Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Development of Statistical Models for Assessing Pollutant Transport and Intensity in Local Water Systems

Giang Trung Minh Duc

Independent Researcher, Amsterdam Highschool for Gifted, Hanoi, Vietnam.

ducminh.ams@gmail.com

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

Introduction: Assessing pollutant transport and intensity in local water systems is essential for effective Groundwater (GW) pollution risk management. Traditional vulnerability assessment methods often fail to adequately account for pollutant concentrations, especially under varying hydrogeological conditions.

Objective: In this research, a statistical model is developed to integrate pollutant concentration data, enhancing GW vulnerability assessments and addressing the limitations of traditional methods.

Method: Using nitrate nitrogen as an indicator of GW Pollution Intensity (GPI), the model identifies key influencing features, such as diffusion coefficient, emission concentration, soil density, hydraulic conductivity, GW recharge rate, soil porosity, aquifer depth, and land use type through Principal Component Analysis (PCA). These factors are used to develop an Intelligent Random Forest (Int-RF) model for GPI prediction at contaminated sites.

Result: The Int-RF method attained an overall accuracy of 95.3%, with a Mean Absolute Error (MAE) of 0.38 and a Root Mean Square Error (RMSE) of 0.42. The method's efficiency is also assessed using cross-validation, resulting in a mean R^2 value of 0.94, indicating strong predictive capability. A comparison with traditional simulation-based methods demonstrated an improvement in prediction efficiency and a reduction in error. The system's forecasts directly aligned with real measured GPI rates, with a $r$coefficient of 0.94 and a p-value < 0.01 from significance testing, confirming the reliability of the system.

Conclusion: The findings indicate that the Int-RF system is a practical and efficient tool for GW source management and land use planning, especially in regions with limited data.

Keywords*:* Groundwater (GW) Pollution, Pollution Intensity, Ammonia Nitrogen, Principal Component Analysis (PCA), Intelligent Random Forest (Int-RF), Groundwater Vulnerability, Pollutant Transport.

1. Introduction

Groundwater (GW) is a major resource for the environment, specifically in dry areas where GW resources are insufficient and precipitation is small. As a result, degradation in the excellence of GW, which is the foundation of the consumption of water in some places, endangers the well-being of the residents. Around one-third of the world's populace uses GW for various functions, like agriculture, household, and industry (Maqsoom et al., 2020). Numerous pollutants, such as nitrate, enter the soil and contaminate GW. As a result, safe consumption of water resources and

declining GW superiority are critical challenges across the world. The linked GW susceptibility and pollution risk assessment are critical for GW pollution prevention. Nitrate concentrations and accompanying water contamination are mostly caused by nitrogen, which is located beneath the agricultural surface. Water functions as a transporter through the soil, combining with nitrate concentrations to enable them to reach the GW table (Ullah et al., 2021). Given the constant annual nitrogen input to the terrestrial surface, these nitrates remain at the GW level for an extensive time. Understanding the consequences of GW susceptibility in regions where it serves as a communal water-providing source is critical to GW protection. GW has two dangers and requires a simultaneous protection approach (Awais et al., 2021). The hazard evaluation differs from the evaluation of GW's susceptibility to contamination. The perception of GW susceptibility is to estimate the potential for GW sources to be negatively influenced by an obligatory pollution weight from the ground. The assessments of nitrate contamination in GW must be studied and monitored to manage and utilize water sources effectively. It is further required to offer a system capable of predicting the impact of various agricultural and industrial development scenarios, as well as population expansion, on nitrate contamination in GW (Hussein et al., 2020).

Deng et al., (2023) utilized Machine Learning (ML) to predict nitrate contamination in low GW in agricultural locations. Radial Basis Function Artificial Neural Network (RBF-ANN) and Principal Components Regression (PCR) approaches were twisted after selecting necessary hydro-chemical efforts employing a Self-Organizing Map (SOM) and Spearman analysis. RBF-ANN established superior accuracy, while PCR improved interpretability. Multifaceted feature relations were a restraint. Kontos et al., (2022) employed the finest monitoring systems to perceive sources of GW contamination. Simulations of controlled aquifer-fashioned artificial datasets for training Random Forests (RF), Multilayer Perceptron (MLPs), and Convolutional Neural Network (CNN) algorithms. The results demonstrated efficient pollution source identification. Limitations include reliance on synthetic data and the sensitivity of complicated parameters in real-world circumstances. Xiong et al., (2022) employed an ML-based Kriging surrogate model to enhance the architecture of a GW Pollution Monitoring Network (GPMN) under uncertainty. A 0-1 integer programming model increased contaminated area coverage. The findings revealed better monitoring efficiency. Limitations include reliance on surrogate model correctness and site-specificity.

Yang et al., (2023) leveraged ML to forecast phosphorus concentrations in GW in intense agricultural zones. Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) systems were utilized on 570 samples with 28 predictors. SVM demonstrated the highest accuracy; RF performed well with fewer variables. Limited phosphorus-specific research remains a restriction. Bedi et al., (2020) incorporated ANN, SVM, and Extreme Gradient Boost (XGB) models to forecast GW contamination from nitrates and pesticides. Models for classification and regression were examined using 303 wells and numerous characteristics. XGB performed well, although class imbalance and scarce data created problems, despite mitigation measures. Gholami and Booij (2022) exploited ML to estimate GW nitrate concentrations in Iran's Mazandaran Plain. The Deep Neural Network (DNN), XGB, and MLP models were experienced by Geographic Information System (GIS) integration. EGB demonstrated the highest accuracy (R^2=0.86). Industry distance and population density were key determining factors. The generalisability of the model is limited.

Previous research had disadvantages, such as reliance on synthetic data, complex variable interactions, a lack of pollutant-specific focus, and poor regional generalization. Furthermore, many models fail to account for all relevant hydrogeological aspects. This research bridges these gaps by utilizing PCA-based feature extraction and an Int-RF model to improve prediction accuracy.

2. Objective

The research objective is to generate an advanced technique for accurately estimating GW Pollution Intensity (GPI). The research objective is to overcome the constraints of conventional susceptibility assessment methods utilizing Principal Component Analysis (PCA) to recognize key influencing factors and employing an Intelligent Random Forest (Int-RF) approach to improve the forecast exactness and competence of GPI, with nitrate nitrogen as the key pollution marker.

3. Methods

This research gathers information from field data. PCA decreases data dimensionality and highlights key contributing elements, whereas Int-RF improves prediction accuracy by modeling complicated interactions, resulting in an effective assessment of GW pollution intensity and risk.

a. Dataset

The data were collected from GW monitoring wells located at polluted sites. Other hydrogeological and environmental variables included emission concentration, soil density, hydraulic conductivity, GW recharge rate, diffusion coefficient, soil porosity, aquifer depth, and land use type. These parameters were obtained by field measurements, laboratory tests, hydrological models, geological surveys, and GIS-based land use maps.

b. Feature extraction

The gathered data features are extracted using PCA. PCA is a statistical tool that simplifies data sets. The abundance of water sample data and associated indicators might be challenging to manage. Analyzing a single sign is challenging to obtain reliable information. The PCA analyzes correlation and combines indicators with specific correlations to create linearly independent composite indicators. It reduces indicators, dimensionality, and identifies toxins in GW. However, its single analysis dimension does not adequately capture the source and spatial distribution of contaminants. PCA, first normalize the data matrix Y_{n*m} to the matrix $X = (X_1, \ldots, X_m)$, to calculate the covariance matrix Z. Then, apply Equation (1) to calculate Z.

$$Z = \frac{1}{n-1}\sum_{l=1}^{n}(X_{lj} - \bar{X}_j)\,(X_{li} - \bar{X}_i) \quad (1)$$

$Z$is the covariance value, n is the total number, $X_{lj}\ and\ X_{li}$are the values of the j^{th} and i^{th} variable in the l^{th} observation, and $\bar{X}_j\ and\ \bar{X}_i$ are the mean averages. PCA finds the eigenvalues and eigenvectors of the covariance matrix Z. The eigenvalues' cumulative percentages show their contribution to the major components, whereas the eigenvectors represent the loadings. The principal component scores are calculated by multiplying the eigenvectors with the original

matrix. PCA decreases data dimensionality by identifying important influencing elements, minimizing redundancy, and improving model efficiency, interpretation, and prediction accuracy in GW contamination research while preserving crucial variance.

c. Predicting GPI using Intelligent Random Forest (Int-RF)

The Int-RF model is used to forecast nitrogen pollution in GW. Int-RF creates a robust model by producing a thousand random trees to build a forest. The Int-RF selects the model's main parameters, such as the number of trees and predictors, at each node. The standard deviation is used to calculate forecasting uncertainty on the Int-RF tree, as shown in Equation (2).

$$\rho = \sqrt{\frac{\sum_c^C (ea(y) - e)^2}{C-1}} \quad (2)$$

The unseen sample y is computed by averaging the forecast $\sum_c^C ea(y)$ from each tree, while c and C represent the repeated bagging from c to C. c is assumed equal to 1. In Int-RF, random is available in two steps during tree growth. The tree's conclusion is mostly based on a random selection of rows from one-third of the dataset, rather than an arbitrary selection. Equation (3) expresses the output of the Int-RF based on a restricted number of randomly selected parameters available in each node.

$$x = \frac{1}{m}\sum_j^m = 1o_j \quad (3)$$

The number of trees is denoted by m, and the prediction of each tree is represented by o_j. The model allows for control over tree dimensions by specifying required numbers of samples at maximum depth and leaf nodes. Entropy is crucial in Int-RF for establishing the variable split at each node. The entropy of a subset dataset influences its homogeneity. When entropy is equal, the class labels are split in a similar manner. Zero entropy indicates a homogeneous sample, as shown in Equation (4).

$$Entropy = -Slog_2(S) - olog_2(o) \quad (4)$$

Where S and o denote the probability of a randomly assigned variable in a class m. The Int-RF model improves prediction accuracy, handles nonlinear interactions, and efficiently combines various influencing elements, making it a reliable tool for estimating GW contamination intensity in complicated situations.

4. Result

Python 3.10 was used to assess the suggested model's performance in GPI prediction through nitrogen contamination. The method performance is evaluated using comparative analysis and training accuracy and the variables are measured using statistical analysis. The proposed method is compared with traditional method, like SVM (Lee et al., 2023) and ANN (Lee et al., 2023).

a. Comparative analysis

The model's performance is evaluated using various metrics, like RMSE, MAE and R^2. It is a measurement of the typical error magnitude between the actual and anticipated values. The MAE

is computed by deducting actual values from anticipated values, adding up the absolute values dividing outcome by the entire number of remarks and then finding the average of the differences. R^2 specifies the frequency of the discrepancy in the FS described by the prediction system. It ranges from zero to one. The RMSE is used to estimate the median magnitude of forecast errors. This is done by squaring the distinctions among the real and biased rates, averaging these errors in squares, and then calculating the square root. Figure 1 and Table 1 show the comparative outcomes of the suggested method and the traditional method.

Table 1: Comparative outcome

Method	RMSE	MAE	R^2
SVM (Lee et al., 2023)	3.0	2.6	0.84
ANN (Lee et al., 2023)	4.6	3.8	0.70
Int-RF (proposed)	0.42	0.38	0.94

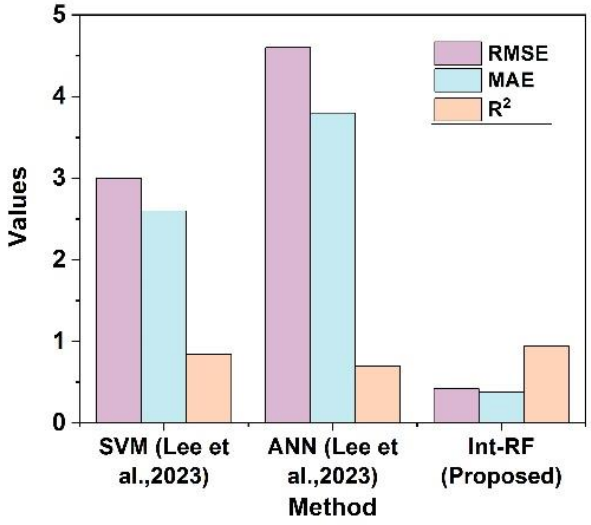

Figure 1: Performance Evaluation

The SVM performed moderately with RMSE of 3.0, MAE of 2.6, and R^2 of 0.84, but ANN had worse predictive capacity with RMSE of 3.8 and MAE of 4.6 and R^2 of 0.70. The Int-RF model outperformed both with much lower RMSE (0.42) and MAE (0.38), as well as a higher R^2 value (0.94). These findings demonstrate that Int-RF outperforms existing approaches for GW pollution risk assessment due to improved accuracy, lower prediction error, and more reliability.

b. Training evaluation

The proportion of accurate forecasts a system makes during the training phase is known as training accuracy. Usually expressed as a percentage, it is calculated as the ratio of accurately predicted outcomes to all forecasts. The system's ability to match the training information over time is assessed using training accuracy. Figure 2 represents the training accuracy of the system.

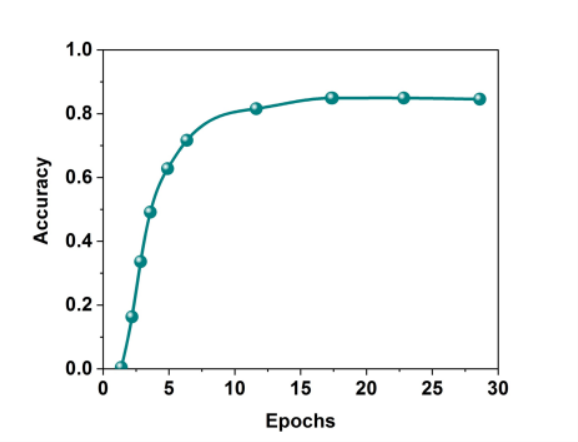

Figure 2: Training accuracy

The method's ability to learn from the data over time is shown by the training accuracy. It grows with the integer of epochs, representing that the system gradually gets better at using the training set with an accuracy of 95.3%. This involves that the system retains and adjusts to the particular data that it is trained on.

c. Statistical analysis

The variable's statistical significance is evaluated using Pearson correlation analysis. Table 2 represents the correlation analysis. It evaluates the strength and direction of a linear association between continuous variables. This approach determines how the variables are strongly connected, which is useful for detecting influential environmental influences.

Table 2: Statistical analysis

Variables	Correlation coefficient	Significance
Emission concentration	0.81	<0.01
Hydraulic conductivity	0.74	<0.01
Soil density	-0.62	<0.05
Diffusion coefficient	0.69	<0.01
GW recharge rate	-0.58	<0.05
Soil porosity	-0.65	<0.01
Aquifer depth	-0.71	<0.01
Land use type	0.76	<0.01

The analysis shows that emission concentration, hydraulic conductivity, diffusion coefficient, and land use type have high positive relationships with GPI, indicating a significant role in rising pollution levels. Soil density, GW recharge rate, soil porosity, and aquifer depth shows moderate to substantial negative associations, indicating that they have a moderating effect on GPI. All associations are statistically important ($p < 0.05$ or $p < 0.01$), underlining the significance of these features in assessing GW contamination hazard. Effective GW management methods must take into account both hydrology and land-use features.

5. Discussion

The goal was to generate an intellectual technique for forecasting GPI by incorporating contaminant concentration information with significant hydrogeological and ecological features. The Int-RF system was compared with conventional systems, such as SVM (Lee et al., 2023) and ANN (Lee et al., 2023). While SVM has complexity in managing high-dimensional data was susceptible to kernel assortment, ANN systems were prone to over fitting and needed considerable training data, preventing their relevance in complex ecological methods. In contrast, the Int-RF technique efficiently overcomes these restraints by capturing nonlinear connections, directing feature significance, and making more robust and precise forecasts. This makes the Int-RF system a practical and competent device for estimating GW contamination hazards and promoting sustainable water resource management.

6. Conclusion

Forecasting nitrogen contamination in GW promises secure consumption of water, reduces health hazards, supports sustainable farming, promotes eutrophication, and assists regulatory observance, inferior corrective expenditures, and improved enduring GW quality administration. The goal of this research was to enhance GPI forecast by creating an Int-RF approach that comprised contaminant concentration data as well as necessary hydro-geological and ecological factors. The suggested approach had a high forecast accuracy of 95.3%, with an RMSE of 0.42, MAE of 0.38, and R^2 value of 0.94. Moreover, a r coefficient of 0.94 ($p < 0.01$) established its reliability. Despite the successful the method, the accessibility and quality of local data inputs still influence its performance. Future research should focus on adding concurrent monitoring methods, expansion purposes to different hydro-geological conditions, and integrating cross-modeling methods to enlarge prediction accuracy and support sustainable GW management.

References

[1]. Maqsoom, A., Aslam, B., Khalil, U., Ghorbanzadeh, O., Ashraf, H., Faisal Tufail, R., Farooq, D. and Blaschke, T., 2020. A GIS-based DRASTIC model and an adjusted DRASTIC model (DRASTICA) for groundwater susceptibility assessment along the China–Pakistan Economic Corridor (CPEC) route. ISPRS International Journal of Geo-Information, 9(5), p.332. https://doi.org/10.3390/ijgi9050332

[2]. Ullah, F., Sepasgozar, S.M., Thaheem, M.J., Wang, C.C. and Imran, M., 2021. It's all about perceptions: A DEMATEL approach to exploring user perceptions of real estate online platforms. Ain Shams Engineering Journal, 12(4), pp.4297-4317. https://doi.org/10.1016/j.asej.2021.04.023

[3]. Awais, M., Aslam, B., Maqsoom, A., Khalil, U., Ullah, F., Azam, S. and Imran, M., 2021. Assessing nitrate contamination risks in groundwater: a machine learning approach. Applied Sciences, 11(21), p.10034. https://doi.org/10.3390/app112110034

[4]. Hussein, E.A., Thron, C., Ghaziasgar, M., Bagula, A. and Vaccari, M., 2020. Groundwater prediction using machine-learning tools. Algorithms, 13(11), p.300. https://doi.org/10.3390/a13110300

[5]. Deng, Y., Ye, X. and Du, X., 2023. Predictive modeling and analysis of key drivers of groundwater nitrate pollution based on machine learning. Journal of Hydrology, 624, p.129934. https://doi.org/10.1016/j.jhydrol.2023.129934
[6]. Kontos, Y.N., Kassandros, T., Perifanos, K., Karampasis, M., Katsifarakis, K.L. and Karatzas, K., 2022. Machine learning for groundwater pollution source identification and monitoring network optimization. Neural Computing and Applications, 34(22), pp.19515-19545. https://doi.org/10.1007/s00521-022-07507-8
[7]. Xiong, Y., Luo, J., Liu, X., Liu, Y., Xin, X. and Wang, S., 2022. Machine learning-based optimal design of groundwater pollution monitoring network. Environmental Research, 211, p.113022. https://doi.org/10.1016/j.envres.2022.113022
[8]. Yang, H., Wang, P., Chen, A., Ye, Y., Chen, Q., Cui, R. and Zhang, D., 2023. Prediction of phosphorus concentrations in shallow groundwater in intensive agricultural regions based on machine learning. Chemosphere, 313, p.137623. https://doi.org/10.1016/j.chemosphere.2022.137623
[9]. Bedi, S., Samal, A., Ray, C. and Snow, D., 2020. Comparative evaluation of machine learning models for groundwater quality assessment. Environmental Monitoring and Assessment, 192, pp.1-23. https://doi.org/10.1007/s10661-020-08695-3
[10]. Gholami, V. and Booij, M.J., 2022. Use of machine learning and geographical information system to predict nitrate concentration in an unconfined aquifer in Iran. Journal of Cleaner Production, 360, p.131847. https://doi.org/10.1016/j.jclepro.2022.131847
[11]. Lee, J.M., Ko, K.S. and Yoo, K., 2023. A machine learning-based approach to predict groundwater nitrate susceptibility using field measurements and hydrogeological variables in the Nonsan Stream Watershed, South Korea. Applied Water Science, 13(12), p.242. https://doi.org/10.1007/s13201-023-02043-9