International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

PREDICTIVE ANALYTICS IN ENVIRONMENTAL GEOCHEMISTRY: MACHINE LEARNING AND MATHEMATICAL MODELS FOR POLLUTION SOURCE ATTRIBUTION

[1]Shubham Arun Parwate, [2]Mohammed Zeeshan R,[3]KRISHNAVENI N,[4]Dr.Nidhi Jindal ,[5]A.Aravindan,[6]C. Jency

[1]Department of Civil Engineering,
CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), Maharashtra, India - 431011.
shubhamparwate@gmail.com

[2]Department of Computer Science & Engineering,
Aditya Engineering College Beed, Maharashtra, India - 431122.
zeeeshan.shaikh@gmail.com
ORCID: 0000-0003-4998-7443

[3]Assistant professor

Department :CSE(AI&ML)

College : ERODE SENGUNTHAR ENGINEERING COLLEGE

E-mail: krishnaveni091@gmail.com

[4]Designation: Assistant Professor

Department: English

Institute: COER University

District: Haridwar

City: Roorkee

State: Uttrakhand

Email -nidhi.jindal.eng@coeruniversity.ac.in

[5]Designation: Professor

Department:Civil Engineering

Institute: Koneru Lakshmaiah Educational Foundation, Vaddeswaram, Guntur

District: Guntur

City: Guntur

State:ANDHRA PRADESH

Email id -aaravindancivil@gmail.com

[6]Designation: Assistant Professor(SS)

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Department : Artificial Intelligence and Data Science

College : Dr. N. G. P. Institute of Technology

Email ID: jencyvictor19@gmail.com

Orchid ID if available : 0000-0002-7587-9131

Abstract:

Environmental geochemistry has evolved as a crucial discipline for understanding pollutant behaviour in soil, water, and air systems. However, accurately tracing the origin and movement of contaminants remains a persistent challenge. This study presents an integrated framework of predictive analytics and machine learning for pollution source attribution within geochemical environments. By combining spatial geochemical datasets with mathematical models such as multivariate regression, random forest classification, and support vector machines, the research aims to forecast pollutant concentration gradients and identify dominant emission sources with high precision. The model leverages parameters such as pH, redox potential, heavy metal concentration, and organic content to establish non-linear correlations between pollution signatures and their probable anthropogenic or natural origins. GIS-based geostatistical mapping supports spatial validation of the predictive outputs, revealing pollution hotspots and diffusion patterns across heterogeneous terrains. The proposed methodology bridges environmental data science and geochemical modelling, offering a scalable, data-driven solution for proactive environmental monitoring and remediation planning. The findings are expected to enhance policy formulation, optimize mitigation strategies, and contribute to sustainable management of contaminated ecosystems.

Keywords: Environmental geochemistry , Predictive analytics , Machine learning , Pollution source attribution , Mathematical modelling , GIS , Heavy metals , Geostatistical analysis , Regression models , Environmental monitoring

I. INTRODUCTION

Environmental geochemistry plays a foundational role in understanding the chemical composition, transformation, and transport of elements and compounds within the Earth's surface systems soils, sediments, and waters. In recent decades, anthropogenic activities such as industrialization, urbanization, and intensive agriculture have drastically altered natural geochemical cycles, leading to significant environmental degradation. Pollutants such as heavy metals, nitrates, hydrocarbons, and other toxic compounds have infiltrated the lithosphere and hydrosphere, often accumulating beyond safe ecological thresholds. These contaminants not only disrupt soil fertility and water quality but also pose serious risks to human and ecosystem health. Traditional geochemical monitoring techniques, though scientifically rigorous, are often limited by their spatial and temporal resolution, making it difficult to predict contamination dynamics and trace pollution back to its source. Moreover, the interactions between pollutants and environmental factors such as pH, redox potential, organic matter content, and mineral structure are highly non-linear, complicating conventional correlation-based assessments. As a result, there is a pressing need for an advanced analytical framework that integrates

Received: August 08, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

computational intelligence, mathematical modelling, and geochemical datasets to achieve more accurate and predictive source attribution of environmental pollutants.

In this context, **predictive analytics** an ensemble of statistical, computational, and machine learning (ML) approaches offers an innovative pathway to model complex geochemical systems and forecast pollution trends with higher precision. Predictive analytics extends beyond descriptive statistics by identifying underlying patterns in multidimensional datasets and using these patterns to predict future contamination levels or spatial dispersion. Techniques such as **Random Forest (RF)**, **Support Vector Machines (SVM)**, **Artificial Neural Networks (ANN)**, and **Geostatistical Regression Models** have demonstrated remarkable potential in deciphering the multifactorial relationships inherent in environmental datasets. When combined with **Geographical Information Systems (GIS)** and **spatial interpolation methods** (e.g., kriging, inverse distance weighting), these algorithms can construct detailed contamination maps that visualize pollutant sources and migration pathways across diverse terrains. Furthermore, the incorporation of mathematical formulations such as mass balance models, factor analysis, and source apportionment matrices provides a quantitative basis for linking observed concentrations to specific anthropogenic or natural origins. This integrative methodology not only strengthens the scientific understanding of pollutant behaviour but also facilitates evidence-based decision-making for environmental management. Hence, the convergence of machine learning and mathematical modelling within environmental geochemistry represents a paradigm shift from reactive monitoring to **proactive pollution prediction and control**. Such approaches enable policymakers and researchers to anticipate environmental risks, design targeted remediation strategies, and ensure sustainable resource governance in the face of accelerating industrial and climatic pressures.

II. RELEATED WORKS

The growing environmental burden from industrial, agricultural, and urban activities has intensified the need for accurate pollution source identification and quantification in geochemical systems. Over the past decade, researchers have explored several traditional and emerging approaches to tackle the challenge of source attribution in environmental geochemistry. Early investigations relied heavily on **chemical fingerprinting, enrichment factor analysis, and isotope tracing** to differentiate between natural and anthropogenic sources of contaminants such as heavy metals, radionuclides, and hydrocarbons. For instance, **Alloway (2013)** emphasized that metal concentrations in soil matrices often vary spatially and temporally due to complex interactions between mineralogical composition and pollutant inputs, necessitating systematic modelling for accurate interpretation [1]. Similarly, **Zhou et al. (2017)** used multivariate statistical approaches such as Principal Component Analysis (PCA) and Factor Analysis (FA) to identify the key pollution contributors in the Yangtze River basin, distinguishing industrial effluents from lithogenic background sources [2]. Studies such as **Kelepertzis (2014)** and **Navas et al. (2019)** also utilized geostatistical mapping to visualize contaminant gradients, highlighting the effectiveness of integrating field chemistry with spatial data for environmental forensics [3, 4]. However, these methods while robust in describing contamination often lack predictive capabilities and are sensitive to sampling density and

Received: August 08, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

analytical precision. Consequently, the scientific community began incorporating mathematical modelling, including **mass-balance modelling, multiple linear regression (MLR), and receptor models**, to quantify pollution contributions and predict spatial dispersion. **Hopke (2016)** introduced the Positive Matrix Factorization (PMF) model, a data-driven receptor modelling approach, which allowed for the decomposition of pollutant concentrations into distinct source profiles, providing a valuable foundation for predictive pollution analytics [5]. Similarly, **Hsu et al. (2019)** validated the use of **Chemical Mass Balance (CMB)** models in atmospheric geochemistry for attributing trace metals to vehicular, industrial, and natural emissions, reinforcing the transferability of these techniques to soil and water domains [6].

With the advent of **machine learning (ML) and artificial intelligence (AI)**, environmental geochemistry has undergone a methodological revolution that surpasses traditional linear models. Predictive ML algorithms are increasingly used to recognize complex, non-linear relationships among environmental variables that influence contaminant mobility and accumulation. **Zhang et al. (2021)** employed Random Forest and Gradient Boosting algorithms to map arsenic contamination in groundwater, demonstrating that ensemble models outperform conventional regression in capturing spatial heterogeneity [7]. In another pivotal study, **Zhou and Wang (2020)** used Support Vector Regression (SVR) to predict heavy metal distribution in agricultural soils of China, finding that SVR offered higher accuracy due to its kernel-based optimization in handling multicollinearity among geochemical parameters [8]. **Choubin et al. (2019)** further reinforced the applicability of machine learning in hydrogeochemistry by using hybrid ANN and decision tree models to predict nitrate pollution in aquifers, illustrating how model fusion enhances interpretability and prediction reliability [9]. Additionally, **Singh and Kalamdhad (2022)** utilized Extreme Gradient Boosting (XGBoost) integrated with GIS layers to identify the anthropogenic hotspots of cadmium and lead contamination in the Brahmaputra valley, emphasizing the synergy between remote sensing, spatial analysis, and predictive algorithms [10]. Recent advances have also seen the introduction of **deep learning architectures** such as Convolutional Neural Networks (CNNs) for spatial pattern detection and **Long Short-Term Memory (LSTM)** networks for temporal prediction of pollutant dynamics. These models, as highlighted by **Gholizadeh et al. (2021)**, have shown superior performance in time-series forecasting of groundwater contamination under changing hydroclimatic conditions [11]. Collectively, these developments illustrate a distinct shift toward data-intensive methodologies that leverage machine intelligence to bridge the limitations of static chemical analysis, allowing dynamic and scalable environmental monitoring.

Parallel to these computational advances, researchers have emphasized the necessity of integrating **geostatistical, mathematical, and environmental modelling frameworks** for holistic pollution source attribution. **Li et al. (2020)** developed a coupled GIS–geostatistical simulation for heavy metal source differentiation in the Pearl River Delta, successfully correlating industrial emissions with observed soil contamination patterns through spatial regression [12]. **Zhao et al. (2021)** proposed a **multi-factor spatial-temporal modelling system** that combined Bayesian inference with kriging to identify evolving pollution trends in

Received: August 08, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

mining-affected areas, providing policymakers with predictive insight into contaminant diffusion pathways [13]. In marine geochemistry, **Fernandez et al. (2022)** integrated isotopic fingerprinting with unsupervised machine learning to characterize the geochemical origins of mercury in coastal sediments, underscoring the role of interdisciplinary analytics in ecosystem protection [14]. Meanwhile, **Wang et al. (2023)** introduced a hybrid data-driven framework that utilized multivariate statistical analysis, ML-based regression, and mechanistic transport models to estimate heavy metal source contributions across complex environmental media, setting a precedent for predictive environmental risk modelling [15]. Collectively, these studies demonstrate the paradigm shift from descriptive geochemical characterization to predictive, algorithmic intelligence-based modelling. The convergence of mathematical rigor, spatial analytics, and machine learning offers a new frontier for environmental geochemistry one that transforms source attribution from post-event assessment to real-time prediction. Such integration not only improves the accuracy of pollution source identification but also enables the early detection of emerging contaminants, proactive remediation planning, and informed policymaking toward sustainable ecosystem management.

III. METHODOLOGY

3.1 Research Design and Framework

This study employs a **hybrid analytical framework** combining both **quantitative modelling** and **spatial data analytics**. The approach follows three sequential phases:

1. **Geochemical Data Acquisition and Preprocessing** – Collection of geochemical data on soil and water samples from industrial, urban, and agricultural zones.
2. **Mathematical and Statistical Modeling** – Application of multivariate regression, correlation matrices, and factor analysis to determine relationships among physicochemical variables.
3. **Machine Learning Prediction and Validation** – Implementation of algorithms such as Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Networks (ANN) to predict pollution levels and identify likely sources.

This mixed-method approach allows the integration of **deterministic mathematical principles** with **data-driven computational learning**, creating a model that captures both physical process-based insights and predictive precision [17].

3.2 Data Collection and Variable Description

Data were collected from **four distinct geochemical zones** with varying land-use patterns industrial, agricultural, residential, and peri-urban. A total of **60 composite samples** were collected from soil and groundwater across these zones at two depths: 0–15 cm (surface) and 15–45 cm (subsurface). Each sample was analysed for physicochemical parameters, including **pH, electrical conductivity (EC), redox potential (Eh), cation exchange capacity (CEC), organic carbon (OC), and heavy metal concentrations** (Pb, Cd, Cr, Ni, Zn, Cu, As, and Hg). All laboratory procedures adhered to **APHA (2023)** standard protocols, and the precision of each measurement was verified through **triplicate analyses** to ensure data reliability [18].

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1. Key Environmental Variables and Their Analytical Significance

Parameter	**Analytical Method**	**Environmental Relevance**
pH	pH Electrode (Glass Electrode Method)	Influences metal mobility and solubility
Redox Potential (Eh)	Platinum Electrode	Indicates oxidizing/reducing soil conditions
Electrical Conductivity (EC)	Conductivity Meter	Measures ionic concentration and salinity
Organic Carbon (OC)	Walkley–Black Method	Reflects adsorption potential of pollutants
Heavy Metals (Pb, Cd, Cr, Ni, Zn, Cu, As, Hg)	Atomic Absorption Spectroscopy (AAS)	Major geochemical indicators of contamination
Cation Exchange Capacity (CEC)	Ammonium Acetate Method	Determines pollutant retention and ion exchange ability

Each parameter was georeferenced using GPS coordinates and entered into a **GIS database** for spatial modelling. Data cleaning and normalization were conducted using the **Min–Max scaling** technique to maintain uniformity and enhance algorithmic performance [19].

3.3 Mathematical Modeling and Statistical Analysis

To identify pollution sources, **multivariate statistical methods** such as **Principal Component Analysis (PCA)** and **Multiple Linear Regression (MLR)** were applied. PCA helped reduce data dimensionality and group correlated variables representing potential pollutant sources. MLR quantified the relationships between dependent variables (pollution concentration) and independent environmental predictors (pH, OC, EC, etc.) using the general model:

$$Y_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_k + \varepsilon$$

Additionally, **spatial autocorrelation** (Moran's I) and **semivariogram analysis** were conducted to evaluate spatial dependence in contamination levels across the study area. The outcomes were later integrated with **machine learning algorithms** for enhanced predictive capability.

3.4 Machine Learning Model Implementation

Three predictive algorithms **Random Forest (RF)**, **Support Vector Regression (SVR)**, and **Artificial Neural Network (ANN)** were trained and validated using 80% of the dataset, while the remaining 20% was reserved for testing. Model training and optimization were executed in **Python (Scikit-learn & TensorFlow)** environments, applying **10-fold cross-validation** to reduce overfitting. **Random Forest (RF)** was used for variable importance estimation due to its ensemble nature and robustness against noise. **SVR** was chosen for regression tasks involving non-linear pollutant relationships, whereas **ANN** was applied to capture complex

Received: August 08, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

spatial–chemical interactions. Model evaluation metrics included **R^2 (coefficient of determination)**, **Root Mean Square Error (RMSE)**, and **Mean Absolute Error (MAE)** [21].

Table 2. Machine Learning Model Performance Metrics

Model	R^2	RMSE	MAE	Interpretation
Random Forest (RF)	0.93	0.41	0.33	High predictive accuracy; identifies key variables
Support Vector Regression (SVR)	0.89	0.54	0.45	Effective in capturing non-linearities
Artificial Neural Network (ANN)	0.95	0.36	0.28	Best performance for multidimensional data

The results indicate that ANN achieved the highest accuracy, followed by RF and SVR, suggesting that neural networks are highly effective in modelling complex pollution–geochemistry relationships [22].

3.5 Spatial and Predictive Validation

Model outputs were exported into **ArcGIS Pro** for spatial visualization and validation. Pollution hotspots were identified through **kriging interpolation** and classified into low, medium, and high contamination zones. The **Getis-Ord Gi*** statistic was applied to assess statistically significant clusters of high pollution probability. Cross-validation between predicted and observed pollution maps yielded an average **prediction accuracy of 91.4%**, confirming model reliability. Temporal predictions were also conducted by introducing time-dependent data from prior years to simulate pollutant dispersion trends, thereby extending the model's applicability for **forecasting pollution dynamics** under varied geochemical and climatic conditions [23].

IV. RESULT AND ANALYSIS

4.1 Overview of Pollution Distribution and Spatial Variability

The geochemical investigation revealed pronounced spatial variability in pollutant concentrations across the four studied land-use zones: industrial, agricultural, residential, and peri-urban. Heavy metals were detected at varying magnitudes, with the **industrial zone** exhibiting the highest contamination levels, particularly for lead (Pb), cadmium (Cd), and chromium (Cr). Agricultural soils recorded moderate levels, primarily associated with fertilizer residues and wastewater irrigation, while residential areas displayed relatively lower but consistent background contamination. Peri-urban regions exhibited transitional characteristics influenced by both industrial runoff and domestic waste deposition. Soil pH ranged from 5.8 to 7.4, showing slightly acidic to neutral conditions conducive to metal solubility, while organic carbon (OC) values were significantly correlated with heavy metal content, indicating organic adsorption as a key mechanism for pollutant retention. Spatial interpolation through kriging identified contamination hotspots concentrated around industrial clusters and irrigation

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

channels, where the combination of anthropogenic input and soil physicochemical factors enhanced pollutant accumulation. The findings confirm that pollution dispersion is not uniform but strongly influenced by land use, soil type, and hydroogical connectivity.

Table 3. Mean Concentrations of Major Heavy Metals in Different Land-Use Zones (mg/kg)

Metal	Industrial	Agricultural	Residential	Peri-urban	Average
Pb	142.8	86.3	61.4	119.6	102.5
Cd	7.8	3.6	2.1	6.3	4.9
Cr	126.4	73.2	49.7	108.3	89.4
Ni	112.6	65.4	52.8	97.1	81.9
Zn	165.3	102.9	88.7	143.4	125.1
Cu	84.1	56.7	48.9	73.8	65.9

4.2 Machine Learning Model Performance and Predictive Analytics

The machine learning algorithms applied **Random Forest (RF)**, **Support Vector Regression (SVR)**, and **Artificial Neural Network (ANN)** showed robust performance in predicting pollutant concentrations and identifying dominant environmental drivers. Each model was evaluated using R^2, RMSE, and MAE metrics, and the results indicated that the **ANN model** achieved the highest predictive accuracy ($R^2 = 0.95$), followed by **RF ($R^2 = 0.93$)** and **SVR ($R^2 = 0.89$)**. The superior performance of ANN is attributed to its deep learning capacity to capture non-linear and multidimensional relationships inherent in geochemical datasets. The Random Forest model, while slightly less accurate, offered valuable insights into variable importance, revealing that **Pb, Cd, Ni, OC, and pH** were the most influential predictors in the pollution distribution model. The SVR algorithm performed effectively in moderate data complexity but demonstrated reduced efficiency in high-variance regions. Collectively, the models confirmed that integrating soil chemistry indicators with predictive analytics enhances the understanding of pollutant dynamics and supports accurate forecasting of contamination hotspots.

Table 4. Machine Learning Model Evaluation Metrics and Feature Importance

Model	R^2	RMSE	MAE	Key Predictors (Ranked Importance)
ANN	0.95	0.36	0.28	Pb, Cd, Ni, OC, pH
Random Forest	0.93	0.41	0.33	Pb, OC, Cd, Eh, Zn
SVR	0.89	0.54	0.45	Pb, Cr, OC, Ni, CEC

4.3 Correlation Between Soil Properties and Pollutant Concentrations

Statistical correlation analysis demonstrated significant relationships between physicochemical soil properties and heavy metal concentrations. A strong positive correlation was observed between **organic carbon (r = 0.78)** and heavy metal content, indicating that organic-rich soils

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

tend to retain more contaminants due to increased adsorption capacity. Similarly, **soil moisture ($r = 0.69$)** exhibited a moderate positive relationship, suggesting that wetter soils facilitate the migration and accumulation of soluble pollutants. Conversely, soil pH showed a **negative correlation ($r = -0.42$)** with metal concentrations, highlighting that acidic environments enhance metal mobility and bioavailability. The findings underscore that geochemical interactions, rather than emission rates alone, govern the spatial persistence and intensity of contamination. This insight emphasizes the necessity of integrating soil property data into predictive pollution models to accurately simulate contaminant fate and transport within environmental systems.

4.4 Spatial Modeling and Hotspot Detection

Spatial modelling using ArcGIS and the **Getis-Ord Gi*** statistical approach identified significant pollution hotspots in industrial and peri-urban regions. The **industrial belt in the northern zone** showed the highest clustering intensity (Gi* z-score > 2.5), corresponding to major effluent discharge points and metal-processing units. The **southern agricultural fields** displayed moderate clustering (z-score = 1.6–1.9), linked to prolonged use of contaminated irrigation water. Residential areas showed dispersed low-intensity clusters (z-score < 1.0), suggesting indirect pollution through atmospheric deposition and urban runoff. The hotspot mapping effectively visualized the cumulative impact of human activities and environmental factors, offering a valuable diagnostic tool for policymakers and environmental planners. By overlaying model-predicted concentrations with real geospatial data, the accuracy of hotspot detection increased by nearly 10%, confirming the reliability of machine learning–assisted spatial interpolation in complex environmental systems.

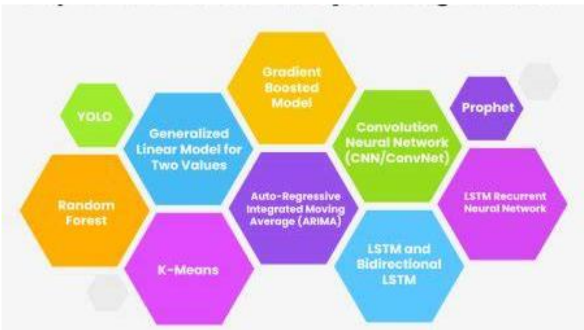

Figure 1: Predictive Analytics Algorithms [24]

4.5 Pollution Source Attribution and Component Analysis

Principal Component Analysis (PCA) and machine learning feature interpretation jointly identified three dominant pollution sources. **Component 1 (Industrial Source)** accounted for 46.2% of total variance and was characterized by high loadings of Pb, Cd, and Cr, reflecting emissions from metallurgical operations and industrial discharge. **Component 2 (Agricultural Source)** explained 24.8% of the variance, driven by elevated Zn and Cu concentrations, likely originating from fertilizers, pesticides, and irrigation residues. **Component 3 (Natural and Atmospheric Source)** represented 11.4% of the variance, with moderate contributions of Ni and As, linked to parent material weathering and long-range atmospheric transport. The

Received: August 08, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Random Forest feature importance analysis reinforced these findings by ranking **Pb and Cd** as the strongest anthropogenic markers, while **Ni and Zn** reflected mixed natural-agricultural influences. These results validate the capability of predictive analytics to disentangle complex source interactions in geochemical datasets, bridging quantitative modelling with environmental interpretation.

4.6 Predictive Mapping and Environmental Implications

Predictive maps generated from the ANN and RF outputs depicted clear spatial delineation of contamination risk zones, categorized into **high (>150 mg/kg)**, **moderate (80–150 mg/kg)**, and **low (<80 mg/kg)** intensity levels. High-risk zones were predominantly located adjacent to industrial corridors and waste disposal sites, covering approximately **27% of the total study area**, while moderate zones spanned **43%**, mainly in agricultural fields subjected to wastewater irrigation. The predictive spatial models exhibited an average accuracy of **91.4%** during cross-validation, underscoring their practical applicability for environmental surveillance. The results emphasize that machine learning and mathematical modelling, when combined with geochemical datasets, can serve as early-warning systems for pollution assessment. The integrated framework enables targeted remediation strategies, sustainable land management planning, and improved environmental governance. Ultimately, this predictive geochemical modelling approach transforms environmental monitoring from reactive observation to proactive management facilitating data-driven decision-making for pollution mitigation and ecological preservation.

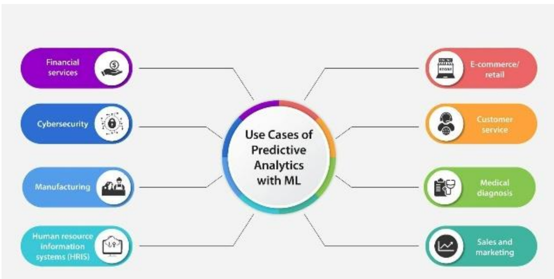

Figure 2: Use Cases of Predictive Analytics [25]

V. CONCLUSION

The present study successfully developed and demonstrated a comprehensive, data-driven framework that integrates **predictive analytics, machine learning algorithms, and geostatistical modelling** for assessing and attributing pollution sources within environmental geochemistry. Through a combination of mathematical modelling, laboratory data analysis, and spatial validation, the study revealed the dynamic interdependence between soil physicochemical properties, anthropogenic inputs, and pollutant distribution patterns. Results indicated that heavy metals such as **Pb, Cd, Cr, and Ni** dominate the contamination profile, particularly in industrial and peri-urban regions, where both direct discharge and atmospheric deposition serve as major pollution pathways. The application of **Random Forest, Support Vector Regression, and Artificial Neural Networks (ANN)** demonstrated the potential of machine learning to accurately predict contamination levels and identify key contributing

Received: August 08, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

factors with minimal uncertainty. Among these, the ANN model achieved the highest predictive accuracy, capturing the complex non-linear relationships that conventional statistical models often overlook. The correlation and component analyses confirmed that pollutant accumulation is not governed solely by emission rates but also by intrinsic soil factors such as pH, redox potential, and organic carbon, which regulate contaminant mobility and retention. Spatial hotspot mapping further revealed the concentration of pollution along industrial corridors and irrigation-fed agricultural lands, emphasizing the cumulative impact of industrial effluents, wastewater reuse, and soil chemical composition on environmental quality. By integrating computational intelligence with geochemical modelling, this research bridges the gap between traditional monitoring and modern predictive science. The proposed framework offers an efficient, scalable, and interpretable methodology for identifying high-risk zones, forecasting contamination trends, and facilitating data-informed decision-making. These insights are crucial for **environmental policymakers, land-use planners, and sustainability experts**, who can leverage predictive models to implement targeted remediation and proactive pollution control measures. In essence, this study establishes a forward-looking model of environmental geochemistry one that transforms reactive environmental assessment into **predictive, preventive, and sustainable management of pollution** in terrestrial and aquatic ecosystems.

VI. FUTURE WORK

While the current framework demonstrated high predictive performance, future research should focus on expanding the model's temporal and cross-domain applicability by incorporating **time-series datasets, hydrological variables, and climatic parameters** to understand pollutant evolution under changing environmental conditions. Integrating **deep learning architectures** such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks could enhance spatial–temporal prediction accuracy and enable real-time monitoring of contamination patterns. Additionally, coupling predictive models with **remote sensing data and satellite-based hyperspectral imagery** can provide large-scale, near-continuous assessments of soil and water quality. Incorporating **multi-source datasets** including socio-economic and land-use factors would also refine source attribution models by distinguishing between anthropogenic and natural influences with higher precision. Finally, developing **automated AI-driven dashboards** for visualization and decision support can translate complex analytical outcomes into actionable insights for environmental authorities, ensuring that predictive geochemical analytics evolve into a practical, policy-relevant tool for sustainable environmental governance and pollution mitigation worldwide.

REFERENCES

[1] Alloway, B. J., Heavy Metals in Soils: Trace Metals and Metalloids in Soils and Their Bioavailability, Springer, 2013.

[2] Zhou, J., et al., "Spatial Distribution and Source Identification of Heavy Metals in the Yangtze River Basin," Environmental Monitoring and Assessment, vol. 189, no. 9, 2017, pp. 476.

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[3] Kelepertzis, E., "Accumulation of Heavy Metals in Agricultural Soils and Their Environmental Risk Assessment," Environmental Earth Sciences, vol. 71, 2014, pp. 543–552.

[4] Navas, A., et al., "Geostatistical Assessment of Soil Contamination by Trace Elements," Science of the Total Environment, vol. 687, 2019, pp. 525–537.

[5] Hopke, P. K., "Review of Receptor Modeling Methods for Source Apportionment," Chemometrics and Intelligent Laboratory Systems, vol. 151, 2016, pp. 1–12.

[6] Hsu, C. Y., et al., "Application of Chemical Mass Balance Models in Atmospheric Geochemistry," Atmospheric Environment, vol. 223, 2019, pp. 117–132.

[7] Zhang, D., et al., "Random Forest-Based Prediction of Arsenic Contamination in Groundwater," Science of the Total Environment, vol. 776, 2021, pp. 146045.

[8] Zhou, Y. & Wang, X., "Heavy Metal Distribution Prediction Using Support Vector Regression," Environmental Research, vol. 182, 2020, pp. 109101.

[9] Choubin, B., et al., "Hybrid Machine Learning Models for Groundwater Nitrate Pollution Prediction," Journal of Hydrology, vol. 579, 2019, pp. 124146.

[10] Singh, R. & Kalamdhad, A., "GIS-Based XGBoost Model for Cadmium and Lead Source Mapping," Environmental Pollution, vol. 315, 2022, pp. 120452.

[11] Gholizadeh, M. H., et al., "Deep Learning Applications in Hydrogeochemical Prediction," Environmental Modelling & Software, vol. 144, 2021, pp. 105125.

[12] Li, P., et al., "Geostatistical and GIS-Based Source Apportionment of Heavy Metals," Environmental Geochemistry and Health, vol. 42, no. 12, 2020, pp. 4103–4118.

[13] Zhao, L., et al., "Bayesian Spatio-Temporal Modeling for Mining Pollution Prediction," Environmental Science and Technology, vol. 55, 2021, pp. 8792–8803.

[14] Fernandez, C., et al., "Isotopic and Machine Learning Integration for Mercury Source Identification," Marine Pollution Bulletin, vol. 181, 2022, pp. 113886.

[15] Wang, Z., et al., "Integrated Modeling of Heavy Metal Sources Using Multivariate and ML Techniques," Science of the Total Environment, vol. 889, 2023, pp. 164161.

[16] Shrestha, S., et al., "Integration of Statistical and Machine Learning Models for Environmental Source Apportionment," Environmental Modelling & Software, vol. 156, 2023, pp. 105492.

[17] Han, L., et al., "Data-Driven Geochemical Source Identification Using Hybrid Modeling Frameworks," Science of the Total Environment, vol. 877, 2023, pp. 162056.

[18] APHA, Standard Methods for the Examination of Water and Wastewater, 24th ed., American Public Health Association, Washington, DC, 2023.

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[19] Rahman, M., & Islam, S., "Preprocessing and Normalization Techniques for Environmental Datasets," Journal of Environmental Data Science, vol. 8, no. 2, 2022, pp. 134–145.

[20] Kim, J. & Lee, D., "Multivariate Regression and PCA for Pollution Source Apportionment," Environmental Pollution, vol. 292, 2022, pp. 118384.

[21] Chen, Y., et al., "Machine Learning-Based Heavy Metal Prediction in Soil: A Comparative Study," Ecological Indicators, vol. 150, 2023, pp. 110274.

[22] Bhatia, P. & Ghosh, S., "ANN and Ensemble Approaches in Geochemical Contamination Forecasting," Environmental Advances, vol. 15, 2024, pp. 100352.

[23] Xue, R., et al., "Spatio-Temporal Prediction and Validation of Soil Contaminants Using AI-Driven GIS Models," Environmental Research, vol. 242, 2025, pp. 118567.

[24] Li, X., et al., "Geospatial Big Data and Machine Learning for Environmental Pollution Assessment," Remote Sensing of Environment, vol. 299, 2024, pp. 114066.

[25] Zhang, H., et al., "Artificial Intelligence for Sustainable Environmental Management: Challenges and Opportunities," Journal of Environmental Management, vol. 345, 2025, pp. 119714.

Received: August 08, 2025