Classification of similar productivity zones in the sugar cane culture using clustering of SOM component planes based on the SOM distance matrix

Miguel A. Barreto S.^{1,2,3} and Andres Pérez-Uribe^{2}.
^{1}Université de Lausanne, Hautes Etudes Commerciales (HEC),
Institut des Systèmes d'Information (ISI).
Miguel-Arturo.Barreto-Sanz@heig-vd.ch
^{2}University of Applied Sciences of Western Switzerland (HEIG-VD) (REDS).
Andres.Perez-Uribe@heig-vd.ch
^{3}Corporacion BIOTEC

Keywords: SOM component planes, distance matrix based clustering of SOM, correlation hunting.

Abstract— A technique called component planes is commonly used to visualize variables behavior with Self Organizing Map (SOM). A methodology to clustering the component planes based on the SOM distance matrix is presented. This methodology is used in order to classify zones with similar agro-ecological conditions in the sugar cane culture. Analyzing the obtained groups it was possible to extract new knowledge about the relationship between the agro-ecological variables and productivity.

1 Introduction

The agricultural productivity of a geographic area depends on many agro-ecological variables like soil and terrain characteristics, climatic constraints, human behavior and management [19]. These agro-ecological variables are interdependent and constantly evolving in time and space. Finding similar productivity zones implies to handle and to analyze a great amount of spatial and temporal data. These data are characterized by their complexity, variability in measurements, and nonlinear relations [11]. The identification of similar productivity zones is difficult without an adequate visualization of the variables and the relations between them.

In classical methods, dependencies between variables can be detected using scatter plots. In addition, when the variables are more than a pair, it is possible to organize a scatter plot matrix with several sub-plots where each variable is plotted against each other variable. However, in this technique the number of pairwise scatter plots increases quadratically with the number of variables [6]. This type of visualization is not practical in applications where the analysis of many variables is necessary.

Moreover, using visualization based on SOM components planes [7], the number of sub-plots will grow linearly with the number of variables. In addition, it is possible to classify variables with similar behaviors. Every SOM component plane is formed by the values of the same component in each prototype vector. Therefore, they can be seen as a sliced version of the map [12]. After plotting all component planes, relations between variables can be observed. The task of organizing similar components planes in order to find correlated components is called correlation hunting [15]. However, when the number of components is large it is difficult to determine which planes are similar to each other. Different techniques can be used to reorganize the component planes in order to aid the correlation hunting. The main idea is to place correlated components close to each other. One of the most used techniques is the projection of the component planes on another plane. The projection could be done using, e.g. another SOM [15].

Diverse authors have reported works related to agro-ecological variables analysis, and the classification of zones and/or patterns in the variables behavior. For instance, Hargrove and Hoffman [5] used principal-component analysis for ecoregionalization. Mingqin and Samal [9] explored the suitability of some fuzzy clustering approaches for agroecozones delineation. Malgrem and Winter [20] presented a climate zonation in Puerto Rico based on principal components analysis and SOMs. Finally Liu et al [21], found sea surface temperature patterns in the West Florida Shelf using Growing Hierarchical SOMs. However, to our knowledge this is the first work attempting to find similar productivity zones by means of bio-inspired techniques.

In this paper we present a methodology for classifying similar productivity zones in the sugar cane (Saccharum *officinarum L.*) culture in the southwest region of Colombia. Each productivity zone was represented by a component plane. In each component plane are showed the agro-ecological variables patterns of the cultivated zone. Component planes with similar patters are grouped. For this aim the methodology presented uses a SOM to project the component planes. This SOM is divided in clusters with a technique based on the SOM distance matrix. Every component plane is labeled with its productivity value. Finally, the clusters were classified in high, medium or low pro-

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

ductivity according to the labels of the component planes belonging to the cluster. Analyzing the clusters of similar productivity zones, it was possible to extract new knowledge about the variables more related to the highest, median and lowest productivity.

This paper presents the following structure. In the next section the methodology is explained. The third section focuses on the application of the methodology to the sugar cane case. Finally, in the section four conclusions and future extensions of this work are presented.

2 Methods

2.1 Self Organizing Map

A SOM [7] is formed of artificial neurons situated on a regular low-dimensional grid. This grid can be in one, two or more dimensions, but generally two are used. The neuron in the grid have rectangular or hexagonal form. Each neuron i represents an n-dimensional prototype vector $mi = [mi_1, \ldots, mi_n]$, where n is equal to the dimension of the input space. In the beginning of the training process the prototype vectors are initialized with random values. On each step of the training, a data vector x from the input data is selected and presented to the SOM. The map's unit m_c closest to x is called: the best-matching unit (BMU). The BMU and its neighboring prototype vectors on the grid are moved in the direction of the sample vector:

$$mi = mi + \alpha(t)h_{ci}(t)(x - mi)$$

where $\alpha(t)$ is the learning rate and $h_{ci}(t)$ is a neighborhood kernel centered on the winner unit c. The learning rate and neighborhood kernel radius decrease monotonically with time.

Through the iterative training, SOM organizes the neurons so that neurons that represent similar vectors in the input space are located on the map in contiguous zones, trying to conserve the linear or nonlinear relations of the input space.

2.2 SOM component planes

SOM allows a straightforward visual inspection because the prototype vectors are organized according to their similarity in a low-dimensional grid. This characteristic is helpful when it is needed to handle large multidimensional vectors.

A way to improve this inspection is by means of the component plane representation. A component plane (CP) is a projection of the same component from each vector prototype in a grid. For example, having the prototype vectors $m1, \ldots, mi$. The first component plane will be formed by $CP_1 = [m1_1, \ldots, mi_1]$ in general $CP_n = [m1_n, \ldots, mi_n]$

Hence, the number of component planes will be equal to the input space dimension. In addition, the component planes are visualized in an identical grid to the SOM. However, the difference between the component plane grid and the SOM grid is that on this new grid each neuron does not plot a prototype vector, instead it represents a component of this vector. Each component in the grid takes the same place that the prototype vector from which it comes. Finally, every component on the component plane is visualized by giving to each neuron a color according to the relative value of the respective component in that neuron. As a result, it is possible to obtain the plots of the component planes in order to compare them and look for relationships between variables.

2.3 Correlation Hunting

The component planes analysis can be a tool for discovering relations between variables. Comparing the planes, it is possible to observe similar patterns in identical positions indicating correlation between the respective components. Even, local correlations can be found if two parameter planes resemble each other in some regions. The process of finding these relationships is called correlation hunting. The expression correlation does not include just linear correlations, but also nonlinear and local or partial correlations between variables [15].

The correlation hunting can be realized manually or automatically. However, in many cases the manual analysis is difficult because usually the component planes are not ordered. In addition the comparison becomes more difficult when the number of components increases. In order to overcome this drawback, it is possible to apply reorganization of the component planes such that similar component planes could be located close to each other [16]. To do this, the component planes can be projected on a plane. The projection could be done using, e.g., Sammon's mapping [10], Curvilinear Component Analysis [3] or another SOM. In this paper SOM was used as projection technique.

The projection process using SOM is the following:

(1) Each component plane vector is normalized, in order to ignore different scaling of components and facilite the comparison of the components.

(2) The vectors are further processed by calculating a measure of distance between them.

(3) The measure of distance between component planes i and j can be defined as the value of the correlation of each map position, formally

$$distCP(i,j) = mc * (CP_i, CP_j)$$

where mc is a suitable measure of correlation, in this paper the Pearson correlation coefficient is used.

(4) A covariance matrix is generated with the obtained distances.

(5) The covariance matrix is used as input to a new SOM.

(6) Each component plane grid from the old SOM is projected by means of the new SOM. This projection is realized locating in the place of the BMUs of new SOM, the respective component planes grids from the old SOM. Hence, planes with high correlation are located near each other.

An advantage of using a SOM for component plane projection is that the placements of the component planes can be shown on a regular grid. In addition, an ordered presentation of similar components is automatically generated. A disadvantage is that the choice of grouping variables is left to the user. This task is complicated when the number of component planes is large.

2.4 Distance matrix based clustering of the SOM

Once we have a projection of component planes in a new SOM, it is possible to use a method to cluster prototype vectors in the new SOM in order to find component planes groups. One might use tradicional clustering algorithms. For example, partitive (e.g., k-means) or agglomerative clustering algorithms (e.g., agglomerative hierarchical clustering) are used to cluster the prototype vectors [17]. Nevertheless, those approaches do not take into account the SOM neighborhoods. To cope this drawback, a cluster distance function can be used to take the neighborhoods into account. The U-matrix [13] had been used as an effective cluster distance function [18]. The U-matrix visualizes distances between each map unit and its neighbors, thus it is possible to visualize the SOM cluster structure. This method is usually applied to select clusters from the map by hand. This selection is normally subjective because it is based on the visual perception of each person. Vellido et al. [14] proposed an algorithm to do distance matrix based clustering automatically. In this algorithm, the U-matrix is used to identify cluster centers from the SOM. The rest of the map units are then assigned to the cluster whose center is closest. The algorithm is the following:

(1) Local minima of the distance matrix are found. This is done by finding the set of map units i for which:

$$f(m_i, N_i) \leq f(m_j, N_j), \forall j \in N_i, \qquad (1)$$

where N_i denoted the set of neighboring map units of the map unit i, $f(m_i, N_i)$ is some function of the set of neighborhood distances $\|m_i - m_j\|, j \in N_i$, associated with map unit i. In the experiments, a median distance was used. The set of local minima may have units which are neighbors of each other. Only one minimum from each such group is retained.

(2) For the initialization, let each local minimum be one cluster: $C_i = m_i$. All other map units j are left unassigned.

(3) Calculate distance $d(C_i, m_j)$ from each cluster Ci to (the cluster formed by) each unassigned map unit j.

(4) Find the unassigned map unit with the smallest distance and assign it to the corresponding cluster.

This algorithm provides an automatic discrimination of clusters which permits an easier exploration of similar component planes.

3 Case study: sugar cane culture

3.1 Problem description

SOMs have proved to be effective for the exploratory analysis of agro-ecologic data and became important technique in ecological modeling [8]. SOMs are recommended in cases when it is essential to extract features out of a complex data set [1]. Moreover, the capability to produce easily comprehensible low-dimensional maps improves the visualization and data interpretation [2, 4]. For these reasons, methodologies based on SOM were selected as tools for exploring the data in this case study. The objective of this case study was to classify similar sugar cane productivity zones located in the southwest region of Colombia. Thus, analyzing the obtained groups should enable us to extract new knowledge about the relationship between the agro-ecological variables and productivity. A more detail description of the problem is presented as follows:

A plant is affected by diverse variables (e.g., climate, soil) during its life. These variables have different effects in the plant at different moments of its development (e.g., germination, flowering). Moreover, the combination and/or change of these variables in certain moments determines development states of the plant. This mixture of factors finally determines the crop production. For example, in the sugar cane case, expert knowledge indicates that the most relevant periods are the beginning and the end of plant development. In the first months (after sowing) the vegetative structure is formed (e.g., leafs grows allowing the photosynthesis process), in this moment the water is very important to improve the development of the plant. During the last months (approximately thirteen months after sowing) the plant accumulates the major part of saccharose. As this stage not much water is essential because the plant is totally developed. Accordingly, to determine how and when the variables affect the plant development would be very helpful to support decision making (e.g. in what moment to seed and/or to harvest in order to obtain a better productivity).

In order to find relations between agro-ecological variables and productivity, it is suitable to study similar productivity zones as a practical framework to model and simplify the complexity of agroecosystems. Thus, analyzing the variables that define these groups, it is possible to extract knowledge about the relationship between the agro-ecological variables and productivity.

3.2 Classification of similar productivity zones (PZs)

The database used was provided by a sugar cane research center (CENICAÑA) located in the region under study. The data base contains information collected during six years (1999 to 2005). The agro-ecological variables used for this experiment are listed as follows. Climate variables are Temperature Average (T), Relative Humidity Average (RH), Radiation (Ra), and Precipitation (P). Soil variables are Order (Ord), Texture (Tex) and Depth (Dee). Topographic variables are Landscape (Ls) and Slope (Sl). Other variables are Water Balance (WB) and Variety (V). Finally, productivity (P) of each cultivated zone. As it was mentioned before, the most relevant periods in the sugar cane are the beginning and the end of plant development. Therefore, it is possible to use the climate data of i Months After Sowing (iAS) and i Months Before Harvest (iBH). In this paper $i = 5$ was used.

For each agro-ecological variable a vector was built, each vector has the value of these variable in the cultivated zone, in total 1328 zones were taken. All the variables were scaled [-1,1] in order to allow their comparison in magnitude. As an example, the vector showed below represents the values of temperature for the first month after seed ($T1AS$).

$$T1AS = (PZ_1, PZ_2, \ldots, PZ_{1328})$$

Then, it was created a matrix ($PZmatrix$) with 54 vectors (one for each variable) composed by 1328 components each one.

The $PZmatrix$ was used as input for a SOM with 1600 neurons (40x40), it was trained with the batch algorithm. With this SOM, it was possible to generate 1328 component planes. Hence, the 1328 component planes represent each one a cultivated zone. The 54 variables were ordered in the component planes thanks to the SOM auto-organization feature. Similar variables as climatic variables were placed in contiguous zones, they present low distance between neighbors (see figure 1). Other variables as the soil, topographic variables and varieties of sugar cane are placed adjacent too, but in this case they show high distance between neighbors, (see figure 1).

The magnitude of the variables in the component plane produced different patterns according to the agro-ecological characteristics of the respective cultivated zone. As an example, in figure 2 is possible to observe how the component planes present different patterns of the agro-ecological variables in two productivity cases.

In order to find agro-ecological variables patterns, the component planes obtained were projected by means of new SOM composed of 1600 neurons (40x40), which was trained with the batch algorithm. The distance matrix based clustering of the SOM technique was used. So, 46 clusters of component maps were obtained with the technique aforementioned.

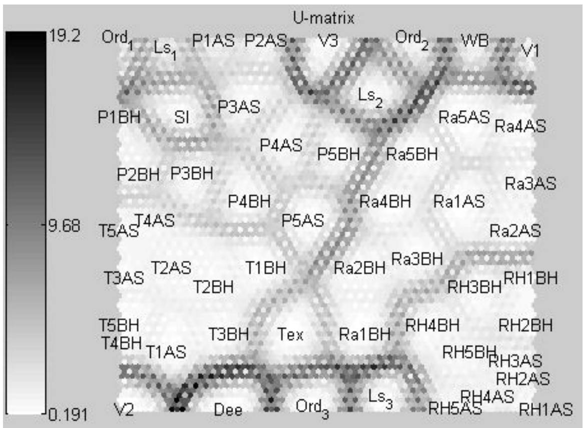

Figure 1: The U-matrix show the distribution of the agro-ecological variables in the component planes

After the clustering process, each component plane was labeled. Productivity was used as a label for the component planes, and thus to compare similar productivity patterns in the clusters with similar agro-ecological variables. Different ranges of productivity were taken in order to discriminate the clusters in the high, median and low productivity. The productivity labels were organized in the range [-10, 10]. Hence, production between -10 and 2 represents low production, between 2 and 5, median productivity, and between 5 and 10, high productivity.

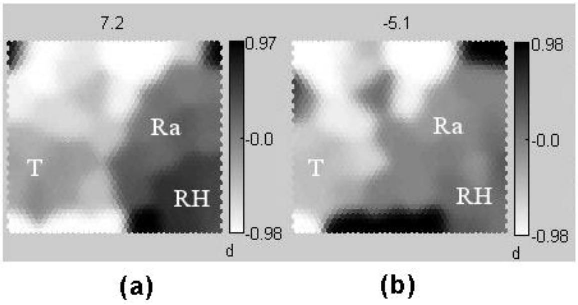

Figure 2: Two component planes selected from high-medium and low productivity clusters. (a) Component plane from high-medium productivity cluster. The patterns of temperature (T), radiation (Ra) and relative humidity (RH) are higher than for the low productivity cluster (b) Component plane corresponding to the low productivity cluster.

Valuable knowledge about the relation between the agro-ecological variables an productivity were found analyzing the clusters, in this paper one of this cases is presented as follows:
In this example we compare the differences between a cluster of high-medium productivity and a cluster with low productivity, in order to discover the difference between patters in zones with different productivity.

For this aim two of the clusters obtained were selected, (see figure 3).

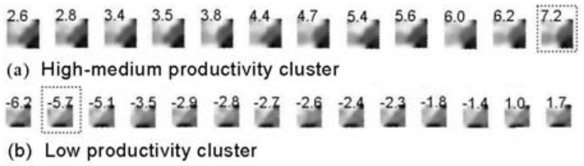

Figure 3: Two selected component planes clusters after apply distance matrix based clustering of the SOM. In dotted lines the planes selected as representation of the clusters for posterior analysis. (a) Component planes with high-medium productivity labels. (b) Component planes with low productivity labels.

In these pair of groups it is possible to see two patterns: low productivity (values between -6.2 and 1.7) and median and high productivity (values between 2.6 and 7.2). With the purpose of to do a detailed analysis, we took the component plane from the cluster 1 with label 7.2, and the component plane with label -5.1 from cluster 2, each one in representation of its group, (see figure 3 in dotted lines and figure 2). To study the patterns in this component planes, it is necessary to know how was organized the variables in the component planes grids. How was aforementioned, in the component planes grids different groups of neurons represent agro-ecological variables. In figure 1 it is possible to observe how they were organized in the U-matrix. This organization of the agro-ecological variables in the U-matrix grid is the same for the component planes. As it was explained in section 2.2, each level of gray in each represents the value of this component in this plane.

Differences between the patterns represented in the two clusters chosen as examples can be observed, (see figure 2).

Three agro-ecological variables were taken as examples to compare the component planes patterns: temperature (T), radiation (Ra) and relative humidity (RH). In a quick inspection of the figure 2, it is possible to observe that the level of gray of T, Ra and RH, are darker in the high productivity cluster than in the low productivity cluster. As an exploratory analysis this quick inspection shows different patterns on the agro-ecological variables for high-medium and low productivity.

To present a more detailed example, radiation (Ra) was analyzed. It is possible to examine the behavior of the radiation for two component planes (previously chosen as example) in a scatter plot. Here, the value of the variables related to the radiation (radiation before harvest and after seed) for each BMU of the component planes were plotted, (see figure 4). In figure 4a it is possible to observe that the two zones present similar values of radiation in the months after seed ($RaAS$), but in the first month after seed ($Ra1AS$) the values are higher for the zones with high-medium production. This behavior shows that the high radiation in the fist month after the seed could be associated with a high-medium production. Because how was aforementioned, in the first months the vegetative structure is formed. But in this analysis is showed that the first month have a more relevant effect than the others months. In addition, in the months before harvest ($RaBH$) the radiation in the high-medium productivity presents the same behavior that the low productivity but with a shift, (see figure 4b). This shift indicate a presence of more radiation in the high-medium productivity zones, although the behavior is similar for the two component planes in both radiations (after seed and before harvest) this shift is no presented in the months after seed. This behavior shows that the high radiation in the months before the harvest is more critical that in the months after seed.

We have shown how to analyze the resulting cluster of SOM component planes. But this paper shows only part of this analysis. Future work will be focus on analyzing other patterns.

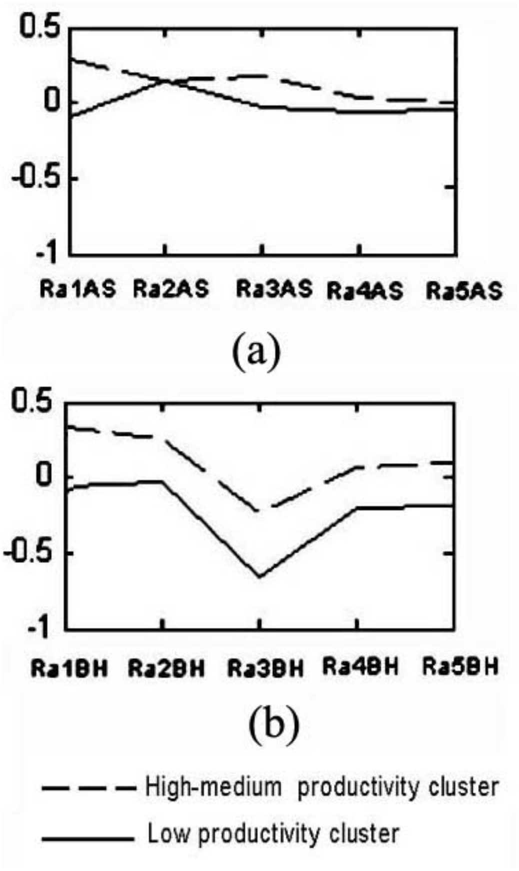

Figure 4: BMUs of the radiation from component planes of high-medium and low productivity. (a) Radiation the five months after seed. (b) Radiation the five months before harvest.

4 Conclusion

This paper presents how to use the clustering of SOM component planes based on the SOM distance matrix methodology applied to agro-ecological modeling. As a case study, this methodology was used in the classification of zones with similar agro-ecological conditions and productivity in the sugar cane culture. By analyzing the obtained groups of agro-ecological variables and cultivated zones, it was possible as an example of the application of the methodology, to find a relationship between the radiation the first month after seed, the months before harvest, and high-medium productivity. More analysis can be made in order to improve the decision support in the sugar cane culture based on the aforementioned methodology. Future work will be focus on the analysis of other patterns.

Acknowledgements

This work is part of a cooperation project between BIOTEC, CIAT, CENICANA (Colombia) and HEIG-VD (Switzerland) named ”Precision agriculture and the construction of field-crop models for tropical fruits”. The economical support is given by several institutions in Colombia (MADR, COLCIENCIAS, ACCI) and the State Secretariat for Education and Research (SER) in Switzerland.

References

[1] Chon, T., Park, Y., Moon, K., Cha, Y. Patternizing communities by using an artificial neural network. Ecological Modelling. Vol. 90 (1996) 69-78.

[2] Chung, H., Hsieh, J., Chang, T. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. Atmospheric Research. Vol. 81 (2006) 124- 139.

[3] Demartines, P., Hrault, J. Curvilinear Component Analysis: a Self-Organizing Neural Network for Nonlinear Mapping on Data Sets. IEEE Transactions on Neural Network. Vol. 8 (1997) 148–154.

[4] Giraudel, J., Lek, S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecological Modelling. Vol 146 (2001) 329-339.

[5] Hargrove, W.W. and Hoffman , F.M., 1999. Using Multivariate Clustering To Characterize Ecoregion Borders Computing in Science and Engineering.:18-25.

[6] Himberg, J. Enhancing the SOM-based Data Visualization by Linking Different Data Projections. Proceedings of 1st International Symposium IDEAL'98, Intelligent Data Engineering and Learning–Perspectives on Financial Engineering and Data Mining (1998) 427–434.

[7] Kohonen, T. Self-Organizing Maps. Springer-Verlag (1997).

[8] Liu, Y., Weisberg, H., He, R. Sea surface temperature patterns on the West Florida Shelf using Growing Hierarchical Self-Organizing Maps. Journal of Atmospheric and Oceanic Technology. Vol. 23 (2006) 325-338.

[9] Mingqin, L. and Samal, A., 2002. fuzzy clustering approach to delineate agroecozones. Ecological modelling, 149:215-228

[10] Sammon, J. A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers. Vol 18 (1969) 401–409.

[11] Schultz, A. and Wieland, R., 1997. The use of neural networks in agroecological modelling. Computers and Electronics in Agriculture, 18:73-90.

[12] Tryba, V., Goser, K. Self-Organizing Feature Maps for Process control in Chemistry. Proc. ICANN, Helsinki (1991) 847–852.

[13] Ultsch, A., Siemon, P. Kohonen's self organizing feature maps for exploratory data analysis. In Proc. INNC'90, Int. Neural Network Conf (1990) 305–308.

[14] Vellido, A., Lisboa, P., Meehan, K. Segmentation of the on-line shopping market using neural networks. Expert Systems with Applications. Vol. 17 (1999) 303–314.

[15] Vesanto, J., Ahola, J. Hunting for Correlations in Data Using the Self-Organizing Map. Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (1999) 279–285.

[16] Vesanto, J. SOM-based data visualization methods. Intelligent Data Analysis. Vol. 3 (1999) 111–26.

[17] Vesanto, J., Alhoniemi, E. Clustering of the self-organizing map. IEEE Transactions on Neural Networks. Vol. 11 (2000) 586–600.

[18] Vesanto, J., Sulkava, M. Distance Matrix Based Clustering of the Self-Organizing Map. ICANN '02: Proceedings of the International Conference on Artificial Neural Networks (2002) 951–956.

[19] Waltman, W.J., Mortensen, D.A., Cassman, K.G., Nelson, L.A., Specht, J.E., Sinclair, H.R., Waltman, S.W., Narumalani, S. and Merchant, J.W., 1999. Agroecozones of Nebraska. In: ASA-CSSA-SSSA (Editor), Annual Meetings-Agronomy Abstracts, Salt Lake City,UT.

[20] Malmgren B. A. and Winter A., 1999. Climate zonation in Puerto Rico based on principal components analysis and an artificial neural network. Journal of climate, 12:977-985

[21] Liu, Y., Weisberg, R.H. and He, R., 2006. Sea surface temperature patterns on the West Florida Shelf using Growing Hierarchical Self-Organizing Maps. Journal of Atmospheric and Oceanic Technology, 23:325-338.