USING SELF-ORGANIZING MAPS FOR INFORMATION VISUALIZATION AND KNOWLEDGE DISCOVERY IN COMPLEX GEOSPATIAL DATASETS

Koua, E.L.

International Institute for Geo-Information Science and Earth Observation (ITC).
P.O.Box 6, 7500 AA Enschede, The Netherlands. E-mail: koua@itc.nl

ABSTRACT

New data acquisition techniques are offering tremendous opportunities that result in more and more geospatial data to be handled. Analyzing these data becomes a difficult task due to the size of datasets, their complexity, scaling problems, and hidden patterns. The complexity of the attribute or feature space in such large datasets as well as related computational burdens, do not always allow traditional deductive and statistically based approaches to analysis to scale well to the data. Like many techniques applied to data analysis (statistical methods, artificial intelligence using rule-based systems and decision trees), Artificial Neural Networks are an emerging solution for data analysis and pattern recognition. Among the Neural Network models, Self-Organizing Map (SOM) is often seen as a promising technique for exploratory analysis of data. There is also an increasing need to organize and support user's information exploration in a way that reduces complexity and facilitates the acquisition of knowledge in such data rich environments. Ideally, users should be able to look at geospatial data in any combination, at any scale, with the aim of seeing or finding spatial relations and patterns. In addressing this issue, the Geo-information science community can learn and use approaches from other disciplines such as Information Visualization, which has the potential to help find information needed more effectively and intuitively. In this paper, we use the Self-Organizing Map algorithm to explore a geospatial dataset. The dataset consist of a collection of socio-economic indicators related to municipalities in a region of the Netherlands. The use of the SOM intends to uncover the structure and patterns from this dataset, and to provide graphical representations that can support understanding and knowledge construction. Spatial analysis, data mining and knowledge discovery methods are combined in this framework, with the goal of portraying the data in a visual form in order to stimulate pattern recognition and hypothesis generation. Some examples of these visual representations (information spaces) based on the SOM are explored: the visualization of clusters and shape of the data using a unified distance matrix visualization, projections (mesh visualization), visualization of component planes (multiple linked views of component planes), 2D and 3D surface plots of the distance matrices. These techniques use spatial metaphors such as distances, regions, and scale, to facilitate the representation of information.

The paper investigates the potential of these representations for exploratory data analysis, information visualization, and knowledge discovery. The ultimate goal is to design tools for visual representations that will allow users view appropriate underlying distributions, patterns, and therefore contribute to enhance the understanding of geospatial analysis results.

Keywords: Visualization, Self-Organizing Map, geospatial data exploration, exploratory data analysis

1. INTRODUCTION

The basic idea in analyzing geospatial data is to find patterns and relationships in data that can help solve a particular geo-problem. Problems are often associated with monitoring, managing, planning and decision-making related to a geo-phenomenon The limitation in human visual and cognitive processing restricted by large volumes of numbers and objects in a two dimensional map are real concerns in analyzing geospatial data as volumes of data increase.

The need for more accurate and flexible data analysis tools is therefore crucial as patterns may be hidden and trends analysis may be difficult under these conditions. It seems today that the geoscientist is more and more dealing with unknown data or data that she or he does not have enough knowledge of underlying relationships (8). As reported in (7), the increasing volume and diverse nature of digital geographic data easily overwhelm mainstream geospatial analysis techniques that are oriented towards the extraction of information from small and homogeneous datasets. Traditional analytical methods applied to geospatial data such as statistical methods have high computational burdens and are confirmatory techniques requiring the analyst to have prior hypotheses. Classification or pattern recognition methods are intended to compare the unknown pattern with all known reference patterns on the basis of some criterion for the

degree of similarity, in order to decide to which class the pattern belongs. In case of unknown data, it is not obvious what mechanisms or rules are behind the actual data or classes of interest. This makes it difficult for these techniques to help discover new knowledge and unexpected patterns, trends and relationships that can be hidden in very large geospatial datasets.

It is commonly argued that Self-Organizing Map (4) can allow the extraction of patterns and the creation of abstractions where conventional methods may be limited for analysis of data because underlying relationships are not clear and mechanisms or rules behind the actual data or classes of interest are not obvious.

This potential of SOM is explored for complex geospatial data in combination with visualization techniques to help in understanding of complexity, and to enhance the overall effectiveness of exploratory data analysis.

2. FRAMEWORK FOR SOM BASED VISUALIZATION TECHNIQUES

The development of the SOM based visualization tool intends to provide additional exploratory data analysis techniques by offering a tool that allows effective extraction and exploration of geospatial patterns. The tool relates different phases of the transformation process from data into information and knowledge within the framework of data mining and knowledge discovery.

2.1 The Self-Organizing Map

The Self-Organizing Map (4) is an Artificial Neural Network used to map high dimensional data onto a low dimensional space, usually a 2D-representation space. The network consists of a number of neural processing elements (neurons or nodes) usually arranged on a rectangular or hexagonal grid, where each neuron is connected to the input. The goal is to group similar nodes close together in certain areas of the value range. The resultant maps (SOMs) are organized in such a way that similar data are mapped onto the same node or to neighboring nodes in the map. This leads to a spatial clustering of similar input patterns in neighboring parts of the SOM and the clusters that appear on the map are themselves organized. This arrangement of the clusters on the map reflects the attribute relationships of the clusters in the input space. For example, the size of the clusters (the number of nodes allotted to each cluster) is reflective of the frequency distribution of the patterns in the input set. Actually, the SOM uses a distribution preserving property which has the ability to allocate more nodes to input patterns that appear more frequently during the training phase of the network configuration. It also applies a topology preserving property, which comes from the fact that similar data are mapped onto the same node, or to neighboring nodes on the map. In other words, the topology of the dataset in its n-dimensional space is captured by the SOM and reflected in the ordering of its nodes. This is an important feature of the SOM that allows the data to be projected onto a lower dimension space while roughly preserving the order of the data in its original space. Another important feature of the SOM for knowledge discovery in complex datasets is the fact that it is an unsupervised learning network, meaning that the training patterns have no category information that accompany them. Unlike supervised methods which learns to associate a set of inputs with a set of outputs using a training data set for which both input and output are known, SOM adopts a learning strategy where the similarity relationships between the data and the clusters are used to classify and categorize the data.

2.2 Design and usability framework

The use of SOM aims at contributing to the analysis and visualization of large amount of data, as an extension of the many geospatial analysis functions available in most GIS software. The design framework includes spatial analysis, data mining and knowledge discovery methods, integrated into an interactive visualization system. Users can perform a number of exploratory tasks to understand the structure of the dataset as a whole and also to explore detail information on individual or selected attributes of the dataset. Graphical representations are used to represent uncovered structure and patterns that may be hidden in the dataset and to support understanding and knowledge construction.

To develop a tool that is useful and appropriate for the user needs and tasks, a usability framework is used to guide the design of the visualization tool. This approach not only uses the techniques, processes, methods, and procedures for designing usable products and systems, it focuses on the user's goals, needs and tasks in the design process (9). User characteristics, visualization tasks and operations are examined to improve user interaction and to support cognitive activities involved in the use of the visualization tool, and in related graphical representations.

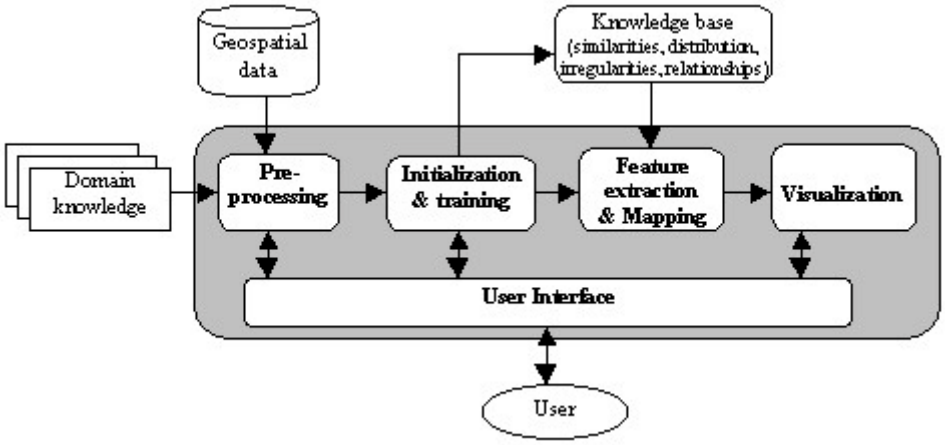

Figure 1. Architecture of the visualization tool

2.3 Architecture

The architecture of the tool includes 4 main components: pre-processing, initialization and training, feature extraction, and visualization (Figure 1). Each of these components corresponds to a logical step in the implementation of the SOM network. The pre-processing consists of the transformation of primary data and the conversion into appropriate format. At this stage, input data are transformed and all components and variables of the dataset are normalized. The initialization and training stage provides the basis of the performance of the network to adapt in solving problems. The visualization component provides tools for visualizing the representation of the data structure and extracted patterns using different techniques.

3. EXPLORING SOM BASED VISUALIZATION TECHNIQUES

The main goals in the acquisition of knowledge as intended by the design of the SOM based visualization tool, are discovery, decision-making, problem solving and explanation. A visualization framework (Figure 2) is used to increase the ability to perform cognitive activities necessary to achieve these goals.

3.1 Information Visualization and Geovisualization

Visualization was defined by (1) as the use of computer-supported, interactive, visual representations of data to amplify cognition. This definition suggests six ways in which visualization can amplify cognition: increasing the memory and processing resources available to the users, reducing the search for information, using visual representations to enhance the detection of patterns, enabling perceptual influence operations, using perceptual attention mechanisms for monitoring, and encoding information in a manipulable medium. Visualization in scientific computing emerged from the computer science field with emerging fields such as Scientific Visualization and Information Visualization. Information visualization research is concerned with graphically representing complex, abstract data domains to facilitate knowledge extraction from very large non-spatial data. It is recognized to have the potential to enable better understanding of complex systems, and to allow the discovery of information that might overwise remain unknown. Representations derived from information visualization research often use geographic metaphor to structure human-computer interaction, and are commonly referred as spatializations or information spaces (2) most of which are generated outside the GIScience and cartography disciplines. New forms of visual representation for geographic data have also emerged with emphasis on interaction and dynamics as an attempt to give a response to the increasing needs of users. In this respect, cartographic research efforts on visualization have been extended to meet other research activities in information science disciplines. This recognition was put forward with the creation of a commission on Visualization in 1995 within the International Cartography Association (ICA), which became later the commission on Visualization and virtual environments (6). The objectives of research within this commission was to cope with the increasing volume of geospatial data by developing theory and practice that facilitate knowledge construction through visual exploration and analysis of geospatial data as well as emphasising on the visual tools necessary to support knowledge retrieval, synthesis and use (6). Geovisualization (visualization applied to geospatial data) can be considered as the core disciple for understanding complex phenomena and processes, structures and relationships in complex geospatial datasets. It includes the use of a variety of techniques from several disciplines such as scientific visualization, GIS and cartography to explore data and to answer questions, generate hypotheses, develop solutions and construct knowledge (3).

3.2 SOM based computational analysis and visualization framework

The use of the SOM is intended to provide additional exploratory data analysis techniques for complex geospatial data. For the user, the main goal is the acquisition of knowledge through discovery, for decision-making, problem solving

and explanation. The SOM is used as a combination of clustering and projection techniques for features extraction, visualization and interpretation of large high-dimensional datasets. The first level of the computation provides ground for extracting patterns from the data. Resultant maps (SOMs) are then visualized using graphical representations. These representations use visual variables (size, value, texture, color, shape, orientation) added to the position property of the map elements. Different visualization techniques are added to enhance the representations, including projections such as Sammon's mapping (10) and Principal Component Analysis, multiple views and 3D views. Metaphors are used to guide user exploration and interpretation of the representation (2). These metaphors are combined with forms of representation and the use of visual variables to enhance exploration and interpretation through user interaction (brushing, viewpoint, focusing…) in the information spaces. The information spaces suggest and take advantage of the metaphor characteristics such as '*near=similar, far=different'* (5), which is epitomized by Tobler's first law of geography (11). Various types of map representations can be used, including volumes, surfaces, points and lines. This allows exploration of relationships between items. A coordinate system enables to determine distance and direction, from which other spatial relationships (size, shape, density, arrangement, etc.) may be derived. Multiple levels of detail allow exploration at various scales, creating the potential for hierarchical grouping of items, regionalization and other types of generalizations. Figure 2 shows the computational analysis and visualization framework.

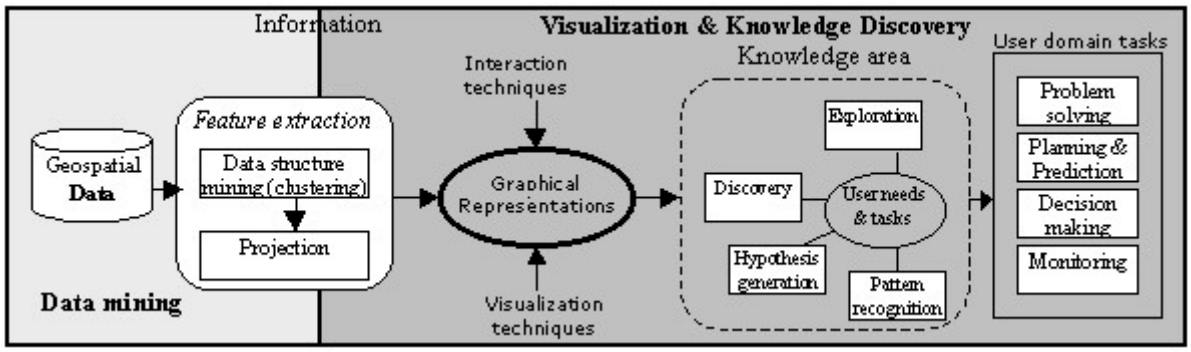

Figure 2. SOM based data exploration and visualization framework

3.3 SOM visualization techniques

Based on the Self-organizing Map algorithm, five main visualization techniques were explored: a cell visualization or U-matrix (a distance matrix visualization), projections (mesh visualization), visualization of component planes (in a multiples linked view), 2D and 3D surface plot of distance matrices. These techniques use spatial metaphors such as distances, regions, and scale, to facilitate the representation of information.

The data used in the training of the SOM network is a collection of socio-economic indicators related to municipalities in Overijssel region in the Netherlands (Figure 3). The idea is to find patterns and relationships among these municipalities. The dataset consist of 29 variables including population and habitat distributions, urbanization indicators, income of inhabitants, family and land data, as well as industrial, commercial and non-commercial services data.

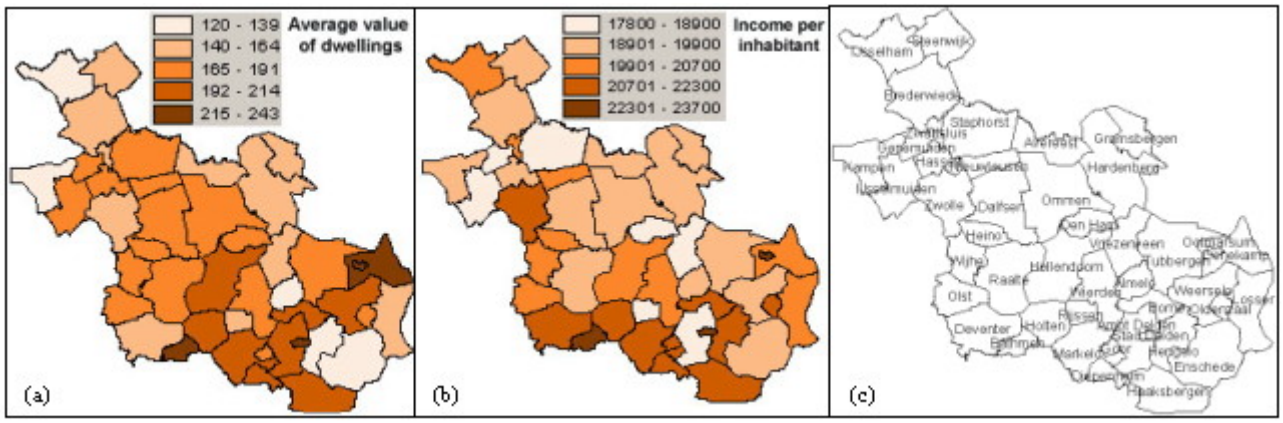

Figure 3. Two attributes of the Overijssel region dataset: (a) average value of dwellings, (b) income per inhabitant and (c) a map with the name of municipalities

3.3.1 Unified Distance Matrix representation for visualizing clusters

The unified distance matrix or U-matrix (12) is a representation of the Self-Organizing Map that visualizes the distances between the network neurons or units. It contains the distances from each unit center to all of its neighbors. The neurons of the SOM network are represented here by hexagonal cells. The distance between the adjacent neurons is calculated and presented with 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 U-matrix representation (Figure 4) reveals the clustering structure of the dataset explored in this experiment. Municipalities having similar characteristics are arranged close to each other and the distance between them represents the degree of similarity or dissimilarity.

For example, the municipality of Enschede is well separated from the rest by the dark cells showing a long distance from the rest of the municipalities. This is the case as this municipality is the largest, the most developed and urbanized in the region. On the top left corner of the map, municipalities Genemuiden, Rijssen, Staphorst, Ijsselmuiden are clustered together. These are small localities that have common characteristics according to the attributes of the dataset. For example, these municipalities turn out to be those with strong protestant religion communities, although the data did not provide variables on religion. This kind of similarities can be composed of a number of other variables provided by the dataset. The U-matrix shows more hexagons than the component planes because it shows not only the distance value at the map units but also the distances between map units.

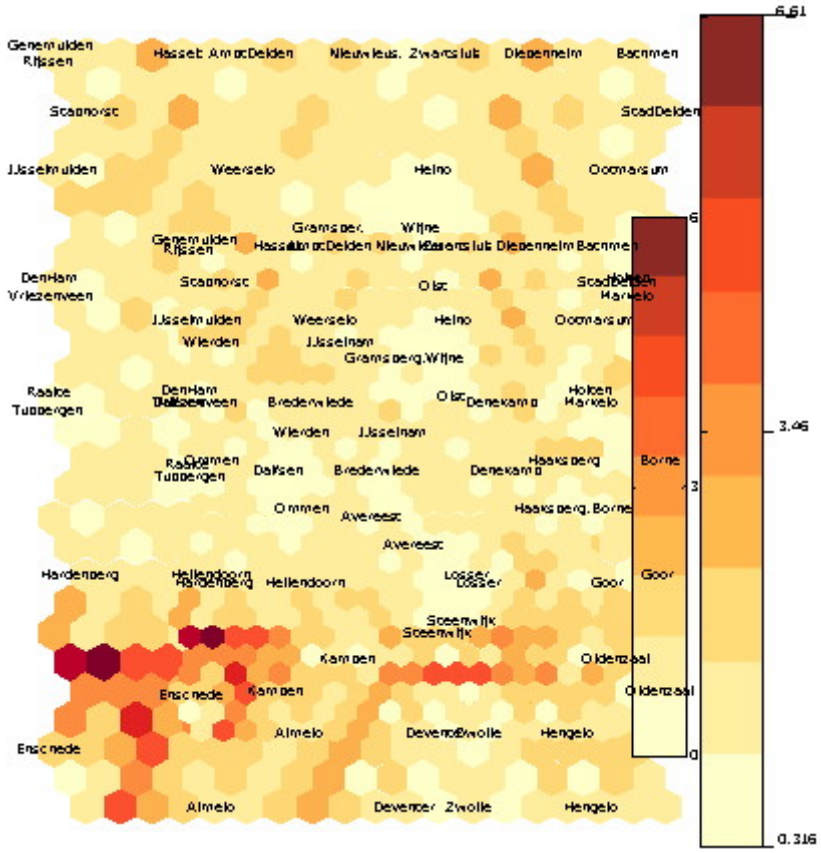

Figure 4. The U-matrix of Overijssel dataset

3.3.2 Mesh visualization (projections)

The SOM does not as other projection methods in general, try to preserve the distances directly but rather the relations or local structure of the input data. To visualize the shape of the SOM in the input space, some projection methods can be used. The mesh visualization (Figures 5 and 6) represents each map unit with a point, (which can be connected to its neighbours with lines) using a projection whereby the distances between data sample pairs are preserved as accurately as possible. It basically visualizes the SOM network and uses the SOM grid to visualize sets of objects each with unique position, color and shape. The projection of a SOM gives an informative picture of the global shape and the overall smoothness of the SOM (Figure 5). A number of visual interaction features can be offered to the user: control on the coordinate of each unit (2D or 3D sapce), the type, color and size of marker used for each map unit and the properties of lines for connecting the map units. As the SOM reduces the input data to a small number of vectors, it can be combined with other projection techniques such as principal component analysis and Sammon's mapping (10) to produce better

projection results while reducing the computation burden related to these projection methods. Figure 5 shows a Sammon's mapping projection of the SOM. The third dimension uses the value (or weight) associated to the map units according to multidimensional attributes.

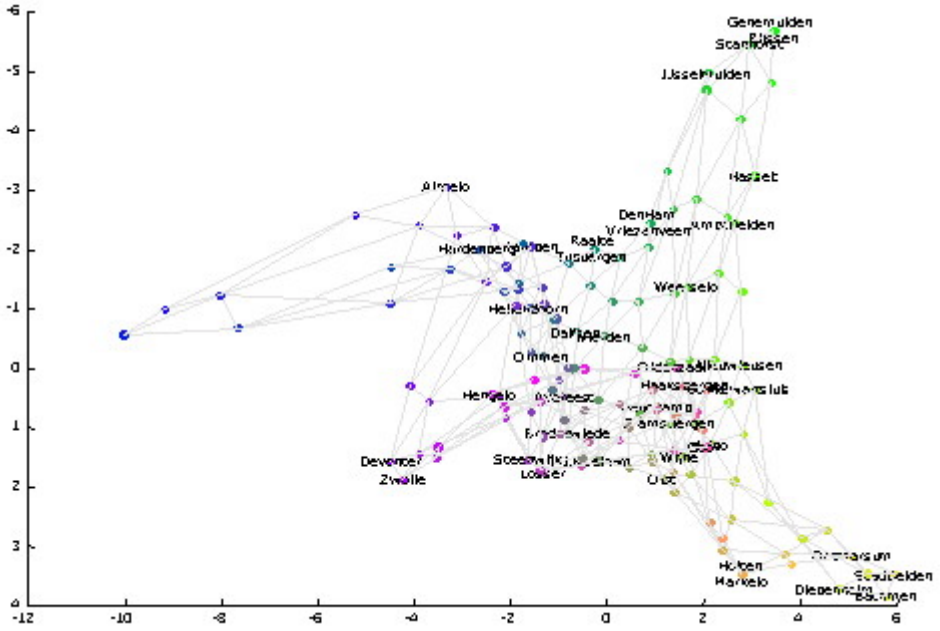

Figure 5. Mesh visualization (projection of the SOM results using Sammon's mapping)

The use of other projection techniques can help optimize the SOM representation while making use of its basic visualization features such as color-coding. Figure 6 shows projections of the SOM with PCA, in 2D and 3D view.

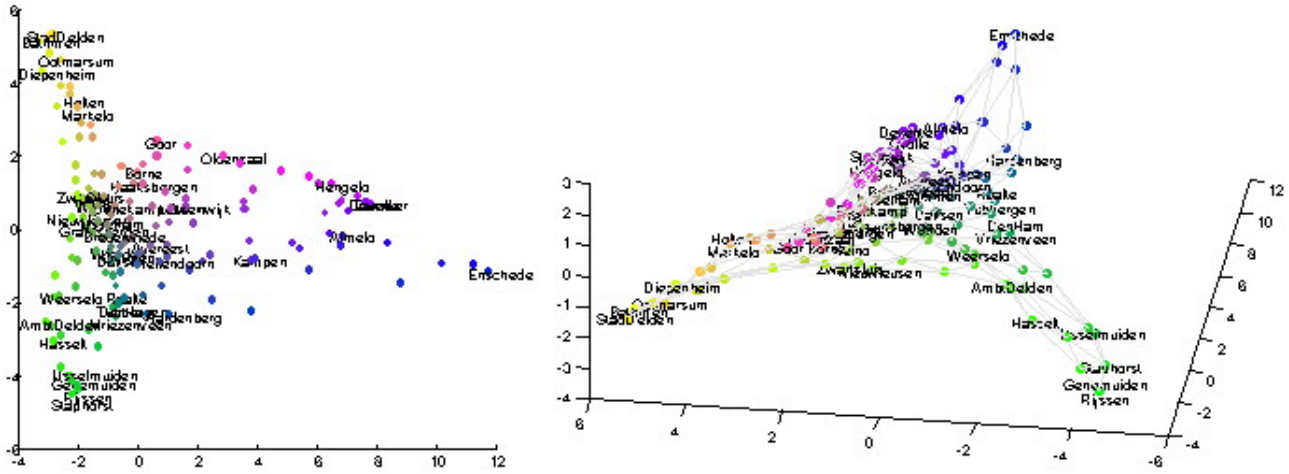

Figure 6. Projectionof the SOM using PCA (in 2D and 3D view)

3.3.3 Visualization of components (Multiple views of component planes)

The visualization of component planes together shows the values of the map elements for different attributes. They show how each input vector varies over the space of the SOM units. Each component plane shows the values of one variable in each map unit using color-coding. This gives the possibility to visually examine every cell (each cell corresponding to each map unit or data item). In Figure 7, each cell corresponds to a municipality in the Overijssel region dataset. By using the position and coloring, relationships between different map units (municipalities in this dataset) can be easily explored.

For example, the two attributes (income per inhabitant and average value of dwellings) presented in Figures 3b and 3c can be easily explored using the SOM representations. The SOM component plane visualization (Figure 7) indicates that the highest dwelling values correspond to municipalities (StadDelden, Bathmen, Diepenheim) where the average income per inhabitant is highest. It also shows the relationships between all the municipalities for the two attributes. By

relating component displays we can explore the dataset, interpretate patterns as indications of structure and examine realtionships that exist. New knowledge can be unearthed through this kind of exploration, the identification of associations between attributes using the various representations, and the formulation and ultimate testing of hypotheses.

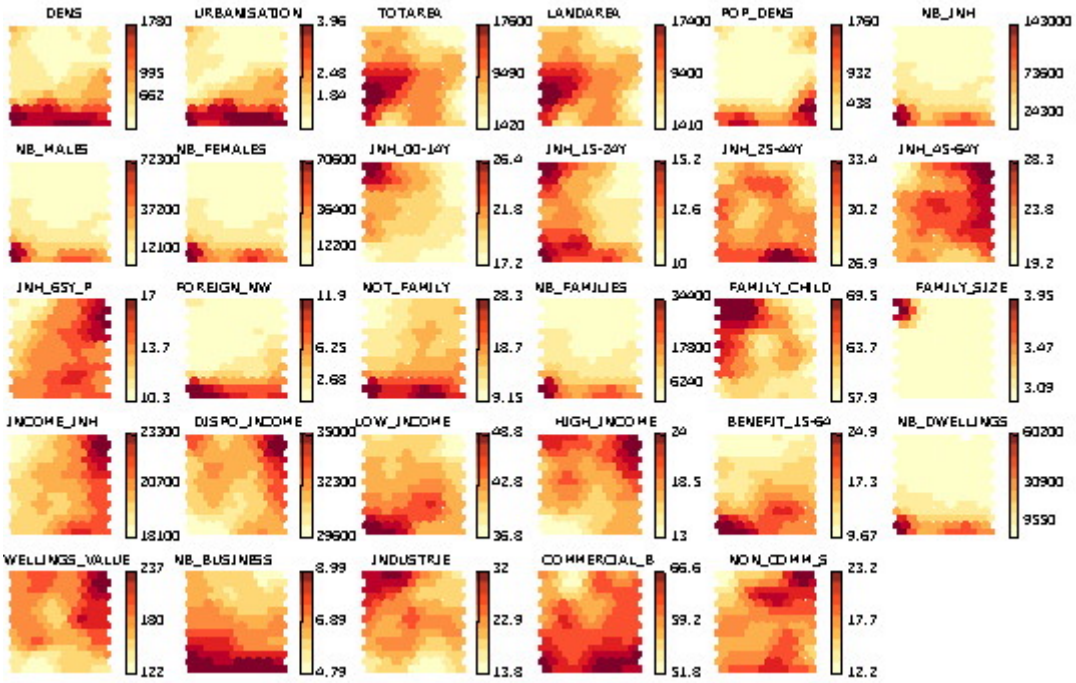

Figure 7. Visualization of component planes

3.3.4 2D and 3D surface plots of distance matrices

The 2D and 3D surface plot of distance matrix (Figures 8a and 8b) use both color and z-coordinate to indicate the average distance to neighboring map units. They use a landscape metaphor to represent the density, shape, and size or volume of clusters. The user has the flexibility to manipulate the coordinates and the view in 2D or 3D space. This visualization can be related to the U-matrix for further clusters investigation.

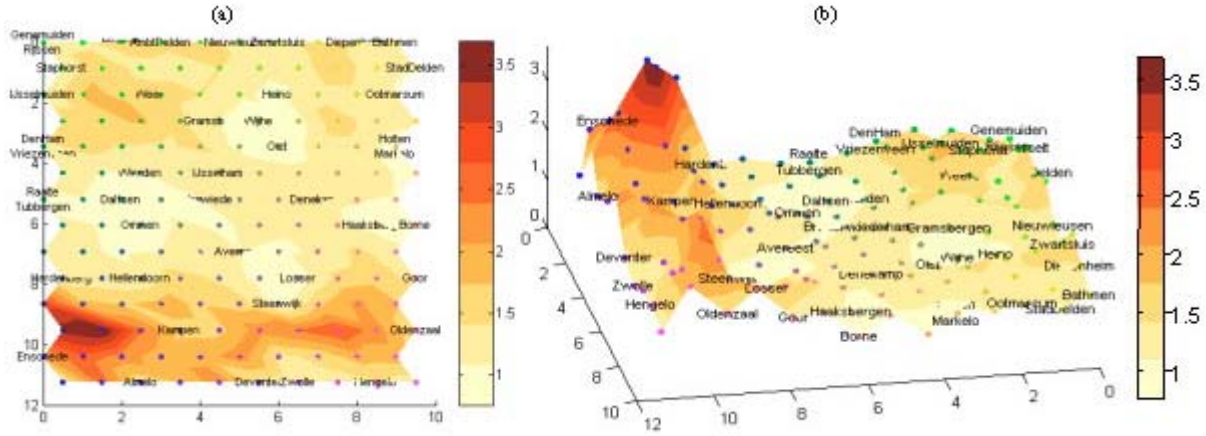

Figure 8. Surface plot of distance matrix in 2D (a) and 3D (b)

4. CONCLUSION

In this paper we have presented some features of the SOM algorithm for exploratory data analysis and visualization. A number of visualization techniques for supporting geospatial data analysis were explored. The overall objective was to explore ways to support users in information extraction from large geospatial datasets and the construction of knowledge through interaction and graphical representation. In this respect, the representations and visualization techniques explored offer opportunities to improve geographical analysis and to support exploration and knowledge discovery in the context of large geospatial datasets. Further development of the tool will integrate these representations in multiple views, linking the geographic space to the information space, for fast and effective exploration of the data.

Important types of geographical processes to investigate include spatio-temporal data and geographic image data classification.

5. REFERENCES

[1] Card, S. K.; Mackinlay, J. D.; Shneiderman, B. (1999). Readings in Information Visualization. Using Vision to Think. San Francisco: Morgan Kaufmann Publishers.

[2] Fabrikant, S. I. (2001). Visualizing region and scale in information spaces. Proceedings of the 20th International Cartographic Conference, August 6-10,2001, Beijing, China.

[3] Kraak, M. J. (2000). About maps, cartography, geovisualization and other graphics. Geoinformatics journal. Volume 3, December 2000.

[4] Kohonen, T (1995). Self-Organizing Maps. Berlin: Spring-Verlag.

[5] MacEachren, A. M.; Wachowicz, M.; Edsall, R.; Haug, D.; Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in databases methods. International Journal of Geographical Information Science, vol. 13, n4, 311-334.

[6] MacEachren, A. M.; Kraak, M.J. (2001). Research challenges in geovisualization, Cartography and Geoinformation science. Volume 28, number 1.

[7] Miller, H. J.; Han, J. (2001). Geographic data mining and knowledge discovery. London: Taylor and Francis.

[8] Openshaw, S.; Openshaw, C. (1997). Artificial Intelligence in geography. Chichester: John Wiley & Sons. pp. 21-25.

[9] Rubin, J. (1994). Handbook of usability testing: How to plan, design, and conduct effective tests. New York: John Wiley & Sons, Inc.

[10] Sammon, J., W. Jr. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401-409, May 1969.

[11] Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography*, 46, 2 (1970), pp. 234-240.

[12] Ultsch, A.; Siemon, H. (1990). Kohonen's self-organizing feature maps for exploratory data analysis. Proceedings International Neural Network Conference INNC'90P. 305 –308, Dordrecht, The Netherlands.

USING SELF-ORGANIZING MAPS FOR INFORMATION VISUALIZATION AND KNOWLEDGE DISCOVERY IN COMPLEX GEOSPATIAL DATASETS

Koua, E.L.

International Institute for Geo-Information Science and Earth Observation (ITC).
P.O.Box 6, 7500 AA Enschede, The Netherlands. E-mail: koua@itc.nl

Etien L. Koua is currently a PhD candidate at the department of Geo-Information Processing, at the ITC (International Institute for Geo-information Science and Earth Observation) in The Netherlands. His research focuses on geovisualization methods and techniques for geospatial data exploration, including the integration of information visualization, knowledge discovery and interaction techniques for patterns extraction, representation and interaction for visualization.

He has a background in Computer Science and Human-Computer Interaction.