UDC 551.32

Characterization of Soil Using Multivariate Analysis and Machine Learning Approaches

W. M. Badawy[1,2], E. S. Mohamed[3], F. I. El-Agawany[4], Maksim G. Blokhin[5]

[1]*Frank Laboratory of Neutron Physics, Joint Institute for Nuclear Research, Dubna, Russian Federation*

[2]*Radiation Protection and Civil Defense Department, Nuclear Research Center, Egyptian Atomic Energy Authority, 13759 Cairo, Egypt*

[3]*National Authority for Remote Sensing and Space Sciences (NARSS), 1564 Cairo, Egypt*

[4]*Department of Physics, Faculty of Science, Menoufia University, 32511 Shebin El-Koom, Egypt*

[5]*Far East Geological Institute of the Far Eastern Branch of the Russian Academy of Sciences, 690022 Vladivostok, Russian Federation*

The Nile Delta is Egypt's main agricultural hub and one of its most densely populated regions, making soil quality and metal pollution central to environmental and public-health management. Comprehensive, statistically robust geochemical assessments, however, have been limited and often focused on local sectors. Recent work on Nile Delta soils has integrated multi-element analyses with multivariate statistics, Bayesian background modeling, and machine learning, advancing beyond earlier studies of Nile sediments and coastal environments that relied mainly on conventional statistics. This short paper summarizes the main methods and findings of that integrated approach and outlines its broader implications for environmental monitoring in agricultural deltas (Fishar, 2018; Hamza, 2009).

Agricultural soils (53 samples) were collected from 10–15 cm depth across the Nile Delta, following standardized protocols, sieved (<2 mm), ground, and digested using mixed acids; Si was determined separately by a fusion–gravimetric procedure. Major elements were analyzed by ICP-AES and trace/REE by ICP-MS, with certified reference materials ensuring accuracy and reproducibility. The 55 element dataset was tested for normality, screened for outliers using Mahalanobis distance probability (Mahalanobis, 1936), and explored via Pearson correlations, variables were min–max scaled before multivariate analysis. Dimensionality reduction used PCA and t-SNE, followed by hierarchical agglomerative clustering evaluated with Silhouette and Davies–Bouldin indices. Bayesian inference (JAGS/MCMC) provided regional background values (posterior $\mu \pm \sigma$) for each element, serving as the basis for single pollution indices, enrichment factors, pollution load index, and total pollution index (Hinton and Roweis, 2003; Lê et al., 2008; Maaten, 2014).

The soils are dominated by crustal elements, with Si and Al as major components and Bi among the lowest trace elements. Comparison with upper continental crust values reveals only slight enrichments for a limited group of elements (Cd, Ti, Cu, P, V, Fe, Ni, Ca), consistent with agricultural and urban influences superimposed on natural backgrounds. REE patterns normalized to UCC and PAAS are broadly similar, while chondrite normalization highlights a europium anomaly that likely reflects combined provenance, weathering, and land-use effects. Element ratios (Zr/Sc vs Th/Sc, Th/U) and ternary diagrams place the samples close to global sediment and crustal compositions, with minimal sedimentary recycling, in line with sediment trapping by the Aswan High Dam. Bayesian background estimates show good agreement with empirical means once outliers are removed.

PCA reduces the dataset to a low-dimensional space where the first two components explain about two-thirds of the variance, while t-SNE offers improved preservation of local structure. In both spaces, hierarchical agglomerative clustering supports a practical four-cluster solution. These clusters correspond to (1) agricultural fields near highways dominated by crustal dust input, (2) rural agricultural areas influenced by local settlements, (3) soil close to major urban and industrial centers, and (4) strongly fertilized fields near Lake Burullus. The agreement between PCA- and t-SNE-based clustering, together with acceptable cluster validity scores, underscores the robustness of this segmentation and its usefulness for focusing monitoring on areas with greater anthropogenic influence

Integrated multivariate and machine-learning analysis of Nile Delta agricultural soils shows that the elemental composition is largely governed by crustal geochemistry and well described by region-specific background values derived from Bayesian inference. Combining high-resolution ICP-AES/ICP-MS data with PCA, t-SNE, and hierarchical clustering yields four geochemically coherent soil groups that reflect differences in land use and geographic setting, such as proximity to highways, rural settlements, major cities, and lacustrine–agricultural zones. This framework demonstrates how regional background modeling and unsupervised learning can reveal subtle structure in complex geochemical datasets and offers a transferable strategy for environmental monitoring and resource management in intensively cultivated deltaic systems.

References:

1. Fishar, M.R., Nile Delta (Egypt), in The Wetland Book: II: Distribution,

Description, and Conservation / C. M. Finlayson, G. R. Milton, R. C. Prentice, and N. C. Davidson, eds. – Springer Netherlands: Dordrecht., 2018. – P. 1251-1260.

2. Hamza, W., The Nile Delta, in The Nile: Monographiae Biologicae / H. J. Dumont, ed. – Springer Netherlands: Dordrecht, 2009. – P. 75-94.

3. Hinton, G., and S. Roweis, Stochastic Neighbor Embedding // In Advances in Neural Information Processing Systems. – 2003. – Vol. 15. – P. 833-840.

4. Lê, S., Josse, J., & Husson, F. FactoMineR: An R Package for Multivariate Analysis // Journal of Statistical Software. – 2018. – Vol. 25 (1). – P. 1-18.

5. Maaten, L. v. d. Accelerating t-SNE using Tree-Based Algorithms // Journal of Machine Learning Research. 2014. – Vol. 15. – P. 1-12.

6. Mahalanobis, P.C. On the generalised distance in statistics // Proceedings of the National Institute of Sciences of India, 1936. – P. 49-55.