Journal of Statistical and Econometric Methods, Vol 13, No 1, 2024, 57-68
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1313
Scientific Press International Limited

Clustering of Countries Based on Fertilizer Consumption

Rajarathinam Arunachalam[1]

Abstract

This study investigates fertilizer consumption patterns across 41 countries from 1961 to 2021. Descriptive statistics reveal substantial variation in average fertilizer use (kg/hectare) and variability between countries. Statistical analyses confirm significant differences in consumption patterns, with fertilizer consumption identified as a major explanatory factor. Cluster analysis further differentiates countries into two groups based on their consumption levels. These findings highlight the need to consider fertilizer use within a country-specific context. Further research exploring the underlying factors driving these variations and their impact on crop yields could inform strategies for optimizing fertilizer use and promoting agricultural sustainability.

JEL classification numbers: E18, HO, I1, J64, J88.
Keywords: ANOVA, MANOVA, Cluster analysis, Inter and intra-cluster distance.

[1] Department of Statistics, Manonmaniam Sundaranar University.

Article Info: Received: March 15, 2024. Revised: April 16, 2024.
Published online: April 22, 2024.

1. Introduction

Fertilizers are crucial in modern agriculture which provides essential plant nutrients to enhance crop yields. Understanding fertilizer consumption patterns across countries is vital for ensuring global food security and promoting sustainable agricultural practices. This study delves into fertilizer consumption data for 41 countries from 1961 to 2021 (the highest fertilizer consumption countries). Fertilizer consumption in this study refers to the quantity of plant nutrients (nitrogenous, potash, and phosphate fertilizers) applied per unit of arable land. Traditional nutrient sources like animal and plant manure are excluded. While some compile data based on the calendar year (January-December), others might utilize a split-year approach.
This study aims to achieve the following objectives:

- Identify distinct clusters of countries based on their fertilizer consumption patterns.
- Determine if there are statistically significant differences in fertilizer consumption across these clusters or different years and countries.
- Analyze the variability in fertilizer consumption within and between clusters.

The findings of this study will contribute to understanding the global fertilizer consumption trends. This knowledge can inform policymakers in developing strategies to optimize fertilizer use for improved agricultural productivity while minimizing environmental impact.

2. Review of literature

Li et al. (2007) proposed a novel clustering method, clustering on Bi-clipped (CBC) stream data for streaming time series data. CBC efficiently handles outliers by combining piecewise aggregate approximation (PAA) for dimensionality reduction and a Bi-clipped process. Experimental results demonstrate CBC's superiority over traditional methods like M-clipped, offering higher-quality solutions in less time, particularly in the presence of outliers. This innovative approach presents a valuable advancement in clustering techniques for time series data streams.
Rajarathinam et al. (2010) an empirical investigation was carried out to study the pattern of variability of five soil parameters simultaneously across the villages and to group these villages having the same pattern by considering data from 47 villages of Kadana taluka (Panchmahal district) of Gujarat state. The data on five soil parameters viz., pH, Electrical Conductivity (EC), Organic Carbon (OC), available Phosphorus (P), and Potassium (K) were obtained from the Soil Health Card (SHC) scheme. These soil parameters were subjected to various statistical analyses. The Analysis of Variance (ANOVA) test indicated that the variation due to pH, EC, OC, P, and K among the villages was highly significant and it indicated that the individual parameters were significantly different across the villages. The Multivariate Analysis of Variance (MANOVA) test revealed a significant

variability pattern between the villages when all the five soil parameters were considered simultaneously. Though all the soil parameters were found individually significant (ANOVA) as well as when considered simultaneously (MANOVA), Ward's clustering technique considered only OC and EC for clustering. Accordingly, three clusters were formed such that there was homogeneity within the clusters and heterogeneity between the clusters. This grouping would help prepare fertility maps as well as implement any soil improvement programme effectively and efficiently

Ferreira et al. (2019) investigate distance measures for clustering remote sensing image time series, focusing on land use and cover monitoring using Self-Organizing Maps (SOM). Evaluating Dynamic Time Warping (DTW), Euclidean (ED), and Manhattan (MD) similarity measures, the study finds that ED and MD outperform DTW in accuracy for this application. The findings offer valuable insights for optimizing clustering techniques in Earth observation-based land use and cover classification, contributing to refining remote sensing data analysis methodologies.

Gangopadhyay et al. (2020) attempt to develop a village-level soil nutrient information system (SNIS) in the Piprakothi block of East Champaran district, Bihar representing the Middle Indo-Gangetic Plains of India. Geographic information system (GIS) based grid sampling was done in the block at 2.5 ha intervals and the nutrient index value (NIV) of each grid soil sample was determined. Moderate to strong alkalinity of soils coupled with a deficiency of available nitrogen (N), phosphorus (P), sulfur (S), zinc (Zn), and boron (B) are the main soil fertility constraints of the block. Correlation studies showed significant positive relationship of organic carbon (OC) with available N (r = 0.63**), K (r = 0.58**), S (r = 040*), Zn (r = 0.39*), Fe (r = 0.58**) and B (r = 0.34*) indicating the role of soil organic matter towards soil fertility. The NIV and pH-based risk zones were developed in the block by giving weightage followed by developing decision trees. The villages namely, Chand Saraia, Dhekaha, Kazipur, Majharia, and Piprakothi were identified under high-risk zones with constraints of low NIV, whereas, Babhnaulia, Belwatia, Jiwdhara, and Serma were marked under high pH risk zones. Villages with similar sets of soil fertility characteristics were grouped using hierarchical clustering towards varying management strategies.

Hayatu et al. (2020) presented a critical issue in agricultural productivity by exploring soil fertility classification in Nigeria's northwest region using advanced statistical techniques. The research unveils significant relationships among key factors like pH, calcium chloride, and others by analyzing soil attributes with the K-means clustering algorithm. Identifying distinct soil fertility clusters offers valuable insights for farmers, potentially improving crop yield and enhancing regional food security. The study's findings provide a foundation for informed decision-making in agricultural practices, emphasizing the importance of soil management strategies for sustainable food production and livelihoods.

Rajarathinam and Ramji (2021) delve into soil data from 47 villages in Palayamkottai taluka, Tamil Nadu, India, sourced from the Soil Health Card scheme. Statistical analyses uncover significant variations across villages for key soil parameters like pH, Electrical Conductivity, Organic Carbon, Phosphorus, and

Potassium. The multivariate analysis underscores the collective significance of these parameters, with Organic Carbon, Electrical Conductivity, and Phosphorus being significant drivers of clustered variation. Ward's method identifies three distinct clusters, showcasing internal homogeneity and external heterogeneity, providing valuable insights for soil management and agricultural planning.

Maniraj and Maran (2022) in their paper, clustering approaches are analyzed for skin lesion segmentation using dermoscopic images. A widely used machine learning approach for image segmentation is Centroid-based clustering (CBC). Fuzzy C-Means Clustering (FCMC), and Expectation-Maximization (EM)–E&M step algorithm. The difference between CBC and FCMC lies in the partitioning method. The former uses hard partitioning, and the latter uses a variable degree of membership. In the EM algorithm, statistical methods are employed for distance calculation whereas, in CBC, the Euclidean distance measure is used. The segmentation results of individual clustering approaches are combined to get the refined skin lesion. Results show that the combined segmentation provides promising results for skin lesion segmentation in comparison with CBC, FCMC, and EM- M step algorithm.

Luo et al. (2023) explained a novel approach to analyzing the COVID-19 pandemic's patterns using time-series clustering. The study identifies four distinct patterns of daily new cases and deaths across various countries by employing dynamic time-warping distance and hierarchical clustering. The findings suggest that while geographic factors play a significant role, the age structure of populations also influences cluster formation. This innovative method offers valuable insights for understanding pandemic dynamics and informs future research in this critical area.

Humretha et al. (2023) utilized K-Means clustering to analyze Banana crop data, aiding yield prediction. Time series data from 2010-2020 sourced from Chennai's Directorate of Economics and Statistics are analyzed. Results reveal significant relationships between environmental factors and Banana production, with K-Means clustering identifying three distinct productivity clusters. The findings offer actionable insights for farmers to optimize crop planting strategies, enhancing overall productivity and yield. This study provides valuable insights into leveraging clustering techniques for agricultural decision-making and crop yield prediction.

3. Materials and Methods

3.1 Material

Fertilizer consumption data was obtained from the World Bank's World Development Indicators database (https://data.worldbank.org). The dataset includes fertilizer consumption data for 41 countries from 1961 to 2021 (The countries are identified by their three-letter ISO country codes). The data measures the quantity of plant nutrients (nitrogenous, potash, and phosphate fertilizers) used per unit of arable land. Traditional nutrient sources like animal and plant manure are not included. It is essential to acknowledge that data collection methods might vary by

country. While some compile data based on the calendar year (January-December), others might utilize a split-year approach. R and Minitab will use statistical software capable of performing cluster analysis, MANOVA, and ANOVA and calculating inter- and intra-cluster distances.

3.2 Methods

3.2.1 Analysis of Variance

An ANOVA technique (Rao, 1952) tests the significance of the variation in each parameter between the countries. We use the following model,

$$y_{ij} = \mu + v_i + e_{ij} \tag{1}$$

i= 1, 2, 3... 60; j=1, 2, 3 … 41

where y_{ij} is the status of the fertilizer consumption in the jth countries of the ith sample, μ is the average status, v_i is the status in the ith country, and e_{ij} is a random error that follows a normal distribution with mean zero and constant variance σ^2.

3.2.2 Multivariate Analysis of Variance

To test the significance of variation among all the different years considered simultaneously, a MANOVA technique (Johnson and Wichern, 2002) is employed. The MANOVA model for comparing the population mean vectors (g=41) is as follows.

$$Y_{ij} = \mu + V_i + E_{ij} \tag{2}$$

i= 1, 2, 3... 41; j=1, 2, 3 … 61

where E_{ij} a vector of random errors is distributed as $N_p(0, \sum)$ (p=1, 2, 3, 4, 5). Here, the parameter vector μ is the overall mean and V_i represents the model's status in (2); each component of the observation vector Y_{ij} satisfies the univariate model (1), and the variance-covariance matrix $\sum$ is the same for all populations.

3.2.3 Cluster Analysis

To address the within-variability of the country's mean values, different year fertilizer consumption values are converted into uncorrelated variables using the pivotal condensation method (Rao, 1952). The transformed uncorrelated variables are used to group the countries with Ward's method, which involves the squared Euclidean distance method. The optimum number of clusters is calculated based on cluster selection criteria enumerated in Aldenderfer and Blashfield (1984).

3.2.4 Inter and Intra Cluster Distance

After forming the clusters, inter and intra-cluster D^2 values are calculated using the averaged individual D^2 values. The square root of this D^2 is used to indicate inter and intra-cluster distances. The cluster means for all characters are computed using the character means for the countries included in the clusters.

3.2.5 Estimation of intra and inter-cluster variance for different characters

An unweighted analysis of variance using the mean values of different characters is implemented (Rao, 1952). The structure of the variance analysis is given below.

Variation type	Degrees of freedom	Mean squares	Expected mean squares
Between countries	(k-1)	MSB (say)	$\sigma_w^{\,2} = m\sigma_b^{\,2}$
Within countries	$\sum_{i=1} n_i - k$	MSW (say)	$\sigma_w^{\,2}$

MSB = Mean square between the countries; MSW = Mean square within-cluster; k = number of clusters; n_i = number of countries in the i^{th} cluster; m is the harmonic mean based on the number of countries in each cluster.

The estimates of inter- and intra-cluster variances (i.e., $\hat{\sigma}_w^{\,2}$ and $\hat{\sigma}_b^{\,2}$) are obtained for each cluster using the mean squares. The ratio of the inter-cluster variances to the total variances is obtained as follows.

$$R^2 = \frac{\hat{\sigma}_b^{\,2}}{\hat{\sigma}_b^{\,2} + \hat{\sigma}_w^{\,2}}$$

The inter-cluster co-efficient of variations is calculated as follows.

$$CV_b = \frac{\hat{\sigma}_b}{\bar{X}} \times 100$$

Note that $\overline{X}$ this is the general meaning of the character.

4. Result and Discussion

The statistical analysis results of fertilizer consumption data from 1961 to 2021, collected from 41 countries, are presented below.

Table 1: Descriptive Statistics

S. No.	Country	Mean	St. Dev.	S. No.	Country	Mean	St. Dev.
1	AGE	231.30	223.50	22	JPN	114.41	22.44
2	ARB	55.95	40.23	23	KOR	141.70	190.70
3	ARG	402.60	243.70	24	MAR	56.18	53.67
4	AUS	197.80	81.80	25	MEX	173.90	104.00
5	BGD	227.30	146.30	26	NAC	72.99	8.85
6	BGR	80.13	30.79	27	NLD	33.33	17.57
7	BRA	308.40	152.30	28	NOR	31.70	10.51
8	CAN	23.73	8.42	29	NZL	188.72	65.06
9	CHINA	124.58	25.69	30	PAK	157.34	59.18
10	COL	428.50	219.40	31	PHL	510.00	939.00
11	DEU	56.77	4.67	32	POL	113.71	43.00
12	DZA	187.60	140.30	33	PRT	110.91	26.73
13	ESP	98.17	17.56	34	ROU	62.89	26.00
14	EUU	88.16	5.15	35	SAS	159.67	32.30
15	FIN	100.03	38.97	36	TUN	18.55	13.61
16	GBR	185.90	96.60	37	TUR	186.50	80.40
17	GRC	165.80	108.40	38	USA	100.07	16.80
18	HUN	166.25	27.07	39	VNM	447.20	299.50
19	IND	158.68	33.95	40	ZAF	144.67	62.64
20	ISR	7.62	5.20	41	ZWE	238.90	167.40
21	ITA	186.90	126.30				

Data on fertilizer consumption for different countries from 1961 to 2021 is reported in Table 1. Each country is assigned a serial number for reference. The countries are identified by their three-letter ISO country codes (e.g., AFE for Afghanistan, ARB for Arab World, ARG for Argentina). This mean column represents the average fertilizer consumption in kilograms per hectare over each country's specified period. For example, Argentina (ARG) had an average fertilizer consumption of 402.6 kilograms per hectare over the period.

The second column indicates the variability or dispersion of fertilizer consumption data around the mean for each country. A higher standard deviation suggests more significant variability in fertilizer consumption over the years. Observe substantial variations in mean fertilizer consumption and variability across different countries. Some countries like Argentina and Colombia show relatively high average fertilizer

consumption, while others like Israel and the Netherlands have lower average consumption levels. Additionally, the standard deviations reveal that fertilizer consumption trends vary widely among countries, suggesting differences in agricultural practices, economic development, and other factors influencing fertilizer use over time.

Table 2: Characteristics of ANOVA

Source of Variation	Sum of Square	Degrees of Freedom	Mean Sum of Square	F test	p-value
Between Groups	33795434.30	40	844885.90	25.75	0.000
Within Groups	80703925.09	2460	32806.47		
Total	114499359.40	2500			

The analysis of variance results presented in Table 2 shows significant variation in the data, with a high F-test value of 25.75 and a p-value of 0.00 for the between-group variation, indicating substantial differences in the means of groups. The sum of squares between groups is much more significant than within groups, suggesting that the majority of the variability in the data is explained by differences between groups rather than within them. Finally, this indicates significant differences in the means of the groups being compared, likely due to some factor(s) influencing the observed values.

The MANOVA results presented in Table 3 indicate a significant effect of the parameter (fertilizer consumption) on the observed data, as evidenced by perfect Pillai's Trace of 0.9894, suggesting complete discrimination among groups. Wilks' Lambda also supports this, with a close-to-zero value of 0.010, indicating strong evidence against the null hypothesis. Hotelling's Trace and Roy's Largest Root further confirm the substantial impact of fertilizer consumption on the dataset, as both values are considerably large, indicating a high level of discrimination among groups. Overall, the MANOVA results emphasize the influential role of fertilizer consumption in explaining the variation observed in the dataset.

Table 3: Characteristics of multivariate statistics

Statistics	Value	Sig.
Pillai's Trace	0.9894	0.000
Wilks' Lambda	0.010	0.000
Hotelling's Trace	11415.58	0.000
Roy's Largest Root	10416.56	0.000

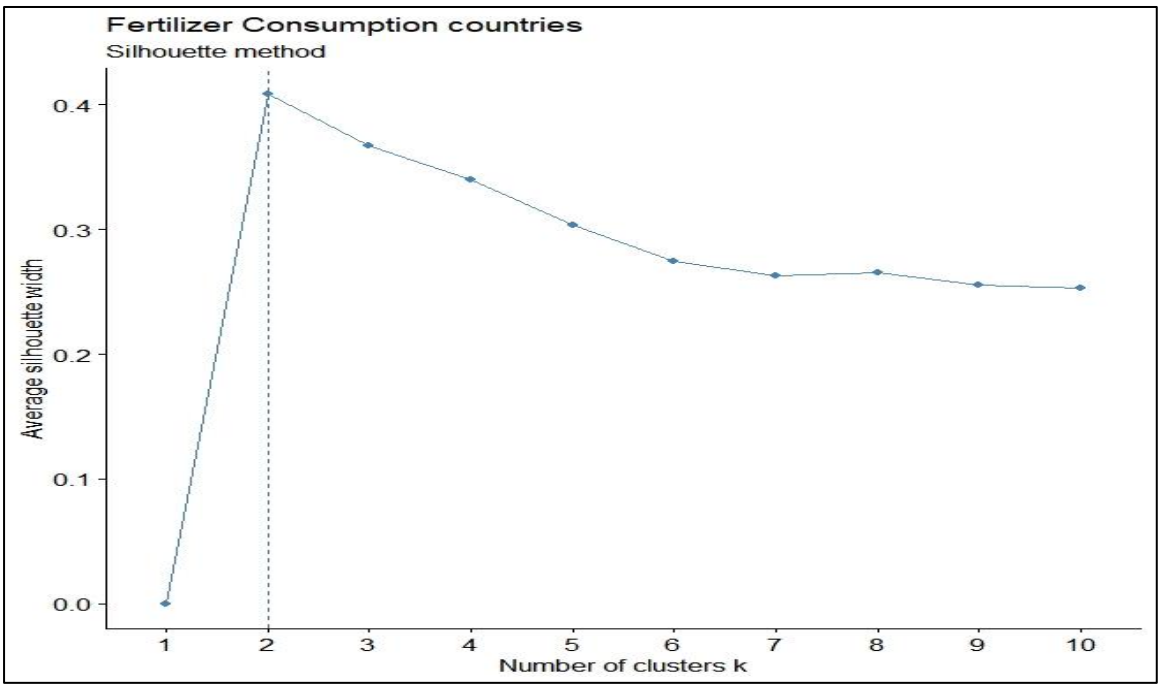

Figure 1: Optimum no of clusters based on the Silhouette method

Figure 1 based on the Silhouette method suggests that the optimum number of clusters is 2, which indicates that the data points are better grouped into two distinct clusters than any other number of clusters. This finding implies a clear separation between the clusters, with well-defined boundaries and minimal overlap. Each data point is closer to other points within its cluster than points in different clusters, resulting in higher silhouette scores. Therefore, utilizing two clusters provides the most cohesive and meaningful data partitioning, allowing for a more straightforward interpretation and analysis of underlying patterns or structures within the dataset.

The Dendrogram produced by Ward's method and the distributions of countries in two clusters are depicted in Figure 2. The distribution of 41 countries in two clusters and the cluster means for each country's fertilizer consumption mean values are presented in Table 4. (Cluster 1 mean value is 216.29 and Cluster mean value is 115.34).

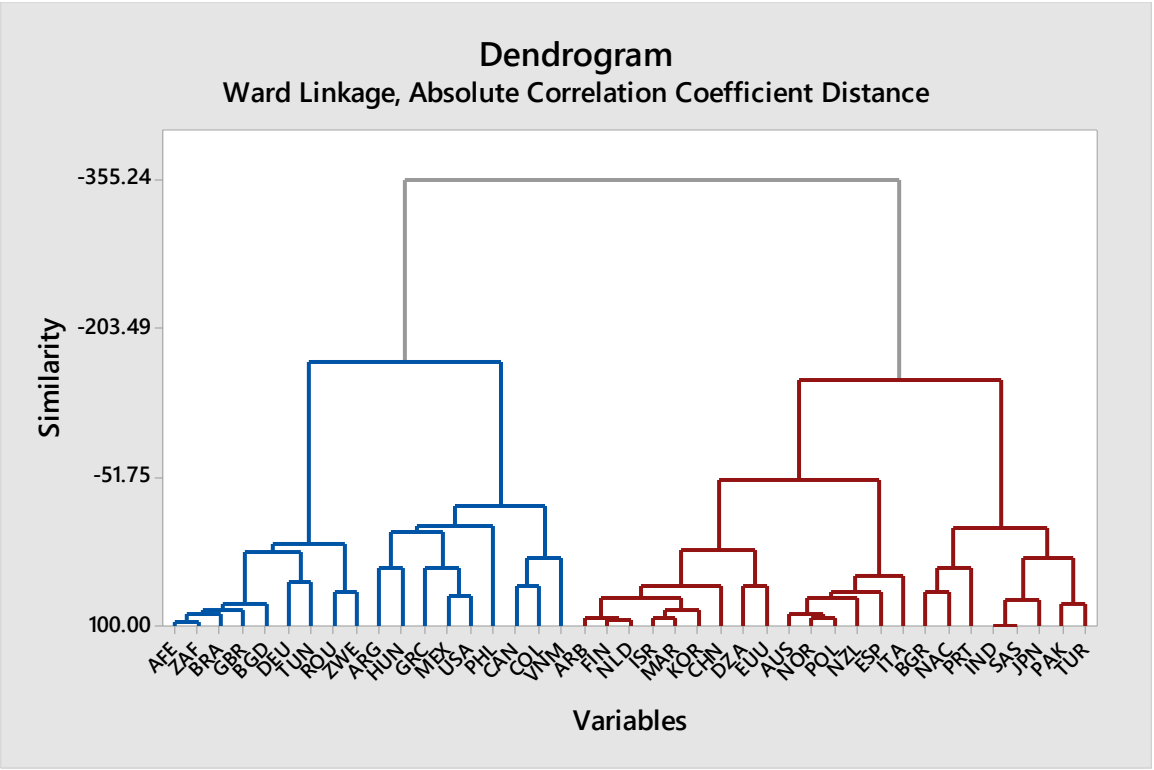

Figure 2: Ward's linkage Dendrogram was formed by the different countries

Table 4: Characterization of cluster composition and mean values based on Ward's method

Cluster	Countries	Mean values
1	**AFG** – Afghanistan, **ARG** – Argentina, **BGD** – Bangladesh, **BRA** – Brazil, **CAN** – Canada, **COL** – Colombia, **DEU** – Germany, GBR - United Kingdom, **GRC** – Greece, **HUN** – Hungary, **MEX** – Mexico, **PHL** – Philippines, **ROU** – Romania, **TUN** – Tunisia, **USA** - United States, **VNM** – Vietnam, **ZAF** - South Africa, **ZWE** – Zimbabwe	216.29
2	**ARB** - Arab World, **AUS** – Australia, **BGR** – Bulgaria, **CHN** – China, **DZA** – Algeria, **ESP** – Spain, **EUU** - European Union, **FIN** – Finland, **IND** – India, **ISR** – Israel, **ITA** – Italy, **JPN** – Japan, **KOR** - South Korea, **MAR** – Morocco, **NAC** - North America, **NLD** – Netherlands, **NOR** – Norway, **NZL** - New Zealand, **PAK** – Pakistan, **POL** – Poland, **PRT** – Portugal, **SAS** - South Asia, **TUR** - Turkey	115.34

Table 5: Characteristics of Mean intra and inter-cluster D^2 values obtained using ward's method

	Cluster 1	**Cluster 2**
Cluster 1	0.0000	8.6297
Cluster 2	8.6297	0.0000

Table 5 displays the mean intra and inter-cluster D^2 values obtained using Ward's method for clustering. Intra-cluster values on the diagonal represent the average dissimilarity within clusters, with Cluster1 having a mean D^2 value of 0.0000 and Cluster2 having a mean D^2 value of 0.0000, indicating that points within clusters are very similar. Inter-cluster values, off the diagonal, indicate the average dissimilarity between clusters, with a value of 8.6297 between Cluster 1 and Cluster 2, reflecting the dissimilarity between these clusters. This suggests that Ward's method effectively partitions the data into clusters with relatively low within-cluster dissimilarity and higher between-cluster dissimilarity.

5. Conclusion

This analysis of fertilizer consumption data from 1961 to 2021 for 41 countries revealed significant variations across nations. The data suggests that fertilizer consumption patterns differ considerably, with some countries exhibiting high average use and others demonstrating lower consumption levels. Statistical tests confirmed that these differences are not random. Multivariate analysis further highlighted fertilizer consumption as a significant factor influencing the observed variations. Additionally, cluster analysis effectively grouped countries into two categories based on their consumption patterns. These findings suggest the importance of considering fertilizer consumption within the context of individual countries. Further exploration into the reasons behind these variations, including agricultural practices, economic development, and government policies, could provide valuable insights for optimizing fertilizer use and improving agricultural sustainability. Analyzing trends over time and comparing consumption patterns with crop yield data could offer an even deeper understanding of fertilizer use effectiveness across different regions.

References

[1] Aldenderfer, M. S. and Blashfield, R. K. (1984). Cluster analysis. SAGE Publications, Inc., https://doi.org/10.4135/9781412983648

[2] Ferreira, K. R., Santos, L. A. and Picoli, M. C. A. (2019). Evaluating Distance Measures for Image Time Series Clustering in Land Use and Cover Monitoring. In MACLEAN@ PKDD/ECML.

[3] Gangopadhyay, S.K., Bandyopadhyay, S., Mukhopadhyay, S., Nayak, D.C., Sahoo, A.K. and Singh, S.K. (2020). Soil Nutrient Information System (SNIS) in Middle Indo-Gangetic Plain of Bihar, India – A Tool for Land Use Planning. Journal of the Indian Society of Soil Science, 68(1), pp 25-33 (2020) DOI: 10.5958/0974-0228.2020.00003.1.

[4] Hayatu, I. H., Mohammed, A., Ismaâ, B. A. and Ali, S. Y. (2020). K-Means clustering algorithm-based classification of soil fertility in North West Nigeria. FUDMA Journal of Sciences, 4(2), pp.780–787.

[5] Humretha, S., Gangaiselvi, R., Kannan, B., Rani, C. I. and Dheebakaran, G. (2023). Implementation of K-Means Clustering Technique in Banana Production of Tamil Nadu, India. International Journal of Environment and Climate Change, 13(9), pp.1439–1446.

[6] Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. Prentice-Hall of India Pvt.Ltd, New Delhi.

[7] Li, W., Liu, L. and Le, J. (2007). Clustering streaming time series using CBC. In Computational Science–ICCS 2007: 7th International Conference, Beijing, China, May 27-30, 2007, Proceedings, Part III 7 (pp. 629-636). Springer Berlin Heidelberg.

[8] Luo, Z., Zhang, L., Liu, N. and Wu, Y. (2023). Time series clustering of COVID-19 pandemic-related data. Data Science and Management, 6(2), pp,79–87.

[9] Maniraj S.P. and Maran, S. P. (2022). Analysis of CBC and FCMC Clustering Approaches for Skin Melanoma Segmentation using Dermoscopic Images. Research Journal of Pharmacy and Technology 2022; 15(10):4807-1. doi: 10.52711/0974-360X.2022.00807.

[10] Rajarathinam, A., Khokhar, A.N. and Dixit, S.K. (2010). Statistical modeling on the grouping of villages of Kadana taluka based on Soil parameters. Int. J. Agricult. Stat. Sci., Vol. 6, No. 2, pp. 423-430.

[11] Rajarathinam, A. and Ramji, M. (2021), villages clustering based on soil parameters. International journal of modern agriculture, 10(2), pp.2625–2635.

[12] Rao, R.C. (1952). Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York, USA.