EXPLORATION OF GEOGRAPHICAL INFORMATION SYSTEMS-BASED MEDICAL DATABASES USING SELF ORGANIZING MAPS (SOM): A CASE STUDY OF ADULT ASTHMA

Tonny J. Oyana[1], Deepthi Boppidi[1], Jun Yan[2], and Jamson S. Lwebuga-Mukasa[3]

1. Department of Geography and Department of Computer Science, Southern Illinois University
1000 Faner Drive, MC 4520, Carbondale, IL 62901-4514
Tel: +1 618-453-3022
Fax: +1 618-453-6465
Email: tjoyana@siu.edu

2. Department of Geography and Geology, Western Kentucky University
1 Big Red Way
Bowling Green, KY 42101

3. Centre for Asthma and Environmental Exposure, Lung Biology Research Program, Pulmonary and Critical Care Division
Department of Internal Medicine
Kaleida Health Buffalo General Division
UB School of Medicine and Biomedical Sciences
Buffalo, NY 14203

Abstract

In recent times, the implementation of both self-organizing maps (SOM) and geographic information systems (GIS) to facilitate visualization, classification, organization, and analysis of the continually growing digital data has gained wide recognition. SOM is a very powerful category of unsupervised neural networks with competitive and cooperative learning abilities. The SOM algorithm is essential for extracting implicit, valuable, and interesting information from vast quantities of data. The principal advantages of SOM include but are not limited to the identification of clusters of similar sequences, projection and visualization of high dimensional data spaces, and the preservation of topological relationships between data vectors. These advantages are essentially valuable to geospatial data which often come with multiple attributes where the dimensionality, complexity and the amount of data is prohibitively large for manual analysis. In this study, we explore the capabilities of both SOM and GIS for potential use in spatially-oriented biomedical databases. These capabilities are illustrated by a case study of adult asthma patient data using a variety of visualization spaces. Extracted features of similar sequences obtained from the experiment are presented.

Key Words: self-organizing maps; SOM; GIS; biomedical geocomputations; spatial databases; clustering; pattern recognition

1. Introduction

The wide acceptance of geocomputational tools to analyze and understand spatial data has gained prominence among the GIS and visualization community and to some extent in biomedical computations. It is therefore vital to explore the potential of geocomputational methods, such as self-organizing maps (SOM) for disease surveillance and the mining of large spatial biomedical databases.

The availability of large amounts of locational information of patients is increasing day by day because of improvements in advanced information technologies. The way data analyst and visualization experts understand spatial data is often determined by the way how that data is easily accessed. New areas of computational development are not only exciting but they also challenge data analyst and visualization experts to develop, at the same time suggest, more sophisticated data exploration tools. The three main purposes of data exploration are to provide a basic understanding of a given dataset, to identify frequency distribution of the variables involved, and to look for any deviants in the data. Data exploration basically provides insights for further analysis of data.

SOM is an effective way to project vast amounts of data in a high dimensional space to a space with low dimensionality. It is a prominent tool in the initial exploratory phase of data mining. SOM is a special architecture of neural networks that cluster the high-dimensional data vectors according to a similarity measure (Kohenen 1982; Nurnberger and Detyniecki 2002). In fact, SOM clusters the data in a manner similar to cluster analysis, but have an additional benefit of ordering the clusters and enabling the visualization of large numbers of clusters (Bock 2004). These clusters are arranged in a low-dimensional topology, usually a pre-defined two dimensional grid structure (Kohenen 1982; Nurnberger and Detyniecki 2002; Jiang and Harrie 2004). Hence this technique is particularly useful for the analysis of large datasets with high

dimensionality where similarity matching plays a very important role (Kohenen 1982; Cuadros-Vargas et al. 2003). It compresses information while preserving the most important topological and metric relationships of the primary data items (Kohenen 1998; Kirk and Zurada 1999). SOM is basically used to classify the objects based on their similarities within groups, thereby discovering the structure of the data hidden in large datasets (Kohenen 1998; Bock 2004; Sugiyama and Kotani 2002). It quantizes the training data into a representative set of prototype vectors and then maps the prototype vectors onto a low-dimensional grid. The clusters are arranged in a grid structure so that the neighborhood relations in the high dimensional data (Costa 1999; Elliman and Pulido 2002; Jiang and Harrie 2004) are preserved in this low dimensional space as well. As a result, SOM is especially useful for data exploration because it has prominent visualization properties and grid foundation. This grid structure simply provides a versatile platform on top of which various data exploration methods can be efficiently constructed and compared. It is important to note that the selection of the size of the map and the parameters used in estimation are primary concerns in the training of a SOM (Kohenen 1998; Bock 2004).

GIS technology and its science are emerging interdisciplinary fields which inquire about the nature of geographic phenomena and of geographic information. The science of GIS (GIScience) seeks to formalize geographic principles using logic and mathematics, to explore scientific, educational, and policy-related uses of geographic information, and to elucidate the complex relationships that individuals and society have with their surrounding environments (Mark 2004). It also provides a framework for scientific and engineering studies of physical and social phenomena (Mark 2004). According to a scientific report in the Nature Publishing Group

released in 2004, the US Department of Labor identified geotechnology as one of three most important emerging and evolving fields, along with nanotechnology and biotechnology.

The integration of SOM with GIS for biomedical applications is noteworthy as noted by Openshaw and Openshaw (1997) and Koua and Kraak (2004). Our earlier communication (Oyana et al. 2005) reported on this issue, the integration of SOM with GIS. In this report, we provide a variety of visualization spaces to illustrate the capabilities of SOM and GIS when used together.

2. Study Objectives and Motivating Factors

The primary objective of this study is to explore and visualize disease data by combining the capabilities of both SOM and GIS. Current demand for novel approaches with a wide potential to visualize or discover unknown facts and knowledge has motivated this work. It is common knowledge that huge data volumes make it extremely difficult to glean any insights from pages and pages of reports. It is usually impossible to either spot trends or understand patterns quickly enough to make the best use of data at hand. SOM algorithm facilitates data reduction method with a primary goal of aggregating and classifying data contained in large datasets into manageable information nuggets. Visual integrations within SOM and GIS are essential in gaining insights into complex spatial relationships of large data sets that consists of many different variables. In fact, visualization can be used effectively to expose associations among operating variables in a large volume of multivariate data for purposes of knowledge construction. In this study, we examine spatially dependent data, which are collected over geographical domains to improve epidemiological and biomedical computations.

3. Experimental Design

The significance of SOM/GIS in exploring data is illustrated using 4910 data points of adult patients diagnosed with asthma or gastroenteritis. Case subjects and control subjects consisted of asthma patients (International Classification of Diseases, 9th Revision [ICD-9] code 493) and gastroenteritis patients (ICD-9 code 558), respectively, residing in Buffalo neighborhoods during the same period. The study was based on a biomedical database that was obtained from Kaleida Health Systems, a major provider of healthcare in western New York. Experimental datasets are available at individual and group (aggregate) level—point and polygon formats. Vectors consisting of five components (namely, X, Y, case_control/code, IN500, IN1000) were visualized using a two-dimensional SOM. Here, X and Y are the coordinates of the patients, the case_control/code indicates whether the patient has asthma (case) or gastroenteritis (control), the IN500 indicates whether the patient is within 500m of the highway, and IN1000 indicates whether the patient is within 1000 m of a pollution source. Using a two dimensional grid, we projected the vectors in the input space onto the output space while preserving the topological relations observed in the input space. We initially chose 100 map units with a map size of 10 x 10, but after several trainings, a 10 x 8 map size was finally selected. Component planes of SOM were visualized further by slicing them to show each component.

The SOM algorithm is applied to analyze the areas affected by the asthma disease and to illustrate the structure of the data hidden in huge datasets. The experiments are conducted in SOM Toolbox 2.0 for Matlab (SOM Project, Hut, Finland), Matlab 7.0 (The MathWorks, Inc., Natick, Massachusetts), and ArcGIS 9.0 (ESRI, Inc., Redlands, California) and SPSS 13 (SPSS Inc., Chicago, Illinois). The development environments support substantial topological data structures which are capable of handling complex geocomputational processes and the integration of separate data sets to produce new spatial information is also possible.

Exploration of potential patterns was achieved through the use of various visualization spaces. Visual methods used for the experiments include several SOM visualization techniques that are based on distance matrices to map the distances between neighboring network units (neurons). The U-matrix shows the distances from each neuron's center to all of its neighbors. In the U-matrix a dark coloring between the neurons corresponds to large distance in the input space while light coloring between neurons specifies that the vectors are close to each other. Hence in the U-matrix a cluster is visualized as a group of units with light coloring surrounded by units with dark coloring.

Several experiments were conducted and training files were constructed using a number of samples ranging from 75% of the available data to 1%. We tested a variety of visualization techniques using three-to-five component planes to gain maximum insights into our data set.

For training the SOM, the learning rate goes from 0.5 in the rough tuning phase to 0.05 in the fine tuning phase. The initial neighborhood radius is equivalent to half of the map size and is gradually reduced during the training phase until it reaches 1. At any instant during the training, the minimum value of the neighborhood radius is 1.

In order to verify the general applicability of the SOM techniques for classifying disease features, the results of the training file (SOM prototype data) were imported into ESRI ArcGIS 9.0 (ESRI, Inc., Redlands, California). Then geographic maps were created and compared with the maps obtained with the original features (Oyana and Lwebuga-Mukasa 2004; Oyana et al. 2004). In these experiments, both maps illustrated similar spatial patterns and distributions of the disease thereby resolving the idea that SOM algorithm captures the dataset effectively and also represents the original data accurately.

The general design approach followed the standard Kohenen's SOM algorithm (Kohenen 1998). Additional details of the SOM algorithm can be found in Openshaw and Blake 1995; Vesanto 1999; Naenna et al. 2003; Jiang and Harrie 2004; Bock 2004. An outline of the key steps in performing this experiment is provided below.

- Establishing the ideal map size. At the beginning, we used a grid of 10 rows by 10 columns for training the test data and after several runs, we settled for 10 rows by 8 columns in the final clustering experiment.
- Initializing the weight vector for each of the neurons (vector quantization). We initialized them in a linear fashion.
- Training the SOM using component planes ranging from three-five of them at ago. The learning rates and neighborhood radius used are already described in the previous Section.
- Updating the weight vector of the winning neuron or the Best Matching Unit (BMU) and that of all the neurons within its proximity (vector projection).
- Evaluating and post-processing the SOM prototype data in GIS and SPSS software.

4. Results

After the training of SOM, a measure of map quality was generated. The maps have two primary quality properties: data representation accuracy and data set topology representation accuracy. The former is usually measured using average quantization error between data vectors and their BMUs on the map. The quantization error gives an idea of the quality of the representation provided by the SOM. This value is calculated for the whole map or for each one of the neurons and each one of the input patterns. For the latter several measures have been proposed, for example the topographic error measure—this is when the percentage of data vectors for which the first- and second-BMUs are not adjacent units. The SOM algorithm generates two types of errors: quantization and topographic. Both measures have been

implemented on the adult asthma feature dataset and the quality measures for the trained map were found to be q = 400.7 (quantization error) and t = 0.043 (topographic error), and before training q = 1.410 X 103 and t = 0. These errors are negligible and it suggests that training was adequate and that the topology is well-preserved. Higher value of the errors indicates that the training is not adequate in folding the map and the problem structure is complex.

4.1 Visualization of component planes

We have utilized multiple visualization spaces to explore and visualize a spatially-oriented biomedical dataset. While several visualization spaces were used during the SOM training sessions this paper reports only a few visualization spaces as shown in Figures 1 through 6.

Figure 1 shows a folding of a 2-D map into 3-D space to cover the whole dataset. The image in the panel represents the distribution of the data in the two dimensional grid with the X-coordinate values range from 6.6 x 105 to 6.9 x 105 and Y-coordinate values range from 4.748 x 106 to 4.764 x 106. The map units are linearly initialized. The middle image in the panel represents the SOM after initialization while the right panel represents the 3-dimensional map after training.

The visualization of a component plane shows the values of the map elements of SOM prototype data for each individual attribute. Together SOM component planes display how each input vector varies over the space of the SOM units. Each component plane only shows the values of one variable in each map unit based on certain color-coding. This makes it possible to visually examine and compare every cell (each cell corresponding to each map unit or data item) across all input dimensions.

4.1.1 Three Component Planes

Figure 2 illustrates the visualization of SOM using U-matrix. The upper left image represents the U-matrix. The primary clusters in the U-matrix are marked by ellipses. There are 7 primary clusters and some small clusters are not marked out in this figure. A dark coloring in 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 boundaries. The upper right image and the two lower images in Figure 2 illustrate the visualization of three component planes. Each of these three components is visualized in the same grid structure. The normalized values of the X-coordinates range between 672000 and 687000. For Y-coordinate, the normalized values range from 4750000 to 4770000. Similarly for case_control/code, values range between 0.0078 and 0.992. The components are visualized using U-matrix and the interpretation is similar to that of the U-matrix explained above. In reviewing these SOM component planes, an interesting pattern seems to be emerging, of the two major clusters at the right corner of U-matrix suggesting that there are two geographic areas in the study region with significant concentration of disease cases. Using the identifier tool, we discovered that one of them corresponds to Buffalo's Westside and the Downtown regions and the other represents Buffalo's Southside. This observation is consistent with earlier findings reported in Oyana and Lwebuga-Mukasa (2004) and Oyana et al. (2004) thus confirming that the SOM prototype data captures spatial structures embedded in the asthma dataset effectively.

Visualization of this dataset can also be achieved by means of a surface plot of the distance matrix. Figure 3 represents a 3D surface plot of adult asthma feature dataset. The 3D surface plot gives a better visual representation (related to three variables X, Y, and case_control/code) of the distance between the clusters in the SOM map units than a 2D surface.

The distance between map units is represented by both color and z-coordinate and indicate average distance to neighboring map units. In this 3-D surface plot, lower values and color coding located in the valleys show cluster similarity. The presence of two noticeable valleys suggests the existence of adult asthma clusters. Once again, this 3-D display confirms further that there exist two major clusters that possibly represent geographic areas with high concentration of asthma cases.

4.1.2 Five Component Planes

Figure 4 illustrates five component planes and cluster analysis examples derived from the SOM prototype data. In this round of SOM training, we added two additional attributes, IN500 and IN1000. The U-matrix suggested 8-to-10 clusters were present in the data while the K-means clustering method identified eight clusters with the best Davies-Bouldin validity index. Looking at the pattern in Figure 4, it is consistent with earlier observations described in Figures 2 and 3. This affirms the statement that SOM is the best suited for extracting the representative features and exploring the data to acquire an understanding and generating hypothesis about the properties of data.

Figure 5 is a geographic map of extracted features of SOM prototype data and final cluster centers displayed over original features of adult asthma data. There were seven final cluster centers. Three of these were major ones and the remaining four were minor ones. Each final cluster center represents case patients of asthma and its surrounding neighborhoods. A further examination of these clusters reveals that Buffalo's Westside and the Downtown regions had the largest cluster sizes. The second and third largest cluster sizes were observed in Buffalo's Southside and Eastside. These findings are consistent with earlier observations reported in Oyana

and Lwebuga-Mukasa (2004) and Oyana et al. (2004) thus confirming further that the SOM prototype data captures the asthma dataset effectively.

4.2 Analysis of SOM Prototype Data with SOM Toolbox and SPSS

We also conducted a comparison of clusters of the prototype SOM data in both SOM toolbox and SPSS software using the K-means clustering approach. In both SOM and SPSS, the largest clusters were matched perfectly at 100%. For the other two major clusters, SPSS classified the cluster class in the east as its second and the one in the south as its third; whereas SOM classified the cluster class in the south as its second and other in the south as its third. All the remaining minor cluster classes were comparable in both cases. Although the classification results in SOM were stronger than the ones in SPSS (this is partly due to the use of best Davies-Bouldin validity index) both SOM toolbox and SPSS software captured the datasets effectively.

We then conducted a cluster analysis using a diagnostic tool as illustrated in Figure 6. The box plot confirmed cluster 6 as the major cluster. There were some variations in clusters 5, 6, and 7, but all the distances were within reason. The box plot confirmed further that all of the clusters representing asthma and its surrounding neighborhoods were within a reasonable distance. The diagnosis also revealed that cluster 4 has an outlier and the remaining clusters were minor ones.

5. Conclusions and Future Work

The SOM provides an excellent visualization and exploration platform for analyzing vast quantities of spatially-oriented biomedical data. These experiments show that when SOM algorithm is combined with GIS methods, there are even more powerful for exploratory analysis than when applied separately. Our experimental results suggest that this hybrid approach to analyze spatially-oriented biomedical data provides a useful exploration tool to support the

formulation of better or new study hypotheses regarding the spatial distribution of a particular disease. Overall, this exploratory work was essential for improving our interpretations of the findings reported in earlier studies (Oyana and Lwebuga-Mukasa 2004; Oyana et al. 2004). We confirmed further based on these experiments that asthma is more prevalent in Buffalo's Westside which is in close proximity to major roadways, Peace Bridge Complex (PBC), and pollution sources.

The quantization and topological errors that indicate the measure of quality of the SOM during training were negligible and there were greater improvements in the error component of trained maps which suggest the training was adequate and topology was well preserved. However, we wish to reduce the error component further by either using principal component analysis or incorporating a mathematical improvement model. In our future work, we are going to make some mathematical improvements to the SOM algorithm. Currently, the SOM algorithm has a number of efficiency and convergence issues that need to be addressed. Some of these issues are (1) speed and quality of clustering, (2) the number of output neurons, (3) the updating procedure for the output neurons, and (4) the learning rate in the SOM model.

Other methodological improvements as was suggested in Oyana et al. 2005 have significant impacts on how data analysts and visualization experts explore spatially-oriented biomedical databases to extract relevant information and gain maximum insights. Improvements can enhance the SOM learning process and provide a better understanding of the biomedical and epidemiological processes of diseases in relation to space, time, and environment.

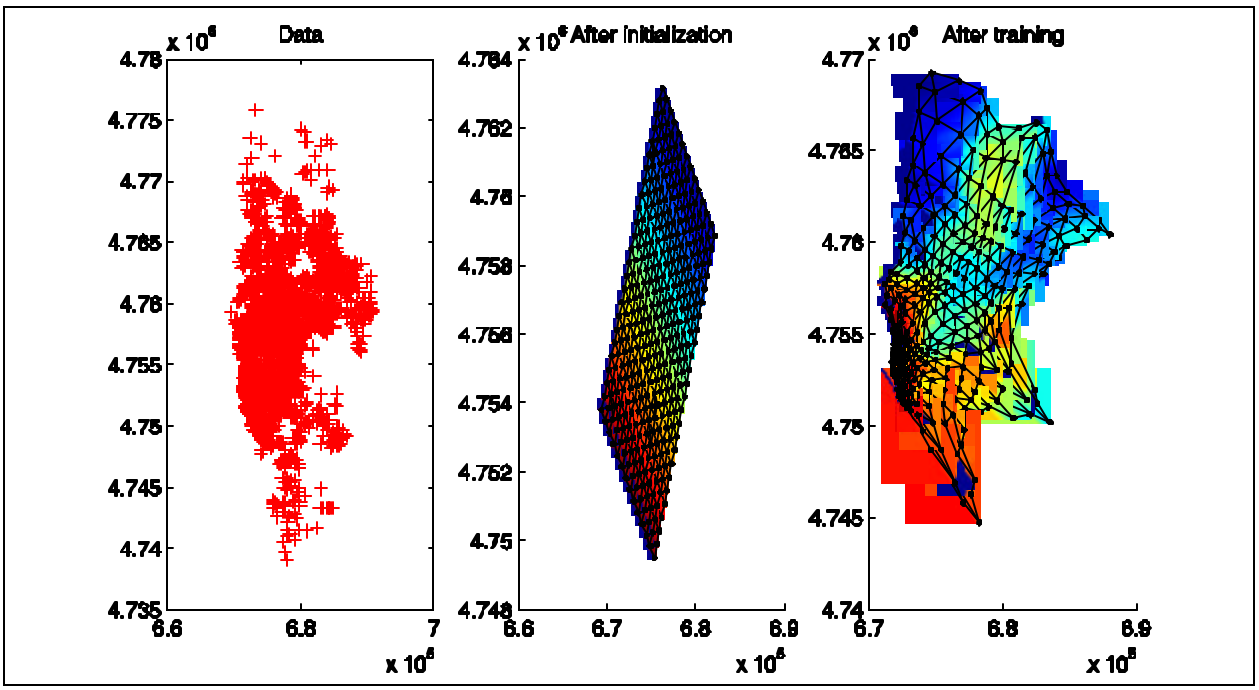


Figure 1: Folding of 2-dimensional map into 3-dimensional space in order to be able to capture the whole data.

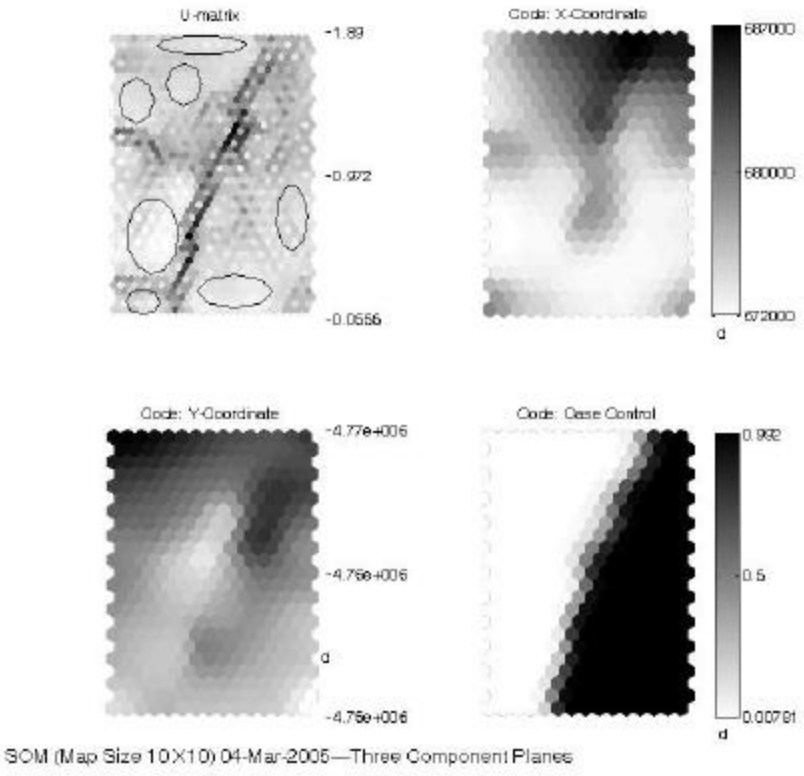


Figure 2: Upper left image shows the U-matrix with seven major clusters. The upper right and the lower images illustrate the visualization of component planes of the adult asthma feature dataset with X-coordinate, Y-coordinate and case_control/ code as variables. Note the very smooth and even component planes among these attributes.

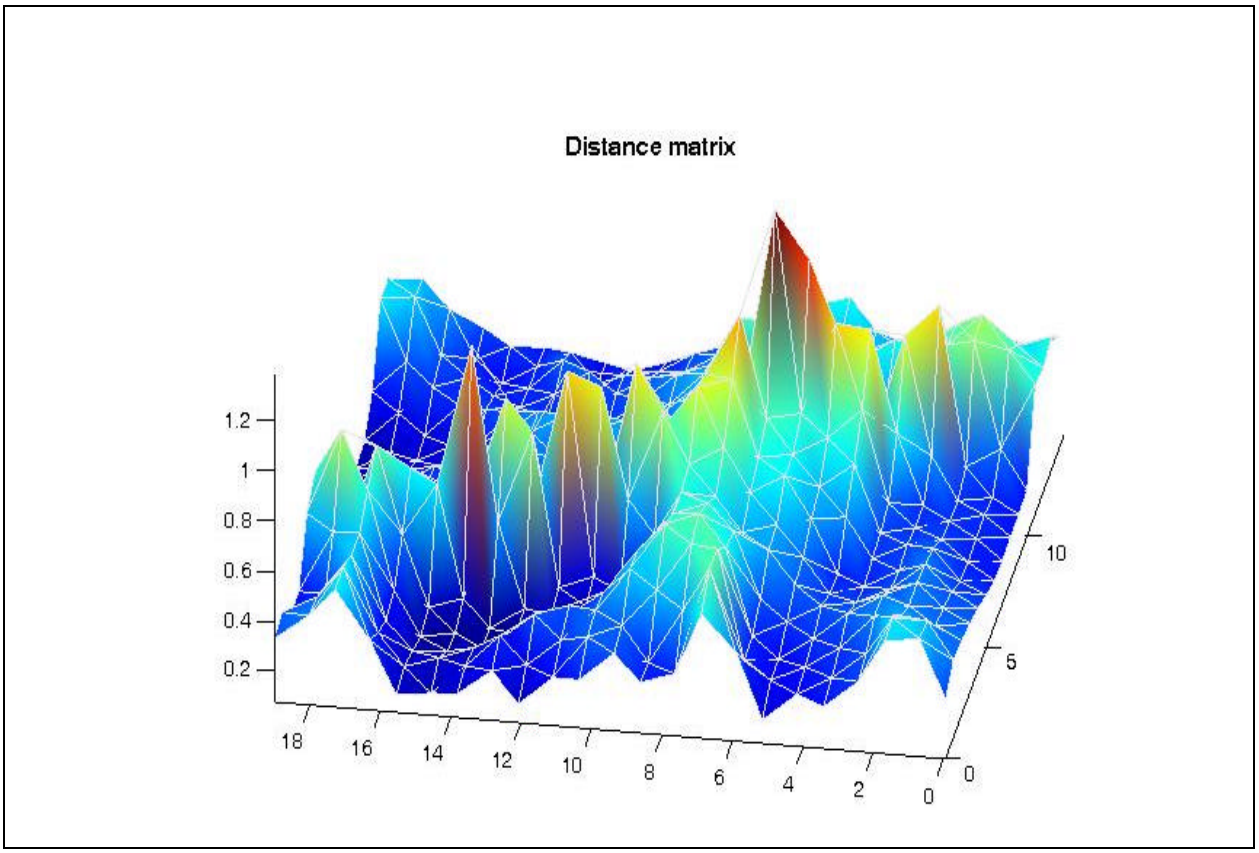
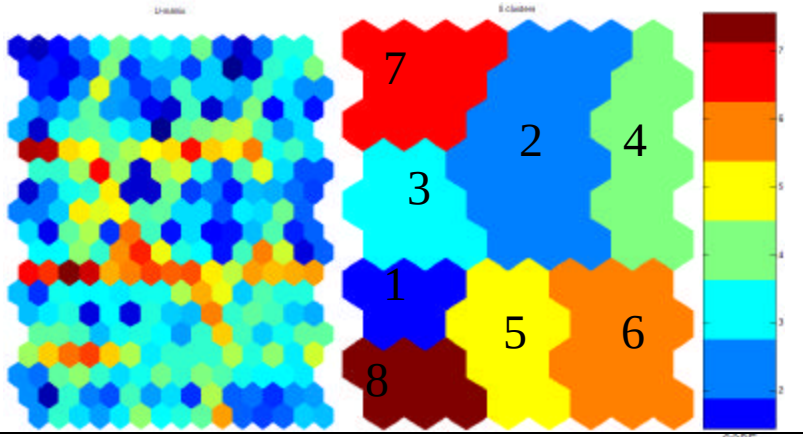
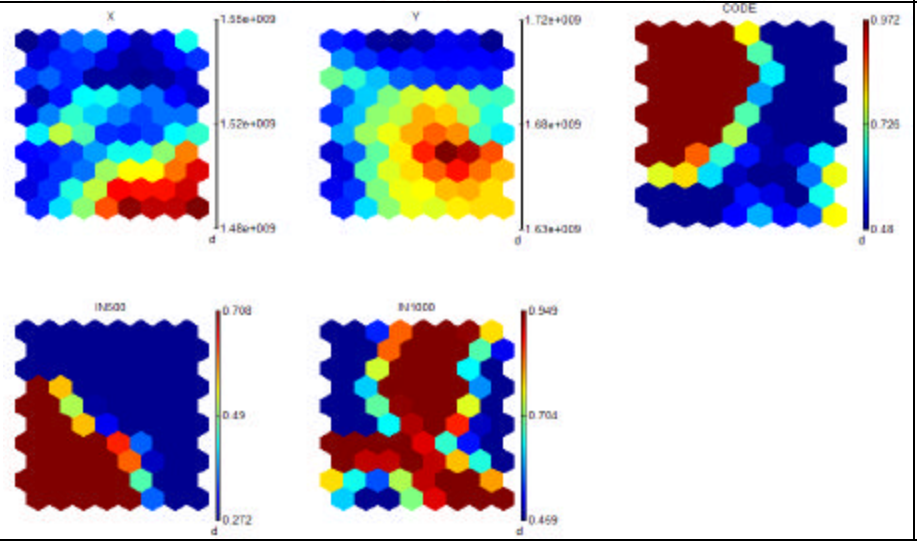


Figure 3: Exploration of adult asthma feature dataset. The 3D surface plot provides a better visual representation of the distance between clusters for this dataset than 2D surface.



Upper panel—the *U*-matrix suggests 8-to-10 clusters when applied to the SOM prototype data (training result) while *K*-means clustering method identifies eight clusters based on the best Davies-Bouldin validity index.



The lower panel shows the five components used for training. X and Y are the coordinates of the cases or controls. Code is the indicator of the cases (1 for case, 0 for control). IN500 is the indicator showing if the point is within 500m of a highway. IN1000 is the indicator showing if the points are within 1000m of a pollution source.

Figure 4: The 10 X 8 SOM in the upper panel illustrates high-dimensional visualization of medical data using the *U*-matrix and *K*-means clustering methods. The five component planes are displayed in the lower panel.

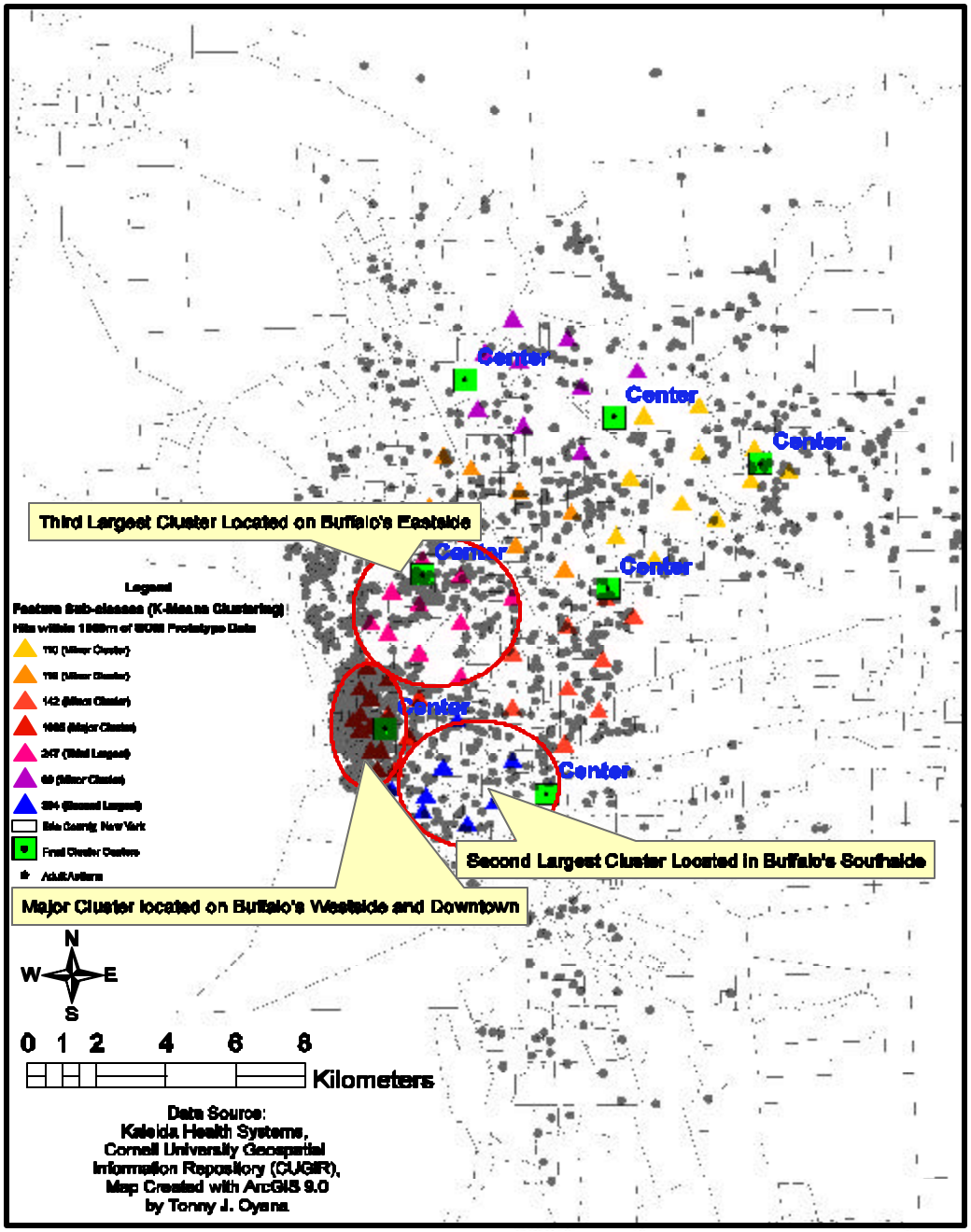

Center
Center
Center
Center
Center
Center
Center
Third Largest Cluster Located on Buffalo's Eastside
Second Largest Cluster Located in Buffalo's Southside
Major Cluster located on Buffalo's Westside and Downtown
Legend
N
W
E
S
0 1 2 4 6 8
Kilometers
Data Source:
Kaleida Health Systems,
Cornell University Geospatial
Information Repository (CUGIR),
Map Created with ArcGIS 9.0
by Tonny J. Oyana

Figure 5: Extracted features of SOM prototype data and final cluster centers displayed over original features of adult asthma data. Each final cluster center represents case patients of asthma and its surrounding neighborhoods.

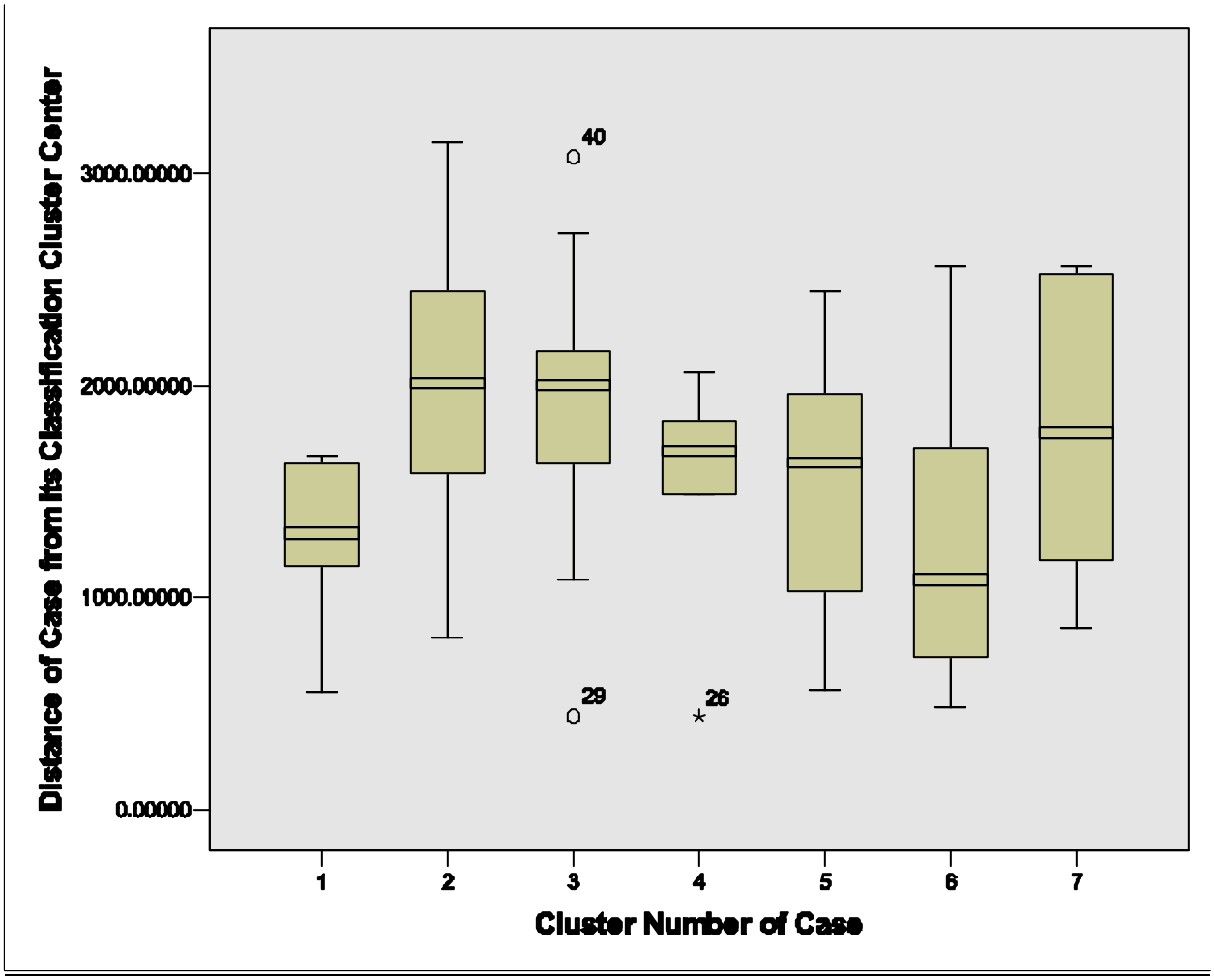


Figure 6: Cluster 6 is the major cluster. For clusters 5, 6, and 7, there are some variations, but all of these distances are within reason. Cluster 4 has an outlier and the remaining clusters are minor ones as validated both by SOM and SPSS methods.

Acknowledgment: Supported in part by SIUC Faculty Start-up Fund and SIUC ORDA Faculty Seed Grant. Ms. Deepthi Boppidi, a Research Assistant to this project was supported by SIUC ORDA Faculty Seed Grant. Dr. Jamson S. Lwebuga-Mukasa, Department of Internal Medicine, UB School of Medicine and Biomedical Sciences, Kaleida Health Systems Buffalo General Division provided the datasets for this study.

Protection of Human Subjects: All research was approved by the Southern Illinois Unversity Carbondale and the University at Buffalo Human Investigation Review Boards in accordance with national and institutional guidelines for the protection of human subjects.

6. References:

1. Andre, T., and Filho, A., 2002. Segmentation of digitized mammograms using self-organizing maps in a breast cancer computer aided diagnosis system. *In Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN 2002),* 0-7695-1709-9/02
2. Bock, T., 2004. A new approach for exploring multivariate data: self-organizing maps. *International Journal of Market Researc*h, **46(2)**:189–203
3. Boudjemai, F., Enberg, P.B., and Postaire, J.G., 2003. Surface modeling by using self-organizing maps of Kohonen Systems, *In Proceedings of the IEEE International Conference on Man and Cybernetics*, 0-7803-7952-7/03.
4. Costa, J.A.F., and Netto, M.L.A., 1999. Cluster analysis using self-organizing maps and image processing techniques. *In Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics*, **5**:367–372.
5. Cuadros-Vargas E., Romero, R., and Obermayer, K., 2003. Speeding up algorithms of SOM Family for Large and High Dimensional Databases. In Yamakawa T., editor, *In Proceedings of the WSOM*, 167–172.
6. Elliman, D., and Pulido, J.R.G., 2002. Visualizing ontology components through self-organizing maps. *In Proceedings of the Sixth International Conference on Informa tion Visualization*, **10(12)**:434–438.
7. Ferreira de Oliveira, M.C., and Levkowitz, H., 2003. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, **9(3)**:378–394.
8. Flexer, A., 2001. On the use of self-organizing maps for clustering and visualization. *Intelligent data analysis*, **5**:373–384.
9. Himberg, J., 1998. Enhancing the SOM based visualization by linking different data projections. *In Proceedings of the First International Symposium on Intelligent Data Engineering and Leaming (IDEAL 1998),* Hong Kong, 427–434.
10. Honkela, T., Kaski, S., Lagus, K., and Kohonen, T., 1996. Exploration of full-text databases with self-organizing maps. *In Proceedings of the 1996 IEEE International Conference on Neural Networks (ICNN),* **1(3-6)**:56–61.
11. Jiang B., and Harrie L., 2004, Selection of streets from a network using self-organizing maps. *Transactions in GIS*, **8(3)**:335–350.

12. Kaski, S., Honkela, T., Lagus, K., and Kohonen, T., 1998. WEBSOM—Self-Organizing Maps of Document Collections. *Neurocomputing*, **21**:101–117.
13. Kirk, J.S., and Zurada, J.M., 2000. A two-stage algorithm for improved topography preservation in self-organizing maps. *In Proceeding of the IEEE International Conference on Systems, Man, and Cybernetics*, **4(8-11)**:2527–2532.
14. Kohonen, T., 1982. Self-Organized Formation of topologically Correct Feature Maps. *Biological Cybernetics*, **43**:59–69.
15. Kohonen, T., 1998. Self-Organization of Very Large Document Collections: State of the Art. *In Proceeding of the International Conference on Artificial Neural Networks (ICANN 1998*), Skovde, Sweden, September 2–4.
16. Kohonen, T., 2001. *Self-Organizing Maps.* 3rd edition (Berlin, Heideberg: Springer Press).
17. Konig, A., 1998. A survey of methods for multivariate data projection, visualization and interactive analysis. *In proceedings of the 5th international conference on soft computing and information/intelligent systems (IIZUKA 1998)*, 55–59.
18. Koua, E.L., and Kraak, M., 2004. Geovisualization to support the exploration of large health and demographic survey data. *International Journal of Health Geographics*, **3**:12.
19. Kraaijveld, M.A., Mao, J., and Jain, A.K., 1995. A nonlinear projection method based on Kohenen's topology preserving maps. *IEEE transactions on neural networks*, **6(3)**:548–559.
20. Manduca, A., 1994. Multiparameter medical image visualization with self-organizing maps. *In Proceeding of the IEEE World Congress on Computational Intelligence and IEEE International Conference on Neural Networks*, **6(27)**:3990–3995.
21. Mao, J., and Jain, A.K., 1995. Artificial neural networks for feature extraction and multivariate data projection. *IEEE transaction on neural networks*, **6(2)**:296–317.
22. Mark, D.M., 2004. IGERT: Integrative Geographic Information Science Traineeship Project, the National Science Foundation (NSF) Award Abstract #0333417.
23. Masters, T., 1993. *Practical Neural Network Recipes in C++* (San Diego, USA: Academic Press).
24. Miller, H.J., and Han, J., 2001. *Geographic data mining and knowledge discovery* (London: Taylor and Francis).

25. Naenna, T., Bress, R.A., and Embrechts, M.J., 2003. DNA classifications with self-organizing maps (SOMs). *In Proceedings of the 2003 IEEE International Workshop on Soft Computing in Industrial Applications (SMCIA 2003*), **23(25)**:151–154.
26. Nurnberger, A., and Detyniecki, M., 2002. Visualizing changes in data collections using growing self-organizing maps. *In Proceedings of the 2002 International Joint Conference on Neural Networks* (IJCNN 2002), **2**:1912–1917.
27. Oja, E. and Kaski, S., (editors). 1999. *Kohonen Maps* (Amsterdam: Elsevier Science).
28. Openshaw, S., Blake, M., and Wymer, C., 1995. *Using Neurocomputing Methods to Classify Britain's Residential Areas.* Available at http://www.geog.leeds.ac.uk/papers/95-1/ [Last accessed on March 31st, 2005]
29. Openshaw, S., and Openshaw, C., 1997. *Artificial Intelligence in Geography* (New York: John Wiley & Sons).
30. Oyana, T.J., Boppidi, D., Yan, J., and Lwebuga-Mukasa, J.S., 2005. Integration of self-organizing maps (SOM) into a geographic information systems (GIS) data model. Workshop on Topology and Spatial Databases. *In Proceedings of Topology and Spatial Databases Workshop, The Department of Geomatics Engineering at University College London, Department of Geography and Geomatics, University of Glasgow and Laser-Scan*, April 5th–8th, 2005.
31. Oyana, T.J., Rogerson, P., and Lwebuga-Mukasa, J.S., 2004. Geographic clustering of adult asthma hospitalization and residential exposure to pollution sites in Buffalo neighborhoods at a U.S.–Canada Border Crossing Point. *American Journal of Public Health*, **94(7)**:1250–1257.
32. Oyana, T.J., and Lwebuga-Mukasa, J.S., 2004. Spatial relationships among asthma prevalence, healthcare utilization, and pollution sources in Buffalo neighborhoods, New York State. *Journal of Environmental Health*, **66(8)**:25–38.
33. Samet, H., 1995. *Spatial data structures* (Reading: Addison-Wesley/ACM).
34. Squyres, J.M., and Lumsdaine, A., 2003. Component Architecture for LAM/MPI. *In Proceedings of 10th European PVM/MPI Users' Group Meeting*, Venice, Italy, 379–387, Lecture Notes in Computer Science, Springer-Verlag. Available: http://www.lam-mpi.org [Last Accessed February 20th, 2005].

35. Suganthan, P.N., 1999. Hierarchical overlapped SOMs for pattern classification. *Neural Networks, IEEE Transactions, 10(1)*:193–196.