\
q_y &= u_R\left(v_R p_o - u_R p_i\right) \Big/ \left(u_R^2 + v_R^2\right) \\
p_i &= x_R u_R + y_R v_R \\
p_o &= x_R v_R - y_R u_R
\end{aligned} \tag{12}$$

This collision avoidance control law activates when the risk is high. In other words, other 2 control laws, infomation gathering and course keeping laws are used when the risk is low. In the following 2 sections, the different types of infomation gathering control laws are introduced depending on the difference of the uncertainties up ahead. The course keeping control law is used to keep the original course, which is not important in this paper, so will not be explained in details.

4. Uncertainty depending on relative position

The information gathering control with uncertainty depending on relative position is introduced in this section. The uncertainty depending on relative position stands for the cases where the infomation that can be obtained are defined as function of relative distance to an uncertainty. This is applicable for the flights in the fog or smokes where the uncertainty differs by the distance, closer you are, clear infomation you can obtain. The control target is fixed wing aircraft and the control input is angular velocity. The control law is designed from fuzzy logic to realize the fuzziness of the infomation. First, the additional parameters of the uncertain infomation is explained in this section. Followed by the control law and control results.

4.1 Uncertain parameters

The control law uses additional parameters for infomation in this section. The basic parameters were I_L and I_R which was explained in section 2. The following parameters quantifying uncertainty are added to design the control law dealing with uncertain information. Uncertainty of information depends on the target existence and location. Therefore, first, uncertainty parameters are defined separately. Then, the uncertainty coefficient for the control law is obtained from the uncertainty parameters for existence and location.

4.1.1 Information probability - I_P

Infomation probability is a parameter describing the probability, possibility or likelihood of the target existence. I_P takes a fixed value from 0 to 1 and is assigned by the user before avoidance. When $I_P = 0$, there is no probability of existence. On the other hand, the target existence is certain when $I_P = 1$.

4.1.2 Information clarity - I_C

Infomation clarity is a parameter describing the clarity of target existence. I_C varies with the quantity of information. When $I_C = 0$, there is no information about existence. On the other hand, information on target existence becomes clear when $I_C = 1$. When the information has uncertainly, I_C varies from 0 to 1 with relative distance between the target and evader. For example, I_C varies as shown in Fig. 4 when visibility is obscured by fog. Information becomes clear ($I_C = 1$) at a certain distance and worsens gradually with further distance. No information is provided ($I_C = 0$) beyond a certain distance. Figure 1 shows an environment when visibility is just barely secured at 5000m. The target is found definitely when the relative distance is less than 4000m. As the relative distance increases, visibility worsens gradually and the target cannot be found when the relative distance exceeds 5000m.

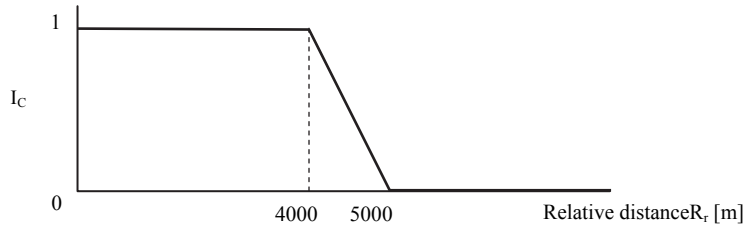


Fig. 4. Information clarity : The figure shows the case when the visibility is 4000 – 5000[m]. The information is clear when the distance is lower than 4000[m] and cannot be obtained when the distance is over 5000[m]. The information clarity changes linearly between two regions.

4.1.3 Information truth - I_T

Infomation truth is a parameter describing truth of the target existence. The value determines whether the target exists or not. It takes a value of either 0 or 1. When $I_T = 0$, there is no target. On the other hand, when $I_T = 1$, there is a target.

4.1.4 Information location accuracy - I_A

Infomation location accuracy is a measure of the area in which the target exists. For example, in a 2D model, I_A is a circle with a radius of arbitrary length. As shown in Fig. 5, it is assumed that I_A depends on the radius of the zone containing the target. I_A takes a value between 0 (existence zone is vast) to 1 (existence zone is very small). In summary, the existence zone is the domain where the target may exist.

4.2 Application to control law

The control law using the uncertain infomation is introduced. From the viewpoint of complexity and difficulty, it is wrong to design a whole new control law adopting

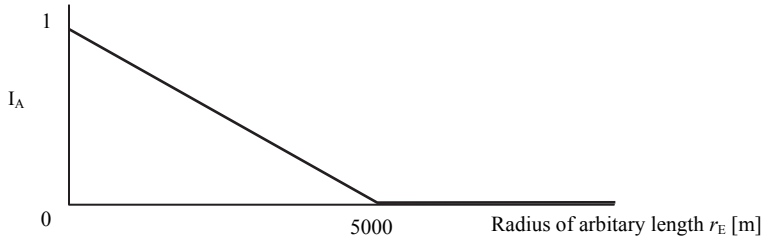


Fig. 5. Information location accuracy: The figure shows that the unknown target is within 5000[m] of radius with highest possibility at the center.

uncertainty. Therefore, the control law to deal with uncertainty simply by introducing technique to the conventional control law with only minor modification is proposed.

4.2.1 Uncertainty coefficient

As a first step in designing a control law to deal with uncertainty, the uncertainty coefficient k based on the uncertainty parameters of existence and location (I_P, I_C and I_T) is brought in. k is a function of these parameters expressed as,

$$k=f(I_P, I_C, I_T) \tag{13}$$

As shown in Fig. 6, k is a coefficient introduced for the following reasons. Under the conventional control law, target information is certain and the evader flies a course for either target existence or target absence. On the other hand, when the information is uncertain, the evader flies somewhere between target existence and target absence. It is assumed that k has three components as follows, depending on relative distance: part based on original estimate (corresponding to $I_C = 0$), part based on information that gradually becomes clear (corresponding to $0 < I_C < 1$), and part based on clear information (corresponding to $I_C = 1$).

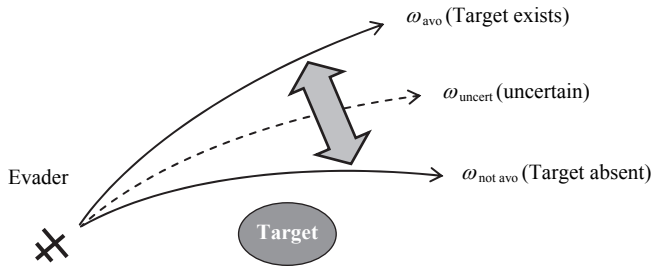


Fig. 6. The trajectory of uncertain condition and input.

In the situation in Fig. 4, when each uncertain parameter is given, the value of k becomes as shown in Fig. 7. k takes a constant value depending on I_P when $R_r > 5000$, because the amount of information quantity does not vary in this condition. Information gradually

becomes clear when $4000 < R_r < 5000$. Variation of I_C reflects variation of k. k takes either 0 or 1 depending on only I_T when $R_r < 4000$, because the information becomes certain. I_T is finally decided whether the target exists or not. However, if it is assumed that I_T approaches a true value gradually as the information becomes clear, k is determined in real time. In such a circumstance, control input (angular velocity) ω_{uncert} takes a value between target existence and target absence. Therefore ω_{uncert} is expressed by the following equation using k,

$$\omega_{uncert} = k\ \omega_{avo} + (1-k)\ \omega_{notavo} \tag{14}$$

where ω_{avo} and ω_{notavo} are the control input to the evader for target existence and target absence, respectively. Avoidance control depends on the area where the target may exist, thus ω_{avo} is a function of I_A.

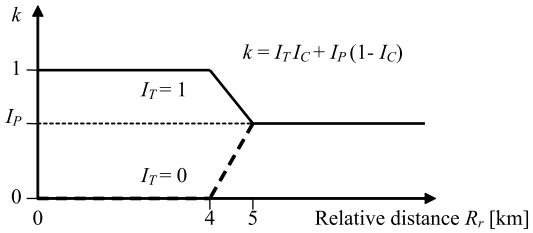


Fig. 7. Image of uncertainty coefficient with uncertain information.

4.2.2 Information acquisition requirement

The I_R explained in section 2 is now brought in for the control. As explained in section 2, the I_R is the requirement of I_L by the user. I_R represents the degree of need to obtain information on the focused area. I_R is determined by the user and takes a value from 0 to 1. When $I_R = 0$, there is no requirement. On the other hand, all information is required when $I_R = 1$. As shown in Fig. 8, the clear region in the focused area becomes small as I_R approaches 0. In contrast, as I_R approaches 1, the cleared part becomes large. If the conventional control law is modified by introduction of the uncertainty coefficient, k, and the information acquisition requirement, I_R, design of a new control law to deal with uncertainty is comparatively easy.

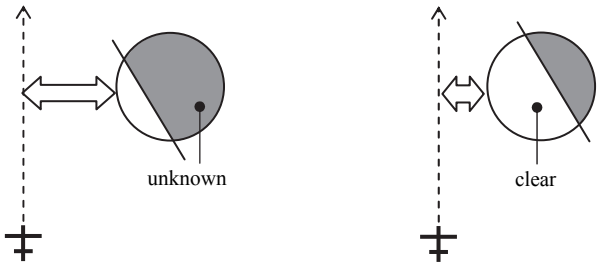


Fig. 8. Image of information acquisition requirement. Left figure shows the case where I_R is small (vehicle passes in far region from the focused area), right figure shows the case where I_R is large (vehicle passes in closer region from focused area).

4.3 Simulation result of relative position - in-fog problem

An example of avoidance problem is uncertainty of information defined in the relative coordinate (body fixed) system is shown. The problem is assumed to be in-fog problem, where there is area where the information is uncertain upahead.

4.3.1 Statement of problem

The evader cannot see the target beyond a certain distance because visibility is obscured by an obstacle like fog. The problem is defined as two-dimensional in the horizontal plane. The evader flies on a straight course with constant velocity towards a target that may exist in existence zone as shown in Fig. 9. Visibility is defined as a function of relative distance from the evader. When the relative distance is smaller than a certain distance, for example 4000m, the evader can see the target clearly. However, the evader cannot see the target when the relative distance is larger than a certain distance, for example 5000m. Visibility changes gradually between these two areas. The information clearness I_C is defined depending on the relative distance to the target as for visibility as explained in Fig.4. The target is close to evader's course, but information about existence and position are uncertain. Therefore, the target existence is given as the information probability, I_P, and the position is given as the target existence zone (circlular region with radius of r_E).

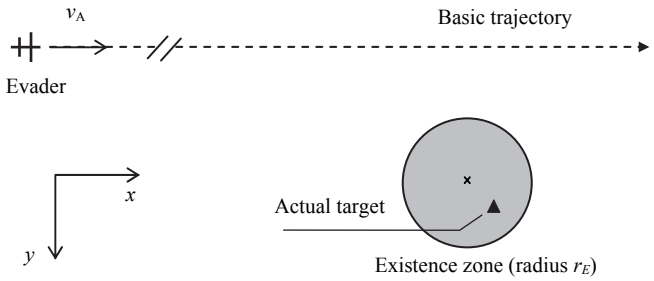


Fig. 9. Image of the problem, the target is assumed to be inside the fog where, the information obtained from evader is uncertain.

4.3.2 Initial conditions and requirement

The initial position of the evader and target existence zone are shown in Fig. 9. Other constants are shown in Table 1. Figure 4 is used for I_C. Also the required separation between the evader and the target is set from 3000m to 4000m.

4.3.3 Initial conditions and requirement

Figures 10 and 11 show the avoidance trajectory and angular velocity, respectively. Solid lines represent the results for the proposed control law; dashed lines represent the results for the conventional control law. The figure shows two cases for the conventional control law: avoidance with correct information; and avoidance with incorrect information where target appears suddenly without information.

The avoidance trajectories in Fig. 10 show that avoidance using the conventional control law with incorrect information causes significant delay because the evader does not avoid until the target is found. On the other hand, the avoidance trajectories produced by the proposed

Parameters	Symbols	Values	Uncertain parameters	Values
Velocity	v_A	250[m/s]	I_P	0.25
Initial position		(0, 0)	I_C	Shown in Fig. 4
Radius of existence zone	r_E	1000[m]	I_T	1
Center of existence zone		(25000, 2000)	I_A	Shown in Fig.5
True position of target		(25500, 2500)	I_R	0.2, 0.9

Table 1. Parameters and constants for the simulation.

control law depend on the value of I_R, because I_R indicates the degree of necessary information. When I_R is large, the evader must fly closer to the target and does not take early avoidance. Therefore, two stage avoidance occurs when I_R = 0.9. The first stage is infomation gathering based on current infomation and information acquisition; the second stage is avoidance after finding the target. Figure 11 shows the angular velocities for avoidance. Information uncertainty reduces the sudden and severe avoidance that occurs using the conventional control law with incorrect information.

Both of the figures show that the proposed control law was able to increase the safetiness and reliability of the flight in uncertain information defined in relative position from the evader.

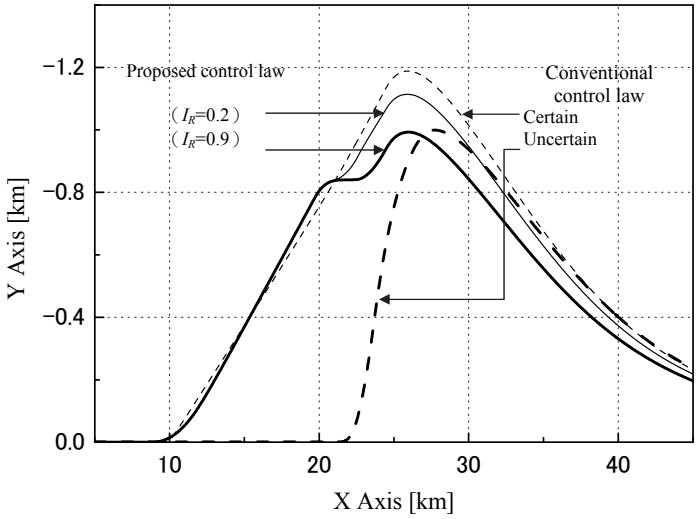


Fig. 10. Avoidance trajectory of the cases with/without proposed control law using uncertain parameters (I_P = 0.25).

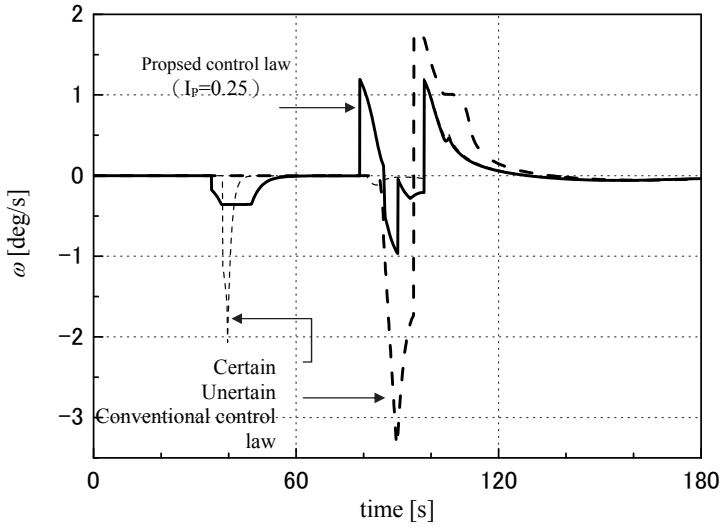


Fig. 11. Angular velocity for avoidance of the cases with/without proposed control law using uncertain parameters (I_P = 0.25).

5. Uncertainty depending on absolute position

In this section, the uncertainty depending on absolute position is treated. Different from the uncertainty that differs by relative position as explained in section 4, the information does not change due to the environment. For example, when aircraft is going around a mountain or a helicopter going around the buildings, the information does not change due to relative position.

The information parameters I_L and I_R which was explained in section 2 is used in this section to see the effect of the information amount. The collision avoidance control explained in section 3 is brought in when the vehicle enters critical condition, otherwise the following either infomation gathering control or course keeping control takes place.

5.1 Design of information gathering control law

Design of information gathering control law is derived using a model in Fig.12 and Fig.13. Figure 12 is the vehicle in the ground fixed coordinate and Fig.13 shows the vehicle in body fixed coodinate. Focused area depends on the speed and direction of the vehicle, so the distance x_P can be set by the users descision. The angle θ_P also depends on the level of safety which can be chosen. The area of focused area is given as follows.

$$S_E = x_P^2 \cdot \tan\theta_P \tag{15}$$

The shadow area cannot be seen from the vehicle. So, the area of shadow area is assumed as the following equation using the focused area.

$$S_B = \frac{1}{2}\left(x_P^2 - x_A^2\right) \cdot \left(\tan\theta_P - \tan\theta_A\right) \tag{16}$$

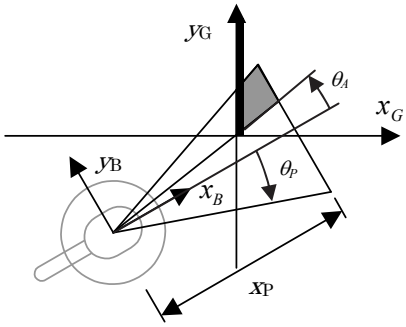


Fig. 12. Vehicle in ground fixed coordinate system.

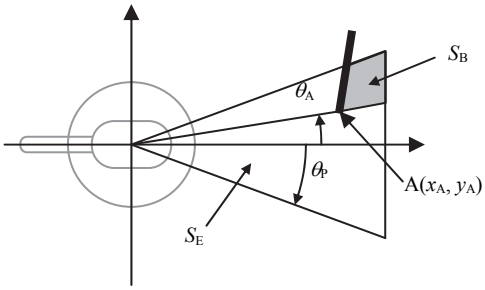


Fig. 13. Vehicle in body fixed coordinate system.

where,

$$\tan\theta_A = y_A / x_A \tag{17}$$

x_A and y_A are relative coordinate in the body fixed frame. The time derivatives of x_A and y_A are given as follows.

$$\frac{d}{dt}\begin{pmatrix} x_A \\ y_A \end{pmatrix} = \begin{pmatrix} y_A \\ -x_A \end{pmatrix}\omega_0 - \begin{pmatrix} 1 \\ 0 \end{pmatrix}V_0 \tag{18}$$

where ω_0 and V_0 are angular velocity and velocity of the vehicle, respectively. To derive the dynamical property of information amount, time derivative of the I_L is derived.

$$\frac{dI_L}{dt} = \frac{S_B\dot{S}_E - \dot{S}_B S_E}{S_E^2} \tag{19}$$

S_B and S_E are given in Eqs.(15) and (16). It is necessary to define the x_P and θ_P of the focused area in order to derive the time derivative. The distance of focused area x_P is proportional to the velocity of helicopter V_0.

$$x_P = V_0\,\Delta t \tag{20}$$

where Δt is a constant with the unit of time. This shows that the vehicle moves the distance of x_P in the period of Δt. The angle θ_P is set as constant, thus the vehicle focuses wide area in case of high speed.
The time derivatives of the S_B and S_E are derived.

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix}S_B\\S_E\end{pmatrix}&=\begin{pmatrix}A & B\\D & 0\end{pmatrix}\begin{pmatrix}\dot{V}_0\\V_0\omega_0\end{pmatrix}+\begin{pmatrix}C\\0\end{pmatrix}\\&=\begin{pmatrix}A & B\\D & 0\end{pmatrix}\begin{pmatrix}a_x\\a_y\end{pmatrix}+\begin{pmatrix}C\\0\end{pmatrix}\end{aligned} \tag{21}$$

where,

$$\begin{aligned}A&=V_0(\Delta t)^2\left(\tan\theta_P-\tan\theta_A\right)\\B&=\left[-x_Ay_A\left(\tan\theta_P-\tan\theta_A\right)+\tfrac{1}{2}\left(x_P^2-x_A^2\right)\left(1+\tan^2\theta_A\right)\right]\Big/V_0\\C&=\left[x_A^2\left(\tan\theta_P-\tan\theta_A\right)-\tfrac{1}{2}\left(x_P^2-x_A^2\right)\tan\theta_A\right]V_0/x_A\\D&=2V_0(\Delta t)^2\tan\theta_P\end{aligned} \tag{22}$$

a_x and a_y are horizontal acceleration in the body fixed frame. Substituting Eq. (21) into Eq.(19), the time derivative of I_L is given.

$$\frac{dI_L}{dt}=E\cdot a_x+F\cdot a_y+G \tag{23}$$

where,

$$\begin{aligned}E&=\left(S_BD-S_EA\right)\Big/S_E^2\\F&=-B/S_E\\G&=-C/S_E\end{aligned} \tag{24}$$

Therefore, the information amount of safety can be changed by a_x and a_y. This shows that the direction of acceleration changes the information amount. The information gathering control law is required to keep the information amount higher than the specified value I_R. Thus the shortage of information amount is defined in the following equation.

$$I_E=\begin{cases}I_R-I_L & (I_R>I_L)\\0 & (I_R\le I_L)\end{cases} \tag{25}$$

In the case of I_L is less than I_R, the controller is required to increase I_L. On the other hand, in the case of I_L is greater than I_R, the high decreasing rate of I_L is not desired. Therefore the following control law satisfies the both cases.

$$\dot{I}_L\ge\left(\dot{I}_L\right)_{\min} \tag{26}$$

where,

$$\left(\dot{I}_L\right)_{\min}=k_E\left(I_R-I_L\right) \tag{27}$$

k_E is a feedback gain. The left hand side of Eq. (26) is given in Eq. (23). Equation (23) has two input variables, a_x and a_y. Thus the minimum norm of input vector is chosen.

$$(a_x)_I = E \cdot u_N, \quad (a_y)_I = F \cdot u_N \tag{28}$$

where,

$$u_N = \left(k_E(I_R - I_L) - G\right) \Big/ \left(E^2 + F^2\right) \tag{29}$$

This control law uses the same feedback gain in the cases of $I_R > I_L$ and $I_R < I_L$. The vehicle is required to keep the desired velocity and direction when the information amount satisfies the requirement. In this case, the following feedback law is used.

$$(a_x)_D = k_V (V_R - V_0), \quad (a_y)_D = k_\psi (\psi_R - \psi_0) V_0 \tag{30}$$

where k_V and k_ψ are feedback gains. Finally, the information gathering control law is given as follows.

i.

$$E \cdot (a_x)_D + F \cdot (a_y)_D + G \geq (\dot{I}_L)_{\min} \tag{31}$$

$$a_x = (a_x)_D, \quad a_y = (a_y)_D \tag{32}$$

ii.

$$E \cdot (a_x)_D + F \cdot (a_y)_D + G < (\dot{I}_L)_{\min} \tag{33}$$

$$a_x = (a_x)_I, \quad a_y = (a_y)_I \tag{34}$$

The $\dot{I}_L$ in the inequality conditions are given in Eq. (26). The control law is called Information Amount FeedBack (IAFB) for it feeds back the information amount as one of the parameters for information gathering.

5.2 Simulation result of absolute position

The simulation result using the collision avoidance control with IAFB is introduced. Two different cases of similation will be shown in this scetion. The first case is the case with the helicopters. The velocity of the vehicle can be changed directly by the control law. The second case is the case with the fixed wing aircraft. The input is given as the angular velocity and the velocity itself is kept as constant.

5.2.1 Simulation result of helicopters

Figure 14 shows the initial condition of the evader and intruder. The intruder cannot be seen from evader at beginning of the control. The intruder is incoming from behind the obstacle with velocity of 10[m/s] and 20[m] away from the obstacle. The evader starts from 150[m] away from the obstacle with various position defined by $y(0)$. Figure 15 shows the simulation result of the avoidance for different initial conditions. Figure 16 shows the case without IAFB for comparison.

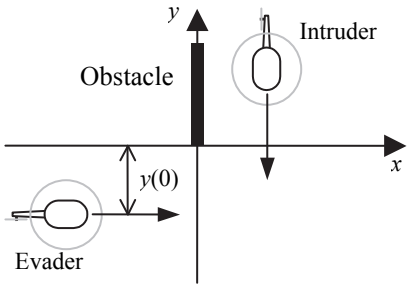


Fig. 14. Initial condition of evader and intruder.

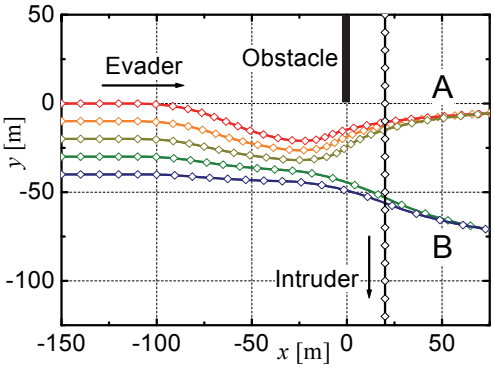


Fig. 15. Trajectory of evader with IAFB starting from different initial positions.

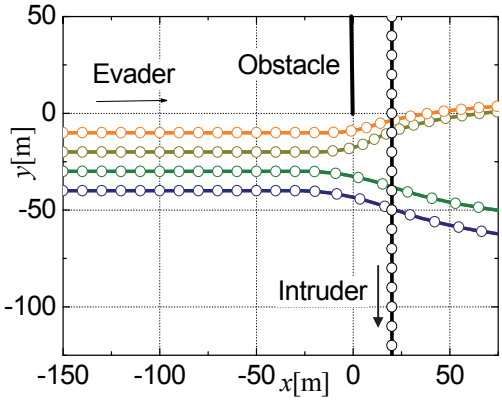


Fig. 16. Trajectory of evader without IAFC starting from different initial positions.

In the first half of the control, the evader starts to obtain the information behind the obstacle. After the intruder is found approaching, the evader decides to evade either in front or back of the intruder depending on the estimated trajectory of the intruder passing in the way. In this case, the results were split into exactly two groups where evader decelerates and passes behind the intruder or accelerates and passes in front of intruder. In Fig.16, the case without IAFB, the first half of the information gathering does not occur, so the helicopter avoids the intruder after they find the incoming vehicle. Figure 17 is comparison of the minimum distance when the two vehicles pass each other. The result with IAFB shows higher level of avoidance due to the earlier motion of gathering information which leads to easier avoidance and faster recognition of the intruder.

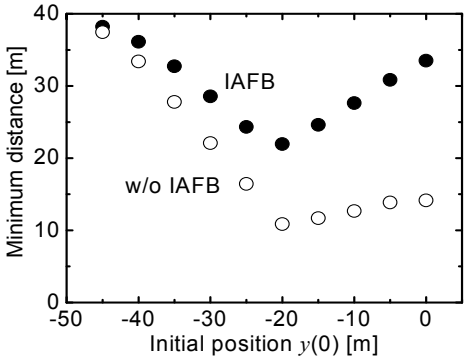


Fig. 17. Trajectory of evader without IAFC starting from different initial positions.

5.2.2 Simulation result of fixed wing aircraft

For fixed wing type aircraft, it is not easy and efficient to change the velocity so often. The control input for these types are changed to angular velocity input. Basic input is same as the one described in section 5.1. Most of the conditions are same as that of the case of the helicopters except that the cruising speed of evader and intruder is 100[m/s] and the results are compared with different I_R and different course of the intruder.

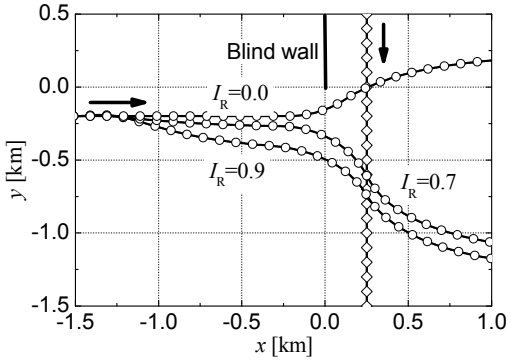


Fig. 18. Trajectory of evader with different I_R.

Figure 18 shows the trajectory of evader with different I_R, in the case of I_R =0 this is same as the case without IAFB, we can see that the trajectory changes by the amount of information required.
Figure 19 shows the time history of the relative distance of the vehicles. The results show that the relative distance decreases very quickly in the case without IAFB and the minimum distance between the vehicles are shorter than the others. This clearly shows the effectiveness of the IAFB.

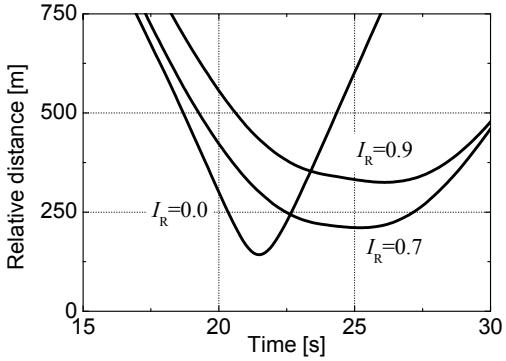


Fig. 19. Time history of relative distance between intruder and evader with different I_R.

Figure 20 shows the trajectory of evader when the intruder starts from different positions. The I_R is set as 0.9 for this simulation. In the cases when the intruder is far from the obstacle, the trajectory is smoother because the intruder is found quicker. The other two cases makes sharp turns due to the slower finding of the intruder.

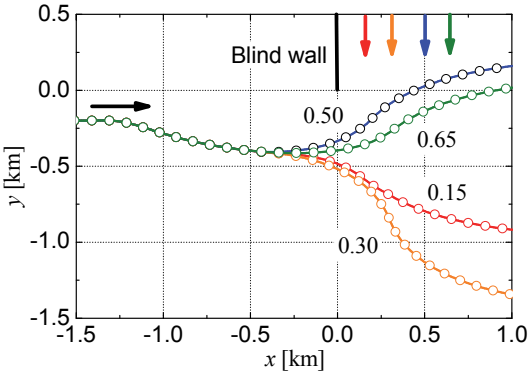


Fig. 20. Trajectory of evader when the intruder approaches from different course.

6. Conclusion

Two cases of collision avoidance control is simulated to see the effect of the information amount as parameter for control. One was that uncertainty of the information changes by the relative position of the evader and the target and the other was that uncertainty of the

infomation is given as absolute position. Both cases have shown smoother and safer trajectories than the conventional control laws. The simulation results have shown that the control laws using information amounts does not rely on the coodinates. The motion of the aircraft show similar trajectories to that of humans to obtain safe margin to gain information when they do not have enough information.

7. References

Ciletti, M.D.; Meza, A.Z. & Takushoku, S. (1997). Collision Avoidance Maneuver for Ships, *Navigation – Japan Institute of Navigation-*, Vol. 54, pp.83-89.

Frazzoli, E.; Mao, A.H.; Oh, J.H. & Feron, E. (2001), Resolution of Conflicts Involving Many Aircraft via Semidefinite Programming, *Journal of Guidance, Control, and Dynamics*, Vol. 24, No. 1, pp.79-86.

Gates, D.J. (2009) Properties of a Real-Time Guidance method Preventing a Collision, *Journal of Guidance Control and Dynamics*, Vol. 32, No. 3, pp. 705-716.

Hiraoka, T.; Tanaka, M.; Kumamoto, H.; Izumi, T. & Hatanaka, K. (2009a), Collision Risk Evaluation Index Based on Deceleration for Collision Avoidance (First Report) : Proposal of a new index to evaluate collision risk against forward obstacles, *Review of Automotive Engineering*, Vol. 30, No. 4, pp. 429-437.

Hiraoka, T.; Tanaka, M.; Takeuchi, S.; Kumamoto, H.; Izumi, T. & Hatanaka, K. (2009b), Collision Risk Evaluation Index Based on Deceleration for Collision Avoidance (Second Report) : Forward obstacle warning system based on deceleration for collision avoidance, *Review of Automotive Engineering*, Vol. 30, No. 4, pp.429-447.

Iwama K. (2008), Study on Collision Avoidance Control Law using Information Amount Feedback, *Master Thesis of Yokohama National University.*

Kubota, N. & Fukuda, T. (1999). An Intelligent Robotic System Based on a Fuzzy Approach, *Proceedings of the IEEE*, Vol. 87, No. 9, pp.1448-1470.

Miele, A.; Wang, T.; Mathwig, J.A. & Ciarcia, M. (2010), Collision Avoidance for an Aircraft in Abort Landing : Trajectory Optimization and Guidance, *Journal of Optimization Theory and Applications*, Vol. 146, No. 2, pp.233-254.

Shioiri, H. & Ueno, S. (2004), Three-dimensional Collision Avoidance Control Law for Aircraft using Risk Function and Fuzzy Logic, *Transactions of JSASS*, Vol. 46, No. 154, pp. 253-261.

Shioiri, H. & Ueno, S. (2004), Collision Avoidance Control Law for Aircraft under Uncertain Information, *Transactions of JSASS*, Vol. 47, No. 157, pp. 209-215.

Slater, G.L.; Byram, S.M. & Williams, T.W. (2006), Collision Avoidance for Satellites in Formation Flight, *Journal of Guidance, Control, and Dynamics*, Vol. 29, No. 5, pp. 1140-1146.

Stipanovic, D.M.; Hokayem, P.F.; Spong, M.W. & Salijak D.D. (2007), Cooperative Avoidance Control for Multiagent Systems, *Journal of Dynamic Systems Measurement and Control –Transactions of the ASME*, Vol. 129, No. 5, pp.699-707.

Introduction to TCAS II (1990), *U.S. Department of Transportation, Federal Aviation Administration*.