International Journal of ChemTech Research
CODEN (USA): IJCRGG ISSN: 0974-4290
Vol.8, No.9, pp 12-19, 2015

Assessing the performance and variability of some Sugar beet varieties using Self-organizing map artificial neural network and Cluster analysis

O. M. Ibrahim[1]*, A.A. Gaafar[2], Asal M. Wali[1] and M.M. Tawfik[3]

[1]Plant Production Department, Arid Lands Cultivation Research Institute, City of Scientific Research and Technological Applications (SRTA-City), New Borg El-Arab, Alexandria 21934, Egypt
[2]Soil Salinity and Alkalinity Research Department, Soil, Water, and Environment Research Institute, Agricultural Research Center
[3]Field Crops Res. Dept., National Research Centre, 33 El Bohouth st., Dokki, Giza, Egypt – P.O.12622

Abstract: The current study aims at evaluate the quality of eleven sugar beet varieties based on five quality parameters (Sucrose as %, K^+, Na^+, Amino-N as mmol/100 gm fresh weight of roots, and sugar yield as kg/ton fresh weight of roots) using self organizing map (SOM) and cluster analysis. The data were obtained from Delta sugar factory, Kafr Alsheikh Governorate after a survey during the seasons of 2012/2013 and 2013/2014 from 15 villages. Distance matrix based on Euclidian coefficient for the 11 sugar beet varieties revealed that dissimilarity ranged from 0.60 between Carola and Raspoly to 7.57 between Atospoly and Top, which reveal the quality diversity among varieties. Both cluster analysis and SOM classified the varieties into three clusters which accounting for 70% of the variation. The clusters in SOM consist of nodes where varieties in the same node are more similar than varieties in different nodes in the same cluster. However, varieties in the same cluster are more similar than varieties in different clusters. The SOM revealed that both Baraka and Top had the highest quality and produced the highest sugar yield because they had the lowest impurities (K^+, Na^+, and Amino-N) even they have less sucrose than Gloria. The results suggested that using self organizing map is helpful to classify sugar beet varieties clearly and more interpretable than cluster analysis and can be used as a tool of classification.
Keywords: Sugar beet varieties, Self-organizing map and Cluster analysis.

Introduction

Artificial neural networks are considering a powerful mathematical modeling technique in the agricultural sciences[1,2,3]. A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional representation of the input space of the training samples, called a map. A main goal in the field of assessment of varieties performance is to extract useful information out of large and usually high-dimensional data sets. In the past, clustering large datasets has been a domain of classical statistical methods. Recently a new approach, Self Organizing Map (SOM), [4] has been proposed in order to classify high dimensional datasets. When it is compared with the other clustering algorithms, SOM is the one that has the greatest visualization capability. In addition, the detailed information

can be determined by using the SOM's outputs due to the easiness of interpretation of the visualized outputs. On the other hand, traditional cluster analysis has very limited visualization property. Clustering methods may involve a variety of algorithms but almost invariably build distinct self contained clusters[5] whereas the neurons of the SOM are not mutually exclusive. This means that the final feature map, instead of showing several distinct clusters with differing characteristics, shows neighboring nodes which have many similar characteristics but differ perhaps on one or two, or in degree of intensity of characteristics[5], therefore if overlapping exists between the clusters, it can be determined through outputs of SOM. The traditional statistical methods can not be sufficient for analyzing the data containing many data cases and large number of variables which describe these data cases, however SOM method is considered as an effective method in dealing with high dimensional data. The traditional cluster analysis methods are designed under strict assumptions of certain statistical distribution functions; however there is no need for making that kind of assumptions in application of SOM. For instance, continuous variables should satisfy normal distribution assumption and categorical variables should satisfy multinomial distribution assumption in order to perform two-step cluster analysis effectively[6]. Furthermore, the number of the clusters should be known at the inception of the K-means clustering method. However, the number of the clusters is not a pre-request at the inception stage of SOM, and the correct number of clusters will be directly shown by the result itself. As mentioned before, the sorting ability of the traditional cluster analysis is an important problem for the reliability of the solutions. Whereas, the SOM can be a remedy for that problem, because The U-matrix does not give any results when there are no obvious clustering relations in the original space, thus, unreasonable arbitrary classification can be avoided [7].

Materials and Methods

During the winter seasons of 2012/2013 and 2013/2014, a field survey was carried out of 11 sugar beet varieties for five quality parameters (sucrose as %, K, Na, Amino-N as mmol/100 gm fresh weight of roots, and sugar yield as kg/ton fresh weight of roots) which measured in Delta sugar factory to describe characteristics of a set of random sugar beet varieties in a total of 15 sampling sites (villages) and a total of 160 samples. Five quality parameters at each sampling site were used as independent variables. They were rescaled within the minimum and maximum range (0–1) and standardized before being provided to the Self Organizing Map (SOM) model and cluster analysis, respectively as inputs.

In this study hierarchical cluster analysis begins by separating each variety into a cluster by itself. At each stage of the analysis, the distance by which varieties are separated is relaxed in order to link the two most similar clusters until all of the varieties are joined in a complete classification tree. The cluster analysis was performed using the Ward method with Euclidean distance coefficient to evaluate dissimilarity among all the surveyed varieties. Before performing the analysis, the data were first standardized by subtracting the mean from each value, and then divided on the standard deviation[8],[9,10].

Self Organizing Map (SOM) is a realistic model of the biological brain function [4]. SOM consists of input and output layers which were composed of neurons serving as the computational units in the network. Input and output layers connected with weight vectors. When the input vector (five quality parameters), *xs*, was given to the network, the distance between the weight vector, w_i, and the input vector, *xs*, was calculated by Euclidean distance, xsw_i. The five quality parameters were given to the SOM model. The output layer consists of seven neurons in a two-dimensional hexagonal lattice connected via weights with input vectors (five quality parameters). Vectors that are close in input space will be mapped to units that are close in the output map [11,12]. Learning of SOM is iteratively and can be conducted with a subset or all data vectors. Prior to learning, the weights of map units (*Wi*) are initialized with random values. During the learning phase each input vector (*xs)* is presented to the network, and Euclidean distances between *xsi* and all vector units or nodes in the network are computed. The node (*q*) with the shortest Euclidean distance commonly known as Best Matching Unit (BMU) is selected as a winner (Fig.1).

$$q_{ji} = \min\left((\Sigma\,(x_{si} - w_{ji})\,2)^{1/2}\right)$$

Where q is winning neuron, *xsi* and *wji* are the *i*th element of the input vector Xs and the *i*th weight of neuron j, respectively.

This winning neuron becomes the centre of an *update neighborhood.* Update neighborhood is an area within which nodes and their associated weights according to Kohonen rule will be updated, such that each

weight vector converges to the input pattern. In this way, the nodes in a self-organizing map compete to best represent the particular input sample.

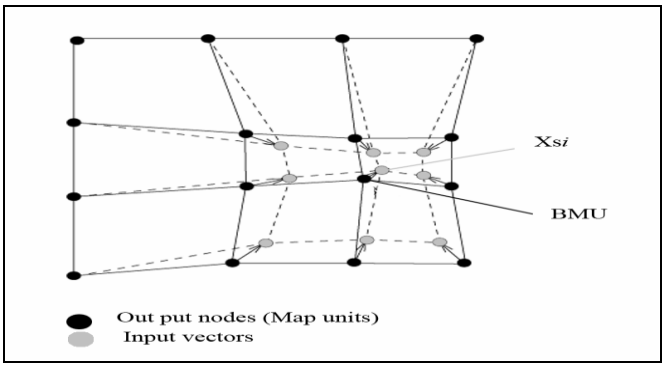

Fig.1. Updating the best matching unit (BMU) and its neighbors toward sample input vector x_{si}. The solid and dashed lines correspond to the situation before and after updating, respectively (modified from[13].

This process is repeated for every input sample as they are passed sequentially to the SOM. During this iterative process, the rate at which the winning nodes converge to the input samples is termed the learning rate *(α^s)*. Throughout learning, the learning rate and size of the update neighborhood (the update radius) decrease, so that the initial generalized patterns are progressively refined. After the learning phase, the SOM consists of a number of vectors, with similar vectors nearby and dissimilar vectors further apart [14,15]. The five quality parameters have been used as input for SOM.

Without normalization, the variable with the largest range will dominate the map organization. All input quality parameters are normalized to the range of 0-1 using a logistic function. Before learning, weights of the map units were initialized randomly. The quality of the results is measured with an average quantization error and a topographic error. Average quantization error is the Euclidian distance between data vectors and best matching unit (BMU) on the map. Topographic error shows the proportion of all data vectors for which first and second BMUs are not adjacent units and is an index for accuracy of the map in preserving topology [4].

Results and Discussions

The five studied quality parameters were used to construct a dissimilarity matrix using the Euclidian coefficient, Table (1) and used to generate dendrogram (tree diagram) showing dissimilarity among all the varieties, figure 2.

Table 1: Distance matrix based on Euclidian dissimilarity coefficient for the 11 sugar beet varieties.

	Gloria										
Gloria	0.00	Toro									
Toro	3.55	0.00	Atospoly								
Atospoly	4.91	4.56	0.00	Carola							
Carola	1.95	2.89	5.52	0.00	Top						
Top	3.76	4.25	7.57	2.33	0.00	Baraka					
Baraka	3.39	4.05	7.48	2.04	0.98	0.00	Cawamera				
Cawamera	4.03	1.75	5.10	2.67	3.95	3.80	0.00	Desrespoly			
Desrespoly	4.86	4.26	2.35	5.62	7.26	7.27	5.43	0.00	Sultan		
Sultan	4.26	1.53	4.54	3.20	4.35	4.42	1.39	4.50	0.00	Farida	
Farida	2.67	1.92	4.67	1.65	3.03	3.11	2.06	4.49	1.85	0.00	Raspoly
Raspoly	2.44	2.74	5.82	0.60	1.94	1.73	2.45	5.81	2.98	1.54	0.00

In Table (1) the distance matrix reveals that dissimilarity ranged from 0.60 between Carola and Raspoly to 7.57 between Atospoly and Top, which reveal the quality diversity among varieties.

Figure 2 displays the tree diagram. The figure provides a graphical view of the clusters. As the number of branches grows to the left from the root, the R^2 approaches 1; the first three clusters (branches of the tree) account for 70% of the variations among all the varieties, In other words, only three clusters are necessary to explain over two-thirds of the variations.

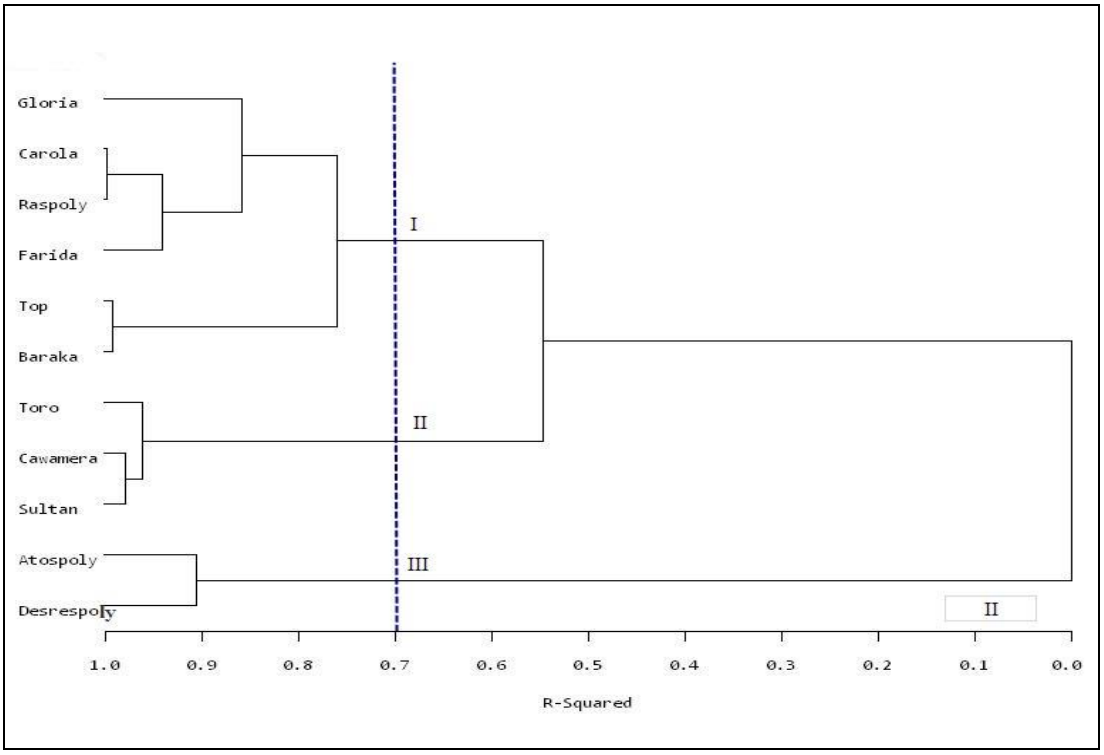

Fig. 2: Dendrogram showing cluster analysis (Ward method) of 11 sugar beet varieties.

Table 2: Quality parameters mean values of sugar beet varieties groups issued from cluster analysis

Clusters	Sucrose, %	K^+	Na^+	Amino-N	Impurities	Quality, %	sugar yield, Kg/tons roots
I	18.89	5.89	2.56	2.67	3.44	81.81	154.54
II	17.62	6.99	4.49	4.18	4.62	73.78	130.01
III	17.19	5.71	2.53	3.08	3.40	80.20	137.88

Based on the cluster analysis in Fig.1, the 11 varieties divided into 3 clusters based on the five studied quality parameters as shown in Table 2 which reveal that the first cluster of varieties (Gloria, Farida, Raspoly, Carola, Top, and Baraka) was the highest in both sucrose (18.89 %) and sugar yield (154.54 kg/ton fresh weight of roots), and the low in impurities (3.44 %). On the other hand, the second cluster (Cawamera, Toro, and Sultan) was the lowest in sugar yield (130.01), this may be due to the highest content of impurities (4.62), yet it has higher sucrose content (17.99) than the third cluster (17.19). The results from table 2 revealed that the first cluster was the best cluster in the quality parameters; however, the second cluster was lowest one. Every cluster can be represented by any variety belonging to that cluster; this will be useful in reducing the number of varieties being tested in the next assessment. Also, the data illustrated that hybridization between distant clusters will resulted in higher genetic variability than within the same cluster. Cluster analysis was approved as a suitable method for data classifying and suggested by [16].

The SOM classified the varieties into three clusters (Fig. 3). The clusters consist of nodes where varieties in the same node are more similar than varieties in different nodes in the same cluster [17]. However, varieties in the same cluster are more similar than varieties in different clusters. In addition, when the line between two nodes is jagged the difference between these two nodes is less than when the line is solid. The SOM revealed that varieties Baraka and Top had the highest quality and produced the highest sugar yield (the node with red color in Fig.4) followed by Gloria, this may be due to they had the lowest impurities (K, Na, and Amino-N), Fig. 5, 6, 7, and 8 even they have less sucrose than Gloria.

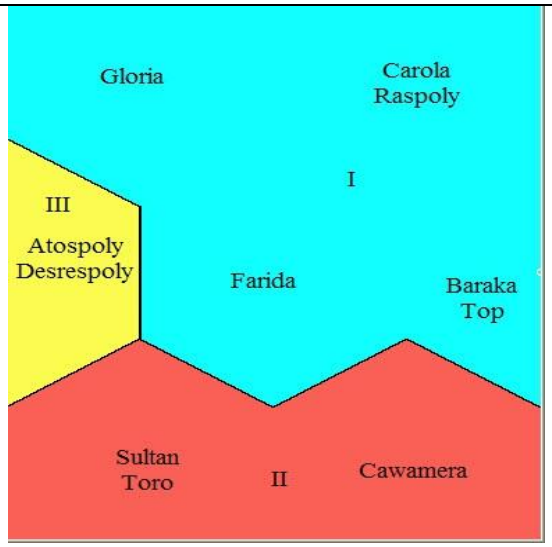

Fig. 3. Clustering of the trained SOM units. The U-matrix and Ward's method were applied to set boundaries on the SOM map. The Latin numbers (I-III) display clusters and the names in each unit of the map represent the varieties.

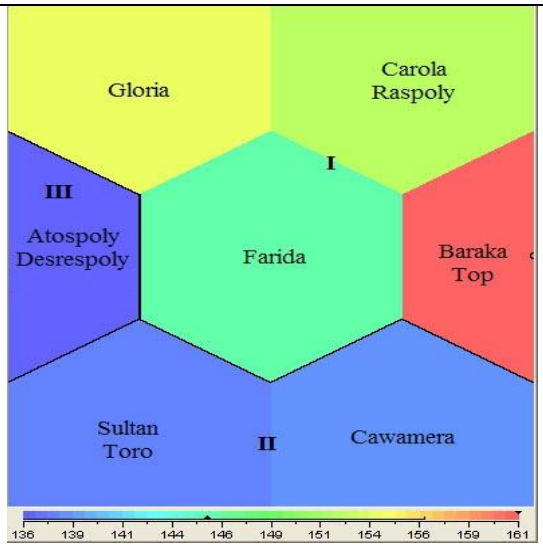

Fig. 4.Visualization of varieties in the trained SOM in color scale according to sugar yield, kg/ton fresh weight of roots. Red color represents high value, whereas blue is low value.

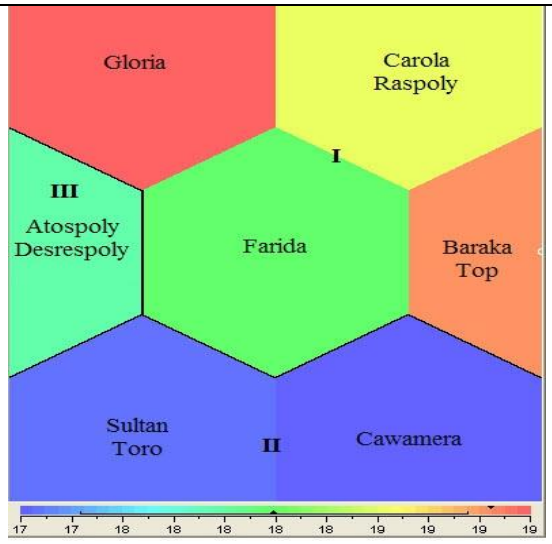

Fig.5. Visualization of sugar beet varieties in the trained SOM in color scale according to Sucrose. Red color represents high values, whereas blue is low values.

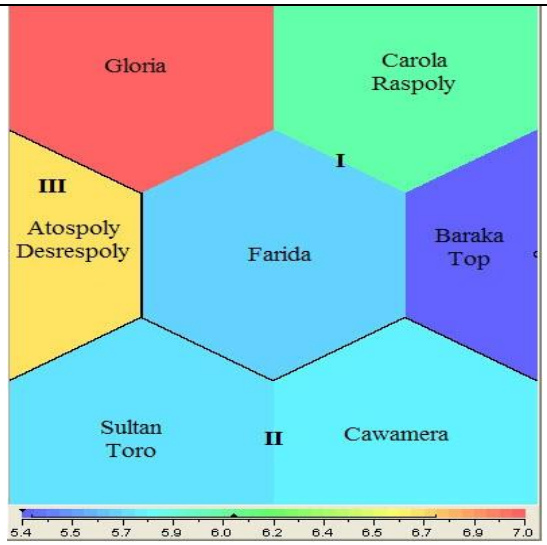

Fig.6. Visualization of sugar beet varieties in the trained SOM in color scale according to K^+. Red color represents high values, whereas blue is low values.

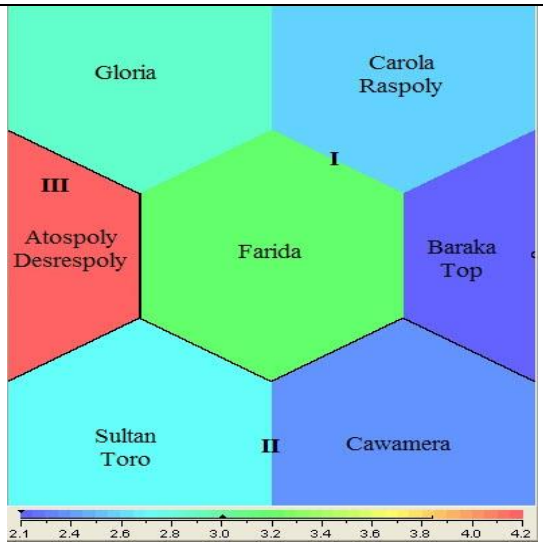

Fig.7. Visualization of sugar beet varieties in the trained SOM in color scale according to Na^+. Red color represents high values, whereas blue is low values.

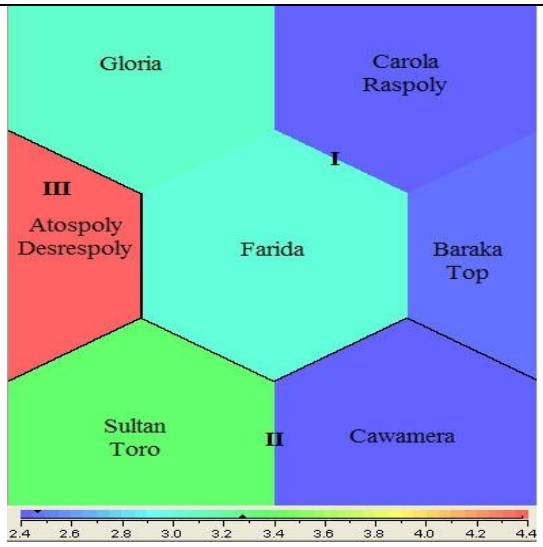

Fig.8. Visualization of sugar beet varieties in the trained SOM in color scale according to Amino-N. Red color represents high values, whereas blue is low values.

However, the varieties Atospoly and Desrespoly had the lowest quality and produced the lowest sugar yield (the node with blue color in Fig.4). The SOM showed high performance in analyzing the relationships among quality variables where varieties with high sugar yield were marked with low impurities (K, Na, and Amino-N) and high sucrose content; this reflects the negative relationship between sugar yield and impurities and the positive relationship between sugar yield and sucrose content. This information can't be quickly extracted from hierarchical cluster analysis neither from the distance matrix nor from the dendrogram until we manually calculate the average of each group for all the quality parameters, however, these information are illustrated visually on the maps of SOM.

The results suggested that using self organizing map is helpful to classify varieties clearly and more interpretable than hierarchical cluster analysis.

Conclusion

Artificial neural network is a biologically inspired computing model formed from several neurons connected with connection weights which constitute the network structure [18].

Two methods of classification, hierarchical cluster analysis and self organizing map (SOM) were utilized to classify 11 sugar beet varieties. Five variables (sucrose as %, K, Na, Amino-N as mmol/100 gm fresh weight of roots, and sugar yield as kg/ton fresh weight of roots) were used in determining the quality of varieties through analysis of the data obtained from Delta sugar factory, Kafr Alsheikh Governorate after a survey of 15 villages.

The SOM showed a high performance for visualization and abstraction of quality data. The trained SOM efficiently classified varieties according to gradients of input quality variables, and displayed a distribution of each component (input quality variables).

Also, the SOM showed high performance in analyzing the relationships among quality variables, and consequently could be used as a tool to extract relationships between quality variables.

The biologically inspired machine learning techniques could be an alternative tool to traditional statistical analysis in fields such as surveying and crop science.

References

1. Ibrahim, O.M. (2012). Simulation of Barley grain yield using artificial neural networks and multiple linear regression models. Egypt. J. Appl. Sci., vol. 27, No.1, p. 1-11.
2. Ibrahim, O.M. (2013). A comparison of methods for assessing the relative importance of input variables in artificial neural networks. Journal of Applied Sciences Research, 9(11): 5692-5700.
3. Ibrahim, O.M. (2015). Evaluating the effect of salinity on corn grain yield using multilayer perceptron neural network. Global journal of advanced research. 2(2):400-411.
4. Kohonen, T. (2001). *Self Organizing Maps*, Springer, New York, USA.
5. Curry, B., F. Davies, P. Evans, and L. Moutinho (2001). "The Kohonen self-organizing map: An application to the study of strategic groups in the UK hotel industry." *Expert Systems*, 18(1):19-31.
6. Norusis, M. (2004). *SPSS 13.0 Statistical Procedures Companion*, Prentice Hall, Upper Saddle-River, N.J.
7. Zhang, X., and Y. Li (1993). "Self-organizing map as a new method for clustering and data analysis." In: *Proceedings of 1993 International Joint Conference on Neural Networks*, IJCNN, Nagoya, 2448-2451.
8. El kramany, M.F; O.M. Ibrahim, El Habbasha, S.F. and N.I. Ashour (2009). Screening of 40 Triticale (*X Triticosecale wittmack*) Genotypes under Sandy Soil Conditions. Journal of applied Sciences Research, 5(1): 33-39.
9. Ibrahim, O.M., Magda H. Mohamed, M.M. Tawfik, Elham A. Badr (2011). Genetic diversity assessment of Barley (*Hordeum vulgare* L.) genotypes using cluster analysis. International Journal *of* Academic Research. 3(2): 81-85.
10. Bakry, A.B., O.M. Ibrahim, T.A. Elewa, and M.F. El-Karamany (2014) Performance Assessment of Some Flax (*Linum usitatissimum* L.) Varieties Using Cluster Analysis under Sandy Soil Conditions. *Agricultural Sciences*, 5, 677-686.
11. Vesanto, J., and E. Alhoniemi (2000). Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks, 11(3): 14.
12. Bação, F., V. Lobo, and M. Painho (2005). The self organizing map, the Geo-SOM, and relevant variants for geosciences. Computers & Geosciences, 31(2): 155-163.
13. Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas (1996). Som toolbox team Helsinki University of Technology,Laboratory of Computer and Information Science, Finland.
14. Richardson, A. J., and C. Risien (2003). Using self-organizing maps to identify patterns in satellite imagery. Progress In Oceanography, 59(2): 223-239.
15. Sueli, A. M., and J. O. Lima (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174(3) 1742-1759.
16. Mohammedi S.A., and B.M. Prasanna (2003). Analysis of genetic diversity in crop plants – Salient statistical tools and considerations. Crop Sci., 43: 1235–1248.
17. Ibrahim, O.M., A.T. Thalooth and Elham A. Badr (2013). Application of Self Organizing Map (SOM) to Classify Treatments of the First Order Interaction: A comparison to Analysis of Variance.World Applied Sciences Journal. 25 (10): 1464-1468.
18. Ibrahim, O.M., Bakry, A. B., Asal, M. Waly and Elham, A. Badr (2014). Modeling grain and straw yields of wheat using back propagation and genetic algorithm. Asian Academic research Journal of Multidisciplinary. 1(28):153-162.

*****