Journal of Environmental Informatics 13(1) 33-44 (2009)

Journal of Environmental Informatics

www.iseis.org/jei

Visualization of High-Dimensional Clinically Acquired Geographic Data Using the Self-Organizing Maps

T. J. Oyana*

Department of Geography, Southern Illinois University, 1000 Faner Drive, MC 4514, Carbondale, IL 62901-4514, USA

Received 14 December 2007; revised 12 October 2008; accepted 20 October 2008; published online 12 March 2009

ABSTRACT. The objective of this study is to visualize high-dimensional data vectors using popular data reduction algorithms. The study reports on the effectiveness and expressiveness of a set of data reduction algorithms in the visualization of geospatial data sets derived from clinical records of patients. The experiments show that when the SOM algorithm is combined with GIS methods together they are even more powerful tools for exploratory analysis than when each is applied separately. The visual approach provides a very useful exploration environment to support the formulation of new and better study hypotheses regarding the spatial distribution of a particular disease. While it was apparent that the spatial distribution and patterns of asthma were predominately located near the major roadways and the Peace Bridge Complex, obstructive sleep apnea is slightly more widespread even in the suburbs and surrounding neighborhoods. The spatial patterns discovered between the original features of adult and childhood asthma are consistent with the SOM-trained data, but a slight difference emerges for the SOM-trained obstructive sleep apnea data set. This study is successful at gaining significant novel insights into the spatial characteristics of patient data in relation to key environmental factors.

Keywords: self-organizing maps, geocomputations, disease, GIS, visualization, data exploration, clustering, pattern recognition

1. Introduction

The recent prominence of new methods for exploring large-scale geospatial data provides both opportunities and challenges. Some of these methods have adequately incorporated spatial techniques [e.g. geographic information systems (GIS)] and pattern recognition algorithms [e.g. self-organizing maps (SOM)] and are currently being used to extract and discover interesting patterns of data. Other opportunities include applications for managing vast amounts of data and for visualizing data vectors in a high dimensional space within a space of low dimensionality, but significant challenges still remain.

A major requirement in exploring locational data is the conduction of some form of clustering. The main goal of clustering is to categorize these data into meaningful groups, which will provide the data analyst three key opportunities: (1) to explore and discover similarities and differences among spatial patterns; (2) to reveal the hidden structure present in the data or to uncover the underlying and interesting patterns of the data; and (3) to derive useful conclusions from the data or knowledge that could be useful in spatial data mining.

In geography, the use of clustering algorithms to solve geographical problems is now widespread (Openshaw et al., 1995; Openshaw, 1998; Murray and Estivill-Castro, 1998; Cuadros-Vargas and Romero, 2002; Guo et al., 2003). There are elaborate studies (Skupin and Fabrikant, 2003; Guo et al., 2004, 2006; Bação et al. 2004, 2005; Cuadros-Vargas and Romero, 2005) in which clustering algorithms have successfully been applied to explore multidimensional large-scale data sets.

Let us briefly focus on the basic mathematical structure of clustering. Consider a number of data points or feature vectors that are to be assigned to clusters. Supposing the feature vectors (data points) in the input data set are denoted by X_i, where $i = 1, \ldots,$ N are members of set X in a multidimensional vector space. Here, members of X must then be assigned to clusters or graphically mapped to a target space of low dimensionality (usually 2-D or 3-D visualization spaces) according to specific algorithmic schemes and criteria.

1.1. The SOM Algorithm

The SOM consists of a regular, usually two-dimensional (2-D), grid of map units, or it can simply be stated as a spatial organization of map units (centroids) (Kohonen, 1982, 1998, 2001; Duin et al., 1999; Vesanto and Alhoniemi, 2000; Flexer, 2001). In the SOM model, the input vector $M = [M_{k=1}, \ldots, M_{k=n}]$ with d dimensions is represented by an input layer W_{ij} containing a grid of units ($m \times n$) with ij coordinates. The main goal of SOM is to combine both analytic and graphical techniques to categorize data into a low-dimensional framework and generate a visual representation of the clusters using different visualization spaces. It possesses both competitive and coope-

[*] Corresponding author. Tel.: +1 618 4533022; fax: +1 618 4536465.
E-mail address: tjoyana@siu.edu (T. J. Oyana).

ISSN: 1726-2135 print/1684-8799 online

doi:10.3808/jei.200900138

rative learning capabilities, and its network typically consists of high-dimensional input and low-dimensional output layers. The SOM network organizes data items in the input space by assigning and adjusting them to output space according to a winning neuron and neighborhood weights. Two of its main properties are the preservation of original topological relations and the maintenance of network relationships among data items.

The SOM learning procedure closely follows a biological understanding of how neurons in the human brain function as they process, organize, and store incoming and outgoing information. The information, obtained through these learning procedures is then used to generalize and can be applied in new or unknown situations. This type of learning procedure is what embodies the SOM algorithm. SOM provides a versatile neural network architecture for data exploration and mining, because it utilizes its neurons efficiently and wastes very few or none when representing any data.

1.2. A Description of PCA, MDS, and Sammon Mapping Algorithms

While clustering reduces the number of data items by grouping them, there are well-known "projection" methods that could be useful for dimension reduction. The main goals of these methods are (1) to represent the input data items in a lower dimensional space in such a way that certain properties of the structure of the data set are preserved as faithfully as possible and (2) to visualize a very large multivariate data set using an output with a reduced dimensionality. In this study, I wish to investigate two specific data reduction methods: the principal component analysis (PCA) and multidimensional scaling (MDS), particularly sammon mapping, for my proposed application domain. These methods are classified as computational algorithms and normally draw from statistical analysis. They have been characterized to possess excellent properties, and I could use the methods to visually explore and analyze data items, detect the data structure, and explore underlying factors in a multidimensional data set. They have also been extensively integrated with SOM, another very popular neurocomputational algorithm (Huang et al., 2005), in several domains.

PCA is a widely accepted technique for dimension reduction (Vesanto and Alhoniemi, 2000; Yang et al., 2003; Huang et al., 2005). The main goal of PCA is to look at the covariance structure of the original and trained SOM data. The principal idea behind PCA is based on establishing the axis direction (eigenvector), which maximizes the explanation of variance in the dependent variable. The basic structure of a PCA assumes a data set X with ($n \times m$) matrix, where n is the number of samples and m is the number of dimensions measured in each sample. With this ($n \times m$) vector matrix, PCA uses the K-leading eigenvectors of the ($n \times n$) covariance matrix as the axes of the lower k-dimensional space and the leading eigenvectors correspond to linear combinations of the original variables that account for the largest amount of term variance (Yang et al., 2003; Huang et al., 2005). According to Yang and colleagues and Huang and colleagues, a major shortcoming of PCA is that it has high memory and computational requirements: it requires $O(n^2)$ memory for the dense covariance matrix, and $O(kn^2)$ for finding the K-leading eigenvectors, where n is the number of data items. These authors further claim that requirements for running PCA could be unacceptably high when the number of data vectors is very large, for example tens of thousands. Another drawback of PCA is that the eigenvectors are usually very difficult to interpret theoretically, although Ding (2000) supports the effectiveness of PCA as illustrated by several empirical studies because of its key role in the reduction of noise, redundancy, and ambiguity.

MDS is a set of mathematical techniques that enable a researcher to uncover hidden structure in the data (Borg and Groenen, 1997; Duin et al., 1999; Huang et al., 2005). Supposedly, I have a set of objects in which a measure of the similarity between objects is known, and using this measure I could determine how similar or how dissimilar two objects are or are perceived to be. Mathematically, I could compute the proximity measure of the two objects in multiple ways, for example, using the correlation coefficient or Euclidean distance from the vector representation of the original set of objects on a feature space of a given reduced dimensionality (Borg and Groenen, 1997; Huang et al., 2005). MDS therefore maps from a higher dimensional space to a lower dimensional space in which each object is represented by a point and the distances between points resemble the original similarity information; that is, the larger the dissimilarity between two objects, the farther apart they should be in the lower dimensional space (Yang et al., 2003; Huang et al., 2005). During the mapping, the MDS algorithm attempts to preserve all interpoint distances and the geometrical configuration of points, consequently revealing the structure present in the data or the hidden structure of the data. This revelation makes it easier to visually explore and analyze the hidden structure of the data. A widely established MDS algorithm is the sammon mapping algorithm. Sammon mapping has been reported to be very efficient and does not generalize because new points are not added to the map without being recalculated. Other types of MDS algorithms include classical scaling and niemann mapping (Duin et al., 1999).

Sammon mapping (Sammon, 1969; Jain and Dubes, 1988) in contrast with the PCA method is a nonlinear projection mapping technique that attempts to optimize a cost function describing how well the pairwise distances in a data set are preserved. It maps vectors in high-dimensional input (original) space to a target space of a lower dimensionality, typically to a plane. The algorithm tries to preserve all distances between input vectors, emphasizing local distances, and arranges vectors in lower dimensionality according to the distance structure between the vectors of the original space by minimizing a quadratic function of the mapping error, also known as stress. The method is iterative and computationally very intensive. However, when applied to the weight vectors of a SOM as opposed to the whole original data set, the computing times of the algorithm stay reasonable.

The introduction to this study describes a set of multivariate data reduction techniques (cluster analysis, SOM, PCA, and sammon mapping), and the remainder is divided into four parts. Part 1 gives the statement of the problem and briefly explains on the need for such systems; Part 2 provides the experimental design; Part 3 presents and illustrates the results; and Part 4 provides some conclusions and directions for future research.

1.3. Rationale for Applying SOM, PCA, MDS, Sammon Mapping, and GIS Methods to Spatially Enabled High-Dimensional Patient Data

While many studies have applied SOM with GIS in other problem domains, there are few studies (Manduca, 1994; Tamminen et al., 2000; Sugiyama and Kotani, 2002; Koua and Kraak, 2004) available in the biomedical and disease informatics subfields. Visual exploration of data within SOM and GIS are essential in gaining fundamental insights into complex spatial relationships of large data sets that may consist of many different variables (Oyana et al., 2005a, b). In fact, visualization can effectively be used to make sense of the data and expose any existing associations among variables in a large volume of multivariate data for purposes of knowledge development and construction. Current demand for novel approaches with a wide potential to visualize or discover unknown facts and knowledge from a very large patient data set has motivated this work. A major contribution of this paper was to integrate a number of popular tools including GIS, SOM, PCA, and MDS with a computational framework and applied to geospatial data analysis. These tools can easily facilitate data reduction by summarizing and classifying large data sets into manageable information nuggets. Indeed, the visual outcomes are indispensable for verifying earlier findings and establishing superior study hypotheses. The methods applied in this study have revealed new interesting spatial patterns and associations from the analysis of subsets (clusters) that were previously unclear or unknown in previous studies (Oyana and Lwebuga-Mukasa, 2004; Oyana et al., 2004; Oyana and Rivers, 2005). The study was able to identify and delineate representative subsets of the original data where fundamental insights regarding the spatial patterns of asthma and obstructive sleep apnea were gained leading to the verification and formulation of superior study hypotheses.

They were two core goals for conducting this study. The first goal was to visualize very large-scale and multidimensional data sets, pre- and post-process SOM data, in a GIS and to apply a powerful graphing method (box plot) to analyze resulting clusters. I visualized *n*-dimensional unit data vectors of spatially dependent data collected over geographic domains to advance epidemiological and biomedical computations. The second goal was to further uncover and confirm core characteristics of the data that are suggestive of any mathematical and statistical properties of the underlying structure in the original experimental datasets. This study also introduced a new significant feature that has not been reported elsewhere: the use of SPSS *K*-means clustering algorithm and applied a graphical technique, a box plot to analyze and compare SOM-trained data with the classes obtained from the SOM toolbox's Davies-Bouldin Validity Index (Davies and Bouldin, 1979).

2. Methods and Materials

The main goal in this experimental phase is two-fold: (1) to use SOM with GIS to visually explore multidimensional data vectors; and (2) to conduct post-processing analysis and to evaluate the quality of SOM-trained data using nonlinear and linear mapping techniques (sammon mapping and PCA) and box plots. Three data sets were used for performing the experiments.

2.1. Data Sets I and II

There are 4,910 and 10,289 data points of adult and children patients, respectively, diagnosed with asthma or gastroenteritis. Case 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 patient database was obtained from Kaleida Health Systems, a major provider of healthcare in western New York. The two data sets are available at individual and group (aggregate) levels — point and polygon vector formats. Vectors consisting of six components (X, Y, case_control/code, IN500, IN1000, and PM1000) were visualized using a two-dimensional SOM. In this case, X and Y represent 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 500 m of the highway; IN1000 indicates whether the patient is within 1,000 m of a pollution source; and PM1000 indicates whether the patient is within 1,000 m of the sampling site of measured particulate matter concentrations.

2.2. Data Set III

The third data set consists of obstructive sleep apnea (OSA) subjects who were identified by the following 9th Revision codes: ICD-9CM 780.51, .53, and .57. There are 3,943 data points of OSA subjects available at individual and group (aggregate) levels — point and polygon vector formats. Vectors consisting of seven components (X, Y, INPOS, INNEG, LOS, INFF, and AGE) were visualized using a two-dimensional SOM. In this case, X and Y represent the coordinates of the patients; the INPOS indicates whether the patient is within 1,000 m of a positive health promotion factor (such as recreational facilities, fitness centers, sightseeing areas, sport clubs and fields, and amusement parks); INNEG indicates whether the patient is within 1,000 m of a negative health promotion factor (such as pubs, airports, nightclubs, fast-food restaurants, and liquor stores); LOS indicates the length of stay at the hospital for the patient; INFF indicates whether the patient is within 1,500 m of a fast-food restaurant; and AGE indicates the patient's age at time of admission to the hospital.

The experiments were conducted in SOM Toolbox 2.0 for Matlab (SOM Project, Hut, Finland), Matlab 7.0 (MathWorks

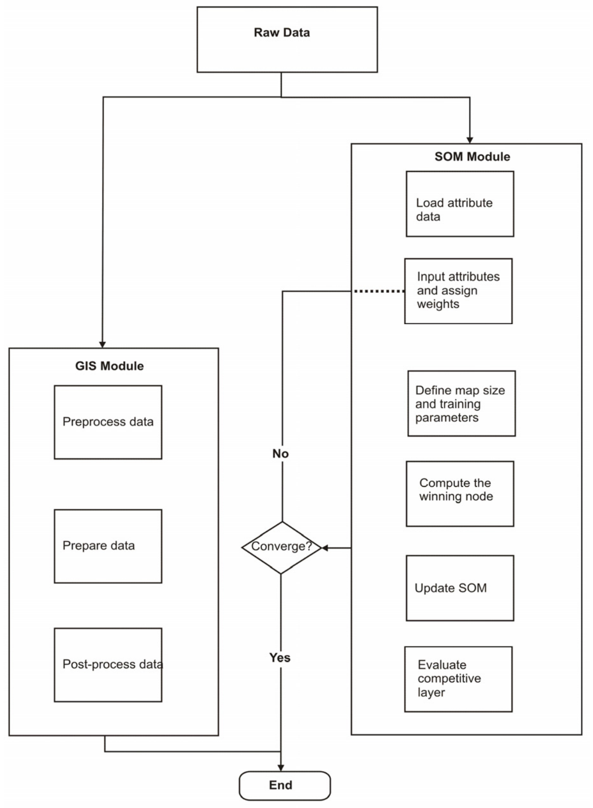

Figure 1. Data flow in the system implemented using a loosely-coupled strategy.

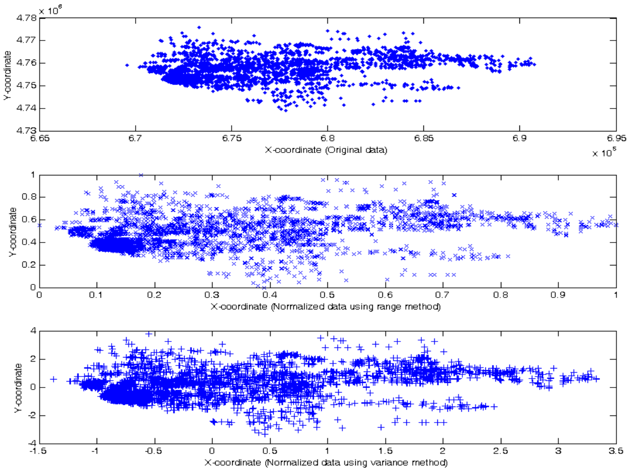

Figure 2a. The spatial distribution of normalized adult asthma data set.

Inc., Natick, Massachusetts), ArcGIS 9.1 (ESRI Inc., Redlands, California), and SPSS 13 (SPSS Inc., Chicago, Illinois). These computational tools 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. The environments also allow with greater ease the formulation and compilation of complex mathematical equations for visual modeling. Figure 1 illustrates the data flow system conceived for visualizing high-dimensional clinically acquired geospatial data.

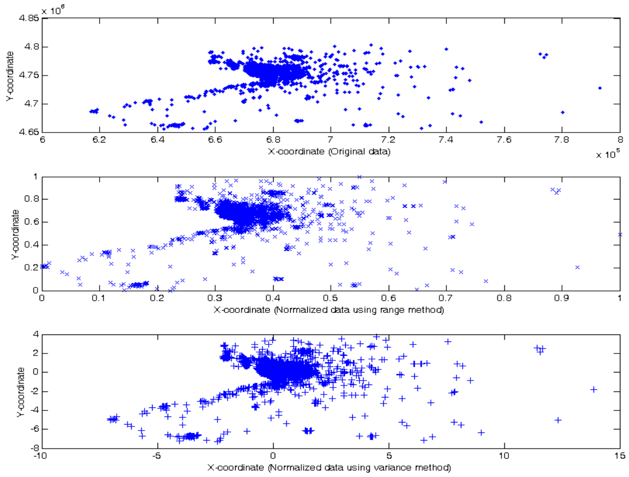

Figure 2b. The spatial distribution of normalized childhood asthma data set.

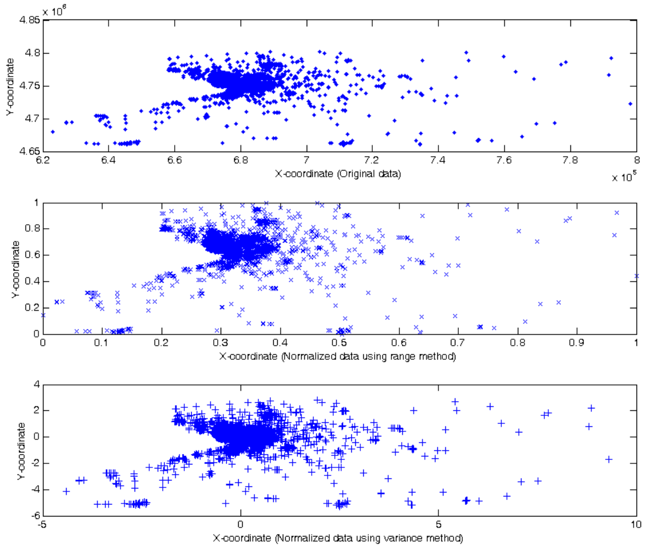

Figure 2c. The spatial distribution of normalized obstructive sleep apnea data set.

The first phase of the experiments was to train the SOM using the experimental data sets; this involved several steps such as normalizing and defining training parameters as illustrated in Figure 1. To determine whether the data sets should be normalized, I tried two methods (algorithms) using the range and variance of the data sets. The range method normalizes values between 0 and 1, while the variance method normalizes to one using a linear operation. In both situations, the data look very stable and well distributed, suggesting that normalized data have a very minimal effect on the outcome. Figures 2a through 2c illustrate the original and normalized data with range and variance algorithms. As these figures illustrate, it does not matter whether experimental data sets are normalized

Table 1. SOM Training Parameters

Data Set	Map size	NhradiusR**	NhradiusF***	Elapsed Time	Qe/Te(linear&sequential)	Qe/Te(linear&batch)*
I—Adult	23X13	5.75	1.4375	1.943 s	433.9; 0.039	211.71; 0.048
II—Child	20X20	5	1.25	4.166 s	942.6; 0.064	521.76; 0.05
III—Sleep	20X20	5	1.25	3.245 s	1433.6; 0.05	722.12; 0.079

*There is a general improvement in map quality (quantization error/topological error (qe/te)) in the last column when separate functions for initialization and training the SOM. This is because I can adjust and specify the radius and training length, whereas in the second last column I used the neighborhood radius and training length defaults as defined in the som_make function.

**Initial neighborhood radius (NhradiusR) for rough-tuning phase = max(msize)/4.

***Initial neighborhood radius (NhradiusF) for fine-tuning phase = (max(msize)/4)/4 until it reaches 1, where max is the maximum value of the map size matrix, for example in data set I it is [23 13], so max is 23.

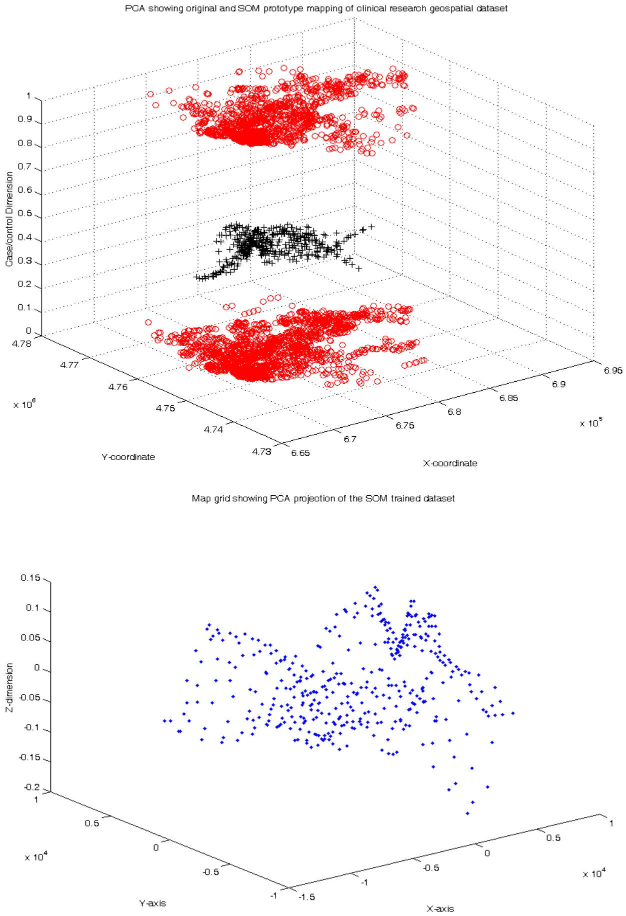

Figure 3a. Projected adult asthma data set with PCA (upper and lower panels).

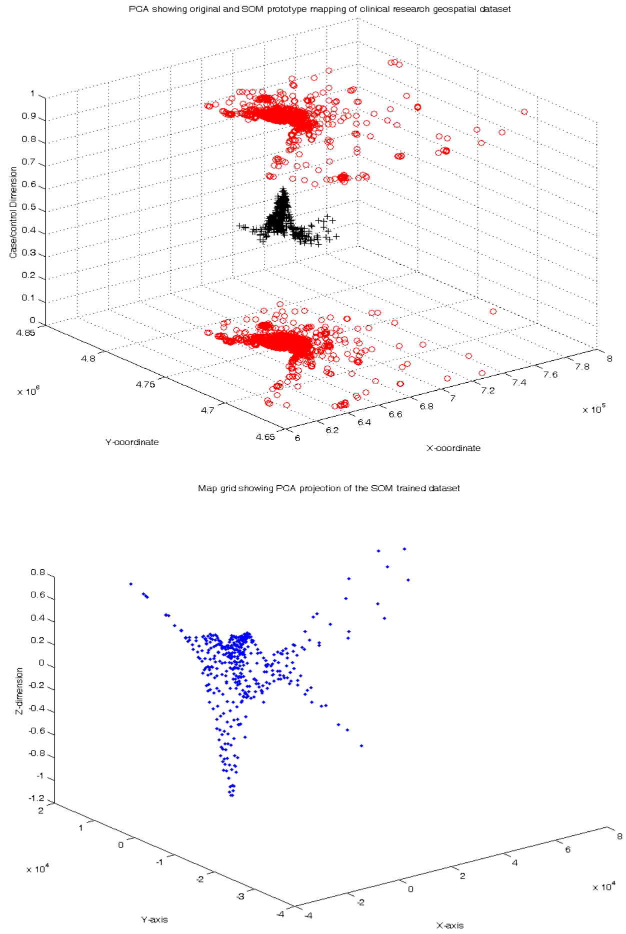

Figure 3b. Projected childhood asthma data set with PCA (upper and lower panels).

or not normalized because the outcomes in both situations were not significantly different.

Using a two-dimensional grid, I projected the vectors in the input space onto the output space while preserving the topological relations observed in the input space. I constructed several experiments using a number of training samples ranging from 75% of the available data to 1%, with the learning rate going from 0.5 in the rough-tuning phase to 0.05 in the fine-tuning phase. The initial neighborhood radius was set to half of the map size and was gradually reduced during the training phase until it reached 1, with the minimum value of the neighborhood radius set at 1 throughout the training. The training regimes followed well-established SOM recommended standards as described in the technical notes of the SOM toolbox.

Visual exploration of potential patterns was achieved through the use of various visualization spaces. The most popular techniques for SOM visualization are based on distance matrices, which map the distances between neighboring neurons, one of these being the U-Matrix. The U-Matrix shows the distances from each neuron's center to all of its neighbors, with the dark coloring between the neurons corresponding to large distance in the input space, while the light coloring between neurons specifies that the vectors are close to each other. A cluster is visualized as a group of units with light coloring surrounded by units with dark coloring as I will illustrate later. To delineate cluster boundaries following the visualization process, I used the *K*-means clustering algorithm with Best Davies-Bouldin Validity Index provided within the SOM toolbox.

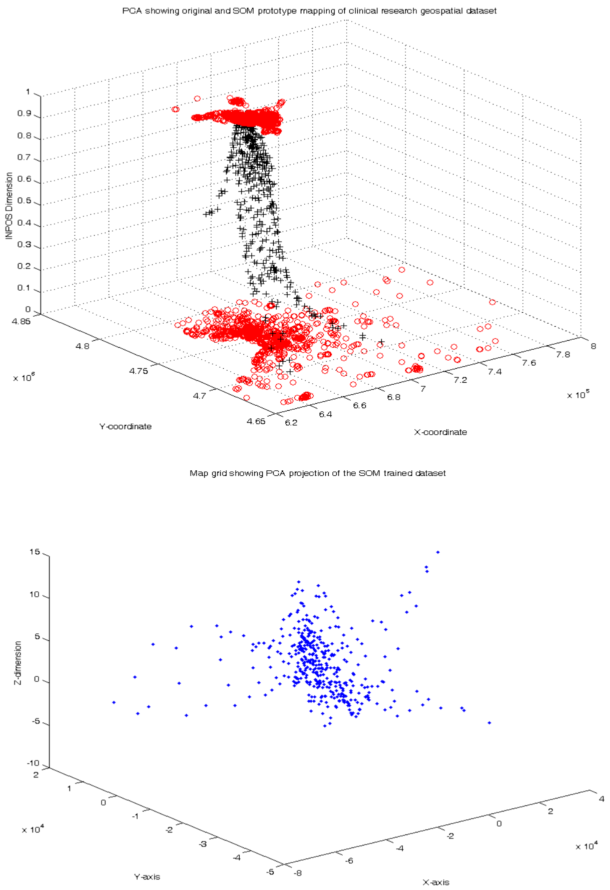

Figure 3c. Projected obstructive sleep apnea data set with PCA (upper and lower panels).

In order to verify the general applicability of the SOM techniques for classifying disease features, the results of SOM-trained data were imported into ESRI ArcGIS 9.1 as illustrated in Figure 1. The pre- and post-processing of geospatial and SOM-trained data were conducted using ArcGIS. The SOM network and the original data files were mapped and compared to visualize and generate the geographic maps pertaining to the spatial distribution of patient data. These geographic maps were compared with the maps obtained from earlier studies (Oyana and Lwebuga-Mukasa, 2004; Oyana et al., 2004; Oyana and Rivers, 2005). The results were consistent with the ones previously obtained, suggesting that these maps from the SOM network clearly illustrated similar spatial patterns and distributions of the disease, thereby resolving the idea that the SOM algorithm captures the data set effectively as well as represents the original data accurately. The final geographic maps illustrate the visual effectiveness of SOM in capturing the structure of the data and new significant insights were also gained from this study.

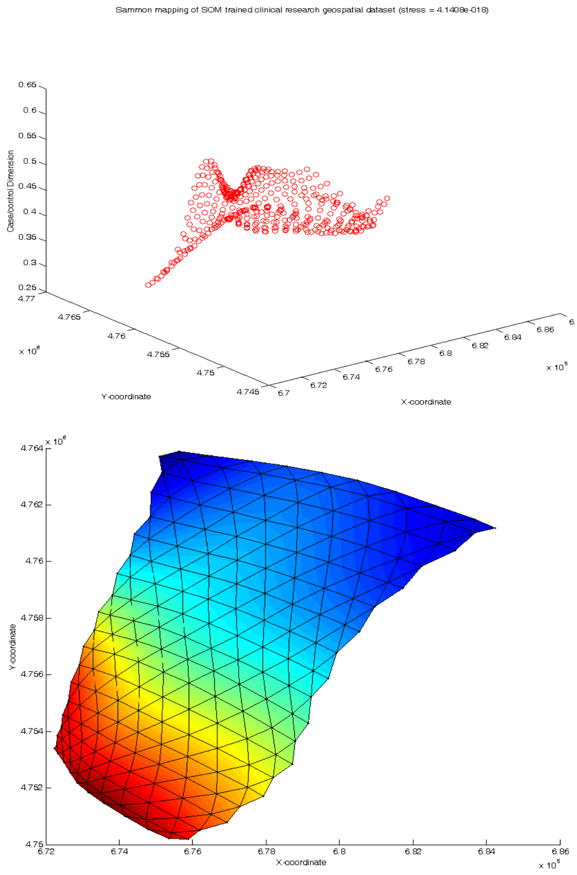

Figure 4a. Projected adult asthma data set with sammon mapping (upper and lower panels).

I used box plots for post-processing analysis and validity purposes. The box plot is a very important graphing method, which allowed me to visually explore general types of statistics within each cluster, for example, the minimum, maximum, median, lower and upper quartiles, and outliers. I was also able to compare different clusters of SOM-trained data. By being able to analyze whether outliers were present in any of these SOM feature subclasses, I was able to determine the appropriateness and validity of each cluster in relation to other classes and separate the good clusters from the bad ones.

3. Results and Analysis of Clusters

The quality of SOM training is summarized in Table 1. The data suggests that map quality for the three data sets was acceptable and the errors were negligible, thus SOM training was adequate, successful, and the topology was well-preserved.

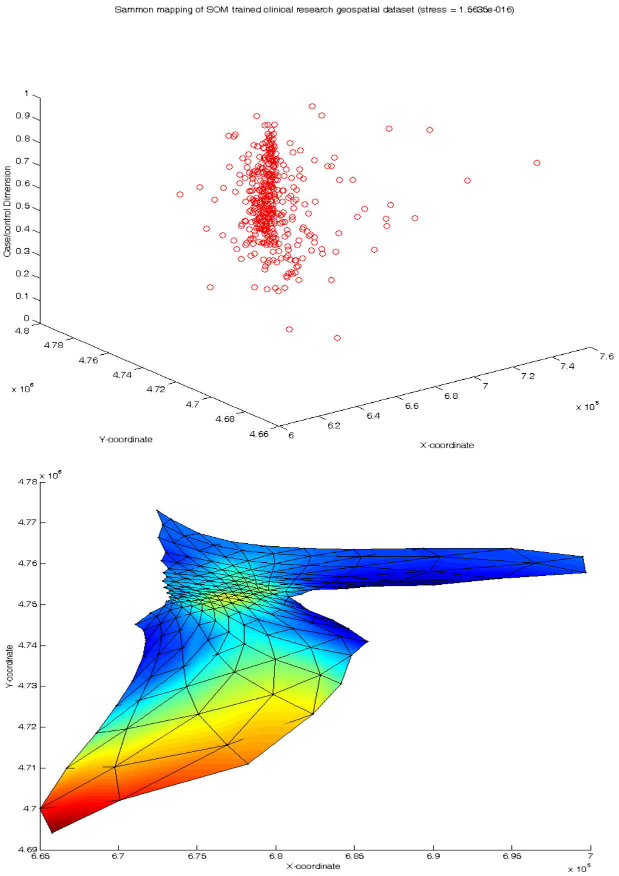

Figure 4b. Projected childhood asthma data set with sammon mapping (upper and lower panels).

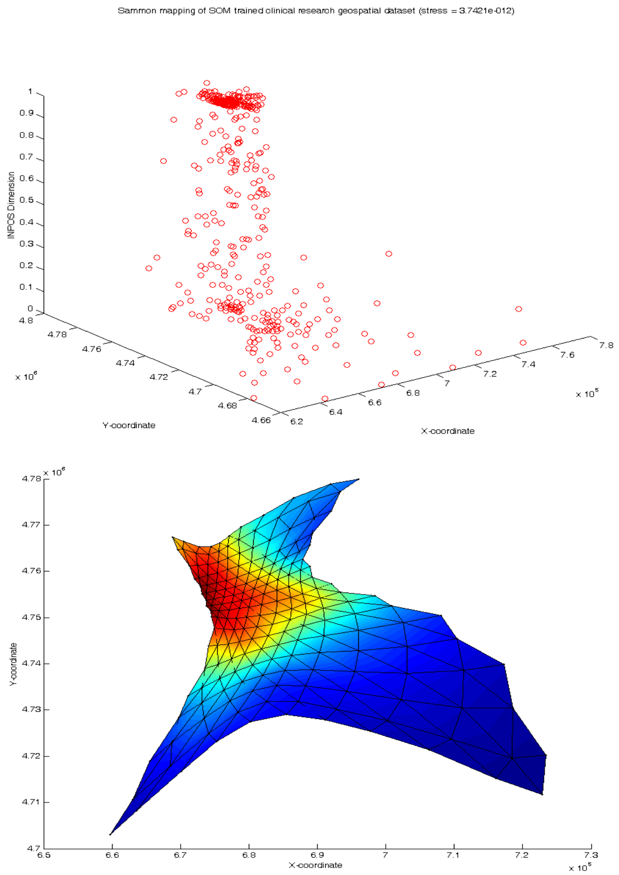

Figure 4c. Projected obstructive sleep apnea data set with sammon mapping (upper and lower panels).

Figure 2a shows spatial distribution of data set I – adult asthma; Figure 2b shows the spatial distribution of data set II – childhood asthma; and Figure 2c shows the spatial distribution of data set III – obstructive sleep apnea. The stability of original data features, as illustrated in Figures 2a through 2c, suggested that normalizing the data sets prior to SOM training would have not made any significant differences.

Figures 3a through 3c illustrate SOM visualization of adult asthma, childhood asthma, and obstructive sleep apnea data sets using the PCA method. The PCA was used to validate and display original data space as a linear projection mapping technique on a subspace of the original data space that best preserves the variance in the experimental data sets. The original data sets are illustrated using the red circles, while prototype SOM data are shown by black cross marks with their corresponding data. The SOM-trained data are within the original data points, and the data clouds are ellipse-like, meaning that SOM training was effective. Each data vector illustrates the principal variables, which significantly contribute to asthma or obstructive sleep apnea at every location. Overall, the PCA method was not only highly effective at revealing the structure of the original and SOM-trained data and accounting for variability, but it also allowed some conduction of onscreen visual inspection and exploration of the data sets. The PCA results (reduced dimensions) were comparable to the results obtained from the SOM clustering algorithm, thus illustrating the effectiveness and expressiveness of both methods by visually providing exploratory knowledge of the data sets.

Figures 4a through 4c illustrate sammon plots for the three experimental data sets. The sammon plots of the trained SOM data revealed that data points, which are close to each other in 2-D visualization space, were also close in the original dimensional space. The grid map connects closer points by lines, and when the distance between dissimilar points is large, I could use these geometrical configurations in lower dimensional space to visually expose the hidden structure of the data set. The stress factors involved in the computation for the adult asthma, childhood asthma, and obstructive sleep apnea data sets were in the order of 4.1408×10^{-18}, 1.5635×10^{-16}, and 3.7421×10^{-12}, respectively. These very low values suggest that the training was very effective. The results from the sammon plots clearly demonstrate successful map folding for all the data sets. Figures 5a through 5p illustrate the U-Matrices and component planes. Figures 5a through 5e represent the adult asthma data set; Figures 5f through 5j represent the childhood asthma; and Figures 5k through 5p represent the

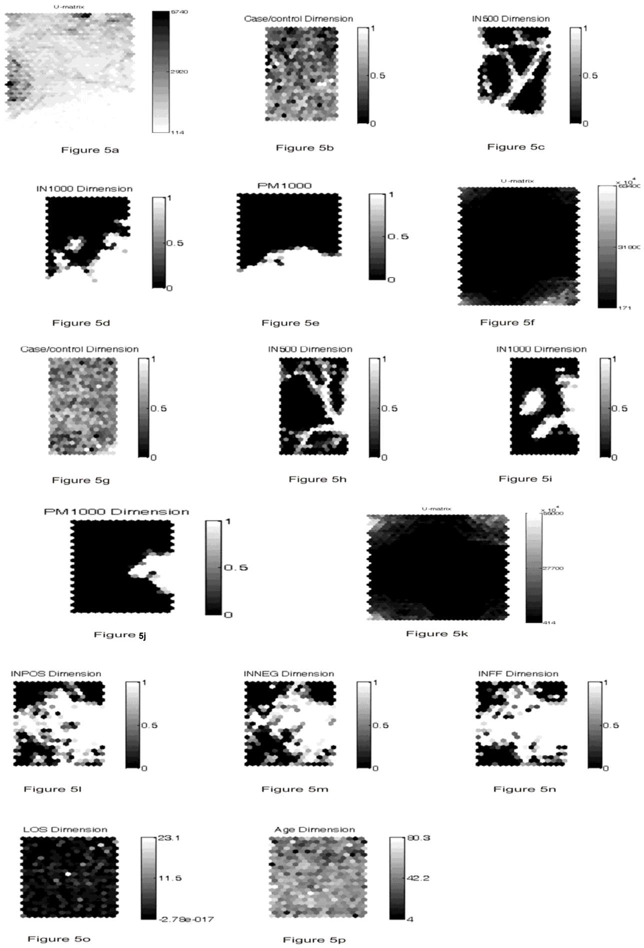

Figure 5. SOM visualization of adult asthma, childhood asthma, and obstructive sleep apnea datasets.

obstructive sleep apnea data sets. Specially, Figures 5a, 5f, and 5k represent the U-Matrices for adult asthma, childhood asthma, and obstructive sleep apnea data sets, respectively. The light blue coloring depicts the clusters in the both the U-Matrices and component planes. The U-Matrices and component planes display how each input vector varies over the space of the SOM units. Each component plane shows only the values of one variable in each map unit based on certain color coding. These SOM visualization techniques make it possible to visually examine and compare every cell (each cell corresponding to each map unit) across all input dimensions. These techniques also allow for visual exploration of clusters in each unit of SOM data. The light blue color tone in Figures 5a through 5p indicates that distances are very close and that these neurons belong to a similar cluster, while the light yellow color (reddish) tone shows a coarse distance with the neurons farther apart, signifying cluster boundaries. Using visual inspection of the U-Matrices, I observed three major adult asthma clusters, two major childhood asthma clusters, and three major obstructive sleep apnea clusters. The component planes for the adult and childhood asthma data sets clearly illustrated strong associations between cases of asthma and proximity to

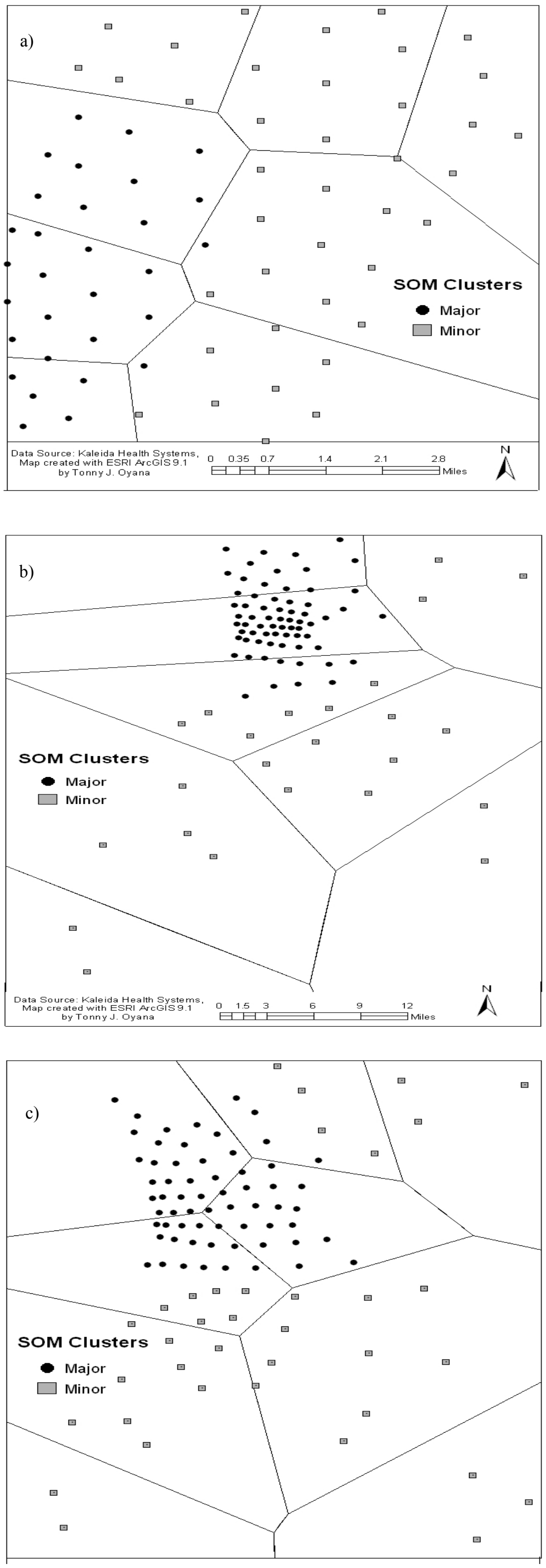

Figure 6. A mapped example of identified spatial patterns of SOM feature subclasses of a) adult asthma data set, b) childhood asthma data set, and c) obstructive sleep apnea data set.

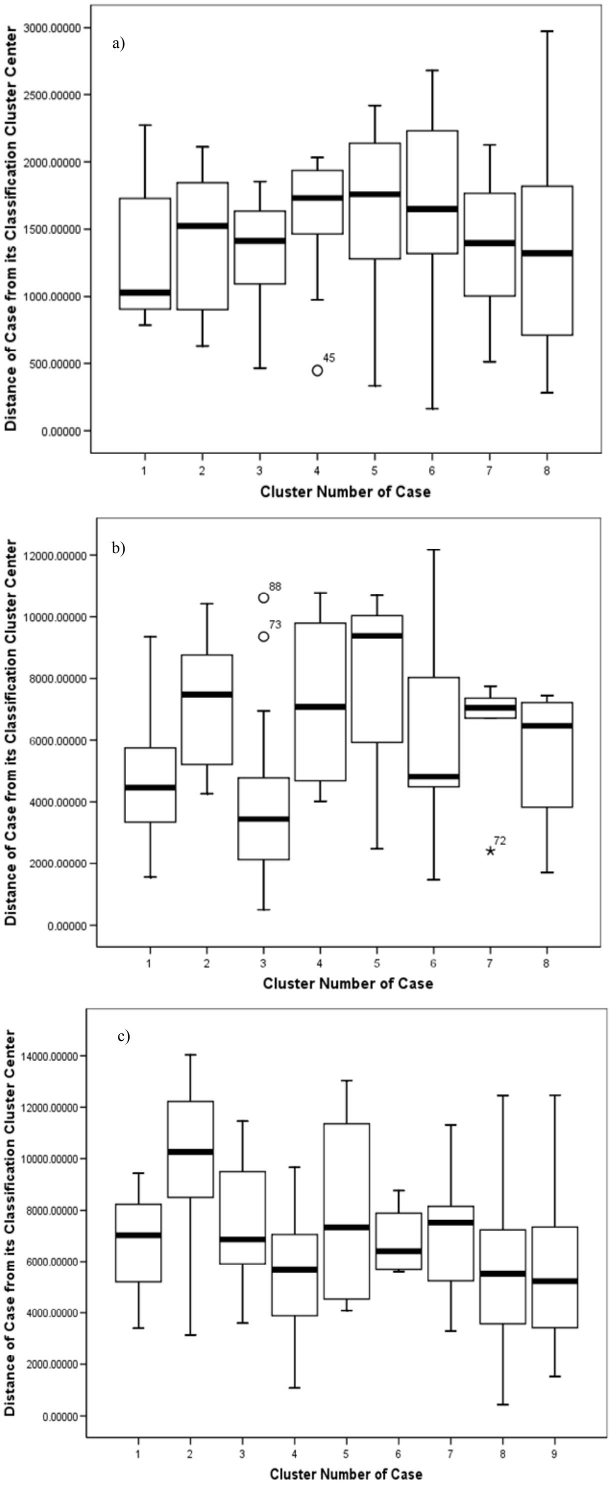

Figure 7. A box plot illustration of cluster analysis of a) adult asthma data set, b) childhood asthma data set, and c) obstructive sleep apnea data set.

major roads and point-source respirable particulate air pollution or field measurements of particulate matter. The associations are particularly evident in the top right corners of Figures 5b through 5e; and in the centers and towards the left corners of Figures 5g through 5j. The obstructive sleep apnea clusters were firmly within the negative health factors, including the fact that most of cases were located in close proximity to fast-food restaurants. The component plane for the length of stay and age of patients for obstructive sleep apnea exhibited two major clusters. The length of stay in one cluster was given at slightly more than 2 days, while the age of patients in another cluster was between 35 and 40 years.

Figures 6a through 6c illustrate extracted feature subclasses of SOM-trained data using a GIS. Figures 6a, 6b, and 6c illustrate major and minor subclasses of adult asthma, childhood asthma, and obstructive sleep apnea, respectively. The GIS maps clearly demonstrated both major and minor feature subclasses of SOM data. These maps have further identified meaningful feature subclasses (geographies) of the adult asthma, childhood asthma, and obstructive sleep apnea data sets. The largest cluster for the SOM-trained adult asthma data set was observed on Buffalo's west side and the downtown areas, while for childhood asthma there was only one very large cluster located in similar geographic settings. The other two clusters located on Buffalo's north side and south side are minor ones if the actual number of SOM hits measured within a distance of 1,000 m is quantified. For the SOM-trained obstructive sleep apnea data set, the largest cluster was observed in Buffalo's west side and the downtown areas. To some extent, the SOM data for asthma and obstructive sleep apnea exhibited a similar spatial distribution and pattern. While it was apparent that the spatial distribution and patterns of asthma were predominately located near the major roadways and the Peace Bridge Complex, obstructive sleep apnea is slightly more widespread even in the suburbs and surrounding neighborhoods. Overall, the spatial patterns discovered between the original features of adult and childhood asthma are consistent with the SOM-trained data, but a slight difference emerges for the SOM-trained obstructive sleep apnea data set. The newly derived SOM feature subclasses both for asthma and obstructive sleep apnea data sets require further evaluation.

I also conducted a comparison of clusters of the SOM-trained data in both SOM toolbox and SPSS software using the K-means clustering approach. Figures 7a through 7c give the results of a cluster analysis using box plots as a diagnostic tool. Figure 7a shows the adult asthma data set with three major clusters (2, 7, and 8). Figure 7b shows the childhood asthma data set with two major clusters (1 and 3) and has an outlier in Cluster 7. Figure 7c shows the obstructive sleep apnea data set with three major clusters (4, 8, and 9). The others clusters are minor ones. Overall, there are some slight variations in clusters, but all of these distances are reasonable. Indeed, according to these box plots, there are two to three very good clusters coming both from the SOM and SPSS data. The SPSS K-mean clustering and SOM toolbox had a perfect match of 100% for the largest clusters. For the other clusters, some slight differences were observed, but the match was still more than 95%. 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 the SOM toolbox and the SPSS software captured the data sets effectively.

The box plot for adult asthma feature data set indicates that Cluster 8 is the largest cluster, followed by Clusters 7 and 3. With the childhood asthma feature data set, the largest cluster is Cluster 3, and the second largest one is Cluster 1. Although the SOM toolbox identified Cluster 6 as another good cluster, further analysis using the box plot does not support this feature subclass as a viable cluster. The childhood asthma feature data set also has an outlier in Cluster 7. In the obstructive sleep apnea feature data set, the largest cluster belonged to Cluster 4, followed by Clusters 8 and 9.

The remaining clusters in the three feature data sets were minor ones; they had slight variations, and the distances between them were reasonable. The box plot confirmed further that all of the clusters representing feature subclasses and surrounding neighborhoods were within a reasonable distance. Overall, the clusters derived using SOM toolbox are consistent to the ones derived using the SPSS *K*-means clustering method.

4. Conclusions

The SOM provides excellent visualization and exploration frameworks for analyzing vast quantities of spatially oriented biomedical data. These experiments show that when the SOM algorithm is combined with GIS methods, they are even more powerful tools for exploratory analysis than when they are applied separately. This novel approach is both robust and superior because it enjoys three potential benefits that are lacking in conventional clustering and visualization techniques. First, both methods provide a platform for the visual exploration of multidimensional data. Second, the SOM algorithm is computationally very powerful and efficient, and it allows for automatic determination of clusters. This algorithm, as I have illustrated in this study, was able to identify clusters of similar sequences; project and visualize high-dimensional data spaces; preserve topological relationships between data vectors during training through the use of neighborhood functions. More important, it was robust regarding weight vectors initialization. The benefits of applying the SOM algorithm to geospatial data are valuable because the data often come with multiple attributes where the dimensionality, complexity, and volume are prohibitively large for manual analysis. Third, the methods in GIS preserve topological data structures (Samet, 1990, 1995) of original spatial features. Through GIS, I mapped both original and SOM data, which revealed the structure of clusters and sub clusters and from these I gleaned fundamental insights and characteristics regarding the three data sets.

The experimental results clearly illustrated the valuable properties of a set of data reduction algorithms to visually explore and analyze spatially oriented biomedical data. The tools provide a very useful exploration environment to support the formulation of new and better study hypotheses regarding the spatial distribution of a particular disease. I gained significant novel insights into the spatial characteristics of patient data and I identified three main subsets (geographies) of asthma and OSA in the study region that require further evaluation. This approach was also essential for improving interpretations of the previous findings reported in Oyana and Lwebuga-Mukasa (2004), Oyana et al. (2004), and Oyana and Rivers (2005). I further confirmed that asthma is more prevalent in Buffalo's west side which is in close proximity to major roadways, the Peace Bridge Complex and pollution sources. The spatial distribution and patterns of asthma and OSA were similar, suggesting that the two diseases track together. Interestingly, I found the clusters to be located at the same geographic locations, supporting the hypothesis that the output of these data reduction algorithms provides the best representative set of the original data features.

The quantization and topological errors indicate that the measure of the quality of the SOM during training was negligible and that there were greater improvements in the error component of trained maps. Such findings suggest that the training was adequate and that the topology was well preserved. However, I would like to reduce the error component further by incorporating a mathematical improvement model. In future work, I am going to investigate mathematical adjustments to the SOM model by improving its learning rate. My experience in applying the SOM algorithm has led to a number of efficiency and convergence issues, which I plan, to address in future studies. These include (1) speed and quality of clustering; (2) the number of clusters; (3) the updating procedure for the output neurons; and (4) the learning rate in the SOM model.

Recent work in applying the SOM model and its suggested variants (Cuadros-Vargas and Romero, 2002; Guo et al., 2003; Skupin and Fabrikant, 2003; Guo et al., 2004; Yang et al., 2003; Bação et al., 2004, 2005; Huang et al., 2005; Cuadros-Vargas and Romero, 2005; Oyana et al., 2005a, b; Guo et al., 2006) have significant implications with regards to how we extract relevant information or gain fundamental insights from very large-scale datasets. These studies along with this one could be valuable in advancing our understanding of the biomedical and epidemiological processes of diseases in relation to space and time.

Acknowledgments. Supported in part by SIUC Faculty Start-up Fund and SIUC ORDA Faculty Seed Grant. Mrs. Damalie Oyana, a computer programmer, was responsible for modifying and compiling the SOM code for these experiments. Thanks to Dr. Jamson S. Lwebuga-Mukasa, Department of Internal Medicine, University of Buffalo School of Medicine and Biomedical Sciences, Kaleida Health Systems Buffalo General Division, for providing the data sets. Thanks to Dr. Chris Lant, Department of Geography and Environmental Resources, Southern Illinois University Carbondale for agreeing to review this manuscript and for his valuable suggestions.

References

Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press, New York, NY, pp. 359.

Bação, F., Lobo, V., and Painho, M. (2004). *Geo-self-organizing map (Geo-SOM) for building and Exploring Homogenous Regions*, In:

Egenhofer, M.J., Freksa, C., and Miller, H.J. (eds.) Lecture Notes in Computer Science 3234, Geographical Information Science, Springer-Verlag, Berlin Heidelberg, pp. 22-37.

Bação, F., Lobo, V., and Painho, M. (2005). The self-organizing map, the Geo-SOM, and relevant variants for geosciences, *Computers and Geosciences*, 31, 155-163, doi:10.1016/j.cageo.2004.06.013.

Bock, T. (2004). A new approach for exploring multivariate data: self-organizing maps, *International Journal of Market Research*, 46(2), 189-203.

Borg, I., and Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer-Verlag, Berlin Heidelberg, pp. 496.

Cuadros-Vargas, E. and Romero, R. (2002). A SAM-SOM Family: Incorporating Spatial Access Methods into Constructive Self-Organizing Maps, *In Proceedings IJCNN'02, International Joint Conference on Neural Networks. Hawaii*, HI. 2002. IEEE Press.

Cuadros-Vargas, E. and Romero, R.A.F. (2005). Introduction to the SAM-SOM* and MAM-SOM* Families*, In Proceedings of the International Joint Conference on Neural Networks (IJCNN'2005)*, Montréal.

Davies, D. L., and Bouldin, D.W. (1979). A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227.

Ding, C.H. (2002). A probabilistic model for dimensionality reducetion in information retrieval and filtering, *In Proceedings of the second IEE International Conference on Data Mining*, December 2002, pp. 147-154.

Duin, R.P.W., Pekalska, E., Ridder, and de D. (1999). Relational discriminant analysis, *Pattern recognition letters*, 20, 1175-1181.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization, *Intelligent data analysis*, 5, 373-384.

Guo, D., Gahegan, M., and MacEachren, A. (2004). *An Integrated Environment for High-dimensional Geographic Data Mining GIScience 2004*, Adelphi, MD, pp.107-110.

Guo, D., Gahegan, M., MacEachren, A., and Zhou, B. (2005). Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach, *Cartography and Geographic Information Science*, 32(2), 113-132, doi:10.1559/152304 0053722150.

Guo, D., Peuquet, D., and Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional Geodata, *GeoInformatica*, 7(3), 229-253.

Hartigan, J. (1975). *Clustering Algorithms. John Wiley and Sons*, New York, pp. 351.

Huang, S., Ward, M.O., and Rundensteiner, E.A. (2005). Exploration of dimensionality reduction for text visualization, *In Proceedings of the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'05)*.

Jain, A.K., and Dubes, R.C. (1988). *Algorithms for Clustering Data,* Prentice-Hall Advanced Reference Series. Prentice-Hall Inc, Upper Saddle River, New Jersey, pp. 334.

Jiang, B., and Harrie, L. (2004). Selection of streets from a network using self-organizing maps, *Transactions in GIS*, 8(3), 335-350, do i:10.1111/j.1467-9671.2004.00186.x.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, 43, 59-69, doi:10.1007/ BF00337288.

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, 2-4 September 1998.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edition, Springer Press, Berlin, Heideberg, pp. 501.

Koua, E.L., and Kraak, M.J. (2004). Geovisualization to support the exploration of large health and demographic survey data, *Int. J. Health Geogr.*, 3, 12, doi:10.1186/1476-072X-3-12.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations, *In the Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, Berkeley, University of California Press ,1, 282-297.

Manduca, A. (1994). *Multiparameter medical image visualization with self-organizing maps*, IEEE World Congress on Computational Intelligence, 27 June-2 July 1994, IEEE International Conference on Neural Networks, 6, 3990-3995.

Murray, A.L., and Estivill-Castro, V. (1998). Cluster discovery techniques for exploratory spatial data analysis, *Int. J. Geogr. Inf. Sci.*, 12(5), 431-443.

Naenna, T., Bress, R.A., and Embrechts, M.J. (2003). DNA classifications with self-organizing maps, *In Proceedings of the 2003 IEEE International Workshop on Soft Computing in Industrial Applications (SMCIA 2003)*, 23(25), 151-154.

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.

Openshaw, S., Blake, M., and Wymer, C. (1995). *Using Neurocomputing Methods to Classify Britain's Residential Areas*, In P. Fisher (ed.) Innovations in GIS, Taylor and Francis, 2, 97-111.

Openshaw, S., and Openshaw, C. (1997). *Artificial Intelligence in Geography*, John Wiley and Sons, New York, Chichester, pp. 329.

Openshaw, S. (1998). *Building automated Geographical Analysis and Exploration Machines*, In: Geocomputation: A primer ,Longley, P. A., Brooks, S. M. and Mcdonnell, B. (eds.), Chichester, Macmillan Wiley, pp. 95-115.

Oyana, T.J., Boppidi, D., Yan, J., and Lwebuga-Mukasa, J.S. (2005a). Integration of self-organizing maps into a geographic information systems 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, 5th-8th April 2005.

Oyana, T.J., Boppidi, D., Yan, J., and Lwebuga-Mukasa, J.S. (2005b). Exploration of geographic information systems-based medical databases with self-organizing maps: A case study of adult asthma, *In Proceedings of the 8th International Conference on Geo-Computation*, 1st-3rd August 2005, Ann Arbor, University of Michigan.

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, *J. Environ. Health*, 66(8), 25-38.

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, *Am. J. Public Health*, 94(7), 1250-1257.

Oyana, T.J., and Rivers, P.A. (2005). Geographic variations of childhood asthma hospitalization and outpatient visits and proximity to ambient pollution sources at a U.S.-Canada border crossing, *International Journal of Health Geographics*, 4(1), 14, doi:10.1186/147 6-072X-4-14.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, London, pp. 416.

Samet, H. (1990). *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, Reading, MA, pp. 493.

Samet, H. (1995). *Spatial Data Structures in Modern Database Systems: The Object Model, Interoperability, and Beyond*, W. Kim, Ed., Addison-Wesley/ACM Press, 1995, 361-385.

Sammon, J.W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computing,* 18, 401-409.

Skupin, A., and Fabrikant, S. (2003). Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization, *Cartography and Geographic Information Science*, 30(2), 99-110, doi:10.1559/152304003100011081.

Sugiyama, A., and Kotani, M. (2002). Analysis of gene expression data by using self-organizing maps and k-means clustering, *In Pro- ceedings of the 2002 International Joint Conference on Neural Networks*, 12-17 May 2002, 2, 1342-1345.

Tamminen, S., Pirttikangas, S., and Roning, J. (2000). Self-organizing maps in adaptive health monitoring, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN) 24-27 July 2000*, 4, 259-264.

Theodoridis, S., and Koutroumbas, K. (2003). *Pattern Recognition*, Second Edition, Academic Press, San Diego, California; London, pp. 689.

Vesanto, J., and Alhoniemi, E. (2000). Clustering of the self-organizing map, *IEEE Transactions on Neural Networks*, 11(3), 586-600.

Yang, J., Ward, M.O., Rundensteiner, E.A., and Huang, S. (2003). Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets, *Joint Eurographics-IEE TCVG Symposium on Visualization (VisSym 2003)*.