FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 4 No. 2, June, 2020, pp 780 – 787
DOI: https://doi.org/10.33003/fjs-2020-0402-363

K-MEANS CLUSTERING ALGORITHM BASED CLASSIFICATION OF SOIL FERTILITY IN NORTH WEST NIGERIA

[1]Hassan Ibrahim Hayatu, [2]Abdullahi Mohammed, [2]Ahmad Barroon Isma'eel, [2]Yusuf Sahabi Ali

[1]Institute for Agricultural Research, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.
[2]Department of Computer Science, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.

*Corresponding Author's Email: ibrogo@gmail.com

ABSTRACT

Soil fertility determines a plant's development process that guarantees food sufficiency through bumper harvests. The fertility of soil varies according to regions, thereby determining the type of crops to be planted. However, there is no repository or any source of information about the fertility of the soil in any region in Nigeria, especially the Northwest of the country. The only available information are soil samples with their attributes which gives little or no information to the average farmer. This has affected crop yield in all the regions, more particularly the Northwest region, thus resulting in lower food production. This has also affected the security of lives and properties as the struggle to identify fertile soil continues. The identification of fertile soil and transmitting such information to farmers will help the country attain food security and enhance the security of lives and properties in the country. Therefore, this study is aimed at classifying soil data based on their fertility in the Northwest region of Nigeria using R programming. Data was obtained from the department of soil science, Ahmadu Bello University, Zaria. The data contain 400 soil samples containing 13 attributes. The relationship between soil attributes was observed based on the data. K-means clustering algorithm was employed in analyzing soil fertility clusters using soil attributes such as Nitrogen (N), Potassium (K), Phosphorus (P), Magnesium (Mg), Sodium (Na), Calcium (Ca), Organic carbon (OC), Electrical Conductivity (EC), Salinity(SL), Clay (CY), Sand (SN), Calcium chloride ($CaCl_2$) and PH contents. Results show that there is a positive relationship between PH and CaCl2, Ca, Mg and EC and also a close negative relationship between SL, SN and CE. The remaining parameters are not related to one another. Additionally, four clusters were identified with cluster 1 having the highest fertility, followed by 2 and the fertility decreases with increasing number of clusters. The identification of the most fertile clusters will guide farmers on where best to concentrate on when planting their crops in orded to imporve productivity and crop yield.

Keywords: Clustering, kmeans, soil, fertility, clustering tendency.

INTRODUCTION

Soil fertility is key to producing productive crops. There is a demand to produce more food as population continues to grow. Continuous cropping for improved yield eliminates large nutrient amounts from the soil, thus affecting the fertility of the soil. Soil fertility is usually determined on the basis of nutrient presence or absence, i.e. macro and micronutrients (Gruhn *et al.,* 2000). Sustainable soil productivity depends on the soil's ability to provide the plants with essential nutrients. So the assessment of soil fertility status is an important feature of sustainable agriculture (Maathuis, 2009).The aim of this work is to use datamining technique to classify soil based on its fertility in northwest Nigeria.

Data extraction is used to evaluate large data sets and identify useful trends in the data. In different fields, data mining techniques are used to find patterns that are used in analysis and prediction. Many studies describe how the classification and clustering techniques are used to analyze agricultural data especially soil information (Hooman *et al.,* 2015; Manjula & Djodiltachoumy, 2017; Muneshwara *et al.,* 2020).The results of soil analysis on various data sets with a variety of data mining techniques may be useful for farmers to gain insight on the soil properties, thus determining the type of crop and fertilizer to use. This knowledge will effectively improve crop yield. The soil analysis may be used in various way such as to protect the environment, diagnosis of crop culture troubles, to identify nutrient deficiencies, energy conversation, and so on (Madhuri *et al.*2018).This study is aimed at classifying soil data based on their fertility in Northwest Nigeria using R programming. The association between soil attributes were studied, so also is the clustering trend of soil fertility according to some parameters (CY, SN, SL, PH, CaCl2, OC, N, Ca, P, Mg, K, Na, and EC) using K-means clustering algorithm. Table 1 describes various research works that are related to soil fertility classification or prediction using different data mining techniques from the literature.

The remaining part of this paper is structured as follows: Section 2 provided review of related literature. Section 3 presents the methods and materials used in the study. In section 4, the result analysis and discussion of finding were presented. Finally, we conclude the paper by providing the summary of our findings and discuss our future directions in section 5.

REVIEW OF RELATED LITERATURE

Table 1: Related Works on Soil Fertility Classification using Data Mining Techniques

S/N	Author(s)	Tittle	Technique Used	Outcome
1.	(Bhagavi & Jyothi, 2011)	Soil classification using data mining techniques: A comparative study	GATree, Fuzzy classification and Fuzzy C-means	Soil texture classification
2.	(Jay, 2012)	Performance Turning of J48 Algorithm for Prediction of Soil Fertility	J48	Prediction of soil fertility
3.	(Rajeswari & Arunesh, 2016)	Analyzing Soil Data Using Data Mining Classification Techniques	JRip, J48 and Naïve Bayes	Soil type prediction
4.	(Manjula & Djodiltachoumy, 2017)	Data Mining Technique to analyze soil nutrients based on Hybrid Classification	Naïve Bayes, Decision Tree and Hybrid of the two	Investigate soil supplement
5.	(Noor, 2017)	A Study of Data Mining Tools and Techniques to Agriculture with Application	ANN, SVM and bi-clustering	Crop yield prediction
6.	(Nikhita & Abhay, 2017)	Application of Data Mining Classification Techniques on Soil Data Using R	ANN and SVM	Soil Classification
7	(Marzieh *et al*., 2017)	Using Self-Organizing Maps for Determination of Soil Fertility	Self-Organizing Maps (SMO)	Determination of Soil Fertility
8.	(Rounak, 2018)	Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land Soil	Naïve Bayes, ZeroR and Stacking	Prediction of Soil type
9.	(Jeyalaksshmi*et al*., 2019)	Data Mining in Soil and Plant Nutrient Management, Recent Advances and Future Challenges in Organic Crops	J48, Naïve Bayes, Random Forest and Hybrid Neural Network	Earth Upgrade Investigation
10.	(Fathima & Sharmila , 2019)	Classification of Soil Based on Fuzzy Logics	Fuzzy Logic	Soil Fertility Classification
11.	(Muneshwara, *et al.,* 2020)	Soil Fertility Analysis and Crop Prediction Using Machine Learning	SVM, KNN and Random Forest	Soil type classification and Crop Recommendation.
12.	(Samundeers *et al.,* 2020)	Soil Data Analysis & Crop Yield Prediction in Data Mining Using R-programming	Decision Tree and C4.5	Crop yield prediction.
13.	(Saranya & Mythili, 2020)	Classification of Soil and Crop Suggestion Using Machine Learning Techniques	KNN, Bagged tree, SVM and Logistic Regression	Prediction of crop to be cultivated.

From the literature review conducted, it was observed that none of work uses soil dataset from northwestern Nigeria and most of them does not bother to check the clustering tendency of their dataset, which this may affect the clustering result. Therefore, this motivate us to conduct this work as it will be very helpful especially at this time that the government and people in the country are going back to farm in order to attained food security and economic development.

MATERIAL AND METHODS

This section describes the source of the soil data used, description of the collected data, pre-processing technique used, the clustering technique used and the programming language used in implementing the proposed study. The flow of the study is shown in Figure 1.

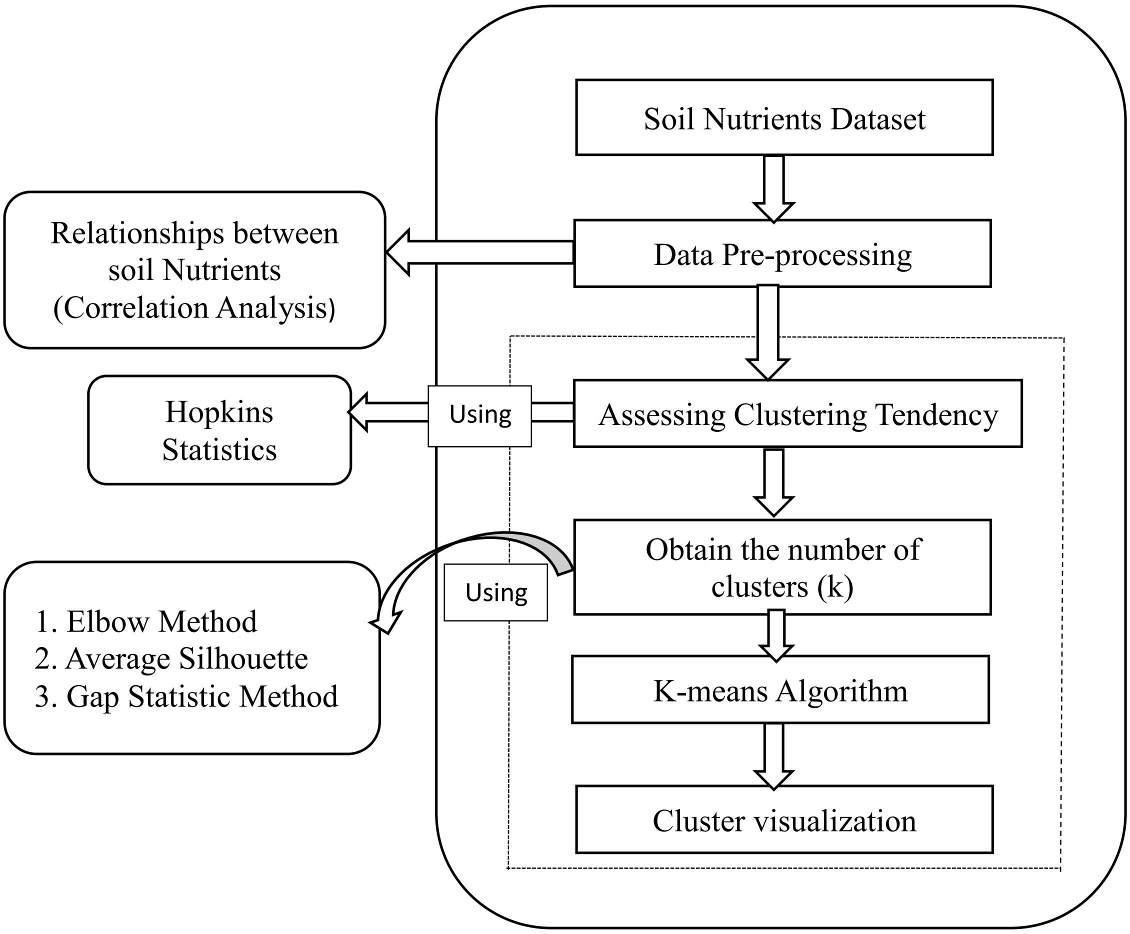

Figure 1: Architecture of the Proposed Study

Data Source and Description

Soil data was collected from Soil Science Department, Ahmadu Bello University, Zaria. The data contains 400 soil samples from the North West zone of Nigeria, it contains 13 attributes which includes CY, SN, SL, PH, CaCl2, OC, N, Ca, P, Mg, K, Na and EC. Table 2 shows the attribute description and Table 3 shows the samples of the dataset with their corresponding percentages of attributes in Table 2.

Table 2. Description of the dataset

Feature	Description
CY	Clay content of the soil (%)
SL	Salinity of the soil (%)
SN	Quantity of sand of the soil (%)
PH	PH value of the soil (ppm)
CaCl2	Calcium chloride content of the soil (ppm)
OC	Organic carbon (ppm)
N	Nitrogen Content of the soil (ppm)
P	Phosphorus content of the soil (ppm)
Ca	Calcium content of the soil (ppm)
Mg	Magnesium content of the soil (ppm)
K	Potassium content of the soil (ppm)
Na	Sodium content of the soil (ppm)
EC	Electrical conductivity of the soil (ppm)

Table 3. Dataset sample

Sample	CY	SL	SN	PH	$CaCl_2$	OC	N	P	Ca	Mg	K	Na	EC
1	9	38	53	6.2	5.6	0.41	0.07	2.8	1.92	0.4	0.19	1.3	4.8
2	9	28	63	6.8	5.7	0.34	0.07	3.33	2.08	0.4	0.14	0.96	4.2
3	17	44	39	6.6	5.6	0.54	0.14	2.63	2.16	0.46	0.12	1.3	6.7
4	17	40	43	6.2	5.5	0.6	0.07	2.9	2.83	0.83	0.09	0.17	5.3
5	15	38	47	6.4	5.8	0.43	0.07	5.08	7.75	4.4	0.19	0.35	14.4
6	21	42	37	6.3	5.4	0.34	0.07	2.98	2	0.7	0.34	0.87	4.6
7	9	42	49	6.5	5.5	0.47	0.14	3.68	1.67	0.82	0.05	0.96	4
8	7	14	79	6.7	5.7	0.36	0.07	4.03	2.46	0.2	0.2	1.3	5.4
9	9	20	71	6.5	5.8	0.41	0.14	5.95	2	0.6	0.34	1.3	4.8
10	11	46	43	6.6	5.9	0.73	0.14	3.85	2.78	0.8	0.07	2.17	6.3

Preprocessing

Data pre-processing is an important stage for handling the data before using it in the data mining algorithms. This process involves various steps, including handling missing values, categorical attribute handling, normalization, feature selection, transformation. In this study, mean imputation technique is used to handle missing values and normalization technique to convert the feature into the same scale as this may improve the performance.

K-means Clustering

K-means clustering algorithm developed by MacQueen (1967) is one of the most widely used unsupervised machine learning algorithms for splitting a dataset into a number of *k* shades (clusters), in which the k denotes the number of clusters mostly provided by the data scientist. It categorizes items or objects in multiple clusters, such that items in the same cluster are related to each other with high intra-class similarity, while items from different clusters are distinct from each other with low inter-class similarity. In k-means algorithm, every cluster is denoted by its centroids which is defined as the mean of points within the given cluster.

In k-means clustering algorithm, the first step is to determine the number of clusters which will be obtained as the final result, which is the parameter *k*. Then *k* items or objects are randomly selected as centroids on the cluster. All remaining items (objects) are assigned to their nearest centroid based on a distance measure (mostly Euclidean Distance Metric). In the next step, the algorithm computes the new mean value of each cluster. To build this step the term "centroid update" cluster is used. Now that the centers are recalculated, each observation is once again tested to see whether it may be closer to a different cluster. All the objects are reassigned using the cluster updated means. The cluster assignment and centroid update steps are repeated iteratively till the cluster assignments cease to change (until convergence criterion is met). That is, in the current iteration, the clusters generated are the same as those obtained in the previous iteration (Hooman *et al.,* 2015).

R Programming

R is considered to be one of the most popular methods for data processing and mining. This promotes mathematical computation, and decreases programming effort. The graphs are easy to map and illustrate. Various statistical and graphical techniques may be implemented with the help of R. Statistical advance and data mining packages are also given in R. Also, R programming software provides us with different packages and built-in functions which makes statistical analysis very simple. R offers well-designed plots, efficient data processing and storage facilities. R is used in pre-processing of data, data visualization, predictive analysis, statistical modeling and deployment (Team, 2013).

RESULTS AND DISCUSSION

Relationship between Attributes

To measure the relationship between the attributes, a Pearson correlation analysis in R was used. The "easystats / correlation package" by Makowski *et al.,* (2019) is used to determine the relationship between the soil's attributes as shown in Figure 2. Although it shows that the attributes does not have much relationships. But, it indicates a strong negative relationship between CY, SN and SL. The result also shows that there is strong positive relationship between PH and CaCl2, Ca, Mg and EC.

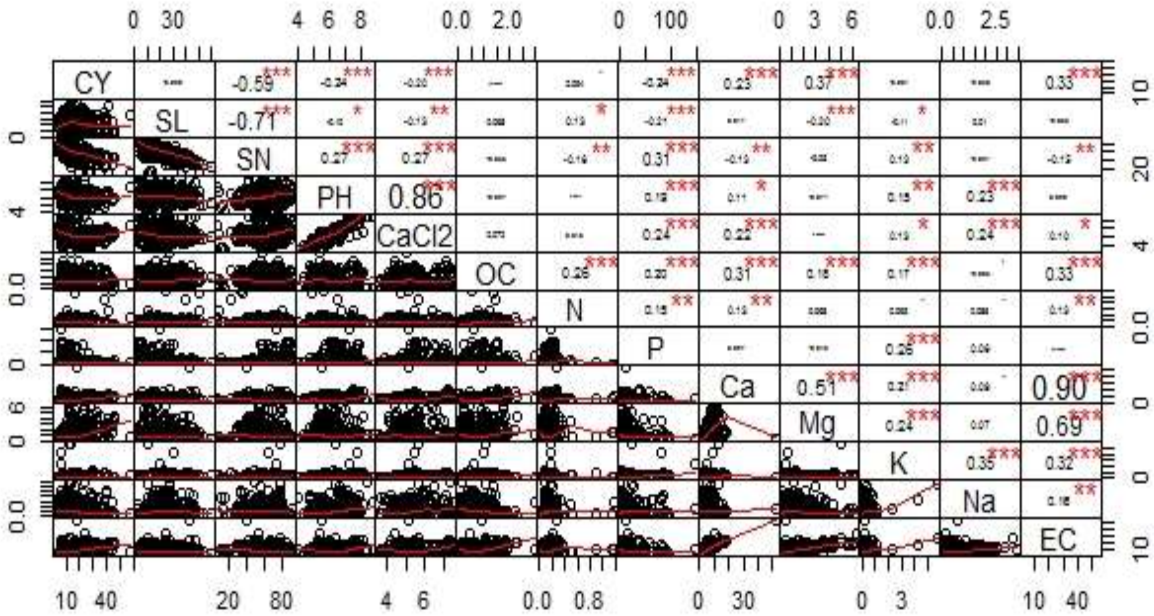

Figure 1: Relationship Between the Soil Attributes

Assesing Clustering Tendency of the Dataset

One big problem in cluster analysis is that clustering methods can return clusters even if there are no clusters in the data. In other words, if clustering techniques are arbitrarily applied to a data set, the data would be separated into clusters.To prevent this, it's important to determine whether or not the datasets contain significant clusters (non-random structures) before applying any clustering method to a data. This approach is what is termed as clustering trend assessment or clustering research feasibility (Han *et al.,* 2012). This study employ the use of hopkins statistics method, which assess the clustering tendency of the dataset by measuring the probability that the given dataset is generated by a uniform data distribution (i. e. it test the spatial randomness of the data) (Andreas *et al.,*2018). Hopkins function from the "clustetend package" YiLan & RuTong (2015) in R was used to calculate the hopkins statistics for the soil dataset. The result shows that the soil dataset used in this work is highly clusterable (H = 0.107, well below 0.5).

Determining the Number of Clusters

Before applying k-means clustering, the first step is the determination of the number of clusters which is to be used in the algorithm. In the literature, various methods were proposed to determine the number of clusters in a dataset. These methods include, but not limited to, the Elbow method, which considers the total intra-cluster variation or total within-cluster sum of squares (WSS) as a function of the number of clusters (Syakur *et al.,*2018). Average Silhouette method, which was presented by Kaufman & Peter (1990). This method calculates the average silhouette of observations for different value of the number of clusters (k) for which the optimum number of clusters k is that which maximizes the average silhouette over some set of possible values for k (Chunhui & Haitao, 2019) and Gap Statistics method, which compare the total intra-cluster variation for different k values with the expected values under the data distribution of null reference. The optimal cluster estimate will be value that maximizes the gap statistics (i.e., that yields the largest gap statistics) (Charrad *et al.,* 2014).

In this study, both methods mentioned above was used to determine the possible number of clusters in the dataset using “factoextra” package in R (Kassambara & Mundt, 2016). The result of this analysis for the Elbow, Silhouette and Gap statistic methods is shown in Figure 3, figure 4 and Figure 5 respectively. Figure 3 shows that the Elbow method selected a maximum of 4 clusters, Figure 4 also shows that the Silhouette method selected a maximum of 4 clusters. However, Figure 5 shows that, the Gap statistics selected a maximum of 10 clusters. Based on this result, 4 clusters is considered to be the optimal number of clusters for the soil data used in this study

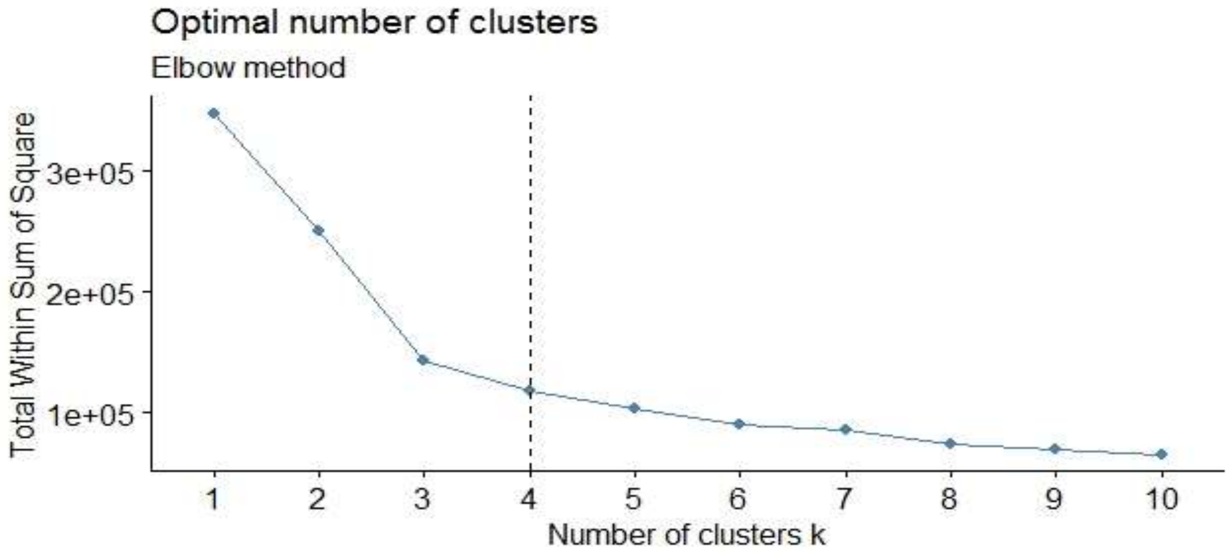

Figure 3: Number of Clusters Selected Using Elbow Method

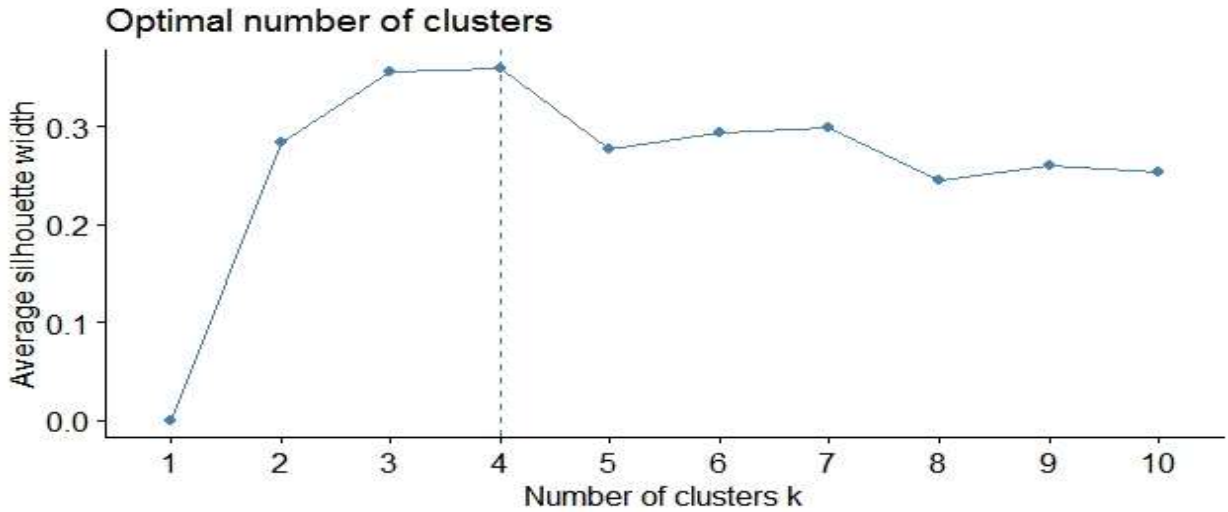

Figure 4: Number of clusters Selected Using silhoutte method

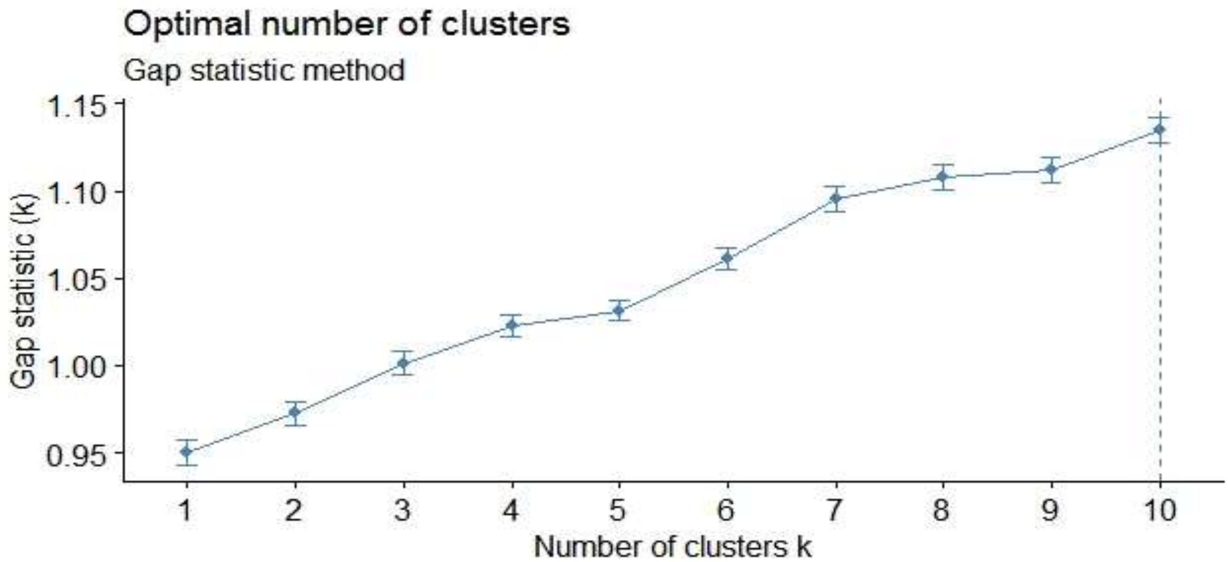

Figure 5: Number of Clusters Seleceted Using Gap statistics

K-means with 4 Clusters.

The K-means clustering was applied to describe the soil fertility status of the given soil using "cluster" package in R developed by Maechler et al. (2019). The result shows that cluster 1 soils are more fertile than the others, followed by cluster 2 soils and so on. From the visual representation of the clusters shown in Figure 5, we can say that most of the soils in the study area belong to cluster two and three.

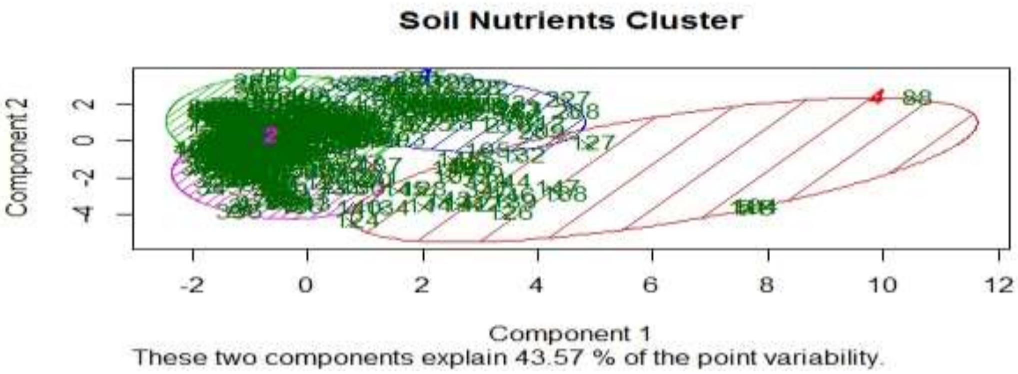

Figure 6: Four Clusters of the Soil Dataset Generated by K-means Algorithm

CONCLUSION

Unlabel soil dataset from Soil Science Department, Ahmadu Bello University, Zaria was used to classify soil fertility using k-means clustering algorithm. Relationship between soil nutrients were studied. The clustering tendency of the soil dataset was also studied before applying the k-means algorithm. The result shows that there is a positive relationship between PH and CaCl2, Ca, Mg and EC for soil fertility and also a close negative relationship between SL, SN and CY. The remaining parameters are not related to one another. It also showed that the soil is classified into four clusters, with cluster 1 having the highest fertility rate, followed by cluster 2, and the fertility decreases with increasing number of clusters. In the future work, the authors intend to develop crop and fertilizer recommendention system as this will help farmers with quick and reliable decision system in order to improve crop production.

REFERENCES

Andreas, A., Margareta , A., & Naomi , C. B. (2018). To Cluster, or Not to Cluster: An Analysis of Clusterability Methods. *arXiv*, 1-30.

Bhagavi, P., & Jyothi, S. (2011). Soil Classification Using Data Mining Techniques: A Comparative Study. *International Journal of Engineering Trends and Technology*, 55-59.

Charrad, M., Nadia , G., Véronique , B., & Azam , N. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 1–36.

Chunhui, Y., & Haitao, Y. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *Multi Disciplinary Scientific Journals*, 1-10.

Fathima, R. M., & Sharmila , K. (2019). Classification of Soil Based on Fuzzy Logics. *International Journal ofadvanced Research in Computer and Communication Engineering*, 131-136.

Firdaus, S., & Uddin, M. A. (2015). A survey on clustering algorithm and complexity analysis. *Intj. Comput. Sci. issues*, 62-85.

Gruhn, P., Goletti, F., & Yudelman, M. (2000). Integrated nutrient management, soil fertility, and sustainable agriculture: current issues and future challenges. *Intl Food Policy Res Inst*.

Han, J., Micheline, K., & Jian , P. (2012). *Data Mining: Concepts and Techniques.* Boston: Morgan Kaufmann.

Hooman , F., Leila , M., & Narsis , Z. (2015). The Application of Data Mining Techniques in Agricultural Science. *Cientia e natura*, 108-116.

Jay, G. (2012). Performance Turning of J48 Algorithm for Prediction of Soil Fertility. *Asian Journal of Computer Science and Information Technology (AJCSIT)*, 251-252.

Jeyalaksshmi, S., Rama, V., & Suseendran, G. (2019). Data Mining in Soil and Plant Nutrient Management, Recent Advances and Future Challanges in Organic Crops. *International Journal of Recent Technology and Engineering (IJRTE)*, 213-216.

Kassambara, A., & Mundt, F. (2016). Extract and Visualize the Results of Multivariate Data Analyses. Version 1.0.3.

Kaufman, L., & Peter , R. (1990). Finding Groups in Data: An Introduction to Cluster Analysis.

M A Syakur, B K Khotimah, E M S Rochman, & B D Satoto. (2018). integration K-Means Clustering method and Elbow Method For Identification of The Best Cluster Profile. *IOP Conference Series: Materials Science and Engineering* (pp. 1-6). IOP Publishing.

Maathuis, F. (2009). Physiological functions of mineral macronutrients. *Current opinion in plant biology*, 250–258.

Madhuri , K., Someswari , P., & Divya , B. Y. (2018). A Survey of using Data Mining Techniques for Soil Fertility. *International Journal of Engineering & Technology*, 917-918.

Maechler, M., Rousseuw, P., Struyf, A., Hubert, M., & Honik, K. (2019). Cluster: Cluster Analysis basis and Extensions. *R package*, Version 2.1.0.

Makowski, D., Ben-Shacha, M. S., M., S. P., & Lüdecke, D. (2019). Methods and Algorithms for Correlation Analysis in R. *Journal of Open Source Software*, 51.

Manjula, E., & Djodiltachoumy, S. (2017). Data Mining Technique to analyse soil nutrients based on Hybrid Classification . *International Journal of Advance Research in Computer Science*, 505-509.

Marzieh, M., Mahdi , N.-G., & Abdol Rassoul , Z. (2017). Using Self-Organizing Maps for Determination of Soil Fertility. *Soil and Water Ress.*, 11-17.

Muneshwara, M. S., Abigail, A. G., Neha, C. G., Preethi, & Akarsh, S. (2020). Soil Fertility Analysis and Crop Prediction Using Machine Learning. *International Journal of Innovative Technology & Exploring Engineering (IJITEF)*.

Nikhita, A., & Abhay, B. (2017). Application of Data Mining Classification Techniques on Soil Data Using R. *International Journal of Advances in Electronics and Computer Science*, 33-37.

Noor, A. (2017). A Study of Data Mining Tools and Techniques to Agriculture with Application. *International Journal of Trend in Research and Development (IJTRD))*, 1-4.

Oteros, J., García-Mozo, H., Hervás, M., & C., G. C. (2013). Year clustering analysis for modelling olive flowering phenology. *Int J Biometeorol*, 545-547.

Rajeswari, V., & Arunesh, K. (2016). Analysing Soil Data Using Data Mining Classification Techniques. *Indian Journal of Science and Technology*, 1-4.

Rani, Y., & Rohil, H. (2013). A study of hierarchical clustering algorithm . *Intl. J. Inf. Comput. Technol*, 1115-1122.

Rounak, J. (2018). Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land Soil. *International Journal for Research in Applied Sciences and Engineering Technology (IJRASET)*, 189-193.

Samundeerswari, K., & Srinivasan, K. (2020). Soil Data Analysis & Crop Yield Prediction in Data Mining Using R-programming. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 1857-1860.

Saranya, N., & Mythili, A. (2020). Classification of Soil and Crop Suggestion Using Machine Learning Techniques. *IJERT*, 671-673.

Taiyun, W., & Viliam, S. (2017). R package "corrplot: Visualization of correlation matrix (Version 0.84). Retrieved from https:github.com/taiyun/corrplot

Team, R. C. (2013). R: A language and environment for statistical computing.

YiLan, L., & RuTong, Z. (2015). Clustertend: check the clustering tendency. *R package version 1.4*.