ISSN Online: 2756-3898, ISSN Print: 2714-500X https://doi.org/10.47514/phyaccess.2025.5.2.008

Cluster Analysis on Radioactivity Levels of K-40, Ra-226, and Th-232 at Chikun Environment of Kaduna State, Nigeria, Using the R Package

Handan T Elisha[1], Sarki S Habila[2], Shirley O Yakubu[2], Umar Isah[3], and Davou H Daze[4]

[1] Department of Statistics, Nuhu Bamalli Polytechnic Zaria, Kaduna State, Nigeria
[2] Department of Science and Laboratory Technology, Federal Polytechnic, Kaltungo, Gombe State, Nigeria
[3] Department of Medical Imaging Technology, Shehu Idris College of Health Sciences and Technology, Makarfi, Kaduna State, Nigeria
[4] Department of General Studies, Plateau State College of Nursing Sciences, Vom, Plateau State, Nigeria

Corresponding E-mail: ethandan14@nubapoly.edu.ng

Received 02-10-2025
Accepted for publication 07-11-2025
Published 10-11-2025

Abstract

Exposure to ionizing radiation from both natural and anthropogenic sources poses serious environmental and health concerns. This study used hierarchical clustering to investigate the spatial distribution of naturally occurring radionuclides K-40, Ra-226, and Th-232 in soil samples collected from 14 locations within Chikun Local Government Area (LGA), Kaduna State, Nigeria. Using the R statistical package, single and complete linkage methods were implemented on Euclidean distance matrices to identify natural groupings among the sampling sites. The optimal number of clusters, determined through the elbow and silhouette methods, was found to be two (2), with further sub-groupings evident in the complete linkage dendrogram. The first cluster comprises the KRPC1 and KRPC2 sampling sites, situated near the Kaduna Refining and Petrochemical Company. These sites exhibited the highest mean activity concentrations of the measured radionuclides, with values of 106.14 Bq/kg for K-40, 265.84 Bq/kg for Ra-226, and 827.52 Bq/kg for Th-232, respectively. The second cluster comprised residential neighborhoods and nearby waste disposal sites, demonstrating comparatively lower activity concentrations of the radionuclides. This spatial distribution suggests that refinery operations exert a pronounced influence on localized radioactivity levels, while the surrounding residential and dumpsite areas exhibit only residual contamination. These results underscore the effectiveness of cluster analysis as a tool for interpreting environmental radioactivity patterns and reinforce the need for sustained monitoring and public health safeguards in industrialized settings.

Keywords: Cluster Analysis; Dendrogram; Linkage; Radioactivity; R package.

I. Introduction

Ionizing radiation is a natural environmental hazard that can cause significant physiological (somatic) and genetic damage. With the discovery of atomic and nuclear energy and advancements in technology, the use of man-made radiation, especially in power production, along with the increased likelihood of human exposure to harmful radiation, has continued to rise [1].

Several studies have been conducted to assess the level of radioactivity of radioactive elements, both within and outside Nigeria, to understand the sources, effects, and risks associated with ionizing radiation. These studies have emphasized the importance of having accurate information about the health effects of exposure to high levels of background radiation [2].

Reference [3] studied the impact of environmental radiation on the incidence of cancer and birth defects in areas with high natural radioactivity and found that natural radioactive elements released from crystalline rocks can enter the food chain, including soil, plants, animals, and humans, leading to increased radiation exposure and potential health effects.

Natural occurring radioactivity in the Earth's crust may be categorized into virgin and modified natural sources. Virgin sources are of cosmogenic or primordial (terrestrial) origin and have been present since the Earth's early geological history. In contrast, modified natural sources result from anthropogenic activities such as mining, fossil fuel combustion, fertilizer production, and the use of naturally derived materials in construction. Gamma emissions from radionuclides such as K-40 and members of the Th-232 decay series, along with their progeny, constitute the principal external source of irradiation to the human body [4].

The application of multivariate analytical techniques has proven valuable for researchers, as it facilitates the identification of underlying patterns and relationships among variables, as well as associations between variables and the corresponding sampling sites [5]. Comparable observations were reported in [6], where the authors employed the R statistical software to investigate the relationships between metal concentrations and their spatial distribution, as well as the associations between the metals and radiological activity levels.

Other multivariate approaches have also been effectively applied within the Nigerian context. For example, [7] examined soils in the vicinity of a coal-fired cement plant and attributed the elevated activity concentrations of natural radionuclides to industrial emissions and combustion residues. Similarly, [8] reported increased levels of uranium- and thorium-series radionuclides in peri-urban Emure-Ekiti, southwestern Nigeria, linking the observed spatial variations in radioactivity to both anthropogenic activities and underlying geological formations.

Reference [1] assessed the activity concentrations of K-40, Ra-226, and Th-232 in soil samples collected from the Chikun area of Kaduna Metropolis using gamma-ray spectrometry and reported elevated radioactivity levels at several locations. However, given that many of these sites are characterized by waste deposits from the Kaduna Refining and Petrochemical Company (KRPC), alongside other industrial and domestic activities, the study did not examine whether shared environmental or anthropogenic factors might explain the spatial variations observed.

The present study, therefore, seeks to determine the presence of natural groupings among the sampling locations in Chikun Local Government Area (LGA), Kaduna State, Nigeria, based on the radionuclide activity concentrations reported in [1]. To achieve this, Complete Linkage and Single Linkage hierarchical clustering algorithms were applied using the R statistical software.

II. Materials And Methods

A. Materials

1) R Package

The primary analytical tool employed in this study was the R statistical software (version 4.5.1), an open-source programming environment recognized for its robustness, versatility, and broad accessibility [9].

2) Data Source

The dataset utilized in this study was derived from the work reported in [1], which determined the activity concentrations of K-40, Ra-226, and Th-232 in soil samples collected from the Chikun area of Kaduna Metropolis using gamma-ray spectrometry (see Table I).

Table I. Result of K-40, Ra-226 and Th-232 Obtained from Soil Samples in Chikun L.G.A [1]

Sample ID	K-40	Ra-226	Th-232
KP1	228.3	28.2	28.8
KP2	517.6	19.6	50.7
KP3	122.2	89.5	147.5
SBR1	29.40	48.3	106.2
SBR2	152.6	33.0	105.1
SBR3	378.1	36.2	37.9
GN1	118.5	73.1	22.1
GN2	175.9	60.1	24.2
GN3	144.9	17.6	84.7
RIDO1	63.00	40.4	107.8
RIDO2	97.20	16.0	25.1
RIDO3	211.0	35.2	133.8
KRPC1	889.73	108.6	243.1
KRPC2	765.3	103.7	288.6

B. Methods

1) Data Importation

The acquired dataset was exported from Microsoft Excel as a comma-separated values (CSV) file before being imported into R for subsequent analysis.

2) Cluster Analysis

Cluster analysis, a data exploration (mining) tool for dividing a multivariate dataset into “natural” clusters (groups) which may exist in a dataset, was used to sort the different objects (location) into groups such that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise [10]. To achieve this, distance measures were computed to construct a proximity matrix from which a specific iterative clustering algorithm is applied. In this case, the single and complete linkage algorithms were employed to comparatively assess areas of similarity by both algorithms.

a) Distance Measures

To initiate a cluster analysis, a proximity matrix, which represents the strength of the relationship in terms of similarities/distance (dissimilarities) between pairs of rows in $y_{p\times n}$ is constructed. As in the case of continuous (interval, ratio scale) variables, the Euclidean distance between two objects, the most common dissimilarity measure, was used [11].

Given an *(n × p)* matrix **Y** with *(*1 × *p)* row vectors y_i, the Minkowski measure between two rows y_r *and* y_s is defined as;

$$d_{rs} = \left[\sum_{j=1}^{p}\left(y_{rj} - y_{sj}\right)^{r}\right]^{1/r} \qquad (1)$$

With this, the study adopts the Euclidean distance measure, which is the most popular variant of the Minkowski measure. When $r = 2$, this gives (2).

$$d_{rs} = \left[\sum_{j=1}^{p}\left(y_{rj} - y_{sj}\right)^{2}\right]^{1/2}$$

$$d_{rs} = \sqrt{[(Y_r - Y_s)'(Y_r - Y_s)]} \qquad (2)$$

The *(n × n)* data matrix $\boldsymbol{D} = [d_{rs}]$ is called the Euclidean distance matrix, and given two objects Y_r *and* Y_s in a p-dimensional space, a dissimilarity measure satisfies the following conditions.

$$d_{rs} \geq 0 \text{ for all objects } y_r \text{ and } y_s \qquad (3)$$

$$d_{rs} = 0 \text{ if and only if } y_r = y_s \qquad (4)$$

$$d_{rs} = d_{sr} \qquad (5)$$

b) Clustering Algorithms

Given the proximity matrix $D_{n\times n} = [d_{rs}]$, the steps for the agglomerative hierarchical clustering algorithm as executed in this research are as follows [10,11]:

Step 1. Begin with n clusters, each containing only a single object.

Step 2. Search the dissimilarity matrix D for the most similar pair. Let the pair chosen be associated with the element d_{rs}, so that objects r and s are selected. In the case of single linkage, the minimum value that connects a pair is considered, whereas the largest value is considered in the case of complete linkage.

Step 3. Combine objects r and s into a new cluster (rs) employing some criterion, and reduce the number of clusters by 1 by deleting the row and column for objects r and s. Calculate the dissimilarities between the cluster (rs) and all remaining clusters, using the criterion, and add the row and column to the new dissimilarity matrix.

Step 4. Repeat steps 2 and 3, $(n - 1)$ times until all objects form a single cluster. At each step, identify the merged clusters and the value of the dissimilarity at which the clusters are merged.

The agglomerative hierarchical procedures were adopted, viz, the single and complete linkage algorithms.

i) Single Link (Nearest-Neighbour) Method

To implement the single-link method, we combine objects in clusters using the minimum dissimilarity between clusters. Then new distances between R and S are calculated using (6).

$$d_{(R)(S)} = min\{d_{rs} \mid \text{r} \in \text{R and s} \in \text{S}\} \qquad (6)$$

ii) Complete Link (Farthest-Neighbour) Method

For the complete link procedure, we also combine objects in clusters using the minimum dissimilarity between clusters. Then, new distances between clusters R and S are calculated using the rule: -

$$d_{(R)(S)} = max\{d_{rs} \mid \text{r} \in \text{R and s} \in \text{S}\} \qquad (7)$$

c) Cluster Optimality

Normally, for a hierarchical clustering algorithm, a visual inspection or a priori information might suffice in determining the number of clusters constituted. However, to quantitatively assess the adequacy of the clustering structure obtained through the hierarchical methods, two validation measures were computed.

i) Elbow Plot

This is a two-dimensional plot of the Within-Cluster Sum of Squares (WCSS) against increasing k-values and looking for a point (the elbow) where the improvement slows down [12].

ii) Silhouette Index Plot

The silhouette index $(s(i)\ )$ measures the degree of separation between clusters and compactness within clusters and is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \qquad (8)$$

Where $a(i)$ represents the average dissimilarity of sample *i* with all other samples in the same cluster, and $b(i)$ is the minimum average dissimilarity of *i* with samples in other clusters.

Hence, our interest was to plot the $s(i)$ values against increasing k-values and spot the highest point to mark the optimal number of clusters.

III. Results And Discussion

This section presents the result of the cluster analysis performed on the data obtained from [1]. More importantly, analyzing the result of dendrograms produced by the single and complete linkage algorithms, given the computed Euclidean distance measures as shown in the proximity matrix in Table II. The Elbow plot and silhouette index plot for cluster optimality are depicted in Figs. 1 and 2, respectively. The elbow plot shows the presence of two elbow-like bends at $k = 2$ and $k = 3$, with the steeper bend at $k = 3$, implying that the optimal number of clusters could be two or three. Hence, given the unclear nature of this check, further investigation using the silhouette index plot in Fig. 2 clearly reveals an optimal cluster number of $k = 2$.

Table II. Proximity Matrix for the Locations in Chikun LGA Computed from their Radioactive Level [1].

Location	GN1	GN2	GN3	KRPC1	KRPC2	KP1	KP2	KP3	SBR1	SBR2	SBR3	RIDO1	RIDO2
GN2	58.89												
GN3	87.73	80.17											
KRPC1	803.05	748.21	766.91										
KRPC2	700.22	647.46	658.70	132.58									
KP1	118.81	61.52	100.96	699.91	601.30								
KP2	403.68	345.11	374.25	428.28	353.59	290.26							
KP3	126.52	137.66	98.13	773.70	658.55	170.60	413.03						
SBR1	125.01	168.30	121.43	873.24	760.19	214.37	492.18	109.61					
SBR2	98.28	88.44	26.69	753.74	643.48	107.59	369.27	76.90	124.15				
SBR3	262.69	204.07	238.58	555.98	466.19	150.29	141.07	283.44	355.53	235.32			
RIDO1	107.21	141.86	88.10	840.50	727.96	183.61	458.64	86.55	34.55	89.95	322.79		
RIDO2	61.02	90.22	76.35	827.17	723.52	131.72	421.19	144.94	110.53	98.78	281.92	92.76	
RIDO3	149.90	117.75	84.20	691.38	579.57	106.65	318.04	104.98	184.15	65.11	192.67	150.36	158.54

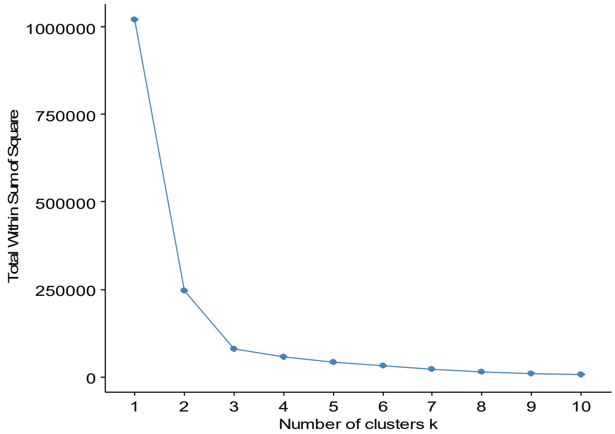

Fig. 1. Elbow plot for the number of clusters.

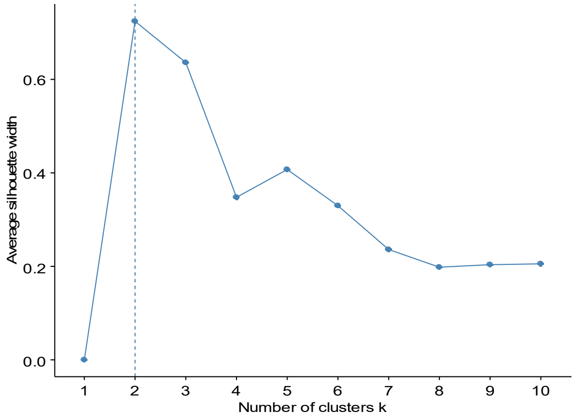

Fig. 2. Silhouette index plot.

Fig. 3(a) and (b) present the dendrograms that were generated using the single and complete linkage algorithms, respectively. As recommended by the silhouette plot, the two clusters consist of KRPC1 and KRPC2 as cluster 1, with mean activities of K-40, Ra-226, and Th-232 at 106.14 Bq/kg, 265.84 Bq/kg and 827.52 Bq/kg, respectively, while every other location belongs to cluster 2.

Analysis of Fig. 4, which provides a visual representation of elemental levels, reveals that Cluster 1 exhibits the highest concentrations across all elements. This finding is unsurprising, as the locations within Cluster 1 are situated around the Kaduna Refinery, which likely contributes to the elevated radiation levels. This observation agrees with [7], who attributed high natural radionuclide concentrations around a coal-fired cement plant to increased industrial emissions.

To further elucidate the similarity patterns within Cluster 2, comparative visual examination of Fig. 3(a) and Fig. 3(b) indicates that, at linkage heights of 150 and 200, respectively, Cluster 2 further resolves into three distinct sub-clusters:

- Sub-cluster A (SBR1 and RIDO1) corresponds to the dumpsite locations, exhibiting mean activity concentrations of 46.2 Bq/kg (K-40), 44.35 Bq/kg (Ra-226), and 107 Bq/kg (Th-232). The elevated levels observed here are likely attributable to the accumulation of heterogeneous waste materials and enhanced organic matter decomposition processes.
- Sub-cluster B (KP2 and SBR3) comprises residential areas situated in proximity to the refinery. These sites display intermediate radionuclide activities, with mean concentrations of 447.85 Bq/kg (K-40), 27.9 Bq/kg (Ra-226), and 44.3 Bq/kg

(Th-232), suggesting secondary dispersion and localized environmental redistribution associated with the refinery's operations.

• Sub-cluster C (KP1, RIDO2, GN1–GN3, SBR2, and RIDO3) represents residential zones located farther from direct industrial influence, characterized by comparatively lower activity concentrations of 156.33 Bq/kg (K-40), 44.09 Bq/kg (Ra-226), and 71.41 Bq/kg (Th-232). These values indicate that radioactivity levels here are predominantly influenced by routine domestic and commercial activities rather than industrial emissions.

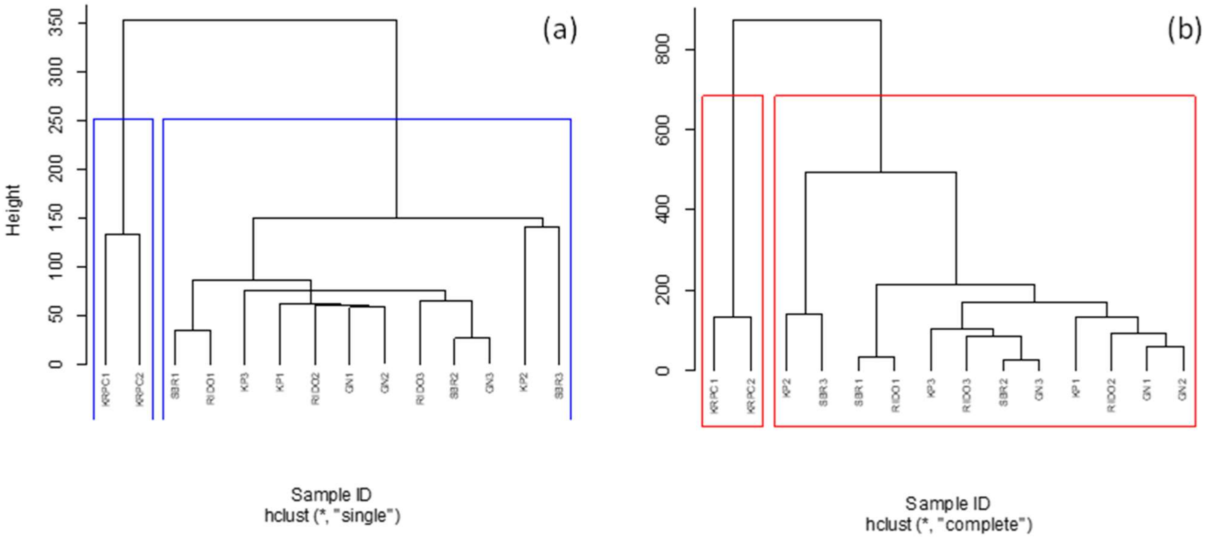

Fig. 3. Dendrograms from the (a) Single and (b) Complete Linkage Algorithms for Chikun LGA.

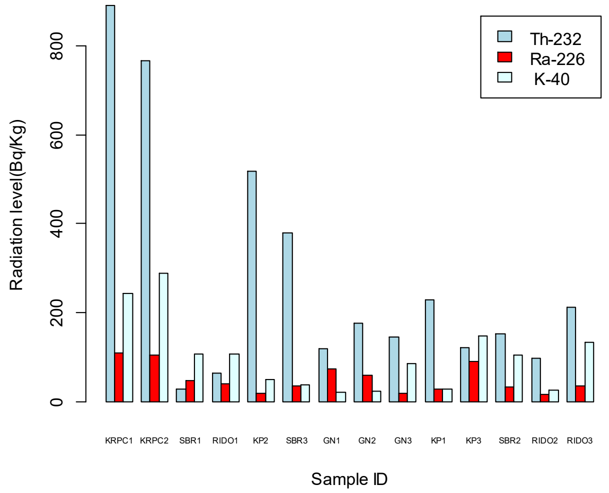

Fig. 4. Barchart showing levels of K-40, Ra-226 and Th-232 in Chikun LGA [1].

IV. Conclusion

This study examined the natural grouping of sampling locations within Chikun Local Government Area (LGA), Kaduna State, Nigeria, based on measured activity concentrations of K-40, Ra-226, and Th-232. Radioactivity data were obtained from fourteen locations across the Chikun environs. Cluster analysis was subsequently applied, and the outcomes of both single-linkage and complete-linkage methods indicate that operations of the Kaduna Refinery exert a discernible influence on radionuclide levels in its immediate surroundings, with diminishing yet notable effects extending to more distant locations. The observed increase in human settlement and activities around these affected zones is therefore a cause for concern and underscores the need for targeted environmental management and public health intervention strategies.

Appendix

R Source Code

```
# DATA IMPORTATION
# ----------------------------------------
radiation.data <- read.csv(file.choose(), row.names = 1)
#Note: all object assignments in the R code block are
done specifically to this research.

# PROXIMITY MATRIX CALCULATION
# ----------------------------------------
# Compute Euclidean distance matrix and Display
proximity matrix
proximity.matrix <- dist(radiation.data, method =
"euclidean")
print(proximity.matrix)

# HIERARCHICAL CLUSTERING - SINGLE LINKAGE
# -----------------------------------------
# Execute single linkage clustering
single.hc <- hclust(proximity.matrix, method = "single")
# Plot single linkage dendrogram
plot(single.hc,
     hang = -1,
     cex = 0.6,
     xlab = "Sample ID",
     main = "Single Linkage Dendrogram")
# Demarcate clusters (k=2)
rect.hclust(single.hc, k = 2, border = "blue")

# HIERARCHICAL CLUSTERING - COMPLETE LINKAGE
# ---------------------------------------
```

```
# Execute complete linkage clustering
complete.hc <- hclust(proximity.matrix, method =
"complete")

# Plot complete linkage dendrogram
plot(complete.hc,
     hang = -1,
     cex = 0.6,
     xlab = "Sample ID",
     main = "Complete Linkage Dendrogram")
# Demarcate clusters (k=2)
rect.hclust(complete.hc, k = 2, border = "red")
# CLUSTER OPTIMALITY - ELBOW METHOD
# ---------------------------------------
library(factoextra)
fviz_nbclust(radiation.data, hcut, method = "wss") +
  labs(subtitle = "Elbow Method")
# CLUSTER OPTIMALITY - SILHOUETTE METHOD
# ---------------------------------------
fviz_nbclust(radiation.data, hcut, method =
"silhouette") +
 labs(subtitle = "Silhouette Method")

# BARCHART
# ---------------------------------------
barplot(t(radiation.data),
beside=TRUE,col=c("lightblue", "red", "lightcyan"),
legend.text=c("Th-232","Ra-226"," K-40"),
cex.names=0.4,
xlab="Sample ID",
ylab="Radiation level(Bq/Kg)",
args.legend = list(x = "topright", cex = 0.6))
```

References

[1] P. M. Gyuk, S. S. Habila, M. D. Dogara, N. Kure, H. I. Daniel, and T. E. Handan, "Determination of radioactivity levels in soil samples at Chikun environment of Kaduna metropolis using gamma ray spectrometry," *Sci. World J.*, vol. 12, no. 2, pp. 52–55, 2017.

[2] S. H. Sarki, S. Abew, M. A. Musa, and A. M. Ibrahim, "Determination of activity concentration level of ^{226}Ra, ^{40}K and ^{232}Th in soil within Igabi Local Government Area of Kaduna State, Nigeria," *Sci. World J.*, vol. 15, no. 1, pp. 113–118, 2020.

[3] A. Zlobina, I. Farkhutdinov, F. P. Carvalho, N. Wang, T. Korotchenko, N. Baranovskaya, and A. Farkhutdinov, "Impact of environmental radiation on the incidence of cancer and birth defects in regions with high natural radioactivity," *Int. J. Environ. Res. Public Health*, vol. 19, no. 14, p. 8643, 2022.

[4] United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), *Sources and Effects of Ionizing Radiation: UNSCEAR 2000 Report to the General Assembly, with Scientific Annexes*, vol. I, United Nations, New York, NY, USA, 2000.

[5] L. L. Harlow, *The Essence of Multivariate Thinking: Basic Themes and Methods*, 2nd ed. New York, NY, USA: Routledge, 2014.

[6] F. Caridi, A. F. Mottese, G. Paladini, L. Pistorino, F. Gregorio, S. Lanza, *et al.*, "Multivariate statistics, radioactivity and radiological hazard evaluation in marine sediments of selected areas from Sicily, Southern Italy," *J. Mar. Sci. Eng.*, vol. 13, no. 4, p. 769, 2025.

[7] M. T. Kolo, M. U. Khandaker, and H. K. Shuaibu, "Naturally radioactivity in soils around mega coal-fired cement factory in Nigeria and its implication on human health and environment," *Arab. J. Geosci.,* vol. 12, no. 15, 2019.

[8] O. J. Popoola, O. E. Olubi, O. S. Adewalure, and A. E. Raphael, "Elevated natural radionuclides in soils and stream sediments: pollution, spatial distribution, radiological hazards, and cancer risks in peri-urban Emure-Ekiti, southwest Nigeria." *Discover Soil,* vol. 2, no. 1, p. 70, 2025.

[9] A. Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, 1st ed., vol. 1. STHDA, 2017.

[10] N. H. Timm, *Applied Multivariate Analysis*. New York, NY, USA: Springer-Verlag, 2002.

[11] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[12] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. New Delhi, India: Pearson Education, 2016.