DETECTING ENVIRONMENTAL CHANGE USING SELF-ORGANIZING MAP TECHNIQUES APPLIED TO THE ERA-40 DATABASE

***Mohamed Gebril*[1], *Eric Kihn*[2], *Eyad Haj Said*[3], *and Abdollah Homaifar*[4*]**

[1,4] *Autonomous Control and Information Technology Center, Department of Electrical and Computer Engineering, North Carolina A & T State University, Greensboro, NC 27411*
**Email:* homaifar@ncat.edu
Email:mmgebril@ncat.edu
[3]*University of Kalamoom, Deratiah, Syria*
Email: ehajsaid@yahoo.com
[2]*Email: Eric.A.Kihn@noaa.gov*

ABSTRACT

Data mining is a valuable tool in meteorological applications. Properly selected data mining techniques enable researchers to process and analyze massive amounts of data collected by satellites and other instruments. Large spatial-temporal datasets can be analyzed using different linear and nonlinear methods. The Self-Organizing Map (SOM) is a promising tool for clustering and visualizing high dimensional data and mapping spatial-temporal datasets describing nonlinear phenomena. We present results of the application of the SOM technique in regions of interest within the European re-analysis data set. The possibility of detecting climate change signals through the visualization capability of SOM tools is examined.

Keywords: Meteorological Database, Data Mining, Clustering, Self Organizing Map, ERA-40

1 INTRODUCTION

There are hundreds of satellites observing and collecting enormous amounts of data about the Earth's land, oceans, and atmosphere. These satellite observations are typically combined with a ground network and models to create a comprehensive view of the state of our planet. Analyzing these data helps us to obtain a global picture and helps to identify indicators of climate change. However, the scientific community faces a serious challenge because the volume of collected climate data has increased rapidly; data generation is growing faster than analysis capabilities because of advances in observational techniques and minimal growth in the number of human analysts. A huge amount of data accumulates in databases and has never been examined by scientists trained to extract relevant information (Kantardzic, 2003).

In large meteorological databases, data mining presents an opportunity to facilitate the automation of feature extraction and pattern classification. Data mining has the potential to develop powerful tools for data analysis and interpolation of extracted patterns. Data mining *itself* can be described as an extraction of implicit, previously unknown and potentially useful information from data in databases (Berry & Linoff, 1997; Datta, 2006). Data mining is one step of knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) where scientific methods are used in an automated fashion. Typical KDD includes a method for extracting and preparing data as well as for making decisions about actions to be taken once the data mining is complete.

Meteorological data mining applications work with huge meteorological databases created by reanalysis projects such as the National Center for Atmospheric Research (NCAR) (Kalnay, Kanamitsu, Kistler, Collins, Deaven, & Gandin, et al., 1996), and the European Centre for Medium-Range Forecasts (ECMWF). In both cases, the total volume of information is of the magnitude of terabytes.

Clustering (Fayyad, 1998) is one of the unsupervised data mining techniques in which data instances are grouped together based on a similarity scheme defined by the clustering system. Clustering can be viewed as unsupervised classification of unlabelled patterns, such as observations, data items, or feature vectors (Forgy, 1965). Since several clustering techniques are available, the results and their interpretations strongly depend on the choice of clustering technique. For example, clustering can be hard or soft, so that every instance belongs to only one group or more than one group respectively. In addition, clustering can be probabilistic, in which an instance is placed in each group depending on its assigned probability. Furthermore, clustering can be hierarchical, such that there is a crude division of the instances into groups at a high level that can then be refined into a finer level. The Self Organizing Map (SOM) method (Kohonen, 2001) was chosen because it is one of the most popular neural network models based on competitive learning (Haykin, 1999) and it is especially suitable for high dimensional data visualization. The self-organizing map is a method that visualizes high-dimensional data in a two-dimensional space. This is done by keeping the topologic and metric relations of the two-dimensional space as close as possible to the relations of the initial space.

The ECMWF Re-Analysis (ERA-40) project (Simmons & Gibson, 2000) has produced a comprehensive global analysis for the 45-year period covering September 1957 to August 2002. The ERA-40 version of the atmospheric model had 60 vertical levels and a reduced Gaussian grid with an approximate uniform spacing of 125 km for surface and other fields. A Gaussian grid is used for scientific modeling on a sphere. The grid is rectangular, with orthogonal coordinates (usually latitude and longitude) chosen such that they can be easily accessed in a fixed array. The longitudes are equally spaced while the latitudes are not equally spaced and are defined by their Gaussian quadrate. There are no grid points at the poles, and the number of longitudes is usually double the number of latitudes. A reduced Gaussian grid is a grid in which the number of grid points in the rows decreases towards the poles, which keeps the grid-point separation approximately constant across the sphere (Hortal, 1991). Data is composed of approximately 45 variables (such as temperature, humidity, pressure, etc.) at 23 different pressure levels of a 360 × 180 nodes grid. The atmospheric model is coupled to an ocean-wave model, which resolves 25 wave frequencies and 12 wave directions at the nodes of a 1.5º grid. Multiple archives of satellite observations are assimilated by the model. These input data streams represent one of the largest and most complete collections of observations ever assembled. These data products are computed on the grid four times per day (00:00, 06:00, 12:00, and 18:00 UTC). Some fields that complement the analyses have been extracted from the 6-hourly forecast data and are included in these products. This technique promotes easier access to groups of related fields and includes all of the chemical transport, net tendency, and some of the surface and single level fields, e.g., convective precipitation and surface sensible heat flux.

ERA-40 data is stored as Network Common Data Format (NetCDF). NetCDF is a set of data formats, programming interfaces, and software libraries that help read and write scientific data files. The NetCDF was developed and maintained at Unidata, which is funded primarily by the National Science Foundation, one of the eight programs in the University Corporation for Atmospheric Research (UCAR) Office of Programs (UOP).

Global atmospheric reanalysis projects (GARPs) are a major advance for studies in meteorology, oceanography, climatology, and other fields. The GARP uses fixed data assimilation systems that eliminate inconsistencies found in archives of operational model outputs caused by changes to the operational models. The GARP supply gridded data for early periods that are deficient (e.g., 1957-1969) or entirely lacking. It also consistently applies advanced data assimilation systems over longer periods than was previously possible. The GARPs have benefited from data recovery efforts undertaken at several institutions, including NCAR. These data sets are, therefore, ideal for the application of data mining tools. The long time period combined with a consistency of data make for a data set well suited for automated information extraction.

In this paper, we analyze ERA-40 data in certain regions of interest by applying clustering techniques to extract new patterns in order to detect climate change over the 40 years. For this purpose, we employ an unsupervised data mining clustering technique known as a Self-Organizing Map (SOM) (Kohonen, 2001).

This paper is organized as follows: in Section 2, we describe our approach using SOM. In Section 3, we show experimental results; and finally in Section 4, we conclude our work and present future plans.

2 APPROACH

While the analysis of large spatial-temporal datasets can be achieved using different linear and nonlinear methods, we consider the SOM to be a very promising tool for clustering. It is preferable to other clustering techniques because the spatial organization of the neurons maps cluster similarity into prototype proximity in the 2D space. The distance among prototypes in the SOM map can, therefore, be considered as an estimate of the similarity between objects belonging to clusters. In addition, SOM is an excellent tool for visualizing high dimensional data and mapping spatial-temporal datasets describing nonlinear phenomena. Figure 1 shows the main stages of our work.

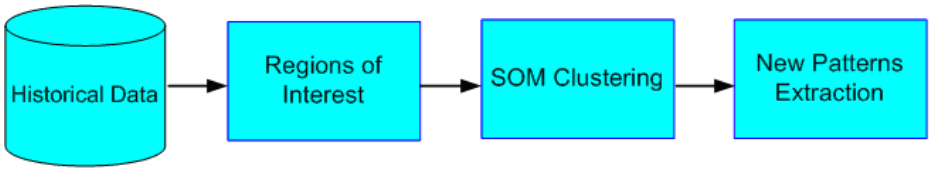

Figure 1. Main Stage of the proposed approach

2.1 Self-Organizing Map

Due to its high-dimensionality, the data was first analyzed and simplified before proceeding with other analyses (Comrie & Glen, 1999). SOM is a special case of a neural network that simultaneously summarizes a set of variables and clusters observations. It can be viewed as a Principal Component Analysis (PCA) combined with a cluster analysis with both procedures influencing each other in the algorithm. The most commonly used algorithm for constructing the SOM follows the work of Kohonen (1995). The Kohonen maps are known to be "topology preserving," i.e., observations with similar multi-dimensional vectors of variables find themselves positioned close to each other on the map, and they are also "self-organizing," i.e., the variables tend to vary along the map in a meaningful way.

Data clustering in a SOM is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The input data for the model is described in terms of vectors, each of which consists of components representing an elementary feature (for instance, sea ice cover, sea surface temperature, etc.) of the data item, expressed as a numeric value. The output is a similarity-based map of the data items, similarity being defined as proximity of items in the feature space. The SOM is a neural model in which the output neurons will compete to be activated or fired. The neuron that wins the competition by having the minimum value using some distance measurement method (e.g., Euclidean distance) is called the "winning neuron." In a "winner-takes-all" (WTA) strategy, only the weights associated with the winner neuron are updated. There is also a winner takes most (WTM) strategy, where many neurons in the neighborhood of the winner neuron adapt their weights in each learning iteration (Brocki, 2007).

The lattice is usually one or two-dimensional. Higher dimensional maps are possible but not as common because the goal of the map is visualization for the applications of this kind of network. Figure 2 shows the structure of the 5x5 lattice in the SOM.

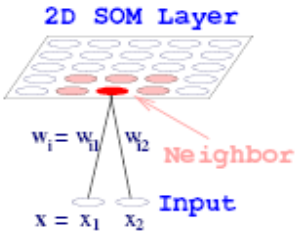

Figure 2. Illustration of the SOM model

The main goal of the SOM is to transform an input signal pattern of arbitrary dimension into a lower (1, 2, or 3 dimensions) dimensional discrete map and to perform this transformation adaptively in a topologically orderly fashion.

The SOM leads to a topological ordering of the feature map in the input space in the sense that neurons that are adjacent in the lattice will tend to have similar synaptic weight vectors. The SOM algorithm consists of 5 steps: 1) choose an input pattern *v* of dimension *m* from the input space randomly; 2) assign the weight vectors $\boldsymbol{w}_i(0)$values randomly; 3) find the *best matching* or *winning* neuron *k* at time step *n* by using the Euclidean minimum-distance criterion as shown in Eq.(1); 4) update the weights using Eq.(2) and the neighborhood kernel function Eq.(3); and 5) repeat the weight updating process until there are no noticeable changes in the feature map.

$$k = \arg \min_i \left\| v(n) - w_i(n) \right\|, where \quad i = 1,...,N \tag{1}$$

Where *n* and N are time step and total number of neurons respectively. Then the weights are updated as:

$$w_i(n+1) = w_i(n) + \eta(n)\, h_{ki}(n)\, [v(n) - w_i(n)] \tag{2}$$

Where η (n) is the adaptation (learning) coefficient, and the neighborhood kernel function ($h_{ki}(n)$) is defined as:

$$h_{ki}(n) = \exp\left[-\frac{\left\| r_k - r_i \right\|^2}{2\sigma^2(n)} \right] \tag{3}$$

Where r_k and r_i are the positions of neuron k and i on the SOM grid, and $\sigma(n)$ is the width of the neighbourhood function (Vesanto & Alhoniemi, 2000). The figures 3a, 3b, and 3c show an example of the SOM where a data set is represented by SOM 10x10 Map nodes. In these figures the red plus signs stand for the trained data while the black circles stand for the map nodes. The map nodes are randomly initialized to values between 0 and 1. The two input nodes represent X and Y values between -1 and 1. Therefore, by training the SOM continuously with random values for X and Y between -1 and 1, the SOM unfolds and flattens out into a grid pattern.

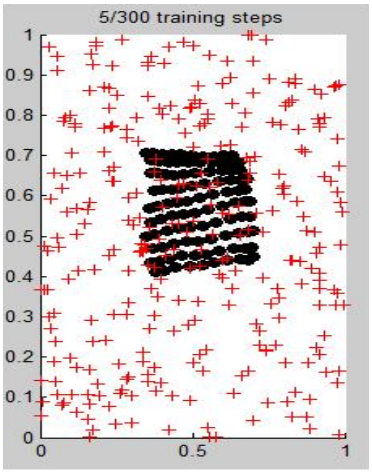

Figure 3a. Initialization

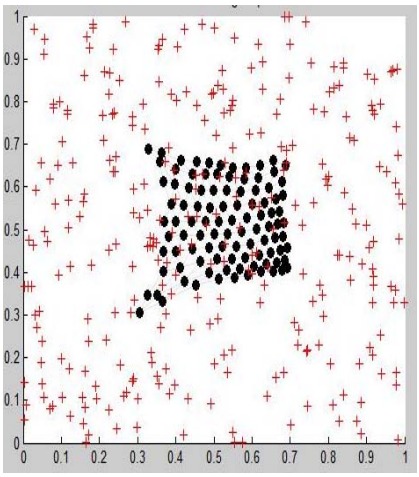

Figure 3 b. Training

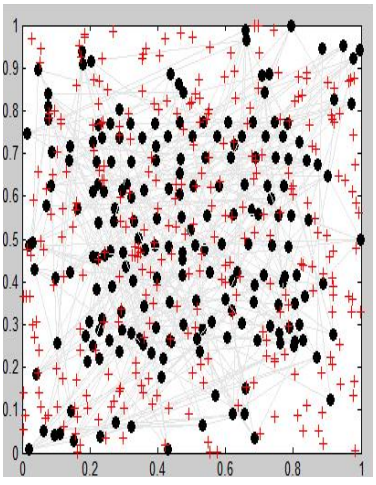

Figure 3c. Unfolding

3 ANALYSIS AND RESULTS

The ERA-40 data set is a climate data set consisting of surface and pressure level variables. In this paper, we pre-selected five variables and two specific grid points from among the 45 surface variables and hundreds of possible grid points. These variables are sea ice cover, sea surface temperature, cloud cover, U-10m wind velocity (i.e., the

east wind velocity 10 m above the sea level), and V-10m wind velocity (i.e., the north wind velocity 10 m above the sea level).

Initially the data need to be normalized to a zero mean value with a variance σ=1. The normalized input values are given random weights. All the units in the input layer are connected to each neuron of the output layer through the weights. At this point, the neurons of the networks are trained. The corresponding weights are modified based on the position of the input data. The aim of the neurons is to preserve the distribution of the input data while reducing the dimension "A salient feature of SOMs is their ability to preserve a dataset's topology" (Villman, Herrmann, & Martinetz, 1997).

The data were in an ASCII file, and the SOM toolbox was used in MATLAB to analyse the data. The data were normalized, initialized linearly, and batch trained. For this research, the grid of neurons was hexagonal. Figure 4 shows the two regions of interest at 76N, 00E and 76N, 134W plotted on the globe for reference. The two locations are close to the North Pole and have a large variation in sea surface temperature. In addition, the chosen surface parameters are the most relevant variables at these locations.

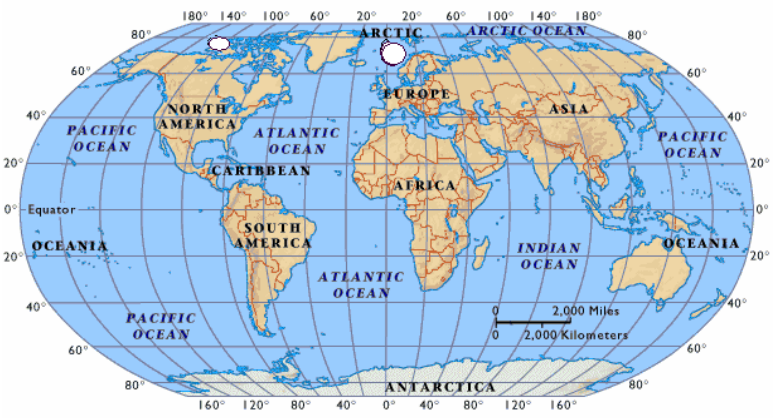

Figure 4. World atlas with black circles indicating areas of interest

For the work in this paper we converted the original Grid (Li, 2003) data files to a fixed latitude/longitude grid in Common Data Format (NetCDF) (Unidata, 2004). The NetCDF is highly suitable for a variety of data types that include single-point observations, time series, regularly-spaced grids, and satellite or radar images. The MATLAB mexnc toolbox was used to read and access the NetCDF formatted data. The SOM toolbox (Vesanto, 1999) in MATLAB was used to analyse and visualize the data. Initially, data were normalized and placed in the hexagonal neuron grid lattice of the SOM. There are then many ways to visualize the data as explained below.

3.1 U-matrix

The U-Matrix shows the unified distance matrix of inter-neurons where the smaller of the distances are of higher density and similarity among the input data and vice versa. It allows the 2-D visualization of the five-dimensional data, and it detects the topological relations among neurons. To show real distribution of the decisions on the map, we added 'hits' on the U-matrix. To visualize them, we used a 'pie' structure. This structure shows percentage of input elements near each map unit. Also, we used a hexagonal map structure with a distribution of labels. Labels were obtained from the final map using a method vote. This method calculates a number of entries of each element in each region near each member unit and then uses the label with the biggest number of entries. Since data is being collected four times a day, the total number of collected data points over forty years is equal to 160 times the number of days in a given month. For example, for the month of February, the total number of measured data is 4x40x28 for each dimension for 40 years. We made 20 experiments with 10 different sized maps. The size of the map influences mainly the scaling factor of the output space representation. Thus, we had to find a compromise between good visual appearance of the map and good representation of the map. After a series of experiments, we found that the default size of the map (21x16) gives a quite good representation of the results. Figures 5

a and b show the U-matrix and its corresponding year label of the SOM network for the month of February at 76N, 00E respectively.

In Figure 5a, the blue color represents the cells that belong to the same cluster while the yellow color represents the boundary between the clustered data. Therefore, we can identify two clusters as circled in Figure 5a. The label matrix represents the cells with the corresponding year in Figure 5b.

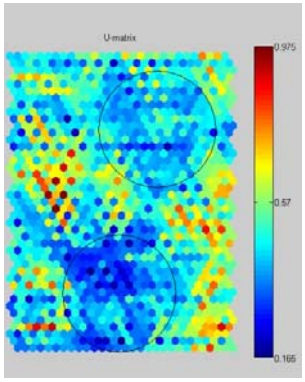

Figure 5a. U-Matrix for February (76N, 00E)

Figure 5b. Label Matrix for February (76N, 00E)

3.2 Component Planes

Component Planes (CPs) can be drawn for each input feature. They show the weights that connect each neuron with a particular input variable (Vesanto, 1999). CPs are used to detect correlations between input variables and to present the same color pattern distribution for all the CPs. CPs with the same color patterns are more strongly correlated, where light blue color represents similarity among data points and light red color represents the barrier between clusters. Figures 6 a and b represent the U-matrix and its CPs for the month of February at 76N, 00E and for January at 76N, 134W to realize the patterns on the SOM`s grid of the distributed data. The CPs have low values on the bottom and high values on top, with some exceptions.

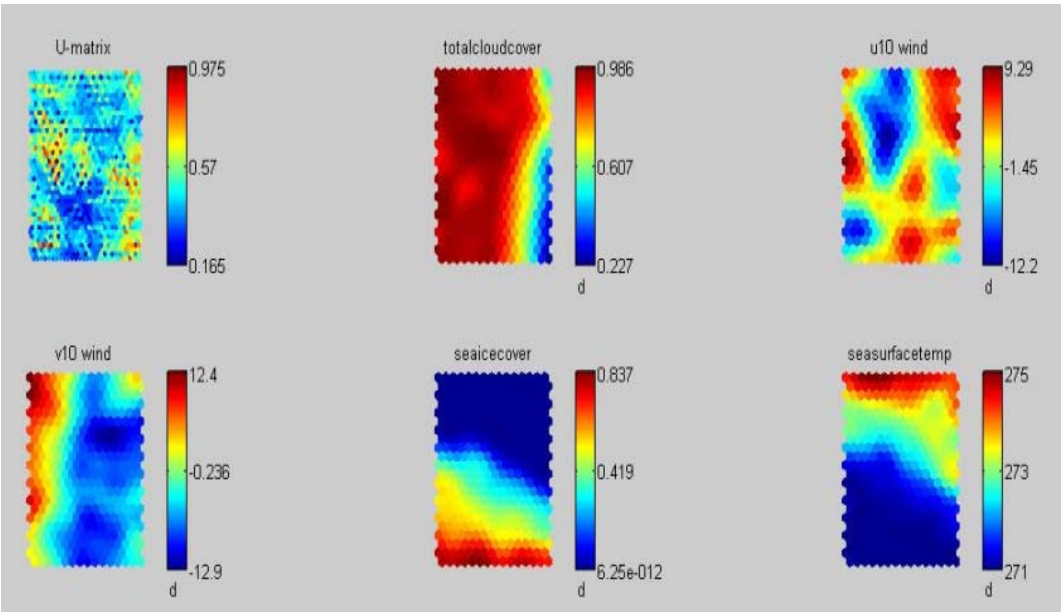

Figure 6a. U-Matrix and the 5 component planes of February at 76N, 00E

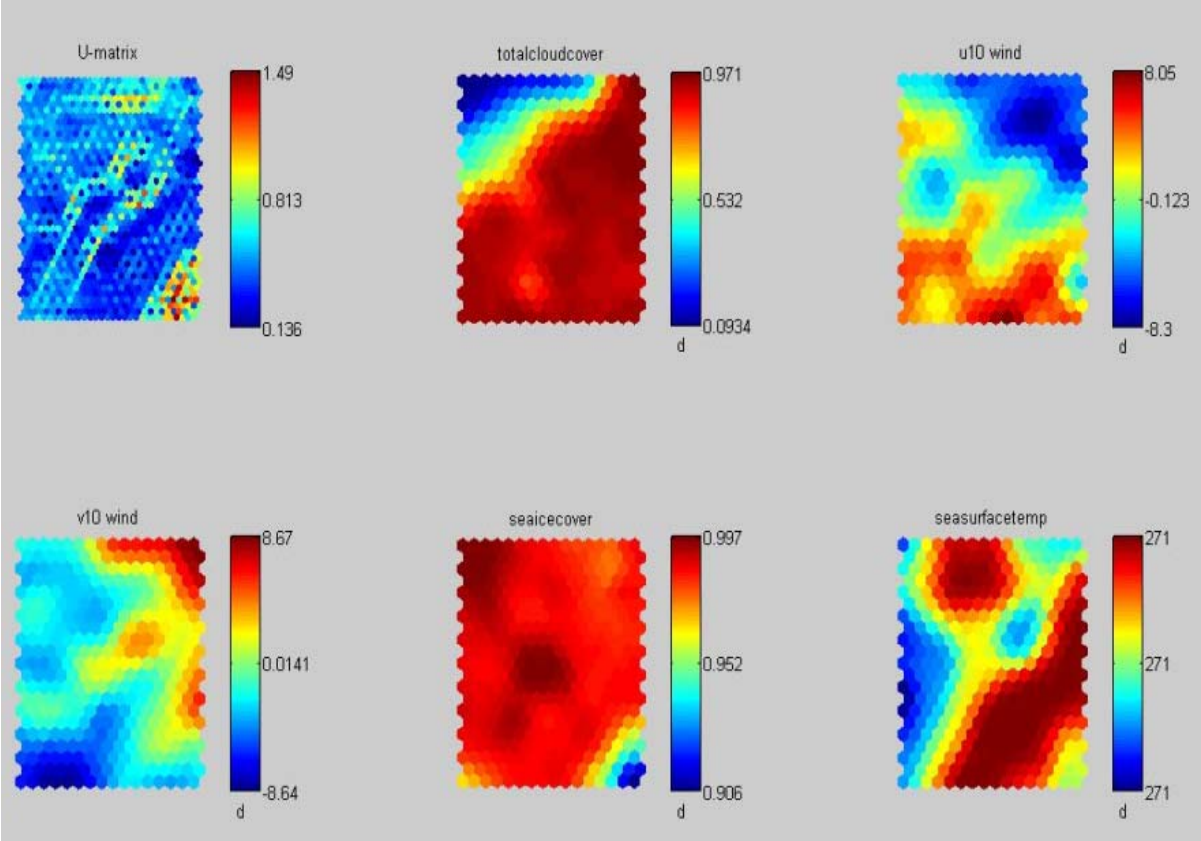

Figure 6b. U-Matrix and the 5 component planes of January at 76N, 134W

Figure 7 shows the clustered data in two groups based on the U-Matrix as shown in Figure 5a. It also shows the corresponding year label matrix.

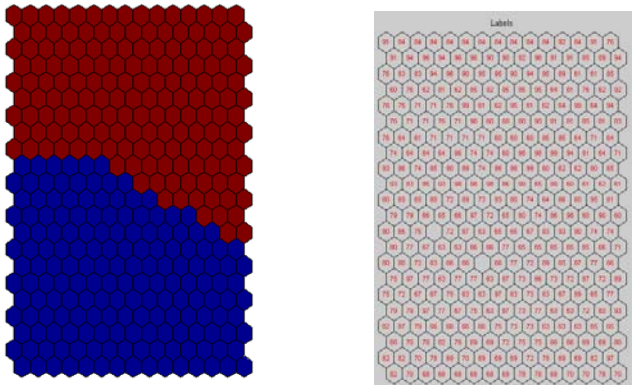

Figure 7. SOM clustered map and the corresponding label matrix

3.3 Principal Component Projections

The Principal Component (PC) projection represents the distribution of units from the output map with a different color; the color represents the distance measure between clusters (units that belong to the same cluster have the same color). Figure 8 shows that there are two big clusters. The top cluster represents the period 1960-1979. The bottom cluster has three sub-clusters. The upper sub-cluster consists of samples from the upper right corner of the U-matrix.

The middle sub-cluster consists mostly of samples from the 1990s. The bottom sub-cluster consists mostly of samples from the 1980s. Despite the noise in the output space, the ERA-40 partitioned samples quite well.

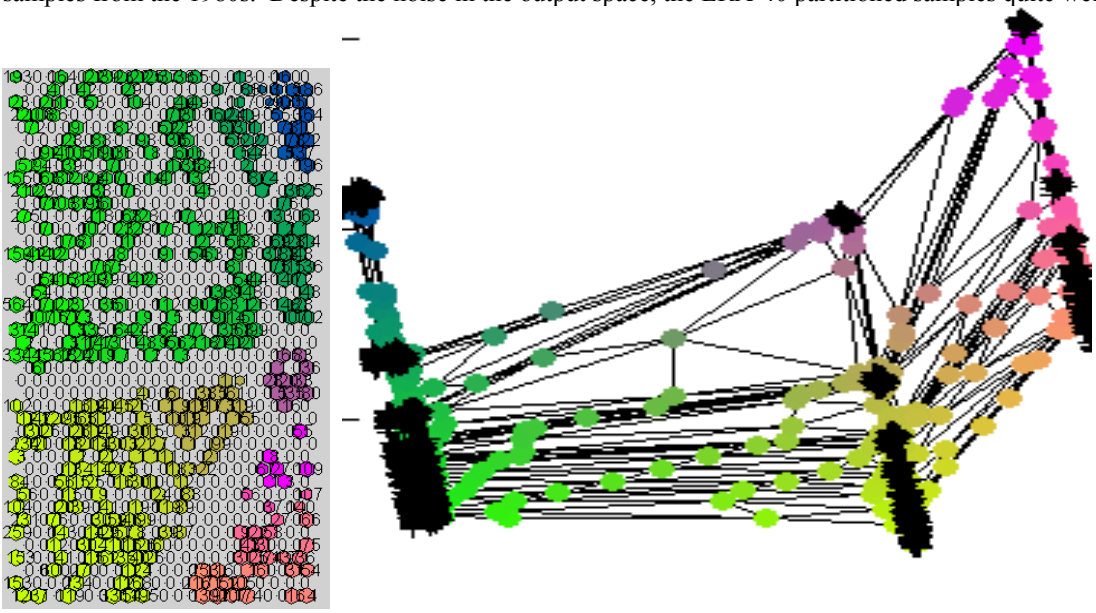

Figure 8. Color code and PC projection

Figure 9 illustrates the distribution of the values of each variable in the output map. This information can be used either as an additional tool for analysing data or for the confirmation of earlier conclusions. This bar plane confirms our conclusions about the ERA-40 data classification.

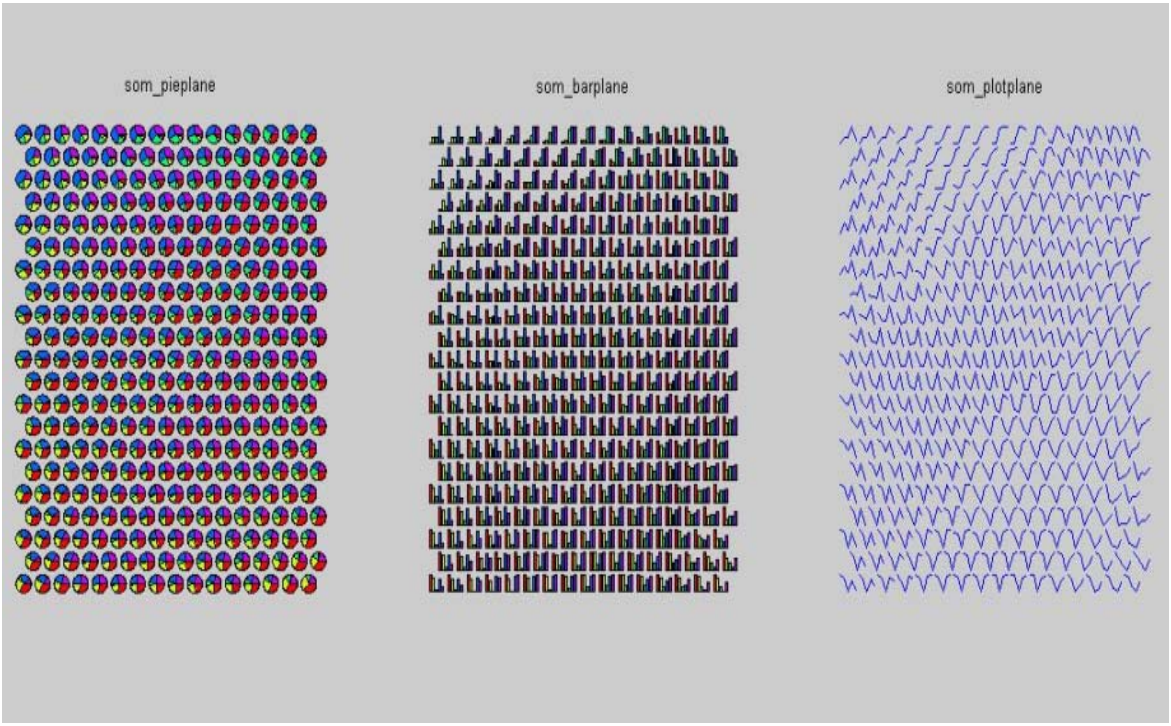

Figure 9. Pipeline, Barplane, and Plotplane of U-matrix

Furthermore, from the year label matrix, we computed the histogram of each cluster based on the number of cells that belong to each year as shown in Figure 10. For example, 12 cells from cluster one were labelled as 1960 in the year label matrix. Figure 10 shows that data in cluster 1 started to appear more often in the early 1960s and after the 1980s. On the other hand, data in cluster 2 appeared more in the period between 1965 and 1970.

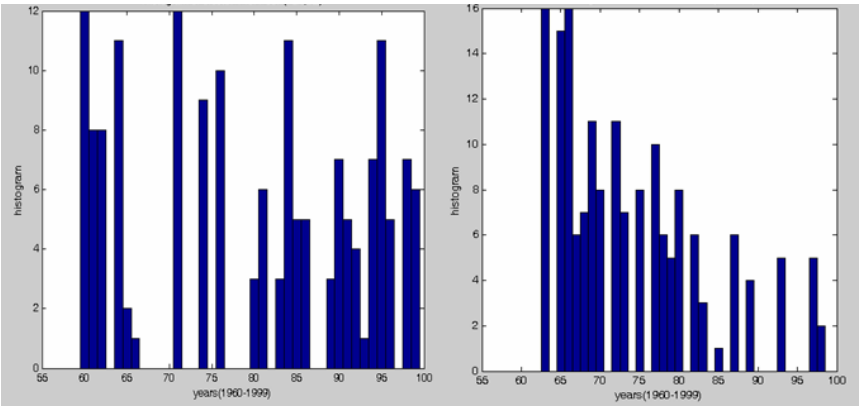

Figure 10. Histogram for cluster 1 (left) and cluster 2 (right)

In order to show that the proposed technique is robust and applicable, the same analysis and procedure explained earlier has been applied to another location, 76N, 134 W. The clustering and histogram distributions are shown in Figures 11, 12, and 13 for the same data variables as for the aforementioned location.

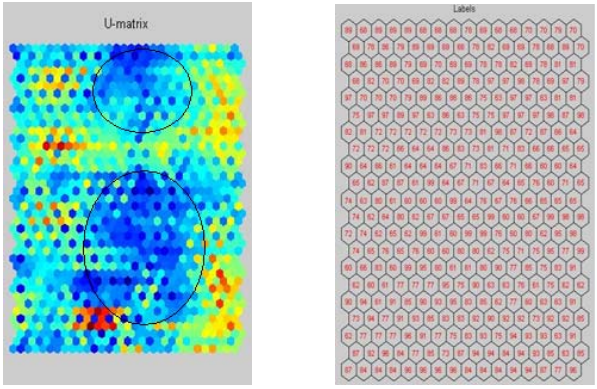

Figure 11. U-Matrix and its corresponding Label Matrix

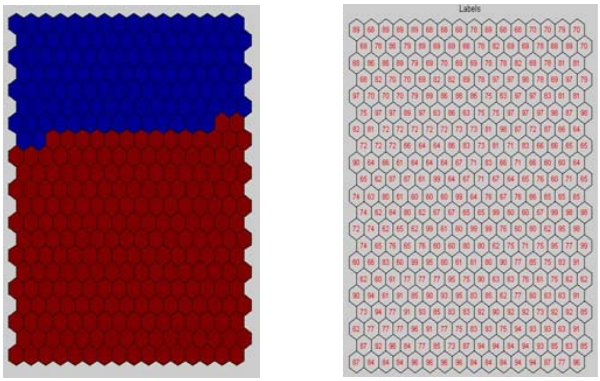

Figure 12. The clustered map and its Label Matrix

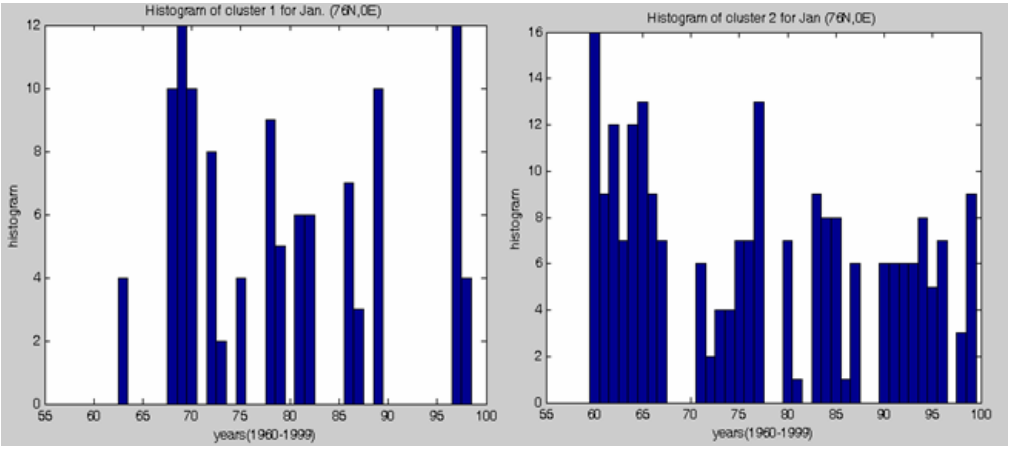

Figure 13. Histogram for cluster 1 (left) and cluster 2 (right)

The results from Figures 11, 12, and 13 show that the data in cluster 1 appeared more frequently in the 1970s and 1980s while data in cluster 2 appeared more frequently in the 1960s. In addition, the total amount of data in cluster 2 is larger than in cluster 1, with the amount of data in each cluster indicated by the frequency of appearance within each histogram.

4 CONCLUSION & FUTURE WORK

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

The SOM method can identify from a vast collection of data changes which states may be indicators of climate change. By looking at changes in clustering over extended periods, multiple variables, and across many grid points, the technique can be used to alert the interested analyst to detected change. Whether this change is attributed to climate, natural cycles, or other causes is a matter of interpretation, but the technique vastly reduces the amount of data to be analyzed.

In this study, data clustering on a SOM has been applied to the ERA-40 database over a 40 year period. The choice of SOM was based on the fact that it is a powerful tool for clustering and visualizing multi-variable data. We have shown that SOM can be used to investigate climate change in an untraditional and automated way and to extract potential new indicators of climate change. We focused on two locations close to the Arctic Ocean that can be extended as regions. The results show that the data can be clustered into two distinct groups. The data in one cluster appear more frequently in the 1970s and 1980s while data in the other cluster appear more in the 1960s. This indicates that there is a temporal component to the climate change or the mode of climate defined by these variables. It is commonly argued that the Self-Organizing Map can allow the extraction of patterns and the creation of abstractions, whereas 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 the understanding of complexity and to enhance the overall effectiveness of exploratory data analysis.

In the future, we plan to further investigate climate change phenomenon at different locations using different variables. This work is fully applicable to different climate databases.

5 REFERENCES

Berry, M. J. A. & Linoff, J. (1997) Data Mining Techniques, New York: John Wiley &Sons Inc.

Brocki, L. (2007) Kohonen Self-Organizing Map for the Traveling Salesperson Problem, Berlin Heidelberg: Springer.

Cavazos, T. (2000) Using self-organization maps to investigate extreme climate event, Journal of Climate 13:1718–1732.

Comrie, A. C. & Glen, E. C. (1998) Principal components based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability, Climate Research 10:201–215.

Datta, S. K. (2006) Means clustering over a large, dynamic network, in Proceedings of 2006 SIAM Conference on Data Mining, Bethesda, MD, April.

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

Edelstein, A. H. (1999) Introduction to Data Mining and Knowledge Discovery. Third Edition. Potomac, MD: Two Crows Corporation.

Fausette, L. (1994) *Fundamentals of Neural Networks–Architecture, Algorithm, and Applications*, Press Location: Prentice-Hall, 1994

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996) From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, pp. 1–30. Menlo Park, Calif.: AAAI Press.

Fayyad, U. (1998) Scaling clustering algorithms to large databases, *Technical Report MSR-TR-98-37*, Redmon, WA: Microsoft.

Forgy, E. (1965) Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications, *Biometrics* 21: 768.Hartigan, J. A., (1975), *Clustering Algorithms*, New York: John Wiley & Sons Inc.

Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, second edition, Delhi, India: Prentice Hall

Han, J. & Kamber, M., (2001) Data mining concepts and techniques, San Francisco: Academic Press.

Heim, Jr., R. R. (2002) a review of Twentieth-Century drought indices used in the United States. Bulletin of the American Meteorological Society, 83:1149-1165.

Hilario, L. G. & González, I. M. (2004) Self-organizing map and clustering for wastewater treatment monitoring, in *Engineering Applications of Artificial Intelligence*, Amsterdam: Elsevier, pp. 215-225.

Hortal, M. (1991) Use of Reduced Gaussian Grids in Spectral Models.Mon.Wea.Rev, 119, 1057-1074.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woolten, J., Zhu, Y., Chellian, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., & Joseph, D. (1996) The NCEP/ NCAR 40-year reanalysis project. *Bull. Amer. Meteorol. Soc.* 77: 437-471.

Kantardzic, M. (2003) *Data mining concepts, models, methods and algorithms*. New York, Wiley–Interscience.

Kaski, S. & Kohonen, T. (1996) Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world: Neural Networks in Financial Engineering. *Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, Singapore, pp. 498-507.

Kohonen, T. (1989) *Self Organization and Associative Memory*, Berlin: Springer-Verlag.

Kohonen, T.,(1995*)* *Self Organizing Maps*, Berlin: Springer-Verlag.

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

Lawrimore, J., et al. (2002) Beginning a new era of drought monitoring across North America. *Bulletin of the American Meteorological Society*, 83:1191-1192.

Li, J., et al. (2003) Parallel netCDF: A high-performance scientific I/O interface, in *Proceedings of the 2003 ACM/IEEE Conference on Supercomputing*, pp. 39–49, Association for Computing Machinery, IEEE Computer Society, Washington, DC, USA, November 15–21, Phoenix, AZ.

Lott, N., & Ross, T. (2000) NCDC Technical Report 2000-02, *A Climatology of Recent Extreme Weather and Climate Events*. Asheville, N.C.: National Climatic Data Center.

Simmons, A. J. & Gibson, J. K. (2000) The ERA-40 Project Plan. ERA-40 Project Report Series.

Svoboda, M., et al. (2002),The Drought Monitor. *Bulletin of the American Meteorological Society*, 83: 1181-1190.

Unidata (2004): http://www.unidata.ucar.edu/packages/netcdf.

Uppala, S. M., et al. (2005) The ERA-40 re-analysis, *Q. J. Roy. Meteor.Soc*., 131:2961–3012.

Vesanto, J. & Alhoniemi, E. (2000) Clustering of the Self-Organizing Map, *IEEE Transactions on Neural Networks* 11 (3):586-600.

Vesanto, J. (1999) Self-Organizing Map for Data Mining in MATLAB: the SOM Toolbox.

Villmann, T., Herrmann, R., & Martinetz, T. (1997) Topology Preservation in Self-organizing Feature Maps: Exact Definition and Measurement. IEEE Trans. on Neural Networks, 8(2), 256-266.

(Article history: Received 28 January 2010, Accepted 17 February 2011, Available online 11 May 2011)