ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(29), DOI:10.17485/ijst/2015/v8i29/84111, November 2015* ISSN (Online) : 0974-5645

A Novel Soil Profile Feature Reduction Model using Principal Component Analysis

D. Ashok Kumar[1*] and N. Kannathasan[2]

[1]Department of Computer Science, Government Arts College, Tiruchirapalli-620022, U.T. of Puducherry, India; akudaiyar@yahoo.com

[2]Department of Computer Science, Kanchi Mamunivar Centre for Post Graduate Studies, Puducherry-605008, U.T. of Puducherry, India; nkthasan@gmail.com

Abstract

Background/Objectives: Data Mining has been used to analyze large datasets and establish useful classification and patterns in the datasets. The efficient analysis of data in different format becomes a challenging work. **Methods/Statistical Analysis:** This work proposed a novel Soil Profile Feature Reduction Model using Principal Component Analysis for data reduction. The proposed model uses the method of k-Means clustering and PC Aapproach for feature reduction which initially applies PCA to acquire reduced uncorrelated attributes showing maximal eigenvalues in the dataset with minimum loss of information. Again proposed model uses k-Means on the PCA reduced dataset to find out discriminative features that will be the most sufficient ones for classification. **Findings:** The weight by PCA generates attribute weights of the soil profile dataset using a component created by the PCA. The component is mentioned by the component number parameter. The normalize weight parameter is usually set to true to spread the weights between 0 and 1. The attribute weights reflect the relevance of the attributes with respect to the class attribute. The higher weight of an attribute is more relevant, it is considered. This is a combination of clustering approach with feature reduction to get a minimal set attributes relating a suitably high accuracy in describing the original features. The result of clustering is same after reducing the attributes using PCA. The experimental results prove that proposed model is reducing number of initial attributes, reducing computational complexity and improving predictive accuracy in High Dimensional Datasets. **Applications/Improvements:** The same soil profile feature is implemented by using the other techniques instead of PCA algorithm in future.

Keywords: Clustering, Feature Reduction, k-Means,Principal Component Analysis, Soil Profile

1. Introduction

The nature of soil varies with the ecosystem and so the productivity of that ecosystem. Soils support all life forms on the planet and play vital role for their existence. The top layer of the soil is a natural filter for many contaminants is shown in Figure 1. Soil has five layers and the property of each layer may vary from region to region. The property is dependent on a variety of environmental parameters of that region is depicted in Figure 2.

The soil properties such as texture, porosity, specific yield depend on the total volume of groundwater recharge, water storage and discharge, also the extent of groundwater contamination[1].

Soil is a mixture of sand, silt and clay and each of them vary in their particle size (Table 1). Soil is made of small mineral particles that differ in size based on the type of soil. Clay, silt and sand are common soils which have an increasing order of particle size respectively. The particles of clay are charged and can attract water molecules.

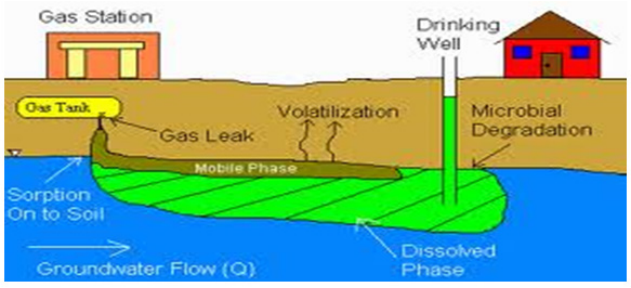


Figure 1. Soil water contamination.

* *Author for correspondence*

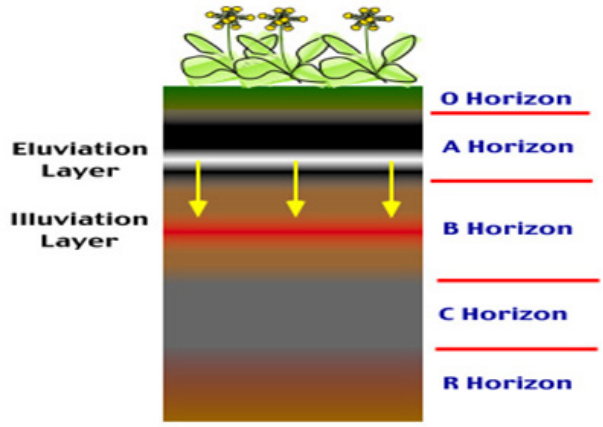


Figure 2. Soil profile.

Table 1. USDA particle size and porosity ranges for sand, silt and clay

Name of separate	Diameter range (millimeters)	Porosity (%)
Sand	2.0 – 0.05	25-50
Silt	0.05 – 0.002	35-50
Clay	Less than 0.002	33-60

Clay is one of the suitable soil for cultivation of crops. It is resistant to erosion, can hold water better than the other soils but the silt of the clay can easily be washed awayand so the nutrients. Whereas the sand allow the water to pass through easily and quickly and it cannot hold water like that of the clayand therefore it cannot support the growth of flora. The sand which is exposed is highly subjected to erosion if the angle and slope of land is too severe.The soil textural triangle gives the textural name, which based on the percentage of sand, silt and clay within the soil sample (Figure 3). The triangle is divided into 10-percent portions of clay, silt and sand. The summation of the three percentages must total 100 percent[2].

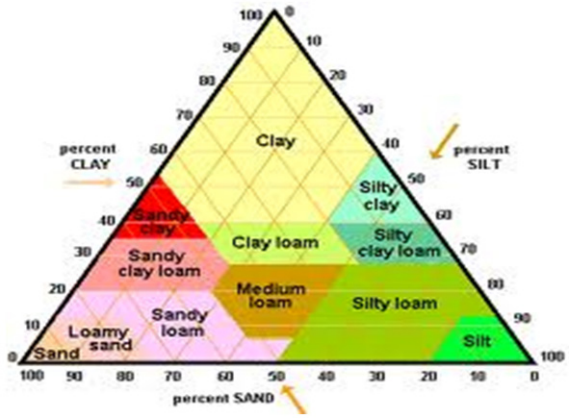


Figure 3. USDA 12 basic soil textural triangle.

Soil system is composed of air, water, dead organic matter and various types of living organisms (Figure 4). It is rich in a number of microorganisms and living matter. The dead and decaying matter of the soil plays a crucial role in formation of humus.

Shape and spatial arrangement of soil particles decide the porosity of that soil. It is the air space or void space between soil particles. Infiltration, ground water movement and storage occur in these void spaces. Porosity of soil typically decreases as pH value increases and also particle size increases because of soil contamination or soil pollution (Figure 5).

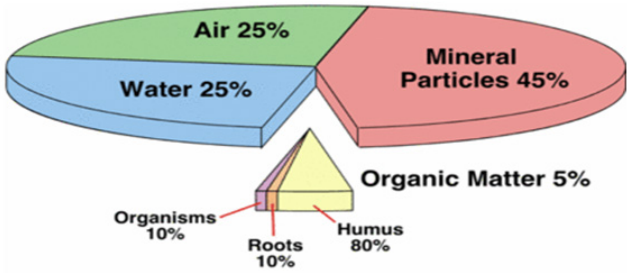


Figure 4. Components of soil.

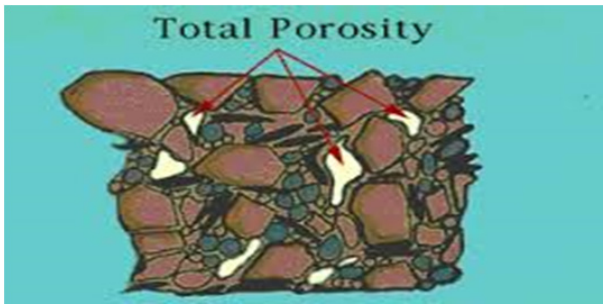
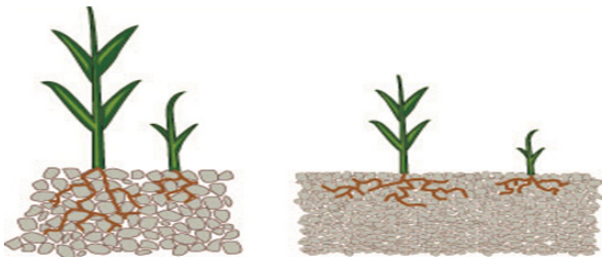


Figure 5. Total porosity of soil.

Bulk density is an important indicator of soil compaction and is given in Table 2. The porosity of soil is governed by soil biota and increases the water and air movement to rhizosphere shown in Figure 6.

Table 2. General relationship of soil bulk density to root growth based on soil textural

Type of Mineral Particle	Ideal bulk densities for Plant Growth (g/cm^3)	Bulk densities that restrict root growth (g/cm^3)
Sand	<1.60	>1.80
Silt	<1.40	>1.65
Clay	<1.10	>1.47



(a) (b)

Figure 6. **(a)** Soil with good structure (pH<=7). **(b)** Soil with poor and dense structure (pH >7).

Soil is polluted by various anthropogenic activities such as addition of industrial wastes, pesticides, fertilizers, domestic sewage etc. to soil. Polynuclear aromatic hydrocarbons and petroleum hydrocarbons and heavy metals are the most common contaminants of soil. Urbanization and industrialization are the major contributors of soil pollution. It leads to numerous health hazards and the chemicals reach human being through contaminated water and plants.

2. Materials and Methods

Regression analysis was used for numeric prediction[7]. Learning is broadly classified into two types: Supervised Learning and Unsupervised Learning. Supervised Learning: The training data are accompanied by the labels representing the class of the observations. New data is classified based on the training set. Unsupervised Learning: The class label of training data is unknown. Given a set of measurements, observations with the aim of establishing the existence of classes or clusters in the data[8–10].

Table 3. List of soil variables their abbreviation and units of measurement

Abbreviation	Description	Units
PHH_2O	Soil reaction in Water	pH units
ECE	Electrical Conductivity of Saturated Paste	dS m^{-1}
ORGC	Organic Carbon	% (mass)
CECC	Cation Exchange Capacity of clay size fraction	$Cmol_c kg^{-1}$
TEB	Total Extractable Bases	$Cmol_c kg^{-1}$
CACO	Calcium Carbonate content	% (mass)
GYPS	Gypsum content	% (mass)
SAND	Sand	%(mass)
SILT	Silt	%(mass)
CLAY	Clay	%(mass)
TOTPOR	Total Porosity	%

Principal Component Analysis (PCA), a mathematical procedure, uses an orthogonal transformation to convert a set of observations of possibly correlated attributes into a set of values of uncorrelated attributes called Principal Components[11]. The test dataset consists of 16383 samples of particular region as in Table 3 using for this research work collected from World Soil Information International Soil Reference and Information Centre[3].

3. Proposed Model using PCA

The Proposed Model consists of three phases such as k-Means Clustering on Original Dataset, to reduce the unnecessary attributes using Principal Component Analysis Feature Reduction method and k-Means Clustering on Reduced Dataset which is shown in Figure 7. The proposed model uses the method of k-Means clustering and PCA approach for attribute reduction, which initially applies PCA to obtain reduced uncorrelated attributes mention maximal eigenvalues in the dataset with minimum loss of information.Then again proposed model usesk-Means on the PCA reduced dataset to find out discriminative features that will be the most sufficient ones for classification.

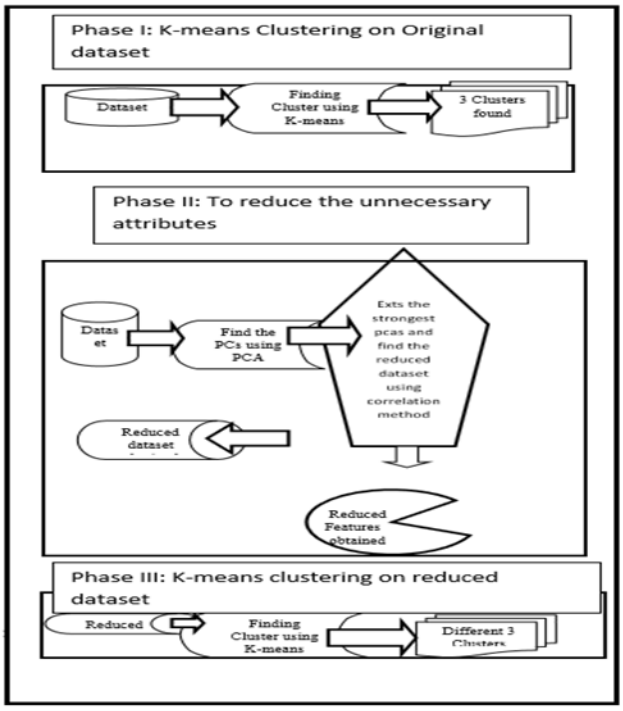


Figure 7. Proposed model.

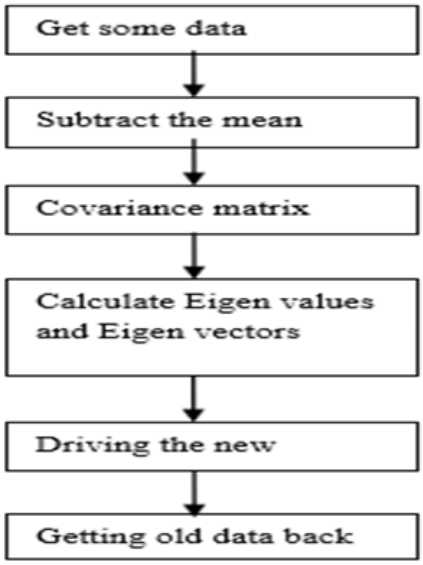


Figure 8. Steps for PCA method.

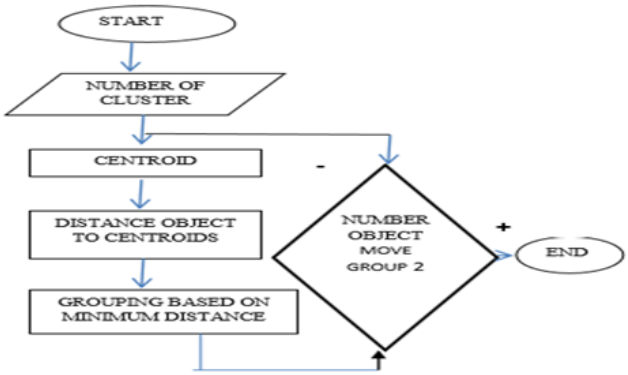


Figure 9. k-means clustering algorithm-process flow.

PCA projects high dimensional data to a lower dimension and projects the data in the least square sense. It captures big principal variability in the data and ignores small variability and reduces the dimensionality of a data set by finding a new set of variables, small than the original set of variables. PCA can also retain most of the sample's information while it extracting relevant information from confusing data sets. It is a simple, non-parametric method for dimension reduction. For this reason it is useful for the compression and classification of data. Principal Component Analysis is used in image process and compression, pattern recognition and other data dimension reduction[6]. Steps for PCA are given below in Figure 8[5].

Cluster analysis is the assignment of a set of observations into subsets so that observations in the same cluster are similar in some sense[4]. The algorithm k-Means is the process of partitioning a group of data points into a small number of clusters. The steps for k-Means algorithm are shown in Figure 9.

4. Experimental Evaluation and Result Analysis

Feature reduction model using Principal Component Analysis is reducing number of initial attributes, reducing computational complexity and improving predictive accuracy. The number of principal components is lesser or equal to the number of original attributes.

Table 4. Cluster in original dataset

Attribute	Cluster_0	Cluster_1	Cluster_2
PHH_2O	-98	6.479	5.130
ECE	-98	0.132	0.100
ORGC	-98	1.645	1.022
CECC	-98	53.252	17.567
TEB	-98	16.453	3.130
CACO	-98	1.206	0
GYPS	-98	0	0
SAND	-98	38.302	53.352
SILT	-98	36.937	19.228
CLAY	-98	24.761	27.420
TOTPOR	-98	49.533	48.668

4.1 Phase-I

Step 1: Collection of dataset: The soil profile data set that is collected is taken.

Step 2: Application of k-Means algorithm: Using the data-set and got the following cluster listing, this is shown in Table 4.

Step 3: Clustering using Centroid Plot View method: Using Centroid Plot View method obtained the 2D as shown in Figure 10.

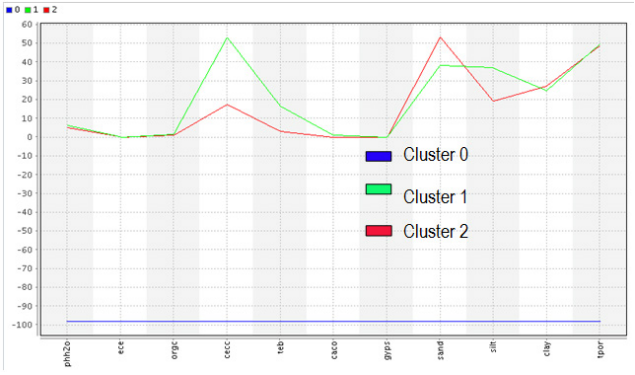


Figure 10. Centroid plot view of original dataset.

4.2 Phase-II

Step 1: Collection of the dataset soil profile: The collected soil profile dataset.

Step 2: Finding the all covariance matrix, Eigen values and Eigen vectors of the dataset: which is shown in Table 5 and Table 6.

Step 3: Clustering using Centroid Plot View method: Using Centroid Plot View method obtained the 2D as shown in Figure 11.

Step 4: Determining the number of meaningful Principal Components: To eliminate the weaker components from the Principal Components set the corresponding Eigen value which is lesser than the other points in that dataset. Consider only the highest values in the dataset those values are taken into account.

Step 5: Finding the reduced data set using the reduced Principal Components: The transformation matrix with reduced Principal Components is applied to the dataset to produce the new reduced projected dataset which can be used for further data analysis, which is shown in Table 7.

Table 5. Eigen values of the covariance matrix

Component	Standard Derivation	Proportion of Variance	Cumulative Variance
PC_1	56.547	0.797	0.797
PC_2	23.429	0.137	0.934
PC_3	14.314	0.051	0.985
PC_4	6.609	0.011	0.996
PC_5	3.088	0.002	0.998
PC_6	1.895	0.001	0.999
PC_7	1.438	0.001	1.000
PC_8	0.519	0.000	1.000
PC_9	0.418	0.000	1.000
PC_10	0.179	0.000	1.000
PC_11	0.033	0.000	1.000

Table 6. Eigen vectors of the covariance matrix

Attribute	PC_1	PC_2	PC_3	PC_4
PHH_2O	0.258	0.099	-0.022	-0.020
ECE	0.241	0.117	-0.033	-0.087
ORGC	0.245	0.101	-0.049	-0.095
CECC	0.405	-0.663	0.362	0.213
TEB	0.297	-0.206	-0.025	0.517
CACO	0.245	0.095	-0.032	0.006
GYPS	0.241	0.117	-0.032	-0.089
SAND	0.327	0.555	0.651	0.008
SILT	0.344	-0.292	-0.170	-0.726
CLAY	0.297	0.209	-0.610	0.359
TOTPOR	0.363	0.157	-0.195	-0.082

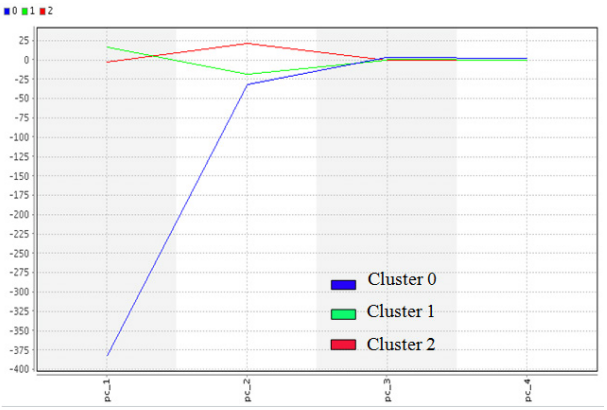


Figure 11. Centroid plot view for reduced dataset.

Table 7. Reduced dataset after application of PCA

Row	PC_1	PC_2	PC_3	PC_4
1	15.888	-10.493	2.224	4.930
2	15.888	-10.493	2.224	4.930
3	15.888	-10.493	2.224	4.930
4	15.888	-10.493	2.224	4.930
5	15.888	-10.493	2.224	4.930
6	15.888	-10.493	2.224	4.930
7	15.888	-10.493	2.224	4.930
8	15.888	-10.493	2.224	4.930
9	15.888	-10.493	2.224	4.930
10	15.888	-10.493	2.224	4.930
11	15.888	-10.493	2.224	4.930
12	15.888	-10.493	2.224	4.930
13	15.888	-10.493	2.224	4.930
14	15.888	-10.493	2.224	4.930
15	15.888	-10.493	2.224	4.930

4.3 Phase-III

Step 1: k-Means cluster analysis using reduced dataset 2D by PCA: This is shown in Table 8.

The Weight by PCA operator produces attribute weights of the given dataset using a component created by the PCA. The component is mentioned by the component number parameter. The normalize weight parameter is usually set to true to spread the weights between 0 and 1. The attribute weights reflect the relevance of the attributes with respect to the class attribute. The higher weight of an attribute is more relevant, it is considered. Weighted for original dataset is shown in Table 9 and Weighted for reduced dataset is shown in Table 10.

Table 8. Cluster in reduced dataset by PCA

Attribute	Cluster_0	Cluster_1	Cluster_2
PHH_2O	-98	6.479	5.130
ECE	-98	0.132	0.100
ORGC	-98	1.645	1.022
CECC	-98	53.252	17.567
TEB	-98	16.453	3.130
CACO	-98	1.206	0
GYPS	-98	0	0
SAND	-98	38.302	53.352
SILT	-98	36.937	19.228
CLAY	-98	24.761	27.420
TOTPOR	-98	49.533	48.668

Table 9. Weighted for original dataset

Attribute	Weight
PHH_2O	0
ECE	0.005
ORGC	0.016
CECC	0.017
TEB	0.017
CACO	0.043
GYPS	0.235
SAND	0.276
SILT	0.540
CLAY	0.936
TOTPOR	1

Table 10. Weighted for reduced dataset

Attribute	Weight
PC_1	0
PC_4	0
PC_2	0
PC_3	1

Clustering was performed by proposed model on the soil profile directly using k-Means. Using PCA the features of the data set were then reduced and clustering was performed again. The clustering results of both the data sets are found to be same. Therefore the model has proved that the PCA is a good feature reduction technique, it can handle data without any loss.

5. Conclusion

The proposed model first finds the cluster of soil profile dataset using the k-Means clustering method and then the PCA approach of feature reduction technique has been implemented on that data set having continuous attribute values. As a result of which a number of uncorrelated and discriminative attributes, more adequate for classification has been obtained. These attributes also specifies the maximal variances among the dataset by retaining the original property of the dataset. Then model applied the k-Means clustering method on that reduced dataset. The same cluster listing obtained in both the methods. So, the experimental results prove that the PCA method is highly useful for getting the same cluster and also takes less time. Therefore, improving the effectiveness of the proposed model in terms of classification accuracy will be investigated in the future using some other feature reduction methods, like Wavelet Transform.

6. References

1. Kumar D, Kannathasan N. A survey on data mining and pattern recognition techniques for soil data mining. IJCSI International Journal of Computer Science Issues. 2011 May; 8(3):422–8.
2. Study guide: Soil mechanics level 1, module 3, unified soil classification system, National employee development staff, soil conservation services. United States Department of Agriculture; 1987.
3. Batjes NH. ISRIC-WISE global data set of derived soil properties on a 0.5 by 0.5 degree grid (version 3.0). Report 2005/08. Washington: ISRIC-World Soil Information.
4. Mishra S, Mishra D, Das S, Rath AK. Feature reduction using principal component analysis for agricultural data set. IEEE 3rd International Conference on Electronics Computer Technology; Kanyakumari. 2011 Apr 8-10. p. 209–13.
5. Smith LI. A tutorial on principal components analysis; 2002 Feb.
6. Xiao B. Principal component analysis for feature extraction of image sequence. 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE); Chengdu. 2010 Jun. p. 250–3.
7. Ashok Kumar D, Kannathasan N. Analysis of linear and segmented linear regression of fruit yield on soil salinity. Proceedings of International Conference on Mathematical Modeling and Applied Soft Computing; 2012. p. 275–82.
8. Kumar DA, Kannathasan N. A study and characterization of chemical properties of soil surface data using k-Means algorithm. Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME); Salem. 2013. p.264–70.
9. Kumar DA, Annie MCLC, Begum TUS. Computational time factor analysis of k-Means algorithm on actual and transformed data clustering. 2012 International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME); Salem, Tamil Nadu. 2012 Mar 21-23.p. 49–54.

10. Thangavel K, Ashok Kumar D. Optimization of code book in vector quantization. Annals of Operations Research. 2006 May; 143(1):317–25.
11. Nachimuthu VV, Robin S, Sudhakar D, Raveendran M, Rajeswari S, Manonmani S. Evaluation of rice genetic diversity and variability in a population panel by principal component analysis. Indian Journal of Science and Technology. 2014 Oct; 7(10):1555–62.