https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

Influence over the Dimensionality Reduction and Clustering for Air Quality Measurements using PCA and SOM

Navya H.N

Bangalore, 560072, India

Abstract—*The current trend in the industry is to analyze large data sets and apply data mining, machine learning techniques to identify a pattern. But the challenges with huge data sets are the high dimensions associated with it. Sometimes in data analytics applications, large amounts of data produce worse performance. Also, most of the data mining algorithms are implemented column wise and too many columns restrict the performance and make it slower. Therefore, dimensionality reduction is an important step in data analysis. Dimensionality reduction is a technique that converts high dimensional data into much lower dimension, such that maximum variance is explained within the first few dimensions.*

This paper focuses on multivariate statistical and artificial neural networks techniques for data reduction. Each method has a different rationale to preserve the relationship between input parameters during analysis. Principal Component Analysis which is a multivariate technique and Self Organising Map a neural network technique is presented in this paper. Also, a hierarchical clustering approach has been applied to the reduced data set. A case study of Air quality measurement has been considered to evaluate the performance of the proposed techniques.

Keywords *— **Air Quality Dimensionality reduction, Hierarchical Clustering,Principal Component Analysis, Self Organising Maps.***

I. INTRODUCTION

Multivariate Statistical Analysis is useful in the case where data is of high dimension. Since human vision is limited to 3 dimensions, all application above 2 or 3 dimension is an ideal case for data to be analyzed through Multivariate Statistical Analysis (MVA). This analysis provides joint analysis and easy visualization of the relationship between involved variables. As a result, knowledge that was unrevealed and hidden among vast amounts of data can be obtained. PCA is a powerful multivariate technique in reducing the dimension of data set and revealing the hidden relationship among the different variables without losing much of information [1].

Artificial Neural Network (ANN) is an information processing computational model based on biological neural networks and is composed of several interconnected processing elements (neurons) that work in parallel to solve a generic problem. In general ANN are used to identify complex relations between input and output or to find patterns in a given data set. ANN is an adaptive system that can change its structure based on the flow of information through the network during the training phase.

The two important learning paradigms in ANNs are supervised learning and unsupervised learning. In supervised learning, there exists a training data that helps in the construction of the model by specifying classes and by providing positive and negative objects that belong to those classes. In unsupervised learning, there is no preexisting taxonomy and the algorithm classifies the output into different classes. Kohonen's Self Organising Map is an unsupervised learning technique that reduces a high dimensional data into a 2 dimensional space. This reduction in dimensionality helps in understanding the relationship quickly and also SOM provides better visualization of components [2].

The rest of the paper is organized as follows. Section 2 discusses related work and their findings. Section 3 briefly explains the existing Principal Component Analysis and Self Organising Map techniques used in our work. Section 4 consists of a case study on air quality to evaluate the performance of the proposed techniques, along with expected results. Finally the concluding points are given in section 5.

II. RELATED WORK

Some of the related work pertaining to Principal Component Analysis and Self Organising Maps are discussed in this section. In [3] multivariate analysis used to analyze and interpret data from a large chemical process. PCA was used to identify correct correlations between the variables to reduce the dimensionality of

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

process data. In [4] Multivariate Statistical Analysis is applied for Dermatological Disease Diagnosis. There are few diseases like psoriasis, seborrhoeic dermatitis, lichen planus, pityriasis, chronic dermatitis and pityriasis rubra piaris in dermatology that share the same clinical features. PCA is applied to identify the relationship between 12 clinical attributes and the circle of correlations depicts patterns of variable associations. Monitoring abnormal changes in the concentration of ozone in the troposphere is of great interest because of its negative influence on human health, vegetation and materials. Modeling ozone is very difficult because formation mechanisms in troposphere are very complex and adding to it is the uncertainty regarding the meteorological condition in urban areas. PCA is used as a data detection method for highly correlated variables in [5]. PCA was applied for the yeast sporulation data for simplification of analysis and visualization of gene expression data. Data was collected for 7 different gene expressions over time and it was observed that much of variability was explained within first two principal components in [6].

In [7] both principal component analysis and self-organizing has been applied to classify and visualize fire risks in forest regions. On application of PCA first two principal component has explained most of the variance but SOM was successful in effective visualization and clustering of nodes to depict fire risks. In [8] PCA, cluster analysis and SOM was applied to a large environmental data set to assess the quality of river water. The results indicated the power of classification of SOM when compared to other traditional methods. In [9] Forest Inventory (FI) contains useful information on forest conditions. To interpret such large data sets, SOM technique is applied that helps in fast and easy visualization, analysis of multidimensional data sets. In [10] PCA and SOM was applied in cellular manufacturing system for visual clustering of machine-part cell formation. PCA was used for reducing the dimensionality of data set and was projected on to 2 dimensional space. The unsupervised SOM technique was applied for data visualization and also to solve the problem of cell formation. In [11] SOM was applied to complex geospatial datasets for knowledge discovery and information visualization. The dataset consisted of socio economic indicators mapped to municipalities in Netherlands. By the application of SOM, structure and patterns of dataset was uncovered and graphical representation helped in discovery of knowledge and better understanding.

III. EXISTING TECHNIQUES

Multivariate Statistical Analysis is useful in the case where data is of high dimension. Since human vision is limited to 3 dimensions, all application above 2 or 3 dimensions is an ideal case for data to be analyzed through Multivariate Statistical Analysis (MVA). This analysis provides joint analysis and easy visualization of the relationship between involved variables. As a result, knowledge that was unrevealed and hidden among vast amounts of data can be obtained.

A. Principal Component Analysis (PCA)

PCA is a powerful multivariate technique in reducing the dimension of data set and revealing the hidden relationship among the different variables without losing much of information [12]. The steps involved are:

1. Calculate the covariance matrix

Covariance is the measure of how one variable varies with respect to other variable. If the variables are more than two, then the covariance matrix needs to be calculated. The covariance matrix can be obtained by calculating the covariance values for different dimensions. The formula for obtaining the covariance matrix is

$$C^{n\,X\,n} = (c_{i\,,j}\,, c_{i\,,j} = cov(Dim_i\,, Dim_j)\,)$$

where $C^{n\,x\,n}$ is a matrix with n rows and n columns

i, j is the row and column indices and covariance is calculated using below formula

$$cov(X,Y) = \sum_{i=1}^{n} [(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)$$

where X_i is the value of X at i^{th} position

Y_i is the value of Y at i^{th} position

$\bar{X}$, $\bar{Y}$ indicates the mean values of X and Y respectively.

2. Find Eigen Vectors for covariance matrix

Eigen vectors can be calculated only for square matrix and not all square matrix has eigen vectors. In case n x n matrix contains eigen vectors, then total number of eigen vectors is equal to n. Another important property is that the eigen vectors are always perpendicular to each other. Eigen values are associated with Eigen vector and describe the strength of the transformation.

3. Sort Eigen vectors in decreasing order of Eigen values

The highest eigen value corresponds to be the principal component which explains the maximum variance among the data points with minimum error. The first few principal components explain most of the variance and eigen values with lesser values can be neglected with very little loss of information.

4. Derive the new data set

The original data will be retained back but will be in terms of the selected eigen vectors without loss of data.

B. Self Organising Map (SOM)

Self Organising Map is a popular technique in Artificial Neural Network (ANN) under the category of unsupervised learning. In conventional ANN approach,

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

the input vector is presented to a multilayer feed forward network and the generated output is compared with the target vector. When there exists a difference, the weights are altered to minimize the error in output. This phase is repeated many times with different sets, until the desired output is achieved. However, SOM does not require a target vector during the training phase. A SOM without any external supervision, learns to classify the training data.

The basic principle of SOM is when the weight of nodes matches the input vector, then that portion of lattice is selectively optimized to resemble the input vector. Each node receives every element of input or training data in vector format one at a time. A calculation is done between the element and weight of the node to determine the fit between them. The calculation performed is to determine the distance between the two and usually it is Euclidian distance or any other distance measure can be used. A winning node that best describes the training element can be obtained and it has the smallest distance between the input element and node's weight. The neighbors of winning node should be identified. Then, these neighbors and winning node is updated to represent the new training element. By this procedure, the map learns through individual elements [13].

SOM Algorithm

1. Initialize the weights for each node either randomly or by pre-computed values.
2. For every input element
   a) Get the input element and convert to a vector
   b) For every node in the map
      i) Compare the input vector with the node
   c) Every node weight is examined to see which one matches and is closest to the input vector. The node with the smallest distance is declared as the winning unit and is commonly referred as Best Matching Unit (BMU).
      Distance between the between input vector and node's weight is calculated using Euclidean Distance formula.

$$Dist = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2}$$

where V is the current input vector
W is the node's weight vector

   d) The radius of neighborhood of the BMU is calculated. All nodes within this radius range will be updated in the next iterations. Initially, the radius will be set to the radius of lattice and it decreases in each step.
   e) The weights of each neighboring nodes are adjusted according to the below equation

$$x(t+1) = x(t) + N(x,t)\,\alpha\,(t)\,(\partial(t) - x(t))$$

where x(t+1) is the next value of weight vector
x(t) is current value of weight vector
N(x, t) is the neighbourhood function, which decreases as a function of time
α(t) is the learning rate, which decreases as a function of time
∂(t) is the vector representing the input document

IV. CASE STUDY

A case study is considered to demonstrate the concept of dimensionality reduction on a data set using Principal Component Analysis and Self Organising Map. Later, hierarchical clustering is applied to the obtained SOM results. The case study considered determines the quality of air in regions of New York city. The data set consists of variables like Ozone, Solar radiation, Wind, Temperature for different regions. Ozone variable consists of numeric values in parts per billion, Solar radiation depicts the radiation in Langleys with a frequency band 4000-7700 Angstroms, Wind variable has numeric values in miles per hour, Temperature indicates in degrees the maximum daily temperature.

A. *PCA Application*

PCA reduces the dimension and explores the hidden information. The first few principal components explain most of the variance. Table 1 indicates the summary of principal component analysis which explains the standard deviation between the principal components, proportion of variance and cumulative proportion.

Table.1: indicates the standard deviation, variance and cumulative proportion associated with principal components

	PC1	**PC2**	**PC3**	**PC4**
Standard Deviation	1.5343	0.9529	0.6890	0.51441
Proportion of Variance	0.5886	0.2266	0.1187	0.06615
Cummulative Proportion	0.5886	0.8152	0.9338	1.00000

The important task in PCA is deciding the number of principal components to be considered for further analysis. Fig. 1. Indicates a scree plot for determining the number of PCs. The elbow point shows the number of components to be considered and beyond this the eigen values are small and indicates negligible data loss.

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

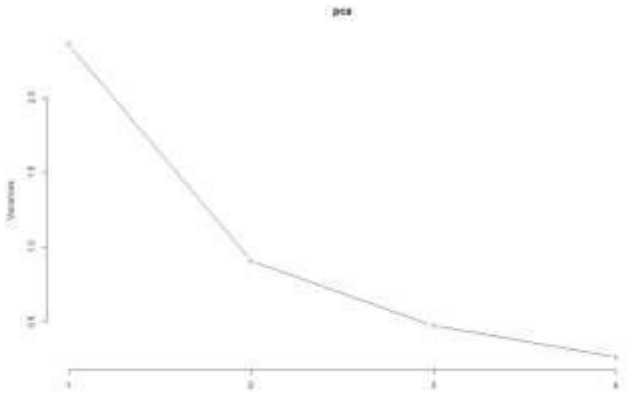


Fig. 1. indicates a scree plot for determining the number of PCs.

The relationship between variables can be identified using biplot. Fig. 2. represents variables and observation of multi-dimensional data. Fig.2 indicates the variables Temperature and Ozone are closely related to each other and Wind is negatively correlated with other variables. The angle between vectors represent an approximation of covariance. A small angle between vectors indicates the variables are highly correlated, an angle of 90 degrees indicates variables are not correlated and 180 degrees represents that the variables are negatively correlated.

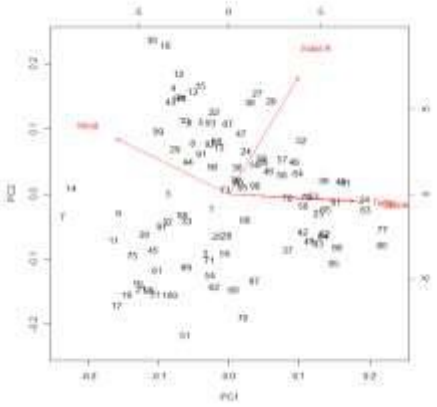


Fig. 2. Represents variables and observation of multi-dimensional data.

B. SOM Application

Before the application of SOM, the data was prepared for analysis by identifying the duplicates, removal of outliers, data conversion, removal of null values and the variables are scaled to provide equal importance during training phase. A SOM grid is created with a size of 10 x 10 and there is no explicit rule for selecting the number of nodes, except that it should allow easy visualization of the SOM. The SOM is trained with a learning rate of 0.05 that gradually declines to 0.01. The radius of neighbourhood can either be a vector or a single number. If it is a single number, the radius will vary from current value to the negative value of the current number. Once, the neighbourhood becomes small and radius is less than 1, only the winning node will be updated.

Fig.3. shows a plot of SOM during training phase after 100 iterations. In this experiment, 100 times the complete data set is presented to the network. The distance between the weight of the node and input sample reduces as the training iterations progress. This distance should ideally reach a minimum value and Fig. 3. shows training progress over time.

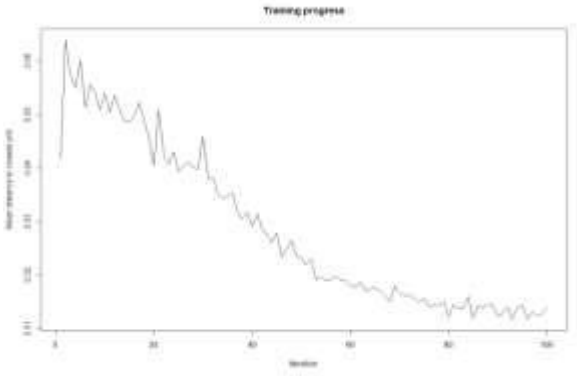


Fig.3. shows a plot of SOM during training phase after 100 iterations.

Unified Distance matrix (U –matrix) is a type of SOM visual representation in a map of 2 dimensional grid as shown in fig. 4. U-matrix indicates the distance between adjacent nodes and it is represented by different grey shades. A dark color indicates a larger distance between the neighboring nodes and represents a gap in input values.

A lighter shade indicates the vectors are close to each other and they indicate cluster themselves, high values represents borders among the clusters.

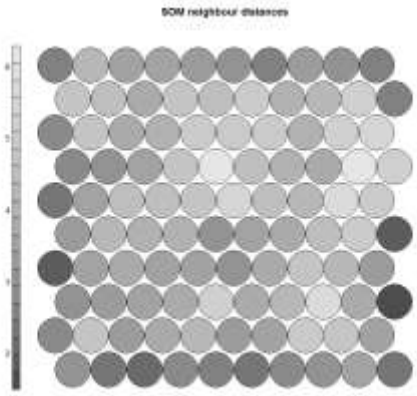


Fig. 4. Indicates the unified distance matrix visualization

Component Planes

Component planes represents clearly the visualization of individual input variables. They are represented in grey scale or a combination of different colors. Fig. 5.

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

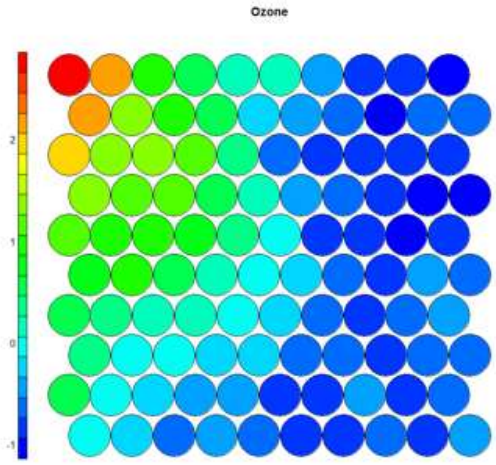
represents component planes for each variable. By inspecting component planes in Fig. 5. it is evident that the variables Ozone, Solar radiation and Temperature are positively correlated and variable Wind is negatively correlated with other variables. Fig. 5. represents component planes for each variable.

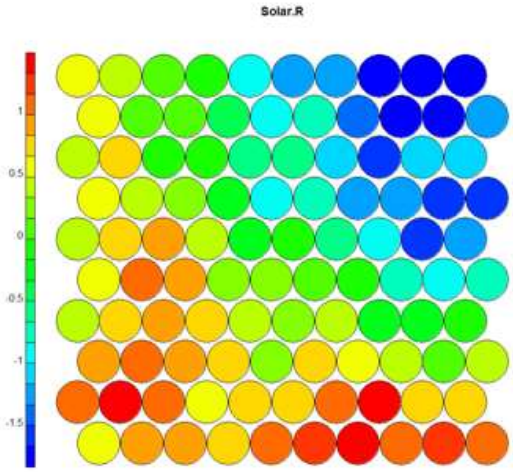
The rate of chemical reactions that produce ozone are affected by solar radiation and temperature. Therefore, an increase in these values increases the chemical reactions and leads to an increase in ozone levels. Wind speed is negatively related to all the variables. The concentration of ozone is high, when there is low wind speeds. The covariance matrix is generated for the input variables and it reflects the same relationship as the component planes.

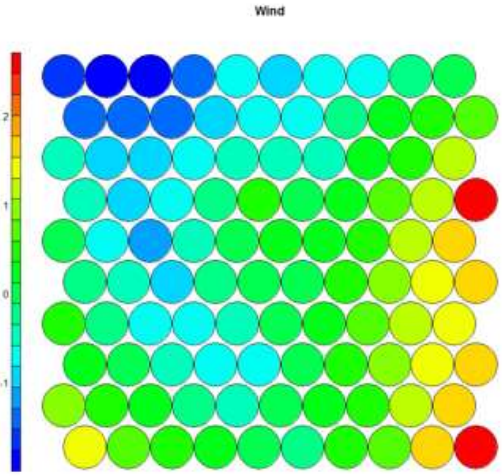
Table 1 represents the correlation coefficient values between the variables. It can be observed that maximum correlation exists between ozone and temperature. Wind is most negatively related with ozone. Component planes help in understanding this concept pictorially, therefore Self Organizing Map helps in easy visualization of relationship between variables.

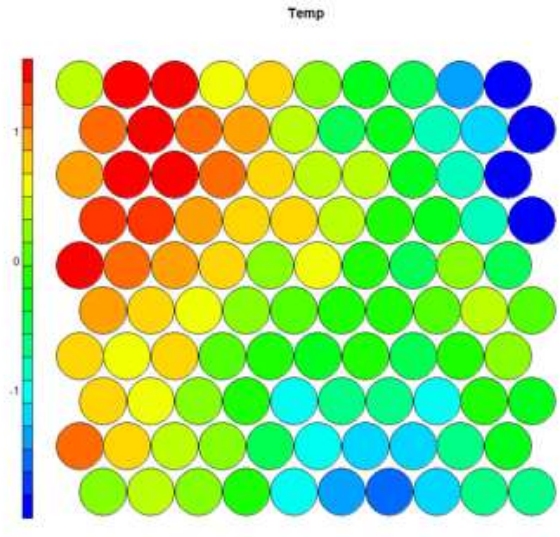
Table 1 represents the correlation coefficient values between the variables

Variables	Ozone	Solar.R	Wind	Temp
Ozone	1.0000			
Solar.R	0.3411	1.0000		
Wind	-0.6265	-0.1147	1.0000	
Temp	0.6938	0.2850	-0.4947	1.0000









Clustering of Self Organising Map

A U-matrix may be appropriate to identify the cluster boundaries. But this may not be effective and therefore, a hierarchical clustering technique is applied after the SOM is trained. Fig. 6. Shows the formation of clusters. Here,

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

there is a formation of 3 clusters – High Ozone, Medium Ozone and Low Ozone.

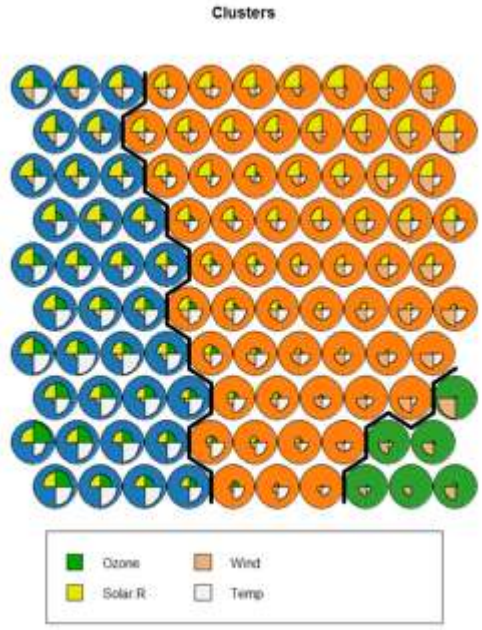


Fig. 6. Blue nodes indicate High Ozone, Orange colour nodes indicate Medium Ozone and Green color nodes indicate Low Ozone.

Fig. 6. shows the formation of clusters into low ozone, medium and high ozone that is depicted by green, orange and blue respectively.

In High Ozone class, the effect of solar radiation and temperature are above average and wind speed has very low effect. The effect of solar radiation and temperature is below average in Low Ozone class and wind speed has significant effect. In Medium Ozone cluster, solar radiation, temperature and wind have about average effect. The clustering results obtained are satisfactory and Self Organising Maps can be used to classify the quality of air.

V. CONCLUSION

In this paper, principal component analysis (PCA) and Self Organising Map (SOM) has been applied to demonstrate the dimensionality reduction of dataset. A case study was considered to visualize and classify air quality data. PCA explained most of the variance in data but it was difficult to interpret the data pattern. However, SOM appears to be the best fit to represent complex data. Especially, the component planes of SOM provides effective visualization and uncovers the correlation between the input variables.

U-matrix can be used for data classification, but may not be a good choice in all cases. Therefore, in this paper hierarchical clustering technique is applied to the SOM results obtained. The quality of air is partitioned into 3 clusters – low, medium and high ozone content. Therefore, it can be concluded that SOM has better resolving power of classification than traditional methods.

REFERENCES

[1] Johnson, Richard Arnold and Wichern, Dean W and others. "Applied multivariate statistical analysis" *s.l. : Prentice hall Englewood Cliffs, NJ,* 1992. Vol. 4.

[2] Gurney, Kevin, "An introduction to neural networks", CRC press, 1997.

[3] Kosanovich, KA and Piovoso, MJ , "Process data analysis using multivariate statistical methods " **,** *IEEE*, pp. 721—724, 1991

[4] Barreto, Alexandre S., "Multivariate statistical analysis for dermatological disease diagnosis", Biomedical and Health Informatics (BHI), *IEEE-EMBS International Conference IEEE,* 2014, pp. 500-504.

[5] Harrou; Fouzi and Nounou; Mohamed Numan and Nounou;Hazem Numan , "Detecting abnormal ozone levels using PCA-based GLR hypothesis testing", *s.l. : IEEE*, 2013, pp. 95-102.

[6] Raychaudhuri, Soumya and Stuart, Joshua M and Altman, Russ B, "Principal components analysis to summarize microarray experiments: application to sporulation time series", Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, NIH Public Access, 2000, pp. 455.

[7] Annas, Suwardi and Kanai, Takenori and Koyama, Shuhei, "Principal component analysis and self-organizing map for visualizing and classifying fire risks in forest regions", *Agricultural Information Research Journal*, 2007,pp. 44-51.

[8] Astel, A., Tsakovski, S., Barbieri, P., & Simeonov, V. , " Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets" *Water* Research,41(19), 2007, pp. 4566-4578.

[9] Klobucar, Damir, and Marko Subasic. "Using self-organizing maps in the visualization and analysis of forest inventory." *iForest-Biogeosciences and Forestry 5.5*, 2012 , pp. 216.

[10] Chattopadhyay, Manojit, Pranab K. Dan, and Sitanath Mazumdar. "Principal component analysis and self-organizing map for visual clustering of machine-part cell formation in cellular manufacturing system." *Systems Research Forum.* Vol. 5., 2011,pp.25-51.

[11] Koua, E. L. "Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets." *Proceedings of 21st*

https://dx.doi.org/10.24001/ijaems.3.11.4 *ISSN: 2454-1311*

International Cartographic Renaissance (ICC), 2003, pp.1694-1702.

[12] Smith, Lindsay I, "A tutorial on principal components analysis" , Cornell University, USA, p. 65, 2002.

[13] Kohonen, Teuvo, et al. "Engineering applications of the self-organizing map."*Proceedings of the IEEE* 84.10,1996, pp. 1358-1384.