Self-Organising Maps

Applications in Geographic Information Science

Editors

PRAGYA AGARWAL

Department of Geomatic Engineering, University College, London, UK

ANDRÉ SKUPIN

Department of Geography, San Diego State University, USA

John Wiley & Sons, Ltd

This page intentionally left blank

Self-Organising Maps

This page intentionally left blank

Self-Organising Maps

Applications in Geographic Information Science

Editors

PRAGYA AGARWAL

Department of Geomatic Engineering, University College, London, UK

ANDRÉ SKUPIN

Department of Geography, San Diego State University, USA

John Wiley & Sons, Ltd

Contents

List of Contributors

Pragya Agarwal   Department of Geomatic Engineering, University College London, London WC1E 6BT, UK

Peggy Agouris   Center for Earth Observing and Space Research, George Mason University, Fairfax, Virginia 22030, USA

Fernando Bação   ISEGI/UNL, Campus de Campolide, 1070-312 Lisboa, Portugal

Irene Casas   Department of Geography and National Center for Geographic Information and Analysis, University at Buffalo, The State University of New York, Amherst, NY 14261, USA

Pete Doucette   Principal Scientist, ITT Corporation, Advanced Engineering and Sciences, Alexandria, VA 22303 USA

Michael Goodchild   National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA 93106, USA

Bruce C. Hewitson   Department of Environmental and Geographical Science, University of Cape Town, Rondebosch 7701, South Africa

Etien L. Koua   International Institute for Geoinformation Science and Earth Observation (ITC), PO Box 6, 7500 AA Enschede, The Netherlands

Menno-Jan Kraak   International Institute for Geoinformation Science and Earth Observation (ITC), Department of Geo-Information Processing, PO Box 6, 7500 AA Enschede, Hengelosestraatgg, The Netherlands

William A. Kretzschmar Department of English, University of Georgia, Athens, GA 30602, USA

Jürgen P. Kropp Potsdam Institute for Climate Impact Research, PO Box 601203, 14412 Potsdam, Germany

Victor Lobo ISEGI/UNL, Campus de Campolide, 1070-312 Lisboa, Portugal and Portuguese Naval Academy, Alfeite, 2810-001 Almada, Portugal

Marco Painho ISEGI/UNL, Campus de Campolide, 1070-312 Lisboa, Portugal

Monika Sester Institute for Cartography and Geoinformatics, Leibniz University of Hannover, 30167 Hannover, Germany

Hans-Joachim Schellnhuber Potsdam Institute for Climate Impact Research, 14412 Potsdam, PO Box 601203, Germany

André Skupin Department of Geography, San Diego State University, San Diego, CA 92182-4493, USA

Anthony Stefanidis Department of Earth Systems and Geoinformation Sciences, George Mason University, Fairfax, Virginia 22030, USA

Jean-Claude Thill Department of Geography and Earth Sciences and Center for Applied Geographic Information Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Jun Yan Department of Geography and Geology, Western Kentucky University, Bowling Green, KY 42101, USA

Xiaobai Yao Department of Geography, University of Georgia, Athens, GA 30602, USA

1

Introduction: What is a Self-Organizing Map?

André Skupin[1] and Pragya Agarwal[2]

[1] *Department of Geography, San Diego State University, San Diego,*
CA 92182-4493, USA

[2] *Department of Geomatic Engineering, University College London,*
London WC1E6BT, UK

1.1   INTRODUCTION

In the quest to understand and address important issues of the modern era, from environmental degradation to economic development, enormous amounts of geographic data are being generated. With the increasing adoption of such technologies as hyper-spectral remote sensing or wireless sensor networks, the growth rate of data volumes continues to rise. Granularity of geographic data is increasing both in geometric space (i.e. more features and finer cell sizes), and in attribute space (i.e. more attributes and finer measurements of attribute values), leaving us with truly *n*-dimensional data. We are thus increasingly faced with a data-rich environment, in which traditional inference methods are either failing or have become obstacles in the search for geographic structures, relationships, and meaning. With respect to statistical analysis, some problems of traditional approaches, especially regarding spatial autocorrelation, are increasingly being addressed (Fotheringham *et al*., 2000, 2002; Rogerson, 2001). However, many see the need for a paradigmatic shift in how geographic data are analysed and this push for a new direction is gaining strength, as indicated by the emergence of such disciplinary labels as *geocomputation* (Fischer and Leung, 2001; Longley, 1998; Openshaw and Abrahart, 2000) or *geographic data mining* (Miller and Han, 2001).

It is this direction, characterized by intense computation applied to large data sets, which is explored in this book. Specifically, it addresses a method known as the Kohonen map or self-organizing map (SOM). It may appear odd to devote a complete volume to a single technique. Indeed, most books on GIS are either textbooks giving an introduction to the overall field or are devoted to a particular application domain, like hydrological modelling. However, those books that explicitly address geo-computation or geographic data mining tend to cover a multitude of very heterogeneous methods and are thus not able to explore each approach in great detail. Very few have limited themselves to a more narrowly defined group of related techniques (Openshaw and Openshaw, 1997). Furthermore, the SOM method was not developed by GIScientists and an excellent monograph already exists that is regularly updated (Kohonen, 2001).

This edited volume aims to demonstrate that there is indeed something special about this method, something that makes it curiously attractive to diverse and sometimes conflicting interests and approaches in GIScience. Those interested in clustering and classification will recognize in it elements of k-means clustering, but with an explicit representation of topological relationships between clusters. Anyone accustomed to dealing with n-dimensional data through a transformation and reduction of variables, as in principal components analysis (PCA) or multidimensional scaling, will tend to interpret the SOM method in that light. The predominantly two-dimensional form of most SOMs means that cartographers and others involved in geographic visualization can readily envision its integration within interactive visualization environments. Those struggling to communicate the results of complex computational procedures to policy-makers and the broader public may find SOMs to be uniquely accommodating in many circumstances. This volume intends to provide a common platform for all those facets of current work in GIScience that pertain to use of SOMs. This is what we hope will separate this volume from others that only allow an abbreviated discussion of the SOM method as one example of artificial neural networks (ANNs) due to broader scope and limited space.

This chapter is aimed at answering basic questions about what a SOM is, how it is created and used, and how it relates to other techniques that readers may be familiar with. All this is done primarily through plain language explanation and visual illustration, as opposed to formulas and the language of mathematics. Kohonen's monograph cannot be beat in the latter regard and is highly recommended to anyone wanting to delve deeper into the inner workings of a SOM (Kohonen, 2001). This chapter also discusses important questions about the relationship between GIScience and the SOM method and finally provides an overview of the other chapters in this book.

1.2   RELATED METHODS

The SOM is part of a large group of techniques known as *artificial neural networks* (ANNs). These have a reputation for performing surprisingly well, while providing little explanation for how results are exactly arrived at. In fact, ANNs are often seen as black-box operations. However, at least in the case of the SOM method, the actual algorithms can be surprisingly simple and the process of self-organization is not beyond comprehension. One quickly realizes that, apart from seeing the SOM only in the context of other ANN methods, depending on its purpose and training parameters one could also interpret it primarily as a *clustering* or *dimensionality reduction* technique. In fact,

the SOM is an ANN method that always performs both clustering *and* dimensionality reduction. The separation invoked in this section is designed to more clearly convey the position of the SOM method in relation to standard statistical and geocomputational approaches.

1.2.1   Artificial Neural Networks

First, it is important to note that ANNs, also known as computational neural networks (CNNs) (Fischer, 2001), are by no means simulations of biological neural networks. At best, one could say that the original idea behind neural computing drew inspiration from biological counterparts, and that most actual implementations are far removed from that inspirational source. What artificial and biological neural networks have in common is that information is not stored in any single location, but rather in a parallel, distributed form, and that certain mechanisms exist in which new information can be 'learned' through changes that potentially affect large portions of the network (i.e. learning rules).

The general structure of an ANN consists of a set of *input nodes* and a set of *output nodes*. Alternatively, these nodes are also known as neurons, processing elements, or computational units (Fischer, 2001). Multivariate data presented to input nodes gets processed such that output nodes are activated according to weights that are associated with each incoming link. Neural network training is largely concerned with setting these weights. To do this, many neural networks contain one or more layers of hidden nodes (Figure 1.1). During training, the weights of incoming connections to these nodes are
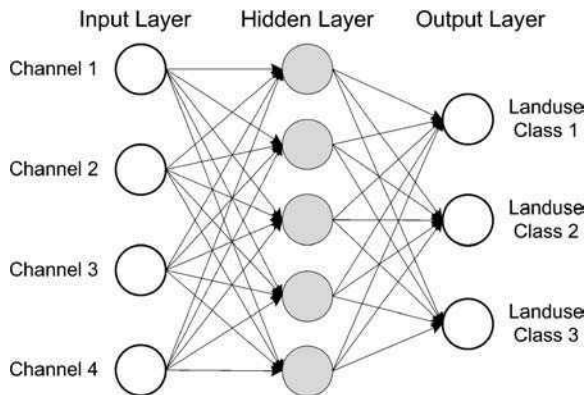
Figure 1.1 *Supervised, feed-forward, neural network trained with multispectral remote sensing data and known landuse classes*

summed up according to a predefined function. Depending on whether its result satisfies a certain threshold function, an outgoing connection can then be activated. The number of hidden layers and type of connections are an important basis for categorizing different ANN types. In addition to the fixed number of layers and fixed network topology found in many neural networks, there are also neuro-evolutionary models, which use genetic algorithms to help shape neural networks during training.

A fundamental distinction can be made between supervised and unsupervised neural networks. In the *supervised* case, input data presented to the network during training

consist of multivariate data with known outcomes or classifications, i.e. input–output pairs. For example, one could train a neural network with land use classes and corresponding multispectral signatures (Figure 1.1). Multispectral data will be presented, in this case, to the input nodes and a land use class is associated with each output node. During training, weights of hidden layers are iteratively adjusted to establish a good fit between multi-spectral values and correct land use classes. After training is complete, new multi-spectral observations can be presented to the input nodes and land use classes predicted. Most awareness of the power of neural networks within GIScience stems from the use of supervised models. When training data are both multivariate and multi-temporal, one can even predict change patterns (Pijanowski *et al*., 2002). Supervised neural networks have also been used for purposes other than classification. For example, regression models could be constructed, when continuous outputs are available.

In *unsupervised* learning, the input vectors do not correspond to classes known a priori. Output nodes compete for the input vectors on the basis of certain similarity functions and the weights of winning nodes are adjusted according the weights of respective input nodes. Due to this competitive learning procedure, input nodes that are quite similar are driving adjustments of similar output nodes. At the same time, dissimilarities in the input data become accentuated. All this supports unsupervised learning's primary role of finding major structures, clusters, and relationships in multivariate data.

In addition to the fundamental distinction discussed above, one can also distinguish neural networks in terms of whether, during training, adjustments made to neuron weights are only fed forward to the following layers or are also having an effect on preceding layers. Accordingly, *feed-forward* and *recurrent* networks are distinguished. Finally, an important concept is that of *back propagation* (Rumelhart and McClelland, 1986), which is used in feed-forward networks and refers to how errors (i.e. differences between known outputs and neural network outputs) are minimized by making adjustments to neuron weights.

Where does the SOM method fall within the overall system of ANNs? The standard SOM algorithm – the most widely known form and used in many popular software packages (e.g. SOM_PAK) – involves an unsupervised neural network with competitive learning and no hidden layers (Figure 1.2). In that traditional form, SOMs have been especially popular for purposes of clustering and visualization. However, there are also
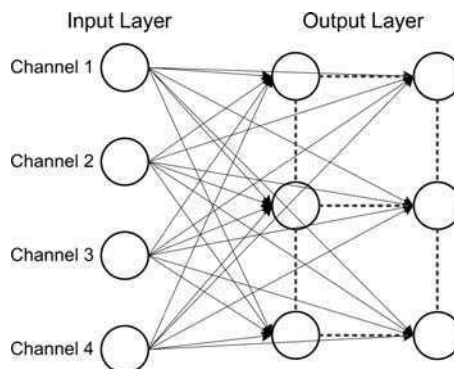
Figure 1.2   *Small SOM trained with multispectral remote sensing data*

supervised variants useful for classification, including Kohonen's own *learning vector quantization* (LVQ) (Kohonen, 2001). In this chapter, and for most of this edited volume, the standard SOM algorithm is the focus of discussion. For detailed coverage of other neural networks readers are encouraged to refer to various surveys of this subject (Gurney, 1997; Hertz *et al*., 1990) as well the growing number of geographically oriented literature (Fischer, 2001; Openshaw and Openshaw, 1997).

1.2.2   Clustering Methods

Discussion of the SOM method in the geographic literature tends to focus on its clustering qualities. The basic idea behind clustering is the attempt to organize objects into groupings based on certain shared characteristics. In spatial clustering, this is typically done in two-dimensional space and thus understood in terms of geometric *proximity*. When applied to feature attributes, clustering may often involve the same Euclidean distance measure, but the results are interpreted as *[dis]similarity*. Clustering involves the search for structures and grouping, and should not be confused with classification, which sorts unknown items into previously defined categories. Since clustering is the most frequent interpretation and implementation of the SOM method, it is useful to compare it to some of the more popular approaches, including hierarchical and *k*-means clustering. In this chapter, the three methods are juxtaposed after being applied to a data set of 32 attributes (mostly derived from population census data) for 50 US States and the District of Columbia (Figure 1.3).

(a) Hierarchical   (b) k-Means   (c) SOM
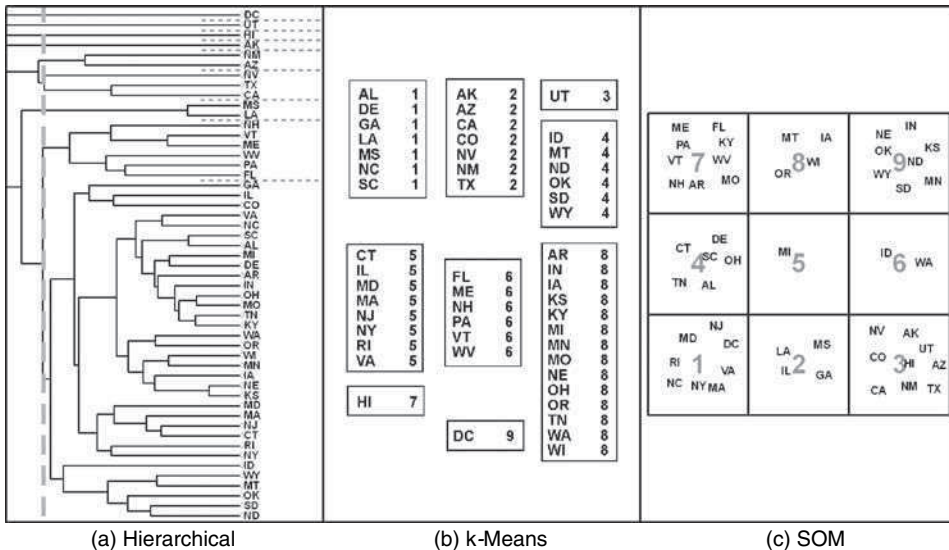
Figure 1.3   *Comparing three different clustering techniques applied to demographic data for US states*

Hierarchical clustering is the most widely known technique. It models distance and similarity relationships between input objects by turning each into a leaf node within

a binary tree structure. This tree is formed either by subdividing the full data set into smaller branch nodes until arriving at individual leaf nodes (divisive clustering) or by merging leaves into larger branch nodes (agglomerative clustering). The exact shape of the clustering tree is affected by the distance criterion used to evaluate candidate branch nodes before each merge (e.g. single-linkage) and by the distance measure used (e.g. Euclidean). The resulting tree structure can be visualized as a dendrogram, a portion of which is shown in Figure 1.3(a), where the average-linkage criterion and Euclidean distance measure are used. It can be seen that the hierarchical clustering tree contains multiple clustering solutions. To allow comparison with the other two methods, one solution (with nine clusters) is emphasized by applying a cut through the tree at the appropriate distance level. Horizontal, dashed lines indicated cluster separations. For example, California, Texas, and Nevada form one cluster, with Nevada joining the other two only just before the nine-cluster split and not long before New Mexico and Arizona are merged at a slightly coarser cluster level.

   One downside of hierarchical clustering is that feature space partitions can be far from optimal, since it attempts to compute all possible granularities at once. Compare this with *k-means clustering*, which looks for a partition based on a given number of clusters (k). As in the hierarchical solution, Utah, Hawaii, and Washington, DC are placed into their own 'clusters', but other states are more evenly distributed across the other six clusters [Figure 1.3(b)]. Like k-means, the standard SOM algorithm also assumes a fixed number of units and uses the same objective function as k-means clustering. However, it creates a topologically ordered partition. For example, the nine-cluster solution derives from a topologically ordered 3×3 grid of neurons [Figure 1.3(c)]. Cluster 1 is an immediate neighbour of cluster 2, while cluster nine is far away from either. Contrast this with the k-means solution, in which no indication of relationships between the nine clusters is given. For in-depth coverage of various clustering techniques, readers are referred to the numerous dedicated volumes on the subject (e.g. Sneath and Sokal, 1973).

1.2.3   Dimensionality Reduction Methods

Creating a topologically ordered partition of n-dimensional data in a form supportive of low-dimensional presentation implies that the SOM method performs some type of dimensionality reduction. This is already apparent in the case of the nine-cluster solution [Figure 1.3(c)], but becomes even more relevant as we move towards larger SOMs consisting of hundreds and even thousands of neurons. In such cases, a SOM will allow mapping out of individual, n-dimensional, data vectors in a low-dimensional display space. It is thus worthwhile to compare the SOM with other dimensionality reduction techniques. PCA is the most frequently used of these methods. The first two principal components often express enough of the multivariate structure of a data set that simple two-dimensional scatter plots are commonly found, illustrated here for the same demographic data used earlier [Figure 1.4(a)]. *Multidimensional scaling* (MDS) is the technique most appropriately fitting into the dimensionality reduction category, as it attempts to preserve high-dimensional distance orderings in low-dimensional space (Kruskal and Wish, 1978). While one can choose the output dimensionality as an input parameter, the two-dimensional form is by far the most common, since it supports
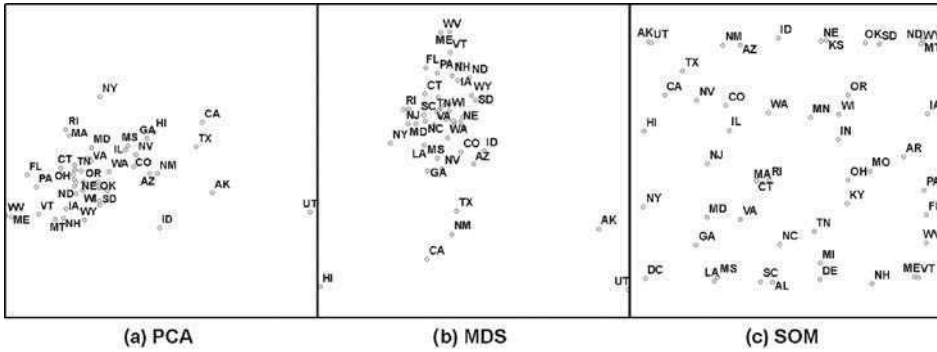
Figure 1.4 *Comparing three different dimensionality reduction techniques applied to demographic data for US states*

many different visualization forms, from traditional print media to interactive exploration [Figure 1.4(b)].

PCA and MDS employ an object-based conceptualization where the original input vectors are interpreted as discrete objects and are the sole basis of computation and visualization, which accordingly almost always consists of labelled point features [Figure 1.4(a) and (b)]. However, the SOM method conceptualizes input vectors not as discrete objects but as representative samples from an n-dimensional information continuum (Skupin, 2002b). During SOM training those samples drive a topologically ordered tessellated representation of that continuum. It is then not surprising that most SOM-based visualizations are constructed from a uniform cell structure resembling raster geometry. The field-like conceptualization implemented in SOM makes it easy to map various other n-dimensional vectors onto a trained SOM, even, and especially, if they were not part of the training data set. In order to allow for comparison with the PCA and MDS solutions, the trained SOM is here applied to the same vectors used during training [Figure 1.4(c)]. Notice especially the differences in the placement of outliers, like Utah or Alaska. The SOM's topology-preserving mapping makes more efficient use of the available space, at the cost of higher distortion of relative feature space distances as compared with PCA and MDS.

1.3 SOM ALGORITHM

Applications of SOM in geographic information science tend to employ the standard algorithm first described by Kohonen (1990). Therefore, it makes sense to spend some time in this chapter on introducing that algorithm. However, there already exist many good, formal descriptions of the algorithm, most notably in Kohonen's own monograph *Self-Organizing Maps* (Kohonen, 2001), and in various journal and conference proceedings articles. Readers are well advised to refer to those sources for the mathematical foundation and physiological justification of the algorithm. This introductory chapter instead presents the SOM method using mostly plain language and graphic illustrations, from pre-processing of training data to using the finished neural network.

1.3.1   Source data

The SOM method can be and has been applied to hugely diverse data sets, as will be evident from the collection of chapters in this book. Broadly speaking, one needs data containing individual items with n-dimensional, quantitative attributes. Raw source data will often already exist in that form. Well-structured data produced through a census of human population or multi-channel remote sensing are prime examples. On the other end of the spectrum, one can even use unstructured data, once they are suitably transformed. For example, a corpus of scientific articles can be turned into a SOM input data set after indexing and construction of a vector space model (Skupin, 2002a). From the analysis of mutual fund performance to classification of human voices, performing these transformations correctly is a crucial part of every analytical procedure. However, since the specific steps are highly dependent on the given subject domain, readers are advised to refer to domain-specific literature.

Pre-processing of n-dimensional vectors resembles typical procedures for other neural network methods. Two main concerns are the existence of skewed distributions and the range of values for each attribute. Neural networks tend to be fairly robust, but being aware of and, if feasible, counteracting the according effects will help to create a more useful model. When encountering highly skewed variables, logarithmic transformation is the first and most obvious choice. It is also a good idea to normalize all values of a given variable to fit into a predefined range, typically between 0 and 1.

1.3.2   Training the neural network

The SOM performs a 'non-linear, ordered, smooth, mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array' of neurons (Kohonen, 2001, p. 106). Each neuron has associated with it an n-dimensional vector of the same dimensionality as the input data. For example, if 32 census attributes for each of the states of the US are used as input, then a 32-dimensional vector is created for each neuron.

The first step in the creation of a SOM is to determine its size and topology type. The SOM's size k is given as the number of neurons to be used in the x and y direction. Thus, a size of two neurons in x and three neurons in y would yield six neurons, while a 100×100 neuron SOM would consist of 10 000 neurons. Two topology types are frequently used. The first is the square topology, where each neuron is connected to four neighbouring neurons. When used for visualization, a 10×10 neuron SOM would thus have a square shape overall [Figure 1.5(a)]. The second, and more frequently encountered, possibility is the use of a hexagonal topology, with six neighbours to every neuron. Given an equal number of neurons in x and y one would observe a rectangular shape, with the longer side along the x-axis [Figure 1.5(b)].

Before training can begin, n weights for each neuron are initialized. In order to later observe true self-organization, one could assign random values. Alternatively, it is possible to help the training algorithm along (and shorten training times) by assigning weights according to a linear estimate, such as the first two principal components derived from the training data. In some software solutions, this is one of the built-in options for SOM initialization (Section 1.4).

(a) Square Topology   *(b) Hexagon Topology*
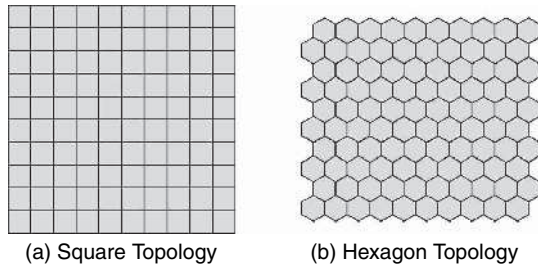
Figure 1.5 *Network size and topology type of a SOM are chosen before training begins. Notice overall shape difference for SOMs with identical size, but different topology type*

Training consists of an iterative process, during which individual input vectors are presented to the neuron grid, the best-matching (i.e. most similar) neuron vector is found and weights of that and other neuron vectors are modified. Understanding the nature of these modifications goes to the heart of understanding self-organization. Once the best-matching neuron is found, its n weights are modified towards an even better match with the input vector. In addition, neighbouring neurons up to a certain distance from the best-matching unit (BMU) are also modified to better fit that input vector. These focal modifications are over the course of many iterations causing similar input data to be associated with closely positioned neurons. In the literature, iterations are alternatively referred to as training steps, runs or cycles.

It is important to understand, especially in comparison with such methods as MDS, that relationships among input data are at no time directly assessed. Instead, topology preservation in a SOM is achieved as a quasi by-product of how weights of neuron vectors are adjusted during training. That is why *self-organization* is an appropriate title. A schematic example should serve to illustrate how this works (Figure 1.6). Let us assume that we were training a 3×3 neuron SOM with only four input vectors (1, 2, 3, 4). In feature space, these four vectors form two distinct clusters (1 and 4; 2 and 3). Starting with the initialized SOM, the first input vector finds the neuron at location ($x = 1$; $y = 2$) to be its BMU. Accordingly, weights of that node are adjusted towards the input vector. In addition, neurons within a single-neuron neighbourhood are slightly adjusted [Figure 1.6(b)]. For the next cycle, the second vector is presented to the neuron lattice, finds the neuron at ($x = 3$; $y = 3$) as its BMU, and adjusts weights for that neuron and its neighbours [Figure 1.6(c)]. Those adjustments cause the third
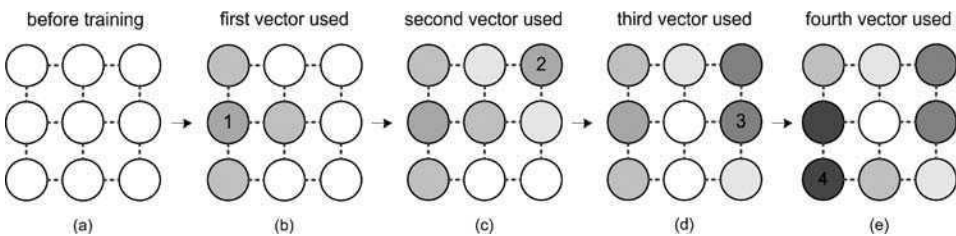
Figure 1.6 *Process of self-organization during SOM training. A 3×3 neuron SOM is trained with four observations representing two distinct groups in attribute space (See Colour Plate 1)*

vector to be drawn to the vicinity of the previous vector at $(x = 3; y = 2)$. Weights of its BMU and neighbouring neurons are modified. However, the single-neuron neighbourhood includes neurons that previously underwent modification on account of the first and second vector. Those neurons now undergo further modifications. Notice how the neuron at $(x = 2; y = 2)$ becomes a separator between cluster regions, since members of the two clusters have attempted to pull it in either direction [Figure 1.6(d)]. Finally, the fourth vector finds its BMU at $(x = 1; y = 1)$. The ensuing weight adjustments finish the self-organization of the SOM into two distinct clusters [Figure 1.6(e)]. This whole process would however be repeated many times over when dealing with real data. As training progresses, an input vector may then be reused and find a BMU that is different from the previous cycle it was involved in. For example, the first input vector may now find the neuron at $(x = 1; y = 1)$ to be a better fit. As a rule though, major global relationships will be established early, followed by a distinction of finer structures late during training.

To look at the training process more formally, let us consider the input data as consisting of n-dimensional vectors $\boldsymbol{x}$:

$$x = [\xi_1, \ldots, \xi_n]^{\mathrm{T}} \in \Re^n \tag{1.1}$$

Meanwhile, each of k neurons has an associated reference vector $\boldsymbol{m}_i$:

$$\boldsymbol{m}_i = [\mu_{i1}, \ldots, \mu_{in}]^{\mathrm{T}} \in \Re^n \tag{1.2}$$

During training, one x at a time is compared with all $\boldsymbol{m}_i$ to find the reference vector $\boldsymbol{m}_c$ that satisfies a minimum distance or maximum similarity criterion. Though a number of measures are possible, the Euclidean distance is by far the most common:

$$\|x - \boldsymbol{m}_c\| = \min_i \{\|x - \boldsymbol{m}_c\|\} \tag{1.3}$$

The best-matching unit and neurons within its neighbourhood are then activated and modified:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t)[x(t) - \boldsymbol{m}_i(t)] \tag{1.4}$$

One of the main parameters influencing the training process is the *neighbourhood function* (h_{ci}), which defines a distance-weighted model for adjusting neuron vectors. Two functions are most popular, the linear and the Gaussian model (shown here):

$$h_{ci}(t) = \alpha(t) \cdot e^{-d_{ci}^2/2\sigma_i^2(t)} \tag{1.5}$$

One can see that the neighbourhood function is dependent on both the distance between the BMU and the respective neuron (d_{ci}) and on the time step reached in the overall training process (t). The maximum of d_{ci} corresponds to the *neighbourhood radius*, which is a training parameter determining the set of reference vectors to be modified around each BMU at a particular time step [$N_c(t)$]. In the Gaussian model, that neighbourhood's size appears as kernel width (σ) and is not a fixed parameter. The neighbourhood radius is used to set the kernel width with which training will start. One typically starts with a neighbourhood spanning most of the SOM, in order to achieve a rough global ordering, but kernel width then decreases during later training cycles. Similarly, the initial

learning rate (α_0) is an input parameter, which is then gradually decreased as t progresses [$\alpha(t)$]. SOM training stops when a predetermined number of training cycles (t_{max}) are completed.

As self-organization progresses during training and neighbourhood radius and training rate slowly decrease, the SOM gradually settles into a stable configuration (Figures 1.7 and 1.8). One way of visualizing this is to show the weights of a particular variable for each neuron and observe changes over multiple training cycles. In Figure 1.7 'vacant housing' as one of 32 census variables is shown on a 20×20 neuron SOM, with snapshots at four time periods. Training begins with randomized weights. Early cycles establish major global relationships, visible here as an almost linear relationship between vacant housing and the *x*-axis after 40 000 cycles. After 80 000 cycles, more detailed structure emerges, with low values of vacant housing in the centre-left portion of the SOM. Finer, local structures emerge during the remaining cycles, with training ending at 100 000 cycles. Alternatively, the training progress could be visualized by plotting individual training vectors onto the trained SOM at chosen cycling intervals according to the location of the best-matching unit. These temporal vertices are then linked to form trajectories (Figure 1.8). With snapshots taken every 10 000 cycles, one can see how the SOM settles towards the end of training, as indicated by a lack of major movement after about 80 000–90 000 cycles.

random initialization   40000 training runs   80000 training runs   100000 training runs
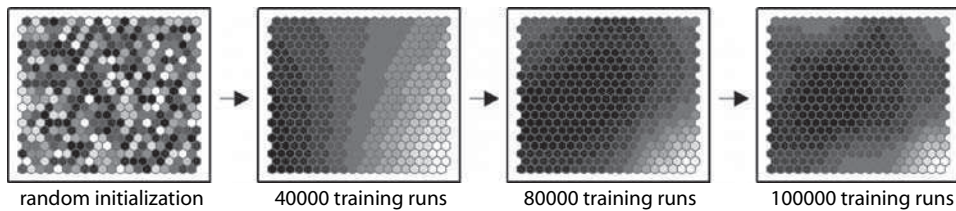
Figure 1.7 *Changes to the component plane for the variable 'vacant housing' during training of a SOM with demographic data for US states. Higher values indicated by lighter shading*

Most of the parameters mentioned here can be specified before beginning the training process, including the network size, topology type, distance function, neighbourhood function, neighbourhood radius, total number of training cycles, and training rate. The ability of directly influencing these parameters constitutes much of the difference between the various SOM implementations, including those found in commercially available software.

1.3.3   Using the trained neural network

Once training is finished, the neural network is ready for use. First, it is advisable to visualize the SOM itself, and sometimes this alone already justifies use of the SOM method. Another main mode of using a trained SOM is to map individual *n*-dimensional features into low-dimensional space by finding the best-matching neuron. For the purpose of interactive, exploratory analysis, SOM can also be linked to other visualizations, including geographic maps.
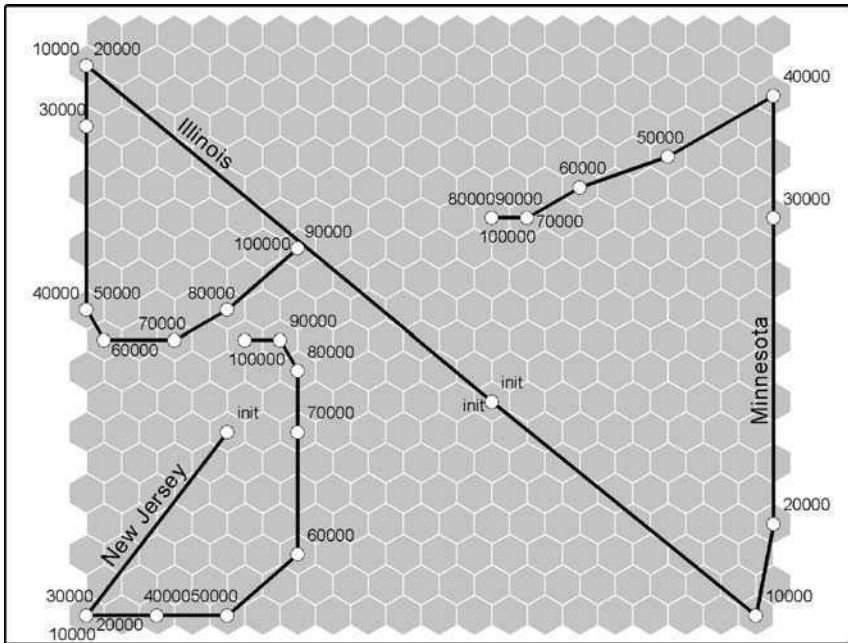
Figure 1.8   *Position changes of three states while training a SOM for a total of 100 000 runs. Starting with a randomly initialized SOM, each state is recorded once every 10 000 runs*

1.3.3.1   Visualizing the SOM

The main methods for visualizing a SOM involve *component planes, distortion patterns*, and *clustering. Component plane* visualization symbolizes neuron weights for individual variables. For example, with census data one could create separate visualizations for each input variable, such as population density, percentage of Hispanic population, and so forth. One of the possible applications is to look for relationships between variables, based on visual similarity of component planes. The most common approach to component plane visualization is to apply colour or grey shading, as seen in the visualization of vacant housing (Figure 1.7). Other possibilities include the use of graduated symbols for individual variables or the placement of bar charts to show the weights for multiple variables at each neuron.

While SOM training has the effect of preserving major topological relationships, geometric proximities can be drastically distorted. This refers particularly to contraction effects observed for sparsely occupied or empty feature space regions and expanded representation of high-density regions (Lin *et al.*, 2000). In more general terms, one can state that low-density and high-density regions in feature space are associated with marked distortion when they are modelled in the low-dimensional space of topologically ordered neurons. One common approach is to visualize the degree of distortion, i.e. the change in relative distance between *n*-dimensional locations and their low-dimensional representation, and treat zones of high contraction as a type of cluster boundary. Identifying these 'clusters' can be rather subjective though, especially when different magnitudes of distortion are encountered in different regions of a SOM, since it

is up to the human observer to decide when visual structure constitutes a cluster boundary. The most frequently used method to visualize distortion patterns is the U-matrix method (Ultsch, 1993), which explicitly symbolizes the n-dimensional distance of neighbouring neurons.

A third approach to visualizing the SOM itself is to treat each neuron as a distinct feature possessing an n-dimensional vector, to which traditional cluster techniques, like hierarchical or k-means, can be applied. Since the neurons are already topologically ordered, one will find that such n-dimensional clusters tend to form regions in two-dimensional SOM space. This can be especially useful with high-resolution SOMs, for example to enable multi-scale, interactive exploration (Skupin, 2002a).

1.3.3.2   *Mapping data onto the SOM*

Visualizing a SOM means exploring the model itself that one has created through neural network training. However, *applying* the model will often involve the mapping of n-dimensional vectors *onto* the trained SOM. The bulk of SOM applications are focused on this aspect of Kohonen's method. For example, in industrial applications, one could track a machine part based on various measurable attributes. In an analysis of voting behaviour of different politicians, one could map individual persons in two dimensions, as an alternative to MDS, which has traditionally been used for this purpose. Geographic objects can also be mapped onto a SOM, as shown in Figures 1.3(c) and 1.4(c). Those figures also illustrate the difference between using a SOM for classification into a limited number of classes [Figure 1.3(c)] and spatial layout with differentiated locations for many features [Figure 1.4(c)]. Speaking of clustering, please note that for supervised classification one should not use the SOM method itself but a related method called *learning vector quantization* (Kohonen, 2001).

When input features can be arranged into meaningful sequences, output locations derived from the locations of best-matching units can be strung together to form trajectories. This has been demonstrated for multi-temporal data, where the same features and attributes are observed for multiple time periods (Deboeck and Kohonen, 1998; Skupin and Hagelman, 2005). Other possibilities include the mapping of space–time paths onto a SOM trained with the attributes of geographic features (see Chapter 6). Even the training process itself can be visualized via trajectories (Figure 1.8).

1.3.3.3   *Linking SOM to other visualizations*

In most circumstances, SOMs will not become the sole analytical tool for investigating an n-dimensional data set. Instead, it constitutes an additional method that will be used in conjunction with other computational and visual tools. Integration of a SOM with other forms of representation is becoming increasingly important, especially when dealing with geographic data, for which a dominant visual form already exists in the form of geographic maps. Integration of most SOMs with more traditional geographic visualizations is straightforward since a two-dimensional SOM can be readily represented using standard GIS data structures. There are obvious advantages to doing this in an interactive setting, but it can even be useful for static cartographic output (Figure 1.9). Notice how
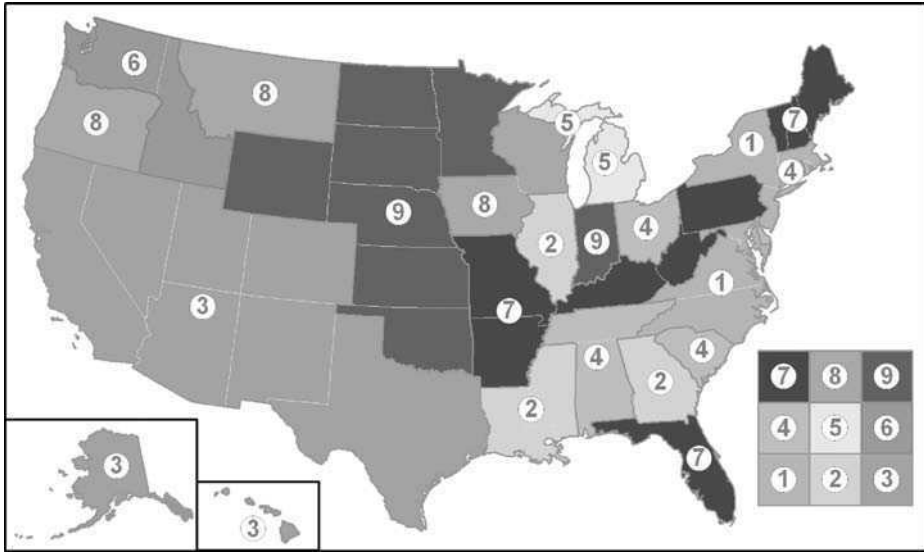
Figure 1.9 *SOM-based clustering of census data combined with colour design informed by network topology. Relationships among clusters are indicated by displaying legend constructed from two-dimensioned SOM geometry (See Colour Plate 2)*

the low-dimensional, topologically ordered set of neurons helps with the design and presentation of a colour legend, in addition to its clustering function.

Side-by-side display of SOM and geographic map in a commercial off-the-shelf GIS was demonstrated as early as 1998 (Li, 1998), but little progress has been made since. In fact, most geographic SOM applications still suffer from a lack of imagination regarding integration, as indicated by the preponderance of run-off-the-mill graphic output from dedicated SOM software (especially the SOM Toolbox for Matlab) in the GIScience literature. One example for a different direction is the integration of SOM training and visualization in the GeoVISTA Studio system (Gahegan *et al*., 2002). An interactive approach mostly unexplored is the linking of different SOM solutions, for example linking a high-resolution SOM to one consisting of a small number of neurons.

1.3.4   Extensions of the SOM method

Beyond the widely used approach presented in the previous sections, numerous extensions and modifications of the SOM method have been proposed. Most of these have not yet found their way into end-user SOM implementations, including those described in this book. Nonetheless, many of these modifications do indeed improve upon Kohonen's original method in a number of ways.

One example is the notion of a *growing SOM* described by Fritzke (1999). Instead of a fixed network size, this approach provides for the insertion of new nodes into the neural network in response to specific nodes not satisfying a certain objective function. For example, one could use the *quantization error* as such an objective function, i.e.

how well a given node actually fits those data items for which it has become the best-matching unit. New nodes would be inserted into the neighbourhood of existing nodes that have high quantization errors. One could envisage numerous variants of this approach, for example using an entropy maximization goal instead of the error minimization described above (Fritzke, 1999). Conversely to the adding of nodes, growing SOMs may also involve the removal of individual nodes, if some objective criterion is thereby improved.

A frequently encountered issue with the traditional SOM algorithm is the appearance of border or edge effects. The most obvious of these is that neurons near the edge of the SOM come to represent larger portions of the n-dimensional input space. In other words, there is a large degree of compression or contraction at the edge as compared with inner neurons. This becomes most obvious during the mapping of large numbers of vectors onto the SOM, where one frequently observes pronounced clustering along the edges. Efforts to address this issue include the growing SOM discussed above and arranging neurons into lattices that are not flat, but curved such that edges are reduced (e.g. forming a cylinder). Arrangement of neurons on a sphere would completely eliminate edges, which is why a number of proposals concerning *spherical SOMs* have been put forward (Ritter, 1999; Sangole and Knopf, 2002; Wu and Takatsuka, 2005).

Other proposals have dealt with how the neighbourhood around the best-matching unit is defined during training. Teuvo Kohonen himself introduced a number of modifications and new methods building on the original SOM approach, many of which are documented in his monograph (Kohonen, 2001). For example, *learning vector quantization* (LVQ) adapts many of the same principles to provide super-vised classification. *Adaptive subspace SOM* (ASSOM) and *hierarchically structured SOMs* are other proposals that have received attention in recent years. Bação *et al*. (see Chapter 2) discuss a number of SOM variants that are of particular interest to GIScience.

1.4   SOFTWARE TOOLS FOR SOMs

There are a number of software tools available for SOMs, in the commercial as well as the public domain. This section will specifically discuss the ones that are commonly employed in geographical analysis, including by the authors in this volume. This is intended as a pointer to the various software options available as a resource for GIScientists and is not meant as a comprehensive, self-help guide to these tools. For this, the reader is advised to refer to documentation and help files, which tend to be freely downloadable for public domain software or are included in the licensed versions of commercial software. Additionally, readers can refer to other volumes, such as Kohonen (2001) and Deboeck and Kohonen (1998), that include overviews of SOM software. This section also intends to provide an updated review of some of the software that was included in these previous volumes, such as SOM_PAK. Most applications described in this book are based on using public domain software, although a few, such as Bação *et al*. (Chapter 2), use self-coded variants of SOM that extend on its basic functionality in order to accommodate the specific nature of geographic data. In some cases, stand-alone software such as SOM_PAK is

used in combination with GIS software for data analysis as well as visualization of SOM output.

1.4.1   Stand-alone Software

One of the most widely used implementations of the basic SOM algorithm is found in the SOM_PAK program (Kohonen *et al*., 1996), which is freely available from the Neural Networks Research Centre of the Helsinki University of Technology (http://www.cis.hut.fi/research/som_pak/). First made public about 10 years ago, many researchers had their first practical experience with SOM using this software. While its command-line interface can at first seem daunting, the involved commands and underlying methods are actually well explained in the accompanying documentation. The simple user interface makes it easy to port the software to various hardware platforms. The complete source code, written in C, is accessible to programmers for noncommercial uses. For example, one could implement alternative similarity coefficients, since only Euclidean distance is built in (Skupin, 2003). Novice, nonprogrammer users will be thankful that executable files for Windows are available for download, because compilation of the source code can be tricky.

Use of SOM_PAK involves four main steps: map initialization, map training (in multiple stages if desired), evaluation of quantization error, and visualization. A major downside to using SOM_PAK is the lack of convenient visualization capabilities beyond the creation of static output in PostScript format. However, the trained SOMs are stored in straightforward text files, known as codebook files (.cod), which can easily be read by other packages. In fact, the codebook format has become a de facto standard for the distribution of SOMs. Many of the software packages listed below offer respective import options (e.g. SOMine, Nenet, SOM ToolBox). It also does not take particularly advanced programming skills to turn the codebook files into something usable within GIS. For example, many of the figures in this chapter were created by transforming the content of SOM_PAK's codebook files into ArcInfo Generate files, before performing all further transformations inside of ArcGIS.

Viscovery SOMine is a commercial product of Eudaptics Inc. (http://www.eudaptics.de) that is distributed as a Windows application in several versions, with different capabilities depending on price. Built around a custom version of the SOM training algorithm it provides for the training and use of SOMs, including clustering, prediction, and exploratory visualization, all in a highly interactive environment. Another interesting SOM tool is *Nenet* (Neural Network Tool), which first became available in 1998, and which provides a full graphical user interface for training and visualization. A limited functionality version remains freely available (http://koti.mbnet.fi/~phodju/nenet/Nenet/General.html), but the current status of the full software is unknown.

1.4.2   Add-In Software and Software Components

When faced with complex exploratory visualization tasks, SOMs are best used in conjunction with other visualization and analysis tools. As an alternative to stand-alone

SOM software, one will often find that the addition of SOM capability to existing analytical tools and integration into existing software development architectures is more useful.

The *SOM Toolbox for Matlab* was developed by researchers at the Helsinki University of Technology in recognition of the large numbers of users performing numerical modelling and analysis in Matlab. While the latter is a commercial product, the SOM Toolbox is public-domain software consisting of a large number of Matlab routines (.m files). Download (http://www.cis.hut.fi/projects/somtoolbox) and installation of the toolbox is straightforward. Some of the most common training and visualization options are accessible through a limited graphical user interface, but full functionality and control over training and visualization require use of Matlab's command-line interface. The sequence of training steps is similar to SOM_PAK, though a number of further options are available. When it comes to visualization, the SOM Toolbox offers an attractive variety of methods and plenty of opportunity for customization. Low cost, ease of SOM training and visualization, and a large Matlab user base have made the SOM Toolbox one of the most popular SOM solutions.

The full potential of the SOM method for interactive, exploratory visualization is most easily tapped when SOM functionality can be integrated into larger visualization software architectures. However, SOM component solutions, for example based on ActiveX or JavaBeans, are still hard to find. One exception is the set of SOM beans included with GeoVISTA Studio (http://www.geovistastudio.psu.edu), an open-source visualization software environment developed at Pennsylvania State University (Gahegan *et al.*, 2002). GeoVISTA Studio comes with a large number of beans, including for data input, preprocessing, numerical analysis, and visualization. More importantly, it provides a code-free programming environment, where the flow of data is directed through visual manipulation of links between different beans. While this process does not involve writing code directly, previous knowledge of Java and JavaBeans greatly helps and novices need significant practice to make meaningful use of GeoVISTA Studio. One notable advantage of the system is that the interactive visualizations constructed through inclusion and wiring of beans can immediately be tested, even before being deployed as applets or applications. GeoVISTA Studio's SOM beans provide training and various visualization methods, including two- and three-dimensional U-matrix visualizations. Integration with other methods, such as geographic maps, scatter plot matrices, or parallel coordinate plots, does not only allow linked selection, but also linked symbolization, where colour choices are shared among multiple methods.

1.5   GISCIENCE AND SOMS

As a method developed outside of GIScience, it is natural to see the relationship between the SOM and GIScience primarily as the latter adopting the former to serve specific analytical purposes. Clustering, visualization, and the range of applications laid out in the various chapters of this book is indeed impressive. However, this does not have to be a one-way relationship. GIScientists may offer a unique perspective and contribute to the further development of this popular neural computing approach. This begins with the storage and manipulation of trained SOMs leveraging the ability of GIS to handle

complex geometric and attribute structures. There has so far been little reflection on how such geographic notions as *scale* become relevant when dealing with nongeographic data or how geographic conceptualizations of *fields* versus *discrete objects* are manifested in different spatializations (Skupin, 2002b). The design of SOM-based visualization stands to gain from traditional cartographic design considerations, for example regarding visual hierarchies and semiotic variables. Meaningful interaction with large SOMs will benefit more from semantically driven notions of scale dependence used in cartography than from the performance-oriented level-of-detail (LOD) approaches common in computer graphics. The notion of GIScience itself was conceived in recognition of the growing need to pursue interdisciplinary strategies, and our community is now actively engaging the SOM method accordingly. This book presents the current state of this endeavour.

1.6   ORGANIZATION OF THE BOOK

The primary aim of this book is to act as a showcase for the valuable role that SOM can play in geographic analysis. This book is about solving academic, theoretical, and applied problems and converting interesting computational methods into useful operational tools. It is also about finding new uses and about providing novel solutions to established problems. The sequence of chapters reflects this view of SOM as a method of many colourful facets.

The current chapter provides an introductory overview of principles, algorithms, and tools associated with the SOM. The method is explored further in Chapter 2 by Bação *et al.*, who demonstrate multiple SOM variants designed to address the specific nature of geographic data and problems. In Chapter 3, Koua and Kraak demonstrate the use of SOM to reveal hidden patterns within an integrated, exploratory visualization environment. In Chapter 4, Thill *et al.* describe work with a linguistic database where the SOM method is used to mine and visualize latent organization rules within the data. That chapter also demonstrates the development of an integrated environment for exploration of SOM output within a standard Windows-based GIS platform. Yan and Thill, in Chapter 5, use SOM as an exploratory data mining tool for spatial interaction data to visualize flows and movements in space within an interactive environment. An extension to the traditional notion of space–time paths is presented by Skupin in Chapter 6, where movement across geographic space is linked to simultaneous movement in n-dimensional attribute space and visualized as a SOM trajectory. Sester, in Chapter 7, is concerned with issues of typification in cartographic generalization, where the density-preserving tendencies of SOM training can be exploited for multi-scale mapping. Hewitson, in chapter 8, demonstrates the diversity of applications supported by the use of the SOM method for climate analysis. Kropp and Schellnhuber, in Chapter 9, then introduce an approach designed to derive global biogeographical prototypes that could be used in climate impact studies. Doucette *et al.*, in Chapter 10, report on experiments aimed at extracting road features from high-resolution, multi-spectral imagery. Finally, an epilogue by Goodchild (Chapter 11) weaves together the diverse strands of SOM applications described in this book, drawing connections to developments in such areas as exploratory spatial data analysis (ESDA) and commenting on how the SOM method relates to the state and future of GIScience.

ACKNOWLEDGEMENTS

REFERENCES

Deboeck, G., and T. Kohonen, eds. 1998. *Visual Explorations in Finance with Self-Organizing Maps*. Berlin, Heidelberg, New York: Springer-Verlag.

Fischer, M. M. 2001. Computational Neural Networks – Tools for Spatial Data Analysis. In *Geocomputational Modelling: Techniques and Applications*, eds M. M. Fischer and Y. Leung, 15–34. Berlin, Heidelberg, New York: Springer-Verlag.

Fischer, M. M., and Y. Leung. 2001. *Geocomputational Modelling: Techniques and Applications, Advances in Spatial Science*. Berlin, New York: Springer-Verlag.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis*. London, Thousand Oaks, CA: Sage Publications.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons, Ltd.

Fritzke, B. 1999. Growing Self-Organizing Networks – History, Status Quo, and Perspectives. In *Kohonen Maps*, eds E. Oja and S. Kaski, 131–144. Amsterdam: Elsevier.

Gahegan, M., M. Takatsuka, M. Wheeler, and F. Hardisty. 2002. Introducing GeoVISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems* 26:267–292.

Gurney, K. 1997. *An Introduction to Neural Networks*. London: UCL Press.

Hertz, J., A. Krogh, and R. G. Palmer. 1990. *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley.

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE* 78(9):1464–1480.

Kohonen, T. 2001. *Self-Organizing Maps*. Berlin: Springer-Verlag.

Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen. 1996. *SOM_PAK: The Self-Organizing Map Program Package, Technical Report A30*. Espoo: Helsinki University of Technology, Laboratory of Computer and Information Science.

Kruskal, J. B., and M. Wish. 1978. *Multidimensional Scaling, Sage University Papers Series. Quantitative Applications in the Social Sciences. 11*. Beverly Hills, CA: Sage Publications.

Li, B. 1998. Exploring spatial patterns with self-organizing maps. Paper presented at GIS/LIS '98, 10–12 November, Fort Worth, TX.

Lin, C., H. Chen, and J. F. Nunamaker. 2000. Verifying the proximity and size hypothesis for self-organizing maps. *Journal of Management Information Systems* 16(3):57–70.

Longley, P. 1998. *Geocomputation: A Primer*. Chichester, New York: John Wiley & Sons, Ltd.

Miller, H. J., and J. Han, eds. 2001. *Geographic Data Mining and Knowledge Discovery, Research Monographs in Geographic Information Systems*. London, New York: Taylor & Francis.

Openshaw, S., and C. Openshaw. 1997. *Artificial Intelligence in Geography*. Chichester, New York: John Wiley & Sons, Ltd.

Openshaw, S., and R. J. Abrahart, eds. 2000. *GeoComputation*. London, New York: Taylor & Francis.

Pijanowski, B. C., D. G. Brown, B. A. Shellito, and G. A. Manik. 2002. Using neural networks and GIS to forecast land use changes: a land transformation model. *Computers, Environment and Urban Systems* 26(6):553–576.

Ritter, H. 1999. Self-Organizing Maps on Non-Euclidean Spaces. In *Kohonen Maps*, eds E. Oja and S. Kaski, 97–109. Amsterdam: Elsevier.

Rogerson, P. A. 2001. *Statistical Methods for Geographers*. London: Sage.

Rumelhart, D. E., and J. L. McClelland. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. 2 vols, *Computational Models of Cognition and Perception*. Cambridge, MA: MIT Press.

Sangole, A., and G. K. Knopf. 2002. Representing high-dimensional data sets as closed surfaces. *Information Visualization* 1(2):111–119.

Skupin, A. 2002a. A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications* 22(1):50–58.

Skupin, A. 2002b. On Geometry and Transformation in Map-Like Information Visualization. In *Visual Interfaces to Digital Libraries (Lecture Notes in Computer Science 2539)*, eds K. Börner and C. Chen, 161–170. Berlin: Springer-Verlag.

Skupin, A. 2003. A novel map projection using an artificial neural network. Paper presented at 21st International Cartographic Conference, 10–16 August, Durban, South Africa.

Skupin, A., and R. Hagelman. 2005. Visualizing demographic trajectories with self-organizing maps. *GeoInformatica* 9(2):159–179.

Sneath, P. H. A., and R. R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco: W. H. Freeman.

Ultsch, A. 1993. Self-Organizing Neural Networks for Visualization and Classification. In *Information and Classification: Concepts, Methods, and Applications*, eds O. Opitz, B. Lausen and R. Klar, 307–313. Berlin, Heidelberg: Springer-Verlag.

Wu, Y., and M. Takatsuka. 2005. Geodesic self-organizing map. Paper presented at Conference on Visualization and Data Analysis, 17–18 January, San Jose, CA.

2

Applications of Different Self-Organizing Map Variants to Geographical Information Science Problems

Fernando Bação[1], Victor Lobo[1,2] and Marco Painho[1]
[1] *ISEGI/UNL, Campus de Campolide, 1070-312 Lisboa, Portugal*
[2] *Portuguese Naval Academy, Alfeite, 2810-001 Almada, Portugal*

2.1   INTRODUCTION

The availability of methods able to perform intelligent data reduction is a central issue in science generically and GIScience is no exception. The need to transform into knowledge the massive digital geo-referenced databases has stimulated work in a number of research areas. It has also led GIScientists to search for new tools, which are able to make sense of such complexity. The field of knowledge discovery in databases (KDD) has proposed a number of tools that may help deal with this problem. However, adapting those tools to the specific context of GIScience remains a research challenge (Openshaw and Openshaw, 1997; Openshaw 1999).

   Self-organizing maps (SOMs) have been proposed as a step forward in the improvement of data reduction tasks (Openshaw and Wymer, 1995) and have been used, with good results, to address different GIScience problems as we shall see later in this chapter. In general, these applications are based on the original SOM algorithm, but another important research problem is to seek ways to adapt the algorithm to the specific needs and paradigms of GIScience. In fact, the possibilities of altering the original SOM

algorithm are numerous, and this flexibility can be used to take into account the particular perspective of GIScience and the special features of geoinformation.

The main objective of this chapter is to present a structured view of the possible modifications of the original SOM algorithm in order to develop SOM variants which are relevant to GIScience.

We shall start by reviewing some important parameterizations of the SOM algorithm, and then proceed to discuss how the different steps of the algorithm may be changed. We will then review some of the most relevant possible variants. Finally, we will explain in detail the Geo-SOM variant, together with an example of its application to an artificial data set and to a geo-referenced data set.

2.2 SOME IMPORTANT PARAMETRIZATIONS OF THE ORIGINAL SOM ALGORITHM

Although they do not constitute true variants of the original algorithm, some parameterization choices can radically change the way the SOM may be used. We will discuss three of them, namely: (1) size of the map; (2) output space dimension; and (3) training schedule.

2.2.1 Size of the Map

There are three major options in terms of the size of the SOM. The first one is to build very large SOMs in which the number of neurons is greater than the number of input patterns (Ultsch and Siemon, 1990; Ultsch and Li, 1993). The second and by far the most common option is to build a medium sized map, smaller than the number of input patterns, but still large enough to have a few units representing each cluster existing in the data (Kohonen, 2001). Finally, there is also the possibility of building small maps where the number of units is drastically smaller than the number of input vectors, usually with only one unit for each expected cluster (Bação *et al*., 2004a). The relevance of the choice of the size of the SOM is such that it can be argued that SOMs of significantly different sizes constitute different tools, which may be used to perform different tasks.

When opting for a larger map the underlying assumption is that we wish to explore in detail the underlying distribution of the data. By using more units than input patterns it is possible to obtain very large U-matrices (Ultsch and Siemon, 1990) on which distances between input patterns can easily be identified. This can be seen as a strictly exploratory exercise. The data reduction, in this case, is solely based in projecting the *n*-dimensional space onto a one-, two- or three-dimensional (1-, 2- or 3-D) space.

The decision to build a medium sized map can be seen as a compromise, in the sense that although reducing the number of dimensions and creating clusters, it still enables the user to understand the basic (or broad) distribution of the data, eventually leading to further and more severe reductions.

Finally, small maps are used when the user is interested in clustering data without concerns about the detailed analysis of its distribution. In this case the primary objective is to form clusters of input patterns which are as similar as possible, aiming at a one step

substantial data reduction. In this context the U-matrix is of little value, and component planes (Kohonen, 2001) become more relevant as they allow a simple description of the resulting clusters.

2.2.2   Output Space Dimension

The output space can have as many, or even more dimensions than the input space. Nevertheless, the output space seldom has more than two dimensions, because it is difficult to visualize high dimensional data. Theoretically, the appropriate dimension of the map should be defined by the intrinsic dimension of the data (Camastra, 2001; Fukunaga and Olsen, 1971). The intrinsic dimension of an n-dimensional data set is m ($m < n$) if it is possible to represent all the data with only m independent variables, i.e. if the data lie on an m-dimensional surface. Thus, by using only the intrinsic dimension of the data we remove redundancy in the number of independent variables used. Estimating the intrinsic dimension is still a largely unresolved problem in most practical cases. Since it is difficult to confirm the true dimensionality of the data, and since the possibility of visualizing the results is very useful, 1-D or 2-D maps are usually preferred.

When the objective is to cluster data based on very small SOMs, the best approach is to use a 1-D SOM. This is due to the fact that the plasticity (Carpenter and Grossberg 1988) of the 1-D map is much greater than a 2-D SOM (Bação *et al*., 2004a). This fact is apparent in the application to the traveling salesman problem (Maenou *et al*., 1997) where 1-D SOMs are preferred to 2-D SOMs. The need to closely represent a number of points that can form complex geometric shapes render inefficient the use of 2-D SOMs. However, if the objective is to obtain a comprehensive visualization of the input space, then a 2-D SOM is to be preferred. The rationale is that a higher level of connectivity will yield a better coverage of the input space.

Finally, it is important to note that any SOM will produce a bias in the representation of the input space. In fact, the distribution of the classification resources (units) will be more than proportional in lower density areas. This effect is usually referred to as the 'magnification effect' (Claussen, 2003; Cottrell *et al*., 1998). The quantification of this effect has proven to be elusive and is still an unresolved issue.

2.2.3   Training Schedule

The first step in building a SOM involves giving initial values to the units. This may be done using completely random values (which usually leads to slow convergence towards the general area of the data), or using values obtained from randomly selected input patterns. This type of initialization will usually produce maps which take a long time to unfold, or may not unfold at all (Kohonen, 2001). Better maps are usually obtained if the units are laid out on a 2-D plane and centred near the mean of the input patterns. The plane may be defined, for example, by the two first eigenvectors of the input patterns.

The counting of training iterations may also vary from one implementation to another. While in the case of SOM_PAK (Kohonen *et al*., 1995) each presentation of an input pattern is counted as an iteration, and the learning parameters are adjusted after each iteration, most implementations count 'epochs' (instead of iterations) presenting the whole

training data set before adjusting the training parameters. In the latter case, the units may be updated after each input pattern is presented. This is termed 'on-line' or 'sequential mode'. Alternatively, changes may be stored and applied only after the whole training set is presented. This is termed 'batch mode'.

The way the learning rate and neighborhood radius varies during training can also be done in different ways. For the SOM to converge to a stable configuration it is necessary to decrease the learning rate to 0. Although this may be done in many ways, we do not know of any conclusive analysis of its impact. Some authors have proposed dynamically changing learning rules that compensate the magnification effect of the standard SOM (Cottrell *et al.*, 1998). In some applications, particularly in on-line system monitoring, it is desirable to maintain some plasticity when using the SOM, and so the learning rate does not converge to 0. The final value of the neighborhood radius can also have a dramatic effect on the final map. If this radius is allowed to decrease to 0, the final stages of training will be equivalent to a k-means algorithm, and thus locally optimal. If instead the radius decreases to 1, then the units will always be pulled away from the local minima, and on the borders of the map they will be pulled towards the center because there are no units pulling them outside the map. Both approaches make sense in different contexts, so care must be taken when choosing these values.

2.3   WAYS TO CHANGE THE ORIGINAL SOM ALGORITHM

Several reviews of the different variants to the standard SOM algorithm (Kangas *et al.*, 1990; Vesanto, 1999, 2000) have been published. We have identified three main areas where changes to the basic SOM algorithm can be made:

1. topology and connections between units;
2. matching and voting mechanism (calculation and voting phases);
3. learning rule (update phase).

We shall now proceed to discuss each of these areas individually. However, it is possible that any single implementation of a SOM will include a combination of changes in these different areas.

2.3.1   Topology and Connections between Units

In the standard SOM the units form some type of regular grid. Neighboring units in this grid are influenced by each other, and thus it can be considered that there is a connection between them. Some variants of SOM alter these connections between neighboring units or eliminate them altogether.

In the Neural Gas Architecture (NGA) (Martinetz *et al.*, 1993) each unit is completely independent of the others. Therefore, the units in NGA do not form a distinct output space. Since there is no output space, this variant cannot be used for mapping or projection purposes, but it can be used for sampling or clustering. The lack of output space forces neighborhoods to be calculated in the input space. This means that during the update

phase of the training algorithm the units are ordered by their distance (in the original input space) to the winning neuron, and updated accordingly.

Another approach is to maintain connections between units, but relax the constraint that they form a regular grid, as happens with the Growing Cell networks (Fritzke, 1991). In this type of neural network units are inserted, one at a time, during the training phase, according to some established criteria. The resulting network can be quite irregular, and since the number of connections in each unit depends on how many units were inserted next to it, no simple output space will be formed. This network gave rise to a family of related architectures, namely the Growing Neural Gas (Fritzke, 1994) and the Growing Grid (Fritzke, 1995) that allow connections between units to be established or broken during training.

Even when the units do form a regular grid, some SOM variants allow growth in the number of units (Almeida and Rodrigues, 1991) or in the number of dimensions (Bauer and Villmann, 1997).

Another form of interaction between units, even stronger than that imposed by the neighborhood effect, is to allow units to receive as inputs the outputs of other units. This happens in SOMs with feedback used in temporal data analysis (Guimarães *et al.*, 2002), where units receive as inputs the delayed outputs. This also happens in hierarchical SOMs, which we will discuss later in this chapter.

2.3.2   Matching and Voting Mechanism (Calculation and Voting Phases)

A large number of variants of the basic SOM relate to the way the matching between input patterns and units is made, and how the best matching unit (BMU) is selected. A trivial way to change the matching mechanism is to use metrics other than the standard Euclidean distance, and many such metrics have been used (Kohonen, 2001). More interesting variants of the basic SOM can be obtained if the units are allowed to have an internal structure that is different from the input patterns. In this case, the units cease to be points in the input space. One such variant is the Adaptive Subspace SOM (ASSOM) (Kohonen, 2001), where the units, instead of being points in the input space, are whole subspaces, i.e. subsets (of lower dimensionality) of the original input space. In this case the matching is done by calculating the distance from the input pattern to the nearest point in that subspace. In temporal SOMs it is relatively common to find delay elements associated with the map units, and matching is done using those delays or past activations of the units (Guimarães *et al.*, 2002). The matching may also be done by finding the fitness of the input pattern to some given criteria that may be stored in the map units.

In some approaches, only a sub-set of units are searched to find the BMU. This happens when spatial or temporal restrictions are imposed (Chandrasekaran and Liu, 1998; George, 2000; Kangas, 1990, 1992), when tree structures are used to accelerate the search, or when certain supervised versions of SOM are used (Buessler *et al.*, 2002; Ritter *et al.*, 1992).

2.3.3   Learning Rule (Update Phase)

In the basic SOM, each unit is updated according to its distance (in the output space) to the BMU, and according to its distance in the input space to the input pattern. The distances

in the input space of units that are neighbors in the output space may vary widely. In particular, if there are large differences in the density of input patterns throughout the input space, there will be regions where neighboring units are close to each other, and others where they are far away. While this is a desirable feature when trying to obtain good U-matrices (Ultsch *et al.*, 1993), it will make the SOM concentrate the units in areas with greater density, leaving very few units to map areas of low data density. This will not be a desirable feature if we want to avoid overspecialization and wish to keep some units available to detect new features or outliers. This line of thought led to the Visualization Induced SOM (ViSom) (Yin, 2001), where a repulsion force is introduced between units, forcing a certain minimum distance between them. It is argued that this approach will provide a 'broader view' of the input space. Clusters will still be detectable in a ViSom by analyzing variations in the number of patters that are mapped to different units.

The direction in which each of the units is updated is usually that of the input pattern. One alternative is to move the unit in the direction of the nearest unit, as was proposed by Lee *et al.* (2001). This resembles the way nodes are pulled in a fisherman's net, and thus this update rule was dubbed 'fisherman's rule'. It has been shown (Lee *et al.*, 2001) that this will improve the convergence speed in the first iterations of the learning phase. The main reason for this is that, at each learning step, the different units will not be attracted in exactly the same direction (the direction of the input pattern), but will instead be pulled in a direction that depends on their immediate neighbors. It is suggested (Lee *et al.*, 2001) that the fisherman's rule be used in the first iterations of the training algorithm, being replaced by the standard rule in the last iterations.

Although the standard SOM is an unsupervised learning algorithm, a number of supervised variants exist. It can be argued that the calibration mechanism (Kohonen, 2001) is in fact a form of supervised learning, but this does not change the way the SOM is trained. The most common way of introducing supervised learning in the SOM training is to change the way units are updated according to the known class of the input pattern. The well known LVQ algorithm (Kohonen, 2001) is one such case, where units are attracted to the input pattern if they have the same class, or repelled if it is different. Other supervised versions of SOM have also been proposed (Buessler *et al.*, 2002).

Finally, the update rule may also be changed because of the particularities of the input space or the metric used. Such is the case when binary features are used (Gioiello *et al.*, 1992; Lobo, 2002; Tanomaru, 1995). In this case, since the only acceptable values for each attribute are 0 and 1, the smooth adaptation required by the standard rule is not possible. In Lobo, (2002) two update rules are proposed that require 'correcting' a number of bits proportional to the desired learning rate. A similar adaptation of the learning rate is required whenever categorical data is used, as shown in Lourenço *et al.* (2004).

2.4   GEO-VARIANTS – INCLUDING GEOGRAPHIC REASONING IN THE SOM

In this section we overview some of the applications of SOM variants to GIScience problems, and will then proceed to cover a selected number of SOM variants which allow for the inclusion of geography within the workings of the SOM.

In Villmann and Merényi (2001) and Villmann *et al.* (2003) SOM variants are used to analyze satellite images. Two different variants are used: the 'GrowingSOM' (Bauer and Villmann, 1997), where the SOM is allowed to grow into an *n*-dimensional hypercube topology so as to accurately match the intrinsic dimensionality of the data; and the SOM with magnification control (Bauer *et al.*, 1996), which tries to distribute the input patterns evenly amongst the map units. An interesting feature of these papers is their use of color coding to extract information from 3-D SOMs. Most other utilizations of SOMs in GISience have used the standard SOM algorithm. These have already been reviewed in Chapter 1.

Next, we discuss variants, which, although in some cases developed for other types of problems, allow for the possibility to explicitly incorporate geo-references and take into account the special features of spatial information into the SOM algorithm. One of the fundamental ideas consists of embedding the first law of geography (Tobler, 1970) into the SOM. This can be achieved through the classification of geographical neighbors in similar areas of the output space. A balance between geographical proximity and attribute proximity should be achieved. Although location can be viewed as just another attribute of a given entity, we make a clear distinction between location and other attributes due to the special role that location has in GIScience. In most scientific fields, classification is solely based on the notion of similarity of attributes, as the idea is to group similar entities. In GIScience, classification can be seen as a compromise between similarity in 'attribute space' and similarity (proximity) in geographic space. We are interested not only in the characteristics of the entities but also on how they interact in geographic space. Generally, it is of little interest to group entities that are far apart even if they have similar attributes. It can be much more informative to evaluate 'local' variation, in other words, assess the degree of similarity of neighboring entities. There are different ways to accomplish this objective, as we will show.

2.4.1   Hierarchical SOMs

Hierarchical SOMs change the normal interconnections between units. They are used in applications fields where a structured decomposition into smaller and layered problems is convenient. One or more than one SOMs are located at each layer, usually operating on different thematic variables (Figure 2.1).

Hierarchical SOMs (Ichiki *et al.*, 1991; Luttrell, 1989) were extensively used in speech recognition, where each layer deals with higher units of speech, such as phonemes, syllables, and word parts (Behme *et al.*, 1993; Jiang *et al.*, 1994; Kempke and Wichert, 1993).

Hierarchical SOMs can have several lower level partial maps that cluster the data according to different characteristics and then pass the results to an upper level SOM, or they may have a lower level global SOM, that acts as a gating mechanism to activate one of several higher level SOMs that specialize in a certain area of the input space.

In terms of GIScience one can envision the usefulness of hierarchical SOM in applications like geodemographics. Hierarchical SOMs allow the creation of purpose-specific or thematic classifications at lower layers which are then composed into a single one. This can constitute a major advantage as it has been noted that the purpose-specific geodemographic classifications constitute more powerful tools than general purpose classifications (Openshaw and Wymer, 1995). For instance, one can envision the creation of different
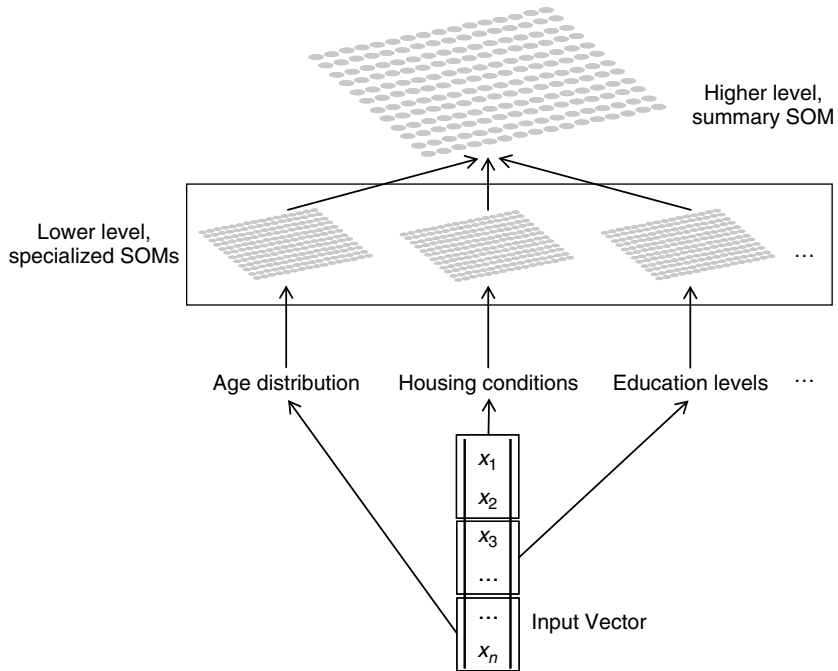
Figure 2.1   *Structure of a hierarchical SOM*

thematic classifications, based on census data, such as 'age distribution', 'employment characteristics', 'housing conditions' and 'education levels'. A higher level SOM will take inputs from each of these thematic SOMs and produce a single classification, for example grouping entities that are similar both in 'age distribution' and 'education levels', even though they might be somewhat different in housing conditions. This higher level SOM may take a number of different types of input information, from the lower level SOMs. This information may be, for example, the coordinates of the BMU or the activation functions of all the units. The exploration of different lower level SOMs can be valuable, especially if done in a computational environment where dynamic linking between SOMs can be set up. This way the interactive exploration of the different classifications can provide insights into the distribution of different geographic features in different thematic classifications. The higher level SOM allows for a general overview of the classification, acting as a summary of the lower-level classifications.

2.4.2   Geo-enforced SOM

One simple way of producing quasi-variants and testing spatial effects is through pre-processing. Instead of altering the basic SOM algorithm the idea is to include spatially relevant variables which are computed as any other socio-economic variable (Lobo *et al.*, 2004). This way there are two major operational decisions to be made. The first has to do with the choice of the spatial variables to use, which depends on the objectives pursued. The second has to do with the weighting that should be attributed to the geographic

variables. Once all the variables are normalized the user has the possibility of deciding how much weight each of the variables will have in the calculations, thus giving more or less importance to the geographic information.

One example would be to include in the data set the geographic coordinates of the centroids of the geographic features along with other attributes (Lobo *et al.*, 2004). This way any clustering solution would be 'affected' by the geographic location of the features, and geographically distant features would be less likely to be in the same cluster. Another option is to include distance measurements. For instance, we can take distance measurements of each geographic feature to important centers and include them as attributes. In Portugal there are two major centers that, due to their economic importance, influence all the regional development (Lisbon and Oporto Metropolitan Area). If one is to develop a regional classification, the distance of the different counties to these important economic centers would help to introduce accessibility information, which might be relevant. This approach can be seen as a pre-processing strategy and does not change the actual SOM training algorithm. Thus, we do not consider it a true variant.

2.4.3 Geographical Hypermap

In this approach, we change the matching and voting phase of the SOM algorithm. The Geographical Hypermap was inspired in the Hypermap architecture which was originally proposed in Kohonen (1991) for speech recognition. In this architecture the input vector is decomposed into two distinct parts, a 'context' vector and a 'pattern' vector. The basic idea is to treat both parts in different ways. The most common approach is to use the context part to select the BMU, and then adapt the weights using both parts, separately or together. However, many other variants exist (Figure 2.2). The Geographical Hypermap implies that the classification of a specific input vector is learned in the context of its geographic BMU. In other words, the Geographical Hypermap will force the classification of input patterns based solely on their geographic location.

In this approach, we change the matching and voting phase of the SOM algorithm.
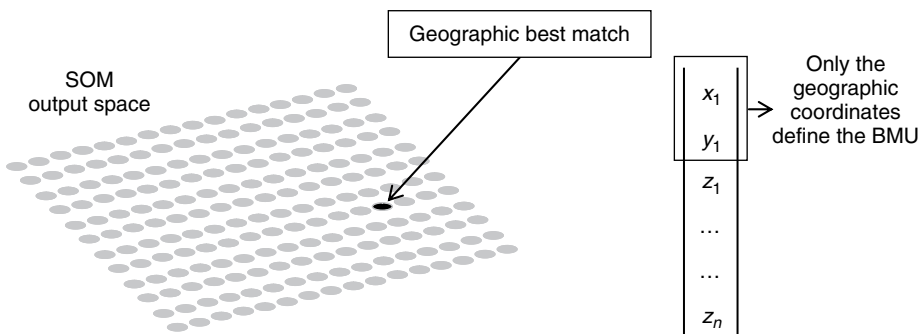
Figure 2.2 *Example of a Geographical Hypermap seen in the output space*

This way, each unit in the SOM will be an average of the non-geographic attributes in the geographic area it covers. The smoothing effect of this averaging depends on the number

of units and the density of input patterns. In Figure 2.3 an example of this averaging process is shown. As can be seen each unit (represented by the large squares) has a geographic location in the map, and the data patterns (represented by the small dots) associated with it are defined by the Thiessen polygon around it. This way, the unit's characteristics are given by the average value of the set of data patterns in those polygons. These polygons may include different numbers of data patterns so the averaging is not of a fixed number of patterns.
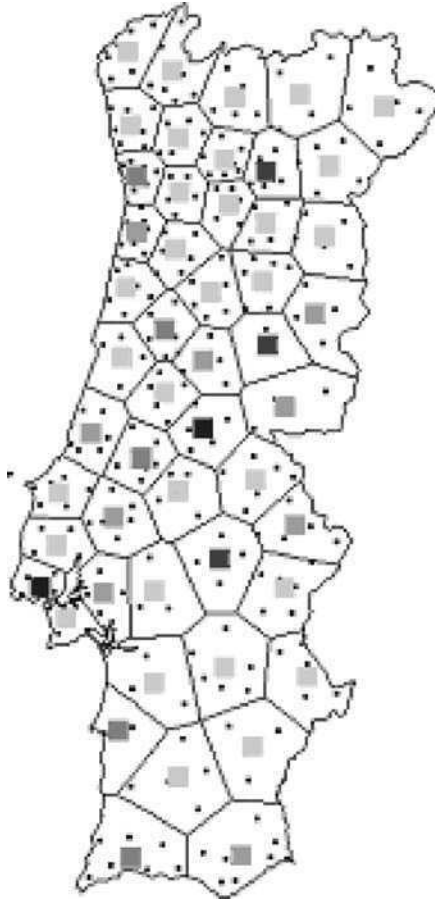
Figure 2.3   *Example of a Geographical Hypermap seen in the input (geographical) space (See Colour Plate 3)*

2.4.4   Spatial-Kangas Map

This approach is yet another method for changing the matching and voting phase of the basic SOM. The Spatial-Kangas map introduced in Lobo *et al.* (2004) is based on the temporal SOM first presented in Kangas (1992) and commonly known as the Kangas map. The Spatial-Kangas map extends the underlying principles of the Geographical Hypermap, in the sense that the BMU is required to be in the geographical neighborhood

of the input pattern. However, in this approach the requirement that the BMU be the geographically closest unit is relaxed, requiring only that it be close (within a certain radius named 'geographical tolerance'). This is done by dividing the search for the BMU into two phases: first establish a geographical neighborhood where it is admissible to search for the BMU (Figure 2.4), and then perform the final search using the non-geographical components of the input pattern. This can be seen as the separation of the vector into two different parts; one that carries the geographic context of the pattern, and the second that provides information for the definition of the BMU within the context.

Figure 2.4 *Spatial-Kangas map structure (Geo-SOM with k > 0)*

2.4.5 Geo-SOM

The Geo-SOM (Bação *et al.*, 2004b) constitutes the generalization of the two previous variants. In fact, the Geographical Hypermap and the Spatial-Kangas map can be seen as particular instances of a more general concept represented by the Geo-SOM.

The geographic neighborhood where we search for the BMU can be controlled by a parameter k, defined in the output space, and called geographical tolerance. If we choose $k = 0$, then the BMU would necessar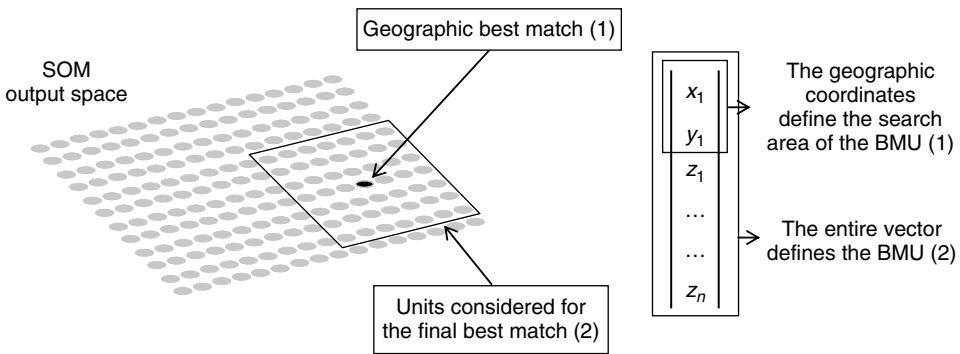ily be the unit geographically closer, which corresponds to a Geographical Hypermap. As k (the geographic tolerance) increases, we consider ever larger geographical regions when looking for BMUs. In Figure 2.5 we give an example of how we select the candidates for BMU when considering a geographic tolerance of 1. Values of k between 1 and the size of the map correspond to different Spatial-Kangas maps. If we allow k to grow to the size of the map then any unit may be selected as BMU, regardless of its geographical coordinates, and a standard SOM will be obtained. It is important to note that although k is called geographic tolerance, it is defined in the output space, and thus the actual geographic proximity, for a fixed value of k, depends on the density of units in that area. This means that in areas where there are many units (areas with a lot of data patterns and variance), a given k will correspond to small geographic distance, whereas in sparsely represented areas, the same k will correspond to a larger geographic distance.

When $k = 0$, the final locations in the input space of the units will be a quasi-proportional representation of the geographical locations of the training patterns (the proportionality is not exact due to the already discussed magnification effect), and thus
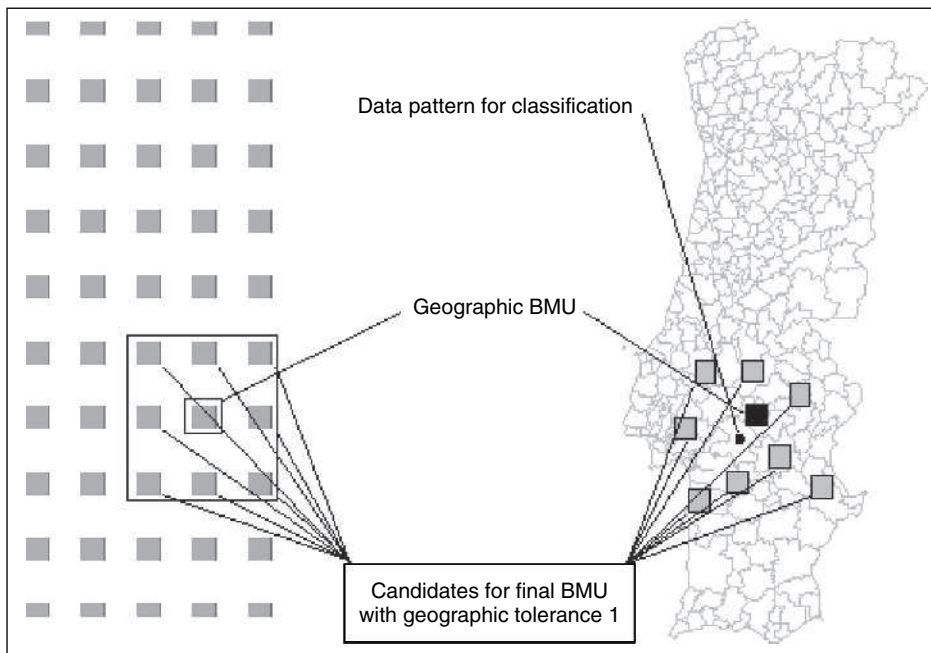
Figure 2.5 *Example of how candidates for BMU are chosen in a Geo-SOM with a geographic tolerance of 1. The actual BMU is chosen amongst these using the using the standard SOM procedure*

the units will have local averages of the training vectors. Exactly the same final result may be obtained by training a standard SOM with only the geographical locations, and then using each unit as a low pass filter of the non-geographic features. The exact transfer function (or kernel function) of these filters depends on the training parameters of the SOM, and is not relevant for this discussion.

Formally, the Geo-SOM may be described by the following algorithm (Bação *et al.*, 2004b):

```
Let
   X be the set of n training patterns x₁, x₂, . .xₙ, each
     of these having a set of components geoᵢ and another set ngfᵢ.
   W be a p × q grid of units wᵢⱼ where i and j are their
     coordinates on that grid, and each of these units having
     a set of components wgeoᵢⱼ and another set wngfᵢⱼ.
   α be the learning rate, assuming values in]0,1[, initialized
      to a given initial learning rate
   r be the radius of the neighborhood function h(wᵢⱼ, wₘₙ, r),
      initialized to a given initial radius
   k be a radius surrounding geographical BMU where the final
      BMU is to be searched
   f be a logical variable that is true if the units are forced
      to remain at fixed geographical locations.
```

```
1 Repeat
2   For m=1 to n
      For all w_ij ∈ W,
4       Calculate d_ij = ||wgeo_m − wgeo_ij||
5       Select the unit that minimizes d_ij as the geo-winner W_winnergeo
6       Select a set W_winner of w_ij such that the distance in the grid
          between W_winnergeo and w_ij is smaller or equal to k.
7       For all w_ij ∈ W_winner, calculate d_ij = ||x_m − w_ij||
9         Select the unit that minimizes d_ij as the winner W_winner
10      If f is true, then
11        Update each unit w_ij∈ W: wngf_ij = wngf_ij +
            α h(wngf_winner, wngf_ij, r)||x_m − w_ij||
12      Else
13        Update each unit w_ij∈ W: w_ij = w_ij+αh(W_winner, w_ij,r)||x_m− w_ij||
14  Decrease the value of α and r
15 Until α reaches 0
```

The Geo-SOM has the potential to organize the SOM output space according to the geographic proximities of the input patterns. This way, areas of the geographic map with similar characteristics will warrant a smaller number of units than the areas of the map where characteristics differ a lot. One of the potential applications for the Geo-SOM is to develop homogeneous zones. Homogeneous region building constitutes a data reduction task, and can be seen as the geographic counterpart of clustering. In fact, as in clustering, the idea is to reduce the number of entities, while losing the smallest amount of information, in order to improve the understanding. Detecting homogeneous regions is in itself knowledge discovery, as it allows the identification of redundancy in the sense that areas that have similar profiles can be managed as one. Contrary to most zone design algorithms (Alvanides and Openshaw, 1999; Horn, 1995; Macmillan and Pierce, 1994; Mehrotra *et al*., 1998), in which the number of zones is pre-defined, the Geo-SOM can be viewed as an exploratory technique to build zones, as will be shown in Section 2.6.

2.5   EXPERIMENTAL RESULTS WITH ARTIFICIAL DATA

In order to assist the comprehension of the major characteristics and properties of some SOM variants, we carry out a set of tests based on artificial data. The objective of using artificial data is to produce a controlled environment where certain features of the variants can easily be understood. We constructed an artificial data set with 5000 3-D points, each of which has geographical coordinates (x and y), and a third variable z that represents a nonspatial attribute. The points follow a uniform distribution in the geographical coordinates, within the rectangle limited by [(0,0), (20, 15)] (Figure 2.6). In the nonspatial dimension there are three zones of high spatial autocorrelation, where the values of z are very similar among neighboring points, with a uniform distribution in the interval [90, 91] in two zones, and in the interval [10,11] in another. There is also one area of 'negative autocorrelation', where half the data points have $z \approx 10$ and the other half have $z \approx 90$. In the rest of the input space z has a uniform distribution in [0,100].
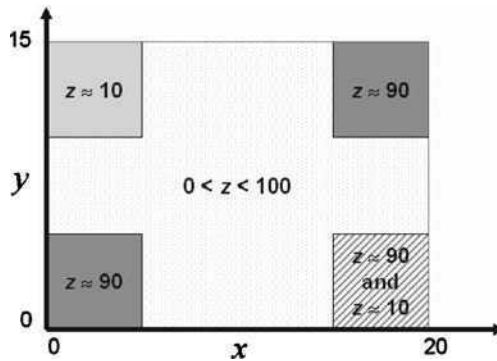
Figure 2.6   *Artificial data set*

Five different SOMs were used to process the data: a standard SOM, a Geo-SOM with $k = 0$ (which is similar to the Hypermap), a Geo-SOM with $k = 1$, $k = 2$ and $k = 4$. The k parameter, which we call geographic tolerance, refers to the size of the neighborhood amongst which the BMU will be searched. In all cases, the SOMs used had 20×15 units with 'bubble' (rectangular) neighborhood functions. In the first training phase (or unfolding phase) we used an initial radius of 15, a final radius of 0, an initial α of 0.3, and 10 epochs. In the second training phase (or fine-tuning phase), we used an initial radius of 3, a final radius of 0, an initial α of 0.1, and 20 epochs.

In order to get a clear image of the error produced by each one of the tested variants we decided to separate the error in geographic error and quantization error. The geographic error computes the average distance between each input pattern and the unit to which it was mapped. This gives a notion of the geographic displacement of the units relative to the input patterns they represent. The quantization error provides an assessment of the distances between input patterns and the unit to which they are mapped in the attribute space, in this case the z variable. The quantization error provides a measure of the quality of the representation of z (non-geographical attribute) achieved by each variant.

The results are quite informative in the sense that they allow a very clear distinction between the behaviors of the different variants. Clearly, the restrictions imposed by Geo-SOM tend to degrade the quantization error and improve the geographic error. In terms of quantization error the highest value is observed, as would be expected, in the Geo-SOM with the smallest geographic tolerance, and decays as k increases, until reaching the minimum with the standard SOM. Conversely, the geographic error decreases as k increases. The actual values are shown in Table 2.1.

Table 2.1   *Average geographical and quantization errors for the artificial data set*

Type of map *vs* Type of error	Geo-SOM $k = 0$	Geo-SOM $k = 1$	Geo-SOM $k = 2$	Geo-SOM $k = 4$	Standard SOM
Geographical error	0.4193	0.8507	1.1800	1.5055	1.6713
Quantization error	21.0690	12.5902	7.1130	2.4440	0.9030

The quantization errors shown in the table are averages for all data patterns, and the individual values vary quite a lot. A close inspection of the way this quantization varies allows us to identify different clusters, which is one of the main purposes of using these techniques. If we calculate the average quantization error of the input patterns that are mapped to each individual unit and plot these values in a color plot, we obtain the results presented in Figure 2.7. In this figure we plot the quantization error in grayscale as a function of the geographical coordinates when using Geo-SOMs with $k = 0$ and $k = 2$, and using the standard SOM.
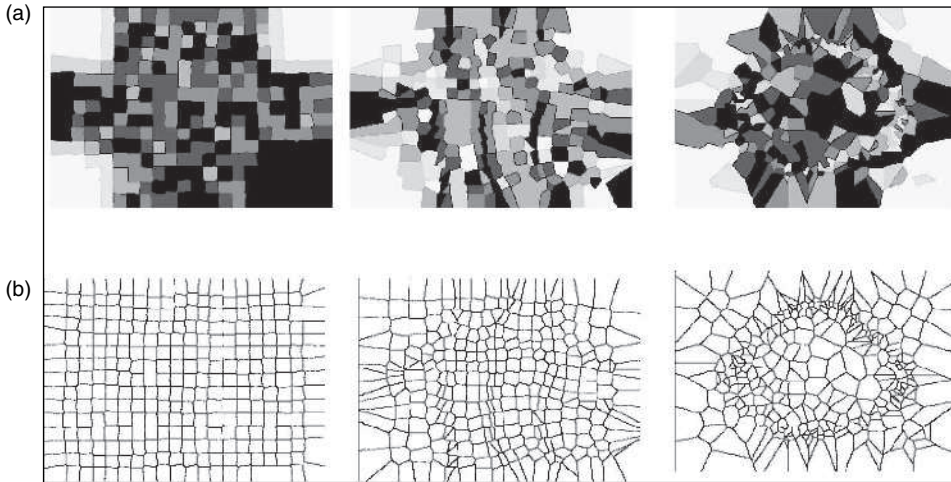
Figure 2.7 *Maps with the average quantization error per unit (a), and geographical coverage of those units (b), using the Geo-SOM with k=0 (left), k=2 (center) and the standard SOM (right)*

With $k = 0$ the Geo-SOM is basically performing local averages. The points where those averages are calculated follow the geographical distribution of the input patterns, which in this case means they are evenly distributed. Areas where 'natural' clusters exist are clearly shown by white areas, where the quantization error is low. Areas where there is less spatial autocorrelation are represented in progressively darker shades of gray, corresponding to increasing quantization errors. From this map little can be inferred about how to define regions in those areas. Thus, the choice of $k = 0$ allows us to identify only the clearly homogeneous areas.

With $k = 2$ the Geo-SOM provides interesting insights into the data. Homogenous areas are still evident, but some new areas with low quantization error appear throughout the map. The lower right corner (shown as a close up in Figure 2.8), where the data follow two distinct behaviors is divided (approximately along its diagonal) into two homogenous areas, one containing each type of data. These are separated by another area that serves as the border, where the quantization error is quite large. The information conveyed is that, in this general area, there are subsets of points which share similar values of z. The added geographic tolerance provided by $k = 2$ allows the identification of a certain

degree of homogeneity, which the Geo-SOM (0) was unable to find. As a conclusion, this map allows us to gain insight into less well structured areas of the data.
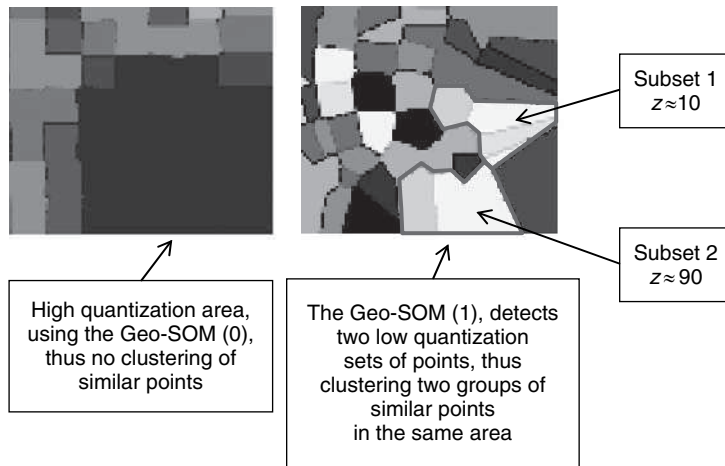
Figure 2.8   *Close up of the lower right corner of the Geo-SOM (0) and Geo-SOM (1) in Figure 2.7 (See Colour Plate 4)*

Finally, when using the standard SOM, the map has little information about the geographical organization of clusters. Since these are defined mostly by non-geographical attributes, their geographical location is basically meaningless and may lead to errors. The lower left corner of the map has basically the same configuration as the other corners even though the data in that corner are significantly different. We may thus conclude that while the standard SOM may be a good clustering tool, it naturally fails to single out the geographical information contained in it.

2.6   EXAMPLE OF THE USE OF THE GEO-SOM

In this section the objective is to show examples of how the Geo-SOM can be used to explore geographic data. Our main concern is not to solve a particular problem but rather to emphasize the potential of the Geo-SOM as an exploratory tool. Nevertheless, data are needed so we chose to analyze census data concerning the Portuguese counties in 2001. These data are publicly available and consist of 70 attribute variables characterizing each one of the 250 counties of mainland Portugal. The variables include a number of socio-demographic indicators such as per capita GDP, purchasing power, age distribution and education levels among others. For the purpose of this analysis all the classification variables were normalized to a distribution of mean 0 and standard deviation 1. The normalization is invariant to size as all the variables are measured as ratios of the population value.

We will now analyze this dataset using the Geo-SOM. To aid the analysis, we present a number of visualization instruments, which we developed based on the Geo-SOM, so as to assist the user in exploring and discovering new patterns in data. The exploration environment was developed based on ArcView®. ArcView was a valuable tool as it allowed the swift development of an exploration environment prototype. Relevant in the development of the prototype was the possibility of linking different files as well as the opportunity of using multiple dynamically linked windows. The result is an environment where the user can shift through different windows probing the available information and building 'what if' scenarios.

We will present the results of two Geo-SOMs using different geographic tolerance parameters ($k = 0$ and $k = 1$) with 50 units each.

The process to set-up the visualization environment requires the representation of the output space (U-matrix and components planes) of the Geo-SOM in ArcView. One of the advantages of the Geo-SOM over the use of geo-enforced SOMs is the fact that the resulting U-matrix maintains a structure which is relatively similar to the geographic map. This way the geographic features from the southern part of the map will be represented in the bottom of the U-matrix and vice-versa. The representation of the U-matrix in ArcView was done through the ascription of fictitious coordinates to the Geo-SOM units, enabling 'mapping' of the output space. Once the output space was available in ArcView we linked the file containing information about the counties with the file of the output space. This was possible due to the fact that the file of the counties had a field corresponding to the unit in which each county was classified.

The typical exploration setting includes a window with a components plane super-imposed on a U-matrix and dynamically linked with a second window which displays the map of Portugal and a database window where the selected elements are displayed (Figure 2.9). In Figure 2.9(a) the matrix represents the U-matrix of Geo-SOM (0).
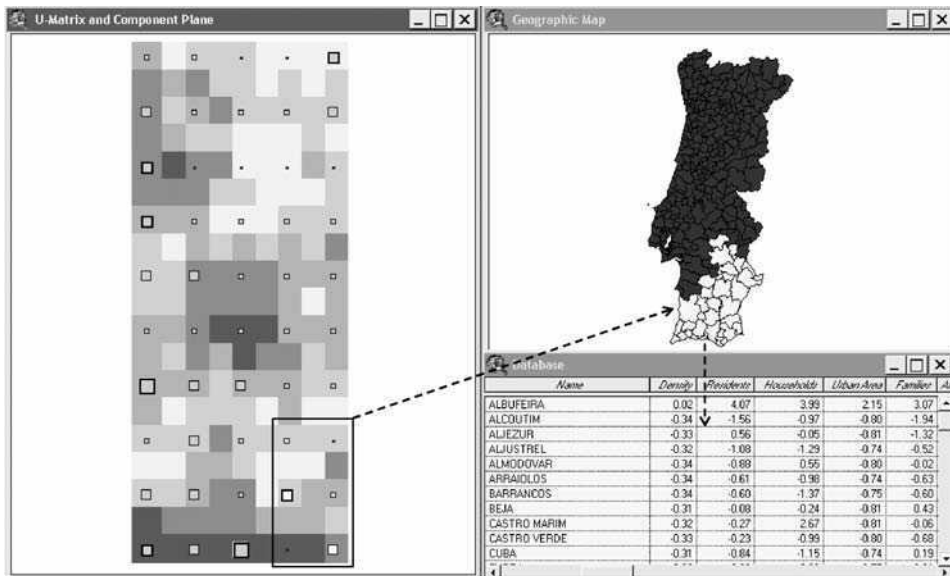

Figure 2.9 *Exploration environment developed to support the Geo-SOM*

The squares superimposed on the U-matrix represent a component plane, in this case the dimension of the square represents the quantization error of that particular unit. In Figure 2.9(b) the geographical map of Portugal's mainland is presented together with a database window where the elements selected in the U-matrix are highlighted.

Using different component planes the user can scan the data looking for unusual patterns which may help understand the data and their geographic distribution. In Figure 2.10 an example of the exploring capabilities offered by the Geo-SOM is presented. In this case, the Geo-SOM (0) was used, and the lower right area of the U-matrix was selected. The selection made in the U-matrix corresponds to the counties highlighted in the geographic map, corresponding to three major cities. Additionally, an *x,y* scatter chart shows the relation between the proportion of college education individuals and purchasing power. In this example we show how simple it is to select different units in the U-matrix and perform additional analysis on the selected individuals.
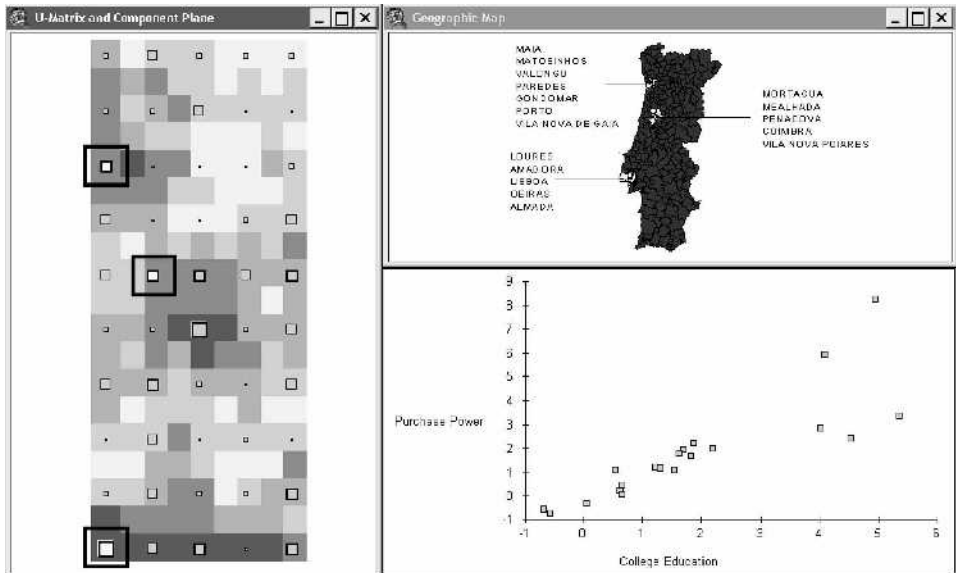
Figure 2.10 *Example of three clusters identified by the Geo-SOM (0) and the possible interaction between the U-matrix and components plane (a), the geographic map (b) and a graph showing the distribution of two particular variables of the selected counties (c)*

The use of the Geo-SOM (1) is more complex as the values of the different units do not involve calculations solely based on the geographically closest neighbors. The workings of the Geo-SOM (1) (and with higher geographic tolerances) can be described as 'averages of similar counties' in the sense that within a geographic tolerance the Geo-SOM will try to group similar counties. This can be viewed as the possibility of lessening the geographic constraint providing the Geo-SOM the possibility of clustering counties with similar profiles and which are located in the same general area. In this case the results are not contiguous regions but sets of areas with similar characteristics that are relatively close in geographic terms.

The following figures present some examples of the type of analysis produced by the Geo-SOM (1). In Figure 2.11 we focus on the cluster highlighted (and identified by a large black square), which groups together what can be considered the three most important industrial areas in the south of the country, especially related to the auto industry. The other three clusters that are represented in this corner of the U-matrix include Lisbon and most of the Lisbon Metropolitan Area counties. Also shown are the values of exports and imports of these three counties. Bearing in mind that the data were standardized to a mean of 0 and a standard deviation of 0, the values attained by these counties in these specific variables are quite impressive.
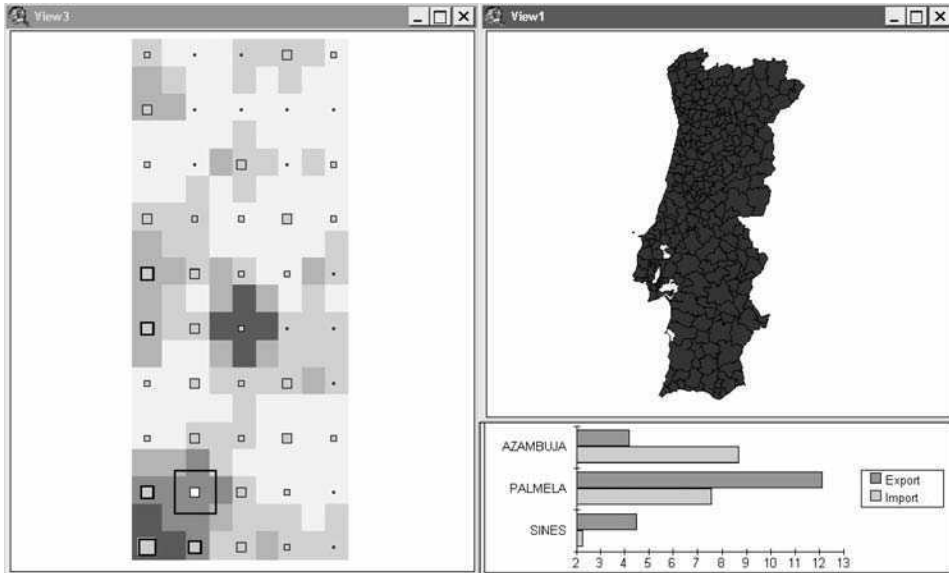
Figure 2.11   *Using the Geo-SOM (1) to detect clusters of specific characteristics*

In Figure 2.12 we compare the membership of a specific county (Braga) in three different SOMs: a mildly Geo-enforced SOM (the *x, y* coordinates of the county's centroids were added to the 70 attribute variables), a Geo-SOM (0) and a Geo-SOM (1). In all three classifications Braga is grouped with different counties. In the Geo-enforced SOM Braga, which is a district capital, is grouped in a cluster which contains most of the Oporto Metropolitan Area, as well as two other district capitals (Viseu and Leiria). Both Viseu and Leiria are located far away from Braga. In the Geo-SOM (0), Braga is grouped in a geographically contiguous set which includes coastal counties north of Oporto Metropolitan Area. Finally, in Geo-SOM (1) only two other counties are grouped with Braga. The contiguous county, Guimarães, can be seen as a twin city as they share a number of administrative services and a university campus. Viana do Castelo, like Braga, is also a district capital. This example shows some fundamental differences between the workings of the different SOM variants. The Geo-enforced SOM clusters with a strong influence of the attribute variables. In the Geo-SOM (0), however, attribute variables are less relevant and geographic location becomes central. Finally, in Geo-SOM (1)

a compromise between attributes and geographic location is achieved. It is probably pointless to argue about the superiority of any of these variants, as the combination of the three analyses produces an improved understanding of the problem. Nevertheless, we argue that from a GIScience perspective it is sensible to use space as a determining factor in the outcome of clustering.

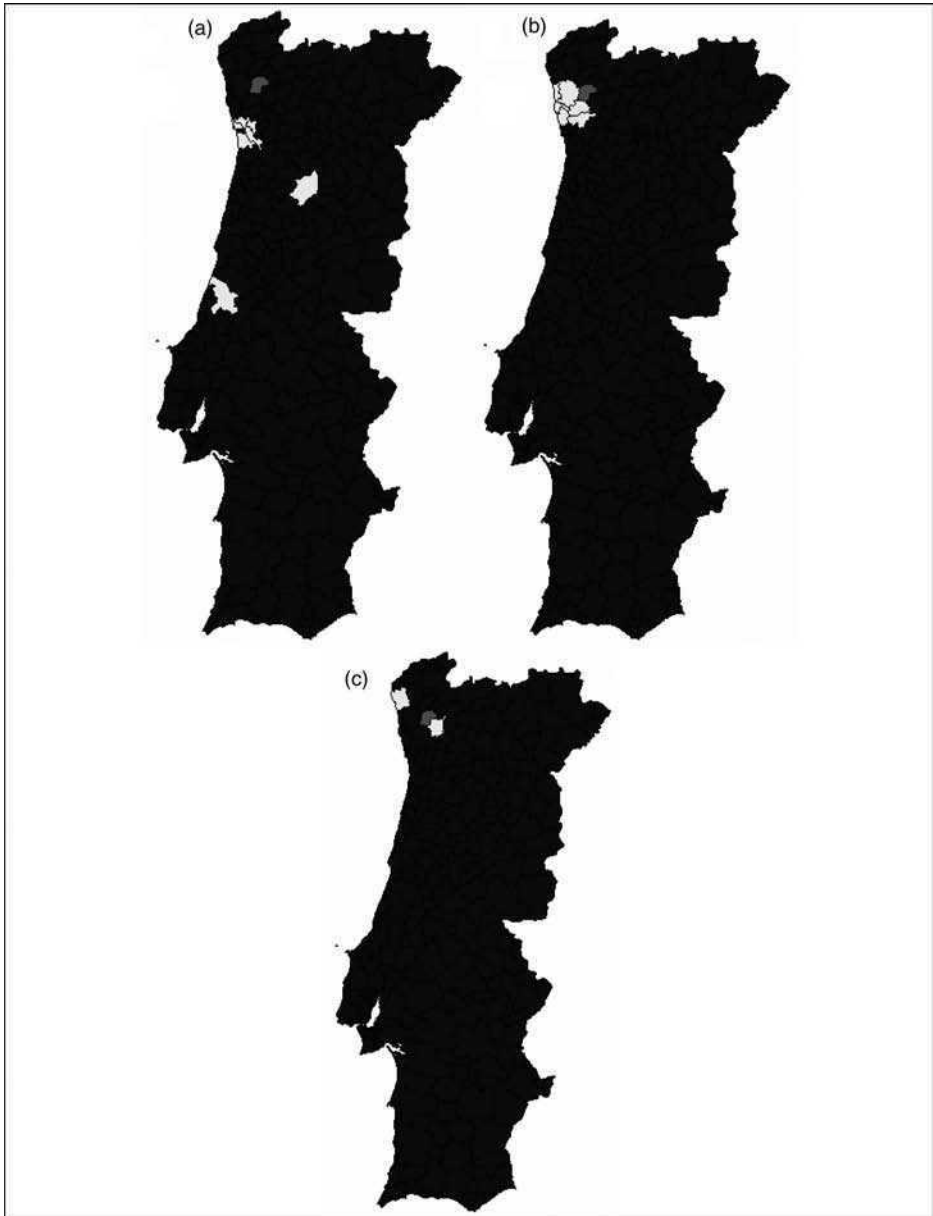**Figure 2.12**   *Comparison of the areas clustered together with Braga using three different SOM variants: (a) a Geo-SOM (1); (b) a Geo-enforced SOM; (c) a Geo-SOM (0)*

2.7 CONCLUSIONS

There are many ways in which the standard SOM can be used in GIScience problems. In this chapter we take a different path, by considering variants to the original algorithm. We propose a number of variants that explicitly take into account location and space in the SOM algorithm. A brief explanation of these different SOM variants was presented. A more detailed explanation of one of the geographically oriented variants, the Geo-SOM, was given. An example of its application to an artificial data set and an actual geographically referenced data set was presented. It was shown that in the latter case, this approach can provide a meaningful insight to the spatial data structure. The Geo-SOM can be thought of as a method which projects multidimensional data into geographic space.

The Geo-SOM should be seen as an effort to adapt a tool, developed in a different scientific context, to the specific needs and reasoning of GIScience. Usually, tools are imported from other areas of knowledge or developed based on specific practical needs posed by specific problems. The Geo-SOM has a different motivation. The Geo-SOM constitutes a theoretical effort in the sense that it is the result of the interaction between a valuable analysis tool (the SOM) and the GIScience perspective of the world. The fundamental assumption of the Geo-SOM is that in spatial analysis, space should take the center stage and attribute variables should be analyzed within their spatial context.

There are a number of issues that remain to be explored in the Geo-SOM. The effect that the relationship between density of input patterns (in the geographic space) and the distance between them (in the variable space) has on the distribution of the units is still an open problem. Another interesting issue to address in future developments is the possibility of using dynamic k values. The idea is to adequate the k parameter according to the specific spatial autocorrelation index of the area of the input pattern. Operationally, the Geo-SOM can benefit a lot if a specific visualization and interactive exploration tool is developed. ArcView served the purposes of prototyping well but there are limitations which hamper the usefulness of the analysis. Such a tool is currently being developed.

Besides Geo-SOM other variants might be of interest in GIScience, such as the hierarchical SOM. The flexibility of the original SOM algorithm and its wide range of application provide the opportunity to adapt the SOM to specific paradigms and problems of GIScience.

All the programming routines for the architectures presented here are available (www.isegi.unl.pt/docentes/vlobo/projectos/programas/programas.html).

REFERENCES

Agarwal, P. (2004). Contested nature of 'place': knowledge mapping for resolving ontological distinctions between geographical concepts. *Lecture Notes in Computer Science*. M. Egenhofer, C. Freksa and H. Miller. Springer-Verlag, Berlin, **3234**: 1–21.

Almeida, L. B. and J. S. Rodrigues (1991). Improving the learning speed in topological maps of patterns. *Neural Networks*: 63–78.

Alvanides, S. and S. Openshaw (1999). Zone design for planning and policy analysis. *Geographical Information and Planning*. J. C. H. Stillwell, S. Geertman and S. Openshaw. Springer-Verlag, Heidelberg: 299–315.

Bação, F., *et al*. (2004a). Clustering census data: comparing the performance of self-organising maps and *k*-means algorithms. KD-Net Symposium 2004, Knowledge-based services for the public sector, Bonn.

Bação, F., *et al*. (2004b). Geo-Self-Organizing Map (Geo-SOM) for building and exploring homogeneous regions. *Lecture Notes in Computer Science*. M. Egenhofer, C. Freksa and H. Miller. Springer-Verlag, Berlin, **3234**: 22–37.

Bauer, H.-U. and T. Villmann (1997). Growing a hypercubical output space in a self-organizing feature map. *IEEE Transactions on Neural Networks* **8**(2): 218–226.

Bauer, H.-U., *et al*. (1996). Controlling the magnification factor of self-organizing feature maps. *Neural Computation* **8**: 757–771.

Behme, H., *et al*. (1993). *Speech Recognition by Hierarchical Segment Classification*. ICANN 93, Springer, Amsterdam.

Buessler, J.-L., *et al*. (2002). Additive composition of supervised self-organizing maps. *Neural Processing Letters* **15**(1): 9–20.

Camastra, F. (2001). Intrinsic dimension estimation of data: an approach based on Grassberger–Procaccia's algorithm. *Neural Processing Letters* **14**(1): 27–34.

Carpenter, G. A. and S. Grossberg (1988). The art of adaptive pattern recognition by a self-organizing neural network. *IEEE Computer Magazine* March: 77–87.

Chandrasekaran, V. and Z.-Q. Liu (1998). Topology constraint free fuzzy gated neural networks for pattern recognition. *IEEE Transactions on Neural Networks* **9**(3): 483–502.

Claussen, J. C. (2003). Winner-relaxing and winner-enhancing Kohonen maps: Maximal mutual information from enhancing the winner. *Complexity* **8**(4): 15–22.

Cottrell, M., *et al*. (1998). Theoretrical aspects of the SOM algorithm. *Neurocomputing* **21**: 119–138.

Fritzke, B. (1991). *Let it Grow – Self-organizing Feature Maps With Problem Dependent Cell Structure*. ICANN 91, Helsinki, Elsevier.

Fritzke, B. (1994). *A Growing Neural Gas Network Learns Topologies*. NIPS, Denver, CO.

Fritzke, B. (1995). Growing grid – a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters* **2**(5): 9–13.

Fukunaga, K. and D. R. Olsen (1971). An algorithm for finding intrinsic dimensional of data. *IEEE Transactions on Computers* **c-20**(2): 176–183.

George, S. (2000). Spatio-temporal analysis with the self-organizing feature map. *Knowledge and Information Systems* **2**: 359–372.

Gioiello, M., *et al*. (1992). *A New Fully Digital Neural Network Hardware Architecture for Binary Valued Pattern Recognition*. International Conference on Signal Processing Applications and Technology, Boston.

Guimarães, G., *et al*. (2002). A taxonomy of self-organizing maps for temporal sequence processing. *Intelligent Data Analysis* **7**(4).

Horn, M. E. T. (1995). Solution techniques for large regional partitioning problems. *Geographical Analysis* **27**(3): 230–248.

Ichiki, H., *et al*. (1991). *Self-Organizing Multi-Layer Semantic Maps*. IJCNN'91, International Joint Conference on Neural Networks, Elsevier, Amsterdam.

Jiang, B. and L. Harrie (2004). Selection of streets from a network using self-organizing maps. *Transactions in GIS* **8**(3): 335–350.

Jiang, X., *et al*. (1994). *A Speaker Recognition System Based on Auditory Model*. WCNN'93, World Conference on Neural Networks, Lawrence Erlbaum, Hillsdale.

Kangas, J. (1990). *Time-Delayed Self-Organizing Maps*. IJCNN'90, International Joint Conference on Neural Networks, San Diego. IEEE Computer Society Press, Los Alamitos, CA, **II**: 331–336.

Kangas, J. (1992). Temporal knowledge in locations of activations in a self-organizing map. *Artificial Neural Networks*. J. T. I. Aleksander. Elsevier, Amsterdam, **2**: 117–120.

Kangas, J. A., *et al.* (1990). Variants of self-organizing maps. *IEEE Transactions on Neural Networks* **1**(1): 93–99.

Kempke, C. and A. Wichert (1993). *Hierarchical Self-Organizing Feature Maps for Speech Recognition*. WCNN'93, World Conference on Neural Networks, Lawrence Erlbaum, Hillsdale.

Kohonen, T. (1991). The hypermap architecture. *Artificial Neural Networks*. T. Kohonen, K. Mäkisara, O. Simula and J. Kangas. Elsevier, Amsterdam, **1**: 1357–1360.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer, Berlin.

Kohonen, T., *et al.* (1995). The Self-Organizing Map Program Package. Laboratory of Computer and Information Science, Helsinki University of Technology, Helsinki, 27.

Lee, J. A., *et al.* (2001). Recursive learning rules for SOMs. *Advances in Self-Organizing Maps*. N. Allinson, H. Yin, L. Allinson and J. Slack. Springer, London: 67–72.

Lobo, V. (2002). Ship noise classification: a contribution to prototype based classifier design. Departamento de Informatica, Universidade Nova de Lisboa, Lisbon.

Lobo, V., *et al.* (2004). *Regionalization and Homogeneous Region Building Using the Spatial Kangas Map*. AGILE 2004, Crete University Press, Heraklion, Greece.

Lourenço, F., *et al.* (2004). Binary-Based Similarity Measures for Categorical Data and Their Application in Self-Organizing Maps. JOCLAD 2004 – XI, Jornadas de Classificação e Análise de Dados, Lisbon.

Luttrell, S. P. (1989). *Hierarchical Self-Organizing Networks*. ICANN 89, London: 2–6.

Macmillan, W. D. and T. Pierce (1994). Optimisation modelling in a GIS framework: the problem of political redistricting. *Spatial Analysis and GIS*. S. Fotheringham and P. Rogerson. Taylor & Francis, Bristol: 221–246.

Maenou, T., *et al.* (1997). Optimizations of TSP by SOM method. *Progress in Connectionsist-Based Information Systems. Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems*. N. Kasabov *et al.* Springer, Singapore, **2**: 1013–1016.

Mark, D. M., *et al.* (2001). Features, objects, and other things: ontological distinctions in the geographic domain. *Lecture Notes in Computer Science*. Springer-Verlag, Heidelberg **2205**.

Martinetz, T. M., *et al.* (1993). Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks* **4**(4): 558–569.

Mehrotra, A., *et al.* (1998). An optimization based heuristic for political districting. *Management Science* **44**(8): 1100.

Openshaw, S. (1999). *Geographical Data Mining: Key Design Issues*. GeoComputation '99.

Openshaw, S. and C. Wymer (1995). Classifying and regionalizing census data. *Census Users Handbook*. S. Openshaw. GeoInformation International, Cambridge, 239–268.

Openshaw, S., *et al.* (1995). Using neurocomputing methods to classify Britain's residential areas. *Innovations in GIS*. P. Fisher. Taylor and Francis, **2**: 97–111.

Openshaw, S. and I. Turton (1996). A parallel Kohonen algorithm for the classification of large spatial datasets. *Computers & Geosciences* **22**(9): 1019–1026.

Openshaw, S. and C. Openshaw (1997). *Artificial Intelligence in Geography*. John Wiley & Sons, Ltd, Chichester.

Ritter, H., *et al.* (1992). *Neural Computation and Self-Organizing Maps: an Introduction*. Addison-Wesley.

Sester, M. and C. Brenner (2000). *Typification Based on Kohonen Feature Nets*. GiScience 2000.

Skupin, A. (2001). *Cartographic Considerations for Map-Like Interfaces to Digital Libraries*. Workshop on Visual Interfaces to Digital Libraries, Roanoke, Va.

Skupin, A. and S. I. Fabrikant (2003). Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science* **30**(2): 95–115.

Skupin, A. and R. Hagelman (2003). *Attribute space visualization of demographic change*. 11th ACM International Symposium on Advances in Geographic Information Systems, New Orleans, LA.

Takatsuka, M. (2001). *An Application of the Self-Organizing Map and Interactive 3-D Visualization to Geospacial Data*. GeoComputation'01, 6th International Conference on GeoComputation, Brisbane.

Tanomaru, J. I. (1995). *A Simple Coding Scheme for Neural Recognition of Binary Visual Patterns*. IEEE International Conference on Neural Networks, Perth.

Tobler, W. (1970). A computer model simulating urban growth in the Detroit region. *Economic Geography* **46**: 234–240.

Ultsch, A., *et al.* (1993). *Knowledge Extraction from Artificial Neural Networks and Applications*. Springer Verlag, Aachen.

Ultsch, A. and H. Li (1993). *Automatic Acquisition of Symbolic Knowledge from Subsymbolic Neural Networks*. International Conference on Signal Processing, Peking.

Ultsch, A. and H. P. Siemon (1990). *Kohonen's Self-Organizing Neural Networks for Exploratory Data Analysis*. INNC90, Paris.

Vesanto, J. (1999). SOM-based data visualization. *Intelligent Data Analysis* **3**: 111–126.

Vesanto, J. (2000). *Using SOM in Data Mining*. HUT, Finland.

Villmann, T. and E. Merényi (2001). Extensions and modifications of the Kohonen-SOM and applications in remote sensing image analysis. *Self-Organizing Maps: Recent Advances and Applications*. U.Seiffert and L. C. Jain. Springer-Verlag, Heidelberg: 121–145.

Villmann, T., *et al.* (2003). Neural maps in remote sensing image analysis. *Neural Networks* **16**(3–4): 389–403.

Yin, H. (2001). Visualization Induced SOM (ViSOM). *Advances in Self-Organizing Maps*. N. Allinson, H. Yin, L. Allinson and J. Slack. Springer, London: 81–88.

3

An Integrated Exploratory Geovisualization Environment Based on Self-Organizing Map

Etien L. Koua and Menno-Jan Kraak
*International Institute for Geoinformation Science and Earth Observation (ITC),
Department of Geo-Information Processing, PO Box 6,
7500 AA Enschede, Hengelosestraatgg, The Netherlands*

3.1   INTRODUCTION

The exploration of patterns and relationships in large and complex geospatial data is a major research area in geovisualization (MacEachren and Kraak, 2001). In such large data sets, the extraction of patterns and the discovery of new knowledge may be difficult as patterns may remain hidden. New approaches in spatial analysis and visualization are needed, in order to represent the data in a visual form that can better stimulate pattern recognition and hypothesis generation, and to allow for better understanding of the geographical processes and support knowledge construction.

More integrated visualization tools are needed for the extraction of patterns and relationships in data. The integration of feature extraction tools with appropriate user interfaces is important to support the user's understanding of underlying structures and processes in geodata.

Information visualization techniques including multidimensional visualization techniques from scientific visualization (Nielson *et al.*, 1997), such as graph visualization, scatterplots, parallel coordinate plots, iconographic displays, dimensional stacking, multi-dimensional scaling techniques and pixel techniques are increasingly used in combination with other exploratory data analysis techniques to explore the structure of large geospatial

data sets. An interesting development in the design of geovisualization environments is the integration of information visualization and cartographic methods for the exploration of geospatial data.

This integration of cartographic methods with information visualization techniques can help provide ways of exploring large geospatial data, and support knowledge construction by offering interactive visual geospatial displays to explore data, generate hypotheses, develop problem solutions and construct knowledge (MacEachren, 1994).

The computational analysis offered by advanced algorithms can be combined with visual analysis methods in a process that can support exploratory tasks. This chapter explores the self-organizing map (SOM) algorithm for such an integration as a means of contributing to the analysis of complex geospatial data. Here the algorithm is used for data mining. Graphical representations such as unified distance matrices, and component planes display are then used to portray extracted information in a visual form that can allow better understanding of the structures and the geographic processes. The design of this visual-computational environment integrates non-geographic information spaces with maps and other graphics that allow patterns and attribute relationships to be explored, in order to facilitate knowledge construction. These graphical representations (information spaces) combine information visualization techniques and cartographic methods to improve the interaction and exploration of extracted patterns by offering visualizations of the structure of the data set (clustering), as well as the exploration of relationships among attributes.

The proposed framework provides a number of steps that underline data mining and knowledge discovery methodology, and an understanding of exploratory tasks and visualization operations are used to guide the user in his hypothesis testing, evaluation and interpretation of patterns from general patterns extracted to specific explorations of selection attributes and spatial locations.

An application of the method is explored using a socio-demographic data set containing relationships between geography and economy, in order to provide some understanding of the complex relationships between socio-economic indicators, locations and, for example, the burden of diseases. This example of exploration of a data set is used to demonstrate the integration of the different graphical representations for the exploration of patterns.

3.2   A FRAMEWORK TO SUPPORT EXPLORATORY VISUALIZATION AND KNOWLEDGE DISCOVERY

3.2.1   Data Mining and Knowledge Discovery

One approach to analyzing large amounts of data is to use data mining and knowledge discovery methods. In geospatial analysis, data mining tools are applied to extract patterns from large data sets and help uncover structures in complex data (Openshaw *et al*., 1990). The main goal of data mining is to identify valid, novel, potentially useful patterns in data, and ultimately to understand them (Fayyad *et al*., 1996). Generally, three general categories of data mining goals can be identified (Weldon, 1996): explanatory (to explain some observed events), confirmatory (to confirm a hypothesis), and exploratory (to analyze data for new or unexpected relationships). Typical tasks for which data

mining techniques are often used include clustering, classification, generalization and prediction. These techniques vary from traditional statistics to artificial intelligence and machine learning. The most popular methods include decision trees (tree induction), value prediction, and association rules often used for classification (Miller and Han, 2001). Artificial neural networks are used particularly for exploratory analysis as nonlinear clustering and classification techniques. For example, unsupervised neural networks such as the SOM are a type of neural clustering technique, and neural architectures using backpropagation and feedforward are neural induction methods used for classification (supervised learning). The algorithms used in data mining are often integrated into KDD (Knowledge Discovery in Databases), a larger framework that aims at finding new knowledge from large databases. While data mining deals with transforming data into information or facts, KDD is a higher-level process using information derived from the data mining process to turn it into knowledge or integrate it into prior knowledge. This process is illustrated in Figure 3.1.
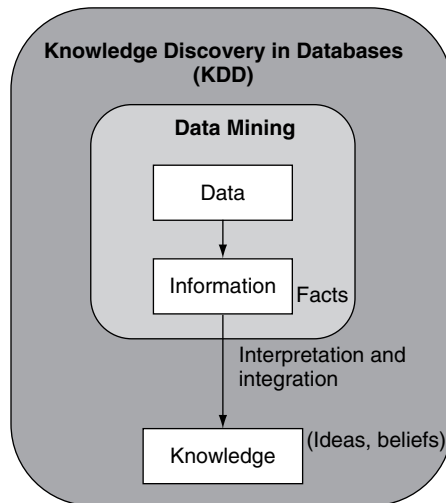

Figure 3.1   *Data mining and knowledge discovery frameworks*

In general, KDD stands for discovering and visualizing the regularities, structures and rules from data (Miller and Han, 2001), discovering useful knowledge from data (Fayyad *et al*., 1996), and finding new knowledge. It consists of several generic steps, namely data pre-processing, transformation (dimension reduction, projection), data mining (structure mining) and interpretation/evaluation.

3.2.2   Data Mining and Geospatial Data Analysis

Recent efforts in data mining and KDD have provided a window for the application of geospatial data mining and knowledge discovery in geovisualization (Gahegan *et al*., 2001; Liu *et al*., 2001; MacEachren *et al*., 1999; Miller and Han, 2001; Roddick and Lees, 2001; Sibley, 1988; Weijan and Fraser, 1996). Geographic data mining and knowledge

discovery methods haves been used in geospatial data exploration (Gahegan *et al.*, 2001; MacEachren *et al.*, 1999; Miller and Han, 2001; Openshaw *et al.*, 1990; Wachowicz, 2000) to discover unexpected correlation and causal relationships, and understand structures and patterns in complex geographic data. The promises inherent in the development of data mining and knowledge discovery processes for geospatial analysis include the ability to yield unexpected correlation and causal relationships. A large proportion of these applications are directed towards spatio-temporal data mining (Roddick and Lees, 2001).

3.2.3   Combining Computational Analysis and Visualization for the Exploration of Geospatial data

The proposed framework explores ways to combine the computational processes provided by data mining and KDD techniques as described above, with appropriate visualization techniques to support the exploration of large geospatial data. In this framework, the first level of the computation provides a mechanism for extracting patterns from the data. The output of this computational process is depicted using graphical representations. Users can perform a number of exploratory tasks not only to understand the structure of the data set as a whole, but also to explore detailed information on individual or selected attributes of the dataset. Figure 3.2 describes the proposed framework.

We propose two levels of exploratory visualization processes closely related to the concept of abduction (Gahegan and Brodaric, 2002). These processes are supported by a number of activities, including selection, analysis, comparison, and the relation of spatial locations or attributes, starting from the general patterns extracted and moving on to more user selection and refinement, which allow the exploration of relationships and the structure of a particular area of interest.

The first level of this framework consists of the visualization of the general structure of the data set (clustering); the second level focuses on the exploration for knowledge discovery and hypothesis generation. These two levels of the visualization process are provided with different representations that can be enhanced using visualization techniques. The fundamental idea of the integration of different visualization techniques is centred around four basic visualization goals, the basis for the exploratory visualization and knowledge discovery process (Weldon, 1996):

- discovering patterns (through similarity representations);
- exploring correlations and relationships for hypothesis generation;
- exploring the distribution of the data set on the map;
- detecting irregularities in the data.

To enhance the exploration of the graphical representations, visualization and interaction techniques, such as brushing, focusing, filtering, browsing, querying, selecting and linking, are used. Projection techniques such as Sammon's mapping and principal components analysis (PCA) are also used to support the different representations. As with maps, these representations use visual variables in addition to the position property of the map elements. Multiple views are used to offer alternative and different views of the data in order to stimulate the visual thinking process that is characteristic of visual exploration.
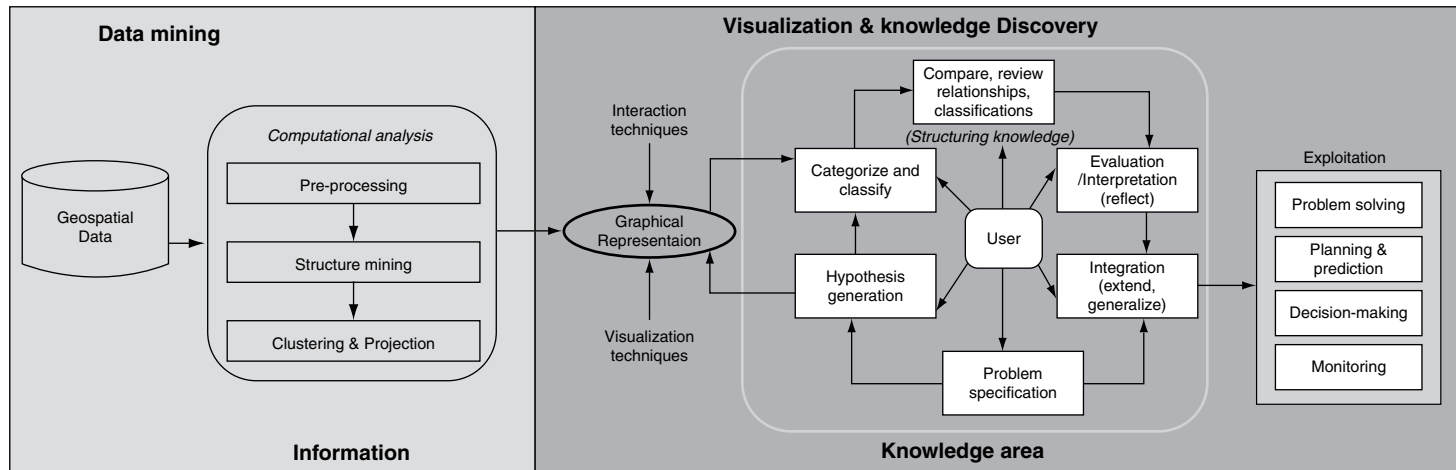
Figure 3.2   *Computation analysis and exploratory visualization framework*

3.2.4   SOM in the Computational Analysis and Visualization Framework

The SOM (Kohonen, 1989) has gained a lot of attention over recent years in geospatial data exploration and visualization. A wide range of SOM applications in geospatial analysis have been explored, including geospatial data mining and knowledge discovery (Gahegan and Brodaric, 2002; Koua, 2002), map projection (Skupin, 2003), and
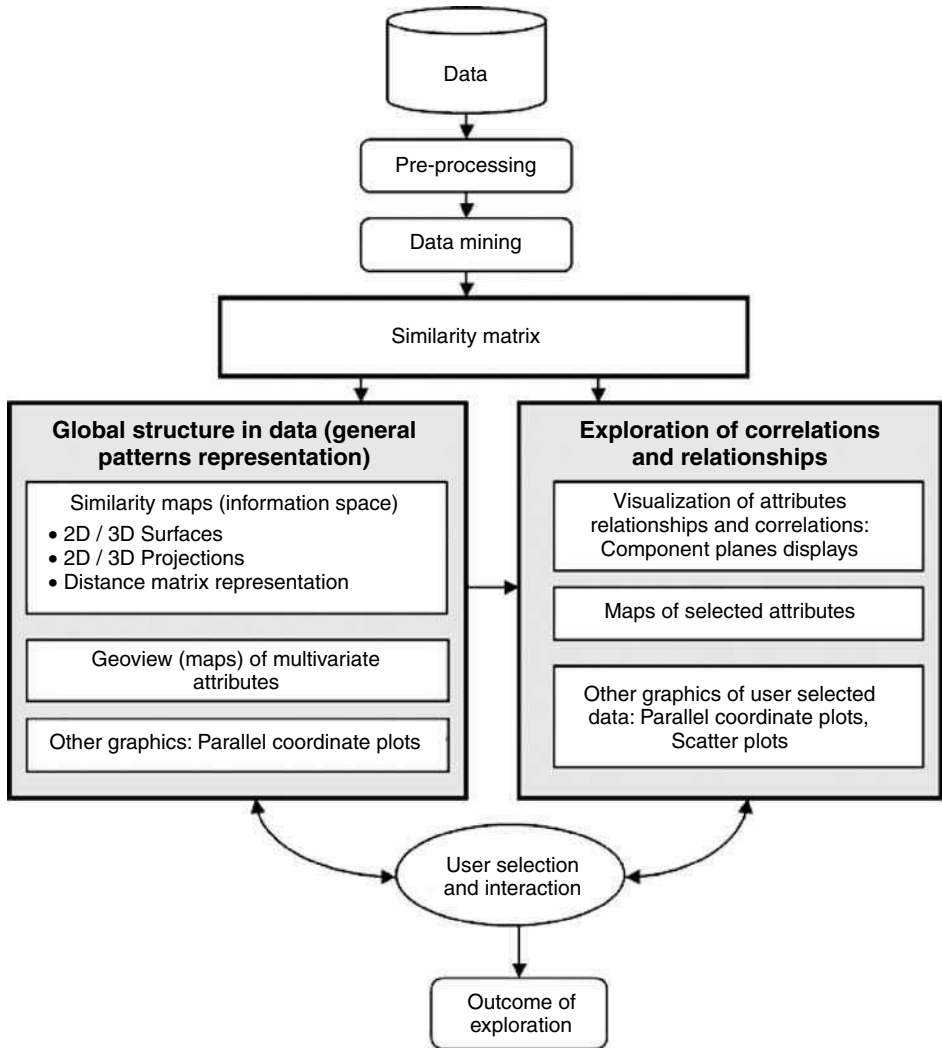

Figure 3.3   *Data exploration framework: from the computational process, global structure and patterns can be visualized with graphical representations and maps of similarity results. Relationships and correlations among the attributes are presented with interactive graphical representations, maps, and other graphics such as parallel coordinate plots*

classification (Gahegan and Takatsuka, 1999; Gahegan, 2000). The growing interest in the SOM for data analysis is due partly to its multidimensional data reduction and topological mapping capabilities. More on this has been discussed in Chapter 1.

The SOM can be a useful KDD method as it follows the probability density function of underlying data. We use the SOM algorithm as a data mining tool to project input data into an alternative measurement space, based on similarities and relationships in the input data, which can aid the search for patterns. As described in Chapter 1, the SOM adapts its internal structures to the structural properties of the multidimensional input, such as regularities, similarities and frequencies. These SOM properties can be used to search for structures in the multidimensional input. Graphical representations are then used to enable visual data exploration, allowing the user to gain insight into the data, evaluate, filter, and map outputs.

The proposed framework explores ways of effectively extracting patterns, using data mining based on the SOM, and of representing the results, using graphical representations for visual exploration. As presented in Figure 3.3, the data mining stage allows a clustering (similarity matrix) of the multidimensional input space to be constructed. From this computational process, the global structure and patterns can be represented with graphical representations and maps (geographical view) of similarity results. Further exploration can be carried out on the relationships and correlations among the attributes. The framework combines spatial analysis, data mining and knowledge discovery methods, supported by interactive tools that allow users to perform a number of exploratory tasks in order to understand the structure of the data set as a whole, as well as to explore detailed information on individual or selected attributes of the data set. Different representation forms are integrated and support user interaction for exploratory tasks to facilitate the knowledge discovery process. They include some graphical representations based on the SOM, maps, and other graphics such as parallel coordinate plots.

Cartographic methods support this design for the effective use of visual variables with which the visualizations are depicted. The graphical representations can be interactively manipulated using rotation, zooming, panning, and brushing.

3.3   A PROTOTYPICALLY EXPLORATORY GEOVISUALIZATION ENVIRONMENT

Based on the conceptual framework described above, we have implemented a prototype geovisualization environment. The visualization environment is intended to contribute to the analysis and visualization of large amounts of data, as an extension of the many geospatial analysis functions available in most GIS software. The objective of the tool is to help uncover structure and patterns that may be hidden in complex geospatial data sets, and to provide graphical representations that can support understanding and knowledge construction. The design of the visualization environment incorporates several graphical representations of SOM output, including a distance matrix representation, two-dimensional (2-D) and three-dimensional (3-D) projections, 2-D and 3-D surfaces, and component plane displays.

3.3.1   Structure of the Integrated Visual-Computational Analysis and Visualization Environment

We have extended the graphical representations of the SOM training results, to highlight different characteristics of the computational solution and integrated them with other graphics into multiple views to allow brushing and linking for exploratory analysis purposes. There are a number of researches reflecting the interest in dynamic displays on the part of experts in cartographic data presentation (Cook *et al*., 1996; Dykes, 1997; Egbert and Slocum, 1992; Monmonier, 1992). Most often they suggest that brushing be applied to a map linked with one or more non-geographical presentations, showing individual values and statistics, and the visualization of neighborhood relationships. We use multiple views to offer alternative and different views of the data in order to stimulate the visual thinking process that is characteristic of visual exploration. Cartographic methods support the design for the effective use of visual variables with which the visualization is depicted. This makes the exploratory geovisualization environment appropriate for relating the position of the map units and the value at the map units represented by color coding, and for exploring correlations and relationships. The design of the interface incorporates several graphical representations that provide ways of representing similarity (patterns) and relationships, including a distance matrix representation, 2-D and 3-D projections, 2-D and 3-D surfaces, and component plane visualization.

The tool was developed based on the integration of Matlab, the SOM toolbox and spatial analysis (Martinez and Martinez, 2002). The main functionality of the visualization system includes pre-processing, the initialization and training of a SOM network, and visualization. Figure 3.4 describes the structure of the geovisualization system. The pre-processing consists of transforming primary data and converting them into an appropriate format. At this stage, input data are transformed and all components and variables of the data set are normalized. After training the network, the visualization component provides features for visualizing the data, using different techniques. A link between the different views is provided for the exploration of relationships.

The SOM network was trained using the SOM toolbox. In the SOM toolbox, the data set is first put in a Matlab 'struct', a data structure that contains all information related to the data set in different fields for the numerical data (a matrix in which each row is a data sample and each column a component), strings, as well as other related information. Since the SOM algorithm uses Euclidean metric distance to measure distances between vectors, scaling of variables is needed to give equal importance to the variables. Linear scaling of all variables is used so that the variance of each is equal to 1. Other normalization methods such as logarithmic scaling and histogram equalization are offered. The original scale values can easily be returned when needed. Missing data are also handled in the SOM toolbox. The input vectors x are compared with the reference vectors m_i, using those components that are available in x.

3.3.2   User Interface

The interface integrates the different representations into multiple views, which are used to simultaneously present interactions between several variables over the space of the SOM, maps and parallel coordinate plots, and to emphasize visual change detection and the
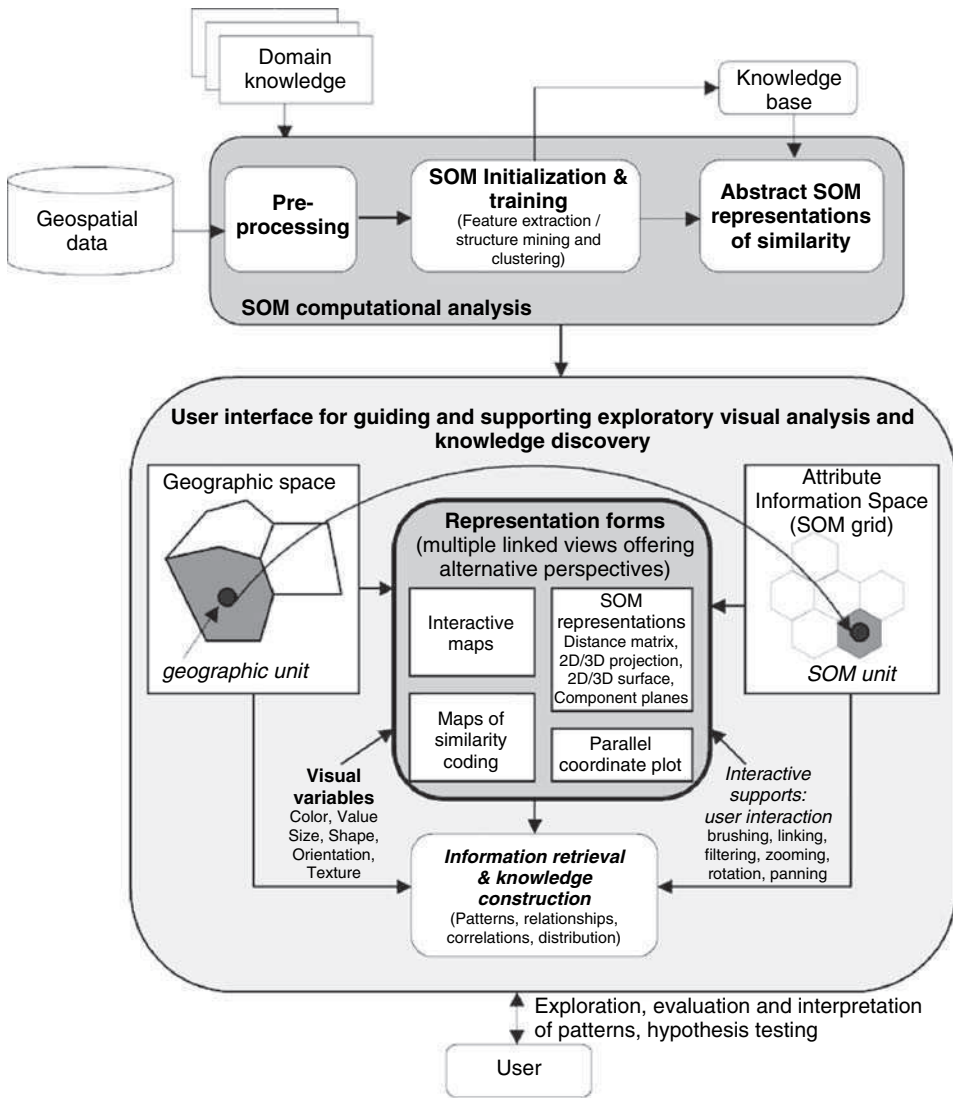
Figure 3.4   *Structure of the geovisualization environment*

monitoring of the variability through the attribute space. These alternative and different views of the data can help stimulate the visual thinking process that is characteristic of visual exploration.

The interface design focuses on three important aspects:

- representation forms (map, grid, surface, projection);
- visualization techniques (distance matrix, component planes display, 2-D and 3-D views of surface plots and projections);
- interaction techniques (brushing, panning, rotation, zooming).

Because the user develops a mental model of the system, it is important that the design helps construct a clear image of the system. For perceptual effectiveness (Eick, 1997), the interface attempts to provide displays in a way that seems natural for interpretation: in a grid, on a map, on a surface, in a 3-D space, with position showing internal relationships. Users have the possibility of visually relating information or aggregations of data to reveal the clustering structure or common visual properties. From the human–computer interaction (HCI) perspective, a number of interaction strategies can help achieve the goals of visual exploration. The interface offers interactive filters for changing the relative positions of elements of the display, changing by rotation the perspective from which it is seen, and displaying detailed information to have access to actual data values on a specific data item of interest. Such transformations of views can interactively modify and augment visual structures, and support the likelihood of emergence (Peuquet and Kraak, 2002). We use different interaction techniques to enhance data exploration, including brushing and linking, panning, zooming, and rotation.

Users can perform a number of exploratory tasks to understand the structure of the data set as a whole and to explore detailed information such as correlations and the relationships for selected attributes of the data set. This is intended to guide them in hypothesis testing, evaluation and interpretation of patterns from general patterns extracted to specific selection of attributes and spatial locations. Other supportive views are provided for further exploration of the displays including zooming, panning, rotation and 3-D view. Figure 3.5 shows the interface of the integrated geovisualization environment. An important issue in the design of geovisualization environments is to provide ways of representing similarity (patterns) and relationships in a way that facilitates the perceptual and cognitive processes involved (MacEachren, 1995). To achieve this goal, cartographic design principles are needed to provide an effective integration of visual variables used in the representation forms, while information visualization techniques provide alternatives for the user interaction necessary to complete the tasks. Bertin's fundamental six visual variables (Bertin, 1983) for graphical information processing can serve as the basis for this integration. These variables (size, value, texture/grain, color, orientation and shape) can be used, either alone or in combination, to depict different arrangements of objects in the graphical representations. For example, size is an effective perceptual data-encoding variable and shape is useful for visual segmentation.

3.3.3   Visual Exploration Support for General Patterns and Clustering

The SOM offers a number of distance matrix visualizations to show the cluster structure. These techniques show distances between neighboring units. The most widely used distance matrix technique is the U-matrix (Ultsch and Siemon, 1990). The four goals of the visualization described in Section 3.2.3 are covered by the different representations of the data. Similarity (patterns) is represented in the distance matrix representation. Relationships are viewed in fine detail with the component plane visualization. The distribution and irregularities are represented in the visualization of the component planes and in the projections.

The default view in the user interface (Figure 3.5) offers after the data has been loaded and the SOM network trained the general clustering structure of the data in different perspectives (maps, projections, unified distance matrix and parallel coordinate plot).
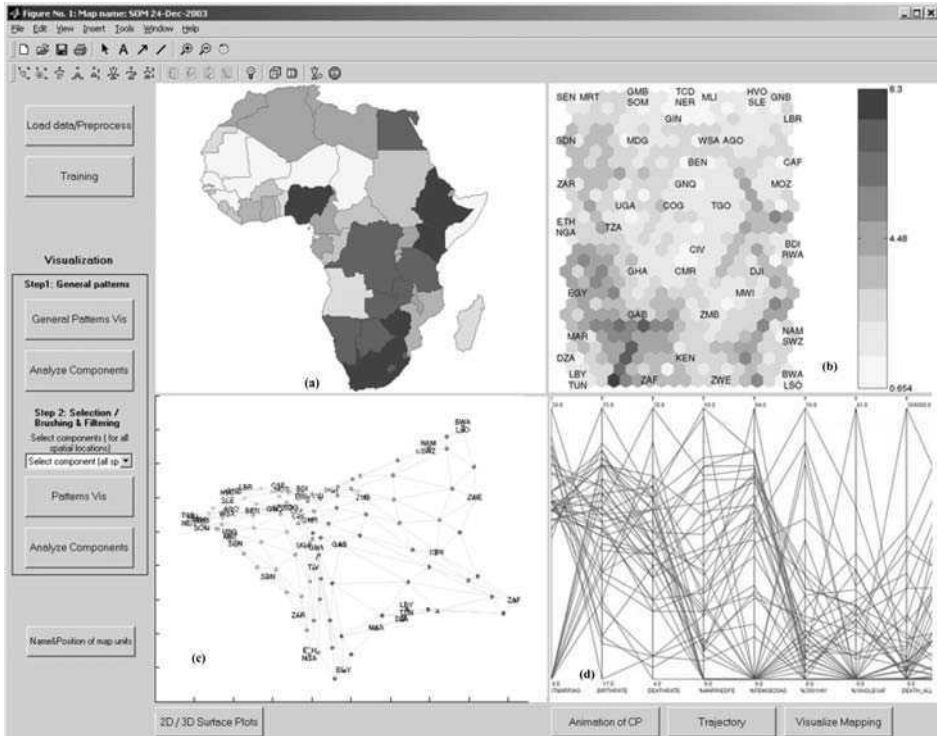
Figure 3.5 *The user interface for the exploratory geovisualization environment in multiple views with the visualization of component planes (bottom left) and map unit labels (bottom right). The default view shows the representation of the general patterns and clustering in the input data: the unified distance matrix showing clustering and distances between positions on the map (b). Alternative representations of the SOM general clustering of the data with projection of the SOM results in 3-D space (c); and a map of the similarity coding extracted from the SOM computational analysis (a), and parallel coordinate plot (d) (See Colour Plate 5)*

This general view implements a number of distance matrix visualizations to explore the SOM results and show the cluster structure and similarity (patterns). The similarity matrix representation visualizes the distances between the network neurons [represented here by hexagonal cells in Figure 3.6(a)]. It contains the distances from each unit center to all of its neighbors. The distance between the adjacent neurons is calculated and presented in different colorings. A dark coloring between the neurons corresponds to a large distance and thus represents a gap between the values in the input space. A light coloring between the neurons signifies that the vectors are close to each other in the input space. Light areas represent clusters and dark areas cluster separators. This representation can be used to visualize the structure of the input space and to get an impression of otherwise invisible structures in a multidimensional data space. The similarity representation shows more hexagons than the actual number of neurons used in the network, because it shows not only the distance value at the map units but also the distances between map units.

The SOM unlike other projection methods in general, tries to preserve not the distances directly but rather the relations or local structure of the input data. While the U-matrix
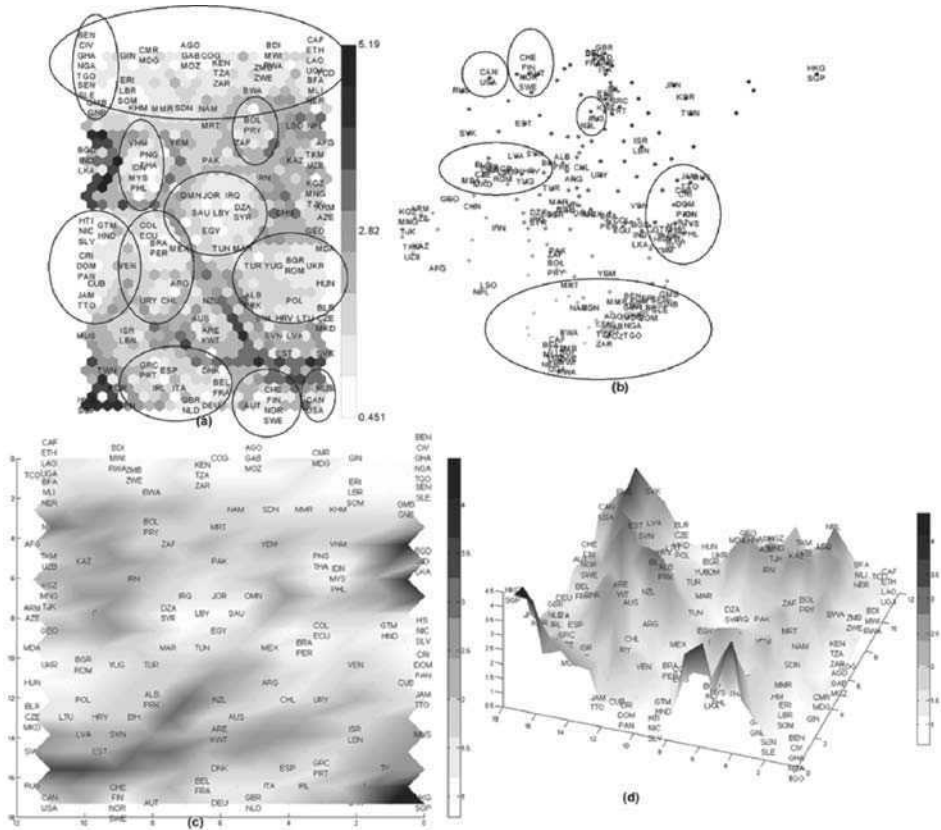
Figure 3.6   *Similarity matrix representation of the data set (a), PCA projection of SOM results (b), 2-D surface plot of distance matrix (c) and 3-D surface plot of distance matrix (d)*

is a good method for visualizing clusters, it does not provide a very clear picture of the overall structure of the data space because the visualization is tied to the map grid. Some projection methods can be used (e.g. Sammon's mapping, PCA) to give a more informative picture [Figure 3.6(b)]. The projection of the SOM results provides freely specified coordinates in 2-D or 3-D space. The third dimension uses the value (or weight) associated with the map units according to multidimensional attributes. Color, size, types of markers used as identifiers of map units, and lines for connecting the map units are used for more interactive exploration.

In the distance matrix [Figure 3.6(a)], countries having similar characteristics based on the multivariate attributes are positioned close to each other and the distance between them represents the degree of similarity or dissimilarity. These common characteristics representations can be regarded as the socio-economic standard for the countries. In Figure 3.6(b), the projection of the SOM offers a view of the clustering of the data with data items depicted as colored. The clustering structure can also be viewed in the interface, as 2-D or 3-D surfaces representing the distance matrix [Figure 3.6(c) and (d)], using color value to indicate the average distance to neighboring map units.

In Figure 3.5, the different views on the general structure of the data set are provided. Alternative representations of the clustering of the data are provided in 2-D and 3-D projections (using projection methods such as Sammon's mapping and PCA), 2-D and 3-D surface plots, and parallel coordinate plot [Figure 3.5(d)]. In Figure 3.5(c), the projection of the SOM offers a view of the clustering of the data with data items depicted as colored nodes. Similar data items are grouped together with the same type or color of markers. Size, position and color of markers can be used to depict the relationships between the data items. This gives an informative picture of the global shape and the overall smoothness of the SOM in 2-D or 3-D space. Exploration can be enhanced by interactive rotation, zooming and selection in 3-D view. Connecting the map units with lines can reveal the shape of clusters and relationships among them. The cluster structure can also be viewed as 2-D or 3-D surfaces representing the distance matrix [Figure 3.6(c) and (d)] using color value to indicate the average distance to neighboring map units. This is an example of spatialization (Fabrikant and Skupin, 2005) that uses a landscape metaphor to represent the density, shape, and size or volume of clusters. Unlike the projection in Figure 3.5(c) that shows only the position and clustering of map units, areas with uniform color are used in the surface plots to show the clustering structure and relationships among map units. In the 3-D surface [Figure 3.6(d)], color value and height are used to represent the clustering of map units according to the multidimensional attributes.

3.3.4   Exploration of Correlations and Relationships

As a second stage of the visualization process, the interface offers options to explore correlations and relationships in the input data. This is implemented by the component plane display (Figure 3.7). As discussed above, here the component planes show the values of different attributes for the different countries. They are used to support exploratory tasks and to facilitate the understanding of the relationships in the data.

3.4   EXAMPLE EXPLORATION OF GEOGRAPHICAL PATTERNS USING THE PROTYOTYPE EXPLORATORY GEOVISUALIZATION ENVIRONMENT

The prototype is used in an example for exploring a data set containing socio-economic indicators. In this section, the data set is explored, and different visualization techniques are used to illustrate the exploration of (potential) patterns within the different options of the interface. This example is used to examine the integration of the different graphical representations in the user interface.

3.4.1   The Data Set Explored

The prototype was used to explored a socio-economic data set related to geography and economic development (Gallup *et al*., 1999) to analyze the complex relationships between
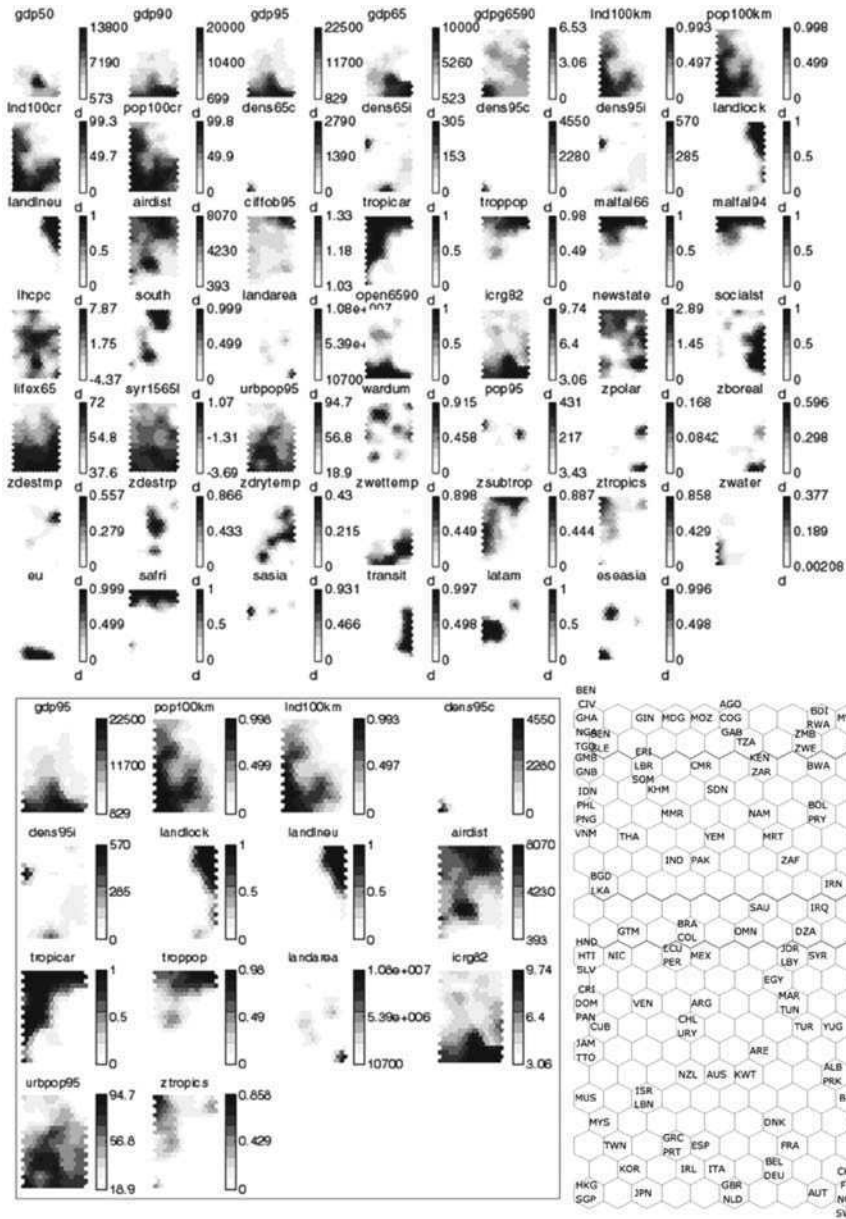
Figure 3.7 *Detailed exploration of the data set using the SOM component visualization: all the components can be displayed to reveal the relationships between the variables and the spatial locations (countries). Selected components related to a specific hypothesis (for example related to economic development at the bottom) can be further explored. Position of the countries on the SOM map (bottom right)*

geography and macroeconomic growth (e.g. how geography may directly affect growth, and the effect location of the countries and climate may have on income levels, income growth, transport costs, disease burdens and agricultural productivity). Additionally, the relationships between geographic regions, whether located far from the coast, and population density, population growth, economic growth and the economic policy itself are other aspects the study of this data set intends to explore. The data set contains 48 variables on the economy, physical geography, population and health of 150 countries (Tables 3.1 and 3.2).

Table 3.1 *Description of the variables of the data set*

Variable	Description	Variable	Description	Variable	Description
gdp50	GDP per capita in 1950	ciffob95	shipping cost, 1995	pop95	population in 1995
gdp90	GDP per capita in 1990	tropicar	% land in geographic tropics	zpolar	% land area in polar non-desert
gdp95	GDP per capita in 1995	troppop	% population in geographic tropics, 1994	zboreal	% land area in boreal regions
gdp65	GDP per capita in 1965	malfal66	malaria index, 1966	zdestmp	temperature desert
gdpg6590	GDP per capita growth from 1965 to 1990	maffal94	malaria index 1994	zdestrp	tropical + subtropical desert
lnd100 km	% land within 100 km coast	lhcpc	log hydrocarbons per capita, 1993	zdrytemp	% land area within dry temperature
pop100 km	% population within 100 km coast	south	southern hemisphere countries	zwettemp	% land area wet temperate
lnd100cr	% land within 100 km coast or river	landarea	land area (sq km)	zsubtrop	% land area in the subtropics
pop100cr	% population within 100 km coast or river	open6590	openness, 1965–1990	ztropics	% land area in the tropics
dens65c	coastal population density, 1965	icrg82	quality of public institutions,	zwater	water (lakes and ocean)
dens65i	inland population density, 1965	newstate	timing of independence	eu	Western Europe
dens95c	coastal population density, 1995	socialist	socialist country, 1950–1995	safri	Sub-Saharan Africa
dens95i	inland population density, 1995	lifex65	life expectancy, 1965 (UN)	sasia	south Asia

Table 3.1   *(Continued)*

Variable	Description	Variable	Description	Variable	Description
landlock	landlocked	syr15651	log years secondary schooling, 1965	transit	transition countries
lnadlneu	landlocked, not west and central Europe	urbpop95	% population urban, 1995 (world bank)	latam	latin America and Caribbean
airdist	km to closest major port	wardum	had external war, 1960–1985	eseasia	east and southeast Asia

Table 3.2   *Countries included in the study*

Code	Country	Code	Country	Code	Country	Code	Country
AFG	Afghanistan	ERI	Eritrea	LBR	Liberia	RUS	Russian Federation
AGO	Angola	ESP	Spain	LBY	Libya Arab Jamahiriya	RWA	Rwanda
ALB	Albania	EST	Estonia	LKA	Sri Lanka	SAU	Saudi Arabia
ARE	United Arab Emirates	ETH	Ethiopia	LSO	Lesotho	SDN	Sudan
ARG	Argentina	FIN	Finland	LTU	Lithuania	SEN	Senegal
ARM	Armenia	FRA	France	LVA	Latvia	SGP	Singapore
AUS	Australia	GAB	Gabon	MAR	Morocco	SLE	Sierra Leone
AUT	Austria	GBR	United Kingdom	MDA	Moldova, Republic of	SLV	El Salvador
AZE	Azerbaijan	GEO	Georgia	MDG	Madagascar	SOM	Somalia
BDI	Burundi	GHA	Ghana	MEX	Mexico	SVK	Slovak Republic
BEL	Belgium	GIN	Guinea	MKD	The fmr Yug. Rep. Macedonia	SVN	Slovenia
BEN	Benin	GMB	Gambia	MLI	Mali	SWE	Sweden
BFA	Burkina Faso	GNB	Guinea Bissau	MMR	Myanmar	SYR	Syrian Arab Rep.
BGD	Bangladesh	GRC	Greece	MNG	Mongolia	TCD	Chad
BGR	Bulgaria	GTM	Guatemala	MOZ	Mozambique	TGO	Togo
BIH	Bosnia and Herze-govina	HKG	Hong Kong	MRT	Mauritania	THA	Thailand
BLR	Belarus	HND	Honduras	MUS	Mauritius	TJK	Tajikistan
BOL	Bolivia	HRV	Croatia	MWI	Malawi	TKM	Turkmenistan
BRA	Brazil	HTI	Haiti	MYS	Malaysia	TTO	Trinidad and Tobago
BWA	Botswana	HUN	Hungary	NAM	Namibia	TUN	Tunisia
CAF	Central African Rep.	IDN	Indonesia	NER	Niger	TUR	Turkey
CAN	Canada	IND	India	NGA	Nigeria	TWN	Taiwan

CHE	Switzerland	IRL	Ireland	NIC	Nicaragua	TZA	Tanzania
CHL	Chile	IRN	Iran	NLD	Netherlands	UGA	Uganda
CHN	China	IRQ	Iraq	NOR	Norway	UKR	Ukraine
CIV	Côte d'Ivoire	ISR	Israel	NPL	Nepal	URY	Uruguay
CMR	Cameroon	ITA	Italy	NZL	New Zealand	USA	United States
COG	Congo	JAM	Jamaica	OMN	Oman	UZB	Uzbekistan
COL	Colombia	JOR	Jordan	PAK	Pakistan	VEN	Venezuela
CRI	Costa Rica	JPN	Japan	PAN	Panama	VNM	Vietnam
CUB	Cuba	KAZ	Kazakhstan	PER	Peru	YEM	Yemen
CZE	Czech Republic	KEN	Kenya	PHL	Philippines	YUG	Yugoslavia
DEU	Germany	KGZ	Kyrgyz Republic	PNG	Papua New Guinea	ZAF	South Africa
DNK	Denmark	KHM	Cambodia	POL	Poland	ZAR	Zaire
DOM	Dominican Republic	KOR	Korea	PRK	Korea Dem. People's Rep.	ZMB	Zambia
DZA	Algeria	KWT	Kuwait	PRT	Portugal	ZWE	Zimbabwe
ECU	Ecuador	LAO	Lao PDR	PRY	Paraguay		
EGY	Egypt	LBN	Lebanon	ROM	Romania		

3.4.2   Exploration Support for General Patterns and Clustering

In the distance matrix [Figure 3.5(b)], countries having similar characteristics based on the multivariate attributes are positioned close to each other, and the distance between them represents the degree of similarity or dissimilarity. For the exploration of the SOM visualizations, some geographic maps of the data set are represented in Figure 3.8 for selected attributes: coastal population density, percentage population within 100 km of coast or river, GDP per capita, distance (km) to closest major port, percentage of land area in the subtropics, and percentage of land in the geographic tropics.

The U-matrix in Figure 3.6(a) reveals commonalities among countries based on the multivariate attributes. At the top of the map, we have the poor economies, mostly the African countries, and at the bottom the rich economies (see Table 3.2).

From this clustering structure, differences can be observed between countries in different parts of the world. A very striking observation is that the clustering somehow reflects the geographic location of the countries. This confirms the general hypothesis suggesting that there is a relationship between the geographic location of the countries and economic growth (Gallup *et al.*, 1999). Even further clustering that reflects the distinct geographic regions is obtained with the similarity matrix representation: West Africa, Southern Africa, the Middle East, Europe, South America, North America (USA and Canada), and Asia. The European countries are in three different clusters next to each other and close to USA and Canada.

A few cases do not reflect this geographic relationship. Laos is found in a cluster with some poor African economies (Central Republic of Africa, Ethiopia, Uganda, Chad, Burkina Faso, Mali, Niger). This may be because Laos's economic characteristics are low compared with those of the other Asian countries and it falls closer to Africa than Asia in this respect. Other countries that have no obvious characteristics in common with
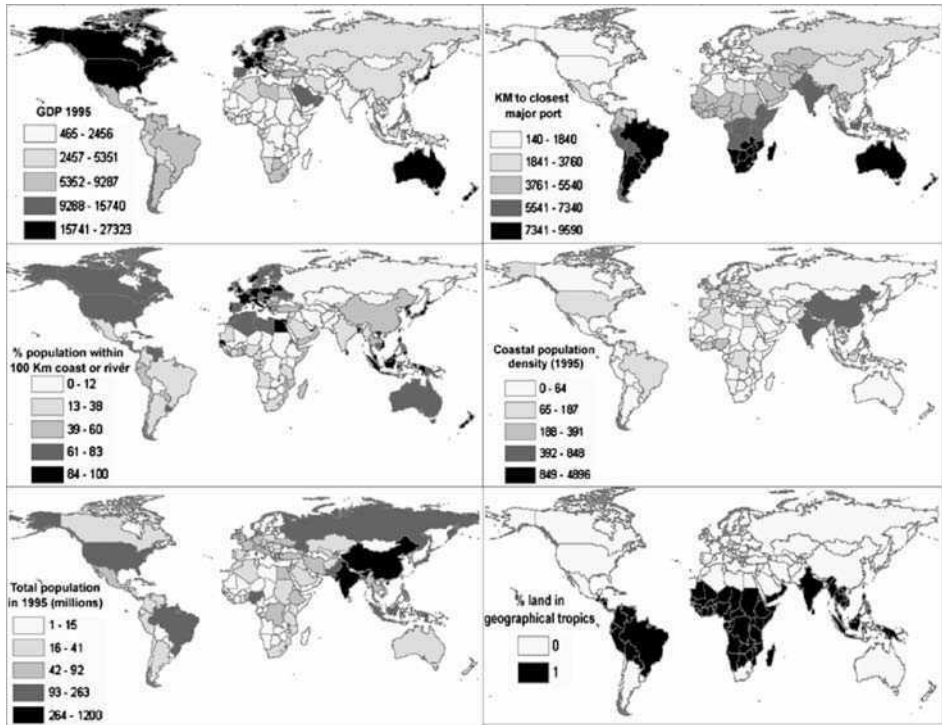
Figure 3.8   *Some maps were created using ArcGIS and represent selected attributes: coastal population density, percentage population within 100 km of coast or river, GDP per capita, distance (km) to closest major (European) port, percentage of land area in the subtropics, and percentage of land in the geographic tropics*

the others in the same geographic region include Mauritania, Yemen, Pakistan, Iran and Mauritius. South Africa has particular characteristics that position the country far away from other African countries and closer to the Middle East, Iran and Pakistan, on the one hand, and close to Bolivia and Paraguay on the other.

The same information provided in the distance matrix can be viewed using 2-D or 3-D surfaces.

3.4.3   Exploration of Correlations and Relationships

Exploration of relations among attributes and map units is primarily based on the visualization of component planes (Figure 3.7), user selection and interaction from the user interface (Figure 3.5). Component planes visualization is used to offer a supportive view that provides exploration of the relationships among different variables for specified locations. The component planes show the values of different attributes for the different map units (countries) represented by hexagonal cells (neurons of the SOM network) and how each input vector varies over the space of the SOM units. In comparison with geographic maps, patterns and relationships among all the attributes can be easily

examined in a multiple views using the SOM component planes visualization. Two variables that are correlated will be represented by similar displays. In Figure 3.7, all the components are displayed and a selection of some of them can be made for further analysis. This kind of visual representation (imagery cues) can facilitate visual detection, and has an impact on knowledge construction (Keller and Keller, 1992).

For this example exploration of relationships and correlations, a summary of the geographic patterns was made based on the average GDP per capita, total population and land area, and several key variables that can be related to economic development: the extent of land in the geographic tropics, the proportion of the population within 100 km of the coastline or within 100 km of the coastline or ocean-navigable river, the percentage of population that lives in landlocked countries, the average distance by air (weighted by country populations) to the closest core economic areas, the density of human settlement (population per square km) in the coastal region (within 100 km of the coastline) and the interior (beyond 100 km from the coastline). The tropical countries were defined as being those that have half or more of the land area in the geographic tropics.

From these patterns the following question can be raised: How great a role has geographic location of the countries played in economic growth, assuming that economic policies and institutions are well established?

This complex linkage between geography, demography, health and economic performance requires closer examination. Using the component planes visualization, we examine two geographic correlates of economic development that were outlined by Gallup *et al*. (1999) and generate other possible hypotheses that the SOM technique allows. The countries in the geographic tropics are nearly all poor. Almost all high-income countries are in the mid and high latitudes. Coastal economies are generally higher income than the landlocked economies.

From the component plane visualization in Figure 3.7, a simple view of the displays allows the attributes to be visually related to the spatial locations. Observed correlations and relationships can help in the understanding of the patterns in the data. To enhance visual detection of the relationships and correlations, the components can be ordered so that variables that are correlated are displayed next to each other in a way similar to the collection maps of Bertin (1981). From the displays in Figure 3.7, relationships among different variables can be observed in one multiple view. For example the poorest economies (reference to the 1995 GDP from the data set) have characteristics such as large proportion of land and population in the geographic tropics, population highly concentrated in the interior, often landlocked, small proportion of land within 100 km of the coast or river, located in the southern hemisphere, and often with tropical or subtropical deserts. Most of these characteristics were identified as closely associated with low income in general (Gallup *et al*., 1999). Other common characteristics of these countries that can be seen as a consequence of the low income are also visualized in the component planes. The poor countries have low life expectancy, high shipping costs, and heavy disease burdens of malaria; they are very far from the closest core markets in Europe, and many have external wars. From these observations, it can be hypothesized that various aspects of tropical geography and public health are vitally important and affect economic growth (Bloom and Sachs, 1998). South Asia, Latin America, the eastern European countries and the former Soviet Union are like Sub-Saharan Africa, with more concentrated in the interior rather than at the coast. Landlocked countries may be particularly disadvantaged by their lack of access to the sea. They all have low income

except those in western and central Europe (integrated into regional European market and associated low-cost trade). High population density seems to be favorable for economic development in coastal regions with good access to internal, regional and international trade. The poorest economies have low urban population density. The urban areas seem to develop more in the coastal regions.

3.5   CONCLUSION AND DISCUSSION

In this chapter we have presented the implementation of an approach to integrate computational and visual analysis into the design of a prototype visualization environment intended to contribute to the analysis of large volumes of geospatial data. This approach focuses on the application of the SOM algorithm to extract patterns and relationships in geospatial data, and the visual representation of derived information. We have presented an application of the SOM algorithm for exploratory visualization, as applied to socio-economic data sets. The SOM demonstrates important capabilities for features extraction, clustering and the projection of the data set. The spatial representation of the SOM (grid) provides opportunities for exploring the attribute space in relation to the spatial locations. A number of visualization techniques were used to explore ways of supporting exploratory tasks and knowledge construction.

A user interface was developed to integrate the different graphical representations and support the exploration process by supporting a number of user activities. The interface is structured to provide a global view and summary of the data as well as tools for detailed exploration of relationships and correlations for exploratory analysis purposes. Interaction was needed to enhance user goal-specific querying and selection from the general patterns extracted to more specific user selection of attributes and spatial locations for exploration, hypothesis generation, and knowledge construction. Interactive manipulation (zooming, rotation, panning, filtering and brushing) of the graphical representations was used provided to enhance user interaction, the objective being to explore ways of supporting visual exploration and knowledge construction.

As such, the SOM can be used as an effective tool to visually detect correlations among operating variables in a large volume of multivariate data. New knowledge can be unearthed through this process of exploration, which can be followed by the identification of associations between attributes, and finally the formulation and ultimate testing of hypotheses. Since the SOM represents the spatial clustering of the multivariate attributes, the visual representation becomes more accessible and easy for exploratory analysis and knowledge discovery. This kind of spatial clustering makes it possible to conduct exploratory analyses to help in identifying the causes and correlates of health problems when overlaid with other data, such as environmental, social, transportation, and facilities data. Such map overlays have also been important hypothesis-generating tools in research and policy-making.

The link between the attribute space visualization based on the SOM, the geographical space with maps representing the SOM results, and other graphics such as parallel coordinate plots in multiple views offers alternative perspectives for better exploration, evaluation and interpretation of patterns, which ultimately supports knowledge construction. These aspects will be the focus of a subsequent usability test to characterize

the overall effectiveness of the representations used in the exploratory geovisualization environment.

REFERENCES

Bertin, J. (1981). Graphics and Graphic Information Processing. Berlin, Walter de Gruyter.

Bertin, J. (1983). Semiology of Graphics: Diagrams, Networks, Maps. Madison, WI, University of Wisconsin Press.

Bloom, D. E. and J. D. Sachs (1998). Geography, Demography and Economic Growth in Africa. Harvard, Center for International Development, Harvard University.

Cook, D. *et al*. (1996). Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data Using Linked Software. Computational Statistics: Special Issue on Computeraided Analysis of Spatial Data **11**(4): 467–480.

Dykes, J. A. (1997). Exploring Spatial Data Representation with Dynamic Graphics. Computers & Geosciences **23**(4): 345–370.

Egbert, S. L. and T. A. Slocum (1992). EXPLOREMAP: An Exploration System for Choropleth Maps. Annals, Association of American Geographers **82**(2): 275–288.

Eick, S. G. (1997). Engineering Perceptually Effective Visualizations for Abstract Data. Scientific Visualization: Overview, Methodologies and Techniques. G. M. Nielson, H. Hagen, H. Müller. Los Alamitos, CA, IEEE Computer Society Press: 191–210.

Fabrikant, S. I. and A. Skupin (2005). Cognitively Plausible Information Visualization. Exploring GeoVisualization. J. Dykes, A. M. MacEachren, M. J. Kraak. Amsterdam, Elsevier: 667–690.

Fayyad, U. *et al*. (1996). From Data Mining to Knowledge Discovery in Databases. Artificial Intelligence Magazine **17**: 37–54.

Gahegan, M. (2000). On the Application of Inductive Machine Learning Tools to Geographical Analysis. Geographical Analysis **32**(2): 113–139.

Gahegan, M. and B. Brodaric (2002). Computational and Visual Support for Geographical Knowledge Construction: Filling the Gaps between Exploration and Explanation. Symposium on Geospatial Theory, Processing and applications, Ottawa, Canada.

Gahegan, M. and M. Takatsuka (1999). Dataspaces as an Organizational Concept for the Neural Classification of Geographic Datasets. 4th International Conference on GeoComputation, Fredericksburg, VA, USA.

Gahegan, M. *et al*. (2001). The Integration of Geographic Visualization with Databases, Data Mining, Knowledge Discovery Construction and Geocomputation. Cartography and Geographic Information Science **28**(1): 29–44.

Gallup, L. J. *et al*. (1999). Geography and Economic Development. Harvard, Center for International Development, Havard University.

Keller, P. and M. Keller (1992). Visual Clues: Practical Data Visualization. Los Alamitos, CA, IEEE Computer Society Press.

Kohonen, T. (1989). Self-Organization and Associative Memory. Heidelberg, Spring-Verlag.

Koua, E. L. (2002). Self-organizing Maps for Geospatial Information Visualization. 98th Annual Meeting of the American Association of Geographers, Los Angeles, USA.

Liu, W., S. Gopal and C. Wood cock (2001). Spatial Data Mining for Classification, Visualization and Interpretation with ARTMAP Neural Network. Data Mining for Scientific and Engineering Applications. R. Grossman. Dordrecht, Kluwer: 205–222.

MacEachren, A. M. (1994). Visualization in Modern Cartography: Setting the Agenda. Visualization in Modern Cartography. D. R. F. Taylor. Oxford, Pergamon: 1–12.

MacEachren, A. M. (1995). How Maps Work: Representation, Visualization, and Design. New York, The Guilford Press.

MacEachren, A. M. and M. J. Kraak (2001). Research Challenges in Geovisualization. Cartography and Geoinformation Science **28**(1).

MacEachren, A. M. *et al.* (1999). Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographical Visualization with Knowledge Discovery in Databases Methods. International Journal of Geographical Information Science **13**(4): 311–334.

Martinez, W. L. and A. R. Martinez (2002). Computational Statistics Handbook with MATLAB. Boca Raton, Chapman & Hall/CRC.

Miller, H. J. and J. Han (2001). Geographic Data Mining and Knowledge Discovery. London, Taylor and Francis.

Monmonier, M. (1992). Authoring Graphics Scripts: Experiences and Principles. Cartography and Geographic Information Systems **19**(4): 247–260.

Nielson, G. M. *et al.* (1997). Scientific Visualization. Overviews, Methodologies, Techniques. Washington, IEEE Computer Society.

Openshaw, S. *et al.* (1990). Building a Prototype Geographical Correlates Machine. International Journal of Geographical Information Systems **4**(4): 297–312.

Peuquet, D., J. and M. Kraak, J. (2002). Geobrowsing: Creative Thinking and Knowledge Discovery Using Geographic Visualization. Information Visualization **1**: 80–91.

Roddick, J. F. and B. G. Lees (2001). Paradigms for Spatial and Spatio-Temporal Data Mining. Geographic Data Mining and Knowledge Discovery. H. J. Miller, J. Han. London, Taylor and Francis: 33–49.

Sibley, D. (1988). Spatial Applications of Exploratory Data Analysis. Norwich, Geo Books.

Skupin, A. (2003). A Novel Map Projection Using an Artificial Neural Network. 21st International Cartographic Conference (ICC), Cartographic Renaissance, Durban, South Africa.

Ultsch, A. and H. Siemon (1990). Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. Proceedings International Neural Network Conference INNC'90P, Dordrecht, The Netherlands.

Wachowicz, M. (2000). The Role of Geographic Visualization and Knowledge Discovery in Spatio-Temporal Modeling. Publications on Geodesy **47**: 27–35.

Weijan, W. and D. Fraser (1996). Spatial and Temporal Classification with Multiple Self-Organizing Maps. Society of Photo-Optical Instrumentation **2955**: 307–314.

Weldon, J. L. (1996). Data Mining and Visualization. Database Programming and Design **9**(5).

4

Visual Exploration of Spatial Interaction Data with Self-Organizing Maps

Jun Yan[1] and Jean-Claude Thill[2]

[1] *Department of Geography and Geology, Western Kentucky University, Bowling Green, KY 42101, USA*

[2] *Department of Geography and National Center for Geographic Information and Analysis, University at Buffalo, The State University of New York, Amherst, NY 14261, USA*

4.1   INTRODUCTION

The analysis of the modalities of spatial interaction has been a long-standing concern among spatial scientists because they are known to be the generating force behind many geographic structures (Gould, 1991) and because of the multiplicity of dimensions along which they can be examined [Figure 4.1(a)]. Although it is still quite difficult to obtain data on movement at the elemental level, e.g. at the person, firm, or vehicle level, many types of aggregate data at a fine geographic resolution have become available. For instance, the United States Bureau of Transportation Statistics (BTS) coordinates the conduct of regular surveys of personal travel on long and short distance, such as the American Travel Survey. The BTS also has responsibility for maintaining geographically disaggregated databases of commercial trade flows between the United States and Canada. In the field of domestic air travel, this agency maintains the Airline Origin and Destination Survey Database Market Table (DB1BMarket), which is derived from a 10 % sample of all airline itineraries issued quarterly by each airline. Because of its fine geographic
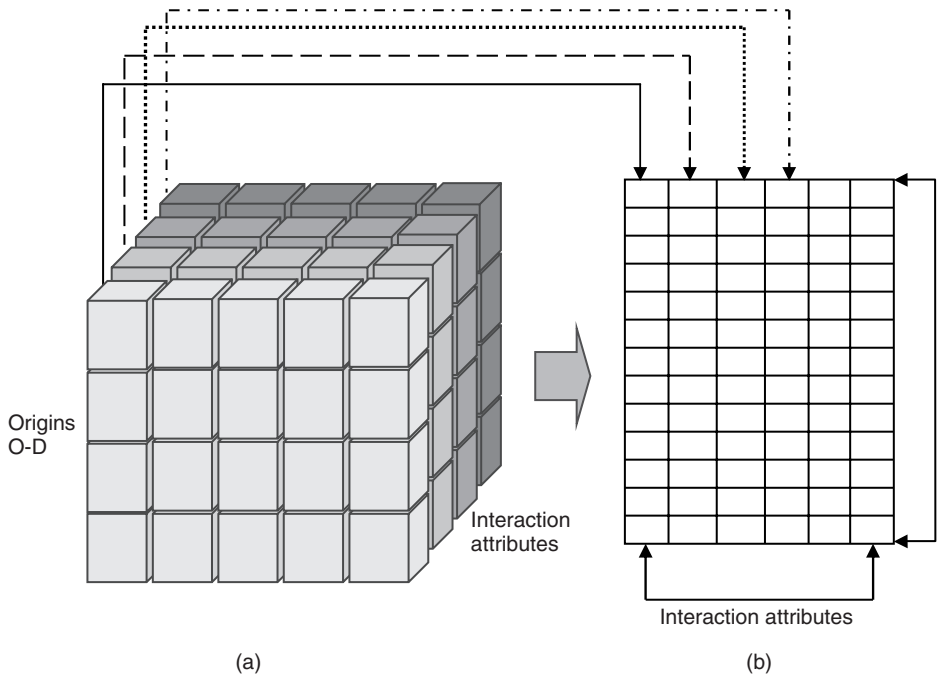
Figure 4.1   *Dimension tube of spatial interaction data and data transformation procedure: (a) O-D matrices; (b) dyadic matrix*

detail (origins and destinations are at the airport level), the DB1BMarket database can reasonably be expected to conceal a wealth of information that could provide unparalleled insights into the formation of airline market structures over time and across space.

In fact, with of the availability of large-scale digital spatial interaction databases rich in attribute information (flow type, transportation mode, timeframe, and others) such as DB1BMarket, the opportunity exists for researchers to examine the formation of different types of spatial interactions as well as their interdependencies by *exploring* the patterns embedded in the data. Methods of flow data compression and of visual exploration that have been proposed so far for this purpose have been found to exhibit serious limitations. In this chapter, a novel *exploratory analysis* approach is introduced to extract significant geographic patterns in large spatial interaction flow databases. The computational method of self-organizing maps (SOMs) is the search engine in this process. To facilitate the data exploration and knowledge discovery process, an interactive visual data mining (VDM) environment is proposed, in which various visualization forms are integrated by implementing a number of interaction techniques. This chapter aims to illustrate the potential usefulness of this integrated visual and computational approach in extracting novel geographic structures from large spatial interaction databases, for which traditional visualization techniques and more conventional data compression techniques are inherently problematic. Findings presented in this chapter come from the study of air travel structures extracted from the 2002 DB1BMarket database.

The chapter is organized as follows. Section 4.2 gives a brief overview of some relevant lines of research on visual exploration and spatial interaction systems. Section 4.3 will present the VDM environment proposed for the exploration of spatial interaction

databases. A selection of results from the US domestic air transportation case study is discussed in Section 4.4. Conclusions are presented in Section 4.5.

4.2   THEORETICAL BACKGROUND

4.2.1   Spatial Data and Visual Exploration

The exploration of spatial data involves some synthetic description of the data through a process of discovery of geographical structures that does not necessitate more than a few a priori assumptions. As in data mining and knowledge discovery in databases (Frawley *et al.*, 1991), new knowledge may be acquired through a highly iterative and interactive process of target data selection, structure extraction/data mining, and evaluation/verification. In geography, the amount of geographically referenced data continues to accumulate with the spread of information technologies, digital mapping, satellite imagery, and the global diffusion of geographic information systems (GIS). This certainly provides us strong rationale to develop a data-driven, inductive approach to geographic analysis and modeling with the objectives of facilitating the creation of new knowledge and aiding the processes of scientific discovery and deductive modeling, as argued by Openshaw (1999).

Many methods of exploratory spatial data analysis (ESDA) have strong ties to visualization. In ESDA, visual methods are not only instrumental in verifying and evaluating results, but also in generating and suggesting patterns and relationships. Data not yet fully understood can be classified, summarized, and formed into high-level structures on the basis of which new concepts can be developed for the benefit of more robust spatial modeling and better spatial theories. Exploring unknown phenomena often requires that certain intuition and background knowledge be incorporated. Visualization happens to be a very powerful strategy for getting high-level human intelligence involved in this process since human vision is extremely effective when it comes to recognizing patterns, relationships, trends, and anomalies (Bailey and Gatrell, 1995; Wachowicz, 2001).

Whether spatial or not, most common visualization techniques have been developed for the exploration of univariate or bivariate data sets (Fotheringham, 1999; Fotheringham *et al.*, 2000; MacEachren and Kraak, 1997). Recent advances in geographic data mining and geographic knowledge discovery (Miller and Han, 2001) fully extend the functionality and applicability of the existing ESDA methods by offering new computational mechanisms to sift through large geographic databases for meaningful information. MacEachren *et al.* (1999) identify some common themes and potentials for the integration of (geo)visualization methods and computational methods by comparing how both are used in the search for patterns in large multivariate spatio-temporal environmental databases. A conceptual framework for this integration is proposed by Wachowicz (2001).

4.2.2   Exploring Interaction Flows

The concept of spatial interaction (SI) is often used to describe the process by which entities at different locations make any form of contact in the geographic space (Fotheringham and O'Kelly, 1989; Roy and Thill, 2004). The entities can be of various natures, i.e. individuals, vehicles, or organizations. The contacts between interacting entities are usually expressed in some sort of transfer of people, materials, information, or energy. In

a general sense, SI can be seen as a general term for describing the movement of people, goods, capital, or information over space.

Roy and Thill (2004) contend that, in spite of the long history of modeling of SI systems and of the processes controlling them, there is room for further enhancements. This is particularly the case when it comes to capturing the role of local contexts, locational patterns in interaction origins and destinations, and spatial interdependencies in framing functional relationships between locales. Such endeavor partakes in the broader trend that has permeated spatial analysis (Fotheringham, 1997) to stress local perspectives. Our ability to develop new deductive modeling paradigms rests squarely on prior knowledge of spatial structures embedded in SI data. To this day, this knowledge has primarily been deductive and little attention has been paid to the exploratory analysis of SI data. Exploratory techniques to representing spatial interaction data date back to Ullman's (1957) seminal analysis of US commodity flows and the Chicago Area Transportation Study (1959), in which movement is represented by 'desire lines' or aggregated to rectangular flow bands with width proportional to flow magnitudes. Movement and flow mapping quickly becomes problematic as the size of the spatial system under study becomes greater than trivially small. The recognition of spatial structures is gravely hindered by purely visual approaches to the exploration of spatial interaction data to the point that some form of data compression is advocated to reduce the apparent complexity found in large flow matrices.

Tobler is one of only a few scholars who have continuously contributed to this field of visualization and exploration of SI data. In the 1970s and 1980s, he published a series of papers that ally mathematical modeling and cartographic mapping methods (Tobler, 1976, 1978, 1981, 1987). The approach serves to visualize the surface of net flows between origins and destinations. The surface is displayed as a field of vectors, which approximate the gradient of a scalar potential computed from the relative net exchanges of flows. Tobler refers to this innovative data exploration method as 'winds of influence', by using the earth science analogy of 'pressure field' that gives rise to winds. The idea of 'vector field' has also been used by some other geographers to analyze both directional and distance components of movement encapsulated in SI data. For instance, Berglund and Karlström (1999) conducted a spatial autocorrelation analysis based on the local G statistic to study the spatial association of flows. Lu and Thill (2003) proposed a more comprehensive approach to detect hot spots of multi-location events and applied it to the study of vehicle theft and recovery in Buffalo, NY.

Marble and colleagues (Marble *et al*., 1995, 1997) also made a significant contribution to the exploratory analysis of spatio-temporal interregional flows with a new approach that makes extensive use of scientific visualization. Their approach implements some dynamic graphics-based tools, which allows the analyst to map various interregional flows, to examine both total set and subset of the flows, and to compare flow volumes with selected characteristics of origins and destinations. However, as the authors point out, cartographic mapping can become unwieldy when the number of pairwise interactions is very large and multidimensional flow matrices are under examination. To facilitate the visualization and analysis of geographic structures in interaction flows, Marble and his colleagues use a typical data projection method, namely, projection pursuit (PP) that reduces the dimensionality of flow matrices.

In fact, as early as the 1960s, some spatial scientists explored the effectiveness of reducing the complexity of SI data to uncover essential relationships (or structures) within

transportation flow matrices. As stated by Smith (1970, p. 411), the intended purpose is to identify 'generic locational characteristics of groups of origins, or of groups of destinations, or of groups of origins and destinations' that are not readily apparent from the inspection of flow matrices or maps thereof. Pioneering work was done by Berry (1962, 1966) along this line of research. He developed three factor analysis (FA) approaches to identify the major commodity flow patterns of India from 63 36×36 commodity flow matrices. The first approach consists in an R-mode analysis, whereby flow destinations are factored to identify clusters of destinations with similar profile of incoming flows. The Q-mode analysis accomplishes the same for flow origins. The third approach extracts structures among thematic dimensions (for instance, commodity types) on each origin-destination pair (dyad). Black (1973) uses the term 'Dyadic Factor Analysis' to describe the latter modality of application of FA. However, methods like PP and FA can only reduce data complexity in the thematic dimensionality of the data by identifying a smaller number of latent components that represent the fundamental structures. Those that can deal with data compression on both data volume (i.e. cases or observations) and thematic dimensionality are much better suited to explore the structures embedded in large SI datasets. The emerging geocomputational paradigm (Openshaw and Abrahart, 2000) and its integration with methods of information visualization into a new visual data mining (VDM) environment (Ferreira de Oliveira and Levkowitz, 2003) offers new ways to address the problem of condensed visualization of essentially relationships within a SI system.

4.3 METHODOLOGY

4.3.1 Limitations of Traditional Data Reduction Methods

In order to identify structures in large geographic databases with high thematic dimensionality, a crucial task is to reduce both the number of attributes (Data Projection) and the number of cases (Data Quantization) without losing too much useful information. This means that we need to filter out uninteresting items or attributes to retain essential structures and group similar cases. A number of conventional multivariate statistical methods such as FA, principal component analysis (PCA), multidimensional scaling (MDS), PP, k-means, and hierarchical clustering address this type of needs, but they share some limitations. To name a few, variables are often expected to be normally distributed; the assumption of linearity between variables is usually necessary and stationary is often required. Furthermore, these methods normally look for general or global relationships, not local structures within data.

These restrictions are hard to overcome in geographically referenced data, which are usually nonlinear, nonstationary, sparse, of no known distribution, often constituted in arbitrary geographic units, and available in large volume. In addition, they can only handle data compression in either data volume or thematic multidimensionality, one at a time. To achieve both, the only solution is to carry out two separate tasks sequentially: reducing the number of variables first, and then the number of data items, or the other way around. The problem with this approach is that the conclusions inferred from the second step are conditional upon the outcome of the first step. Consequently, dependencies between spatial structures and other thematic aspects of interaction systems (transport mode, transportation service providers, and so on) are investigated only in one direction,

while controlling for the other. This calls for methods in which the two tasks interact with each other in a single process.

4.3.2   Self-Organizing Maps

The method of SOMs can be considered as a combination of data projection and data quantization. It is a special kind of competitive neural networks. The principle behind it is rather simple: neurons in the output layer compete with each other and the winner earns the right to represent the input data vector on the basis of some measure of dissimilarity in the attribute space (Kohonen, 2001). SOM allows the winner node, as well as the nodes in its neighbor, to learn the new input-node match and adapt so that each node gradually specializes to represent similar inputs. It preserves the natural order in the input attribute space of the data. As a result, data vectors with more similarity are close to each other on the feature map grid and essential relationships in the input data set can be visualized in a condensed form.

   The principles of the SOM method are not discussed here as Chapter 1 gives a detailed account of it. It suffices here to highlight the properties of SOM that make it well suited for SI data reduction. In comparison with traditional clustering methods, SOM has a big advantage in that it relies on the continuous 'learning' of all input data. This is a radical departure from the *k*-means method, which uses only the nearest distance for clustering. In addition, SOM offers some powerful visualization tools for data exploration (Vesanto, 1999). In fact, the regularly shaped projection grid greatly facilitates the comparison of different visualization forms. It provides a valid platform for user interaction and control, which is an essential part of visual exploration in large complex spatial data. Lastly, probably the most compelling argument in favor of SOM for exploring SI systems is that it is capable of both data projection and data quantization *simultaneously*. Many SI systems, such as the domestic air traffic system, are often constituted in *large databases* with *high thematic dimensionality* and thus both data quantization and data projection are essential. This enables the exploration of all possible structures, without any preconceived views on influential relationships. As shown in Figure 4.1, in addition to the geographic origins and destinations, many SI data are complicated by the third dimension, which could include flow type, transport mode, transport time, or any other quality or quantity of SI. In many cases, we are in fact more interested in how groups of origins, groups of destinations and various interaction attributes work together to shape SI systems in the geographic space over time.

4.3.3   Visual Data Mining

VDM is a collection of interactive reflective methods that support the exploration of data sets by dynamically adjusting parameters to display how they affect the information being presented (Ferreira de Oliveira and Levkowitz, 2003). This emerging area in exploratory and intelligent data analysis and mining draws on concepts from visualization and data mining. The effectiveness of visualization techniques in displaying geographic data provides us a solid base to integrate geovisualization methods with computational methods and data mining techniques.

   With SOM as the core data mining engine, an integrated approach is developed for analyzing large SI databases. Following the framework suggested by MacEachren *et al*. (1999), a prototype interactive VDM environment is developed using ESRI's ArcObjects and Visual Basic, in which different visualization forms are integrated through interaction forms. Each visualization form can be seen as a different view of the actual SI data. For instance, a 'cartographic map' displays the geographic distributions of flows among sets of origins and destinations, while a 'SOM component plane' identifies the properties of clustered structures in the input attribute space by means of a two-dimensional grid where each node takes the value of a selected component of the prototype data. By linking the component plane and the cartographic map together, it is possible to examine how the clustered structures detected by SOM are geographically defined. Besides linking, some other interaction forms are implemented in this VDM environment, including assignment, color-map manipulation, focusing, and brushing (Figure 4.2).

   A loosely coupled integration strategy is used, in which SOM training is conducted using SOM Toolbox 2 (http://www.cis.hut.fi/projects/somtoolbox/) in UNIX workstation.
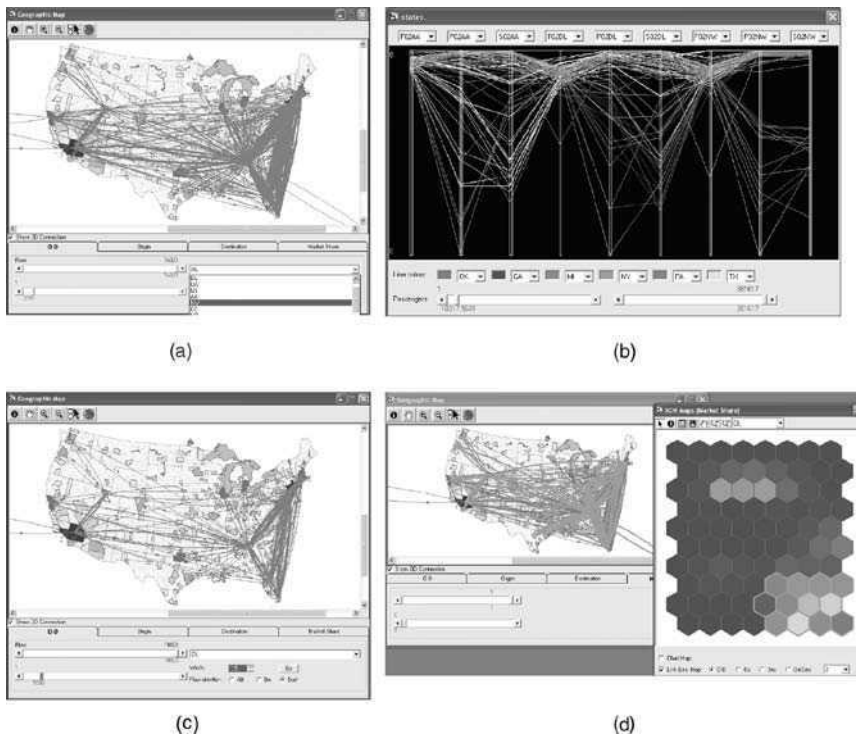
Figure 4.2   *Selected user interaction forms in the VDM environment. (a) Assignment: allows the analyst to visualize the passenger volume of different airlines. (b) Color manipulation on parallel coordinates plots: colors are used to draw the markets that originate in different geographic regions. (c) Focusing: two scroll bars are used so that the upper bound and lower bound of passenger flow can be changed dynamically in display. (d) Linking and brushing: highlighted in the geographic map are the markets represented by the neurons selected in the component plane (See Colour Plate 10)*

SOM Toolbox 2 is freeware, developed in Matlab 5 (http://www.mathworks.com) by a research group with the Laboratory of Computer and Information Science at the Helsinki University of Technology, Finland. In a second stage, training results are brought into the VDM environment for post-training evaluation and interpretation. In addition to SOM maps and geographic maps, the prototype VDM environment also includes other common techniques of exploratory data analysis (Fotheringham *et al.*, 2000) to assist the understanding of original data and the evaluation of the results, such as scatter plot, star coordinate plot, and parallel coordinate plot.

Figure 4.3 illustrates the linkage of various visualization forms in the prototype VDM environment. The linkage of multiple SOM component planes is straightforward as a
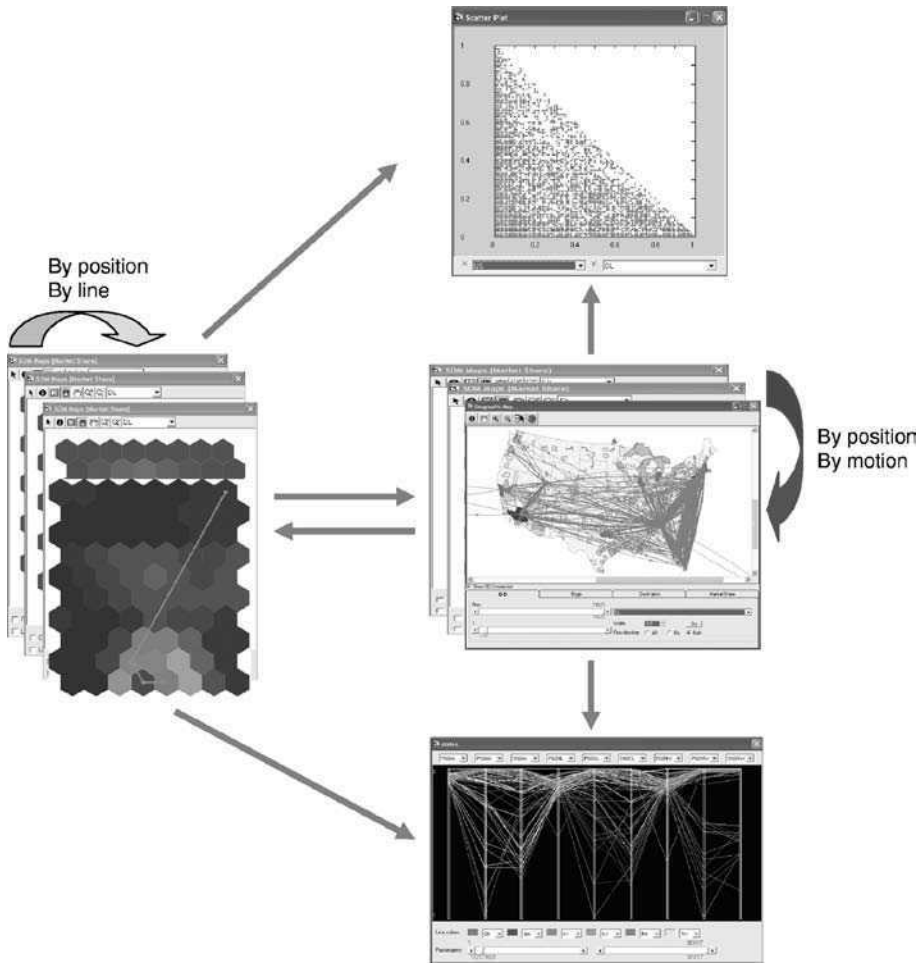
Figure 4.3 *Framework of the integrated VDM environment. The linkages among different visualization forms can be implemented mainly through four ways: by position (the position of data items remains fixed across visualization forms); by color (the same color is used for the same group of data items); by line (the same data items are connected by explicit lines); by motion (groups of data items are displayed one after another using animation) (See Colour Plate 11)*

result of the identical positions of the nodes across all these planes. Likewise, various geographic maps are connected via position as well. In addition, an animation technique is also used which allows the analyst to examine the step-by-step changes of a flow map by automatically incrementing either upper or lower bound on the interaction attribute. Finally, dynamic linking is implemented among different types of visualization forms in this prototype VDM environment.

4.4 CASE STUDY: US DOMESTIC AIRLINE MARKET

4.4.1 The State of the Industry

Ever since the enactment of the Airline Deregulation Act of 1978, the US domestic airline industry has changed rapidly from a regulated and stable system to an unregulated, turbulent, and dynamic situation. Nationwide, the market has become highly concentrated as the number of major airlines has declined. Although average airfare has dropped, consumers in markets with high concentration have actually experienced an increase in airfare and a loss of service due to lack of competition. This is especially the case for small and middle size communities in the Southeast and the upper Midwest (Goetz and Sutton, 1997). The recent trend suggests that further consolidation in the US domestic airline industry is still underway.

Geographers have shown considerable interest for the geography of the US airline industry post deregulation. Most previous studies were conducted at the point market level; air travel was aggregated to cities by singling out either the origin or the destination (Goetz and Sutton, 1997; Vowles, 2000). Results at this level can be very misleading because they smooth the differences between city-to-city markets. Therefore travel itineraries[1] in terms of pairs of origin and destination cities are more appropriate for studying essential geographic relationships in the US domestic airline industry, since they represent the scale of the actual decision making process. After all, city-pair markets are the specific products that consumers purchase. Patterns and dynamics at this level are closer to the reality of the underlying processes of demand and supply if one seeks to understand the underlying processes of the formation of spatial interaction.

4.4.2 Data and their Representation

The VDM approach is applied to an SI database, called the Airline Origin and Destination Survey Market Database (DB1BMarket), which is derived from a 10 % sample of airline tickets sold quarterly by each large certified air carrier operating scheduled passenger service in the US (GPO, 2003). Attributes in this database include number of passengers, airfare, and miles flown. It is the most significant data source for analyzing structure in the US domestic airline industry. In the next section, data representation of spatial interaction is briefly discussed. To illustrate the effectiveness of the integrated

---

[1] This notion is different from the route, which is a point-to-point segment of the air transportation system, without any transfer airport in-between.

SOM-VDM approach to knowledge discovery, findings from 2002 data are presented in Section 4.4.3.

SI data are most commonly represented by a matrix construct called an origin-destination (O-D) matrix [Figure 4.1(a)] where each row is indexed by an origin and each column by a destination (Black, 2003). Each element of the O-D matrix corresponds to a certain measure of interaction intensity or quality (for example, the flow volume of a certain type) taking place between the respective origin and destination. The diagonal elements can be used to represent intra-regional interactions (flows within a region) or simply be left out if intra-regional relationships are ignored or nonexistent, as is the case with air travel. Basic O-D matrices can also be transformed into a so-called *dyadic* matrix [Figure 4.1(b)]. This representation is more table-like: each row corresponds to an O-D pair while each column represents a particular interaction attribute. A dyadic matrix is used in this research because multiple O-D matrices need to be handled at once.

Data in the 2002 DB1BMarket database are transformed into a dyadic matrix wherein each column containing either the share of the market held by a particular airline (in %) or the average market airfare charged by this airline (in current $). Airports located in the same metropolitan area are pooled together. Only markets between the 278 metropolitan areas in the contiguous US are considered in this study. A total of 34 certified air carriers operated inside the US in 2002; as a result, there are 68 data fields in the final input data table.

4.4.3   Results and Discussion

Intensive training and parameter testing were conducted to arrive at the results presented here. A two-stage training process is followed, namely, a rough training stage followed by a fine-tuning stage. Both training stages are based on a 10×8 hexagonal SOM lattice. In the rough training stage, the initial neighborhood radius is set to 3, the initial learning rate is 0.5, and training is conducted over two epochs; in the fine-tuning training stage, the initial neighborhood radius is reduced to 1, the initial learning rate is 0.05, and training extends over 20 epochs. In both stages, the final neighborhood radius of 1 is used, the learning rate function is linear, and similarity between data vectors is measured by the Euclidian distance. All input variables are normalized to a [0 1] interval. The reverse normalization procedure is also applied to the SOM prototype data so that the final results can reflect the actual scale of data variation.

A critical issue is whether SOM can pick out the structures embedded in the input SI data. SOM has quite a few visualization methods for detecting the clustered structures and revealing the overall data shape. Most notably, the very principle of SOM is based on the projection of each original data vector to a node that is its BMU. With this property, the structures identified in the output neurons can be assumed true to the original data as well. In SOM, the technique of 'distance matrix' can be used to visualize the patterns in SOM prototype data qualitatively. In principle, colors (or other visual variables) can be assigned to each node on a SOM feature map on the basis of a certain statistic of inter-node distances, e.g. minimum, median, or maximum

of the distances to its neighbors. In Figure 4.4(a) (U-matrix), the color of map units (hexagons) denotes the distance of SOM nodes to their respective neighbors. Red corresponds to the largest distance while blue represents the smallest distance. A cluster of similar markets can often be identified as an area with low distance values delineated by a border of nodes with high values (thus large distances between nodes). Several clusters can be recognized in Figure 4.4(a). Figure 4.4(c) and (d) (basic distance matrix) exhibit the same patterns as the U-matrix in (a), except that the mean distance between each node and its adjacent neighbors is used. They are also smoothed linearly in order to enhance visual effect (two- and three-dimensional surfaces). This spatial pattern is more formally revealed in Figure 4.4(b) by the k-means clustering solution computed on the SOM prototype data. Of all the possible numbers of clusters, the clustering with $k = 9$ is reported here since it results in the best Davies–Bouldin
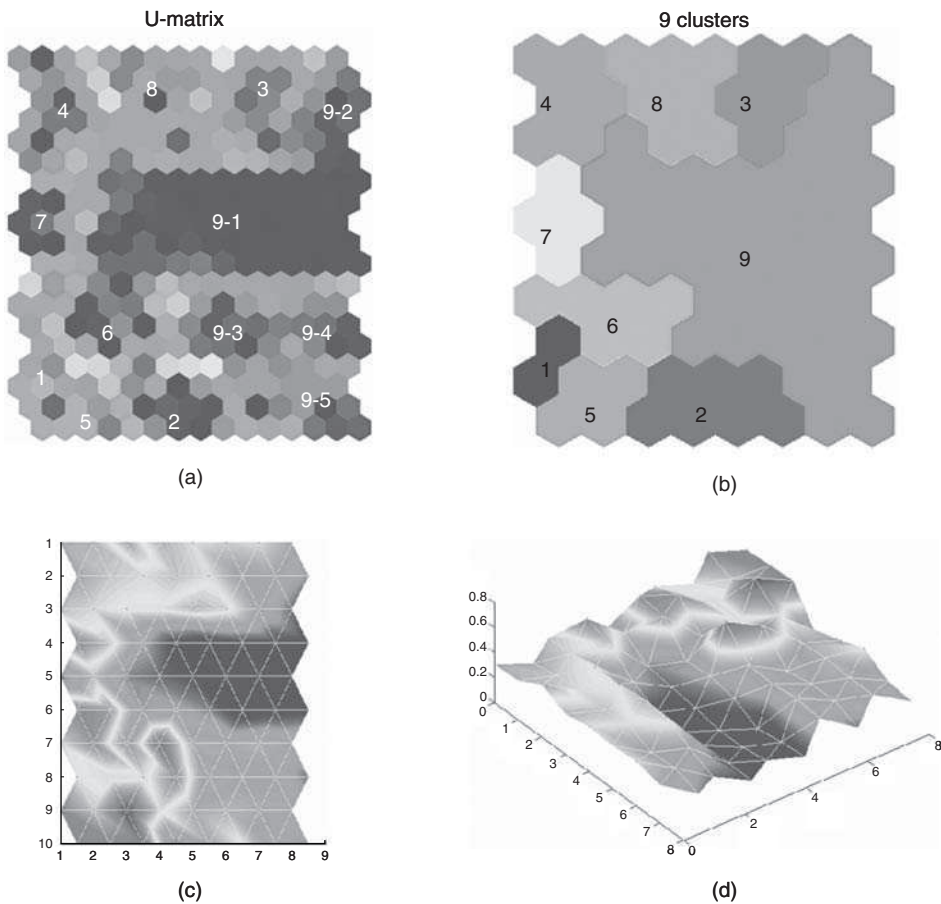
Figure 4.4 *Distance matrices and clusters based on 2002 market share information: (a) U-matrix; (b) clusters identified by k-means; (c) distance matrix (two-dimensional); (d) distance matrix (three-dimensional) (See Colour Plate 12)*

index (Davies and Bouldin, 1979). The Davies–Bouldin index is a validity index for evaluating clustering results. It is a function of the ratio of the sum of within-cluster difference to between-cluster separation. The lower the index value the better the clustering.

The SOM method extracts high-level structures marked in the clustering of similar prototype data of the feature map. Both the U-matrix and the basic distance matrices suggest that certain structures exist in the US domestic airline market. However, a more vital question is what factors can be attributed to each cluster. This can be answered by directly visualizing the distribution of the value and importance of each component with SOM component planes [Figure 4.5(a)]. SOM prototype data have the exact same number of components as the number of interaction attributes of each input case. By visually comparing distance matrices to component planes, we can find out which component or combination of components contributes the most to a particular cluster. For instance, cluster 2 identified by *k*-means is associated with high values of market share held by US Airways (US airline code). Hence, the nodes that form cluster 2 can be interpreted as the directional city-pair markets dominated by US Airways. Following this example, we can mark each cluster in Figure 4.4(b) by certain properties, that is, the components (airlines) that contribute to it more visibly. These properties are reported in Table 4.1.
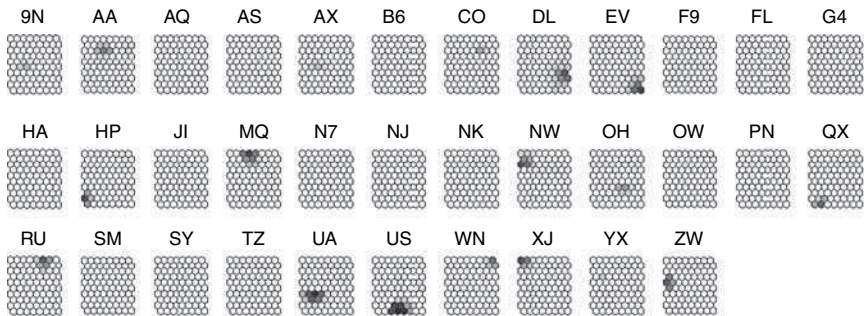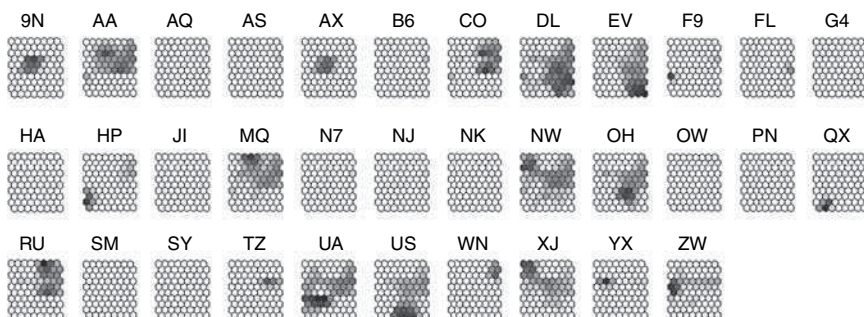
Figure 4.5 *SOM component planes (the darker the larger the value): (a) market share; (b) average airfare*

Table 4.1   *Node clusters based on 2002 market share information*

Cluster no.	Cluster property (airline)
1	America West (HP)
2	US Airways (US)
3	Continental Airlines (CO), Continental Express (RU)
4	Northwest (NW), Mesaba (XJ)
5	Horizon (QX)
6	United (UA)
7	Air Wisconsin (ZW)
8	American (AA), American Eagle (MQ)
9-1	No dominant airlines
9-2	Southwest Airlines (WN)
9-3	Comair (OH)
9-4	Delta Air Lines (DL)
9-5	Delta Air Lines (DL), Atlantic Southeast (EV)

Let us examine more closely the properties of some clusters appearing in Figure 4.4(b) and Table 4.1. Interestingly, cluster 8 is shared by two airlines, namely, American Airlines (AA) and American Eagle (MQ). This result should not be a surprise since American Eagle is a regional affiliate of American Airlines. The same occurs to cluster 3, in which Continental (CO) and its Continental Express (RU) subsidiary define most heavily, and cluster 4, where both Northwest (NW) and Mesaba (XJ), a regional affiliate of Northwest, have high market share. Cluster 7, representing the markets of Air Wisconsin (ZW, another subsidiary airline of Northwest), is just located below cluster 4 in the feature map. All these observations indicate that the nodes representing similar markets are indeed close to each other on the SOM feature map.

The U-matrix also recognizes more detailed structures than the specific 9-means solution can capture. Cluster 9 could be further divided into five sub-clusters. Cluster 9-1 denotes the markets without any dominant airlines. Cluster 9-2 represents markets dominated by Southwest (WN) while Comair (OH) has high values of market share in cluster 9-3. Clusters 9-4 and -5 are dominated by Delta (DL) and Atlantic Southeast (EV), a regional subsidiary of Delta.

Component planes also successfully capture the magnitude of market share and market scope of many small airlines, including Allegiant Air (G4), Frontier Airlines (F9), Airtran Airways (FL), Midway (JI), National (N7), Vanguard (NJ), Spirit (NK), Executive (OH), Sun Country (SY), Mesaba (XJ), and Midwest (YX). They operate on few markets and thus, fail to contribute in a meaningful way to any essential relationship embedded in the SI data. The market share component planes of Aloha Airlines (AQ), Alaska Airlines (AS), and Hawaii Airlines (HA) also exhibit low values. This is understandable given that this case study is confined to the markets within the contiguous states, while these airlines mostly serve the markets in and out of Alaska or Hawaii.

Component planes can also be used to examine the associations among components (airlines in terms of both market share and average airfare). At first glance, the overall patterns in Figure 4.5(a) (market share) and Figure 4.5(b) (airfare) may seem quite different. However, when the attention is on the most meaningful information, close correspondence exists. Invariably, nodes with high to moderately high market share also

exhibit rather high fares. This suggests that airlines tend to set high airfare in the markets where they have dominance. This is consistent with many previous studies and points to the considerable impact that the level of competition has on the practice of setting fare in the US domestic markets.

Another interesting observation derived from the analysis of component planes is that airlines can still charge high fares in some markets even if they have relatively low market share. For instance, American (AA) has two areas of high airfare as seen in the component planes [Figure 4.5(b)]: one corresponds to markets that it has high market share; in the second area American does not but Delta (DL) happens to have high market share. Collectively American and Delta may have total control of these markets and thus, both can set high fares. Also notice that Atlantic Southeast (EV) and Air Wisconsin (ZW, a regional subsidiary of Northwest) charge their customers highest airfare since they receive relatively very low competition from other airlines in their respective markets. The markets within cluster 9-1 have the lowest airfare due to the lack of dominance of any particular airlines. This further proves that high level of competition leads to low pricing. At last, travelers obviously benefit a lot in markets served by low-fare airlines, such as Southwest (WN), JetBlue (B6), Vanguard (NJ), Spirit Airlines (NK), and Airtran Airways (FL), even when these low-fare airlines have relatively low market share. Thanks to the competition of low-fare airlines, even full-service airlines are forced to set lower fare in these markets too. The city-pair markets projected at the upper right-hand corner of component planes are typical examples of the so-called 'Southwest effect'.

To further illustrate the effectiveness of the VDM approach, let us take a close look at all city-pair markets originating in the Buffalo metropolitan area. In Figure 4.6, these markets are mapped to match with SOM prototype data. At first glance, no structure is apparent because these markets are scattered all over the entire SOM feature map without any regularity [Figure 4.6(a)], a reflection of the fact that Buffalo does not operate as a hub for any airline. Once we add to the SOM map a locational descriptor of the destinations matched to each output node, i.e. state [Figure 4.6(b)], and pinpoint airline dominance in component planes [Figure 4.6(c)], patterns in terms of destinations and airlines emerge. City-pair markets from the Buffalo metropolitan area to Southeastern states such as Georgia (GA) and Florida (FL) are controlled by Delta (DL); on the East coast, by US Airways (US); to Southwestern states such as Texas (TX), by American (AA) and Southwest (WN); to the upper Midwest by both Northwest (NW) and United (UA); to the Pacific Northwest, by United (UA). This structure is also confirmed in the geographic map of dominant flows [Figure 4.6(d)].

Finally, a question of great pertinence to spatial scientists is how structures embedded in the attribute space are distributed in the actual geographic space. In the prototype VDM environment, flow maps and SOM feature maps can be directly connected via dynamic linking and brushing. Figure 4.7 illustrates this functionality. In Figure 4.7(a) nodes with high values of JetBlue (B6) market share are highlighted on the component plane. The city-pair markets assigned to these nodes are highlighted on the map of market served by JetBlue [Figure 4.7(b)]. These markets include the city-pair markets centred on the area where JetBlue operates most, that is, the New York City region. This example
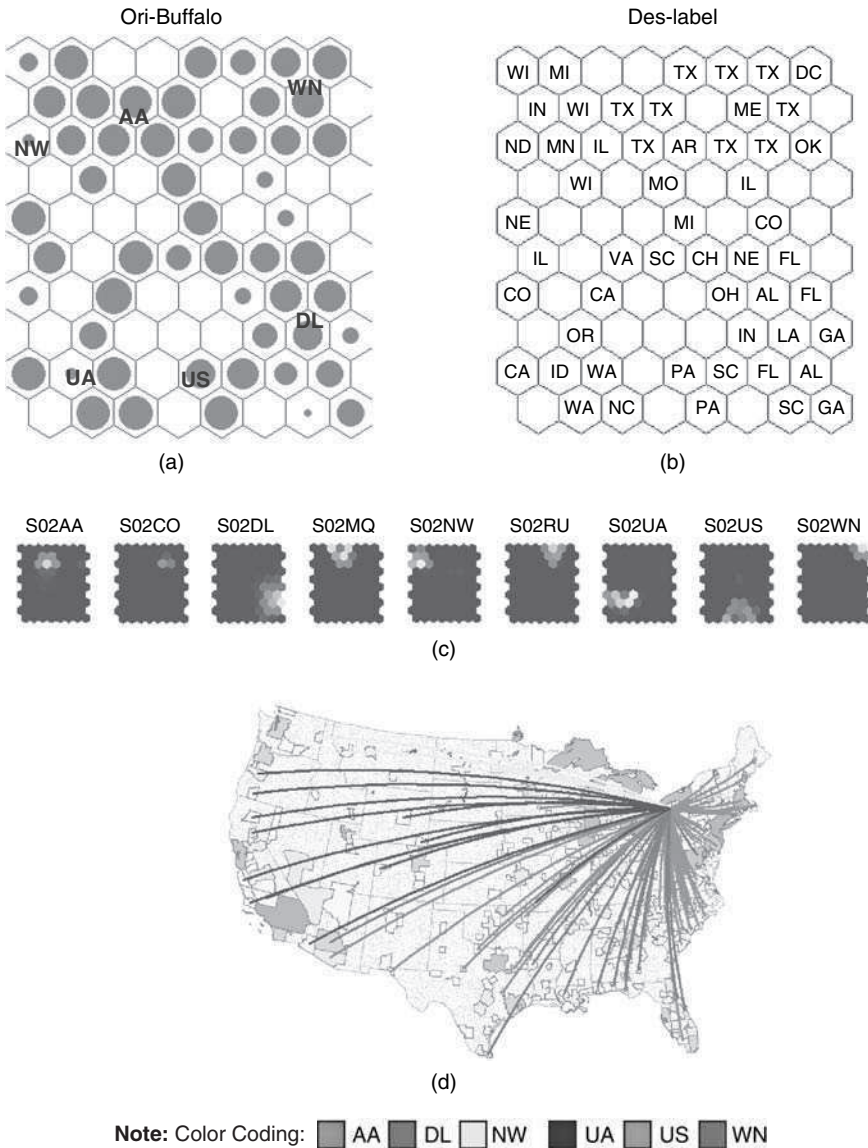
Ori-Buffalo

Des-label

(a)

(b)

S02AA   S02CO   S02DL   S02MQ   S02NW   S02RU   S02UA   S02US   S02WN

(c)

(d)

Note: Color Coding: ☐ AA ☐ DL ☐ NW ■ UA ☐ US ☐ WN

Figure 4.6 *Markets originating in the Buffalo metropolitan area: (a) hit counts; (b) most frequent destination state; (c) selected SOM component planes; (d) markets with airline market share > 50% (See Colour Plate 13)*

underscores the ability of SOM in capturing the finer structures embedded in SI data in addition to essential relationships. This ability afforded by the VDM environment to link various visualization forms, here flow maps and SOM feature maps, provide us a rather flexible but powerful way to evaluate and make sense of the results.

(a)   (b)
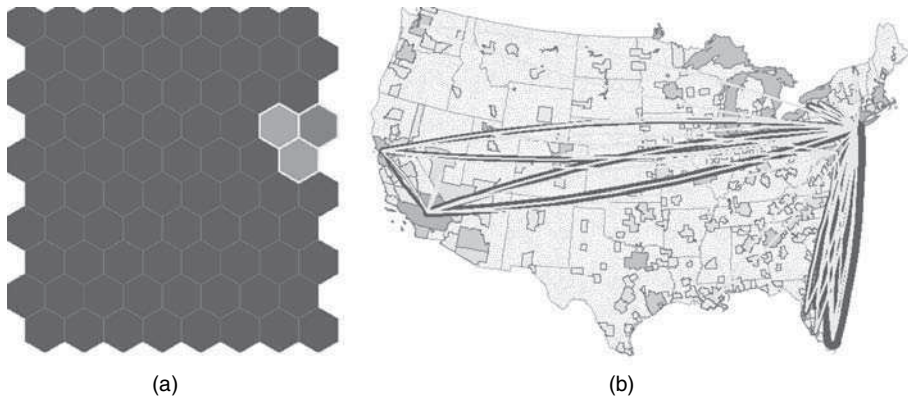
Figure 4.7   *Markets of JetBlue identified by SOM: (a) market share component plane; (b) flow map (passengers $\geq$ 20 000) (See Colour Plate 14)*

4.5   CONCLUSIONS

When dealing with large SI databases with high dimensionality, it is often a prerequisite to reduce the data complexity before any effective data analysis can be conducted. The SOM method has the advantage of collapsing origin-destination information and interaction attribute information simultaneously to retrieve essential and other well-defined relationships embedded in the data. The consistent geometric properties of the output of SOM training provide for powerful visualization tools for visual data exploration, validation and evaluation. In addition, native SOM visualization forms can advantageously be integrated in an interactive VDM environment, as illustrated in this study.

Findings from the case study of US domestic air travel suggest that SOM is capable of identifying clustered structures in large SI data sets. Through SOM training, we are able to obtain an overview of a large data set. In the US domestic airline industry, airline carriers tend to serve markets in different US regions in order to reduce direct competition. In the markets where the level of competition is high, airfare tends to stay low. The SOM method is also instrumental in uncovering other interesting and less obvious relationships. For instance, our finding indicates that if in a market where a major airline serves and its competitor is another major airline, the airfare usually remains at a relatively high level, while if the competition is from a low-fare carrier, the price is often driven down significantly. Many of these findings are confirmed by a number of reports published by US General Accounting Office (GAO). In summary, all of these suggest that SOM is capable of locating rather localized, focused, or partial structures as well as essential relationships carried by the entire dataset.

Finally, our research examines the roles of visualization in the exploration of spatial interaction data. Visualization is commonly realized by using visual variables to describe the different kinds of information in the data. Usually only a limited number of visual

variables can be applied to a single visualization form; otherwise it will become too complex to comprehend. In many cases, a complex data set often contains so many clues that it is unlikely to show all of them in one single visualization method. The idea is, in addition to reducing data complexity, to adopt multiple visualization forms so that the number of visual variables can be multiplied and information in display can be greatly increased. In our prototype VDM environment, a variety of visualization methods are implemented and linked together. As a result, various aspects of spatial interaction can be cross-examined intensively in order to fully understand the data. The improved interactivity of visualization is useful since the purpose of data exploration is, after all, to suggest hypothesis. For instance, linking SOM feature maps with geographic flow maps simply offers us a way to take a look at how the structures revealed by SOM are geographically defined. The findings in the case study indicate that it could be a possible way to incorporate geographic properties with the existing data mining tools without adding them directly into the algorithms.

ACKNOWLEDGEMENTS

REFERENCES

Bailey, T.C. and A.C. Gatrell. (1995). *Interactive Spatial Data Analysis*. Harlow: Longman.

Berglund, S. and A. Karlström. (1999). Identifying Local Spatial Association in Flow Data. *Journal of Geographical Systems* 1: 219–236.

Berry, B.J.L. (1962). *Structural Components of Changing Transportation Flow Networks*. Fort Eustis, VA: US Army Transportation Research Command.

Berry, B.J.L. (1966). *Essays on Commodity Flows and the Spatial Structure of the Indian Economy*. Department of Geography, Research Paper No. 111. Chicago: University of Chicago Press.

Black, W.R. (1973). Toward a Factorial Ecology of Flows. *Economic Geography* 49: 59–67.

Black, W.R. (2003). *Transportation. A Geographical Analysis*. New York: Guilford Press.

Chicago Area Transportation Study (1959). *Final Report*, Vol. I, Survey Findings, Chicago, IL, pp. 96–99.

Davies, D.L. and D.W. Bouldin. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(4): 224–227.

Ferreira de Oliveira, M.C. and H. Levkowitz. (2003). From Visual Data Exploration to Visual Data Mining: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 9(3): 378–394.

Fotheringham, A.S. (1997). Trends in Quantitative Methods I. Stressing the Local. *Progress in Human Geography* 21: 88–96.

Fotheringham, A.S. (1999). Trends in Quantitative Methods III. Stressing the Visual. *Progress in Human Geography* 23: 597–606.

Fotheringham, A.S. and M. O'Kelly. (1989). *Spatial Interaction Models: Formulations and Applications*. Boston: Kluwer.

Fotheringham, A.S., C. Brunsdon and M. Charlton. (2000). *Quantitative Geography. Perspectives on Spatial Data Analysis*. London: Sage Publications.

Frawley, W.J., G. Piatetsky-Shapiro and C.J. Matheus. (1991). Knowledge Discovery in Databases: An Overview. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley (eds), pp. 1–27. Menlo Park, CA: AAAI/MIT Press.

Goetz, A.R. and C.J.Sutton. (1997). The Geography of Deregulation in the US Airline Industry. *Annals of the Association of American Geographers* 87: 238–263.

Gould, P. (1991). Dynamic Structures of Geographic Space. In *Collapsing Space and Time: Geographic Aspects of Communications and Information*, S.D. Brunn and T.R. Leinbach (eds), pp. 3–30. New York: HarperCollins.

GPO. (2003). *Code of Federal Regulations. Title 14 – Aeronautics and Space, Part 241*. Washington, DC: Government Printing Office.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edition. Berlin: Springer.

Lu, Y. and J.-C. Thill. (2003). Assessing the Cluster Correspondence between Paired Point Locations. *Geographical Analysis* 35(4): 290–309.

Marble, D.F., Z. Gou, and L. Liu. (1995). Visualization and Exploratory Data Analysis of Interregional Flows. In *Proceedings of the 1995 Conference on Geographic Information Systems in Transportation (GIS-T)*, pp. 128–136. American Association of State Highway and Transportation Officials (AASHTO): Washington, DC.

Marble, D.F., Z. Gou, L. Liu, and J. Sauders. (1997). Recent Advances in the Exploratory Analysis of Interregional Flows in Space and Time. In *Innovations in GIS 4*, Z. Kemp (ed.), pp. 75–88. London: Taylor & Francis.

MacEachren, A. and M.-J. Kraak. (1997). Exploratory Cartographic Visualization: Advancing the Agenda. *Computers & Geosciences* 23(4): 335–343.

MacEachren, A., M. Wachowicz, D. Haug, R. Edsall and R. Masters. (1999). Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographic Visualization with Knowledge Discovery in Database Methods. *International Journal of Geographic Information Science* 13(4): 311–334.

Miller, H.J. and J. Han (eds). (2001). Geographic Data Mining and Knowledge Discovery: An Overview. In *Geographic Data Mining and Knowledge Discovery*, pp. 3–32. London: Taylor & Francis.

Openshaw, S. (1999). Geographical Data Mining: Key Design Issues. In *Proceedings of GeoComputation 99*. http://www.geovista.psu.edu/sites/geocomp99/Gc99/051/gc_051.htm. Accessed on 06/02/04.

Openshaw, S. and R. Abrahart. (2000). *GeoComputation*. New York: Taylor & Francis.

Roy, J.R. and J.-C. Thill. (2004). Spatial Interaction Modeling. *Papers in Regional Science* 83(1): 339–361.

Smith, R.H.T. (1970). Concepts and Methods in Commodity Flow Analysis. *Economic Geography* 46: 404–416.

Tobler, W. (1976). Spatial Interaction Patterns. *Journal of Environmental Systems* 6: 271–301.

Tobler, W. (1978). Migration Fields. In *Population Mobility and Residential Change*, W. A. V. Clark and E. Moore (eds), pp. 215–232. Studies in Geography No. 25, Department of Geography, Northwestern University, Evanston, IL.

Tobler, W. (1981). A Model of Geographic Movement. *Geographical Analysis* 13: 1–20.

Tobler, W. (1987). Experiments in Migration Mapping by Computer. *The American Cartographer* 14(2): 155–163.

Ullman, E.L. (1957) *American Commodity Flows: A Geographical Interpretation of Rail and Water Traffic Based on Principles of Spatial Interchange*. Seattle, WA: University of Washington Press.

Vesanto, J. (1999). SOM-Based Data Visualization Methods. *Intelligent Data Analysis* 3(2): 111–126.

Vowles, T.M. (2000). The Geographic Effects of US Airline Alliances. *Journal of Transport Geography* 8: 277–284.

Wachowicz, M. (2001). GeoInsight: An Approach for Developing a Knowledge Construction Process Based on the Integration of GVis and KDD Methods. In *Geographic Data Mining and Knowledge Discovery*, H.J. Miller and J. Han (eds), pp. 239–259. London: Taylor & Francis.

This page intentionally left blank

5

Detecting Geographic Associations in English Dialect Features in North America within a Visual Data Mining Environment Integrating Self-Organizing Maps

Jean-Claude Thill[1], William A. Kretzschmar Jr[2], Irene Casas[1] and Xiaobai Yao[3]

[1] *Department of Geography and Earth Sciences and Center for Applied Geographic Information Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA*

[2] *Department of English, University of Georgia, Athens, GA 30602, USA*

[3] *Department of Geography, University of Georgia, Athens, GA 30602, USA*

5.1   INTRODUCTION

Until recently, the study of variation within languages, under the general rubric of dialectology, has been carried out using qualitative research approaches. The main tool of dialectologists, the geographic concept of isoglosses (boundary lines between regions in which different variants of a particular linguistic feature are found, such as *pail* in one area and *bucket* in another), has primarily been applied through qualitative and subjective assessment of the available data (Kretzschmar, 1992; Schneider, 1988). The digital storage of linguistic databases such as the Linguistic Atlas of Middle and South Atlantic States (LAMSAS) has enabled a more comprehensive analysis of lexical and

---

pronunciation variants in large population groups [Kretzschmar and Schneider, 1996; see Kretzschmar (1996a) for development of different quantitative analyses].

Linguistic research conducted over several decades indicates that American English has a strong geographic component (Kurath, 1949; Labov *et al.*, 2006). Field work as well as carefully crafted statistical tests of spatial autocorrelation point in that direction. Geography, significantly location, is the primary variable to be considered, because people tend to talk like the people they talk with: linguistic features must be *available* to people before they can decide to use them [see, e.g., the ideas of Saussure, as reported in Kretzschmar (1998)]. People can only talk like people with whom they have had some contact, whether personally or, much less importantly, through some passive medium like reading or listening to the radio or watching television and movies [see Chambers (1998) on the 'myth' that media such as television strongly affect people's speech]. Social variables such as education, social circumstances, age, and sex are then influential in guiding the decisions that people make about the use of the linguistic features available to them.

While it may be tempting for geographers to do so, we should not be trying to find well-defined 'regional dialects' of English, i.e. a complete set of linguistic features which is held in common by most or all of the speakers from some geographical area. Regional differences in speech can be perceived by most people, but previous research by Heeringa and Nerbonne (2001), Kretzschmar (2003), and others contends that such neatly compartmentalized sets do not exist. The geographic expression of linguistic variations is far more complex than can be captured by crisp regional boundaries. The regional differences that we perceive have so far not been properly characterized in geographical terms, nor correlated with social, demographic, and economic profiles of individuals.

Previous methods of quantitative spatial analysis have not extracted meaningful relationships from the LAMSAS databases for a number of reasons: data are often sparse, have a skewed distribution and are highly multidimensional (Kretzschmar, 1996a). This chapter presents a new approach to the spatio-linguistic analysis of the variations of word usage and pronunciation in the North American Middle and South Atlantic region that integrates Kohonen's (2001) data reduction technique of self-organizing maps (SOMs) and tools of visual data mining in a seamless environment. This approach to knowledge discovery is well suited to the properties of large multidimensional linguistic databases and opens new horizons for the reassessment of concepts and theories that were devised decades ago on the basis of ad hoc qualitative research techniques.

Section 5.2 gives a brief overview of the study of dialectal variations in American English. This section also provides a background description of the nature of LAMSAS data sets, which are the primary source of data in this research, and reviews earlier efforts at the quantitative analysis of LAMSAS data. Section 5.3 presents the knowledge discovery environment that was designed to reveal meaningful relationships from the LAMSAS data sets, using the computational data mining capability of the SOM algorithm and other interactive visualization tools. The functionality of the system is illustrated in Section 5.4 on a small set of lexical variations commonly expected to discriminate major dialectal variations along the eastern seaboard. Conclusions are presented in Section 5.5.

5.2 DIALECTOLOGY AND LAMSAS

The American Linguistic Atlas Project (ALAP), sponsored by the American Council of Learned Societies, was begun in 1929 under the direction of Hans Kurath in New England (Kurath, 1939–43). ALAP was conceived as a large-scale survey about the words and pronunciation of everyday American English, and has been extended across most of the United States with ALAP survey methods. ALAP interviews were conducted with speakers of different ages and social circumstances, in cities as well as rural areas, including extensive interviews with African Americans, so the ALAP surveys include a wide spectrum of American speech. In creating this data set, ALAP adopted an integrated cultural approach and divided the American English speakers into three cultural 'types': Type I, 'folk' speakers, who were old, uneducated, and unconnected with community affairs; Type II, 'common' speakers, who were younger, better educated, and with more connections in the community; and Type III, 'cultivated' speakers, the best educated, well-connected and high-culturally aware. The primary analytical tools in the main works to exploit these data (Atwood, 1953; Kurath, 1949; Kurath and McDavid, 1961) were maps based on isoglosses, boundaries of use by ALAP informants of individual linguistic features. Weighting adverbs such as 'regularly', 'frequently', or 'occasionally' were employed with reference to the intensity of the use of a feature, but no quantifying statements were made.

The LAMSAS is the largest single survey of regional and social differences in spoken American English (Kretzschmar *et al.*, 1993). Of all the existing regional atlases, so far it is the only one created as a geographic data library and for which a substantial portion of linguistic attributes and features have been coded and stored electronically. LAMSAS consists of interviews, transcribed in fine phonetic notation, with 1162 informants from 483 communities within a region that stretches from New York State south to Georgia and northern Florida, from the eastern coastline as far west as the borders of Ohio and Kentucky. Within the communities two speakers were normally selected as representative of the community because of life-long residence there, one a member of the oldest living generation with little education or compensating experience, one younger and better educated with a less insular outlook. Interviews targeted 800 words or short phrases designed to reveal regional and social differences in everyday vocabulary, grammar, and pronunciation, through indirect elicitation of responses. Multiple responses from each speaker were permitted for any given target. Field work was conducted from 1933 to 1974, but was largely complete by 1949. The database of communities and informants has been geo-coded and can be queried interactively through a map interface (http://us.english.uga.edu/lamsas/). Each of the questions available in digital form is independently valuable as a research target since each question sought a particular variable pronunciation, lexeme, or grammatical point; all questions offer pronunciation evidence, since responses were recorded in phonetic notation.

While traditional dialectology was descriptive and qualitative, newer methods embrace quantitative analysis in order to comprehend the complex multidimensionality of language variation. Statistical and computational data reduction techniques not only shed better light on classical linguistic problems, but they also 'suggest avenues for exploring the question at more abstract levels, and perhaps for seeking the determinants of variation' (Nerbonne and Kretzschmar, 2003, p. 248), which

is the argument put forth in this chapter. Over the past few years, LAMSAS has been at the forefront of the computational movement in dialectology, having been subjected to a variety of inferential statistics and analytical procedures of spatial analysis in particular (Kretzschmar, 1996a; Kretzschmar and Lee, 1993; Kretzschmar and Light, 1996; Kretzschmar and Schneider, 1996).

Common multivariate statistics, including discriminant analysis, logistic, and log-linear procedures, have been applied with rather limited success. All yield results, often indicating interactions between variables (which one might expect from the plurality of significant univariate tests), but the reliability of the results has been in question. Some tests have proven sensitive to skewed data sets–and the majority of features in the LAMSAS data sets occur in skewed proportions. A further problem is that the number of cells for multivariate analysis by all categories of interest exceeds the number that can reasonably be attempted given even the large size of the LAMSAS data set. Some tests have no adequate measures of reliability.

However, spatial-analytic techniques have been applied with great success to LAMSAS data. Joint-count statistics of given lexical variants were statistically significant in about three-quarters of normally distributed variants, those elicited from between 20 % and 80 % of all communities (Kretzschmar and Lee, 1993). Further statistical evidence that all speech has a strong geographic component was obtained by quadrat analysis (Kretzschmar and Lee, 1993) and density estimation techniques (Kretzschmar and Light, 1996). In the latter publication, probability maps were created for all lexical variants available. As with the multiple comparison techniques, density plots generally correspond to isoglosses posited by traditional, subjective methods, but always reveal more detailed insights than previously available. Nerbonne and Kleiweg (2003) used multidimensional distances reflecting dissimilarities between informants on the basis of lexical uses. Hierarchical clustering of informants according to Ward's variance minimization criterion produced dialectal areas that mirror Kurath's (1949) qualitative geographic partitioning of the eastern seaboard into three main dialect areas. However, results of the INDSCAL variant of multidimensional scaling applied to the same data pointed out that boundaries between adjacent dialect areas are rather soft, instead of crisp (Heeringa and Nerbonne, 2001). These findings indicate that language variation across the region forms a continuum, instead of the sharply bounded set of internally coherent dialect regions predicted by traditional linguistic theory and suggested by the subjective qualitative analyses of Kurath and other traditional dialectologists.

5.3   DIALECT KNOWLEDGE DISCOVERY SYSTEM

5.3.1   System Design and Implementation

We propose an interactive system of exploratory spatial data analysis (ESDA) to detect geographic and socio-economic associations in English dialect features. The system seamlessly integrates multiple functions of dialect knowledge discovery in the linguistic space, and seeks relationships between significant dialectal structures in the linguistic space and the distributions in the geographic space; socio-demographic characteristics of informants can then be associated with distributions in geographic space. The system is

implemented in the Windows operating system as an application developed using ESRI's MapObjects, Visual Basic, and C++.

The core data mining engine consists of Kohonen's (2001) SOM algorithm. SOM serves the purpose of reducing the dimensionality of multidimensional linguistic data sets, identifying latent organization rules, and classifying surveyed informants into larger features exhibiting similar linguistic features. The computational algorithm of SOM is integrated with different visualization forms to involve high-level human intelligence and knowledge at two levels of the process of knowledge discovery. First, visual methods participate in the verification and the evaluation of significant structures embedded in the linguistic space as revealed by SOM's data reduction algorithm. Second, they enable recognition of relationships between linguistic 'meta'-structures and patterns displayed in the geographic and socio-demographic spaces. The visual method of exploration constitutes a Visual Data Mining (VDM) environment that interactively supports the process of knowledge discovery.

The hybrid SOM-VDM knowledge discovery environment is designed to answer a number of specific questions on the linguistic and dialectologic reality in the Middle and South Atlantic States, including:

1. How are linguistic features (whether words, grammatical constructions, pronunciations, or combinations of the latter) distributed over geographic areas, if not in relatively uniform patterns of complementary distribution (e.g. *pail* in one area, and *bucket* in another)?
2. Since individual linguistic features can be shown to have specific distributions in geographic space, are there some relatively small groups of linguistic features that can be perceived as most 'salient' in our identification of regional differences?
3. What is the congruence between isoglosses posited by traditional, subjective methods for simple linguistic features (lexical, pronunciation, or grammatical) and the 'fuzzy' multidimensional clusters derived from SOMs?

Finally, the function of the VDM to permit association of socio-demographic information with the results of application of the SOM algorithm may lead to address one further research question: What is the interaction of social variables such as sex, age, ethnicity, occupation, and education, with geographic variables?

5.3.2 LAMSAS Data Mining

Word usage and pronunciation vary within all known languages spoken by living human populations. All linguistic representations of even the most simple aspects of daily life, such as the item in LAMSAS to elicit the name of the utensil in which one commonly carries water from a well (i.e. pail or bucket), are characterized by specific lexical variants, pronunciation variants, and various ways of participation in grammatical constructions. The combinatorics and multidimensionality of patterns embedded in language data have been a significant obstacle to the implementation of analytical techniques in dialectology. Lexical data from large surveys such as the LAMSAS also present various other challenges to conventional quantitative techniques that stem from survey design considerations. The data are multiple response data, so for example some speaker might have said

both *skeeter hawk* and *snake doctor*; thus the same speaker can be marked as positive in more than one of the lexical files for any item. The data are clearly not normally distributed. Furthermore, data sets are inherently sparse because of the large number of lexical variations per word and the selectivity exhibited by informants in their use of lexical variations [for distributional properties of such language data, see Kretzschmar and Tamasi (2003)].

The LAMSAS survey design emphasized the indirect elicitation of responses from informants in order to reduce the formality of the interview situation and increase the spontaneity of responses. However, as pointed out by Nerbonne and Kleiweg (2003), different field workers implemented the prescribed design in different ways at different times, in different locales. The rate of failure to elicit a response for the use of specific lexical variations strongly depends on the field workers' own practices. Hence, for each informant, there are three coding possibilities for each linguistic feature under analysis: a '0' if the speaker did not use the feature, a '1' if the speaker did use it, and '99' if the item was not covered with the speaker. The last option is missing data, which turns out to be a rather frequent and biased occurrence. The inherent properties of the LAMSAS data set partially or completely invalidates a number of conventional methods of data analysis. However, the SOM neural network has been shown to be robust while dealing with such data (Openshaw and Openshaw, 1997). The conceptual and technical foundations of the application of the SOM algorithm are presented in Chapter 1.

Aside from the contributions in this book, relatively few geographic applications of SOM have so far been reported in the literature. The SOM algorithm has successfully been used for the classification of remotely sensed data (Byrne *et al.*, 1994; Chen and Shrestha, 2000; Hara *et al.*, 1994; Villmann *et al.*, 2003). In all these works, SOM is used as an unsupervised classifier, working on the multi-spectral information in remotely sensed data. Openshaw and Wymer (1991) tested an application of the SOM algorithm against a *k*-means classification on census data in the UK. Outside the application of SOM to remotely sensed or census data, a handful of studies of geographic feature identification have been conducted with the SOM algorithm. An early case study by Kaski and Kohonen (1996) applied SOM to a data set of 39 welfare statistical indicators of countries. Himanen *et al.* (1998) explored the applicability of SOM in identifying daily travel patterns in a disaggregate travel diary data set. In 2001, Nerbonne and Heeringa (2001) analyzed similarities between Dutch dialects with the SOM algorithm based on Levenshtein distances, in which the authors interpreted proximity on the feature map in relation to geographic proximities of communities surveyed.

Application of the SOM algorithm in a data mining process involves multiple steps, which are depicted in Figure 5.1. Data, stored in a relational database to facilitate geovisualization, are passed on to the SOM application. Pre-processing of original data is necessary before cases are presented to the SOM algorithm for training. Two data configuration options are available to reflect alternate views on the meaning of missing information from a linguistic perspective and on the relative importance to impart to positive and negative answers to the use of each lexical variation. In the first pre-processing modality, a single input node is associated with each lexical variation: positive, negative, and missing answers are respectively recoded x, y, and z, where $0 \leq x \leq z \leq y$. This modality allows for flexibility in handling missing data. If it is believed that a missing answer for a certain lexical variation is likely to signify that it is not used, the analyst can elect to set $z = x$. As the degree of confidence in this position drops, z should
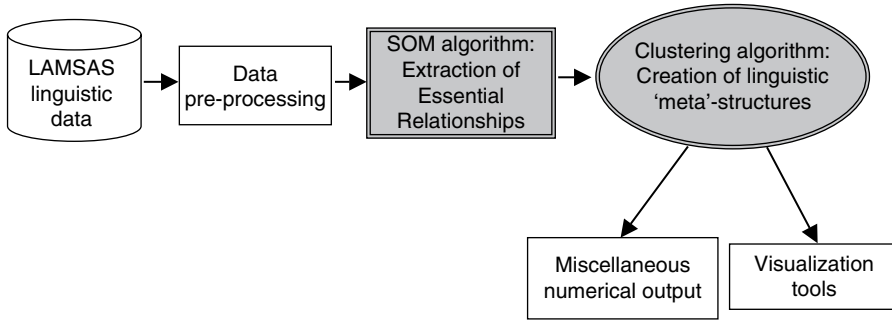
Figure 5.1 *SOM data mining process applied to LAMSAS data sets*

be increased. For $z = (x + y)/2$, it is assumed that there is not sufficient information to make any inference on missing answers; z values greater than $(x + y)/2$ are very improbable given the nature of the survey design.

Each lexical variation is decomposed into three input nodes under the second pre-processing modality, one node for each of the three possible answers (positive, negative, missing or unreported). A positive answer is coded '1' on the positive node, '0' on the other two; a negative answer is coded '1' on the negative node and '0' on the other two; finally, a missing answer is coded '0' on both the positive and negative nodes while some pre-specified value $0 \leq z \leq 1$ is assigned to the 'missing' node. The adjustable z value serves to reduce the contribution of missing answers to the training of the SOM network. At the limit, when $z = 0$, the network is trained only on positive and negative answers. This flexible coding has the advantage of not discarding an entire informant's data vector just because part of it was not recorded.

The Kohonen output layer generated by the SOM algorithm consists of a square lattice with variable numbers of rows and columns. Proximity between input and output nodes is measured by the Euclidean distance. The neighborhood function is taken to include neighbors in a square region centred on the winning node and shrinking linearly with the number of iterations in the training process. Weights to winning nodes and their neighbors are updated at each iteration with an error-adjustment coefficient:

$$0 < \eta(t) = \frac{e^{-d^2/0.15}}{4t + 1} < 1 \tag{5.1}$$

where t is time defined as current_iteration/total_iterations and d is the Euclidean distance between each node in the neighborhood and the input vector whose weights are to be updated. The number of epochs and initial search radius are inputs supplied by the analyst at the time of network training.

Once the SOM network has been trained and essential relationships in the linguistic input data vectors have been identified, output nodes are classified into n number of regions by means of a clustering algorithm using the centroid method (Sokal and Michener, 1958). With this algorithm, clusters are considered to be dissimilar in accordance with the squared Euclidean distance between cluster centroids. The data mining application creates up to 10 partitions of the Kohonen output layer with different numbers of clusters specified by the user. These solutions are graphically displayed in a sequence

of color-coded square lattice maps. Other graphical outputs of the application include the component planes and the U-matrix of the trained network. Non-graphical outputs include a statistics file with the average U-matrix statistic and the average input–output quantization error at each iteration, and a report file with estimated weights and the distribution of linguistic profile of informants mapped to each output node.

5.3.3   Visual Data Mining

One of the problems in attempting to visualize LAMSAS lexical and pronunciation data is that there is usually so much information that it is impossible to show it all in a single figure. As the SOM algorithm reduces the dimensionality of a data set, it permits visualization of essential relationships embedded in the original data as a series of abstractions ('meta'-structures) in different graphical forms under the control of the analyst. When utilized interactively, the multiple visualization forms associated with the SOM application (feature maps, component planes, and U-matrix map) enhance the power of the method to discover new knowledge in complex data sets (Vesanto, 1999). In addition, other visualization techniques can show relationships between linguistic data, on the one hand, and geographic and social constructs on the other, such as brushing and focusing, along with dynamic linking. The interactive role of the analyst in the use of all of the visualization tools associated here with application of the SOM algorithm, creates a VDM environment that complements computational data mining techniques such as SOM networks [Ferreira de Oliveira and Levkowitz, 2003; for VDM conceptual considerations see MacEachren *et al*. (1999) and Wachowicz (2001)].

The VDM environment developed for the exploration of LAMSAS data sets is sketched in Figure 5.2. It enables the user to interact dynamically with three sets of linked visualization forms, namely visualization forms associated with the prototype data generated by the SOM data mining method at en earlier stage (feature maps, component planes, and U-matrix map), geographic maps of LAMSAS survey informants, and plots and charts
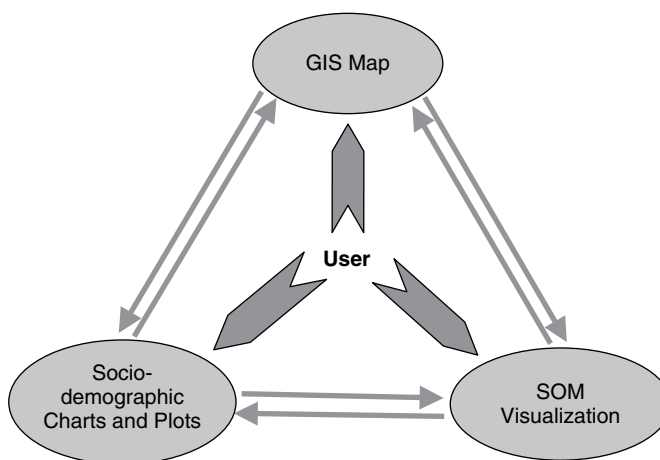
Figure 5.2   *VDM process applied to LAMSAS data sets*

of socio-demographic characteristics of informants (age, gender, ethnicity, educational attainment, and others). With the power to navigate between these three spaces, the analyst can explore geographic and social correlates of complex linguistic constructs. Brushing and focusing techniques are enhanced by dynamic linking of the different types of visualization forms in the VDM environment. For instance, when the user selects the nodes belonging to a certain cluster on the SOM feature map, the locations of respondent mapped to these nodes are highlighted on the GIS map. The reserve linking of selections is also possible.

5.4   UNDERSTANDING SOUTHERN ENGLISH

The problem with conventional techniques for analysis of the most distinctive regional variant of American English, Southern English, is that they have yet to get much beyond mere perceptual description of Southern English (Kretzschmar, 2003). It is generally accepted that Southern English is somewhere below the Mason–Dixon Line (Preston 1997), but we all tend to mean something different by the label 'Southern English'. Of course, it has long been known that there is a great deal of variation in the English spoken by Southerners, measurably more variation than is found in the North (Kretzschmar, 1996b). The only safe conclusion from such varying perceptions and varying Southern linguistic features is that Southern English must really be a high-level abstraction, rather than any specific and systematic collection of linguistic features shared by a community of speakers. Furthermore, the abstraction is one that even specialists construct somewhat differently, so that we fool ourselves if we think that we are all talking about the same thing when we talk about Southern English.

This section is intended to illustrate the application of the knowledge discovery environment developed around the SOM algorithm to shed new light on the long-standing debate over Southern English and its geographic imprint. Given the space limitation, no discussion of socio-demographic correlates will be provided, and the number of input files and the size of the square lattice have been severely limited. The study is based on a selection of six lexical data files for input: *quarter of* (for telling time), *pretty day* (for good weather), *dog irons* (for what you put the wood on in a fireplace), *shelf* (for the shelf over a fireplace, a mantel), *kindling* (for the wood used to start a fire), and *blinds* (on a roller, for window privacy). These particular response types are selected for this experiment because they all have between 100 and 600 occurrences in the data set, out of 1162 speakers; experience indicates that words with a moderate frequency level, as opposed to a very high or very low frequency level, work better for finding regional distributions.

Intensive training and testing of different sets of parameters were conducted to arrive at the stable and fairly consistent results presented here. We represent each lexical variation by three input nodes (the second pre-processing modality discussed in Section 5.3.2), so that the training data consist of 1162 vectors, each being 18-dimensional (6×3). We have neutralized the impact of missing data by coding '0' the missing node of each lexical variation. The network is based on a 5×5 square lattice and the initial search radius is commensurably set to 2 to capture enough of the local neighborhood effects and avoid the effects spanning the entire grid. Training is conducted over 30 epochs, beyond which

weight stability is achieved. Four clustering schemes of nodes are requested, ranging from three to six clusters each.

The component planes of each input node are depicted in Figure 5.3. In this figure, the planes are grouped by the lexical variation they index: in each triplet, the leftmost plane is for positive responses, the central plane is for negative, and the rightmost plane is for missing responses. Given the '0' coding on missing nodes, the third component plane of each lexical variation is uniformly zero. The component planes reveal that each input vector has its own signature on the output nodes and that there is little redundancy in the data set. Some similarities exist between occurrence of the words 'blinds' and 'pretty day', the former appearing in a subset of cases where the latter is used.
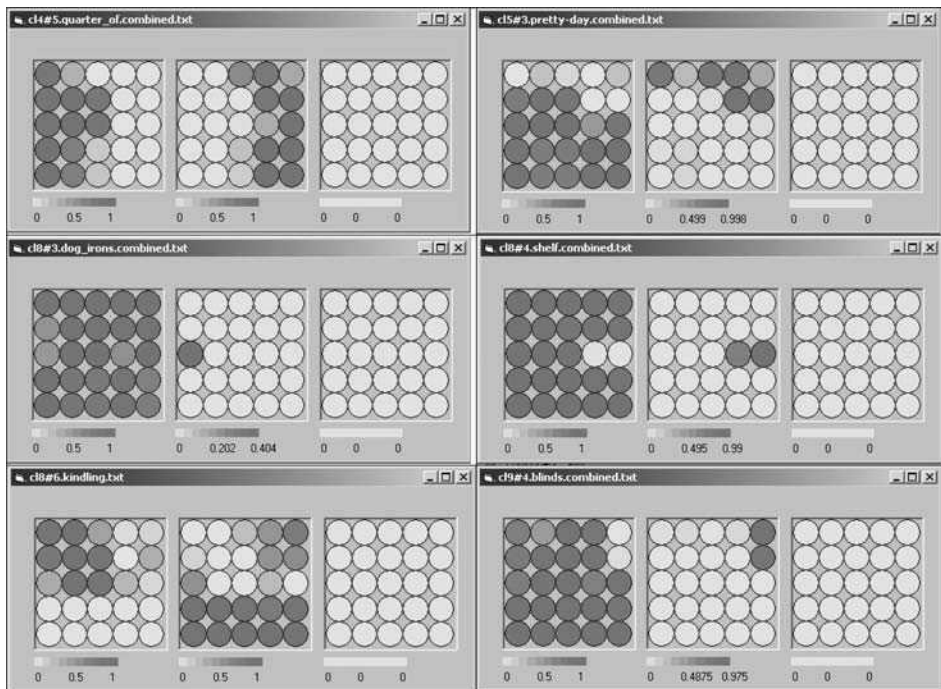
Figure 5.3   *Component planes of the node triplets associated with each lexical variation of the experimental SOM run (See Colour Plate 6)*

Let us now consider the three-cluster solution illustrated in Figure 5.4. It is appropriate for this experiment to consider quartiles for interpretation of weights and thus of the component words: 75 % or better can be taken as selection of the presence or absence state of the word; 50–75 % can be taken as a trend for the presence or absence state of the word; the third number is zero since it stands for missing data whose impact of network training has been eliminated, and finally, if no number is reported above 50 %, the word should be considered as having no clear-cut trend. Cluster 1 selects for the lack of usage of any of the words analyzed; dynamic linking between the feature map and the GIS map indicates that, while the 649 informants associated with these

Average Cluster Weights	Clusters		
	1	**2**	**3**
Quarter of / positive	0.84	0.65	0.01
Quarter of / negative	0.02	0.07	0.82
Quarter of / missing	0.00	0.00	0.00
Pretty day / positive	0.73	0.82	0.53
Pretty day / negative	0.17	0.07	0.39
Pretty day / missing	0.00	0.00	0.00
Dog irons / positive	0.85	0.90	0.88
Dog irons / negative	0.11	0.00	0.06
Dog irons / missing	0.00	0.00	0.00
Shelf / positive	0.97	1.00	0.80
Shelf / negative	0.02	0.00	0.17
Shelf / missing	0.00	0.00	0.00
Kindling / positive	0.87	0.00	0.14
Kindling / negative	0.10	0.98	0.62
Kindling / missing	0.00	0.00	0.00
Blinds / positive	0.84	0.99	0.77
Blinds / negative	0.04	0.00	0.21
Blinds / missing	0.00	0.00	0.00

Input Nodes (row label); Legend (above table)

Figure 5.4 *Feature map of the three-cluster solution (See Colour Plate 7)*

nodes, that is, informants who use none of the target words, are peppered throughout the LAMSAS survey area, they are most strongly concentrated south of the Mason–Dixon Line as well as in Western Pennsylvania. In order to evaluate the significance of cluster 1, we can compare its geographic expression to Kurath's map of dialect regions, which he first prepared on the basis of multiple selected lexical isoglosses (Kurath, 1949), and later largely confirmed through isogloss analysis of pronunciation features (Kurath and McDavid, 1961). It is clear that the territory where linguistic cluster 1 prevails (Figure 5.5) does not match any of Kurath's three primary dialect regions. Its densest pattern roughly corresponds to dialect subregions labelled 10–13 and 15–18 in Figure 5.6.

Cluster 2 identifies the usage of *kindling* and the lack of usage of *pretty day, dog irons, shelf*, or *blinds*, and a tendency towards not using *quarter of*. Only 64 informants are mapped to this cluster. The majority of these informants live in Georgia, South Carolina, and Northern Florida, with smaller groups in western Virginia and in the Chesapeake Bay area, while the rest are scattered as far north as Rochester, NY (Figure 5.7). Once again, there is no correspondence with Kurath's dialect regions. However, Georgia, South Carolina, Northern Florida, and Upstate New York are just the places where one of the two main LAMSAS field workers (Raven McDavid) conducted interviews. Cluster 2
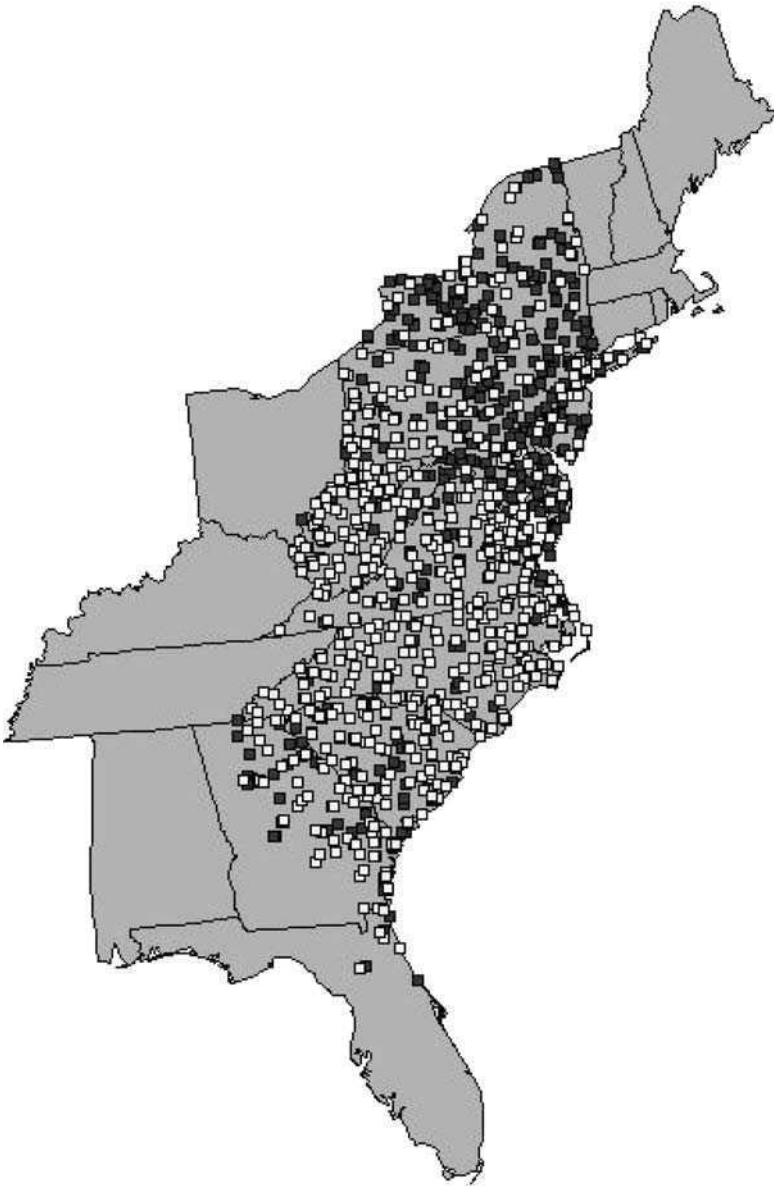
Figure 5.5   *Geographic distribution of informants mapped to cluster 1 (clear squares)*

thus may be influenced by how the survey was conducted, as much as it is associated with the presence or absence of linguistic features.

Cluster 3 selects for the usage of *quarter of*, a tendency to use *kindling*, the lack of usage of *dog irons, shelf*, or *blinds*, and a tendency not to use *pretty day*. The 449 informants mapped to cluster 3 (Figure 5.8) most often live north of the Mason–Dixon Line, excluding the western half of Pennsylvania. The geographic area dominated by the

WORD GEOGRAPHY OF THE EASTERN STATES

THE NORTH

1. Northeastern New England
2. Southeastern New England
3. Southwestern New England
4. Upstate New York and w. Vermont
5. The Hudson Valley
6. Metropolitan New York

THE MIDLAND

7. Delaware Valley (Philadelphia Area)
8. Susquehanna Valley
9. Upper Potomac & Shenandoah Valleys
10. Upper Ohio Valley (Pittsburgh)
11. Northern West Virginia
12. Southern West Virginia
13. Western North and South Carolina

THE SOUTH

14. Delmarvia (Eastern Shore of Maryland and Virginia, and Southern Delaware)
15. Virginia Piedmont
16. Northeastern North Carolina (Albemarle Sound and Neuse Valley)
17. Cape Fear and Peedee Valleys
18. South Carolina

Figure 5.6 *Kurath's map of dialect regions. Reprinted with permission from Hans Kurath,* A Word Geography of the Eastern United States, *Ann Arbor: The University of Michigan Press, 1949, Figure 3*

densest pattern of this cluster encompasses Kurath's Northern dialect region, but it also includes Midland regions 7, 8, and 14.

All three dialect clusters identified by the centroid algorithm applied to the trained network show a geographic focus. This focus is not exclusive or crisp, however. This result is in agreement with earlier studies that suggested that Kurath's delineation of dialect region is more a construct of convenience than a reality, and that the linguistic space is a continuum (Heeringa and Nerbonne, 2001; Kretzschmar, 1992). The results from SOM show a Northern plus Midland pattern in cluster 3 (Figure 5.8), and a Southern plus Midland pattern in cluster 1, which addresses the northern reaches of Southern English, and especially whether and how Kurath's Midland region separates the North from the South. Finally, Southern English is found to be very heterogeneous, in that cluster 1 is negatively defined by the absence of any of the target features, not by the presence of particular target features. As a result, communities in Georgia, South Carolina, and northern Florida may be linguistically as distinct from other southern communities as they are from northern ones.

The VDM environment is most effective in the comparison of the sequence of clustering solutions of SOM prototype data, and so the best analysis would examine a sequence
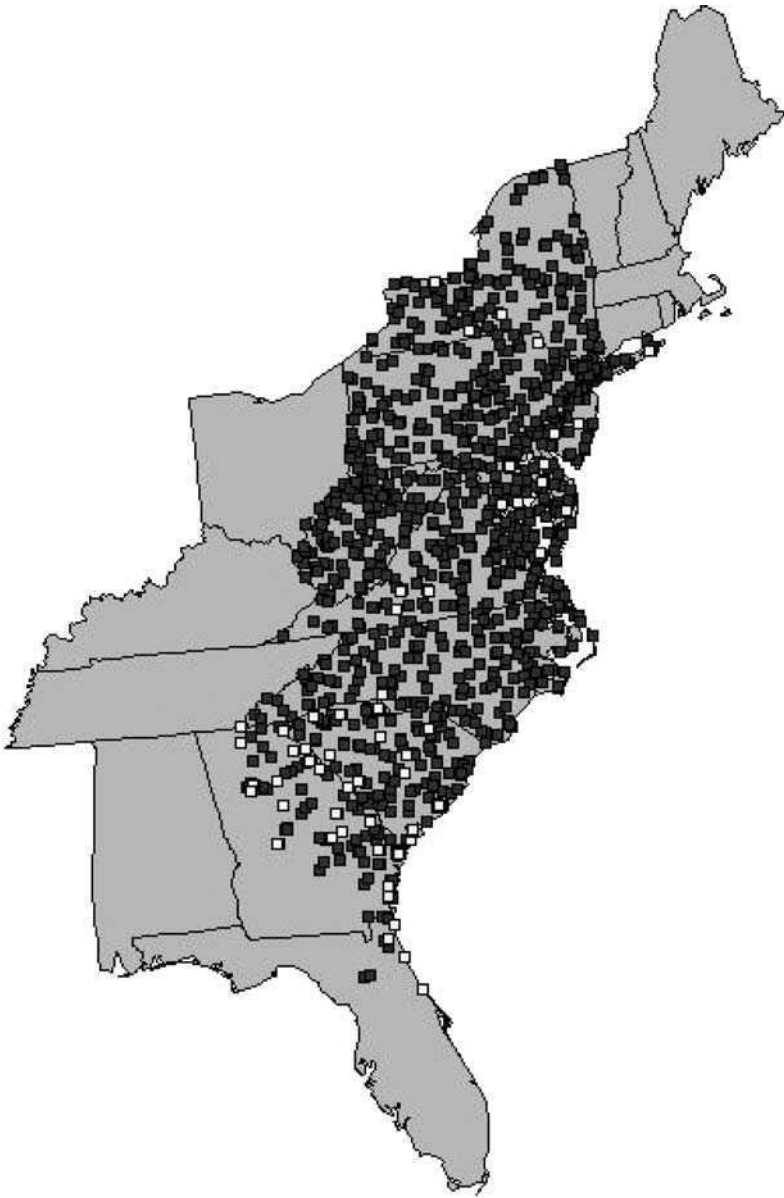
Figure 5.7   *Geographic distribution of informants mapped to cluster 2 (clear squares)*

of SOM solutions in detail (especially when more input files and a larger square lattice is specified). Due to space limitations, we present only one other solution, the feature map with four clusters (Figure 5.9). Geographic maps of the four clusters are compiled in Figure 5.10. It can be seen that cluster 4 of the four-cluster solution is generally similar to cluster 3 of the three-cluster solution; however, the other clusters are configured in fundamentally different ways. Cluster 1 now roughly corresponds to Kurath's regions
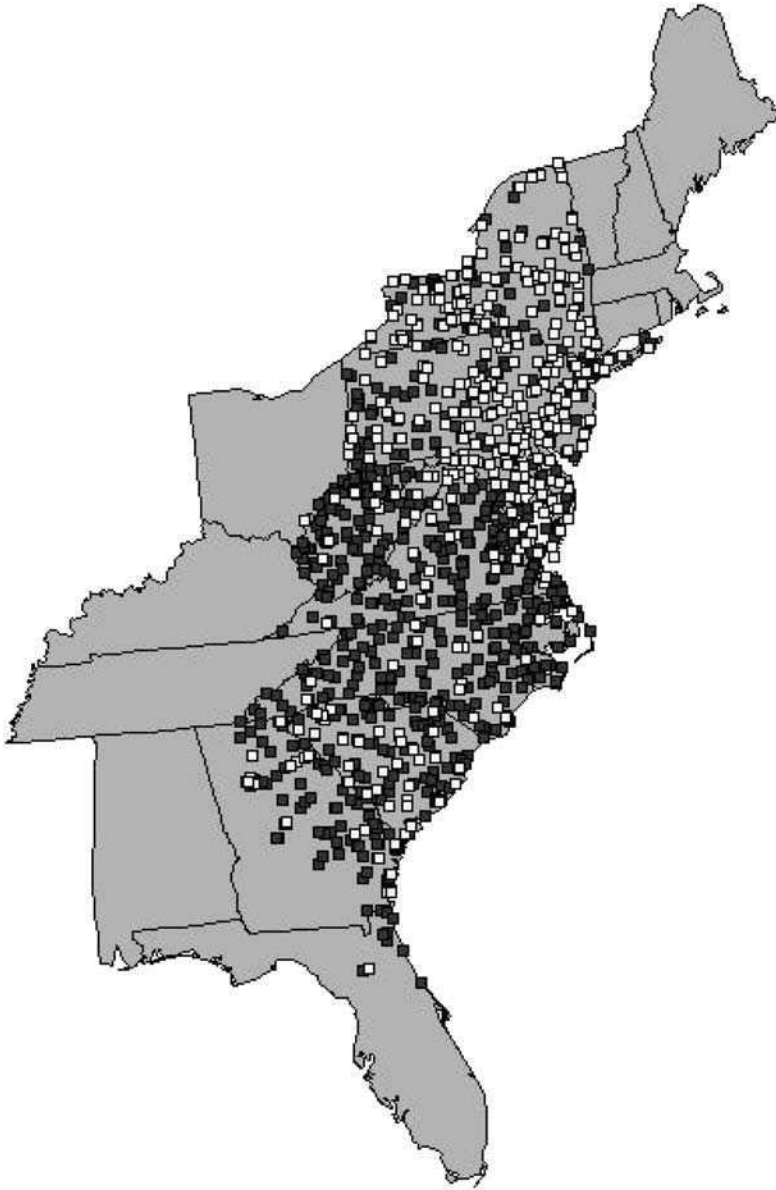
Figure 5.8   *Geographic distribution of informants mapped to cluster 3 (clear squares)*

10 and 11, a West Midland grouping, with more scattered informants to the east and south. Cluster 2 comes close to a general Southern distribution in its densest pattern, with scattered informants further north. Cluster 3 has its densest pattern in the East Midland area, with scattered informants elsewhere. Finally, it should be said that the cluster 4 appears to be less dense and coherent than the others, so that it cannot really be claimed to be a clear Northern complement to the Southern pattern of cluster 2.

	Clusters			
Average Cluster Weights	**1**	**2**	**3**	**4**
Quarter of / positive	0.69	0.83	0.01	0.00
Quarter of / negative	0.07	0.02	0.74	0.89
Quarter of / missing	0.00	0.00	0.00	0.00
Pretty day / positive	0.83	0.70	0.08	0.90
Pretty day / negative	0.07	0.19	0.81	0.03
Pretty day / missing	0.00	0.00	0.00	0.00
Dog irons / positive	0.85	0.90	0.93	0.83
Dog irons / negative	0.06	0.06	0.03	0.10
Dog irons / missing	0.00	0.00	0.00	0.00
Shelf / positive	1.00	0.97	0.95	0.67
Shelf / negative	0.00	0.02	0.02	0.30
Shelf / missing	0.00	0.00	0.00	0.00
Kindling / positive	0.06	0.93	0.22	0.07
Kindling / negative	0.92	0.03	0.55	0.68
Kindling / missing	0.00	0.00	0.00	0.00
Blinds / positive	0.96	0.85	0.57	0.93
Blinds / negative	0.02	0.04	0.42	0.03
Blinds / missing	0.00	0.00	0.00	0.00

The left side of the table is labelled "Input Nodes" and the legend header reads "Legend".

Figure 5.9   *Feature map of the four-cluster solution (See Colour Plate 8)*

Comparison of the three-cluster solution with the four-cluster solution validates the sequential, interactive analysis of the VDM. The clusters for each solution do not appear to be independent, but instead, when viewed in sequence, reveal different relationships in the data. Sequential analysis shows that it would be difficult to uphold any strong, sharp distinction between the South and the North; that the eastern and western parts of Kurath's Midland region pattern quite differently; that smaller, more local areas may emerge from the clustering (such as western Virginia or the Chesapeake); and that the SOM application is sensitive to aspects of survey design and execution. Just as earlier quantitative approaches have generally confirmed Kurath's findings, but at the same

Figure 5.10 *Geographic distribution of informants mapped to the four clusters (clear squares) of the four-cluster solution (See Colour Plate 9)*

time have shown linguistic reality to be more complex than Kurath's isogloss method and linguistic model would allow, SOM in this VDM environment offers yet more opportunities for the skilled analyst to see and understand geolinguistic relationships in the rich, multidimensional data. Use of the VDM to associate socio-demographic characteristics with the SOM network adds yet more opportunities.

5.5   CONCLUSIONS

Linguistic data sets represent a major challenge to spatial scientists and dialectologists because of the inherent properties of the data collected: data are often sparse, have a skewed distribution and are highly multidimensional. We proposed in this chapter an integrated data mining approach that draws on the computational power of the SOM algorithm and the heuristics of visualization to explore essential relationships in large, geo-referenced data sets such as LAMSAS. A sample experiment on six lexical variations illustrated the ability of this framework to go beyond existing methods of data analysis, to modify and enrich previous understanding of the data and to discover new relationships of interest to dialectologists. The limited results presented here support the idea that traditional theories on the territoriality of dialect constructs are not clearly founded on the empirical evidence. Lexical variations combine to form complex dialectal structures that can be elicited by SOM-based data mining and explored through interactive visualization techniques.

   While our experiments with the analysis of LAMSAS data sets using SOM are indicative of the usefulness of this approach in linguistic research, limitations are also apparent. SOM is particularly effective for detecting geographic and spatial clusters that may not be composed of contiguous locations or coherent patterns, which has usually not been possible with other techniques. However, one should be fully aware that the patterns extracted are dependent on the parameterization of the SOM algorithm in the calculation of the SOM network, as well as on data pre-processing. Therefore, extensive experimentation is necessary to identify best practices in SOM net training that apply to a specific application domain. Nonetheless, even the limited experiment presented here has successfully answered the research questions that we earlier proposed. We have been able to demonstrate how linguistic features (in this case words, but potentially and transparently also grammatical constructions, pronunciations, or combinations of the latter) can be distributed over geographic areas in the LAMSAS survey region, not in uniform patterns of complementary distribution, but in complex distributions that appear to change from the different viewpoints of the different cluster solutions. We have, moreover, demonstrated that individual linguistic features contribute differentially to the cluster solutions of the SOM algorithm, which speaks to the issue of the salience (or lack of it) of particular linguistic features or groups of features for regional patterns. Finally, we have also shown that the 'fuzzy' multidimensional clusters derived from the application of SOM are not wildly different from earlier, more subjective analysis, but instead enhance and enrich what we already thought we knew with new insights. We look forward, in another place, to demonstration of the value of association in the VDM of socio-demographic information with the geographic and spatial clusters.

REFERENCES

Atwood, E.B. (1953). *A Survey of Verb Forms in the Eastern United States*. Ann Arbor: University of Michigan Press.
Byrne, W., K. Mastrogiannis and G.F. Meyer. (1994). Classification of Multi-spectral Remote Sensing Data with Neural Networks: A Comparative Study. *IEEE Colloquium on 'Applications of Neural Networks to Signal Processing' (Digest No. 1994/248)*: 51–52.

Chambers, J.K. (1998). Myth 15: TV Makes People Sound the Same. In *Language Myths*, L. Bauer and P. Trudgill (eds), pp. 123–131. Harmondsworth: Penguin Books.

Chen, C.H. and B. Shrestha. (2000). Classification of Multi-sensor Remote Sensing Images Using Self-Organizing Feature Maps and Radial Basis Function Networks. *International Geoscience and Remote Sensing Symposium (IGARSS)* 2: 711–713.

Ferreira de Oliveira, M.C. and H. Levkowitz. (2003). From Visual Data Exploration to Visual Data Mining: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 9(3): 378–394.

Hara, Y., R.G. Atkins, S.H. Yueh, R.T. Shin and J.A. Kong. (1994). Application of Neural Networks to Radar Image Classification. *IEEE Transactions on Geophysics and Remote Sensing* 32: 100–111.

Heeringa, W. and J. Nerbonne. (2001). Dialect Areas and Dialect Continua. *Language Variation and Change* 13: 375–400.

Himanen, V., T. Järvi-Nykänen and J. Raition. (1998). Daily Travelling Viewed by Self-Organizing Maps. In *Neural Networks in Transport Applications*, V. Himanen, P. Nijkmap and A. Reggiani (eds), pp. 85–110. Aldershot: Ashgate.

Kaski, S. and T. Kohonen. (1996). Exploratory Data Analysis by the Self-organizing Map: Structures of Welfare and Poverty in the World. In *Neural Networks in Financial Engineering*, A.-P. N. Refenes, Y. Abu-Mostafa, J. Moody and A. Weigend (eds), pp. 498–507. World Singapore: Scientific.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd Edn. Berlin: Springer.

Kretzschmar, W.A., Jr. (1992). Isoglosses and Predictive Modeling. *American Speech* 67: 227–249.

Kretzschmar, W.A., Jr. (1996a). Quantitative Areal Analysis of Dialect Features. *Language Variation and Change* 8: 13–39.

Kretzschmar, W.A. Jr. (1996b). Foundations of American English. In *Focus on the USA*, E. Schneider (ed.), pp. 25–50. Philadelphia: John Benjamins.

Kretzschmar, W.A., Jr. (1998). Analytical Procedure and Three Technical Types of Dialect. In *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*, M. Montgomery and T. Nunnally (eds), pp. 167–185. Tuscaloosa: University of Alabama Press.

Kretzschmar, W.A., Jr. (2003). Mapping Southern English. *American Speech* 78: 130–149.

Kretzschmar, W.A., Jr and J. Lee. (1993). Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Geographical Information Systems* 7: 541–560.

Kretzschmar, W.A., Jr and D. Light. (1996). Mapping with Numbers. *Journal of English Linguistics* 24: 343–357.

Kretzschmar, W.A., Jr and E.W. Schneider. (1996). *Introduction to Quantitative Analysis of Linguistic Survey Data*. Thousand Oaks: Sage Publications.

Kretzschmar, W.A., Jr and S. Tamasi. (2003). Distributional Foundations for a Theory of Language Change. *World Englishes* 22: 377–401.

Kretzschmar, W.A. Jr, V.G. McDavid, T.K. Lerud and E. Johnson (eds). (1993). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Kurath, H. (1939–43). *Linguistic Atlas of New England. 3 vols*. Providence: Brown University, for ACLS.

Kurath H. (1949). *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Kurath, H. and R. I. McDavid. (1961). *The Pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.

Labov, William, Charles Boberg, and Sherry Ash. (2006). *Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.

MacEachren, A., M. Wachowicz, D. Haug, R. Edsall and R. Masters. (1999). Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographic Visualization with Knowledge Discovery in Database Methods. *International Journal of Geographic Information Science* 13(4): 311–334.

Nerbonne, J. and W. Heeringa. (2001). Computational Comparison and Classification of Dialects. *Dialectologia et Geolinguistica* 9: 69–83.

Nerbonne, J. and P. Kleiweg. (2003). Lexical Distance in LAMSAS. *Computers and the Humanities* 37: 339–357.

Nerbonne, J. and W.A. Kretzschmar, Jr. (2003). Introducing Computational Techniques in Dialectology. *Computers and the Humanities* 37: 245–255.

Openshaw, S. and C. Wymer. (1991). A Neural Net Classifier System for Handling Census Data. In *Neural Networks for Statistical and Economic Data*, F. Murtagh (ed.), pp. 73–86. Dublin: Munotec.

Openshaw, S. and C. Openshaw (1997). *Artificial Intelligence in Geography*. Chichester: John Wiley & Sons, Ltd.

Preston, D. (1997). The South: The Touchstone. In *Language Variety in the South Revisited*, C. Bernstein, T. Nunnally and R. Sabino (eds), pp. 311–351. Tuscaloosa: University of Alabama Press.

Schneider, E.W. (1988). Qualitative Methods of Area Delimitation in Dialectology: a Comparison Based on Lexical Data from Georgia and Alabama. *Journal of English Linguistics* 21: 175–212.

Sokal, R.R. and C.D. Michener. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin* 38: 1409–1438.

Vesanto, J. (1999). SOM-Based Data Visualization Methods. *Intelligent Data Analysis* 3(2): 111–126.

Villmann, T., E. Merenyi and B. Hammer. (2003). Neural Maps in Remote Sensing Image Analysis. *Neural Networks* 16(3–4): 389–403.

Wachowicz, M. (2001). GeoInsight: An Approach for Developing a Knowledge Construction Process Based on the Integration of GVis and KDD Methods. In *Geographic Data Mining and Knowledge Discovery*, H. J. Miller and J. Han (eds), pp. 239–259. London: Taylor & Francis.

6

Self-Organizing Maps for Density-Preserving Reduction of Objects in Cartographic Generalization

Monika Sester

Institute for Cartography and Geoinformatics, Leibniz University of Hannover, 30167 Hannover, Germany

6.1 INTRODUCTION AND OVERVIEW ON RELATED WORK

Generalization is a process used for reducing the volume of data of a spatial data set while preserving important structures. Generalization operations can be distinguished into selection, simplification, classification, enhancement, aggregation, displacement, and typification, and are generally employed by cartographers when designing maps in smaller scales. Depending on the spatial situation, the target scale, the application, and the objects involved, these operations are applied with different parameters and in different sequences of the operations.

During the last 40 years, considerable attempts have been made to automate this process of generalization (Meng, 1997). They have been partly successful in designing a number of automated operations that allow for solving dedicated problems in generalization. Among those, algorithms for line simplification are very prominent (Douglas and Peuker, 1973).

Typification is the process of reducing the number of similar objects while preserving their spatial arrangement and density. For instance, a group of islands or a group of buildings can be reduced to a new group with less objects, while preserving the spatial distribution. The number of objects in the target scale can be determined by analysing the relation of space needed by the objects in the source scale and the target scale, e.g. using Töpfer's 'radical law' (Töpfer, 1976), that computes the number of objects from a function of the quotient of the two scales – typically the square root. It can also be calculated by considering the black-and-white ratio in the source and target scale, i.e. the ratio of objects vs background in the map, which should be preserved. The decision regarding which specific objects to preserve is more difficult. A mere random selection can lead to the required reduction. However, this will not necessarily preserve the spatial distribution of the objects, or other constraints that are inherent with the object. In this chapter there is a concentration of approaches that treat point and polygon objects, but the typification of linear objects is not treated here. That problem requires different approaches, as the target function is different, e.g. in the case of reduction of streets the connectivity of the road network has to be preserved. Typification of point or polygon objects involves a structure recognition process that determines homogeneous groups of objects, within which individual objects can then be selected and rearranged according to the then known underlying structure. This leads to the following steps:

1. a reduction rate is given;
2. the structure of the object distribution is recognized, in terms of identifying groups of objects with similar characteristics in spatial proximity;
3. a reduction of the number of objects and possibly a rearrangement of those objects is done within the groups.

The main difficulty in these steps is the determination of the original object structure, especially the identification of homogeneous structures, i.e. substructures of similar objects located in relative proximity. To this end, Töpfer (1976) proposes to identify so called 'Kleinkomplexe' (small arrangements) from which representative objects then can be selected. Anders and Sester (2000) use a clustering approach that is based on an hierarchical structure of neighbourhood graphs to determine such groups. Regnault (1996) reduces the problem of generalizing buildings in a settlement area to a one-dimensional case. He justifies this by the fact that buildings are typically arranged linearly along roads or streets. After identification of the arrangement of the buildings using Minimum Spanning Trees, he is able to do the reduction and rearrangement step by newly positioning the original buildings along the road, taking the old distribution into account.

As structure recognition is a complicated problem, researchers have sought for approaches that are able to avoid this recognition process. Müller and Wang (1992) use a raster based approach by applying morphological operations that emphasize large objects while reducing small ones. Bjørke (1996) proposes an entropy based method for feature elimination. Cecconi *et al*. (2005) used Mesh Simplification algorithms from Computer Graphics to perform a density preserving reduction of building objects.

Højholt (1995) first elaborated on the use of Kohonen Self-Organising Feature Maps for typification, which was later extended by Sester and Brenner (2000). The main advantage of this approach is the fact that no explicit identification and representation of

the underlying structure is needed, but it is respected implicitly in the process. Thus, in the above described list of necessary steps, the second one can be omitted.

In this chapter, the application of SOM to typification will be discussed. The necessary specifications and adaptations are described which lead to the core typification algorithm, which is presented and verified with some examples. In Section 6.3, an extension of this core algorithm for the typification of buildings is outlined in detail. A range of examples will demonstrate the potential of this approach. An evaluation of the procedure and the results is given in Section 6.4, and Section 6.5 summarizes and concludes the chapter.

6.2   SOM FOR TYPIFICATION

The principles of self-organizing maps (SOMs) are discussed in Chapter 1. Therefore, in this section only the adaptations for the application in typification are described.

SOMs have the characteristic that they can approximate the density in the input space: this means that in areas with many stimuli also many neurons will accumulate, whereas areas with less objects in input space will also be represented with less neurons. This characteristic can be exploited beneficially for typification.

The aim of typification is to reduce the number of objects for presentation in a smaller scale while preserving their spatial distribution and density. In order to apply SOM for typification, the following assumptions are made: the original objects in the source scale represent the *input space*, whereas the reduced number of objects in the target map represent the neurons in the *map space*. The reduction rate can be determined using Töpfer's radical law and the objects are selected randomly. The objects are points that represent objects of the same type – this assumption can be partially relaxed, which is shown in Section 3.

Thus, the map space of the neurons is composed of a subset of the objects in the input space. This leads to some consequences concerning the structure of the map: the topology of the neurons is known and given by their neighbourhood. It can be determined by a Delaunay Triangulation, thus leading to a nonregular grid. The inital weights of the neurons are the positions of the corresponding original objects. As in the case of typification the rearranged neurons will be in the vicinity of the original objects, good approximate values for the position are already given, which leads to the following constraints and simplifications concerning the selection of the neighbourhood range and function:

- The degree of neighbourhood is set to one, i.e. only the immediate neighbours of the neurons are considered.
- The movement of the neurons – even in the initial phase – is relatively restricted, i.e. the neighbourhood function h yields low values also in the beginning of the iteration.
- The topology is not rebuilt, even when in the course of the iterations due to the movement of the neurons the Delaunay criterion is violated. This is reasonable as the initial neighbourhood relations have to be preserved.

This approach has been implemented in a program written in $C++$. The Gaussian function was chosen as neighbourhood function h. The parameters needed for the process are learning rate and number of iterations. Those parameters have been empirically determined and then fixed for all experiments: the number of iterations is 50, the learning rate

was reduced linearly with the number of iterations from 1 to 0.001. No pre-processing steps are needed. The only parameter that has to be specified by the user is the reduction rate. The input and output format for the objects is ESRI Shapefiles (ESRI, 1998).

In the following, the process as well as the characteristics of the approach are visualized with the help of several examples. The iterative adaptive process in the course of the learning is visualized in Figure 6.1: the stimuli, i.e. the original objects, are represented with small dark dots, the neurons are visualized as light circles.
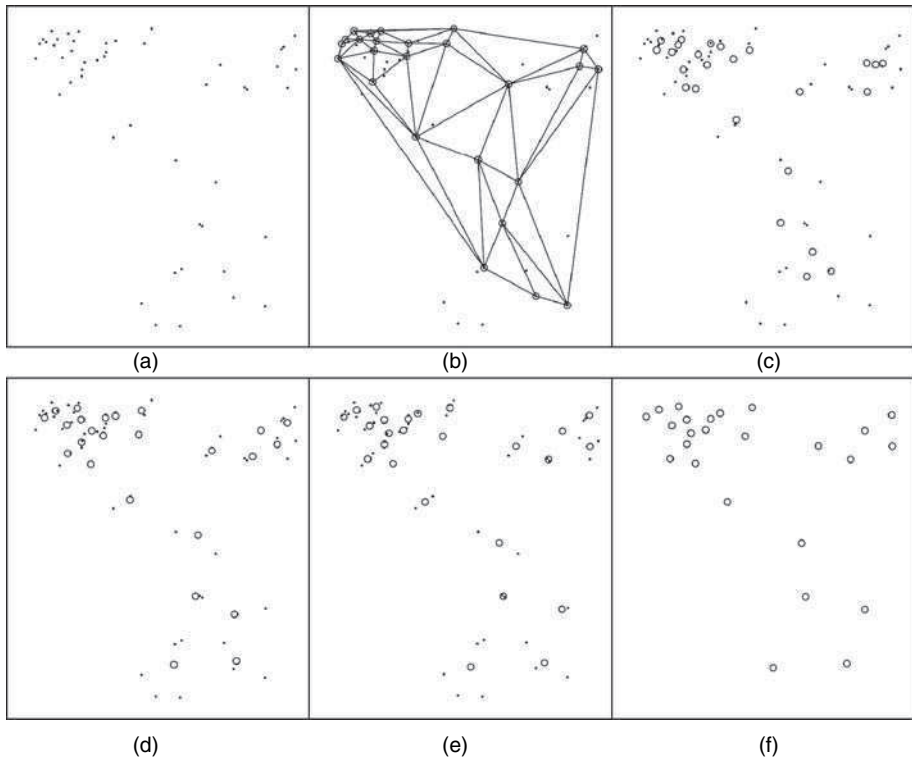
Figure 6.1   *Input space with stimuli (dark dots) (a); map space with selected neurons (light circles), triangulated (b); iterative adaptation of neurons to stimuli (c–e); result of adaptation process (f) (reduction rate 50 %)*

Figure 6.1(a) shows the initial situation of the process, namely the stimuli. From those, a certain percentage, here 50 %, is selected randomly which then represent the neurons. These neurons are linked with a Delaunay triangulation [Figure 6.1(b)] from which their neighbourhood can be determined. In Figure 6.1(c)–(e) it can be observed that during the iterations the neurons first tend to move towards the centre in order to determine the coarse structure of the map. In later iteration steps, the local adaptation to the stimuli is done. The result can be seen in Figure 6.1(f). It is clearly visible that the density and distribution of the original situation is preserved by resulting high-density areas in the upper left, as well as the low-density areas in the lower right.
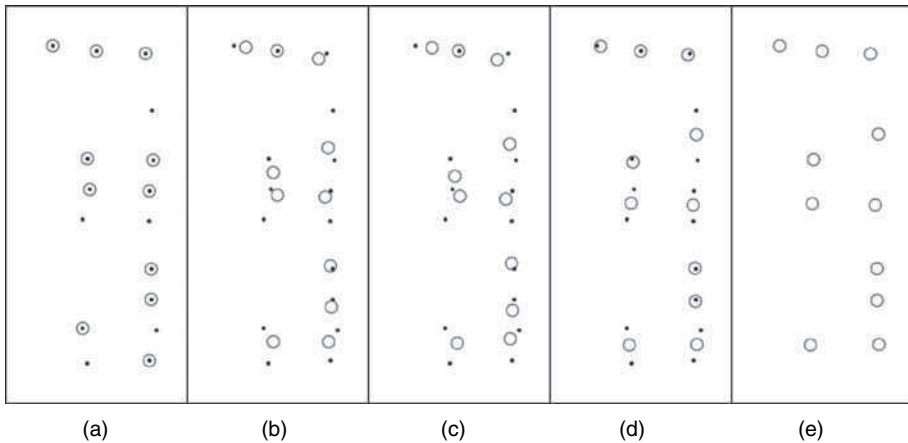
Figure 6.2 *Typification of a primarily linear structure (reduction rate 70 %): (a) start situation with dark points as stimule and circles as neurons; (b–d) intermediate situation of neuron distribution; (e) final distribution of neurons*

Figure 6.2 shows that the algorithm reacts to a regular, primarily linear, spatial structure, consisting of two vertical rows of objects. The random selection of 70 % of the objects leads to the initial situation in Figure 6.2(a). In the course of the iterations the neurons adapt to the stimuli. In the end [Figure 6.2(e)] both the linear structure and the different densities in the original structure are nicely preserved.

The preservation of the regularity in Figure 6.2, i.e. the linear structure, is achieved only implicitly in the process. It is due to the fact that the distances within the rows are smaller than those across the rows. This leads to the desired effect that the neurons move to the closest stimuli in their vicinity. However, given an equally spaced grid of stimuli, the algorithm cannot preserve this regularity, as the neurons move into the middle of the neighbouring stimuli (Figure 6.3).
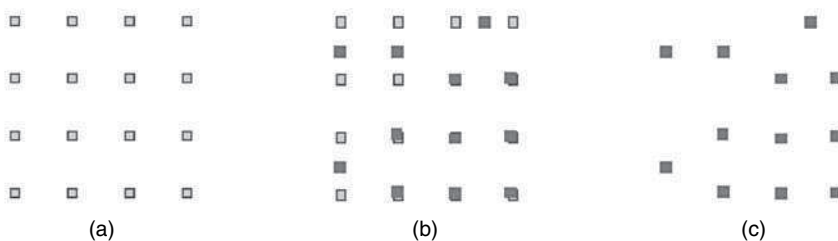
Figure 6.3 *Regular grid structure of objects, which cannot be preserved: initial situation (a); overlay of initial situation and result (b); result (c)*

In Figure 6.4 an example is given that shows three rows of points: whereas the second and third row are equally spaced, the first row exhibits two different densities. The result of two executions of the algorithm is shown in the two rows of Figure 6.4; the reduction rate was 70 %. The result verifies that the different densities can be preserved, even when there are slight differences in the two experiments. Due to the initial random selection
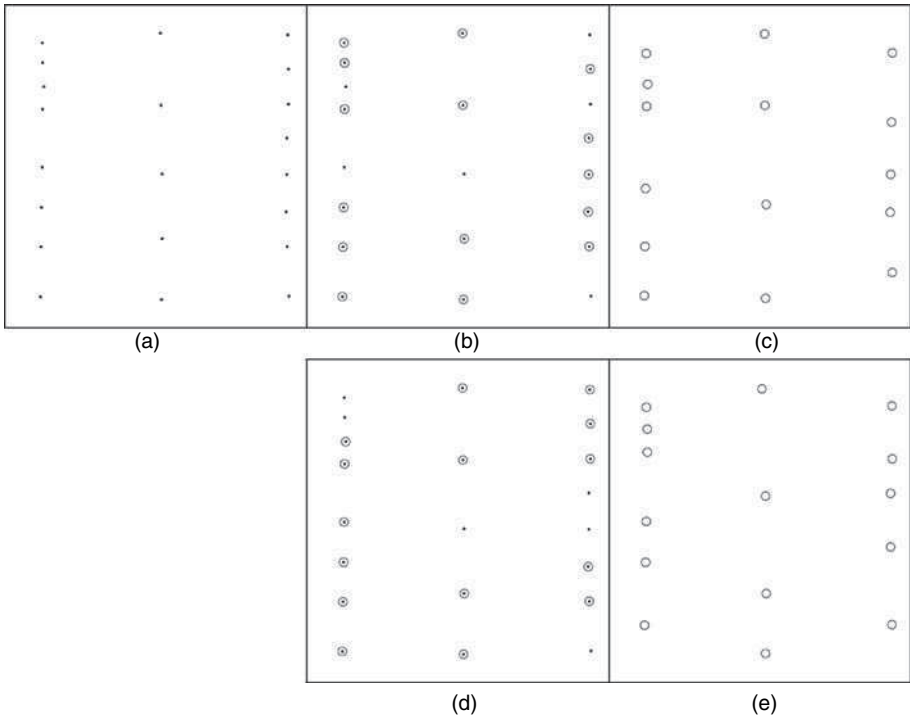
Figure 6.4   *Result of two runs on a data set consisting of three rows of objects with different spacing (a): initial selection of neurons in runs 1 and 2 (b and d); result of the two runs (c and e). Note that in the second run one of the neurons in the first row moved to the top cluster due to the attraction of those stimuli (reduction rate 70 %)*

process the approach leads to different results when processed repeatedly. However, the influence of the selection process is reduced, as it is followed by he rearrangement of the objects with the SOM.

Figure 6.5 shows a larger data set, where different reduction rates are applied, leading to representations with less objects {reduction to 60 % [Figure 6.5(b)] and to 40 %
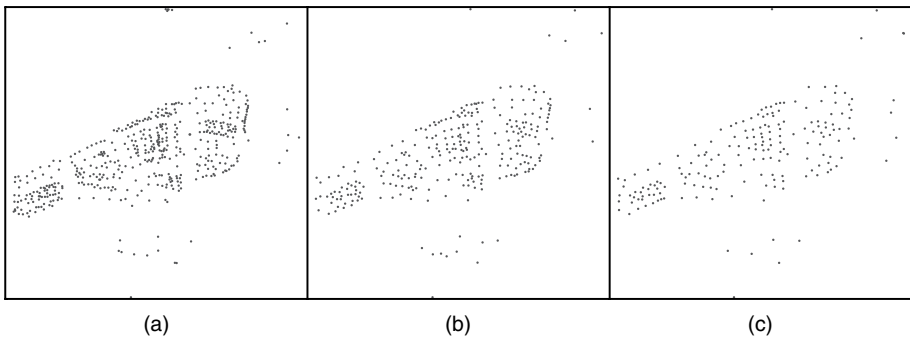
Figure 6.5   *Reduction of a larger data set (a): reduction to 60 % (b) and 40 % (c)*

[Figure 6.5(c)] of the original data set}. In all the representation the structure of the original data set is represented very well with increasingly less points.

The process described above constitutes the core algorithm applicable for point objects of similar type. In order to apply it for the generalization of buildings in smaller scales some adaptations have to be made, which are described in Section 6.3.

6.3   EXTENSION OF CORE ALGORITHM FOR THE GENERALIZATION OF BUILDINGS IN SMALL SCALE REPRESENTATIONS

In small scale representations of digital and analogue maps buildings are typically no longer presented as individual objects but as objects that are placeholders for the original objects. The core algorithm described in the previous section works on groups of similar point objects. Buildings are polygon objects of different type and size. In traditional maps the placeholders are mostly symbols, usually squares, but they can also be individually shaped buildings, if they exceed certain sizes. Thus the algorithm has to be adopted to work on non-point objects which, in addition, are no longer of exactly the same type. Another necessary adaptation relates to the number of objects processed at a time: when a whole city or even a whole map sheet has to be processed, it makes no sense to apply the core algorithm to the whole data set but on meaningful partitions, as the generalization has to act locally. Such a partition can be generated by the road network.

The procedure described in this section is designed for the generalization of buildings in maps of scales less than 1:25 000. The workflow constitutes the following steps:

1. partitioning the whole data set into generalization meshes;
2. selection of a target scale and a reduction rate;
3. random selection of buildings from the set of all buildings;
4. conversion of building polygons to their centroids;
5. typification of the selected points using the core typification algorithm;
6. reassignment of building shapes to the rearranged points;
7. resolution of spatial conflicts;
8. optional re-iteration.

In the following, these processing steps will be explained in more detail.

6.3.1   Partitioning into Generalization Meshes

The partitioning is done based on the road network using a topology-building algorithm, that looks for closed meshes in the given linear street objects. The typification is then executed for each individual mesh and the resulting typified buildings are merged to one data set in the end. This leads to the effect that buildings within one street mesh are generalized separately from buildings in another mesh. Besides speeding up computing time it is also necessary in order to restrict the movement of the objects locally.

6.3.2   Specification of Target Scale and Reduction Rate

The user specifies the target scale of the new object representation, e.g. 1:30 000 or 1:70 000. Furthermore, also the reduction rate can be explicitly given in terms of percentage of objects to be retained in the target scale (values between 0 and 1, where 1 means to preserve all objects). Although the reduction rate could be determined from the target scale using Töpfer's law, it still makes sense to give the user the option to choose the reduction rate himself. The target scale also determines the size of the building symbols (0.5×0.5 mm in map).

6.3.3   Random Selection Prioritizing Larger Objects

The goal of typification of buildings is a presentation which is similar to the original presentation, but with less objects. Human cartographers tend to prioritize large buildings in the target map. This, however, does not mean that *only* large buildings should be presented. A mere random selection treats all objects as equal. Therefore, a prioritization of large objects was implemented, leading to the desired effect, that typically the large objects survive. This was achieved using the building size as weight in the random selection process.

6.3.4   Conversion of Building Polygons to their Centroids

As the core algorithm needs point objects as input, the buildings have to be converted to points, which is achieved by calculating the centroid as building representatives.

6.3.5   Typification of the Building Centroids

The building centroids are the input to the core algorithm. The only parameter transferred to the core algorithm is the reduction rate, all the other parameters are set as described above in section 6.2. The result is a reduced and rearranged set of point objects.

6.3.6   Reassignment of Areal Objects

The result of the typification algorithm in step 5 are the representatives of the buildings for the new representation in the smaller scale. As they are, however, only point objects, they have to be assigned areal building objects again. There are different options to do so, e.g. to use the representative as such and use the shape of the original object. This could, however, lead to irritations when the object has been shifted considerably in step 5: the map reader would not expect to see a distinct object at a different location. Another option is to look in the vicinity of the rearranged neurons and take the original building which is closest. This option was chosen, as it can cope with larger movements of the neurons and presents that stimulus, that attracted the neuron most.

   Then, two cases have to be distinguished depending on a threshold concerning the size of the building. This threshold corresponds to an area of 0.5 mm $\times 0.5$ mm in the target

scale: if the original building is smaller than this threshold, it is too small to be legible, and it is symbolized by a square of 0.5 mm × 0.5 mm. The square is oriented according to the orientation of the original building. The orientation is the main direction of the building determined by an algorithm presented by Doytsher (1988). If the building is larger than the threshold, it is represented in its original shape.

6.3.7   Resolution of Spatial Conflicts

After the assignment of area shapes to the point shaped building centroids, the newly symbolized objects might get into spatial conflict with each other as well as with other objects in the map, especially with the roads. Therefore, all the objects are processed with a displacement procedure developed at the Institute for Cartography and Geoinformatics, called PUSH (Sester, 2005). This holistic displacement approach is based on Least Squares Adjustment. It leads to a minimization of all spatial conflicts while enforcing legibility constraints in the whole scene. Conflicts are solved by moving and deforming objects, depending on their object-specific parameters. Thus, in this process, adequate distances between all the objects are enforced in order to guarantee legibility.

6.3.8   Optional re-iteration

If the buildings cannot be placed without creating spatial overlaps with each other or with the neighbouring objects, the reduction rate is diminished by a pre-defined factor (5 %) and the whole process is repeated. This is done iteratively until a conflict-free solution is found or a minimal threshold for the reduction is reached.

6.3.9   Examples

The process can be applied to generate small scale maps of buildings. The original buildings are from cadastre (i.e. scale 1:2000), the roads are from scale 1:25 000. Figure 6.6 shows how buildings for the target scale 1:50 000 can be automatically generalized: it presents two extracts of a map with the original situation on the left and the corresponding generalization on the right. It can be seen that the distribution of original buildings is preserved after the generalization. In that target scale of 1:50 000, almost all the buildings are replaced by square symbols; only a few are large enough to still be represented in their original shape.

   With this algorithm generalizations for arbitrary target scales can be generated fully automatically. This is shown in Figure 6.7. Again, the presentations were generated from cadastral buildings and the streets of the 1:25 000 road network that composed the generalization meshes. Three different scales were generated and presented in appropriate sizes: 1:25 000 [Figure 6.7(a)]; 1:50 000 [Figure 6.7(b)]; and 1:75 000 [Figure 6.7(c)]. Observe that for the smaller scale maps the road network would also need a reduction (not done here).
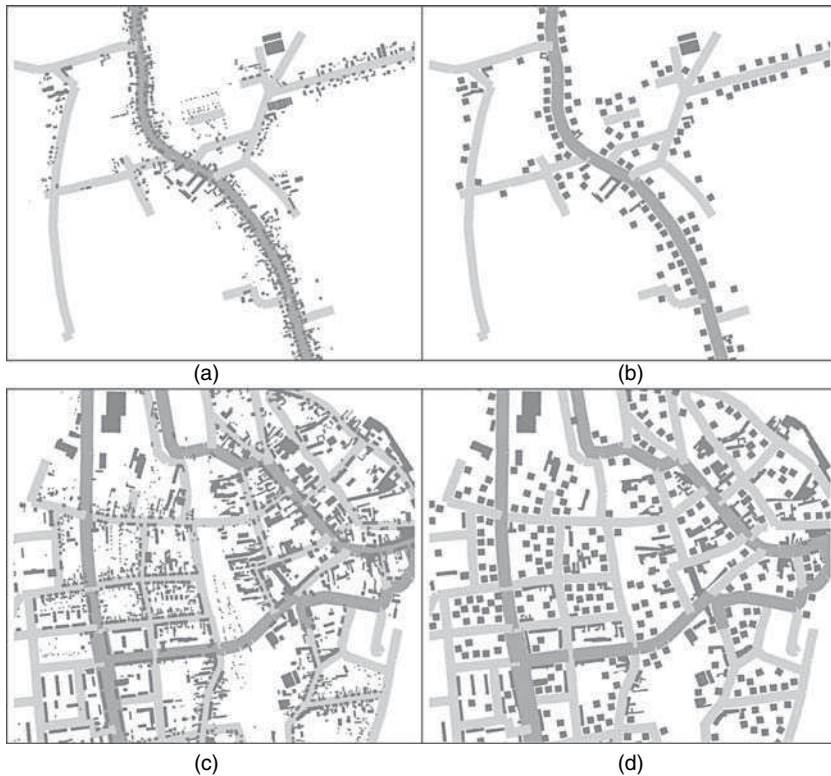
Figure 6.6 *Examples for application of typification for the generalization of buildings: original situation with cadastral buildings (1:2000) and roads (1:50 000) (a, c) and result of typification for target scale 1:50 000 (b, d)*

6.4 EVALUATION OF THE PROCEDURE AND THE RESULTS

As stated above the examples were all processed with the same parameters. This is an indication that the results are not sensitive to those values. The initial weights of the neurons as well as the topology of the net are given with good approximate values, thus no foldings ('butterfly effect') occurred, which are quite frequent in other applications of SOM. As obvious from the knowledge about the underlying process and visible from the above examples, the approach works well with randomly distributed objects.

There are some disadvantages of the approach which were also presented and explained in the examples above. As shown in Figure 6.3, the process cannot guarantee to preserve strong regularities in the data, e.g. rectangular grids. A human cartographer would simplify a regular mesh of 4×4 buildings, by a mesh of 3×3; this cannot be achieved with this method. Regular structures can only be preserved if they are distinctly clustered and separated (as shown in Figures 6.2 and 6.4).

Furthermore, the approach is not deterministic, which means that different runs of the algorithm lead to different results. This is due to the random selection of the neurons at the beginning of the process. However, the results are still reflecting the density and

Figure 6.7 *Automatic generation of different target scales from cadastral building data: 1:25 000 (a); 1:50 000 (b); 1:75 000 (c)*

distribution of the original situation, and thus are in accordance with the goal of the generalization task.

The visual impression of the results is pleasing, and they are similar to solutions given in topographic maps. Besides visual comparisons, no systematic evaluation of the results have been done by the author, mainly due to lack of adequate measures for evaluating the quality. However, the process has been tested and is now being used for the production of the map 1:50 000 by two State Mapping Agencies in Germany (in Lower Saxony and Brandenburg). The results have been evaluated by human cartographers, who in most cases were satisfied with the results. For their map production only minor corrections have been done manually (Wodtke, 2004).

Major critiques of the cartographers from the mapping agencies concern issues that are beyond the scope of the current algorithm and thus are subject to further investigations and improvements. One topic relates to the handling of objects in densely populated city centres: a human cartographer would use a space filling mesh instead of presenting individual buildings, as those tend to be squeezed together by the road network [e.g. in the meshes north-east of the centre of Figure 6.6(d)]. As the program records the proportion of object vs background area of each individual generalization mesh, this could be used as a starting point to decide if a mesh has to be filled with a complete signature or by individual buildings. A second issue is the use of squared building symbols: in some cases human cartographers would also employ a rectangular symbol when the original building resembles more that type of geometric object. In order to extend the current procedure, the building shape of the original buildings would have to be classified into square or rectangle and applied accordingly.

The algorithm is run as a batch process. The processing time for a whole map sheet of a map 1:50 000 (size in reality 20×20 km), containing e.g. 25 110 buildings is processed in approximately 15 min (on a standard PC with 600 MHz).

6.5   SUMMARY AND OUTLOOK

It has been shown in this chapter that SOM are well suited for reflecting and reproducing spatial structures. The main characteristics of this approach is that the spatial structures can be preserved without having to identify them beforehand. This is a great advantage opposed to other approaches that have to rely on the interpretation of homogeneous groups before a reduction of the data set can be done. The results are visually pleasing and convincing, which can be read from the fact that the procedure is currently being used by the Mapping Agency in Germany for map production.

One limitation of the approach lies in the fact that the preservation of dominantly regular structures as grid-like building arrangements cannot be guaranteed. A recent approach to especially recognize and treat grid structures is given by Anders (2006). Furthermore, as the method incorporates a random selection of objects, it is non-deterministic, i.e. different runs of the algorithm lead to different results. However, all the results fulfil the demand that they reflect the spatial distribution of the objects. In general, cartographic generalization is a task for which there typically are no 'model' solutions; also the work of different cartographers leads to (slightly) different results, which, however, can all be considered as valid (Spiess, 1995).

The application domain of the current version of the algorithm is urban and suburban areas, where the underlying prior assumption of the algorithm more or less holds, namely the presence of groups of similar objects. In rural areas, other approaches might be necessary that rely on an aggregation and simplification of adjacent buildings in order to present typical building shapes, e.g. farmhouses (Revell, 2004).

The approach described can be extended to generalize other kinds of objects as well. It is obvious that all kinds of point objects can be generalized, e.g. wells. For the generalization of polygonal objects the task is to determine how after the rearrangement of the objects in the SOM the objects are assigned an area object again. In principle, a similar procedure can be applied as in the case of buildings, however, there might be specific problems to solve arising from the fact that larger differences in size might occur, e.g. in the case of generalizing a set of lakes or islands.

REFERENCES

Anders, K.-H. (2006) Grid Typification. In: Riedl, A., Kainz, W. and Elmes, G. A. (Eds) Progress in Spatial Data Handling, 12th International Symposium on Spatial Data Handling, Springer-Verlag, Heidelberg, pp. 633–642.

Anders, K.-H. and Sester, M. (2000) Parameter-Free Cluster Detection in Spatial Databases and its Application to Typification. In: International Archives of Photogrammetry and Remote Sensing, Vol. XXXIII, Amsterdam, The Netherlands, pp. 75–82.

Bjørke, J. (1996) Framework for Entropy-Based Map Evaluation. Cartography and Geographic Information Science 23(2), 78–95.

Cecconi, A., Burghardt, D. and Weibel, R. (2005) Integration von Multirepräsentationsdatenbanken in den Generalisierungsprozess für die Internetkartographie (Integration of Multirepresentation Databases and Cartographic Generalization for Web Mapping, in German). Kartographische Nachrichten 55(1), 11–17.

Douglas, D. H. and Peucker, T. K. (1973). Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. The Canadian Cartographer 10, 112–122.

Doytsher, Y. (1988). Defining a Minimum Area Rectangle Circumscribing Given Information. The Cartographic Journal 25, 97–103.

ESRI (1998) ESRI Shapefile Technical Description, ESRI White Paper, July; http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Højholt, P. (1995) Generalization of Built-up Areas using Kohonen-Networks. Proceedings of Eurocarto XIII, Joint Research Centre, European Commission, Ispra, pp. 177–182.

Meng, L. (1997) Automatic Generalization of Geographic Data, Technical Report, http://www.vbbviak.sweco.se/Research net/preport/fm9706.htm, Sweden.

Müller, J. and Wang, Z. (1992) Area-Patch Generalization: A Competitive Approach. The Cartographic Journal 29, 137–144.

Regnault, N. (1996) Recognition of Building Clusters for Generalization. In: Kraak, M. and Molenaar, M. (Eds) Advances in GIS Research, Proceedings of 7th International Symposium on Spatial Data Handling (SDH), Vol. 1, Faculty of Geodetic Engineering, Delft, The Netherlands, pp. 4B.1–4B.14.

Revell, P. (2004) Building on Past Achievements: Generalising OS MasterMap Rural Buildings to 1:50 000. Sixth Workshop on Progress in Automated Map Generalization, Leicester.

Sester, M. (2005) Optimizing Approaches for Generalization and Data Abstraction. In: International Journal of Geographic Information Science, 19(8–9), pp. 871–897, 2005.

Sester, M. and Brenner, C. (2000) Kohonen Features Maps for Typification, Proceedings of the 1st GIScience Conference, Savannah, GA, USA, University of California Regents, Santa Barbara, CA.

Spiess, E. (1995) The Need for Generalization in a GIS Environment. In: J. C. Müller, J. P. Lagrange and R. Weibel, (Eds) GIS and Generalization – Methodology and Practice, Taylor & Francis, Bristol, pp. 31–46.

Töpfer, F. (1976) Ein Auswahlprogramm für punktförmige Objekte. Vermessungstechnik 24(11), 417–420.

Wodtke, K.-P. (2004) Die neue DTK50 – Umsetzung des AdV-Konzepts in Niedersachsen. In: Der X-Faktor – Mehrwert für Geodaten und Karten, Symposium, Königslutter am Elm, Kartographische Schriften, Vol. 9, Kirschbaum Verlag, Bonn, pp. 171–184.

7

Visualizing Human Movement in Attribute Space

André Skupin

Department of Geography, San Diego State University, San Diego, CA 92182-4493, USA

7.1   INTRODUCTION

Imagine the following scenarios:

(1) You are on holiday driving on a country road somewhere in central Texas. As you come through a small town, you say to your passenger: 'You know, I've never been here before, but what I saw for the last couple of miles looked familiar.' Then, you give a voice command to the vehicle's navigation system: 'Match last ten miles to similar locations outside of Texas!' The system responds with a list of five counties, one of which you recognize: 'Ah yes, that's where I spent the summer of '93.'

(2) You are a human geographer interested in finding out about differences in how men's and women's life experiences are shaped by their daily spatial routine. You know that researchers have used positional tracking with GPS (Global Positioning System) to capture and compare space–time paths. However, you would like to have a convenient way of directly comparing paths captured in *different* cities. Luckily, a major GIS (geographic information system) software company has just released a product that allows such holistic comparison, generating a display of GPS tracks that looks a lot like a map, but is not based on geographic space. The direct visual comparison of male and female tracks will not only inform your research conclusions, but will also

make for great poster presentations at professional meetings and may even catch the eye of policy makers.

(3) You are teaching an introductory college course in urban geography. You want to give your students some experiential sense of the structure of the city of New Orleans. A popular approach for doing this is to take students on an inexpensive city tour using public transport. Ideally, you would like to take a bus that starts at the Mississippi river, and runs northward, crossing the Irish Channel, the Garden District, and Central City in quick succession. These are three adjacent, yet racially and economically extremely diverse areas, illustrating the socio-economic patchwork that is typical for this city. There is just one problem: you and your students are in Philadelphia! To look for a solution, you start out with a 'normal' geographic map display of New Orleans and draw a line following your chosen path. Then you turn to the same GIS software product mentioned in the previous scenario. It generates a single 'map' showing both your New Orleans path and all the bus lines running in Philadelphia. From this map, you chose the bus path that provides the best visual match. For a more authentic New Orleans experience, you have the heat turned up in the bus, even in the middle of summer. As the bus drives through New Orleans in Philadelphia, you make sure to point out not only similarities but also differences between the two.

In all of these scenarios, one recognizes elements of contemporary geographic inquiry and one can imagine certain approaches to partially implement them. However, different methods for locating geographic features and performing computations on them are here combined in a novel way. The basic premise of this chapter is that as one moves across geographic space, one simultaneously passes through an *n*-dimensional attribute space of the geographic features encountered along the way. It is posited that explicitly visualizing these attribute–time paths (ATPs) as trajectories in a spatialization may be of value in the investigation of moving entities.

First, I will discuss some of the important developments within geographic information science informing this new way of looking at spatio-temporal trajectories. These range from early thoughts about time geography to its recent re-emergence in the context of network accessibility modeling and feminist visualization. On the other hand, these scenarios only sound viable in the context of such computationally intense methods as artificial neural networks, Bayesian networks, or genetic algorithms. These methods are indicative of a growing awareness of a need to deal with high-dimensional attribute data beyond approaches rooted in the data-poor environment of traditional statistical inference (Openshaw, 2000). The chapter argues that great synergistic potential may lie in a combination of time geography with methods designed to deal with high-dimensional attribute spaces. To that end, I first give a brief overview of some related techniques. After outlining a methodology aimed at combining space–time paths (STPs) with self-organizing maps (SOMs), two implementations are discussed and illustrated.

7.2   RELATIONSHIP TO OTHER WORK

The last decade has seen a revived interest in early work on time geography (Häger-strand, 1970; Pred, 1977), which deals with the movement of individuals in space over

time. Hägerstrand and his contemporaries laid out the foundations of time geography with such notions as STPs and prisms, and envisioned a number of interesting applications of these concepts. However, technologies for detailed capture of STPs and their computational modeling were either not yet developed or were missing crucial components. By the early 1990s GIS had developed to a point were many of the database requirements and modeling aspirations of time geography could be supported. Harvey Miller's work on modeling network accessibility with space–time prisms exemplified this (Miller, 1991).

It also became possible to deal with large amounts of disaggregate data, for example travel diaries, including the places of residence, employment, and other activities (Kwan, 2000b). Toward the end of the 1990s, consumer-grade GPS receivers became available that made it feasible to capture detailed paths of individuals. It is not surprising that, at a time when many postmodern and feminist geographers looked upon maps, mapmakers, and mapmaking technology with great suspicion, similar criticism was extended to the integration of GIS and GPS in the implementation of time geography. Partly designed as constructive response to rightful social critique of unquestioned use of geospatial technology, a growing number of geographers have in recent years advanced geographic information science by actively engaging it from within, mostly under the heading of participatory GIS. In the context of time geography, Mei-Po Kwan's work on the development of 'feminist visualization' has been particularly significant (Kwan, 2000a, 2002), and is quite compatible with the methodology described later in this chapter.

Evidence for the resurgence of time geography can also be found in the evolution of the concept of 'geospatial lifelines' towards real-world application (Sinha and Mark, 2005). As technology for capturing geographic location moves beyond dedicated devices (i.e. GPS receivers) towards ubiquity (e.g. in mobile phones), STPs will likely become an integral part of location-based services (Mountain and Raper, 2001).

Apart from the ability to capture STPs, the scenarios described earlier make both overt and implicit reference to a capacity to assess *similarity* of STPs. The type of similarity referred to here is not based on low-dimensional, geometric characteristics, like shape. Instead, the focus is shifted to the attributes of geographic features. Most efforts at modeling similarity are purely computational (as opposed to involving a visualization component) and restricted to the spatial domain, with the temporal domain only gaining prominence recently (Yuan, 2001). It is still rare to see the attribute domain explicitly considered. Indeed, while one would expect 'multidimensional' modeling to include the added dimensionality of attributes, it typically refers to the combination of three spatial dimensions and one temporal dimension (Raper, 2000). In the context of this chapter, the most important observation is that STPs have rarely been linked to representations of the attribute domain, even within the growing area of geographic data mining and knowledge discovery (Miller and Han, 2001).

When looking for visual representations of the attribute domain, the SOM is an obvious candidate. Most implementations of SOM trajectories involve objects whose attributes are changing and are therefore changing position with respect to a SOM in which each neuron has a fixed set of weights, one for each attribute. This has frequently been used in stock market and other financial analysis (Deboeck and Kohonen, 1998; Kohonen, 2001). In the context of spatio-temporal data, this approach has been used to depict counties as trajectories based on multi-year census attributes (Skupin and Hagelman, 2005).

7.3   METHODOLOGY FOR VISUALIZING MOVEMENT IN ATTRIBUTE SPACE

As an STP runs through and past features located in geographic space, it can be conceptualized as simultaneously passing through and past these same features located in an *n*-dimensional attribute space as given by *n* attributes known for each feature. We can refer to the resulting trajectory as an ATP. An STP can be easily displayed in either three-dimensions (within a space–time cube) or two dimensions (when the cube is viewed orthogonally to the two spatial dimensions). However, an ATP cannot be directly displayed, since *n* will typically far exceed the number of available display dimensions. It is proposed here to first spatialize the attribute data and then project the ATP onto the spatialization to form a spatialized attribute–time path (SATP). Figure 7.1 illustrates this schematically with a trajectory traversing an area tessellated by polygonal features. Attributes of these features are spatialized using any suitable dimensionality reduction technique (e.g. SOM, MDS, PCA). Since every attribute has only one value for every polygon, each polygon becomes an individual point object in the spatialization. Polygons that are actually traversed become SATP vertices in the order of traversal (Figure 7.1). Notice how polygons E and G form the beginning and end points of the path, but are actually located relatively close in the spatialization. In other words, the SATP describes a circular route caused by the relative attribute similarity between those two polygon features.
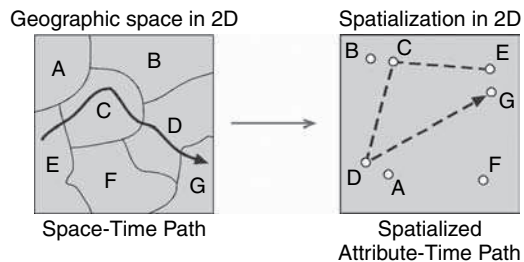
Figure 7.1   *STP transformed into an ATP traversing a spatialization of attributes for polygon features*

When spatializing in two dimensions, the third dimension remains available to represent time, thus forming a spatialized attribute–time cube (SATC), which we will not deal with further in this chapter.

The following describes a specific methodology for implementing SATPs, as pursued in this chapter (Figure 7.2). Spatialization of individual ATPs is based on a single spatialization derived from a large number of geographic objects and their attributes. Choosing the geographic type, extent, and granularity of geographic objects is a crucial first step. Geographic type refers in particular to differences between objects conceptualized as points, lines, or areas. One could even spatialize individual cells or pixels, as provided, for example, by multispectral remote sensing. In this chapter, all examples are based on polygon objects. Specifically, we completely tessellate a given study area via administrative or enumeration areas (i.e. counties, census block groups, etc.), thereby
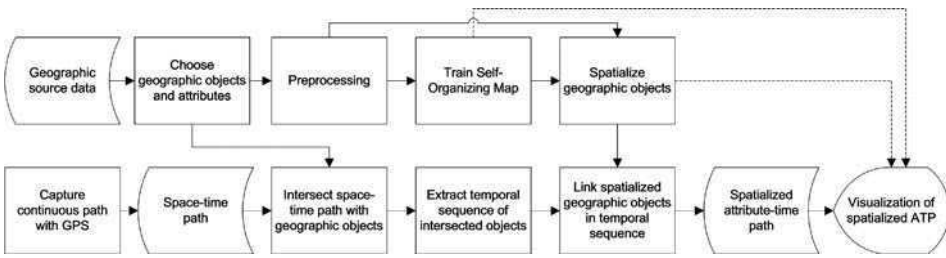
Figure 7.2   *Methodology for creating a SATP using GPS, GIS and SOM*

allowing unequivocal association of path vertices with geographic objects. Point and line objects could of course also be used, within certain proximity constraints. Objects to be spatialized must at least cover the expected extent of STPs, but one may want to go much beyond that in order to allow future paths to be easily spatialized, especially since one of the prime goals of this approach is to facilitate comparison of paths traversing different geographic areas. The granularity or density of geographic objects must be matched against characteristics of the captured paths and the purpose for spatializing them. For example, spatialization of paths based on counties (i.e. STPs spatialized as temporal sequence of counties) may be interesting for regional analysis. However, this would likely be too coarse when one wants to link an STP to the visual experience of someone following it on the ground.

When spatializing geographic objects, it is natural to want to include a great number of attributes, especially in an exploratory setting. Depending on the specific application, one may find it useful to include demographic, economic, or physical attributes. Those choices will often be limited by the actual availability of such attribute data, especially when dealing with a large geographic extent and fine granularity, as discussed above. Socio-demographic data, as published by the US Census Bureau, are a rare exception, with dozens of attributes readily and at little cost available at multiple granularities. That was the main reason for using census attributes in the experiments described in this chapter.

The purpose of preprocessing is to turn raw attribute data into something suitable for neural network training using the SOM method. This may involve, for example, logarithmic transformation for highly skewed distributions and normalization of attribute ranges. After SOM training is completed, the same input data or other data (not illustrated in Figure 7.2) are mapped onto it to derive point coordinates for each input feature.

GPS is a logical choice for capturing STPs. Among dedicated devices, even consumer-grade receivers can now capture quasi-continuous paths with great spatial and temporal resolution. Standard GPS protocols, like NMEA, provide time stamps in Greenwich Mean Time for every observation. GIS overlay can be used to match an STP to the geographic objects encountered. This can be based on an exact or proximal match. After extracting the temporal sequence of objects, their corresponding point locations are found in the spatialization and linked to form a SATP. Various layers could now be displayed within the same two-dimensional geometric space that originated with the SOM. Apart from the SOM and its immediate visual derivatives (e.g. U-matrix, component planes, neuron clustering), one can display the SATP and the point locations of spatialized geographic objects simultaneously or in sequence.

7.4   EXPERIMENT 1: TRAVEL ON INTERSTATE HIGHWAYS

This section describes a first experiment for implementing the methodology laid out in the previous section. Traveling on US Interstate highways, especially in the western states, provides ample time for contemplating the geographic space one traverses. While traveling the United States by car, detailed geographic trajectories were captured by GPS, totaling over 6000 miles in length. The hardware used consisted of a Compaq iPAQ PocketPC paired with a CompactFlash GPS card with external antenna and accompanying software, which stored track coordinates as a text file.

The chosen granularity of geographic base data was at the county level. For each of the 3140 counties, 40 socio-economic variables from the 1990 census were used, with a focus on race, marital status, age structure, and housing characteristics. Then a high-resolution SOM consisting of 10 000 neurons was trained (Figure 7.3). A selection of 12 of the 40 component planes are shown here. As is typical with this form of SOM visualization, one can recognize major relationships between variables and one can also observe how prevalent certain portions of a variable's range are. For example, in the population density variable, few neurons have very high values. However, white population percentage shows high and medium values throughout, except in areas with large black population percentage and especially in SOM areas with a high percentage of households with children headed by a female (i.e. single mothers).
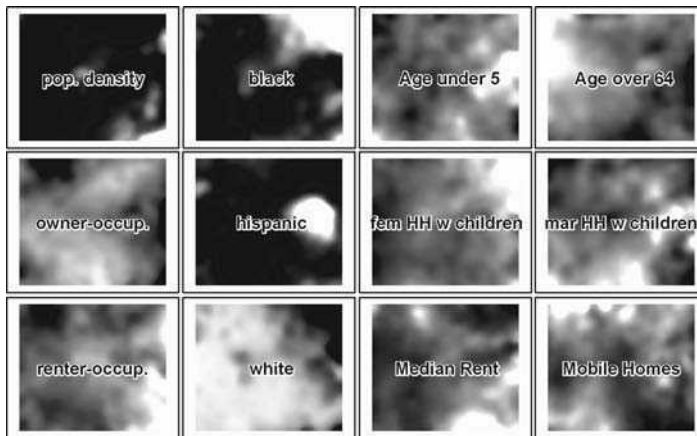
Figure 7.3   *Several component planes from a* 100 × 100 *neuron SOM trained with socio-economic data for all US counties. Lighter shading indicates higher values for a component layer*

A SOM with a relatively large number of neurons allows discerning finer structures in the input space that would be lost to the aggregating effects of a coarser SOM. When *n*-dimensional observations are then mapped onto such a SOM, the resulting two-dimensional locations are spread throughout the finely grained display space. This is advantageous whenever geometric operations on individual objects are desired, for example to place multivariate point symbols or perform selections. Choosing SOM size in this experiment was thus driven by the goal of ideally establishing a unique two-dimensional location for each county. Despite the 3:1 neuron-to-county ratio (10 000

neurons versus 3140 counties), some neurons became associated with multiple counties. To counter this remaining clustering effect, counties were randomly distributed within hexagonal polygons spanned around each node in the SOM (Skupin, 2002). This allows generating unique county coordinates while still maintaining unequivocal links between neurons and counties.

Figure 7.4 shows one of the GPS tracks, in which a drive from Santa Barbara to New Orleans via San Francisco was documented with more than 25 000 vertices. GPS tracks were overlaid with county maps to produce a sequence of traversed counties and spatialized on the basis of that sequence (Figure 7.5).
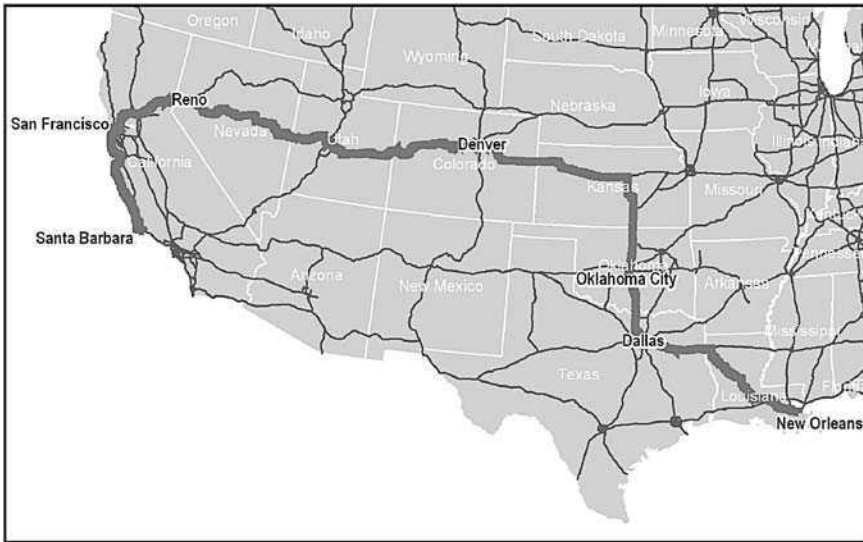
Figure 7.4   *Experiment 1: Overview Map. Traveling from Santa Barbara to New Orleans, a track consisting of* 25 000+ *vertices was captured with a GPS receiver*

As different as such cities as San Francisco and New Orleans might be and as far apart in geographic coordinate space they are, when arriving at one of these from the other, one realizes that – relative to the rest of the country as expressed by the involved attributes – one is back to where one started! The proposed method allows to spell this out, albeit visually, with the two cities appearing as neighbors in the SOM (lower right corner in main map in Figure 7.5). Notice how some geographically close portions of the path correspond to relatively compact portions of attribute space. One such region is entered when crossing from Smith County into Gregg County in Texas (see upper insert map in Figure 7.5). The path only leaves that region when crossing from St Charles Parish into Jefferson Parish (not labeled here), just outside New Orleans.

Time stamps provided by GPS allow mapping the amount of time spent at certain locations, indicated here through graduated circles (Figure 7.6). Despite traversing huge portions of a very large country, the resulting visualization indicates that most time, and presumably money, was spent in a limited portion of attribute space (compare also Figure 7.3).

Figure 7.5   *STP for travel from Santa Barbara to New Orleans projected onto SOM of 3140 counties*

Figure 7.6   *Visualization of time spent in each county during a multi-day drive from Santa Barbara to New Orleans*

7.5   EXPERIMENT 2: JOURNEY TO WORK

One major goal driving the notion of ATPs and their spatialization has been to allow exploring possible links between the *experience* of geographic space and the attributes of geographic features encountered along a trajectory. Ultimately, one would like to see (in

a spatialization) trajectories that readily evoke the notion of traveling through attribute space. However, as one travels across geographic space, one should be able to experience patterns in attribute space as corresponding patterns in geographic space. County-level granularity combined with movement on the Interstate highway system (see Section 7.4) does not really allow this, owing to the large size of counties and the homogenizing effects of Interstate highway routing.

For the second experiment, much finer granularity and shorter, urban paths were chosen. Census block groups, which typically contain around 500 persons, provide that fine granularity, yet their geometry and census attributes are readily available for the whole country, which allows keeping the geographic extent at the national level. The census data used here contained 208 671 block groups from the 2000 census, together with 31 socio-demographic attributes (Table 7.1). Because many of the raw attributes were to be divided by either population size or household size, those block groups containing no population or no households were removed, yielding a final input data set

Table 7.1   *Experiment 2: Variables for* 200 000+ *census block groups used as input to SOM training*

	Variable	Normalized by
1	Population size	Area
2	White population	Population size
3	Black	Population size
4	American Indian / Eskimo	Population size
5	Asian	Population size
6	Hawaiian / Pacific Islander	Population size
7	Other	Population size
8	Multi-race	Population size
9	Hispanic	Population size
10	Males	Population size
11	Females	Population size
12	Age <5	Population size
13	Age 5–17	Population size
14	Age 18–21	Population size
15	Age 22–29	Population size
16	Age 30–39	Population size
17	Age 40–49	Population size
18	Age 50–64	Population size
19	Age ≥65	Population size
20	Median age	—
21	Average household size	—
22	Households with 1 male	Households
23	Households with 1 female	Households
24	Households married with children	Households
25	Households married without children	Households
26	Male head of household with children	Households
27	Female head of household with children	Households
28	Average family size	—
29	Vacant housing units	Housing units
30	Owner-occupied housing unit	Housing units
31	Renter-occupied housing unit	Housing units
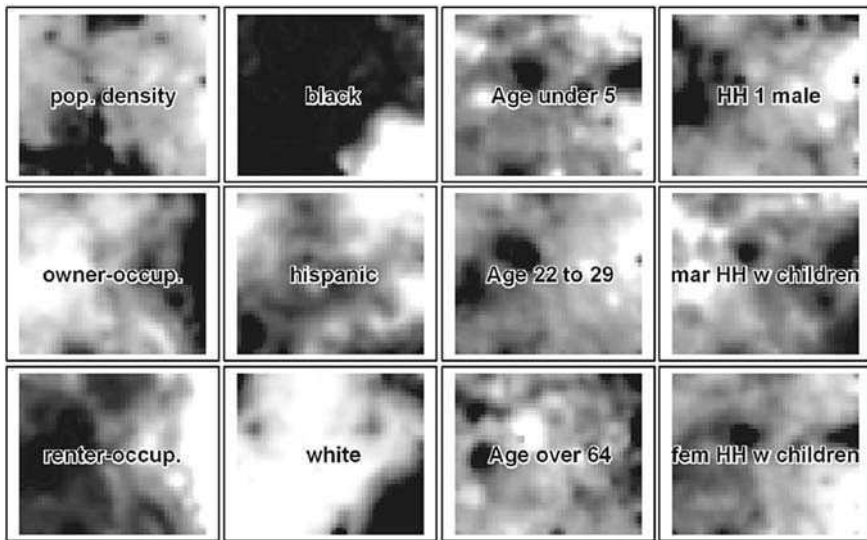
Figure 7.7   *Several component planes of a SOM trained with socio-economic attributes for 200 000+ US census block groups. Lighter shading indicates higher values*

of 207 933 block groups. Some highly skewed variables were logarithmically scaled and all variables eventually fitted into a 0–1 range. Given the large number of block groups and the goal of creating a point location for each of them (as discussed in Section 7.4), a SOM consisting of 250 000 (500×500) neurons was created (Figure 7.7). Training took 92.5 h (wall clock time) on a 2.8 GHz Xeon PC. Mapping of all 207 933 block groups onto the trained SOM took another 123 min.

Recent implementations of time-geography concepts have generally focused on urban environments, with travel on city streets. In deciding on a specific type of path to be captured, inspiration was drawn from the kind of socially critical analysis pursued by Mei-Po Kwan (Kwan, 2002). Journey-to-work paths are a particularly worthwhile subject of inquiry, since the vast majority of employed persons have to travel a certain distance from their residence to the place of employment. Differences in the mode, duration, and routing of these paths provide an interesting subject of study, reflecting society's organization along lines of gender, race, age, and other factors. Travel mode, duration, and routing are of course interrelated, as already noted by Hägerstrand: '. . . the car-owner, because of his random access to transport, has much greater freedom to combine distant bundles than the person who has to walk or travel by public transportation' (Hägerstrand, 1970). When pursuing the quickest route to work, private vehicles will tend to provide a more straightforward path and shorter overall travel time than public transport, at least in the New Orleans metro area. Perhaps more important with respect to the method proposed in this chapter is that different paths taken between residence and place of employment may entail differences in the geographic environment *experienced* en route.

The author's previous places of residence (Mid-City neighborhood in New Orleans) and employment (University of New Orleans) were chosen as origin and destination, respectively. Journey-to-work paths were captured using GPS on two subsequent mornings.

Photos were also taken along the journey, to later allow juxtaposing visual impression (as one element of en route experience) with attribute space location. On the first day, a private vehicle was taken and the quickest route through the street network followed (from here on referred to as 'private path'). On the next day, public transport with busses of the Regional Transit Authority was utilized and the path captured (from here on referred to as 'public path'). Both tracks were started at approximately the same time of day (Figure 7.8). As expected, the private track was shorter in both space and time, running from Mid-City through the Bayou St John neighborhood, then along City Park and the Mirabeau Gardens neighborhood, reaching the destination within about 18 min. The public path involved taking two buses, one connecting Mid-City with the Central Business District and the edge of the French Quarter, the second bus running first parallel to the Mississippi river and then following a straight northward path, toward Lake Pontchartrain. Following this path took 65 min, including transfers.
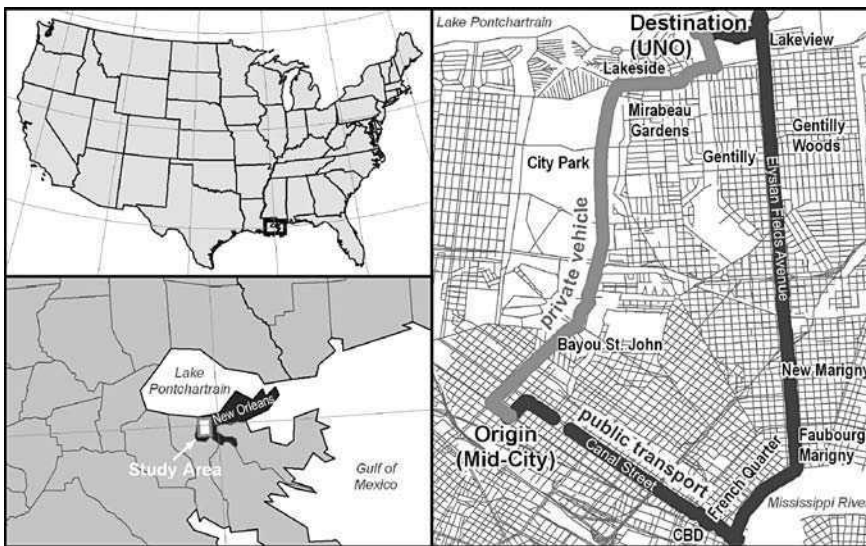
Figure 7.8 *Experiment 2: Overview of study area. Two different journey-to-work paths were collected between origin and destination*

After intersecting the private and public paths with census block groups, the corresponding sequence of block groups was mapped onto the SOM (Figures 7.9–7.11). A total of 31 and 12 different block groups were traversed on the public and private path, respectively. In Figure 7.9, block groups are labeled in the order of traversal. The origin in Mid-City is labeled '1' for both paths and the final vertex as '37' for the public path and '13' for the private path. Note that a new identification is created every time a census block boundary is crossed. Multiple entries into the same block group are possible, depending on how block groups and paths are shaped. The resulting duplicate labels for some block groups are kept in Figure 7.9, in order to allow tracking of the exact vertex sequence. Figure 7.10 shows both paths together and with respect to the complete two-dimensional SOM space. Finally, Figure 7.11 seeks to identify some of the specific
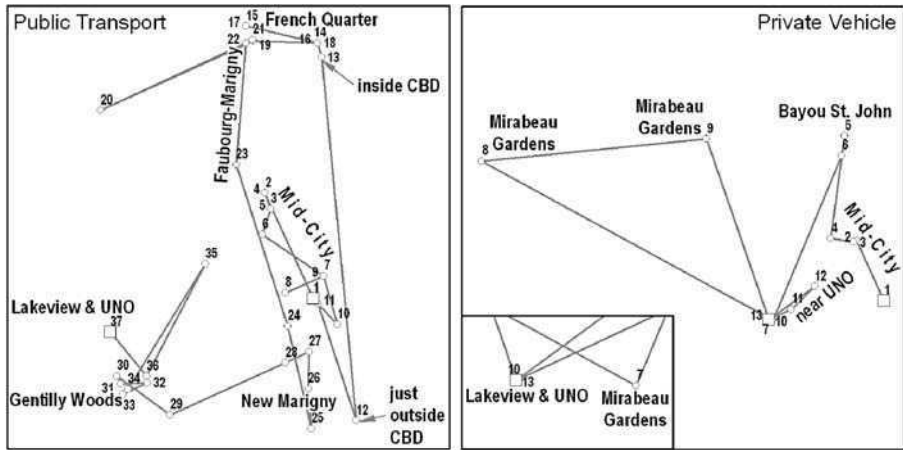
Figure 7.9 *Journey-to-work paths traveled with public transport and private vehicle and visualized on spatialized block groups. Census block groups are labeled in order of traversal. Neighborhoods are also labeled*
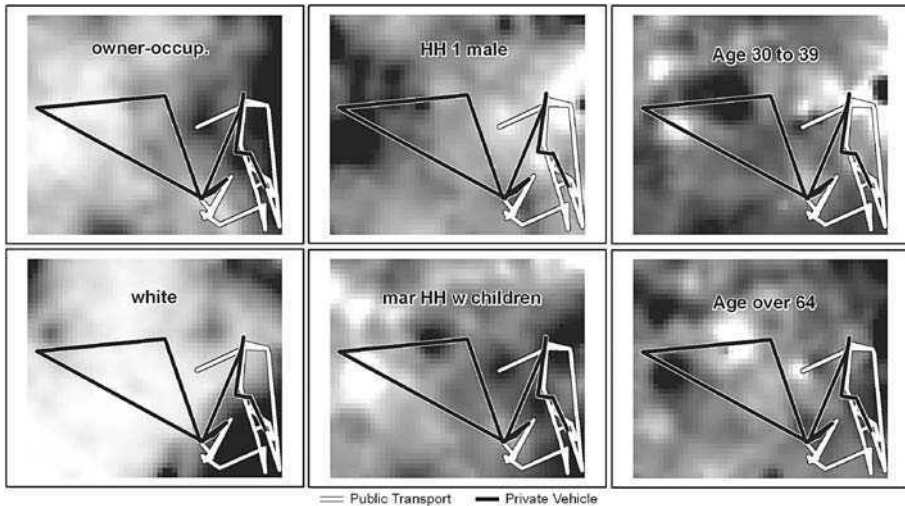
Figure 7.10 *Journey-to-work paths overlaid on a spatialization of census block groups. Lighter shading indicates higher values in component planes*

attribute patterns common to neighborhoods along the public path. It shows the extreme diversity of neighborhoods encountered. Summary statistics for urban counties (as used in the first experiment) tend to hide internal urban heterogeneity. New Orleans, for example, can best be characterized as a patchwork of often extremely different socio-demographic zones. The Mid-City origin of the public path is a bit of an exception, as it is actually quite integrated, thus mirroring a possible summary view of the city. However, once
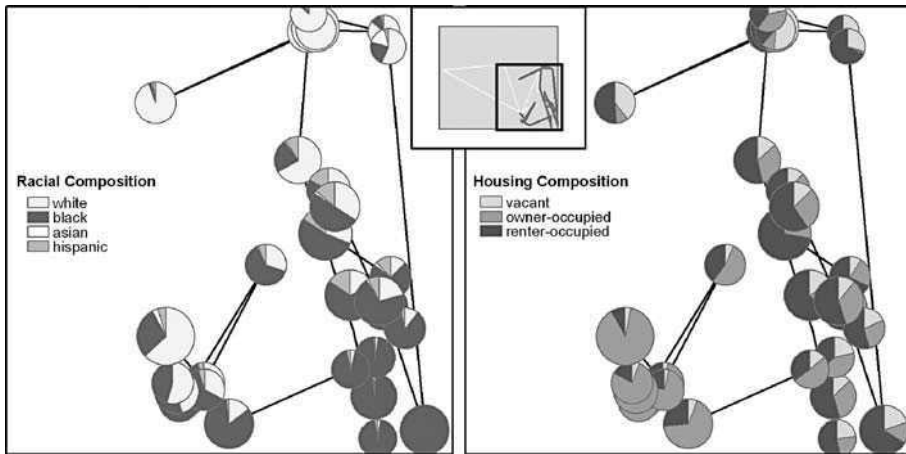
Figure 7.11   *Visualization of attributes of block groups traversed during journey-to-work using public transport*

moving south along Canal Street, the city's extremes become more apparent, at first in terms of gradually increasing percentage of black population. Just before reaching the CBD, this movement towards the extreme lower right corner of the SOM ends in a block group with 100 % black population. Entering the CBD corresponds to a large jump upwards along the SOM's right edge, followed by traversal of block groups on the edge of the French Quarter, and so forth.

What both Figure 7.9 and 7.11 illustrate is that named neighborhoods become manifested as regions in the SOM. For example, along the public path the French Quarter is the region with by far the highest percentage of white population (left portion of Figure 7.11), and large proportions of vacant housing, of households consisting of single males, and of persons in the 30–39 year age range (Figure 7.10). Compare this to Gentilly Woods, which is a middle-class area, with mixed racial composition and mostly owner-occupied housing. Traversing a geographic path means to either move within one of the neighborhoods or to move between them. Moving between named neighborhoods can occur rapidly, as seen when entering the CBD coming from Mid-City (see left portion of Figure 7.9), or it can involve intermediate block groups. Examples for the latter are seen in vertex '23' linking Faubourg-Marigny and New Marigny or vertex '29' between New Marigny and Gentilly.

7.6   CONCLUSIONS

This chapter argues that adding an attribute space representation to the mix of Häger-strand's original ideas with GPS, GIS, and geographic visualization may be an interesting and useful endeavor. While the early examples shown here are meant to illustrate the potential of this approach, they also convey a sense of the issues to be explored in future work. One of these relates to the choice of geographic data with which STPs

are to be matched in order to generate ATPs. While both examples used census data, the methodology accommodates other types of data. For example, when mapping out hiking trails in attribute space, one would want to focus on physical attributes, such as vegetation cover or slope steepness. With the emergence of wireless sensor networks, the on-the-fly 're-routing' of ATPs based on changes in environmental factors (e.g. temperature, humidity) may become a valuable option. Today, hikers may look at Web sites displaying loops of NEXRAD data. Tomorrow, they might also see a looped animation of an SATP, possibly indicating a slow drift towards a danger zone.

For much of this chapter, SATPs were treated (processed, stored, visualized) similar to STPs. Of course, there are important differences that remain to be investigated. For example, with STPs the notion of *bundles* (Hägerstrand, 1970) has tangible, common sense implications. In a bundle, different STPs meet in geographic space for a period of time, the persons associated with them are enabled to directly communicate and interact. Similarly, making a phone call establishes a temporary bundle of trajectories in the virtual space of the phone system. But how are we to interpret a bundle of SATPs? What does it mean when two people moving through different cities are 'meeting' in attribute space? Assuming that a sufficiently rich set of attributes drives the creation of a spatialization, SATP bundles may correspond to similar impressions and experiences. In turn, similar (or different) experiences may become manifested in similar (or different) social attitudes.

Whether or nor these speculations about SATPs hold true remains to be seen. In this context, it may be worthwhile linking ATPs and their spatialized form to the investigation of activity spaces. Similarly, one might ask to what degree such notions as *domains* or *constraints* (e.g. those shaping space–time prisms) are transferable to ATPs, thereby answering recent calls to rethink the concept and implications of individual accessibility in the light of technological advances and societal change (Kwan and Weber, 2003). In approaching any of these issues, a major aim of future spatializations must be to incorporate multiple paths taken by multiple persons in multiple geographic areas, which was not demonstrated in this chapter. Such ability to visually compare paths covering separate study sites would truly demonstrate the usefulness of this method for time geography.

Some might argue that using a SOM for deriving a spatialization from only the non-spatial attributes of geographic features ignores important spatial relationships (e.g. topology, distance in geographic space) that may be very relevant for understanding a given domain. That is a valid argument, whenever such relationships are indeed ignored during training and use of a SOM. This is the case when individual geographic features are visualized as points in a spatialization or when trajectories are generated for features that are spatially fixed, but whose attributes are changing over time (Skupin and Hagelman, 2005). However, the ATPs described in this chapter are different in that neighboring vertices within a path correspond to topologically connected features in geographic space. Therefore, the length of a line segment in the spatialization gives some indication of spatial autocorrelation. Due to the distortion of n-dimensional proximities, this is only a rough approximation and quite dependent on the exact parameters of the spatialization method. The exact nature of the relationship between spatial autocorrelation and proximity in a spatialization is an interesting subject for future research.

REFERENCES

Deboeck, G. and T. Kohonen, eds. 1998. *Visual Explorations in Finance with Self-Organizing Maps*. London: Springer.

Hägerstrand, T. 1970. What About People in Regional Science? *Papers of the Regional Science Association* 24 (7–21).

Kohonen, T. 2001. *Self-Organizing Maps*. Berlin: Springer-Verlag.

Kwan, M.-P. 2000a. Gender Differences in Space–Time Constraints. *Area* 32 (2):145–156.

Kwan, M.-P. 2000b. Interactive Geovisualization of Activity–Travel Patterns Using Three-Dimensional Geographical Information Systems: a Methodological Exploration with a Large Data Set. *Transportation Research Part C* 8:185–203.

Kwan, M.-P. 2002. Feminist Visualization: Re-envisioning GIS as a Method in Feminist Geographic Research. *Annals of the Association of American Geographers* 92 (4):645–661.

Kwan, M.-P. and J. Weber. 2003. Individual Accessibility Revisited: Implications for Geographical Analysis in the Twenty-first Century. *Geographical Analysis* 35 (4):341–353.

Miller, H. 1991. Modelling Accessibility Using Space–Time Prism Concepts within Geographic Information Systems. *International Journal of Geographical Information Systems* 5:287–303.

Miller, H. J. and J. Han, eds. 2001. *Geographic Data Mining and Knowledge Discovery, Research Monographs in Geographic Information Systems*. London and New York: Taylor and Francis.

Mountain, D. and J. Raper. 2001. Modelling Human Spatio-Temporal Behaviour: a Challenge for Location-Based Services. *Geocomputation 2001*, 24–26 September University of Queensland, Brisbane.

Openshaw, S. 2000. GeoComputation. In *GeoComputation*, eds. S. Openshaw and R. J. Abrahart, 1–31. London and New York: Taylor and Francis.

Pred, A. 1977. The Choreography of Existence: Comments on Hägerstrand's Time Geography and Its Usefulness. *Economic Geography* 53 (2):207–221.

Raper, J. 2000. *Multidimensional Geographic Information Science*. London and New York: Taylor and Francis.

Sinha, G. and D. M. Mark. 2005. Measuring Similarity Between Geospatial Lifelines in Studies of Environmental Health. *Journal of Geographical Systems* 7 (1):117–136.

Skupin, A. 2002. A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications* 22 (1):50–58.

Skupin, A. and R. Hagelman. 2005. Visualizing Demographic Trajectories with Self-Organizing Maps. *GeoInformatica* 9 (2):159–179.

Yuan, M. 2001. Representing Complex Geographic Phenomena in GIS. *Cartography and Geographic Information Science* 28 (2):83–96.

This page intentionally left blank

8

Climate Analysis, Modelling, and Regional Downscaling Using Self-Organizing Maps

Bruce C. Hewitson

Department of Environmental and Geographical Science, University of Cape Town, Rondebosch 7701, South Africa

8.1 INTRODUCTION

8.1.1 Synoptic Climatology

The climate system, a complex interaction of the atmosphere with the terrestrial and ocean sub-systems, operates on a multitude of timescales from seconds to millennia, and spatially from the molecular to global scales. However, in terms of societal experience of the climate system, the synoptic scale is the dominant scale of concern; the scale of the major atmospheric pressure systems that condition the weather experienced on a day-by-day basis. The climate, in this context, is the aggregation of daily synoptic weather. A climatology, the characteristic state of the climate system, has a strict definition of the 30 year mean behaviour of some given parameter. However, in common practice climatologies of different variables will often span a wide range of time periods.

Of particular interest to the climate research community are the dynamics of the climate system, historical variability and trend, and projecting the future evolution of the climate. The latter issue is particularly pertinent on the short term (days to months) for management of societal activities, and on longer time frames in the context of human induced climate change and the resulting impacts on natural and societal systems. For

those activities where society is strongly impacted by the climate, and where resources limit response options, society can be said to be vulnerable – a persistent state for most of the developing nations of the world. Research activity addressing these issues is a major consumer of global computing capacity, with climate simulation being one of the single biggest users of supercomputers, and reflects the intrinsic dependence of society on the climate system.

While the dynamics of instantaneous synoptic fields is well understood, the analysis of the long term collection of synoptic events that make up the climate presents a major challenge, be it analysis of historical data or the output from simulations of climate models; data volumes are very large and multidimensional, there is significant spatial and temporal autocorrelation, yet each weather event is a unique permutation across some continuum of states. Conventionally time–space averaging is often as far as the analysis of such large data sets goes. Beyond this the classical approaches are based on techniques such as Empirical Orthogonal Functions (EOFs), cluster analysis, or other similar (usually linear) approaches in order to generalize and gain insight into the data space.

8.1.2   Self-organizing Maps for Climate Analysis

Self-organizing maps (SOMs) bring a new approach to the analysis of climate data that circumvents many of the shortcomings of more traditional approaches. The first use of SOMs in this manner is likely to have been the application of SOMs to evaluate seasonal cycles in a graduate thesis (Main, 1997). In subsequent years there have been only a few studies employing SOMs in climate work (Cavazos, 1999, 2000; Hudson, 1998), and SOMs had little visibility in the climate community. In 2002 Hewitson and Crane (2002) published an article exploring the utility of SOMs in climate studies, showing a number of applications amenable to SOM analysis. Since then the technique has begun to see a broader adoption in a number of climate applications, adding value for a broad range of topics spanning the analysis of present day climate dynamics, interpolation of precipitation data, inference on climate change processes, climate model validation, and development of regional climate change scenarios. The increased adoption of the technique has, in part, been due to the unique attributes of SOMs that compensate for shortcomings in more traditional techniques. In particular the SOM has proven to be exceptionally valuable by allowing a powerful insight into the structure and characteristics of the *n*-dimensional data space. Techniques such as EOFs [or Principal Component Analysis (PCA) in the language of other disciplines] are powerful in reducing the dimensionality of a system, but are problematic to interpret, while imposing a linear filter on the processing of the data. The interpretation difficulty in PCA is that it reduces the data to spatial patterns of variance which are not always readily explainable in terms of physical process. Reusch *et al*. (2005) comprehensively compare the use of PCA and SOMs in an application to climatologically data, and demonstrate how PCA, even with component rotation, can fail to adequately extract the known spatial patterns, while mixing patterns into single components. In contrast, the SOMs-based analyses are shown to be more robust isolating patterns with correct attribution of variance. With PCA, it was difficult, if not impossible, to detect pattern mixing without prior knowledge of the patterns being mixed.

Similarly, cluster analysis, while commonly used and valuable in some respects, suffers from widely variable results simply as a function of the chosen cluster algorithm (e.g. Key and Crane, 1986), and is less conducive to visualization or examining the continuum of data.

While a SOM implicitly allows clustering, the SOM does not in principle cluster the data, but rather finds a representative subset of the continuum as characterized by the source data with the clustering being a post-processing step of mapping data to associated nodes. This attribute of spanning the continuum in particular makes the procedure attractive to climate research. Furthermore, relative to other methodologies, the procedure is intuitive, interpretable in terms of the native characteristics of the data, places no assumptions of linearity, and generates results that are robust across a wide range of generalizations of the data space as the SOM matrix size is changed. More than just facilitating a valuable representation of the data space the SOM can make possible powerful analyses, such as examining the temporal evolution of the n-dimensional system. As with any methodology new to a discipline, adoption of the technique requires some demonstrable improvement over established approaches. Nonetheless, in many cases the simple notion of finding generalized archetypes of the data set, along with the advantages as outlined above, is often enough to convey the concept to new users, and SOM papers and conference presentations in the climate arena are becoming relatively common.

A particular attribute in the SOM analysis of atmospheric fields is the inherent nature of the SOM to span the data space and thus facilitate the visualization of the continuum. With the most dissimilar atmospheric states being located on the most distal nodes of the SOM array, and similar stated on adjacent nodes, the projections of the high-dimensional space onto the SOM node array affords great clarity into the behaviour of the high-dimensional system. A potential limitation of SOMs in this regard is that a regular two-dimensional array of SOM nodes provides four vertices. The SOM characteristically places the most opposing states of the data space onto the vertices, representing the primary modes of the data space. The remaining modes of the atmosphere are then mapped as transitions between these principal vertices. Thus there is a possible constraint with only four vertices; where significantly more than four primary modal states of circulation exist, these would be forced to be located in the transition nodes between the vertices leading to less coherent mapping of the data vectors to the archetype states. With most atmospheric data this does not appear to be too problematic in practice, and univariate fields do appear to form a relatively simple continuum without distinct disjuncture in the data space. However, with other forms of climate data or with multivariate combinations this could be a problem, leading to a reduction in the clarity of archetypes.

Some exploratory work to investigate this problem has been undertaken with a toroidal SOM array, as offered in the SOM Toolbox for Matlab[1] or as used by Ultsch (2003).[2] In this approach the ends of the two-dimensional node array are wrapped to form a continuous surface. This, however, presents significant visualization difficulties, and does not appear to improve the final SOM mapping, at least in this application arena. More useful, although still somewhat problematic for visualization, has been to use n-dimensional arrays of the SOM nodes. A three-dimensional array of nodes would offer

---

[1] See http://www.cis.hut.fi/projects/somtoolbox/
[2] Available online at http://www.mathematik.uni-marburg.de/~databionics/downloads/papers/ultsch03maps.pdf

eight vertices, and a four-dimensional array, 16 vertices, etc. The three-dimensional array with eight vertices readily accommodates the characteristics of most climate data and may still be fairly easily visualized as a set of two-dimensional layers. However, within the climate community this is still exploratory and published results remain focused on the use of two-dimensional node arrays.

For the remainder of this chapter a number of SOM applications with climate data are explored, demonstrating typical use and how the SOM adds valuable insight. Common to each application is the analysis of the continuum of space–time fields of atmospheric variables. In most cases these are gridded fields of, for example, sea level pressure on a time interval of, say, 6 or 12 h. In some cases the data spans decades, in others it may only be one season. In each case the SOM is used with a number of specific objectives in mind, such as:

- Identifying the archetypes representative of the continuum of events.
- Evaluating and comparing the frequency of occurrence and characteristics of the archetypes between two data sets or time periods – in effect comparing the two-dimensional histogram of archetypes as represented by the SOM node array.
- Investigating the time-evolution of the climate system in the reduced dimensionality of the SOM node array.
- Using associations between SOM-derived archetypes and variables at other spatial scales.
- Exploring the characteristics of the data probability distribution function as represented by the SOM.

With the assumption that readers are cognizant of the basic SOM conceptual approach (detailed explanation of the SOM concept is left to Chapter 1), the balance of this chapter now presents examples of how the SOMs can be used to gain insight into the climate system. Five applications are focused on, each presenting a different use of the SOM:

(1) assessing the variability of circulation modes;
(2) temporal trajectories of seasonal evolution;
(3) downscaling local climate response to synoptic scale circulation;
(4) evaluating stationarity of circulation modes;
(5) conditioning spatial interpolation synoptic circulation modes.

8.2   APPLICATION EXAMPLES IN CLIMATE STUDIES

8.2.1   Analysis of Circulation Variability

One of the classic arenas of climate analysis is the sub-discipline of 'synoptic climatology'; defined first by Barry and Perry (1973) as 'obtaining insight into local or regional climates by examining the relationship of weather elements, individually or collectively, to atmospheric circulation processes.' Foundational to this approach is the creation of weather types – generalized characteristic weather states – for relating to other variables, be it precipitation or even more indirect responses to atmospheric forcing such as storm surge or pollution episodes.

In this example we use the SOM to assess the circulation controls on precipitation in central Pennsylvania, USA, following the study by Hewitson and Crane (2002). The source data in this example are gridded fields of January sea level pressure (SLP) spanning 1958–1997, with a data grid every 12 h over a domain spanning the eastern continental USA. Expressing each 12-h grid as a row vector, this creates a matrix of 2480 rows spanning 40 years.

The SOM is first used to assess the archetypal circulation modes for the region. A SOM matrix of 5×7 nodes is established, allowing for 35 archetypes. Different node array sizes were explored but all show the same broad patterns (as shown to be robust in Crane and Hewitson, 2003). The 5×7 array is analogous to using 35 clusters in more traditional methodologies and other synoptic studies over this domain have used a similar number of synoptic types (e.g. Comrie 1992).

The node vectors are intialized with the diagonals related to the first two eigenvectors of the data set, and then trained with the 40-year time series of data. The implementation used the basic freeware package SOM_PAK v3.2.[3] The training parameters were a learning rate of 0.05 (parameter 'alpha' in SOM_PAK), with an initial update radius of 5 (the smaller of the two array dimensions) reducing to 1 by the end of the training. The training ran for 250 000 iterations ($\sim 100 \times$ the number of data vectors), although the SOM converged on a stable solution rapidly. After training each node represents an archetype, with all nodes spanning the circulation continuum. As each node vector is, in fact, a spatial grid, these can be mapped. Figure 8.1 shows the identified January archetype circulation patterns on each node.

Using the trained SOM a wide range of further analyses may now be undertaken, a few of which are discussed here with full details found in Hewitson and Crane (2002). One of the first useful steps is to assess the histogram of circulation by mapping each data grid in the 40-year time series to the SOM, and determining the frequency of occurrence on each node. After undertaking this mapping a number of two-dimensional plots of the node array may be created. For example, Figure 8.2(a) shows the histogram of the full 40 years of data, with Figure 8.2(c) and (d) showing the histograms for low and high precipitation months, and Figure 8.2(b) showing the trend over 40 years of frequency of occurrence on each node.

The analysis gives insight into the historical nature of the circulation that is not readily accessible through traditional methodologies. With Figure 8.1 as a reference, from Figure 8.2(a) it can be clearly seen that January has two nominally preferential modes of the circulation, one characterized by high pressure to the southeast of the domain, and one with a continental high-pressure system with a low pressure to the north-east. However, the frequency distribution across the nodes is not exceptionally varied, with the frequency of occurrence showing an approximately 2:1 range from the most frequent to the least frequent node. Figure 8.2(c) and (d) show the histogram associated with a particularly wet and dry January, respectively, where greater disparity in frequency of occurrence is apparent. The two years differ strongly in the projection on the SOM array, with the dry year showing a strong preferential occurrence of nodes associated with strong high-pressure systems – systems that do not produce much precipitation.

[3] http://www.cis.hut.fi/hynde/lvq/
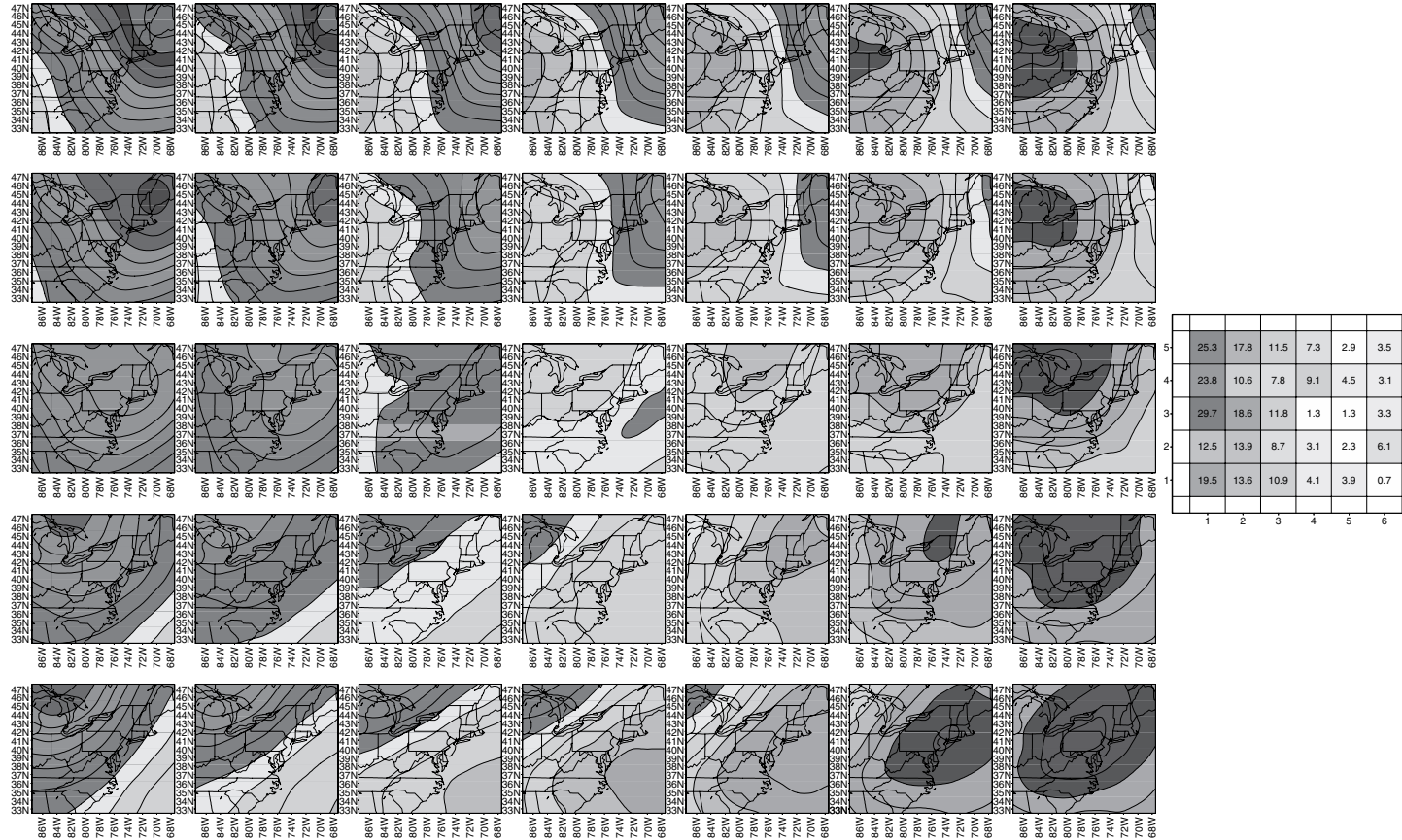
Figure 8.1 *The 5 × 7 array of SOM node vectors of January sea-level pressure (SLP) for the north-east United States. Blues represent relatively low pressure, while reds indicate high pressure. The plot to the right displays the mean precipitation (mm) for each synoptic state represented in the SOM array. (After Hewitson and Crane, 2002) (See Colour Plate 15)*

(a)

	1	2	3	4	5	6	7
5	2.56	2.94	3.52	3.23	3.23	2.40	2.81
4	2.32	2.77	2.94	3.43	3.18	2.61	2.36
3	2.44	2.40	2.61	2.48	2.89	2.61	2.19
2	1.53	2.32	3.10	3.14	3.06	2.52	3.18
1	2.56	2.52	3.43	4.01	3.72	3.31	3.68

(b)

	1	2	3	4	5	6	7
5	0.91	0.71	-0.77	-3.09	-5.04	-3.78	-2.19
4	0.82	-0.05	-0.76	-1.63	-2.65	-2.86	-1.22
3	-0.50	-1.03	-0.69	0.06	-0.93	-0.16	2.19
2	-0.31	-0.53	-0.17	-0.15	-0.36	2.50	6.17
1	1.10	0.90	0.00	-1.06	0.43	4.11	7.33

(c)

	1	2	3	4	5	6	7
5	6.45	1.61	8.06	8.06	3.23	6.45	6.45
4	1.61	1.61	3.23	1.61	0.00	1.61	3.23
3	1.61	1.61	3.23	1.61	4.84	3.23	0.00
2	3.23	0.00	0.00	0.00	8.06	3.23	0.00
1	0.00	0.00	3.23	0.00	3.23	6.45	3.23

(d)

	1	2	3	4	5	6	7
5	0.00	1.61	6.45	3.23	4.84	3.23	8.06
4	0.00	8.06	0.00	3.23	6.45	3.23	1.61
3	6.45	3.23	9.68	6.45	4.84	0.00	4.84
2	0.00	0.00	6.45	1.61	0.00	0.00	0.00
1	3.23	1.61	0.00	1.61	0.00	0.00	0.00

Figure 8.2 *Histograms of the node frequencies of the SOM matrix in Figure 8.1. (a) Frequency of occurrence (%) over 40 years. (b) The trend expressed as the change in mean frequency of occurrence (days) over 40 years. (c, d) Frequency of occurrence (%) for the dry 1978 (c) and wet 1981 (d) Januaries*

Notice also that in each of the two example years there are a number of nodes with zero frequency of occurrence.

From this analysis it is also relatively simple to take all days mapping to a node and determine the mean associated precipitation at a location. In this example this has been done (Figure 8.1) for precipitation of an observing station in the central region of the domain, effectively identifying the precipitation that is characteristically generated by each circulation mode. In this case it is found that the precipitation is not always related to the most frequent modes of circulation.

With this knowledge in hand it is illuminating to consider Figure 8.2(b), the trend in frequency of occurrence. The strongest change over the 40-year period is the positive trend for the nodes in the lower right of the SOM node array. These circulation modes are also characterized by the lowest precipitation. On the basis of this it would be reasonable to infer that over the 40-year period one would anticipate a reduction in January precipitation for the region. However, contradicting this inference is the fact that the January precipitation of the region is actually increasing (not shown) over the time period. The apparent contradiction is readily explained by the fact that the nature of the

precipitation that arises from a given circulation mode is itself changing. Thus, in this case, the low precipitation modes which are increasing in frequency have themselves a positive trend over time in the amount of precipitation received from these modes.

This is a valuable insight that is in line with physical understanding of global warming, where the primary response is anticipated to be a more moist atmosphere, and a strengthening of the mid-latitude high pressure systems – the mode of circulation represented by the increasing frequency modes in the SOM. These results thus add to the ever growing accumulation of indicators for climate change, and are in accord with physical understanding of how this is likely to be manifest.

8.2.2   Time Evolution of the Seasonal Climate

Following the above it is apparent that a logical extension to the analysis of circulation modes is the temporal evolution of the system. In this approach one may use the mapping of each data time slice (in the above example, daily data) and the associated node coordinates to map the trajectory of the climate evolution across the SOM node array – effectively mapping the time evolution and trajectory of the high dimensionality data space in a two-dimensional projection that is more readily interpretable. Main (1997) undertook such an analysis to inter-compare different Global Climate Model (GCM) simulations of the climate system. Hewitson and Crane (2002) extended the basic idea to evaluate the transition matrix of the SOM nodes, while Gutowski *et al*. (2004) followed with a SOM-based diagnosis of, in this case, the somewhat problematic precipitation fields produced by a Regional Climate Model (RCM). This latter application is discussed here.

Precipitation is a derived quantity in GCMs and RCMs; the basic thermodynamic state of the simulated atmosphere is used with relatively crude parameterization schemes to determine the generated precipitation. This is a particular weak attribute of climate models in general. Complicating the already (by necessity) simplistic parameterization issue is the difficulty in evaluating the model precipitation, especially when convective precipitation is a significant component of the total. SOMs offer a particularly valuable means of assessing this, and especially the seasonal development of the precipitation fields that, arguably, should be a primary attribute well simulated by a model if the model is to be deemed credible.

While there are a range of methods for examining the dimensionality space of models (e.g. Govindan *et al*., 2002), Gutowski *et al*. (2004) effectively use SOMs to examine the time series of observed precipitation fields spanning the continental USA compared with the equivalent precipitation derived from a RCM simulation. Using 10 years of monthly precipitation fields the SOM effectively maps the range of seasonal mean precipitation fields (not shown). Following this, the centroid coordinates on the SOM node array for each month are determined, i.e. the centroid of the node coordinates for all Januaries, all Februaries, etc. This gives the projection of mean seasonal position from the high-dimensional data space onto the two-dimensional SOM array. Following the evolution of months maps the mean seasonal trajectory in time of the observed climate system on the one hand, and of the simulated climate on the other.

Figure 8.3(a) shows the respective observed and simulated seasonal cycles across the Sammon map (Kohonen, 1995; Sammon, 1969) of the SOM node array. Apparently, the simulated data diverges strongly from the observed climate during winter months,
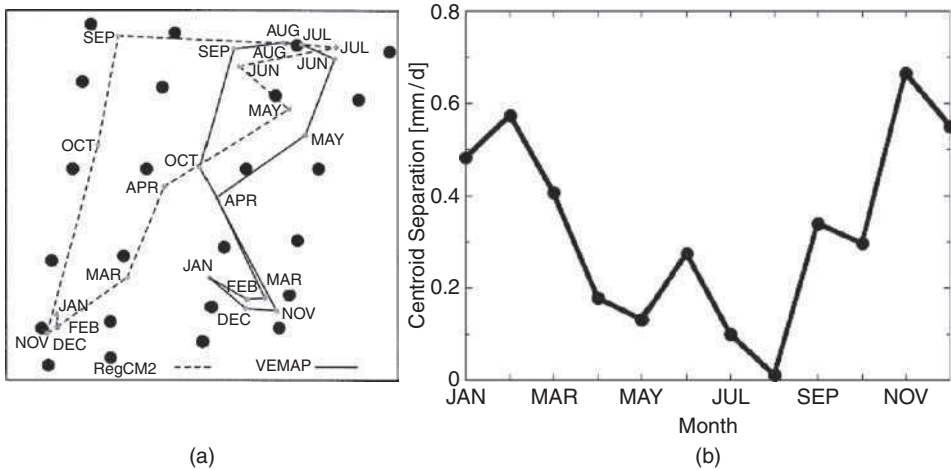
Figure 8.3 *Seasonal evolution across the SOM node array. (a) Sammon map of the SOM node array, and the seasonal trajectory for the observed (VEMAP) and simulated (RegCM2) precipitation fields as defined by the centroid of each month's frequency distribution. (b) Inter-centroid distance between the observed and simulated data. (From Gutowski et al., 2004, Diagnosis and attribution of a seasonal precipitation deficit in a US regional climate simulation. J. Hydrometeorol., 5: 230–242)*

and least during the summer. The divergence between the observed and simulated mean state is shown in Figure 8.3(b). The insight given through mapping the data onto a SOM is valuable, and the degree of divergence between the data sets, and the nature of the divergence in moving through the seasonal cycle, is now readily apparent. Whether viewed in the form of inter-centroid distances, or as a trajectory across the SOM array, this contrasting of the data is not readily achieved through the traditional techniques used by the climate modelling community. For example, the maximum divergence in November identifies a climate seasonal state where the climate model clearly struggles to capture the changing nature of the dynamics as the climate system transitions into the core winter processes. This provides a valuable diagnostic insight for the model developer.

8.2.3   Climate Downscaling

One of the greatest challenges facing climate change research is to translate the knowledge of future climate change at the global scale to how this may manifest at the regional scale of relevance. For example, to say that the globe is warming by, say, 3 °C over the next 100 years is next to useless for a water resource manager in a catchment area, or for developing agricultural response policies in a region, or for assessing future changes in extreme events to which a locality may be vulnerable (e.g. tropical storms). The primary tool for projecting future climate changes is the GCM; unfortunately the GCM has a spatial resolution too coarse to allow for skill in simulating the local scale features of interest. However, the synoptic scale circulation of the atmosphere largely conditions local climate. Hence the concept of downscaling, using cross-scale relationships to infer local climate response from the synoptic scale, has become a standard approach to translate GCM-simulated climate change at the synoptic scale to the spatial resolution of relevance.

Downscaling requires some form of cross-scale function. A regression type function in principle satisfies this; however, this ignores the fact that some of the local climate is a stochastic response only partly conditioned by the synoptic scale, and is not an exact causal response to synoptic scale features alone. Consequently, regression-based approaches to downscaling underestimate the variance.

The SOM provides a valuable approach to compensate for the limitations of regression (and other) downscaling techniques: by characterizing the continuum of synoptic states using observed historical data, one may then derive for each sub-space (each node archetype) of the continuum a probability distribution function of the observed local climate response – a response distribution conditioned by synoptic state. Using the SOM to determine the synoptic scale archetypes, and associating the local response to these provides the means to condition the stochastic response arising from other factors not represented in the continuum of synoptic circulation. To generate a downscaled local climate response one then maps the future climate synoptic circulation state simulated by the GCM to the SOM node array, and for the given node, stochastically sample the associated probability distribution function (PDF) of the local climate response.

In an application focused on Africa, one of the regions of the world most vulnerable to climate change, Hewitson and Crane (2005, 2006) apply a SOM-based downscaling approach to generate scenarios of local climate change. The SOM is initially trained on a multivariate suite of atmospheric variables at the synoptic scale. The variables chosen reflect the thermodynamic and kinematic state of the atmosphere in three dimensions for a 1500 km × 1500 km domain centred on the local target region of interest (a 10 km × 10 km area) – effectively characterizing the synoptic scale state of the atmosphere. After training with 25 years of historical daily data the SOM distinguishes the continuum of states into 35 archetypes (using a 5 × 7 node array). For each node the associated values of daily precipitation are then determined using all the days mapping to the node, and the precipitation values used to construct the PDF response to the synoptic state associated with the node. The approach is very effective in isolating the range of precipitation responses. Figure 8.4 shows the PDF derived from just two of the nodes of the SOM array, nodes on the opposing ends of one diagonal of the SOM node array.

In the downscaling application the GCM simulation data of the future climate is mapped to the SOM trained on the historical data. For each day of the GCM-simulated climate the associated SOM node is identified, and from the PDF associated with this node a random value of precipitation selected. This approach provides temporal sequencing that is consistent with the synoptic scale forcing, while the stochastic element maintains appropriate variance arising from local scale forcing unrelated to the synoptic scale. In practice this is very effective. Figure 8.5(b) shows the climate change projection using the native GCM grid cell precipitation for southern Africa, and showing the effect of the GCM low spatial resolution. Remembering that individual grid cells taken in isolation are not the skilful attribute in a GCM, this is of little value to the regional resource manager, policy developer or impacts researcher. However, Figure 8.5(a) shows results of SOM-based downscaling from the synoptic scale circulation fields to all 10 km × 10 km target locations over South Africa and Namibia. Of note is the high spatial resolution, and significantly, that even if these were area averaged to the GCM grid scale, the results are different.
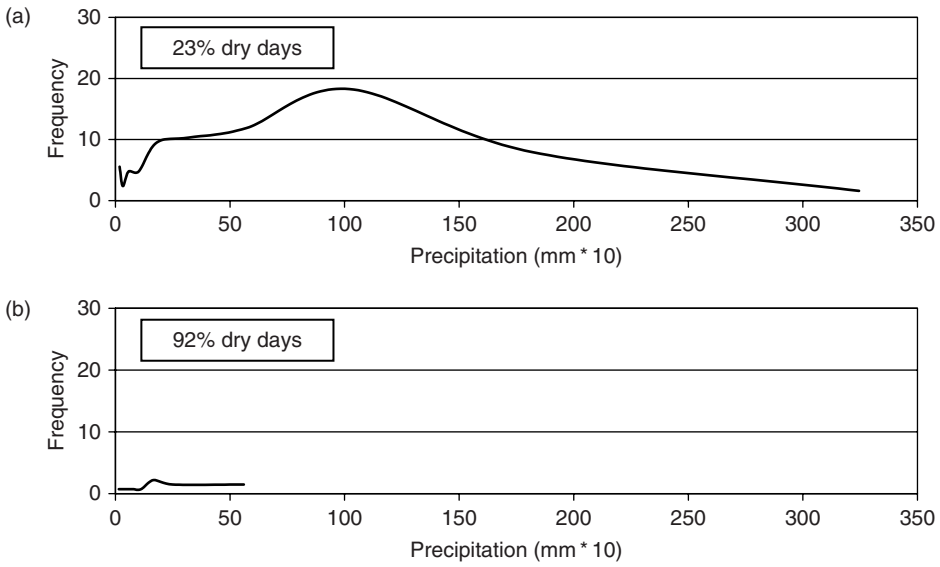
Figure 8.4 *PDF of precipitation associated with two nodes on the SOM array of atmospheric synoptic scale circulation. (a) PDF related to circulation conducive to rainfall. (b) PDF related to circulation leading to dry conditions*
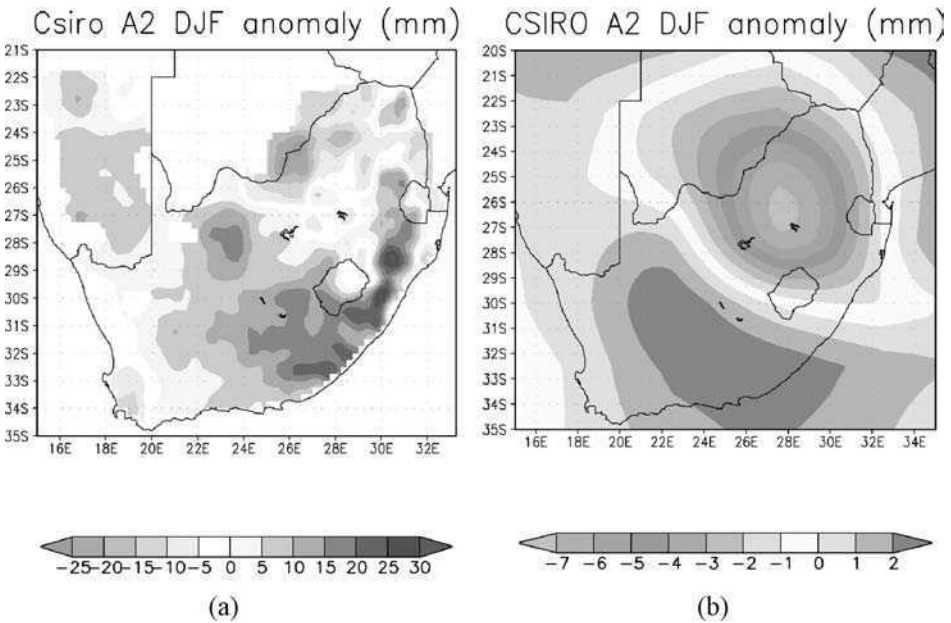
Figure 8.5 *SOM-based downscaling (a) and raw GCM (b) precipitation anomalies of climate change projections for the period 2071–2100 over South Africa (See Colour Plate 16)*

Remembering that the GCM precipitation parameterization is relatively crude and results in a commensurately low-resolution climate change projection, the downscaled solution offers a valuable enhancement. The downscaling is inherently consistent with the foundational thermodynamic and kinematic fields of the GCM in that it is constrained by the real world response to the range of dynamic states, and, as such, presents a more robust and credible solution for the climate change projection.

8.2.4   Stationarity

Based on the example presented in the previous section it appears that there is a fundamental weakness to downscaling: implicit in the use of atmosphere archetypes with related local climate response is the assumption that the relationship between the large scale and local scale in the future remains constant – that the relationship is statistically stationary. This cannot be demonstrated in advance, and raises a degree of uncertainty over the validity of the downscaled data. However, it is arguable that if the full suite of synoptic scale forcing variables are included in the SOM, if the local-scale factors (land use practice, etc.) remain nominally the same, and if the historical data used to train the SOM spans the range of natural variability, then the stationarity question is not a problem. To some degree this is unlikely to be the case.

In the downscaling example used above the variables used effectively encompass the dominant synoptic forcing on precipitation. Further, the local forcing attributes such as topography or land use are either a constant or unknown, and thus already accommodated or else intractable. This leaves for examination the question of whether the cross-scale relationship is stable, in that do the future circulation states fall within the realm of the historical data; in essence, that future climate change is dominantly in terms of changes in frequency of occurrence, persistence, and temporal sequencing of pre-existent events. Assessing the stationarity of the climate forcing thus resolves to assessing whether the n-dimensional envelope of the current climate contains the envelope of the simulated future climate, or at least does so to an acceptable degree.

Using a data set that characterizes the current climate envelope, a simple test of the future climate envelope in relation to the current climate is to determine the percentage of occurrences by which the future climate data exceed the, say, 90th percentile distance from the centroid of the current climate data space. The 90th (or other) percentile is used as one anticipates that some percentage of the historical climate events are unique and not characteristic of the mean climate envelope. However, in n-dimensions this is not a readily tractable task as the shape of the climate envelope is not necessarily some simple spherical shape amenable to defining – in reality this is likely some complex shape.

The SOM offers a valuable solution here; each node in a SOM trained on the current climate represents some finite, relatively homogenous, sub-space of the full data space. Mapping the future climate onto the SOM will map each future climate data vector to some node of the SOM array. Since the node represents a relatively homogenous sub-space of the current climate, the error with which the future climate data maps to this node is a reflection of the degree of agreement with the archetype represented by the node, and which can then be compared with the error of the current climate mapping to the nodes.

To approach the question of stationarity we train a SOM with 30 years of daily data over a 1500 km × 1500 km spatial window for the current climate. For each node we then determine the 90th percentile of the Euclidean distance from the node vector for all days mapping to the node. Then, simulated daily data for the future climate are mapped to the SOM. For a given node the number of occurrences is determined where the Euclidean distance between each future climate day mapping to the node and the node vector exceeds the 90th percentile distance of the days in the control climate. This exceedance factor then represents the degree to which the future climate envelope (for this node-defined subspace) can be said to be non-overlapping with the present day climate envelope. Accumulated over all nodes the exceedance factor gives an estimate of the overall non-overlap between the future and present day climate envelopes. Hewitson and Crane (2006) use this approach to assess the stationarity of three global climate model simulations of future climate, and demonstrate a degree of non-stationarity. Even so, it is shown that this does not preclude the validity of the climate projection, but rather puts a conservative constraint on the degree of the climate projection.

Using the SOM analysis as outlined above, a high exceedance factor indicates potential stationarity problems for any downscaling – and potential problems for the region concerned! Furthermore, by examining the exceedance factor according to each node the non-stationarity of the future climate can be attributed to particular synoptic states, allowing further diagnosis of the dynamics underlying the change in climate. The above analysis may also be extended to all 1500 km × 1500 km windows over the globe, the results of which can be used to generate a global spatial plot of the non-stationarity of the simulated future climate, alerting the researcher to regions where the downscaling may be tractable, or problematic.

Applying this stationarity test to the future climate simulations from different GCMs gives (encouraging) results showing notable agreement between the GCMs. While there are regionally dependent differences, in general the GCMs indicate future climate stationarity to not be a large problem for much of the mid-latitudes, but there are some regions of the tropics that are cause for concern – unfortunately the tropics contain some of the more vulnerable regions to climate change.

8.2.5 SOM-based Conditional Interpolation

One of the more notable difficulties in any climate analysis is the problem of irregularly spaced point observations. Much of the climate system is monitored through weather stations measuring the basic weather variables such as temperature and precipitation. These observations reflect the local response to the synoptic scale forcing. However, the distribution of stations is highly variable, and their degree of spatial representivity largely unquantified. Traditionally, interpolation techniques such as Kriging (e.g. Biau *et al.*, 1999) or Cressman (1959) are used to spatially interpolate from the station point specific observations, but can additionally introduce extrapolation beyond the magnitude bounds of the observed data. These techniques have inherent a number of assumptions about the data, some of which are highly questionable. For example, interpolation of precipitation in such a manner assumes a continuous response surface; in fact precipitation is a bounded continuum – positive precipitation over some given area but bounded by zero. A second

problematic assumption is that the interpolation parameters are constant in time – again an assumption likely to be wrong.

Hewitson and Crane (2005) investigate these issues in greater depth, but we review here how SOMs are used to manage these difficulties. This example uses station data across South Africa, a region characterized by strong climate gradients with a station presence that is highly variable in space and time; in other words a region highly problematic for interpolation.

The approach to interpolation begins with determining, using SOMs, the time-dependent inter-station interpolation parameters. For each of the $\sim$ 3000 stations a SOM array of nodes is trained using the daily data for the station in question, and also all
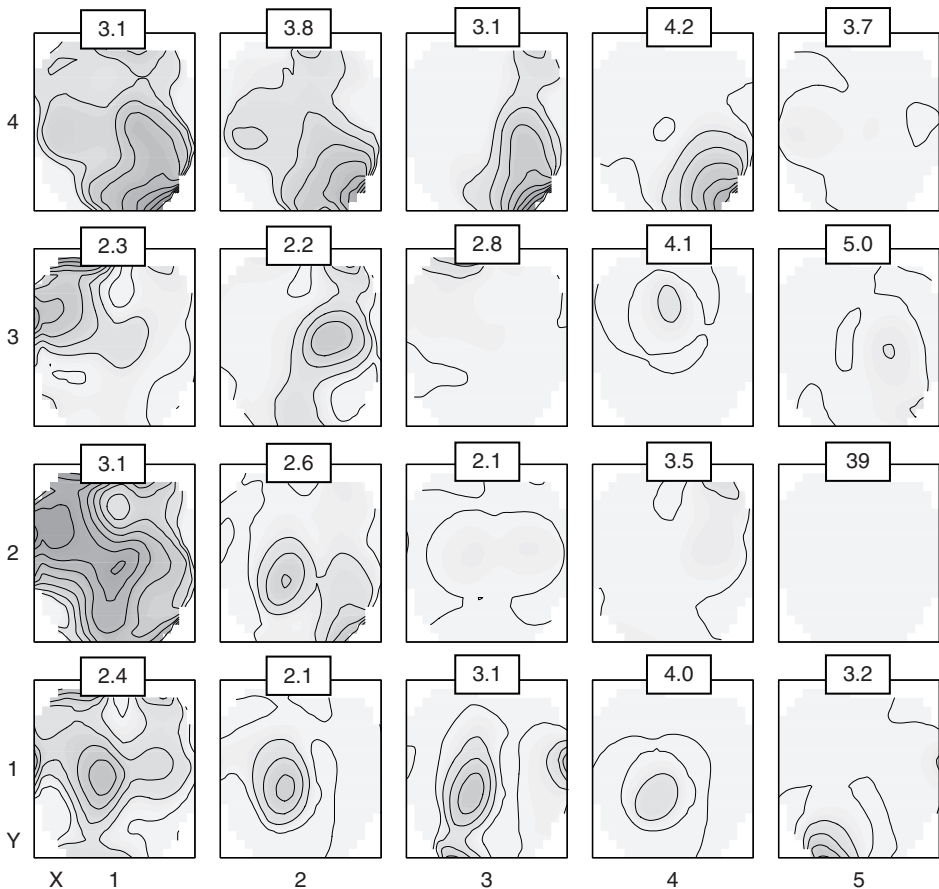
Figure 8.6 *SOM derived archetype precipitation fields in relation to a recording station in the centre of the domain and its neighbours. Darker shades are higher precipitation, and the boxed numbers on each node reflect the frequency of occurrence (%) of days. Node (x = 5, y = 2) is the dry state across the entire region, and hence has a very high frequency of occurrence in comparison with the other nodes. (From Hewitson and Crane, 2005).*

station neighbours within a finite radius (0.75° used). Each day's observations are thus a vector of station precipitation values for all stations within the radius, and each node in the trained SOM reflects an archetypical spatial pattern of precipitation around the target station in question. Figure 8.6 shows the maps of archetype spatial patterns of precipitation as reflected by each SOM node. Of significance here is that the inter-station relationships are highly variable and conditional on the state of the atmosphere.

Using the SOM-based disaggregation of the inter-station relationships of precipitation, subsequent interpolation can be intelligently designed to accommodate these day-to-day changes in inter-relationships. The results when applied to interpolation of the station data onto a regular grid are startling; when comparing the station observation data with the SOM-derived conditionally interpolated data and also the interpolated data from a commonly used interpolation scheme (Cressman, 1959), the conditional interpolation shows large improvements in quality of the interpolated product. For example, considering if the interpolated value is wet when it should be dry, the conditional interpolation overestimates wet states by only 0.8 %, while the Cressman scheme overestimates wet states by 48.4 % largely as a consequence of the inherent assumptions outlined earlier.

The SOM, in this application, thus proves to be very beneficial, notably because it allows the development of interpolation parameters that span the continuum of a very complex inter-station relationship function. The subsequent interpolated results are arguably far closer to reality (which is unknown in its full detail) than any of the other of the commonly used interpolated precipitation data sets in common use.

8.3   CONCLUSIONS

The examples used above express some generic aspects of the application of SOMs in climate analysis. The SOM, however, is finding more and more applications and a full range of other uses of the SOM for climate analysis is not covered here. For example, Crane and Hewitson (2003) use the SOM to define regional boundaries of climate systems, and use this as a basis to scale up from point observations. Tennant and Hewitson (2002) use SOMs to explore the dynamic controls on intra-seasonal variability of precipitation. Meanwhile, Eckert *et al*. (1996) used SOMs to classify members of ensemble forecasts. In this approach multiple forecasts are generated with a climate model, each with slightly different initialization conditions to allow one to explore the envelope of possible climate evolutions inherent in a chaotic system. The SOM is then used to group the individual forecasts (ensemble members) to identify characteristic climate evolution pathways.

Common to these and other applications in climate research is the attractiveness of the SOM in ease of application and accessible visualization of a number of attributes of a complex multidimensional system. In particular, the SOM results seem, at least in climate applications, to be particularly robust in the face of choices over node array dimensions, training iterations (including resilience to over-training), and choice of training parameters. Crane and Hewitson (2003) and Reusch *et al*. (2005) expand on these issues.

Implicit in many of the applications presented here are elements of cluster analysis, but the underlying approach to the analysis is often fundamentally different. In particular is the acknowledgement that the climate system is a continuum, with no categorical states. As such, SOM provides an effective means to reduce the very large data volume

representing the possible state-space of the climate system into readily interpretable modes that represent the continuum.

Frequently the SOM is used as a means for further analysis. In much the same way EOFs are used to perform dimensionality reduction, SOMs achieve the same result with distinct advantages of interpretability, and re-representing the data in a way such as to facilitate a broad range of additional analysis; for example temporal evolution, frequency of occurrence, and of growing importance, the comparison and evaluation of data sets that do not necessarily have a 1:1 correspondence. This latter issue is one of growing importance where climate models are used to simulate the climate behaviour, but where a simulated day in the model may have no direct correspondence to a given day in the observed world. The task then is to assess how, for example, the days of January in the model compare with the days of January in the real world. By mapping the days into the data space represented by a SOM, these forms of assessments can be readily undertaken.

In other forms SOMs present a distinctly new approach to problems, such as the downscaling example above, and especially the issue of stationarity that is becoming a growing concern on a number of fronts in climate change research.

SOMs are thus proving viable, valuable, and are growing in popularity among climatologists. The number of presentations and journal papers including some application of SOMs is increasing, and SOMs are increasingly being taught as part of graduate climate curricula. In short, the future for SOMs as a core methodological tool in climate research looks positive, and a strong increase in usage and visibility in the literature is anticipated in years to come.

REFERENCES

Barry, R.G., Perry, A.H., 1973: Synoptic Climatology: Methods and Applications. Methuen & Co. Ltd, London.

Biau, G., Zorita, E., von Storch, H., Wackernagel, H., 1999: Estimation of precipitation by kriging in the EOF space of the sea level pressure field. J. Clim., 12:1070–1085.

Cavazos, T., 1999: Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. J. Clim., 12:1506–1523.

Cavazos, T., 2000: Using self-organizing maps to investigate extreme climate events: an application to wintertime precipitation in the Balkans. J. Clim., 13:1718–1732.

Comrie, A.C., 1992: A procedure for removing the synoptic climate signal from environmental data. Int. J. Climatol., 12: 177–183.

Crane, R.G. and Hewitson, B.C., 2003: Upscaling of station precipitation records to regional patterns using Self-Organizing Maps (SOMs). Clim. Res., (25): 95–107.

Cressman, G.P., 1959: An operational objective analysis system. Mon. Wea. Rev., 87:367–374.

Eckert, P., Cattani, D., Ambühl, J., 1996: Classification of ensemble forecasts by means of an artificial neural network. Met. Applicat., 3: 169–178.

Govindan, R.B., Vyushin, D., Bunde, A., Brenner, S., Havlin, S., Schellnhube, H., 2002: Global climate models violate scaling of the observed atmospheric variability. Phys. Rev. Lett., 89.

Gutowski, W.J., Otieno, F.O., Arritt, R.W., Takle, E.S., Pan, Z., 2004: Diagnosis and attribution of a seasonal precipitation deficit in a US regional climate simulation. J. Hydrometeorol., 5: 230–242.

Hewitson, B.C., Crane, R.G., 2002: Self organizing maps: application to synoptic climatology. Clim. Res., 22: 13–26.

Hewitson, B.C., Crane, R.G., 2005: Gridded area-averaged daily precipitation via conditional interpolation. J. Clim., 18: 41–57.

Hewitson, B.C., Crane, R.G., 2006: Consensus between GCM climate change projections with empirical downscaling. Int. J. Climatol., 26: 1315–1337.

Hudson, D.A., 1998: Antarctic Sea ice extent, southern hemisphere circulation and South African rainfall. PhD thesis, University of Cape Town.

Key, J., Crane, R.G., 1986: A comparison of synoptic classification schemes based on 'objective' procedures. J. Climat., 6:375.

Kohonen, T., 1995: Self-Organizing Maps. Springer-Verlag, Heidelberg.

Main, J.P.L., 1997: Seasonality of circulation in southern Africa using the Kohonen self organizing map. MS thesis, Department of Environmental and Geographical Science, University of Cape Town, 84 pp.

Reusch, D.B., Alley, R.B., Hewitson, B.C., 2005: Relative performance of Self-Organizing Maps and Principal Component Analysis in pattern extraction from synthetic climatological data. Polar Geogr., 29(3): 188–212.

Sammon, J.W., 1969: A nonlinear mapping for data structure analysis. IEEE Trans. Computers, C-18(5): 401–409.

Tennant, W.J., Hewitson B.C., 2002: Intra-seasonal rainfall characteristics and their importance to the seasonal prediction problem. Int. J. Climatol., 22: 1033–1048.

Ultsch, A. 2003: Maps for the visualization of high-dimensional data spaces. In Proc. WSOM'03, Kyushu, Japan, pp. 225–230.

This page intentionally left blank

9

Prototyping Broad-Scale Climate and Ecosystem Classes by Means of Self-Organising Maps

Jürgen P. Kropp and Hans Joachim Schellnhuber

Potsdam Institute for Climate Impact Research, 14412 Potsdam, PO Box 601203, Germany

9.1 INTRODUCTION

In earth systems science processing and interpretation of large amounts of data has become one of the most important research tasks. Machine learning techniques such as artificial neural networks (ANNs) have several advantages in this context, not only because they are able to replicate the computational power of their biological examples. Other essential attributes are their ability to represent nonlinear relationships, their adaptive capacity if new information is fed in, and their robustness in handling noisy data. On the downside ANNs need large amounts of homogeneous data and the operator has to provide plausible explanations about why they approximate a solution.

   In geosciences, climate, and biogeographical research neural networks can improve the knowledge background fundamentally, in particular for data rich situations. The main ANN applications are pattern recognition, forecasting, or a combination of both. The most common ANNs are supervised algorithms, e.g. the feedforward network (multi-layer perceptron) with backpropagation learning rule. Özesmi and Özesmi (1999) and Aitkenhead *et al*. (2004) apply this network type for regional habitat identification and structuring. A similar approach is used by Moldenhauer and Lüdeke (2002) and Grieger

(2002) for prediction of terrestrial net primary production and reconstruction of paleo-biomes, respectively, by using high resolution climate data as training data. Recently, this network type was also used for risk assessments. Ermini *et al*. (2005), for instance, apply it for an assessment of landslide susceptibility by using lithology, slope angle, and land cover data as training input, whilst Hanewinkel *et al*. (2004) utilise it for the identification of forest stands susceptible to wind throw. ARTMAP networks, which are based on adaptive resonance theory, are a further class of ANNs also used for regional vegetation classification (Carpenter *et al*., 1999). Unlike these widely used ANN types, the self-organising map (SOM) is quite different (for details of the algorithm cf. Chapter 1). SOMs are inspired by their biological example: the brains of mammals. In areas of the brain neocortex the neurons are organised in ways that reflect some physical characteristics of signals stimulating them (Bauer *et al*., 1996). Signals received from adjacent peripheral receptor fields are also processed in neighbouring domains, which is similar to a topology preserving mapping. The SOM extracts in a self-supervised, nonlinear and topology preserving way the essential structural information from numerical data, as opposed to memorising all of it (Kohonen, 2001). It seeks for samples with similar attributes and in an ideal case it forms topologically ordered groups of archetypal pattern by approximation of the density of the data set. These features of SOMs are of specific interest when a classification scheme is not known a priori. SOMs are used by Malmgren and Winter (1999) and Kropp (1999a) to detect regional or global climate zonation schemes, by Foody (1999) for a regional classification of vegetation, whilst Crane and Hewitson (2003) apply them to perform an up-scaling and clustering of station precipitation data to regional patterns. Kropp *et al*. (2006b) have used them to assess vulnerability of German communities to weather extremes.

Here we use SOMs to examine the relationships between climate, soils, and globally distributed vegetation clusters, since Epstein *et al*. (2002) assert that temperature, precipitation, and the moisture regime are major controls over the distribution of vegetation. Nevertheless, the estimation of a necessary and sufficient number of classes and global distribution of ecosystem complexes and climate typologies is still a challenging task, since the majority of approaches differ widely and mainly use a priori assumptions about the class structure and number.

The concept of ecosystem structuring, i.e. biogeography, was founded by Alexander von Humboldt. Early concepts of biographic regions were based on patterns of species (flora and fauna) and higher taxa distribution, although on land ecogeographic clusters are often defined by their dominant vegetation. In practice, the definition of ecosystem complexes is difficult. For instance, can a forest definition be based on a threshold for the percentage of canopy closure, or should only trees be considered as woody vegetation $>5\,\text{m}$ height? Such questions imply that any kind of structuring depends on its purposes, e.g. on specific questions, the size of region observed, and also on the background of the analyst. These difficulties may be one reason why vegetation classifications differ. Very early on, climate was related to these kinds of classification issues. However, static model approaches employing biogeographical classifications are limited with the hypothesis of 'dynamic equilibrium' (climax hypothesis), i.e. they assume a quasi-equilibrium in both climate and vegetation. Nevertheless, it is often mentioned that climate and vegetation interact closely (cf. Bailey, 1995; Eyre, 1963; Jäger, 1997; Mckenzie *et al*., 2003; Simmons, 1979). Thus, the construction of static vegetation

models follows a very simple algorithm in order to assess potential climate impacts: the boundaries of actual natural phenomena are recorded – normally the vegetation – and subsequently related to climatic isolines (e.g. Tivy, 1996; Walter and Breckle, 1991a). An example which is in major use today is the famous Köppen (1918) scheme, which related climate and vegetation based on a finite set of rules, i.e. limit values for monthly means of temperature and precipitation and named these clusters (between 15 and 35 types) regarding to dominant vegetation types. Thornthwaite (1948) developed a classification scheme based on annual pattern of soil-moisture conditions regarded as depending in a sophisticated manner on the monthly input of rain and on the output of evapotranspiration indicated by temperature. In contrast, Holdridge's Life Zone Classification (37 life zones) depicts major types of global climate to terrestrial ecosystems using simple functions of measured bioclimatic variables (Holdridge, 1947).

These approaches have successively been extended including simple ecophysiological rules, but they are still equilibrium models. Examples are the so-called plant functional type (PFT) models. PFT aggregates the variety of species into characteristic functional groups, mostly representing their physiognomic and morphologic features and defines bioclimatic envelopes. One of the earliest of these models was developed by Box (1981), who defined 90 PFTs. Their distribution depends on eight climate variables, for instance, moisture or frost frequency. More recent examples are the so-called BIOME1-2 models developed originally by Prentice *et al.* (1992). The distribution of 17–19 biomes is predicted on a highly resolved spatial scale ($0.5° \times 0.5°$) as a function of environmental constraints, such as cold tolerance, chilling, heating and moisture requirements. These biomes are also based on PFTs. Extended versions of this model type consider competition between plant functional types and biogeochemical fluxes (BIOME3-4, Haxeltine and Prentice, 1996; Kaplan *et al.*, 2003). The latest generation of vegetation models (DGVMs) are dynamic and integrate nutrient cycles and behaviour of vegetation in response to climate (cf. Foley *et al.*, 1996; Friend *et al.*, 1997; Roelandt, 2001; Sitch *et al.*, 2003; Woodward *et al.*, 1998). The coupling with climate models (AGCMs) allows to model vegetation's response to climate change in a transient mode. Criticism is expressed regarding these approaches from distinct sides. Roelandt (2001) mentions that small scale and long-term prognoses rapidly increase computational costs, so that they are rather unsuitable for such tasks. An alternative approach prefers earth system models of intermediate complexity, but biogeographers criticise that these models involve over-simplification (Cramer *et al.*, 2001). Concerning these discussions some questions clearly remain:

- Almost all vegetation models assume a strong response to climatic constraints, but how large is this influence with respect to the formation of vegetation clusters? Is it possible to relate a structuring in climate data to vegetation mappings which implies a local self-controlling of climate and vegetation?
- How many biogeographical units (classes, biomes, PFTs, etc.) are sufficient for the discussion of an ecogeographical organisation on the global scale? The current classifications are a mix between heuristic and empirical knowledge. Also, the most recent models discuss this question only in passing.
- Finally, what are the implications of an ANN model for future vegetation modelling and how can it be used to analyse the impacts of regional global warming?

Using high-resolution climate data as training data, we have tackled these questions with SOMs. The analytical approach presented also involves an algorithm providing a quantitative measure for the topological ordering (SOMTOP model). The topology reveals an additional level of information, which is highly relevant for neighbourhood sensitive classification tasks. This approach supplies a number of objectified classes as well as an optimal dimension for data representation by identifying inherent feature types hidden in the training data. The clusters obtained are related to the broad-scale distribution of ecosystem complexes and are used to estimate the regional impact of climate change.

We start with a description of the methodology in Section 9.2. Data exploration and statistical measures valuable to assess the quality of obtained results are reported in Section 9.3. The results and implications regarding their links to biome types are presented in Section 9.4. Subsequently we highlight the ANN results in the scope of climate change. In Section 9.5 we outline out the advantages of ANN in climate and vegetation modelling and suggest further extensions of this approach. We conclude with a summary and a general outlook (Section 9.6).

9.2   METHODOLOGICAL CONCEPT AND DATA

9.2.1   The SOMTOP Neural Network Model

We implement a neural network consisting of: (i) a SOM (SOM-, cf. Chapter 1); and (ii) an algorithm providing a quantitative measure of the distortion of neighbourhoods (-TOP, cf. Section 9.2.1.2). This method together with additional optimisation criteria (cf. Section 9.3) reduces a flood of data to the essential information, a problem which occurs frequently both in geographical investigations as well in technical applications. Although the SOM per se is topology preserving, errors may be caused by the stochastic training process and unsuitable training parameters or grid configurations, particularly if the dimensions of training data and SOM differ. The topographical product is employed to quantify such errors. Especially regarding the form of the SOM arrays several different strategies are proposed with respect to the choice of the grid configuration. Usually rectangular or hexagonal grids are applied. If the dimension is equal for each direction the map is called *d*-cubic. In a hexagonal grid each node has six immediate neighbours, in a rectangular array four. This holds, in particular, for grids with cyclic boundary conditions, e.g. toroid maps. For maps of planar topology the number of adjacent nodes is less at the border of the network. This may increase the probability of topology errors. The choice of the network topology depends mainly on the training data set, e.g. *d*-cubic grids or boundless toroid maps does not favour a specific direction as much as a rectangular one. Thus, for these maps it may be difficult to find a stable orientation in the data space, which implies that a rectangular grid configuration can be advantageous for heterogeneous data. A further point is related to the number of nodes. Some authors find the use of small SOMs problematical (e.g. Ultsch and Mörchen, 2005), since the topology preservation is of little use and the number of nodes is considered as equal to the number of clusters expected to be found in the data set (similar to *k*-means clustering, i.e. the number of classes is defined a priori). However, the analytical strategy of the SOMTOP has quite a different background. The topographical product makes it possible to determine the accuracy of the topology preservation independent of the size

of the SOM grid. Thus, small grids can be used for analytical purposes, and in particular, computing time can be saved.The number of classes will be fixed after several network runs and additional statistical tests (cf. Section 9.3 and the following).

9.2.1.1   Data and learning process

The training data comprises 62 483 items (land cover excluding Antarctica) made up of 37 input variables (dimensions), i.e. for a resolution of $0.5° \times 0.5°$, representing a global climate for the mid-twentieth century. For each grid cell the monthly means of temperature, precipitation and sunshine hours were calculated from historical weather records – obtained from 7000 weather stations worldwide (CLIMATE2.1, W. Cramer, pers and communication, Leemans and Cramer, 1991). The means are associated with the respective grid cell if the empirical data at least cover 5 years during the period between 1930 and 1969 otherwise they are interpolated. The sunshine hours are converted into values of photosynthetically active radiation (PAR). The database is topography sensitive, i.e. the dependence of temperature on elevation is considered by application of an adiabatic lapse rate of 0.6°C/100 m. In order to suppress the influence of different seasonality between the northern and southern hemisphere, the data set for the southern hemisphere is shifted by 6 months. The 37th component of the data set characterises the soil-moisture properties in terms of water storage capacity, based on information from the FAO soil maps (Fᴀᴏ/Uɴᴇsᴄᴏ, 1974). The grid cells are geo-referenced, but this information is not used during the network training. The data are investigated in detail with respect to the questions provided in Section 9.1 by feeding them into SOM networks characterised by different size and network geometry. The SOMs require no further *ex ante* knowledge in contrast to supervised network training [e.g. Grieger (2002) uses climate variables obtained from ECHAM3 runs and data for 11 predefined biomes to train a feedforward ANN in order to relate unknown climate data to biomes].

   The input data manifold V comprises the elements $\nu_i = (v_1 \ v_2 \ldots v_d)^T$, $i \in \{1, \ldots, 62483\}$, where $d = 37$ denotes the dimensionality of the data vectors (input vectors). The different SOMs used employ a set of neurons (nodes) A, which are arranged in a regular network of dimensionality $d' \leq d, d' \in \mathbb{N}$. The synaptic strength of a neuron $j \in A$ is given by its associated reference vector $\omega_j = (w_1 \ w_2 \ldots w_d)^T$. For a more detailed description of the SOM algorithm we refer to Chapter 1 (cf. also Kohonen, 2001). During the SOM's iterative formation process the input data from the continuous input space V is mapped randomly onto the discrete network A (output space):

$$\Phi_{V \to A} : \nu \in V \longmapsto j(\nu) \in A \tag{9.1}$$

Learning can be terminated if certain quality criteria are approached, in accordance with the requirements of the user. Here we have two aims: first, a most representative data description with respect to the mapping error; and secondly, a topology preserving embedding.

9.2.1.2   Measurement of topology distortions

In the course of the learning process there may be topological (neighbourhood) distortions due to the randomness of the training process and the discrepancy between the topology
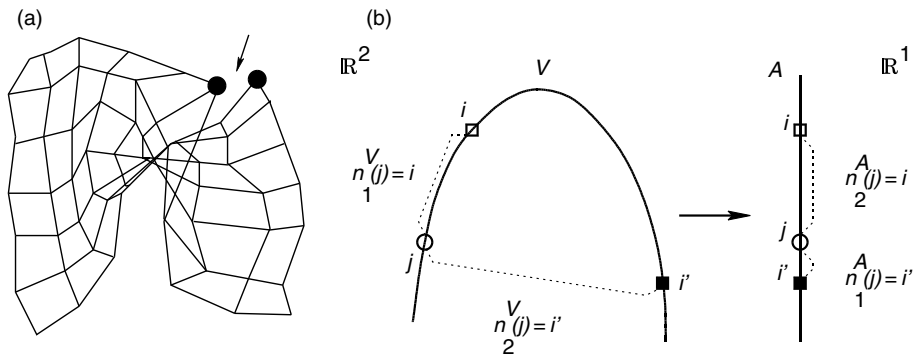
Figure 9.1   *(a) Topology distortion of a two-dimensional map. Due to the twisted map two nodes (black bullets) become neighbours which are clearly separated in a planar network. (b) Measurement of the distances from point j to the next neighbours of order one and two, if the points lying in $\mathbb{R}^2$ are mapped onto $\mathbb{R}^1$*

of the input data and the neural network [cf. Figure 9.1(a)]. Thus, the latter has to be considered when discussing the quality of the results. Typically this context is neglected and only 2-d SOM are applied, which may be related to the problem of visualising high-dimensional SOMs. But the a priori assumption that the training data can be represented by a 2-d SOM can lead to misinterpretations, in particular in cases where topology preservation is important. Climatic zones as well as vegetation are characterised by typical neighbourhood relationships, e.g. the tundra is adjacent to the ice desert but not to tropical rainforests. Consequently, a proper choice of the training parameters and a well-defined SOM array are essential in order to minimise the error caused by the stochastic training process and to achieve a sound result (Ritter and Schulten, 1988). In general the higher the degree of neighbourhood preservation the higher is the accuracy of the mapping. The quantitative measure employed to assess the quality of the topological ordering in the course of the learning process is the so-called topographical product (Bauer and Pawelzik, 1992). It provides a measure of topology distortions in maps between spaces of possibly different dimensionality. According to Bauer and Pawelzik (1992), two distance ratios first have to be defined:

$$Q_1(j, k) = \frac{D^V(\omega_j, \omega_{n_k^A(j)})}{D^V(\omega_j, \omega_{n_k^V(j)})} \tag{9.2}$$

and

$$Q_2(j, k) = \frac{D^A[j, n_k^A(j)]}{D^A[j, n_k^V(j)]} \tag{9.3}$$

where $n_k^V(j)$ and $n_k^A(j)$ denote the kth order (next) neighbour of the point j in the input and output space, respectively [cf. Figure 9.1(b)]. The distance between the points is measured in the input space (D^V) and output space (D^A) by using the node coordinates j and and the reference vectors (ω_j). For an illustration refer to the neighbours of j shown in Figure 9.1(b). In the $\mathbb{R}^2$ it is given by $n_1^V(j) = i$ and in the $\mathbb{R}^1$ by $n_1^A(j) = i'$. For the distance ratio measured in V we obtain $Q_1(j, 1) > 1$, because $D^V(j, i') > D^V(j, i)$. For

the output space A, it follows analogously that $Q_2(j, 1) < 1$, indicating the neighbourhood distortion as a consequence of the mapping from $\mathbb{R}^2$ to $\mathbb{R}^1$. Only when $Q_1 = Q_2 = 1$ do the points in the output and input space coincide and the topology is preserved. In order to suppress effects of local magnification factors the logarithm of the product of the combined ratios [Equations (9.2) and (9.3)] is averaged, i.e.:

$$P = \frac{1}{N(N-1)} \left\{ \sum_{j=1}^{N} \sum_{k=1}^{N-1} \log \left[\prod_{l=1}^{k} Q_1(j, l) Q_2(j, l) \right]^{\frac{1}{2k}} \right\} \qquad (9.4)$$

The topographical product P measures the preservation of the neighbourhood for all orders between the neural units j in the output space A and the weight vectors pointing into the input space V. If $P > 0$ the dimension is too large, if $P < 0$ it is too small and if $|P|$ approaches zero the output space A matches approximately the topology of the training data.

The importance of neighbourhood distortions can be illustrated for a few SOM array configurations with only a small number of nodes. Figure 9.2 shows a cube (dotted line) with a stochastically distributed scatter-plot of 120 points around each corner. Considering the neighbourhood relationships it is possible to distinguish which network type fits the data distribution best. For a linear chain of eight nodes [Figure 9.2(a)] two nodes at the end of the network represent neighboured data, although they are not
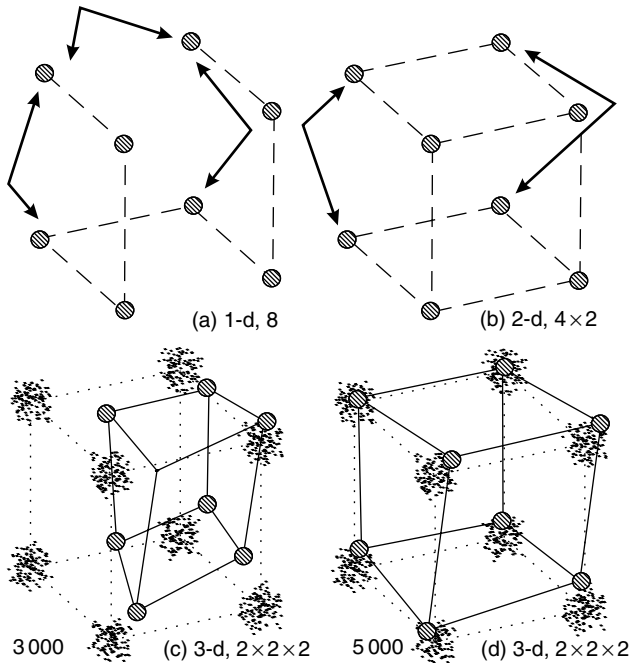
Figure 9.2 *Possible optimum solutions achieved by a 1-d (a) and 2-d (b) network for the scatter-plot shown in (c,d). Neighbourhood distortions are indicated by the arrows. Learning progress of a $2 \times 2 \times 2$ SOM (solid lines) after 3000 and 5000 iterations (cf. Table 9.1)*

adjacent nodes. A 4×2 network fits better than the linear network, but again non-adjacent nodes represent neighboured data clusters [Figure 9.2(b)]. For a linear network (dashed line) and a $2 \times 2 \times 2$ network (solid line) the adaptation process is shown in detail in Figure 9.2(c,d) (solid line). Obviously the 3-d network provides the best representation of the input data. Concerning the $2 \times 2 \times 2 \times 2$ network (not shown), geometric considerations make clear that this is not a suitable solution. The accuracy of mapping can be measured properly by the topographical product P (Table 9.1), which shows that a linear chain, the two- and four-dimensional networks are unsuitable to represent the training data. In the presented approach P is calculated regularly during the convergence phase in order to assess the training results (cf. Section 9.3).

Table 9.1 *Topographical product for the training of certain networks on the described data set for certain network geometries. Five runs carried out for each network configuration. The results show that a $2 \times 2 \times 2$ network is a suitable solution*

d'_A	Network geometry	Topographical product P	Number of iterations
1	8	-0.076945 ± 0.009546	$15\,120 \pm 29$
2	4×2	-0.038619 ± 0.009297	$9\,280 \pm 13$
3	$2 \times 2 \times 2$	0.001960 ± 0.000298	$17\,890 \pm 34$
4	$2 \times 2 \times 2 \times 2$	0.128865 ± 0.069812	$13\,264 \pm 298$

9.3 STATISTICAL TESTING AND SENSITIVITY ANALYSIS

As outlined above, a major goal of the examinations was to ensure optimum categorisation under the precondition of the best data generalisation (approximation of the probability density function). In addition, we want to identify the smallest total number of nodes necessary to best represent the data under topological considerations. As we discuss in detail below, training of SOMs and the calculation of the topographical product and mapping error at intervals are adequate strategies to provide this information, but for large databases and high dimensional data input the computational costs are very high. Therefore preprocessing is necessary to decide the range in which SOMTOP simulations have to be performed. First, we test whether the 37-dimensional data space can be embedded in a space of a clearly smaller dimension. In general, the data vector for each $0.5° \times 0.5°$ grid cell represents a complex climate and soil pattern, implying that it also could include dependent information. Thus, one can assume that the data really lies on a a sub-manifold characterised by $d' \ll d$, which offers several advantages. It is cheaper to store information in lower dimensions, data visualisation is only intuitive in two or three dimensions, and fewer features reduce the complexity of an underlying model (degrees of freedom). But the reduction of dimensions is far from trivial. Initially we employ two methods to estimate an upper and lower border for the embedding space. Principal component analysis (PCA) is a linear strategy to determine the intrinsic dimension of a database. It can provide a clue for the upper border of the embedding dimension (see e.g. Joliffe, 1986). Using PCA yields $d'_{PCA} = 4$ (Eigenvalues: $\lambda = 21.3, \ 8.1, \ 3.3, \ 1.0$, explaining 91% of the overall variance) for the dimension of the embedding space. The PCA result additionally shows a pronounced variance in one

direction, so we employ rectangular SOM grids. For consideration of nonlinear inhomogeneities, scaling analysis is used additionally to determine the capacity dimension (fractal dimension) of the database, which can indicate the lower border. Analysis after Grassberger and Procaccia (1983) yields $d'_{GP} = 1.8$. Considering these results we have to carry out network simulations for a one- to four-dimensional output space. An iterative strategy can save additional computing time. Concerning the empirical and theoretical approaches discussed in Section 9.1, it seems evident that we have to search for a classification scheme within an interval of $[15 \ldots 90]$. Several grid configurations were tested for each dimension and improper networks were excluded from the further analysis.[1]

In total, we performed five runs each for 24 array configurations on a massive parallel computing system (Power 3-based IBM SP-2). The time point for finalisation of the learning process is estimated by calculating both the topographical product and the quantisation error at intervals during the convergence phase, i.e. first-time after $l \cdot 10$ learning steps, where l indicates the number of training vectors. When the difference quotient becomes approximately constant the learning process is terminated. Table 9.2 shows some of the calculations for P making clear that for any number of nodes the 2-d case is topologically the best. Additional statistical tests are applied in order to estimate which total number of nodes provides an adequate classification (cf. Kropp, 1999b), because neither the topographical product nor the quantification error are sufficient criteria to achieve the goals defined above. Therefore, a sensitivity analysis is carried out to compare the solutions found for several network topologies and total number of nodes by examining the scattering of vectors between the classes. By calculating the joint probability that each data vector was sorted into the same class for the different runs performed for a defined grid form one can decide whether the solution is 'stable' or 'unstable'. Assuming that a classification exists in the database, the algorithm identifies the clusters (global minimum) only in the case of an optimum configured network array. Such a strategy produces maps as presented in Figure 9.3. For instance, the outcomes

Table 9.2 *Examples for results of the topographic product for different network geometries (five runs for each network geometry, overall simulations are performed for 24 network configurations) providing an overview of efficiency and precision of this approach*

d'_A	Network geometry	Topographical product P	Optimum solutions
1	16	-0.0223 ± 0.0012	
2	4×4	0.0076 ± 0.0019	$\leftarrow$
3	$4 \times 2 \times 2$	0.0215 ± 0.0031	
4	$2 \times 2 \times 2 \times 2$	0.0628 ± 0.0047	
1	20	-0.0291 ± 0.0025	
2	5×4	-0.0026 ± 0.0013	$\Leftarrow$
3	$5 \times 2 \times 2$	0.0189 ± 0.0023	
1	24	-0.0257 ± 0.0004	
2	6×4	-0.0055 ± 0.0019	$\leftarrow$
3	$4 \times 3 \times 2$	0.0135 ± 0.0021	
4	$3 \times 2 \times 2 \times 2$	0.0418 ± 0.0025	

[1] With the help of the topographical product it is easy to decide which grid configurations are suitable. For instance, within the group of two-dimensional arrays a network of 10×2 nodes has an approximately 'linear' structure what can be measured by P.
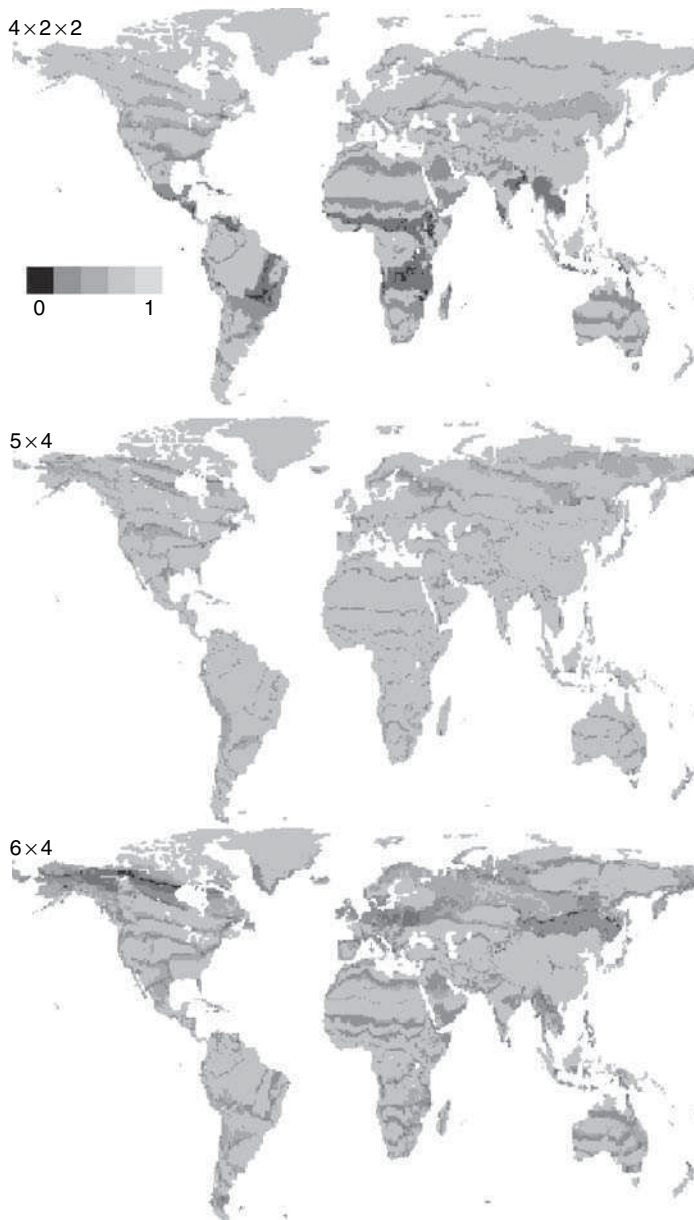
Figure 9.3   *Normalised joint probability that for five runs the data vectors are mapped on the same class (node) calculated for different network geometries. Black indicates no whilst light grey indicates complete coincidence between the learning results. For the 5 × 4 network the transition zones (regions with vector fluctuations) are the smallest*

indicate that $4 \times 2 \times 2$ and 6×4 arrays are inappropriate representations, because the areas showing distinct attributions of data vectors are large. Considering additionally the results of P, strong evidence is provided that a $5 \times 4 = 20$ node network supplies a suitable classification of climate and soil data (Table 9.2 and Figure 9.3). Only in this case the transition zones indicating the class borders (in ecogeography described as intermediate habitats or ecotones) are small. Further examinations are carried out, e.g. the comparison of density distributions in the proximity of the discriminating hyperplanes with equal distributions (cf. Kropp, 1999b) in order to assess the class separation showing also that the 2-d, 5×4 case is the most suitable solution. Therefore, a detailed analysis of the categorisation with respect to the questions outlined in Section 9.1 is carried out on the basis of this network geometry.

9.4   GLOBAL PROTOTYPES OF CLIMATE AND THEIR RELATIONSHIP TO VEGETATION

After the training process we obtain an associated node (class) number for each grid cell. By using the geo-coordinates for each $0.5° \times 0.5°$ cell we get a geographically explicit map of the data categorisation as determined by the SOM, although the geographic reference was not used during the learning process (Figure 9.4). The categories are intuitively well distributed regarding their climatological characteristics. This is supported by the topological ordering on the network (Figure 9.4, inset), i.e. tropical regimes are neighboured by other tropical divisions and not by polar regimes. However, comparing Figure 9.4 with thematic maps representing the vegetation distribution or vegetation models similarities and differences can be observed (for further results, cf. Kropp, 1999b). Thus, a detailed analysis is required. Regarding the number of generalised classes/vegetation types the BIOME model shows the best agreement with the learning results obtained by SOM training. For comparison an output of 19 biomes is used. Statistical tests show the different nature of the two approaches. The BIOME model is driven by ecogeographically motivated functions in which climate plays an important role. The SOMTOP algorithm learns only with respect to the statistical properties of the input data. Therefore the classification achieved by the neural method is characterised by an average variance for each variable and cluster which is approximately 26% smaller than for the corresponding biome types. In the following, similarities and differences between the neural and the process based approach will be examined by the discussion of two examples. The first one is located in the polar zone, which is divided by biogeographers into the biome types 'tundra' and 'ice-desert'. Focusing on the tundra biome (dark grey areas, node 1, in Figure 9.4) empirical investigations have estimated their northern limit along the 2°C July isotherm (Aleksandrova, 1977). The average July temperature of the border cells detected by the SOMTOP model was $3.9 \pm 1.9°C$. The southern borderline of the tundra is placed along the 10°C July isotherm (Walter and Breckle, 1991b). The neural model has fixed the transition zone at the $9.7 \pm 1.6°C$ line. The latter indicates that the results obtained by an autonomous learning algorithm are reasonably close to the empirical values. However, further examinations are necessary regarding the tundra structuring of the BIOME model [Figure 9.5(a)]. In this model, the Tibetan and parts of the Andean highlands as well as the tundra itself are associated with the biome type
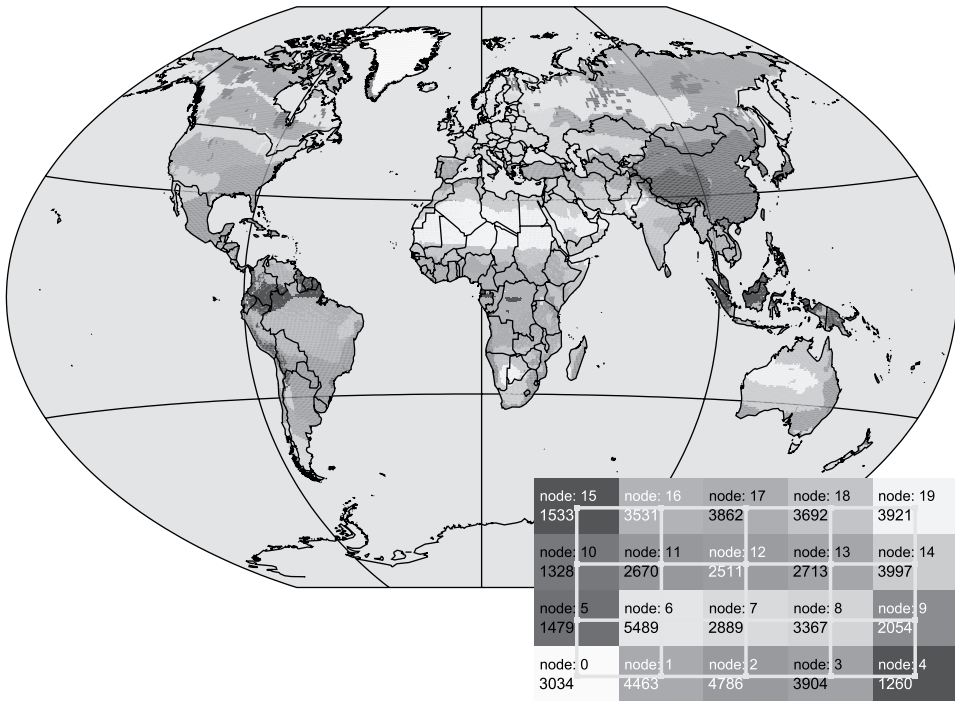
Figure 9.4   *Global distribution of classes (represented by the different nodes) obtained after simulation with the SOMTOP algorithm. The topological arrangement of the classes on the network is shown by the inset. The colour coding used in the inset corresponds to that shown on the map and the node numbers are equivalent to the class numbers used in the text. The additional number indicates the quantity of associated input vectors (See Colour Plate 17)*

'tundra', whereas the SOMTOP model has separated highlands and tundra in two different classes (nodes 1 and 4) [Figure 9.5(a), inset]. But which of these two categorisations is more appropriate with respect to the ecogeographic ordering? From a climatological point of view one can argue that a variety of differences are observable between the tundra regions and the Tibetan and Andean highlands, which the latter characterised by the absence of the polar night, a significantly higher annual precipitation (233 mm vs 492 mm), and a higher annual mean temperature (-13.4 vs -1.5°C). Literature surveys support these facts, e.g. Walter and Breckle (1991b) have mentioned that an adaptation of vegetation to these environmental constraints is probable. In addition, empirical investigations for Tibet indicate that most plants are more closely related to the vegetation of the surrounding areas than to the tundra vegetation (Walter and Breckle, 1991b). This is not surprising, because plants migrated from the neighbouring deserts and semi-deserts to the Tibetan plateau after the last glacial period of the Pleistocene. Thus, according to these arguments and the additional support of the SOMTOP results this structuring should be revised in process models.

The second example focuses on habitats in the tropical regions [Figures 9.4 and 9.5(b)]. Although the SOMTOP model has produced a very good classification reflecting the transition from rainforests via seasonal rainforests and tropical dryforests to savanna and
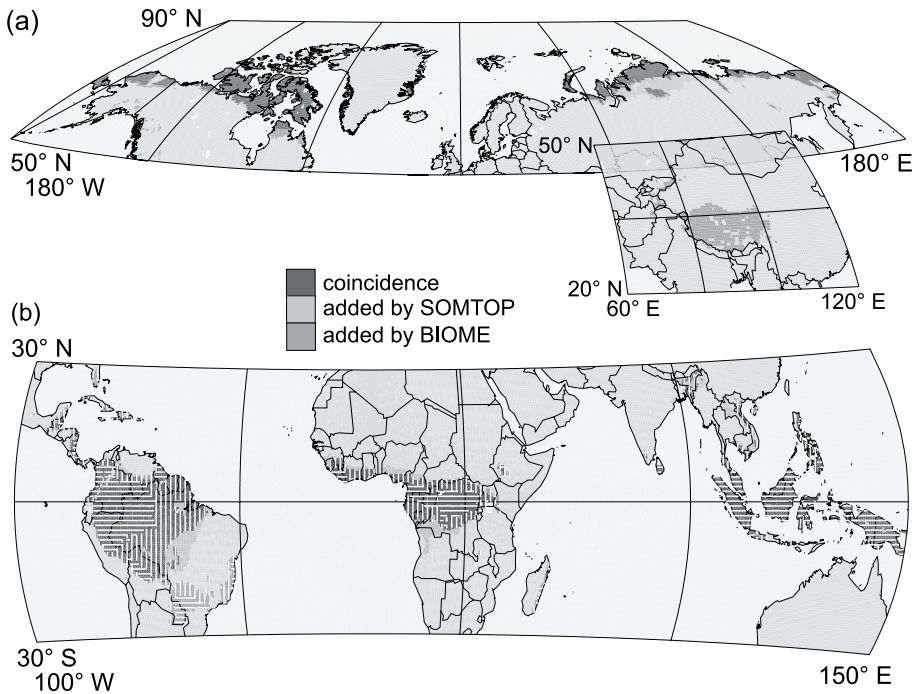
Figure 9.5 *Comparison between SOMTOP categorisations and biome types in the arctic and tropics domain. (a) For class #1 (SOMTOP) and the 'tundra' biome – the inset represents the highlands (node 4), and (b) for the aggregated rainforest division: SOMTOP classes #15 and #16; BIOME 'tropical rainforest' (horizontal hatched) and 'seasonal rainforest' (vertically hatched). Focusing on the single archetypes some larger differences are apparent in South America (regarding the SOMTOP classification in these regions compare with Figure 9.4) (See Colour Plate 18)*

desert, in terms of the quality of separation of the classes it is weaker, as for the previous discussed example. Consequently a more in-depth analysis is required of how well the network has approximated the density distribution of the data and whether there exists ambiguity with respect to no-man's land between the two classes. Additional measures were defined and calculated in order to assess the point density in the boundary areas between neighbouring classes (cf. Figure 9.4, inset; Kropp, 1999b). The results show that for nine of the 31 network edges the defined criteria are not sufficient (for details cf. Kropp, 1999b). Thus, we have to ask whether additional evidence can be drawn from the feature spectra assessing the quality of the class segmentation. For explanation we compare one of these nine cases, i.e. the separation of the two neighboured classes #15 and #16. The results indicate that the SOMTOP approach allows a subtle classification which depends in this case on only a few prominent features [Figure 9.6(a)]. For the two discussed categories which can be related to the vegetation types 'tropical rainforest' (class #15) and 'seasonal rainforest' (class #16) the temperature, radiation pattern, and the value for the water storage capacity of soils are more or less equivalent [Figure 9.6(a)]. But the rainfall patterns show markedly different seasonality [Figure 9.6(a)]. Looking into details the averaged monthly minimum precipitation amounts to 200 mm and the
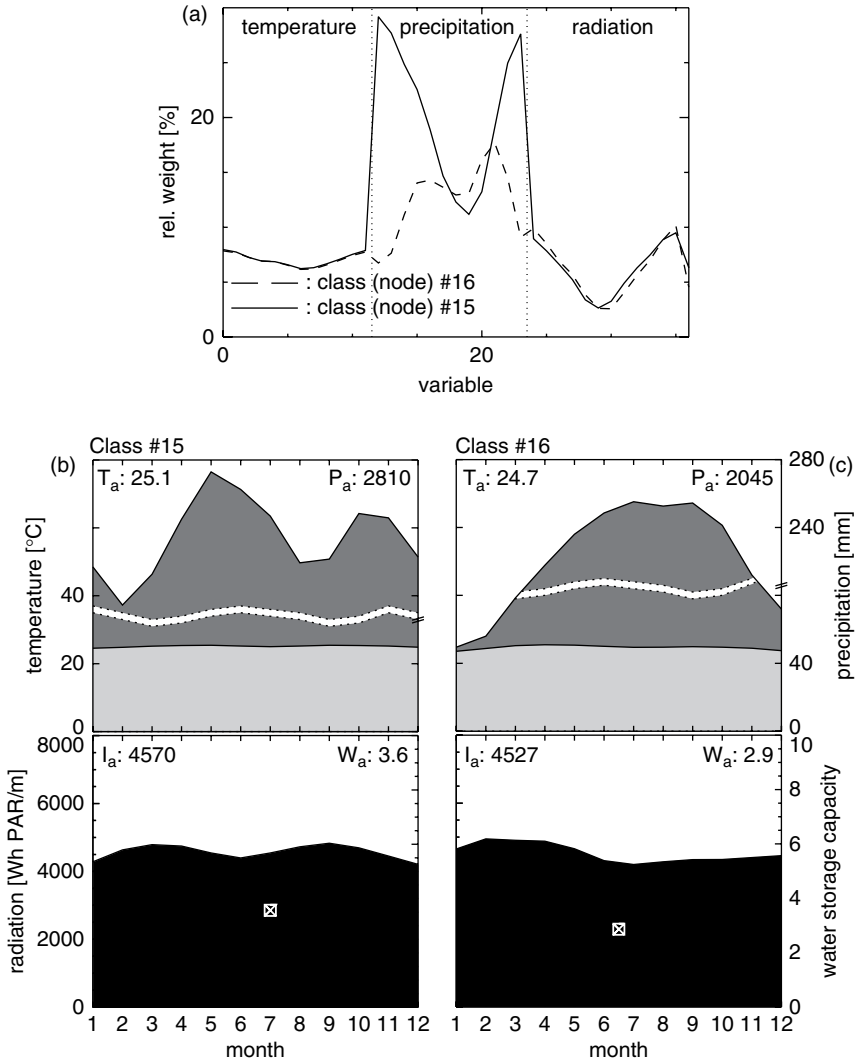
Figure 9.6 *(a) Significance (weight on the representing node vector) of the monthly means for temperature, precipitation, radiation and water storage capacity of soils for classes represented by nodes 15 and 16 (tropical zones); cf. Figure 9.4. (b) Climatological diagram of class #15 and (c) of class #16. Light grey indicates the temperature, medium grey the precipitation, black the insolation, and the boxed cross the water storage capacity of soils (W_a). T_a is the average monthly mean temperature, P_a the averaged annual sum of precipitation, and I_a the average monthly mean radiation for the grid cells associated with the respective classes*

averaged annual precipitation is 2800 mm for class #15 with two peaks in spring and fall [Figure 9.6(b)]. Together with a monthly mean temperature of about 25.1°C these are typical conditions for the existence of tropical rainforests (e.g. Malaysian archipelago, Columbian coast, Amazonian and Congo basins). In comparison class #16 is characterised

by an annual precipitation of about 2000 mm and a monthly mean temperature of 24.7°C, but shows only one rainy season and a short dry season in winter ($P_{min} \leq 50$ mm) [Figure 9.6(c)]. In general, these conditions are sufficient for the existence of rainforests, but the short dry season has the consequence that one can observe so-called 'seasonal rainforests' (for a detailed discussion of the distribution of rainforests, see e.g. Park, 1995; Tivy, 1996; Walter and Breckle, 1991c; Whitmore, 1999). Considering the significance of the input variables on the node vectors [Figure 9.6(a)] the distinction between classes #15 and #16 seems somewhat astonishing, because the SOMTOP approach is able to detect climatic features with respect to their seasonality very sensitively, even though it uses only an integral measure (i.e. Euclidean distance) to distinguish the differences between clusters. Nevertheless the class separation is logical with respect to empirical results.

Exploring the output of the BIOME model for these regions for the aggregated case [Figure 9.5(b)] it is different only with respect to a tiny fraction. Looking into detail, i.e. differentiating between the two classes, the spatial distribution of the biome types 'tropical rainforest' and 'seasonal rainforest' in particular differs to those of classes #15 and #16 mainly in South America and Africa, whereas in SE Asia both are entirely equivalent. These differences are related to an underestimation of the precipitation component, especially its seasonality, by the BIOME model. This has led to an association of dryer regions to the corresponding biomes. In contrast the SOMTOP approach has delivered a clear-cut class separation with respect to this feature (Figure 9.6). This outcome is supported by empirical investigations on climatic constraints for the occurrence of rainforest (see above). Here the SOMTOP approach has provided a more precise classification, matching exactly the climatological limits of rainforest distribution.

9.4.1 Assessing the Impacts of Climate Change

It now seems feasible to predict the impacts of climate change on the basis of the classification scheme obtained above. We have applied ECHAM3 climate change scenarios (IS92a, $2 \times CO_2$), using the anomalies (Cubasch *et al.*, 1996) and calculating the class membership to the climate categories based on the trained network. This examination suggests which regions could be affected the by impacts of climate change, i.e. vegetation change. In comparison with other approaches this strategy has several advantages: (i) the regional impacts can be computed fast and easily; and (ii) other climate variables apart from temperature can also be taken into account in an assessment. We illustrate potential impacts again for the example of the Tibetan highlands. Figure 9.7 shows that approximately 20 % of the highlands change the class membership (hatched areas). Two-thirds of these regions become more arid and therefore are associated with the Central Asian deserts (Figure 9.4, node 9) (mean temperature increases about 4.2°C, but precipitation increases only about 5.9 mm). These regions are located in the north and north-western parts of the Tibetan plateau. One-third of the highlands, mainly located at the eastern margins, becomes warmer and more humid and is associated with class #10 (cf. Figure 9.4). Here the temperature increases about 3.5°C, but the precipitation by about 57.5 mm. Figure 9.4 shows that the affected regions vary in width between 50 km and 200 km. Comparing this with the impact in other regions (e.g. arctic zones) this value can expand to 300–500 km (Kropp, 1999b). This implies – under the assumption of a linear change – that vegetation
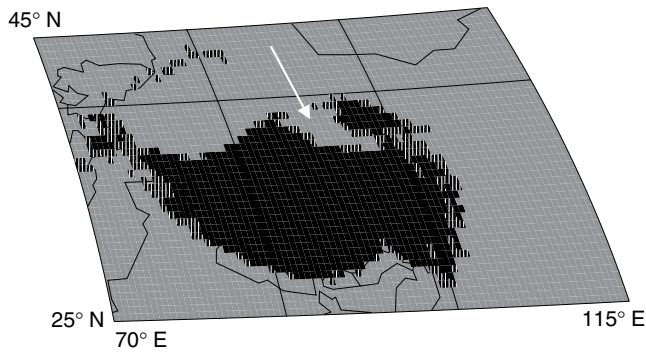
Figure 9.7   *Magnified representation of the Tibetan plateau (node 4, cf. Figure 9.4). The hatched areas are sensitive to climate change. The arrow indicates the Qaidam depression, which is more arid and is therefore clearly separated from the highland areas*

requires migration rates between $500\,m\ year^{-1}$ and $2000\,m\ year^{-1}$ for adaptation. Yet it might be that shift rates of approximately $5\,km\ year^{-1}$ are necessary in some regions. However, this is only the half the story, because migrating species must take advantage of gaps in the existing vegetation. Delcourt and Delcourt (1991) have estimated migration rates between $120\,m\ year^{-1}$ and $170\,m\ year^{-1}$ for the last post-glacial period, Clark (1998) for trees less than $1000\,m\ year^{-1}$, the IPCC (1997) projects a shift rate of about $7\,km\ year^{-1}$ for some species, while Malcolm *et al.* (2002) estimate $> 1000\,m\ year^{-1}$ for plant species to keep pace with human induced global warming during the twenty-first century. Thuiller (2007) estimated that plant and animals are shifting northwards approximately 6 km per decade. However, temperature increase changes might be even larger as it is currently assumed. The extreme summer in Europe 2003 showed what might happen (Rebetez *et al.*, 2006). Obviously there is some uncertainty about the migration rates, which undoubtedly depend on the observed species, but the SOM results are in reasonable correspondence with these studies.

9.5   DISCUSSION

As shown, the SOMTOP methodology offers a variety of advantages for data rich research in climatology and biogeography. It fits ideally the necessities of the 'geo-cybernetic enterprise' which needs smart and advanced methods for understanding the complexity of the entire earth system (cf. e.g. Bunde *et al.*, 2002; Kallache *et al.*, 2005; Kropp and Scheffran, 2007; Schellnhuber and Kropp, 1998). Employing the advantages of process based model approaches and neural networks in an intelligent and effective way could fundamentally extend the model family in earth systems science. In particular, the SOMTOP approach might fill the gap between high-resolution models (e.g. AGCM/DGVMs) and earth system models of intermediate complexity. This holds especially if more detailed and homogeneous information can be used for network training, e.g. about the biological constraints of the biome/PFT existence, competition and/or dominance.

As reported, objective knowledge can be derived from data, which will substantially improve our understanding of the ordering of climate and vegetation. Strong evidence is provided that the large-scale distribution of ecosystems is mainly determined by an 'abiotic' climate and soil inventory. Nevertheless, discussion will continue about what this classification has to do with the output of BIOME models. Biogeographers (e.g. Cramer *et al*., 2001) mention that a variety of processes and phenomena have to be taken into account when estimating vegetation types and their distribution which is one goal of modern vegetation models. The BIOME model, as introduced here, is based in its start-up setting on long-term climate means and parameters characterising soil properties. These inputs are used to calculate the modern vegetation cover on earth and sets the baseline for climate impact studies. Comparing the biome structuring and SOMTOP classification this implies that also the BIOME model is simply forced by climate and that ecogeographically motivated rules potentially play a minor role, i.e. dominance, competition, or other physiological information might be over-interpreted with respect to the biome formation. Obviously such details are not necessary for the decision of the distribution of broad-scale ecosystem complexes. However, SOMTOP supports the more than century-old insight that equilibrium vegetation and climate are closely connected. Further analytical efforts (for details see Kropp, 1999b) have shown that gradients can be detected in the database used. In a variety of cases they exist only with respect to single variables, but the SOMTOP approach is able to determine such gradients sensitively. In the context of global vegetation distribution one can speculate about the genesis of these gradients, since at first glance the climate space should be continuous. Thus, the question regarding the source of the structuring is not finally resolved. Such a classification may be a result of a self-control of the local biosphere/climate system, i.e. equilibrium vegetation has a significant influence on climate such that gradients emerge leading to temporarily stable situations. The latter could have serious consequences, since vegetationally stabilised climate margins may be highly vulnerable to land use change. This conjecture clearly requires some further examination.

9.6   SUMMARY

The objective of this study was to examine the structuring in spatially resolved climate and soil databases and relate the deduced knowledge to common model approaches using vegetation classifications on the global scale. The SOMTOP architecture we use consists of a SOM and an algorithm allowing assessment and control of the topology preservation during the simulations. The major advantages of the method are threefold: (i) it needs no *ex ante* assumptions about a hidden classification scheme; (ii) small SOM networks can be applied which considerably reduces computing time; and (iii) an effective dimension can be calculated for a high-dimensional database.

By using these features the SOMTOP approach becomes an attractive alternative for ecogeographical studies, since it allows high accuracy mappings. As shown, the outlined results are crucial to enlighten the major forcing factors for global broad scale ecosystem complex formation and distribution. SOMTOP provides objective interpolation of 20 climate and soil classes. Detailed examinations make clear that these findings are in a good accordance with empirical and model results. For example, for certain transition

zones an intriguing coincidence with other approaches was shown, providing strong evidence that climate is indeed the dominant factor for the formation of ecosystem complexes. Differences can also be well motivated either from a statistical or from an ecogeographic point of view. According to these results the classification scheme can be considered as equivalent for vegetation structuring on the global scale. For example, it can be shown that in the tropics only a few key variables, such as seasonality and amount of precipitation, are relevant for the classification scheme. This underlines the sensitivity of the method presented here. Finally, the SOMTOP results could prove useful in various ways for vulnerability studies. With the help of climate change scenarios, regions can be identified which are most susceptible to climate change, i.e. where a vegetation change may result.

Summing up, we feel that the results obtained allow new insights in biogeography. The philosophy of using available data for an autonomous, objective and self-organised systems analysis opens a promising road towards further progress in a system-based examination of vegetation distribution and climate change impacts, in particular for data rich situations. The approach reduces uncertainties, i.e. regarding the question of the extent to which vegetation clusters are meaningful on the global scale. This knowledge should be considered in process based vegetation models. The method presented here is open for further applications, e.g. it can be coupled to a global circulation model for a subsequent computation of characteristic climates. Further, it can be used as emulator, e.g. for the calculation of a reference climate which may be used as input for other models.

SOMTOP can significantly improve our knowledge background, in particular where topographical information is essential. The advantages of the methodology will be used in future work directed mainly to the analysis of spatio-temporal climate data. Using the the SOMTOP algorithm it may be feasible to detect changes in a climate regime by the topographical product, e.g. for annual databases.

REFERENCES

Aitkenhead, M. J., Mustard, M. J., and McDonald, A. J. S., 2004 *Using neural networks to predict spatial structure in ecological systems. Ecological Modelling* 179(3): 393–403.

Aleksandrova, V. D., 1977 *Geobotanische Gliederung der Arktis und Antarktis*, 188 ff. Komarov – Vorträge XXIX, Nauka, Leningrad.

Bailey, R. G., 1995 *Ecosystem Geography*. Springer Verlag, New York.

Bauer, H.-U. and Pawelzik, K. D., 1992 *Quantifying the neighborhood preservation of self-organizing feature maps. IEEE Transactions on Neural Networks* 3(4): 570–579.

Bauer, H.-U., Geisel, T., Pawelzik, K., and Wolf, F., 1996 *Selbstorganisierende Karten. Spektrum der Wissenschaft* (4): 38–47.

Box, E. O., 1981 *Macroclimate and plant forms: an introduction to predictive modelling in phytogeography*. Junk, The Hague.

Bunde, A., Kropp, J. P., and Schellnhuber, H. J. (eds), 2002 *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*. Springer Verlag, Berlin.

Carpenter, G., Gobal, S., Macomber, S., Martens, S., Woodcock, C. E., and Franklin, J., 1999 *A neural network method for efficient vegetation mapping. Remote Sensing of Evironment* 70: 326–338.

Clark, J. S., 1998 *Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. American Naturalist* 152: 204–224.

Cramer, W., Bondeau, A., Woodward, F. I., Prentice, C. I., Betts, A. R., Brovkin, V., Cox, P. M., Fisher, V., Foley, J. A., Friend, A. D., Kucharik, C., Lomas, M. R., Ramankutty, N., Sitch, S., Smith, B., White, A., and Young-Molling, C., 2001 *Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. Global Change Biology* 7: 357–373.

Crane, R. G. and Hewitson, B. C., 2003 *Clustering and upscaling of station precipitation records to regional patterns using self-organizing maps (SOMs). Climate Research* 25: 95–107.

Cubasch, U., von Storch, H., Waszkewitz, J., and Zorita, E., 1996 *Estimates of climate change in Southern Europe derived from dynamical climate model output. Climate Research* 7: 129–149.

Delcourt, H. R. and Delcourt, P. A., 1991 *Quarternary Ecology: a Paleoecological Perspective*. Chapman and Hall, London.

Epstein, H. E., Gill, R. A., Paruelo, J. M., Lauenroth, W. K., Jia, G. J., and Burke, I. C., 2002 *The relative abundance of three plant functional types in temperate grasslands and shrublands of North and South America: effects of projected climate change. Journal of Biogeography* 29: 875–888.

Ermini, L., Catani, F., and Casagli, N., 2005 *Artificial neural networks applied to landslide susceptibility assessment. Geomorphology* 66(1–4): 327–343.

Eyre, S. R., 1963 *Vegetation and Soils*. Aldine, Chicago.

Fao/Unesco, 1974 *Soil map of the world*. 1:5 000 000.

Foley, J. A., Prentice, C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S., and Haxeltine, A., 1996 *An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. Global Biogeochemical Cycles* 10(4): 603–628.

Foody, G. M., 1999 *Applications of the self-organising feature map neural network in community data analysis. Ecological Modelling* 120: 97–107.

Friend, A., Stevens, A. K., Knox, R. G., and Cannell, N. G. R., 1997 *A process based, terrestrial biosphere model of ecosystem dynamics* (Hybrid *v3.0*). *Ecological Modelling* 95: 249–287.

Grassberger, N. and Procaccia, I., 1983 *Measuring the strangeness of strange attractors. Physica D* 9: 189–208.

Grieger, B., 2002 *Interpolating paleovegetation data with an artificial neural network approach. Global and Planetary Change* 34: 199–208.

Hanewinkel, M., Zhou, W., and Schill, C., 2004 *A neural network approach to identify forest stands suceptible to wind damage. Forest Ecology and Management* 196: 227–243.

Haxeltine, A. and Prentice, I. C., 1996 Biome3: *an equilibrium biosphere model based on ecophysiological constraints, resource availability and competition among plant functional types. Global Biogeochemical Cycles* 10: 693–709.

Holdridge, L. R., 1947 *Determination of world plant formations from simple climatic data. Science* 105: 367–368.

IPCC, 1997 *The Regional Impacts of Climate Change. An Assessment of Vulnerability*. Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge.

Jäger, E. J., 1997 *Allgemeine Vegetationsgeographie*, pp. 511–582. Perthes Verlag, Gotha.

Joliffe, I. T., 1986 *Principal Component Analysis*. Springer Series in Statistics, Springer Verlag, New York.

Kallache, M., Rust, H., and Kropp, J. P., 2005 *Trend assessment: applications for hydrology and climate research. Nonlinear Processes in Geophysics* (12): 201–210.

Kaplan, J. O., Bigelow, N. H., Prentice, I. C., Harrison, S. P., Bartlein, P. J., Christensen, T. R., Cramer, W., Matveyeva, N. V., McGuire, A. D., Murray, D. F., Razzhivin, V. Y., Smith, B., Walker, D. A., Anderson, P. M., Andreev, A. A., Brubaker, L. B., Edwards, M. E., and Lozhkin, A. V., 2003 *Climate change and arctic ecosystems: 2. modeling, paleodata-model comparisons, and future projections. Journal of Geophysical Research* 108(D19): 8171.

Kohonen, T., 2001 *Self-Organizing Maps*. Springer Series in Informations Sciences, Springer Verlag, New York.

Köppen, W., 1918 *Klassifikation der Klimate nach Temperatur, Niederschlag und Jahresablauf. Petermanns Geographische Mitteilungen* 64: 193–203, 243–248.

Kropp, J. P., 1999a *A neural solution: a data driven assessment of global climate and vegetation classes*. In T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck, and K. Fukushima (eds), *6th International Conference on Neural Information Processing*, vol. I, pp. 279–285. Ieee Service Center, Piscataway.

Kropp, J. P., 1999b *Neuronale Netze und unscharfe Wissensbasen in der integrativen Umweltsystemanalyse*. Verlag im Internet GbR: 'dissertation.de', Berlin.

Kropp, J. P. and Scheffran, J. (eds), 2007 *Advanced Methods for Decision Making and Risk Management in Sustainability Science*. Nova Science Publishers Inc., New York.

Kropp, J. P., Block, A., Reusswig, F., Zickfeld, K., and Schellnhuber, H. J., 2006b *Semiquantitative assessment of regional climate vulnerability: The North Rhine – Westphalia study. Climatic Change* 76: 265–290.

Leemans, R. and Cramer, W. P., 1991 *The* Iiasa *database for mean monthly values of temperature, precipitation and cloudiness on a global terrestrial grid*. Technical Report RR-91-18, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Malcolm, J.R., Markham, A., Neilson, R.P., and Garaci, M., 2002 *Estimated migration rates under scenarios of global climate change. Journal of Biogeography* 29(7): 835–849.

Malmgren, B. and Winter, A., 1999 *Climatic zonation in Puerto Rico based on principal components analysis and an artificial neural network. Journal of Climate* 12: 977–985.

McKenzie, D., Peterson, D. W., Peterson, D. L., and Thornton, P. E., 2003 *Climatic and biophysical controls on conifer species distributions in mountain forests of Washington State, USA. Journal of Biogeography* 30: 1093–1108.

Moldenhauer, O. and Lüdeke, M. K. B., 2002 *Climate sensitivity of global terrestrial net primary production (NPP) calculated using the reduced-form model NNN. Climate Research* 21: 43–57.

Özesmi, S. L. and Özesmi, U., 1999 *An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecological Modelling* 116: 15–31.

Park, C. C., 1995 *Tropical Rainforests*. Routledge, London.

Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R., Monserud, R. A., and Solomon, A. M., 1992 *A global biome model based on plant physiology and dominance, soil properties and climate. Journal of Biogeography* 19: 117–134.

Rebetez, M., Mayer, H., Dupont, O., Schindler, D., Gartner, K., Kropp, J. P., and Menzel, A., 2006 *Heat and drought 2003 in Europe: a climate synthesis. Annals of Forest Science* 63: 569–577.

Ritter, H. and Schulten, K., 1988 *Convergence properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection. Biological Cybernetics* 60: 59–71.

Roelandt, C., 2001 *Coupled simulation of potential natural vegetation, terrestrial carbon balance and physical land-surface proporties with the* Albioc *model. Ecological Modelling* 143: 191–214.

Schellnhuber, H. J. and Kropp, J. P., 1998 *Geocybernetics: controlling a complex dynamical system under uncertainty. Naturwissenschaften* 85(9): 411–425.

Simmons, I., 1979 *Biogeography: Natural and Cultural*. Edward Arnold Ltd, London.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K., and Venevsky, S., 2003 *Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Global Change Biology* 9: 161–185.

Thornthwaite, C. W., 1948 *An approach toward a rational classification of climate. Geographical Review* 38(1): 55–94.

Thuiller, W., 2007 *Climate change and the ecologist. Nature* 448: 550–552.

Tivy, J., 1996 *Biogeography: A Study of Plants in the Ecosphere*. Longman, Harlow.

Ultsch, A. and Mörchen, F., 2005 *ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM*. Technical Report, Deptartment of Mathematics and Computer Sciences No. 46, University of Marburg, Marburg.

Walter, H. and Breckle, S.-W., 1991a *Die Ökologie der Erde: Ökologische Grundlagen in globaler Sicht*, vol. 1. Fischer Verlag, Stuttgart.

Walter, H. and Breckle, S.-W., 1991b *Die Ökologie der Erde: Spezielle Ökologie der gemäßigten und arktischen Zonen Euro-Nordasien*, vol. 3. Fischer Verlag, Stuttgart.

Walter, H. and Breckle, S.-W., 1991c *Die Ökologie der Erde: Spezielle Ökologie der tropischen und subtropischen Zonen*, vol. 2. Fischer Verlag, Stuttgart.

Whitmore, T. C., 1999 *Arguments on the forest frontier. Biodiversity and Conservation* 8: 865.

Woodward, F. I., Lomas, M. R., and Betts, R. A., 1998 *Vegetation-climate feedback in a greenhouse world. Philosophical Transactions of the Royal Society of London* 353: 29–39.

This page intentionally left blank

10

Self-Organising Map Principles Applied Towards Automating Road Extraction from Remotely Sensed Imagery

Pete Doucette[1], Peggy Agouris[2] and Anthony Stefanidis[3]

[1] *Principal Scientist, ITT Corporation Advanced Engineering and Sciences, Alexandria, VA 22303 USA*

[2] *Center for Earth Observing and Space Research, George Mason University, Fairfax, Virginia 22030, USA*

[3] *Department of Earth Systems and Geoinformation Sciences, George Mason University, Fairfax, Virginia 22030, USA*

10.1   INTRODUCTION

The extraction of geospatial features from remotely sensed imagery remains the primary means to create or update geospatial databases. Presently, most feature extraction or updating is done manually by humans in typical production settings. Given the labor costs involved with manual extraction, efforts to automate extraction to some extent have been the subject of considerable research activity. The growing availability of high spatial resolution imagery further drives the demand for timely and increasingly accurate feature data. The volume of high (spatial and spectral) resolution imagery collected by a growing number of space borne sensors has overwhelmed traditional manual image analysis and extraction processes. Detailed road networks represent a feature type in high demand for use in GIS databases. A pressing need exists for robust automated road extraction

algorithms. The goal of this work is to leverage self-organising learning principles toward automating road centerline extraction from high spatial resolution images.

In this chapter we present a compilation of work that considered the application of self-organising map (SOM) learning principles towards the road extraction problem. A distinctive aspect of this approach versus others was the use of SOM node-based representations to perform the 'organisational' analysis of road networks in remotely sensed images. Node regions in images were used to represent a higher level of abstraction for 'perceptual organisation' versus image pixels. In recent years, there has been an increase in research activity to better identify and demonstrate aspects of perceptual organisation in computer vision (Jacobs and Lindenbaum, 2003). Analogies have often been drawn between neurobiological learning and self-organising learning principles (Haykin, 1999; Kohonen, 2001). To that end, our interest was to explore how SOM principles could bring new insight to the problem of automated road extraction.

The sections are laid out in chronological order of how the research evolved. Material is excerpted from earlier published works (Agouris *et al*., 2001a; Doucette, 2002; Doucette *et al*., 1999, 2000, 2001, 2004). Section 10.2 provides a synopsis of road extraction methods in the literature. Section 10.3 is an overview of our early work with the original SOM algorithm for application to the road extraction problem. Section 10.4 describes modifications made to the updating process of original SOM. Section 10.5 describes how the node neighborhood mechanism of the SOM was modified to construct road topology. Section 10.6 represents a culmination of our automated road extraction work with SOM principles, in which spectral information is leveraged into the analysis.

10.2   A SYNOPSIS OF ROAD EXTRACTION METHODS

Automated road extraction methodology has traditionally modeled roads according to image properties related to edges, geometry, radiometry, and texture. At high spatial resolution (e.g. 1 m per pixel or better), spectral content takes on more significance. Although color aerial imagery has long been available at sub meter spatial resolutions, acquisition is costly and time consuming. However, the reality of space borne multi-spectral imagery at spatial resolutions of 1 m per pixel or better is presumably inevitable.

Road extraction strategies in the literature are often categorised according to their degree of automation. In this chapter, methods will be classified into two operational modes: on-line or off-line. In either mode, human oversight is usually required to some extent. On-line methods are designed to assist the human operator on a 'per road segment' basis. For example, a human operator marks start and end (or more) points between two road intersections to initialise the algorithm. The process is repeated for each road segment, with manual editing performed as needed for each segment extracted. Well known genres of on-line methods include *road trackers* (Barzohar *et al*., 1997; Mckeown and Denlinger, 1988; Vosselman and Knecht, 1995) and *snakes* (Gruen and Li, 1997; Trinder and Li, 1995). By contrast, off-line methods typically require a one-time initial-isation of algorithm parameters. Automatic extraction is performed over the entire scene of interest without interruption, followed by manual editing (Baumgartner *et al*., 1999; Doucette *et al*., 2004; Harvey, 1999; Havercamp, 2002; Hinz and Baumgartner, 2003; Price, 2000; Shackelford and Davis, 2003). When (outdated) GIS data exist, 'update'

strategies are used as a means of guiding an off-line extraction (Agouris, 2001b; Bordes *et al*., 1997; Doucette *et al*., 1999; Walter and Fritsch, 1998; Zhang *et al*., 2001). Recent works that provide a broad overview of the state of the art in road extraction research include Saleh (2004) and Mena (2003).

The road extraction algorithm presented in this chapter is an off-line strategy. Off-line algorithms are less constrained by computational complexity than on-line algorithms. Because on-line methods require recurrent human inputs in real-time, they must be computationally fast to maintain efficient interaction with a human operator. The degree of computational complexity of an on-line algorithm is therefore dependent upon the capabilities of the computing hardware. By contrast, off-line algorithms can be more computationally complex because they do not require continual real-time attention. This allows off-line algorithms to address the more difficult extraction problems found with more complex images (e.g. more urban versus rural). Off-line extraction methods that use a bottom-up approach can be generally broken down into three stages of computational complexity: (1) low-level detection that generates initial hypotheses for candidate road components, (2) mid-level grouping of road components, and (3) high-level reasoning for road network completion. The complexity of implementation increases with each stage.

10.3 SELF-ORGANISED MAPPING OF ELONGATED REGIONS

In high spatial resolution imagery, roads manifest as *elongated regions* (ERs). Early work with the original SOM (Kohonen, 1982) was based on mapping the medial axes of ERs from which to derive road centerlines. This was accomplished by fitting a SOM network to the input space (i.e. road pixels) onto a SOM network space (i.e. a one-dimensional chain of nodes). This mapping could then be used as a delineated road centerline suitable as GIS input.

Figure 10.1 demonstrates a one-dimensional SOM network fitting of a synthesised ER curving from the upper left to lower right. The input is the (*x, y*) coordinates for each point sample in the ER. Random initialisation of a 10 node SOM chain is shown in Figure 10.1(a). Node ordering progresses in figure 10.1b and 10.1c, and refinement and convergence is achieved in figure 10.1d. The SOM neighborhood function begins with a simple step function of two nodes on either side of a winning node (i.e. ordering phase). Over time, the neighborhood function shrinks to one node, and then to zero (i.e. refinement and convergence). Similarly, the learning rate function approaches zero over time to promote a stabilised convergence. Convergence is declared when the cumulative updates to the node adjustments fall below an empirically set threshold.

Applying the original SOM technique to update existing GIS roads and rivers is demonstrated in Figure 10.2. The existing GIS data are shown as white lines super-imposed on a high spatial resolution multispectral image in Figure 10.2(a). A single river feature is in the upper left, and the remaining features are roads. A maximum likelihood classification (MLC) of the multispectral image was performed to create input spaces for road and river pixels. The road input space is shown in Figure 10.2(b). Classification errors were left in the input space, since the objective was to determine the extent to which the SOM could tolerate noise. One-dimensional SOMs were used for linear road sections. However, in order to map road intersections, a two-dimensional SOM topology
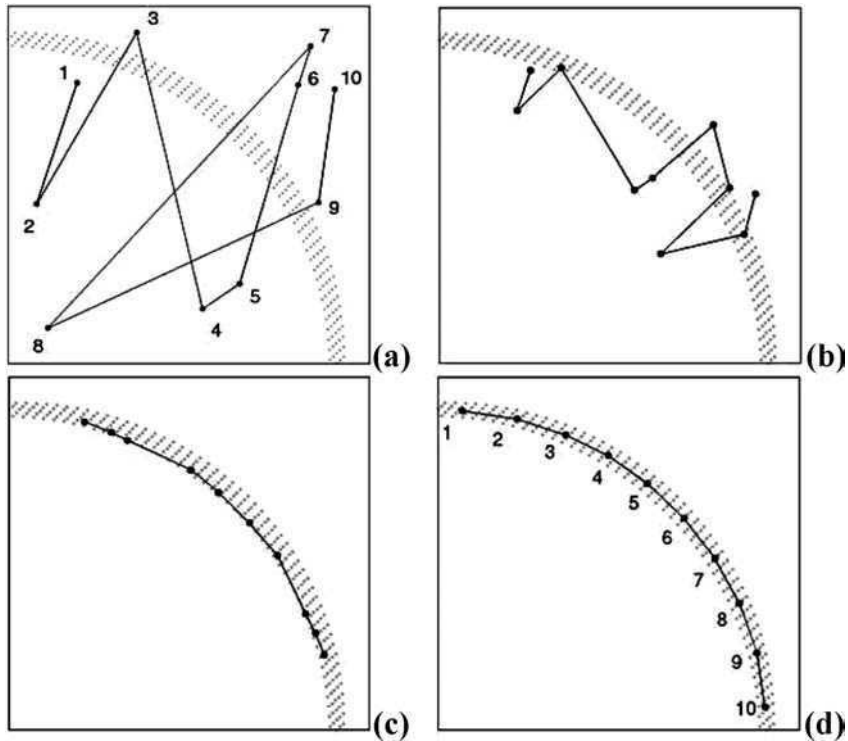
Figure 10.1   *A one-dimensional SOM to fit an elongated region: (a) random initialisation; (b) after a few iterations; (c) ordering achieved; (d) convergence. (Reproduced from Doucette et al., Self-Organised Clustering for Road Extraction in Classified Imagery, ISPRS Journal of Photogrammetry and Remote Sensing, 55(5–6), pp. 347–358. Copyright 2001, Elsevier)*

was needed (e.g. 'cross' or 't' shapes). Figure 10.2(c) shows the results of using one- and two-dimensional SOMs to map the medial axes of the input spaces.

This SOM application was representative of an off-line update method, where existing GIS information is used to initialise the SOM. However, significant limitations existed, which included: (1) requiring initialisation of the SOM relatively close to the features of interest in the input space; (2) high sensitivity to noise in the input space; (3) approximate estimation of the number of nodes to be used in the SOM network; and (4) the requirement for a MLC classification to create an input space for roads. In work that followed, several important modifications to the original SOM algorithm were implemented to mitigate these limitations.

10.4   THE WINNER-TAKE-ALL APPROACH

The original SOM algorithm is based on *competitive learning* principles. When a SOM network neighborhood of competing nodes is collapsed to zero, the result is the degenerate case referred to as a 'winner-take-all' (WTA) network. An analogous situation exists
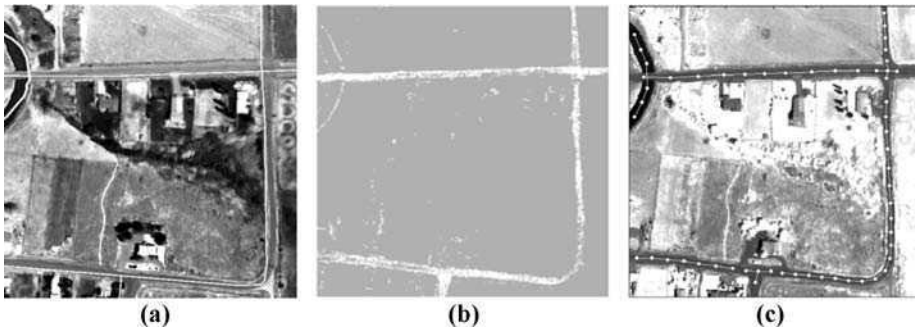
Figure 10.2 *Early results from SOM for road extraction: (a) image overlaid with coarse resolution vector data (white lines); (b) classified road pixels; (c) SOM convergence for one- and two-dimensional road components. (Reproduced from Doucette et al., Automated Extraction of Linear Features from Aerial Imagery Using Kohonen Learning and GIS Data, Lecture Notes in Computer Science, Springer-Verlag, Portland, ME, Volume 1737, pp. 20–33. Copyright 1999, Springer-Verlag)*

with the traditional iterative clustering techniques of *K*-means (MacQueen, 1967), the Generalised Lloyd Algorithm (Lloyd, 1982), the Linde–Buzo–Gray algorithm (Linde *et al.*, 1980), and vector quantisation (Gersho and Gray, 1992). The term 'cluster center' used in the context of traditional clustering techniques is synonymous with use of the term 'node' (also neuron) for the SOM. In this chapter, the term 'node' is used for either case for consistency. Traditional clustering algorithms use a 'batch' updating technique, in that *all* input samples and nodes are visited before each update. By contrast, the SOM uses sequential updating, i.e. one sample at a time per node update.

The *K*-means algorithm can be viewed essentially as a batch-updating WTA version of the SOM algorithm as demonstrated by Lutrell (1989). The *K*-means can also be configured with a SOM style neighborhood function, transforming it into a 'batch map' (Kohonen, 1993). The main advantages we considered for preferring a *K*-means updating approach over the original SOM include: (1) no learning rate parameter required, (2) no sensitivity to the order in which samples are presented, and (3) convergence is generally faster more reliable (Doucette *et al.*, 2001).

10.4.1   The *K*-means Approach

With a *K*-means approach, topological relationships among nodes (e.g. neighborhood function) are not used as with the original SOM. Our strategy was to define topological relationships among nodes *after K*-means convergence. A description of this strategy is provided in Section 10.5. With this strategy in mind, finding appropriate node convergence patterns suitable for constructing road topologies was needed. For example, the median statistic was less sensitive to noise than the mean for mapping the medial axes of elongate regions. The tradeoff was that the accuracy of the median is limited to the nearest half pixel, and the median required more computation time [e.g. $O(N{\cdot}\log N)$ for the median versus $O(N)$ for the mean]. Finally, the fact that the median could not be calculated in a sequential updating procedure (such as in the original SOM), necessitated the use of a batch updating method. Therefore, a median calculation was substituted for the mean in the *K*-means algorithm.

10.4.2   Node Validation

Automating node validation is the classic problem of determining how many nodes are needed to adequately represent an input space when using an iterative clustering method. In addition to node quantity, the convergence patterns from iterative clustering are dependent upon node initialisation in the input space. This problem was addressed by applying the node merging technique of the ISODATA (iterative self-organising data analysis technique) algorithm (Hall and Ball, 1965). ISODATA is an extension of *K*-means in which node merging and splitting is allowed. For example, nodes that move within some empirically defined minimum spacing between each other from one update to the next, are merged (e.g. one of them is deleted).

The goal of node merging was to establish a node convergence pattern suitably spaced for road centerline construction. Node merging enabled nodes to be better distributed along the medial axes of elongated regions, as opposed to off-axis representations. For example, Figure 10.3 shows node convergence patterns before and after node merging. The input space is represented by the white elongated regions (with omission noise introduced). In Figure 10.3(b), nodes 12, 14, 16, and 18 have been deleted via merging. The dotted polygons are the Voronoi regions associated with each node. The input samples contained within each node's Voronoi region are closest to that node.
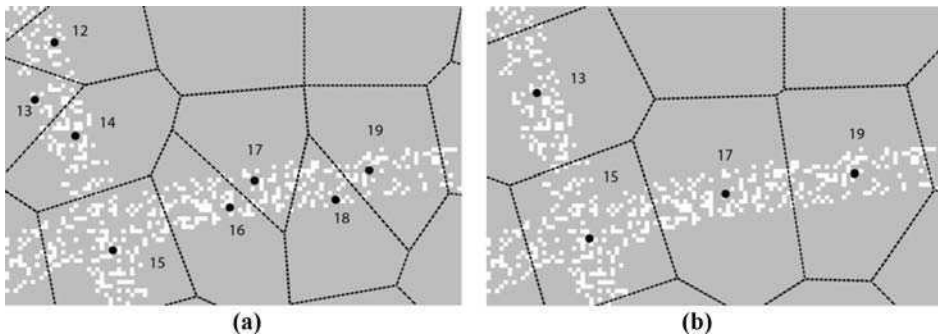
Figure 10.3   *Node merging for noisy elongated regions (white pixels): (a) prior to merge; (b) after merge (nodes deleted: 12, 14, 16, and 18)*

To further facilitate suitable node spacing at convergence, a regular grid pattern was used for node initialisation. The initial grid spacing was set slightly larger than the node merging threshold. To speed up convergence time, nodes that did not win competitions (i.e. dead nodes) were automatically identified and eliminated after each iteration.

10.4.3   The SORM Algorithm

The self-organised road mapping (SORM) algorithm (Doucette *et al.*, 2001), represented the integration of several modifications to the original SOM for application toward automated road extraction. The following description of the SORM algorithm is excerpted from Doucette *et al.* (2004). It is an adaptation of the *K*-means algorithm as defined by Tou and Gonzalez (1974).

- *Step 1 (Initialisation)*. Initialise K nodes, $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$, in the input (raster) space. The nodes are arranged as a regular grid, with node spacing defined by β_{grid}. The node spacing is an empirically derived quantity that is based on the approximate width of roads. Its value can be preset and adjusted automatically, or modified manually at run time.
- *Step 2 (Determine sample–node associations)*. At the nth iteration, determine which node each sample vector $\mathbf{x}$ (x and y coordinates of raster samples) is closest to according to:

$$\mathbf{x} \in S_j(n) \quad \text{if } ||\mathbf{x} - \mathbf{c}_j(n)|| < ||\mathbf{x} - \mathbf{c}_i(n)|| \quad (i = 1, 2, \ldots, K \text{ and } i \neq j) \qquad (10.1)$$

where $S_j(n)$ represents the set of samples whose closest node is $\mathbf{c}_j(n)$. Ties are resolved by the order rule. If $S_j(n)$ is less than a minimum threshold number of samples (determined empirically), then node $\mathbf{c}_j(n)$ is deleted.
- *Step 3 (Update node positions)*. Compute the median of the samples associated with each node from step 2, and let each median represent the updated location of the node according to:

$$\mathbf{c}_j(n+1) = \begin{cases} \text{sort}(\mathbf{x})_{(N_j+1)/2}, & \text{if } N_j \text{ is odd} \\ \frac{1}{2}\left[\text{sort}(\mathbf{x})_{N_j/2} + \text{sort}(\mathbf{x})_{(N_j/2)+1} \right], & \text{if } N_j \text{ is even} \end{cases} \quad (j = 1, 2, \ldots, K)$$
$$(10.2)$$

where N_j is the number of samples in $S_j(n)$, $\text{sort}(\mathbf{x})$ represents the sorted samples for a given node, and $\mathbf{x} \in S_j(n)$. The median statistic is used because it is less sensitive to noise than the mean.
- *Step 4 (Node pruning)*. Let the Euclidean distance between any two nodes i and j be defined by d_{ij}, and a minimum distance allowance between nodes be defined by β_{min}. If $d_{ij} \leq \beta_{\text{min}}$, then nodes i and j, are 'merged' (e.g. by deleting or 'pruning' node j). If node pruning occurs, then return to step 2.
- *Step 5 (Convergence test)*. Iterate from step 2 until all node positions remain unchanged such that:

$$\mathbf{c}_j(n+1) = \mathbf{c}_j(n) \qquad (j = 1, 2, \ldots, K) \qquad (10.3)$$

10.4.4   Techniques for Speeding up Convergence

The computational requirements for iterative optimisation techniques such as clustering can be substantial. The computational complexity of the SORM algorithm is *O(mNKT)*, where m is the number of dimensions of the input space, N is the number of samples, K is the number of nodes, and T is the number of iterations needed for convergence. 'Early stopping' is perhaps the simplest technique used to speed up convergence. This can be accomplished either by raising the stopping criterion threshold, or limiting the number of iterations T to some maximum, T_{max}. Since it is not uncommon for small node adjustments to fluctuate needlessly during the latter stages of network refinement, early stopping is a practical idea. However, the challenge is to avoid premature stopping.

Focused searching of the input space represented a more sophisticated approach for algorithm speed-up. For example, step 2 of the SORM algorithm consisted of *NK* possible

distance calculations, which accounted for a significant bottleneck. Two focused searching techniques were investigated in this work: (1) local space searching; and (2) node activity searching. In both cases, the goal was to confine the search of the input space to local regions.

10.4.4.1   Local space searching

The objective of local space searching was to reduce the sample space searched for any given node. The brute force approach was to determine node–sample distances for *all* of the input space samples. The local space search considered only those samples defined in a local neighborhood of the closest nodes for a given node. This is similar in principle to the 'shortcut winner search' (Kohonen, 2001), but applied to the SORM algorithm. In local space searching, the neighborhood sample set was determined by a node–sample association index established from the previous iteration. Therefore, one complete iteration through the input space was required. To determine the closest neighboring nodes to a given node, the inter-node distances had to be calculated after each iteration. The computational complexity of the inter-node calculation was $O[\frac{1}{2}(K^2 - K)]$, which turned out to be inconsequential compared with the *NK* calculation when N was much larger than K.

10.4.4.2   Node activity searching

The goal of node activity searching was to avoid re-visitation to regions of nodes that had quickly settled into position. Nodes that showed little or no sign of adjustment between successive updates were flagged as being temporarily 'inactive'. These were usually nodes that settled into stable positions faster than others, which was dictated by the distribution of the input space. Only the 'active' nodes, along with their neighbors, were considered in the *NK* calculation of step 2 of SORM. Neighboring nodes, which could be active or inactive, were included to allow for their adjustment relative to the adjustments made by active nodes. This mechanism allowed for temporarily inactive nodes to be reactivated. The local neighborhood nodes were identified in a manner identical to that of local space searching. The activity threshold was determined empirically.

   The reduction of the SORM step 2 calculations went from *NK* to NK_{active}, where K_{active} is the number of active nodes, plus the nodes in the activity neighborhoods. Initially, node activity was high, and so no algorithm speed-up was apparent. However, as node activity levels receded over time, algorithm acceleration became apparent. As in local space searching, node activity searching could only occur after the first iteration, and inter-node distances had to be calculated for each iteration. The kind of speed-up from activity searching differed from local space searching, in that it was an acceleration effect. By contrast, the speed-up in local space search is essentially constant with each iteration. In general, algorithm speed-up could be substantial as the number of nodes used increased. However, convergence results could vary considerably depending upon initial conditions.

10.5   DEFINING NODE TOPOLOGY IN THE INPUT SPACE

The original SOM algorithm uses a predefined network topology among nodes, which is preserved as it maps an input space. However, it is also possible to define the topological relationships among the nodes *in the input space* instead of in the network (Kangas *et al*., 1990; Kohonen, 2001). The adaptability allowed by defining network topology in the input space proved to be better suited for constructing road network topology. Topological relationships among nodes were defined by metrics derived from the input space. The input space metrics were used in a fuzzy inference model, which was used to validate the construction of node topologies that were consistent with road networks (Agouris *et al*., 2001a; Doucette, 2002).

10.5.1   Input Space Metrics for Node Topology

The three primary input space metrics that were used to guide topology construction among the nodes were *proximity, link angle*, and *orientation angle*. Node proximity is simply the Euclidean distance between two given nodes in the input space. The link angle is the geometric angle formed between two given nodes. The node orientation angle is measured from the spatial distribution of pixels contained within a single node's Voronoi region. A derivative metric is the *deflection angle*, which is a measure of the difference between orientation and link angles for two given nodes. For example in Figure 10.4, for two nodes i and j, with orientation angle θ_i for node i, and link orientation angle θ_L, a deflection angle between node i and θ_L is defined as:

$$_i\theta_D = |\theta_i - \theta_L| \tag{10.4}$$

where $|0° \leq {}_i\theta_D \leq 90°|$.

Similarly for node j, the deflection measure, $_j\theta_D$, is computed in the same manner for the same node link.
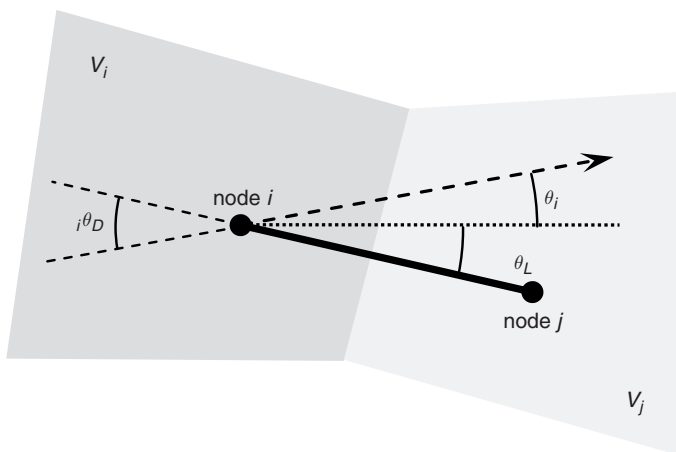
Figure 10.4   *The deflection angle between node orientation and node link*

The orientation angle is derived from the input space for a given Voronoi set V_i, of image pixels associated with node i. Two different orientation angle measures were investigated: principal component and Hough transform.

10.5.1.1   *Principal component orientation*

This value is derived from the principle component transformation calculation, defined as:

$$\theta_P = \frac{1}{2} \tan^{-1} \left(\frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2} \right) \tag{10.5}$$

The standard deviations σ_x and σ_y, are in x and y directions of the pixels contained in Voronoi set V_i, for node i, and σ_{xy} is the covariance.

10.5.1.2   *Hough transform orientation*

The Hough transform (HT) is used to determine the maximum pixel count occurring on a straight line within Voronoi pixel set V_i for node i. The HT scans are run locally for each node's Voronoi pixel set in the input space. The HT scans occur over a radial range defined by ρ (e.g. maximum dimension for a given Voronoi region), and angular range of θ (e.g. 0 to 180°) (Gonzalez and Woods, 2002). The resolution of the HT is determined by radial and angular quantisation intervals ($\Delta\rho$ and $\Delta\theta$, respectively).

The computational complexity of the HT is $O(\rho\theta)$, and depending upon selection of $\Delta\rho$ and $\Delta\theta$, computation of the HT orientation angle, θ_H, can be considerably more expensive than the principal component angle computation, θ_P. However, θ_H is generally less sensitive to outliers than θ_P. In effect, the HT orientation is comparable with the mode angle, where the principal component orientation to the mean angle. The HT quantisation parameters ($\Delta\rho$ and $\Delta\theta$) are determined empirically.

10.5.2   Node Topology Construction

A fuzzy inference system based on fuzzy set theory (Zadeh, 1965) was developed to model node topologies that were consistent with road network topology. Fuzzy modeling was used as a convenient tool to develop smooth similarity measures, rather than using hard thresholds. The process referred to as 'fuzzy organisation of elongated regions' (FOrgER) represented a weighted graph-theoretic method for linking nodes into progressively larger curvilinear components and networks (Doucette, 2002). The motivation for FOrgER originated from Gestalt grouping principles (e.g. *proximity, continuity, context*, and *closure*) used to link nodes into distinctive features (Zahn, 1971). A conceptually similar fuzzy-based implementation for grouping road components is in Steger *et al.* (1997).

FOrgER captured the relationships defined among the node proximity, link, and orientation metrics in fuzzy membership functions. The objective of FOrgER was to use fuzzy inference rules to construct and validate node topology from low level to high level

grouping. An example of fuzzy inference rules using proximity and deflection angles metrics for low level grouping, was:

1. If (*Deflect Angle i* is **small**) AND (*Deflect Angle j* is **small**) AND (*Proximity* is **close**), then (*LinkStrength i:j* is **high**).
2. If (*Deflect Angle i* is **large**) OR (*Deflect Angle j* is **large**), AND (*Proximity* is **close**), then (*Link Strength i:j* is **possible**).
3. If (*Deflect Angle i* is **large**) AND (*Deflect Angle j* is **large**), AND (*Proximity* is **close**), then (*Link Strength i:j* is **low**).
4. If (*Proximity* is **far**) then (*Link Strength i:j* is **improbable**).

In mid level grouping, link angles were tested between successive links for *continuity* of direction. *Contextual* validation for a given link was tested against the strength of neighboring links. An example of fuzzy inference rules to validate a link between nodes *i* and *j* with node neighbors *h* and *k* on either side, was:

1. If (*Link Strength h_i* is **high**) AND (*Link Strength j_ k* is **high**), then (*Link Strength i_j* is **high**).
2. If (*Link Strength h_i* is **high**) OR (*Link Strength j_k* is **high**), then (*Link Strengthi i_j* is **possible**).
3. If (*Link Strength h_i* is **low**) AND (*Link Strength j_k* is **low**), then (*Link Strength i_j* is **low**).

In high level grouping, a network *closure* process tested the linking of end nodes of linked network components. Closure analysis was performed mainly for road intersections, bridging gaps caused by occlusions in the image (e.g. tree shadows), and sharp bends in roads.

10.6   INTEGRATING AUTOMATED SPATIAL AND SPECTRAL ANALYSIS

Integrating spatial and spectral information into the road extraction process led to the final algorithm modification called 'self-supervised road classification' (SSRC). The operational flow of the entire procedure is shown in Figure 10.5. The approach integrates processes consisting of: (1) edge analysis (low level process); (2) self-organisation (mid level process); (3) topology construction (mid and high level processes); and (4) spectral analysis (low level process).

The first process consists of identifying candidate road centerline pixels in an image through edge analysis. In the second process, a node-based representation of the candidate centerline pixels is generated via the SORM algorithm (described earlier). In the third process, road topology is constructed by linking the SORM nodes in the input space using a fuzzy grouping model (FOrgER). The fourth process is a refinement feedback loop in which spectral information for roads is derived from the SORM nodes. The procedure has the option to iterate as desired to refine road extraction results.
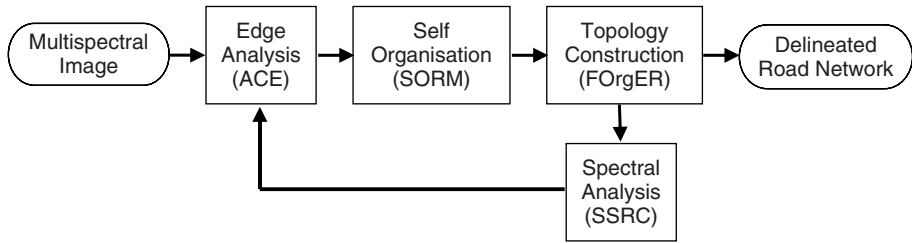
Figure 10.5 *The operational flow that integrates spatial and spectral analysis processes for road extraction*

10.6.1   Edge Analysis

The edge analysis process finds candidate road centerline pixels based on the technique of anti-parallel edge detection (Nevatia and Ramesh, 1980; Zlotnick and Carnine, 1993). Candidate centerline pixels are detected from a single layer image with the 'anti-parallel edge centerline extractor' (ACE) algorithm (Doucette *et al.*, 2004). ACE uses the Canny technique to extract edges, and the Sobel technique to determine edge orientation. The resulting Canny and Sobel images are scanned to detect the centerlines of elongated regions. The input parameters for ACE are: (1) minimum and maximum feature width (w_{min} and w_{max}); (2) maximum allowed deflection angle between anti-parallel edge orientations (α_{max}); and (3) minimum number of pixels required per connected component (cc_{min}). The deflection angle (α) between edge orientations is calculated as:

$$\alpha = |180° - |\varphi_p - \varphi_q|| \tag{10.6}$$

where φ_p and φ_q are the gradient orientation angles at pixels p and q, respectively.

   Figure 10.6 demonstrates the edge analysis, self-organisation, and topology construction processes with an image. Figure 10.6(a) is a 1 m per pixel multispectral image of suburban streets. To enhance the contrast of the roads for a single layer representation, the multispectral image was preprocessed with a principal component analysis (PCA). The second principal component layer (PC2) was empirically selected as the single layer input for the ACE algorithm. Figure 10.6(b) shows the results from ACE using parameters of $w_{min} = 5\,\text{m}$, $w_{max} = 15\,\text{m}$, $\alpha_{max} = 45°$, and $cc_{min} = 3$ pixels.

   It is clear from these results that ACE is effective to the extent that: (1) roads can be described by anti-parallel edges; and (2) anti-parallel edges are exclusive to roads. Errors result when these assumptions break down, which is to be expected from remotely sensed imagery. Therefore, the goal of ACE is to use low level analysis to generate initial hypotheses for the locations of roads. The goal of topology construction is to manage the errors generated by ACE through the higher level processing.

10.6.2   Self Organisation and Topology Construction

The ACE results of Figure 10.6(b) were used as the input space for the SORM algorithm. Figure 10.6(c) shows the SORM node convergence pattern. The links between nodes in Figure 10.6(d) were the result of the topology construction process from the FOrgER
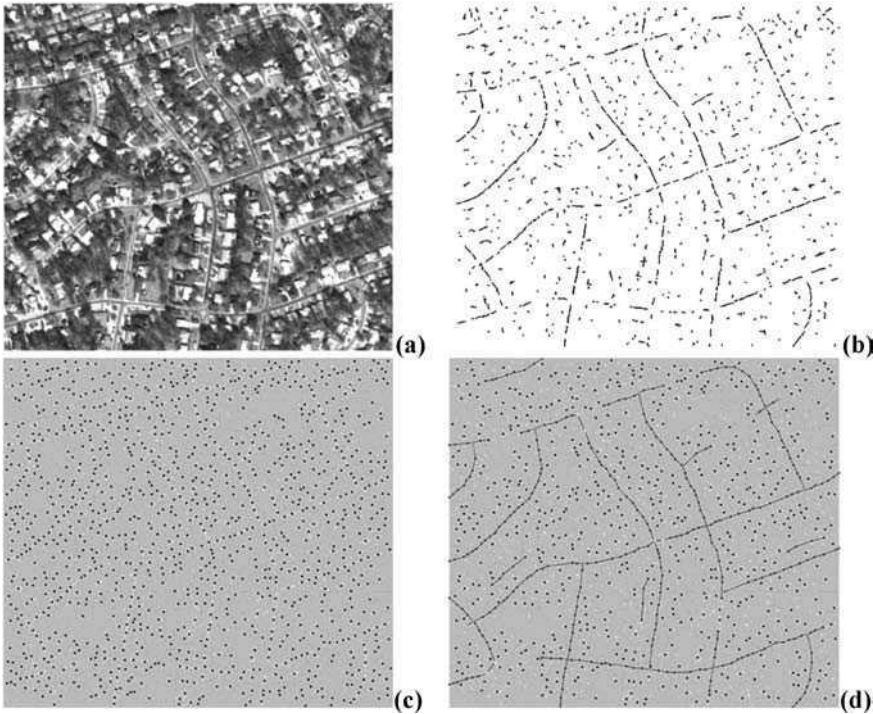
Figure 10.6 *Spatial analysis processes applied to extract roads: (a) 1 m per pixel image of suburban streets; (b) edge analysis results from ACE algorithm; (c) self-organised node convergence pattern from SORM algorithm; (d) topology construction results from FOrgER algorithm. (Reproduced with permission of the American Society for Photogrammetry and Remote Sensing, from Doucette et al., 2004)*

algorithm. The automated road extraction performance was evaluated quantitatively against manually extracted roads. The primary metric used for the evaluation was the quality, Q, which is defined as $Q = tp/(tp + fp + fn)$, where tp = true positives, fp = false positives, and fn = false negatives (Wiedemann *et al.*, 1998). Q represents a normalised measure of omission and commission errors, and takes on values between 0 and 1. For the image in Figure 10.6, the road extraction quality went from $Q = 0.29$ for ACE in Figure 10.6(b), to $Q = 0.67$ for topology construction in Figure 10.6(d).

10.6.3 Spectral Analysis

A goal of spectral analysis is to automate the process of training sample selection and refinement for roads. Training samples from the image are gathered from spatial neighborhoods centered on each node, and for all nodes. Spectral statistics (mean and variance) for the road class are derived from these training samples. Spectral statistics for non-road classes are derived from a *K*-means clustering. This result provides for a multi-class spectral distribution of the non-road candidate classes.

A maximum likelihood classification (MLC) is then performed to spectrally separate roads from non-roads. Because the MLC is automatically or 'self' supervised, the accuracy of the classification is not assessed in the conventional sense. Rather, the objective here is to provide sufficient separation between only roads and non-roads, for subsequent spatial analysis of the classification results.

This is followed by automatic morphological filtering applied to the road class to fill gaps and reduce noise. With the road class in hand, the ACE and topology construction process can be repeated to further refine the results, i.e. the self-supervised road classification (SSRC) feedback loop in Figure 10.5. The assessment of the final road extraction results demonstrates the extent to which spectral separation is sufficient.

The two scenes in Figure 10.7 present results from a sample demonstration of the SSRC approach for automating road extraction. Both scenes represent a 1 m per pixel multispectral image of suburban street networks. To demonstrate the effects of different levels of occlusion, Figure 10.7(b) contained considerably more tree shadows than Figure 10.7(a). The road centerlines extracted from SSRC (following two iterations)
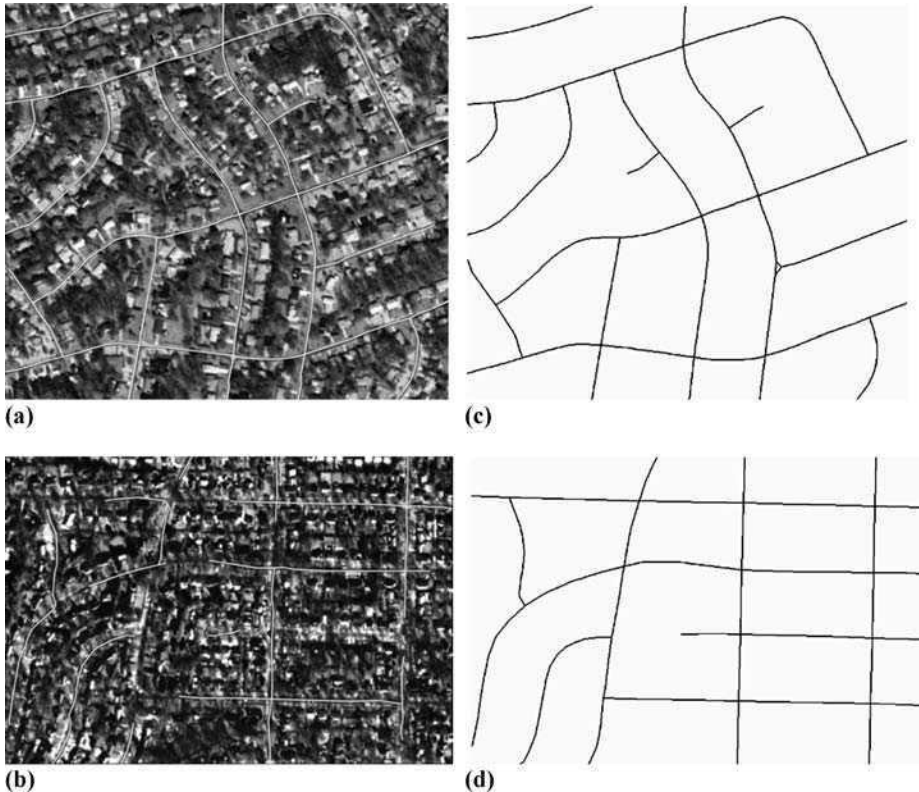
Figure 10.7 *Results from SSRC: (a, b) road extraction results overlaid on scenes (white lines); (c, d) human delineated ground truth used for algorithm evaluation. (Reproduced with permission of the American Society for Photogrammetry and Remote Sensing, from Doucette et al., 2004)*

are depicted as white lines superimposed on the images in Figure 10.7(a) and (b). Figure 10.7(c) and (d) shows the manually extracted roads for comparison.

Experiments showed significant increases in road centerline extraction performance through two SSRC iterations. For the scene in Figure 10.7(a), the road extraction quality metric went from 0.67 (prior to running SSRC), to 0.83 on the first SSRC iteration, and 0.87 on the second. For the scene in Figure 10.7(b), the quality metric improved from 0.08, to 0.37, to 0.63 through two iterations (Doucette *et al.*, 2004). Dramatic improvement of the quality metric was tangible evidence to validate the potential benefits from the SSRC approach. Further observations revealed that the quality metric generally stabilised after two iterations of SSRC. Open research questions include: (1) determining how to modify parameters between iterations; and (2) establishing optimal convergence for SSRC iterations.

A limitation of the current method is that it assumes roads possess a spectrally unique signature throughout the input image. This signature can be confounded by errors introduced by topology construction, shadowing effects from trees and buildings, or variations in road composition. For example, a bimodal spectral signature for the road class could result if dirt and paved roads occur within the same scene. In this event, the spectral classifier would need to use an appropriate bimodal distribution model for roads to avoid suboptimal classification results. For future research consideration, we speculate that a supervised neural network classifier (e.g. radial basis function) may provide better results than statistical methods (e.g. MLC) to deal with effects from multimodal spectral signatures.

10.7   SUMMARY

The extraction of geospatial features from remotely sensed imagery remains the primary means to create or update geospatial databases. The growing availability of high spatial and spectral resolution imagery is driving a pressing need for robust automated image analysis algorithms. In this chapter we presented a compendium of work that considered the application of SOM learning principles toward the problem of automating road extraction.

Early work considered application of the original SOM algorithm. The objective was to use SOM topology that could represent elongated regions (ERs) by a mapping of their medial axes. In high spatial resolution imagery, the medial axis of an ER corresponded to a road centerline. This SOM application had significant limitations, which included high sensitivity to initial conditions and noise.

The self-organising road map (SORM) algorithm included modifications to the original SOM. The most important variation was the adoption of a batch updating *K*-means method. The *K*-means algorithm was shown to be essentially a winner-take-all (WTA) batch version of the SOM. The simpler *K*-means algorithm provided less sensitivity to network initialisation. The median statistic was used in place of the mean to better deal with noise in the input space. The classic problem of node validation was addressed with on-the-fly node pruning and merging. The use of a regular grid for node initialisation in the input space promoted good distribution of nodes, and fast convergence. Techniques

for speeding up convergence that were explored included local space searching and node activity searching.

Further modifications considered a variant of the original SOM that allowed node neighborhoods to adapt to the input space. Node metrics in the input space were used in a fuzzy inference system to construct and validate node topologies that were consistent with road network. Creation of fuzzy membership functions and inference rules were motivated by the Gestalt grouping principles of proximity, continuity, context, and closure.

The final version of SOM adaptations culminated with the self-supervised road classification (SSRC) algorithm. The goal of SSRC was to improve upon SORM-based extraction by incorporating spectral information into the process automatically. Training samples were automatically gathered from road topology nodes. The spectral statistics were used in a maximum likelihood classification. The entire process would repeat with edge analysis of the new road class. Experiments with images demonstrated significant increases in road centerline extraction performance over the first two iterations of SSRC. Determining how to establish optimal convergence remained an open research question.

A distinctive aspect of this SOM-based approach to road extraction was the use of node representations upon which to base topology construction. Node regions in images were useful in representing a higher level of abstraction for perceptual organisation. Self-organisation principles allowed for a dynamic partitioning of node regions, versus using fixed partitioning methods of the image space. Limitations of using SOM principles included sensitivity to the initial conditions, and relatively high computational requirements. An objective for future research is more detailed sensitivity analysis among the different processes of edge analysis, self-organisation, topology construction, and spectral analysis.

ACKNOWLEDGEMENTS

REFERENCES

Agouris, P., Doucette, P., Stefanidis, A. (2001a) Spatiospectral Cluster Analysis of Elongated Image Regions. In: International Conference on Image Processing, IEEE Signal Processing Society, Thessaloniki, Greece, pp. 789–792.

Agouris, P., Stefanidis, A., Gyftakis, S. (2001b) Differential Snakes for Change Detection in Road Segments. Photogrammetric Engineering & Remote Sensing, 67(12), pp. 1391–1400.

Barzohar, M., Cohen, M., Ziskind, I., Cooper, D. (1997) Fast Robust Tracking of Curvy Partially Occluded Roads in Clutter in Aerial Images. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), Gruen, A., Baltsavias, E., Henricsson, O. (eds), Birkhauser Verlag, Ascona, Switzerland pp. 277–286.

Baumgartner, A., Steger, C., Mayer, H., Eckstein, W., Ebner, H. (1999) Automatic Road Extraction Based on Multi-Scale, Grouping, and Context. Photogrammetric Engineering & Remote Sensing, 65(7), pp. 777–785.

Bordes, G., Giraudon, G., Jamet, O. (1997) Road Modeling Based on a Cartographic Database for Aerial image Interpretation. In: Semantic Modeling for the Acquisition of Topographic Information from Images and Maps, Birkhauser Verlag, Basel, pp. 123–139.

Doucette, P. (2002) Automated Road Extraction from Aerial Imagery by Self-Organisation. PhD Dissertation, Department of Spatial Information Science and Engineering, University of Maine, 238 p.

Doucette, P., Agouris, P., Musavi, M., Stefanidis, A. (1999) Automated Extraction of Linear Features from Aerial Imagery Using Kohonen Learning and GIS Data. Lecture Notes in Computer Science, Springer-Verlag, Portland, ME, Volume 1737, pp. 20–33.

Doucette, P., Agouris, P., Musavi, M., Stefanidis, A. (2000) Road Centerline Vectorization by Self-Organised Mapping. International Archives of Photogrammetry and Remote Sensing, 33(B3), pp. 246–253.

Doucette, P., Agouris, P., Stefanidis, A., Musavi, M. (2001) Self-Organised Clustering for Road Extraction in Classified Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 55(5–6), pp. 347–358.

Doucette, P., Agouris, P., Stefanidis, A. (2004) Automated Road Extraction from High Resolution Multispectral Imagery. *Photogrammetric Engineering & Remote Sensing*, 70(12), pp. 1405–1416.

Gersho, A., Gray, R. M. (1992) Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston.

Gonzalez, R., Woods, R. (2002) Digital Image Processing. Upper Saddle River, NJ.

Gruen, A., Li, H. (1997) Semi-Automatic Linear Feature Extraction by Dynamic Programming and LSB-Snakes. *Photogrammetric Engineering & Remote Sensing*, 63(8), pp. 985–995.

Hall, D., Ball, G. (1965) ISODATA: A Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute, Menlo Park, CA.

Harvey, W. (1999) Performance Evaluation for Road Extraction. *Bulletin de la Societe Francaise de Photogrammetrie et Teledetection*, 153(1999-1), pp. 79–87.

Havercamp, D. (2002) Extracting Straight Road Structure in Urban Environments Using IKONOS Satellite Imagery. *Optical Engineering*, 41(9), pp. 2107–2110.

Haykin, S. (1999) Neural Networks. Upper Saddle River, NJ.

Hinz, S., Baumgartner, A. (2003) Automatic Extraction of Urban Road Networks from Multi-View Aerial Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1–2), pp. 83–98.

Jacobs, D., Lindenbaum, M. (guest eds) (2003) A Special Issue on Perceptual Organization in Computer Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4) and 25(6).

Kangas, J., Kohonen, T., Laaksonen, J. (1990) Variants of the Self-Organising Map. *IEEE Transactions on Neural Networks*, 1(1), pp. 93–99.

Kohonen, T. (1982) Self-Organised Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, pp. 59–69.

Kohonen, T. (1993) Things You Haven't Heard About the Self-Organising Map. In: Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, pp. 1464–1480.

Kohonen, T. (2001) Self-Organising Maps. Springer-Verlag, Berlin.

Linde, Y., Buzo, A., Gray, R. (1980) An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, COM-28, pp. 84–95.

Lloyd, S. (1982) Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28, pp. 127–135.

Lutrell, S. (1989) Self-Organisation: A Derivation from First Principles of a Class of Learning Algorithms. In: International Conference on Neural Networks, IEEE Washington, DC, pp. 495–498.

MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability, Berkeley, CA, 1, pp. 281–296.

Mckeown, D., Denlinger, J. (1988) Cooperative Methods for Road Tracking in Aerial Imagery. In: IEEE Proceedings of Computer Vision and Pattern Recognition, Ann Arbor, MI, pp. 662–672.

Mena, J. (2003) State of the Art on Automatic Road Extraction for GIS Update: a Novel Classification. *Pattern Recognition Letters*, 24, pp. 3037–3058.

Nevatia, R., Ramesh, B. (1980) Linear Feature Extraction and Description. *Computer Vision, Graphics, and Image Processing*, 13, pp. 257–269.

Price, K. (2000) Urban Street Grid Description and Verification. In: Proceedings of the 5th IEEE Workshop on Applications of Computer Vision, Palm Springs, CA, pp. 148–154.

Saleh, R. (guest ed.) (2004) Special Issue on Linear Feature Extraction from Remote Sensing Data for Road Network Delineation and Revision. *Photogrammetric Engineering & Remote Sensing*, 70(12).

Shackelford, A., Davis, C. (2003) Urban Road Network Extraction from High-Resolution Multi-spectral Data. In: 2nd GRSS/ISPRS Joint Workshop on Data Fusion and Remote Sensing over Urban Areas, Berlin, Germany, pp. 142–146.

Steger, C., Mayer, H., Radig, B. (1997) The Role of Grouping for Road Extraction. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), Gruen, A., Baltsavias, E., Henricsson, O (eds), Birkhauser Verlag, Ascona, Switzerland, pp. 245–256.

Tou, J., Gonzalez, R. (1974) Pattern Recognition Principles. Addison-Wesley, Reading, MA.

Trinder, J., Li, H. (1995) Semi-Automatic Feature Extraction by Snakes. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images, Gruen, A., Kuebler, O., Agouris, P. (eds), Birkhauser, Basel, pp. 95–104.

Vosselman, G., Knecht, J. (1995) Road Tracing by Profile Matching and Kalman Filtering. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images, Gruen, A., Kuebler, O., Agouris, P. (eds), Birkhauser, Basel, pp. 265–274.

Walter, V., Fritsch, D. (1998) Automatic Verification of GIS Data Using High Resolution Multispectral Data. International Archives of Photogrammetry and Remote Sensing, 32(3), pp. 485–489.

Wiedemann, C., Heipke, C., Mayer, H., Jamet, O. (1998) Empirical Evaluation of Automatically Extracted Road Axes. In: Empirical Evaluation Methods in Computer Vision, Bowyer, K., Phillips, P. (eds), IEEE Computer Society Press, Santa Barbara, CA, pp. 172–187.

Zadeh, L. (1965) Fuzzy Sets. Information and Control, 8, pp. 338–353.

Zahn, C. (1971) Graph-Theoretic Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, C-20(1), pp. 68–86.

Zhang, C., Baltsavias, E., Gruen, A. (2001) Knowledge-Based Image Analysis for 3D Road Reconstruction. *Asian Journal of Geoinformatics*, 1(4), pp. 3–14.

Zlotnick, A., Carnine, P. (1993) Finding Road Seeds in Aerial Images. *Computer Vision, Graphics, and Image Processing*, 57(2), pp. 243–260.

11

Epilogue: Intelligent Systems for GIScience: Where Next?
A GIScience Perspective

Michael Goodchild
NCGIA, University of California, Santa Barbara, CA 93106, USA

The self-organizing map (SOM) algorithm was originally designed to explore complex multidimensional data spaces describing objects that may or may not be located in geographic space, and whose locations may or may not be known. The subject matter of GIScience is thus only a small subset of the subject matter of SOM, since GIScience starts with the assumption that all objects of interest are georeferenced, and goes on to address fundamental issues underlying such information: its nature, representation, storage, handling, analysis, visualization, and modeling (for a recent review of the definitions and content of GIScience see Mark, 2003).

The chapters of this book have explored many facets of what is clearly a complex relationship between SOM and GIScience. In Chapter 2, Bação *et al*. show how the basic SOM algorithm can be modified to recognize location explicitly, and used to solve certain longstanding problems in GIScience, including regionalization and service area design, through modifications to the basic algorithm. Other authors see SOM in much the same way as its originator and early proponents, as a method for exploring complex multidimensional data sets, and for operationalizing the fundamental scientific task of classification. If the basic objects of analysis are georeferenced, then classes can be mapped, as Kropp and Schellnhuber do, for example, in Chapter 9.

Other authors see SOM as a means of *spatializing* data, in other words organizing objects in a space defined by their similarities, in effect enlisting SOM as a technique for

adding locational references to data that are not inherently spatial, and thus extending the techniques of GIScience to spaces other than the geographic. Such applications represent a remarkable implementation of Tobler's First Law of Geography (Tobler, 1970), namely the empirical tendency for objects that are nearby in geographic space to be more similar than objects that are distant in geographic space, by insisting that an information space created by spatialization have the properties that we as humans recognize to be true of geographic space. Montello *et al*. (2003) have proposed a First Law of Cognitive Geography, based on the observation that people *think* things are more similar if they are near each other, a finding that provides direct support for the logic of spatialization as a tool for visualizing complex multidimensional data sets.

Many years ago Tobler and Wineberg (1971) provided yet another motivation for techniques like SOM that position objects in a readily visualized space. Their interest lay in the locations of ancient settlements in Cappadocia, some of which were known and some unknown. Measures of interaction between settlements were available, in the form of counts of the numbers of tablets found in one settlement that mentioned another settlement. On the basis that interaction would decline systematically with distance once appropriate normalizations had been applied, Tobler and Wineberg were able to use metric scaling methods to estimate the missing locations.

In the late 1960s and early 1970s social scientists were just beginning to appreciate the power of digital computers to support a vast range of new, more complex, and more powerful methods of analysis. Before that time, methods of analysis were essentially manual, with the aid of printed tables of standard statistical distributions, electric calculators, and mechanical sorting machines. Multivariate methods such as factor analysis had been invented many decades previously, but their methods were fundamentally compromised by the lack of powerful machines to do the necessary matrix inversions. Rigid assumptions, such as the normality of distributions, had to be imposed to make analysis tractable, whether or not they were supported by the data; and metrics such as variance were preferred over possibly more useful alternatives simply because they made the analysis feasible with the computational resources of the time.

Today, of course, we are blessed with an abundance of techniques that exploit cheap, powerful computing capabilities to do things with data that were scarcely conceivable before 1970. We also have geographic information systems that make it easy to acquire, store, analyze, and visualize georeferenced information. More fundamentally, perhaps, these new methods have shifted the paradigms of science significantly, towards larger data sets, and a greater emphasis on the exploratory methods and induction over confirmatory methods and deduction, as the chapters in this book make clear. All of this is consistent with a widespread belief that the simple problems of science have been solved, and that further progress will require a new kind of science that emphasizes collaboration between disciplines, fueled by a search for elusive patterns in complex, multidimensional data sets.

A similar series of transitions are evident in the evolution of GIS. In the early days of the 1960s and 1970s, the focus of developers was on the data structures needed to represent the contents of maps in computers, and in the simplest kinds of analysis – measurement of area, for example. Later, the functionality of GIS expanded to include a substantial fraction of the known methods of spatial analysis, such as metrics for the measurement of spatial autocorrelation, tests of the randomness of point patterns,

methods of spatial interpolation, and methods of density estimation (Longley *et al*., 2005). The ideas of exploratory spatial data analysis (ESDA) originated in the late 1980s and early 1990s, and were implemented in specialized packages such as Regard (Anselin, 1999; Unwin, 1994) and more recently GeoDa (Anselin *et al*., 2006; geoda.uiuc.edu). However, even today the mainstream GIS products support only a small fraction of these ideas.

Essentially, ESDA seeks to create an intuitive, easy-to-use interface to geographic information that encourages exploration, and makes it possible for users to discover patterns and anomalies in data that would not otherwise be apparent. As such, the tests of its success seem to have much in common with other mainstream software environments, and little with traditional GIS, which is known for its complexity and the long training needed to extract useful results. Indeed, ESDA may resemble the kinds of GIS-derived products that are now available in the general marketplace, such as Google's Keyhole (http://earth.google.com/) whose user interfaces have more of the look and feel of a video game than a piece of software designed for serious scientific research, and on which designers expect users to move from complete ignorance to mastery in a few minutes, rather than years.

With this background, it is possible to see a little further into the future of SOM and GIScience. Tools such as SOM, suitably adapted for regionalization, zone design, or classification, or simply for the exploration and visualization of structure within massive data sets, would be valuable additions to the GIS toolbox, and would help in a more general process of moving to simpler, more intuitive user interfaces. At the same time, they raise issues of fundamental significance that are logically part of the GIScience research agenda: what are the appropriate methods of representation of SOM inputs and results; what are the appropriate designs of user interfaces and visualization methods; how should one deal with time, dynamics, and uncertainty; what is the appropriate approach to effects of scale; and how can SOM results be made available for further analysis as part of the GIS database? A strong thread running through the chapters of this book suggests that SOM provides more than a single addition to the spatial analytic toolbox, but instead reflects an entirely new paradigm for ESDA and spatial data mining. If so, does this suggest that it should be implemented in a stand-alone toolbox rather than as part of GIS functionality, and how does the general trend in the GIS software industry to component ware impact this issue? In short, the advent of SOM, and the thinking that lies behind the chapters of this book, raise important questions for GIScience and add significantly to the GIScience research agenda.

REFERENCES

Anselin, L., 1999. Interactive techniques and exploratory spatial data analysis. In P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, editors, *Geographical Information Systems: Principles, Techniques, Management and Applications*. New York: John Wiley & Sons, Ltd, pp. 253–266.

Anselin, L., I. Syabri, and Y. Kho, 2006. GeoDa: an introduction to spatial data analysis. *Geographical Analysis* 38(1): 5–22.

Longley, P.A., M.F. Goodchild, D.J. Maguire, and D.W. Rhind, 2005. *Geographic Information Systems and Science*, 2nd Edition. New York: John Wiley & Sons, Ltd.

Mark, D.M., 2003. Geographic information science: defining the field. In M. Duckham, M.F. Goodchild, and M.F. Worboys, editors, *Foundations of Geographic Information Science*. New York: Taylor & Francis, pp. 3–18.

Montello, D.R., S.I Fabrikant, M. Ruocco, and R.S. Middleton, 2003. Testing the first law of cognitive geography on point-display spatializations. In W. Kuhn, M. Worboys, and S. Timpf, editors, *Spatial Information Theory: Foundations of Geographic Information Science*. Lecture Notes in Computer Science 2825. Berlin: Springer, pp. 316–331.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.

Tobler, W.R. and S. Wineberg, 1971. A Cappadocian speculation. *Nature* 231(5297): 39–42.

Unwin, A., 1994. REGARDing geographic data. In P. Dirschedl and R. Ostermann, editors, *Computational Statistics*. Heidelberg: Physica, pp. 345–354.

Index

Plate 1 *Process of self-organization during SOM training. A 3 × 3 neuron SOM is trained with four observations representing two distinct groups in attribute space (See Figure 1.6)*
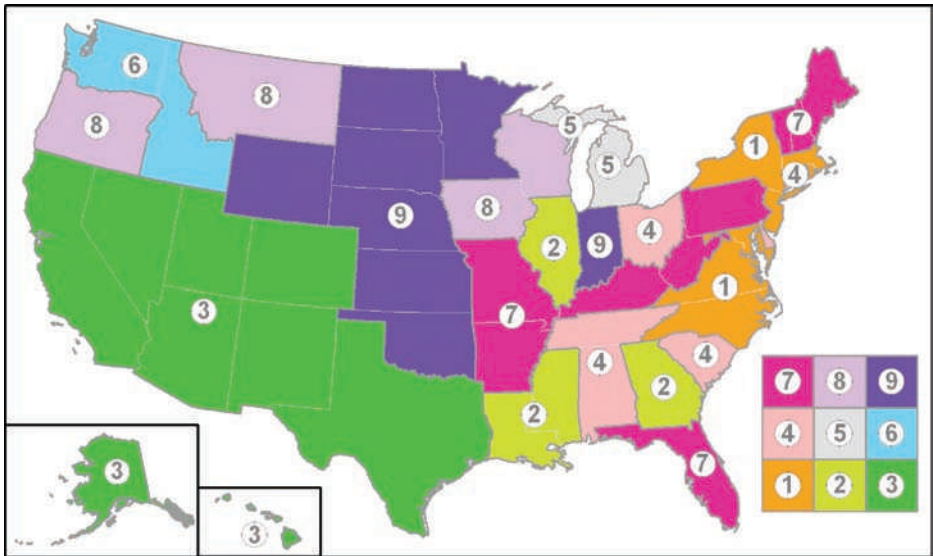
Plate 2 *SOM-based clustering of census data combined with colour design informed by network topology. Relationships among clusters are indicated by displaying legend constructed from two-dimensioned SOM geometry (See Figure 1.9)*
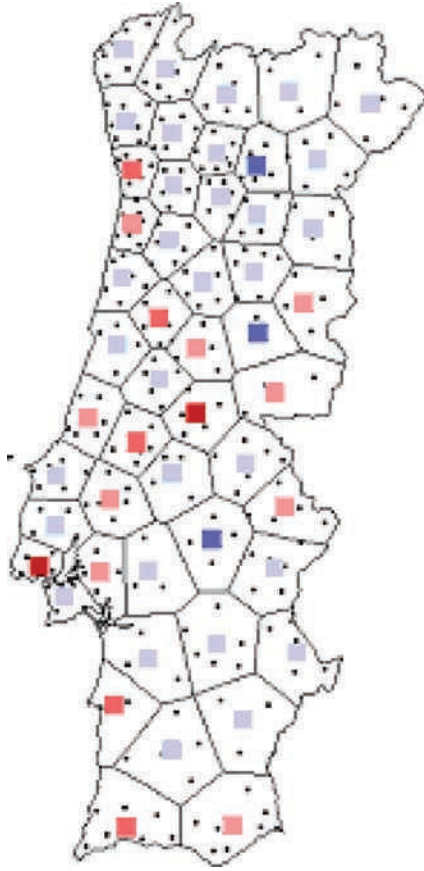
Plate 3 *Example of a Geographical Hypermap seen in the input (geographical) space (See Figure 2.3)*

Subset 1
$z \approx 10$

Subset 2
$z \approx 90$

High quantization area, using the Geo-SOM (0), thus no clustering of similar points

The Geo-SOM (1), detects two low quantization sets of points, thus clustering two groups of similar points in the same area
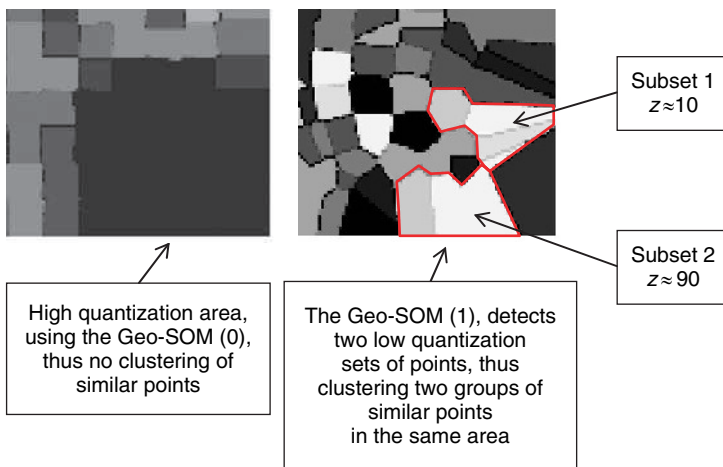
Plate 4 *Close up of the lower right corner of the Geo-SOM (0) and Geo-SOM (1) in Figure 2.7 (See Figure 2.8)*
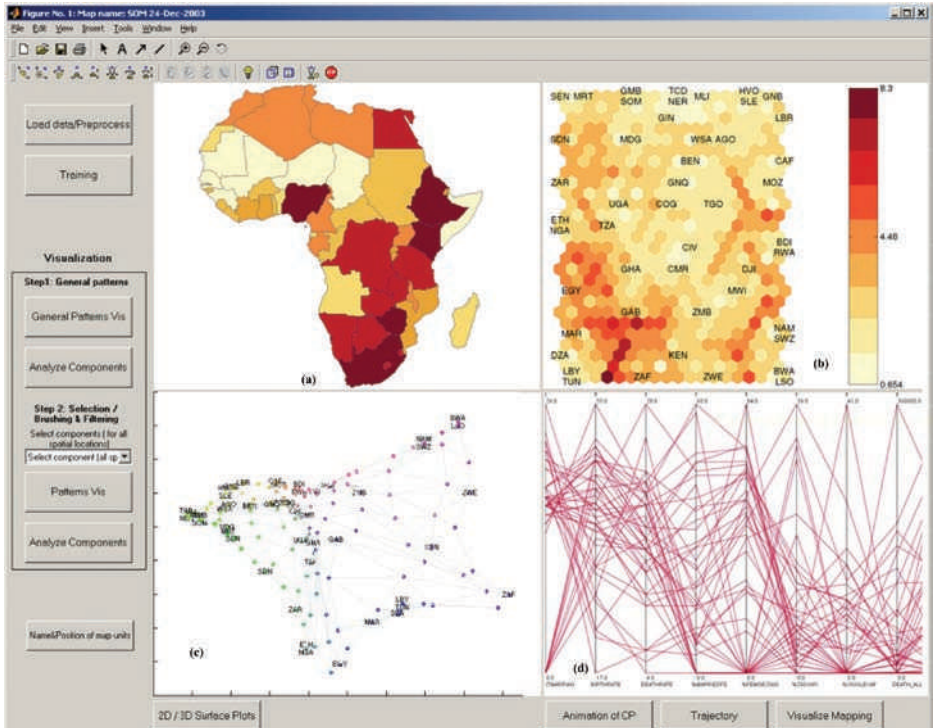
Plate 5 *The user interface for the exploratory geovisualization environment in multiple views with the visualization of component planes (bottom left) and map unit labels (bottom right). The default view shows the representation of the general patterns and clustering in the input data: the unified distance matrix showing clustering and distances between positions on the map (b). Alternative representations of the SOM general clustering of the data with projection of the SOM results in 3-D space (c); and a map of the similarity coding extracted from the SOM computational analysis (a), and parallel coordinate plot (d) (See Figure 3.5)*

Plate 6   *Component planes of the node triplets associated with each lexical variation of the experimental SOM run (See Figure 5.3)*
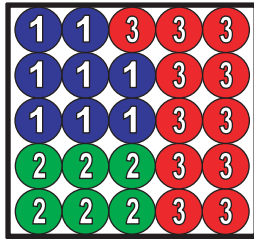
Legend

Average Cluster Weights	Clusters		
	1	**2**	**3**
Quarter of / positive	0.84	0.65	0.01
Quarter of / negative	0.02	0.07	0.82
Quarter of / missing	0.00	0.00	0.00
Pretty day / positive	0.73	0.82	0.53
Pretty day / negative	0.17	0.07	0.39
Pretty day / missing	0.00	0.00	0.00
Dog irons / positive	0.85	0.90	0.88
Dog irons / negative	0.11	0.00	0.06
Dog irons / missing	0.00	0.00	0.00
Shelf / positive	0.97	1.00	0.80
Shelf / negative	0.02	0.00	0.17
Shelf / missing	0.00	0.00	0.00
Kindling / positive	0.87	0.00	0.14
Kindling / negative	0.10	0.98	0.62
Kindling / missing	0.00	0.00	0.00
Blinds / positive	0.84	0.99	0.77
Blinds / negative	0.04	0.00	0.21
Blinds / missing	0.00	0.00	0.00

Input Nodes (row label)

Plate 7 *Feature map of the three-cluster solution (See Figure 5.4)*

Legend

	Clusters			
Average Cluster Weights	**1**	**2**	**3**	**4**
Quarter of / positive	0.69	0.83	0.01	0.00
Quarter of / negative	0.07	0.02	0.74	0.89
Quarter of / missing	0.00	0.00	0.00	0.00
Pretty day / positive	0.83	0.70	0.08	0.90
Pretty day / negative	0.07	0.19	0.81	0.03
Pretty day / missing	0.00	0.00	0.00	0.00
Dog irons / positive	0.85	0.90	0.93	0.83
Dog irons / negative	0.06	0.06	0.03	0.10
Dog irons / missing	0.00	0.00	0.00	0.00
Shelf / positive	1.00	0.97	0.95	0.67
Shelf / negative	0.00	0.02	0.02	0.30
Shelf / missing	0.00	0.00	0.00	0.00
Kindling / positive	0.06	0.93	0.22	0.07
Kindling / negative	0.92	0.03	0.55	0.68
Kindling / missing	0.00	0.00	0.00	0.00
Blinds / positive	0.96	0.85	0.57	0.93
Blinds / negative	0.02	0.04	0.42	0.03
Blinds / missing	0.00	0.00	0.00	0.00

(Left axis label: Input Nodes)

Plate 8   *Feature map of the four-cluster solution (See Figure 5.9)*
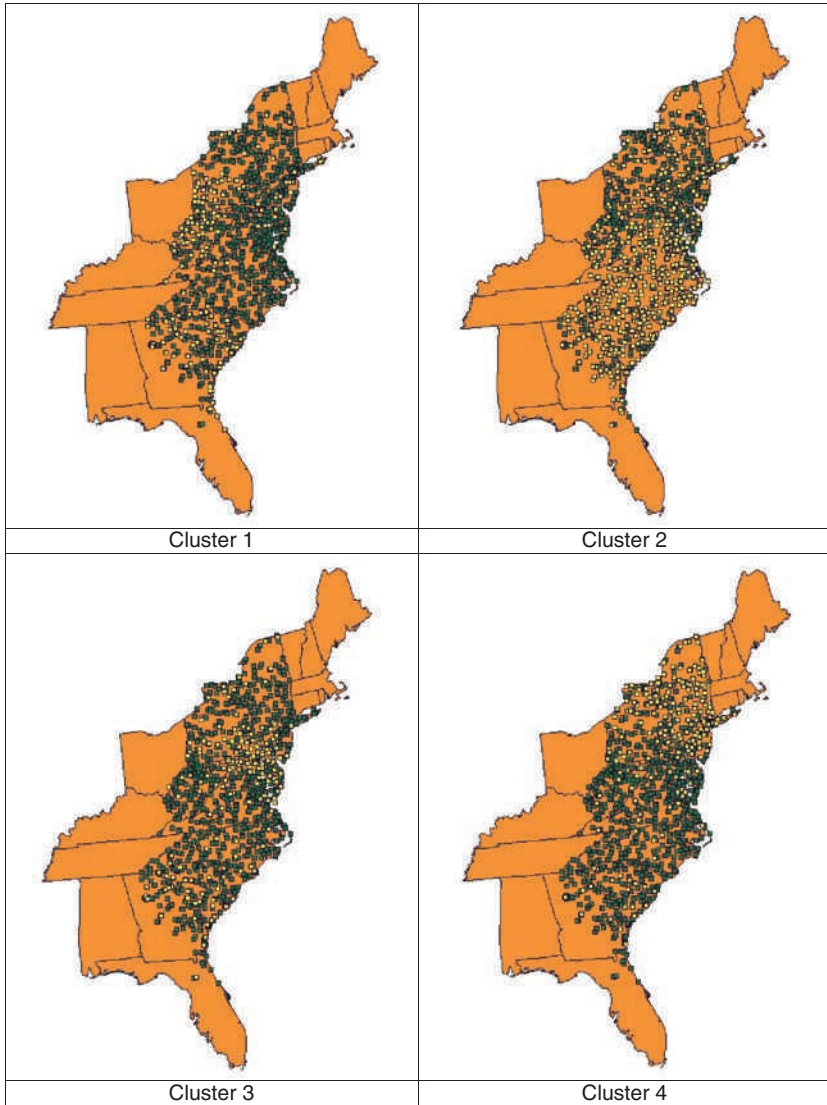
Plate 9 *Geographic distribution of informants mapped to the four clusters (clear squares) of the four-cluster solution (See Figure 5.10)*
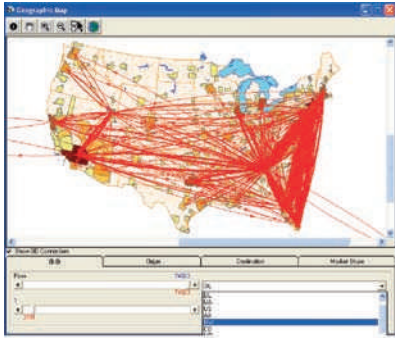
(a)
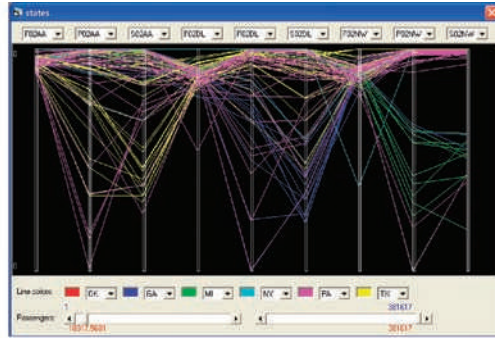
(b)
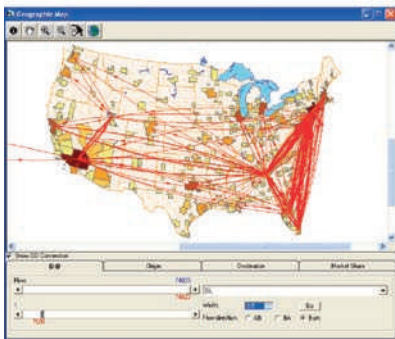
(c)
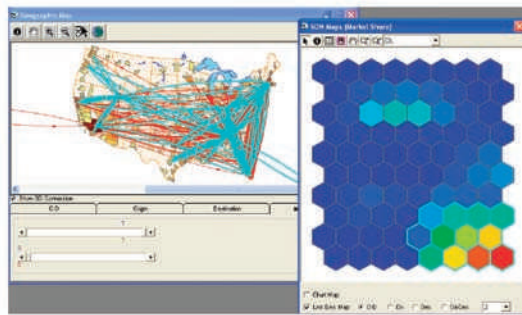
(d)

Plate 10 *Selected user interaction forms in the VDM environment. (a) Assignment: allows the analyst to visualize the passenger volume of different airlines. (b) Color manipulation: colors are used to draw the markets that originate in different geographic regions. (c) Focusing: two scroll bars are used so that the upper bound and lower bound of passenger flow can be changed dynamically in display. (d) Linking and brushing: highlighted in the geographic map are the markets represented by the neurons selected in the component plane (See Figure 4.2)*
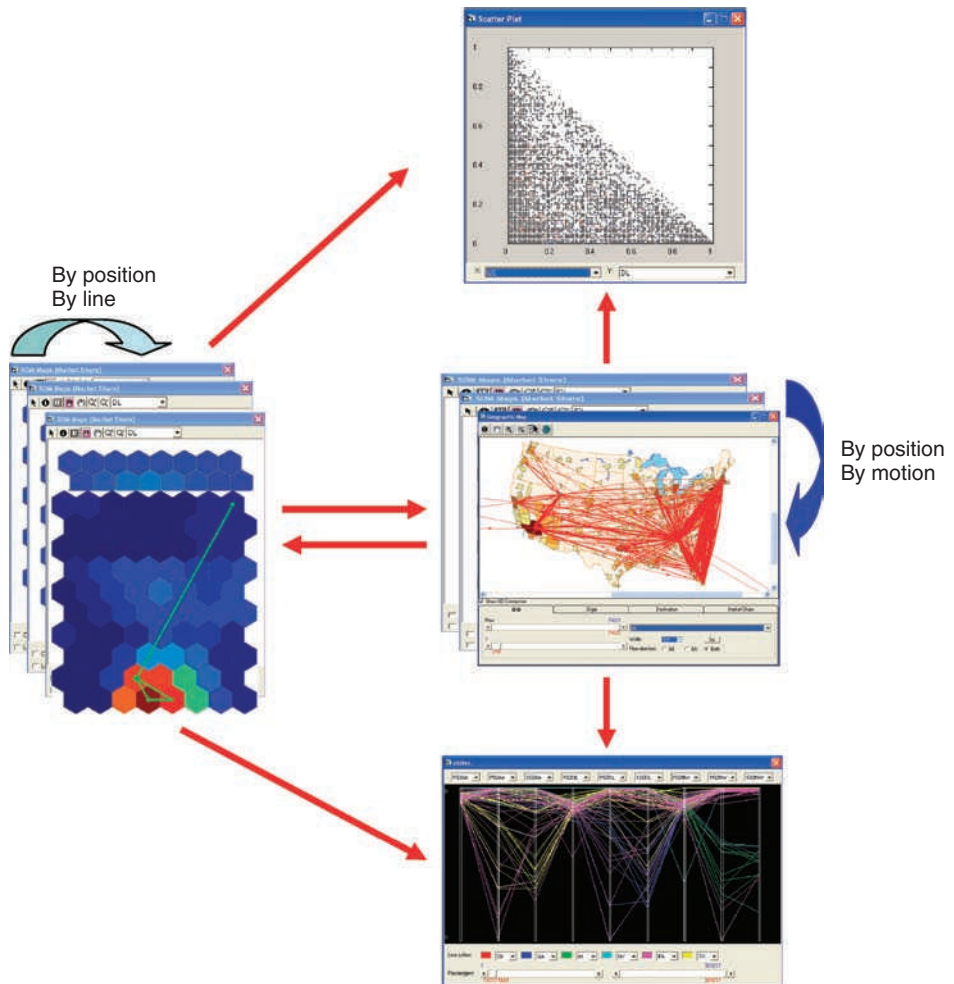
Plate 11 *Framework of the integrated VDM environment. The linkages among different visualization forms can be implemented mainly through four ways: by position (the position of data items remains fixed across visualization forms); by color (the same color is used for the same group of data items); by line (the same data items are connected by explicit lines); by motion (groups of data items are displayed one after another using animation) (See Figure 4.3)*
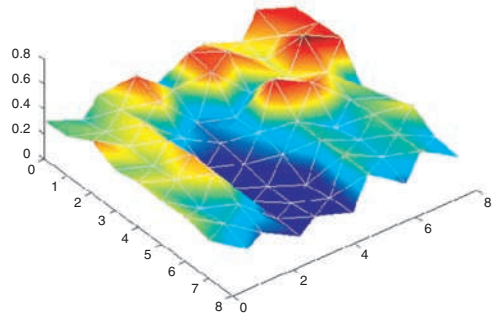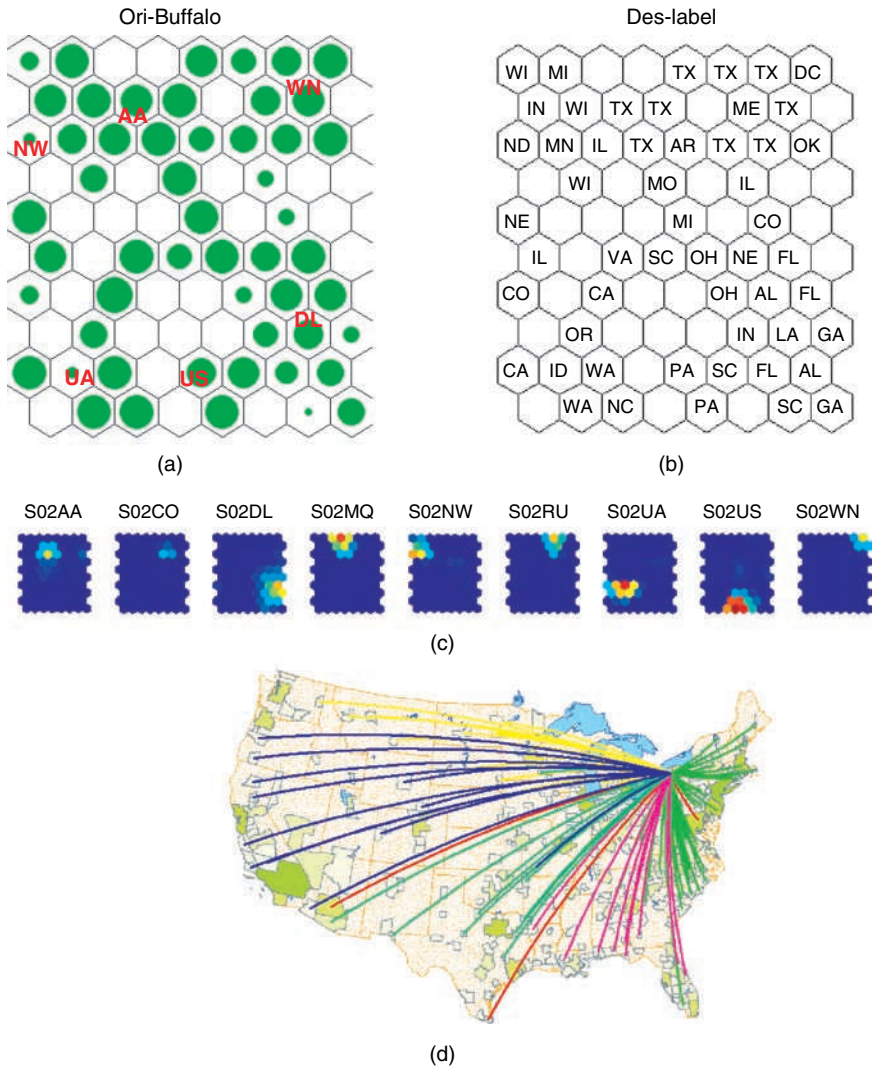
U-matrix

4   8   3   9-2

7   9-1

6   9-3   9-4

1

5   2   9-5

(a)

9 clusters

4   8   3

7   9

6

1

5   2
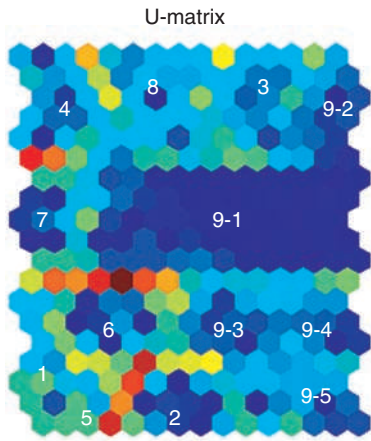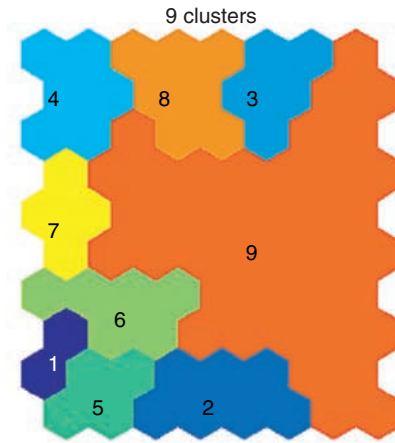
(b)
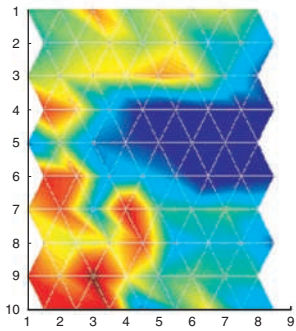
(c)

(d)

Plate 12 *Distance matrices and clusters based on 2002 market share information: (a) U-matrix; (b) clusters identified by k-means; (c) distance matrix (two-dimensional); (d) distance matrix (three-dimensional) (See Figure 4.4)*
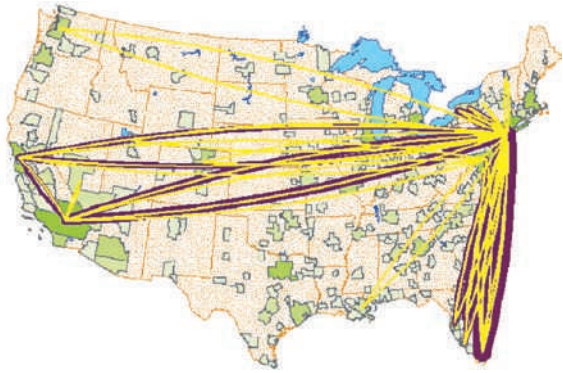
Ori-Buffalo

Des-label

(a)

(b)

S02AA S02CO S02DL S02MQ S02NW S02RU S02UA S02US S02WN

(c)

(d)

Note: Color Coding: AA DL NW UA US WN

Plate 13 Markets originating in the Buffalo metropolitan area: (a) hit counts; (b) most frequent destination state; (c) selected SOM component planes; (d) markets with airline market share > 50% (See Figure 4.6)
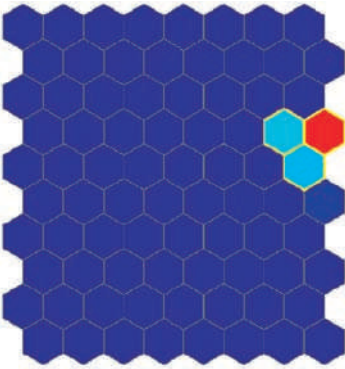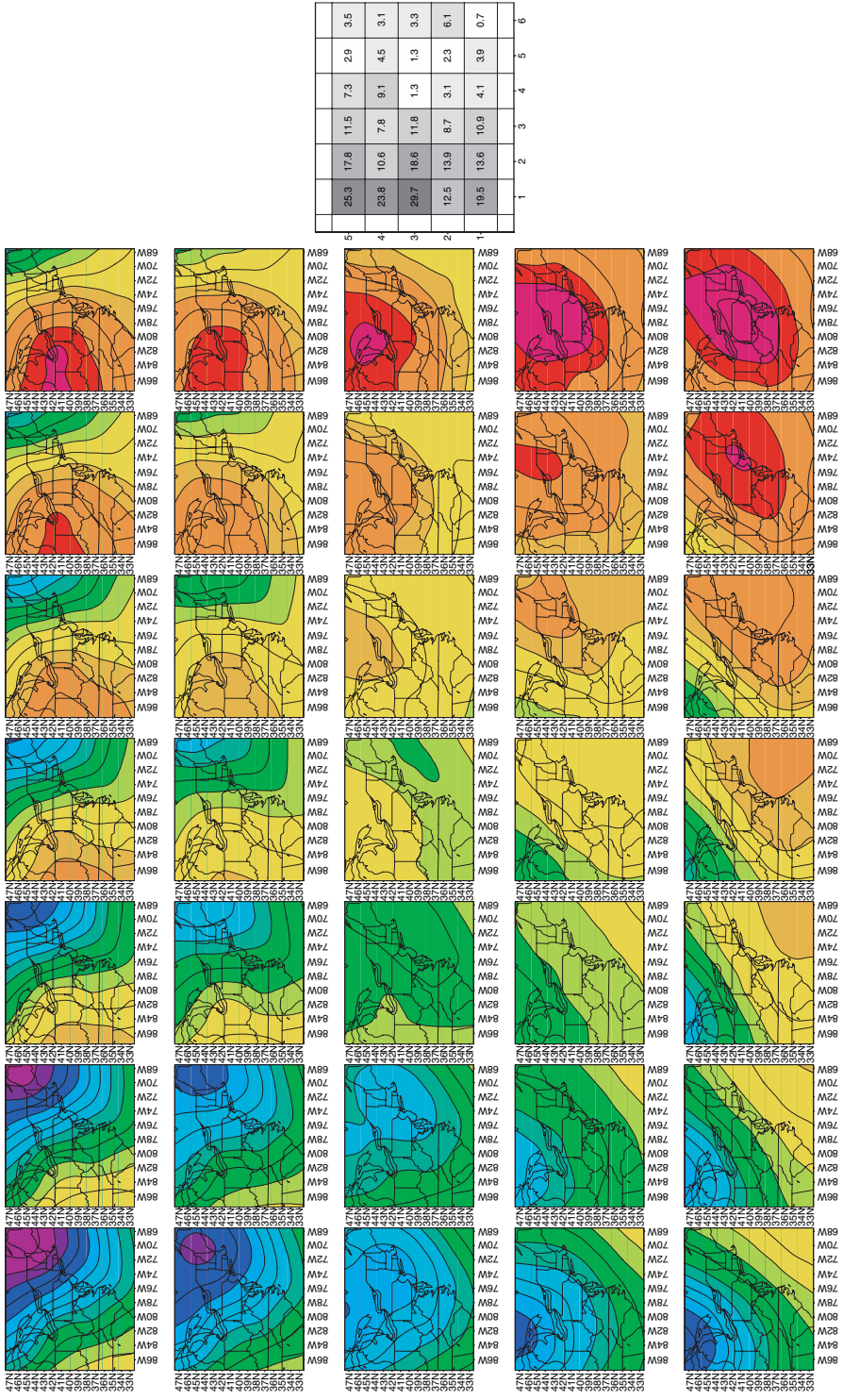
(a) (b)

Plate 14 *Markets of JetBlue identified by SOM: (a) market share component plane; (b) flow map (passengers ≥ 20 000) (See Figure 4.7)*

	1	2	3	4	5	6
5	25.3	17.8	11.5	7.3	2.9	3.5
4	23.8	10.6	7.8	9.1	4.5	3.1
3	29.7	18.6	11.8	1.3	1.3	3.3
2	12.5	13.9	8.7	3.1	2.3	6.1
1	19.5	13.6	10.9	4.1	3.9	0.7

Plate 15 *The 5 × 7 array of SOM node vectors of January sea-level pressure (SLP) for the north-east United States. Blues represent relatively low pressure, while reds indicate high pressure. The plot to the right displays the mean precipitation (mm) for each synoptic state represented in the SOM array. (After Hewitson and Crane, 2002) (See Figure 8.1)*
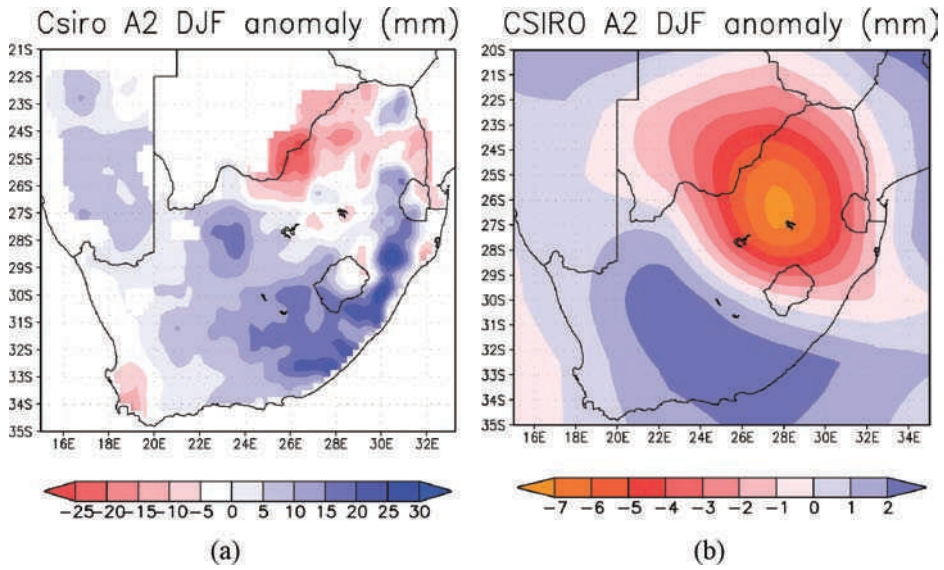
Plate 16 *SOM-based downscaling (a) and raw GCM (b) precipitation anomalies of climate change projections for the period 2071–2100 over South Africa (See Figure 8.5)*
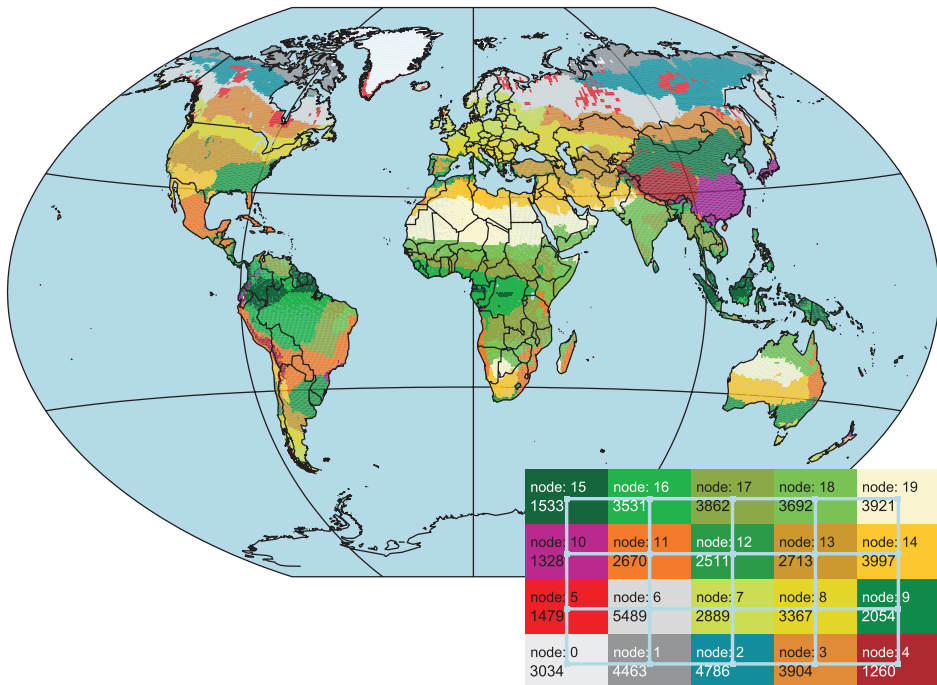

Plate 17 *Global distribution of classes (represented by the different nodes) obtained after simulation with the SOMTOP algorithm. The topological arrangement of the classes on the network is shown by the inset. The colour coding used in the inset corresponds to that shown on the map and the node numbers are equivalent to the class numbers used in the text. The additional number indicates the quantity of associated input vectors (See Figure 9.4)*
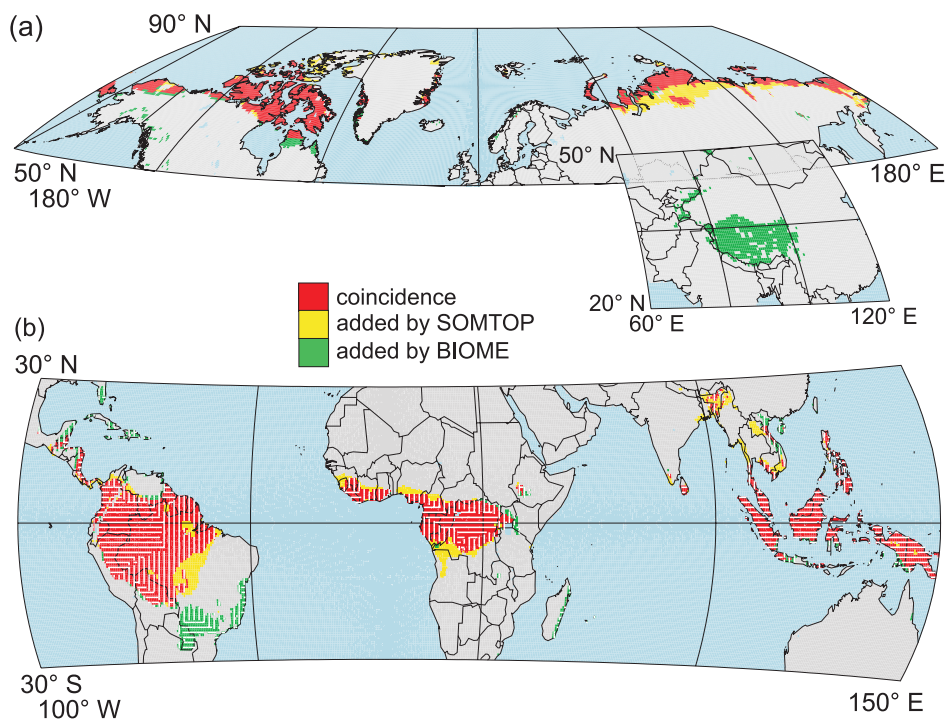
Plate 18 Comparison between SOMTOP categorisations and biome types in the arctic and tropics domain. (a) For class #1 (SOMTOP) and the 'tundra' biome – the inset represents the highlands (node 4), and (b) for the aggregated rainforest division: SOMTOP classes #15 and #16; BIOME 'tropical rainforest' (horizontal hatched) and 'seasonal rainforest' (vertically hatched). Focusing on the single archetypes some larger differences are apparent in South America (regarding the SOMTOP classification in these regions compare with Figure 17) (See Figure 9.5)