Journal of NAMP 67, 2 (2024) 207-218

The Nigerian Association of Mathematical Physics

Journal homepage: https://nampjournals.org.ng

COMPARISON OF STATISTICAL MODEL AND RANDOM FOREST FOR GROUNDWATER CONTAMINATION PATTERNS.

[1]Udegbe, S. O. [2] Ukaoha, K. C

[1]Department of Computer Science, National Open University of Nigeria.
[2] Department of Computer Science, Pan-Atlantic University, Lagos

ARTICLE INFO

Article history:
Received xxxxx
Revised xxxxx
Accepted xxxxx
Available online xxxxx

Keywords:
Groundwater modelling, Contamination; Heavy Metals, Multiple Linear Regression, Random Forest Model.

ABSTRACT

In this study, a random forest model was compared to a statistical model for predicting heavy metal concentrations in groundwater in Edo State, Nigeria. The pH of groundwater samples was determined using a pH meter, and heavy metal concentrations were measured with Atomic Absorption Spectrophotometer (AAS). Pearson Correlation Coefficient was used to evaluate correlations between heavy metal concentrations. Both Random Forest Model (RFM) and Multiple Linear Regression (MLR) were employed to model these concentrations, with goodness of fit assessed via R-squared and root mean square error (RMSE). Results showed that heavy metal concentrations, except for lead, were generally within acceptable limits. The RFM outperformed MLR in predicting iron and lead concentrations but was less effective for arsenic. Python was used for modelling and data extraction. Both models are suitable for predicting groundwater contamination, with RFM showing better overall performance.

1. INTRODUCTION

Groundwater, commonly referred to as the "lifeblood of ecosystems," is essential for human survival, agriculture, and the environment. Researchers, legislators, and environmental scientists have paid particular interest to groundwater quality and pollution patterns because they have a direct impact on human health and ecological stability. Contaminants in groundwater are often caused by dissolving of mineral deposits in the Earth's crust [1],[2],[3],[4],[5]. However, as the world's population, urbanisation, industrialization, agriculture, and economy grow, anthropogenic toxins now pose a threat to the environment [5] investigated the appropriateness of groundwater for drinking and irrigation in Akhmim, Egypt. The researchers reported that about 95% of the gathered groundwater samples are significantly polluted.

*Corresponding author: Udegbe, S. O
E-mail address: nou224139979@nou.edu.ng
https://doi.org/10.60787/jnamp.v67i2.371

[6] Performed a physiochemical assessment of groundwater pollution resulting from the discharge of industrial wastewater in Faisalabad, Pakistan and provided data of the levels of contamination of groundwater in Faisalabad.

Groundwater is a critical resource in the agricultural, civil, and industrial sectors. Groundwater pollution caused by diverse chemical components must be predicted in order to plan, policy, and manage groundwater resources effectively. This will help to mitigate the complex and numerous issues associated with addressing groundwater contamination. To accurately analyse contemporary groundwater pollution patterns, new techniques and methodologies are required. Statistical models such as multiple linear regression (MLR) have long been a cornerstone of environmental research, providing structured frameworks for data analysis and insights into complex phenomena; however, in the last two decades, the use of machine learning (ML) techniques for groundwater quality (GWQ) modelling has increased exponentially [7].

Machine learning (ML) is the algorithmic study of how computers replicate or apply human learning behaviour. Machine learning algorithms are intended to accurately anticipate patterns in multivariate data [8]. They are frequently utilised in a variety of applications, including pattern recognition, anomaly detection, and classification. Many writers have tried and analysed the use of machine learning (ML) techniques such as Logistic Regression and Random Forest to forecast water quality and groundwater contamination [9].

Random Forest (RF) is an ensemble classification/regression method that trains many classifiers and then combines their findings via a vote procedure [10], [11]. It is a method in which a large number of decision trees are generated and trained on the original training data, with the output class selected by a majority vote of the trees [9]. The random forest model excels at managing big, missing, and outlier data because to its robust model structure. [12] used Random Forest regression to estimate groundwater contamination in Africa and compared its results to a multiple linear regression model. [13] utilized artificial neural network techniques to forecast the likelihood of groundwater arsenic contamination in Cambodia, Laos, and Thailand and provided data and information on the potential health hazards it poses to nearby communities.

The present study aims to develop an efficient predictive model for groundwater contamination using Random Forest (RF) and comparison its effectiveness with Multiple linear Regression (MLR).

2.0 MATERIALS AND METHODS

2.1 Sample Collection

Groundwater samples were collected from 14 locations or stations (Table 1) across 3 Local Government Areas (LGAs) of Edo State and a control sample, a bottled water sold within the Benin City Environ was purchased.

The pH of the groundwater samples was determined using a pH meter while the concentrations of the heavy metals were determined using Atomic Absorption Spectrophometer (AAS) after digestion. The results obtained will be compared to **[14]** and **[15]** standards of portable water.

2.2 Statistical Analysis

Regression and correlation analysis was employed to determine the mathematical relationships and closeness between the values of the variables understudy. Results of descriptive statistics for the relevant variables used to predict groundwater contamination will also be provided which include mean and standard deviation.

Table 1: Sampling locations across 3 Local Government Area (LGA) of Edo State

LGA	Location	Stations
Oredo	GRA1`	1

	New Benin	2
	Ibiwe	3
	Ihogbe	4
	Ekehuan	5
	Oreoghene	6
Egor	Uselu	7
	Uwelu	8
	Okhoro	9
	Ugbighoko	10
Uhunwhonde	Obadan	11
	Ogheghe	12
	Igieduma	13
	Ehor	14
	Control	15

2.3 Multiple Linear Regression (MLR) and Random Forest Model (RFM)

The general methodology of machine learning includes (a) data preparation, analyses, and visualization; (b) normalization; (c) model selection and implementation; and (d) performance metrics. In this study, the following steps were implemented using "Python" Code within "Anaconda Notebook": (i) import the required libraries, (ii) import the "-.csv" file containing the dataset (concentrations of Mn, As, Fe, Pb, and Zn variables and pH), (iii) statistical analyses and data visualization, (iv) Multiple Linear Regression and Random Forest models, and (v) performance metrics.

A comparative analysis of the predictive performance of both MLR and RF models were utilized for forecasting concentrations of three selected heavy metal species, namely Fe, Pb and Pb. This evaluation was based on key model evaluation metrics outlined earlier in the methodology, specifically focusing on the R-squared (R^2) value and Root Mean Square Error (RMSE). Both models were belt on same size train test split, which is 70% train to 30% test splits. The RF model run under the following parameters: criterion="entropy," n_estimators = 10, and random_state = 0. By comparing the R-squared values and RMSE obtained from both the RFM and MLR models, their relative effectiveness in predicting the concentrations of heavy metals in the groundwater samples were assessed. The model with higher R^2 value and lower RMSE value was considered more reliable and accurate in predicting metal species concentrations in groundwater. The flowchart of the MLR and RFM prediction of the heavy metal concentrations in groundwater is presented in Figure 1.

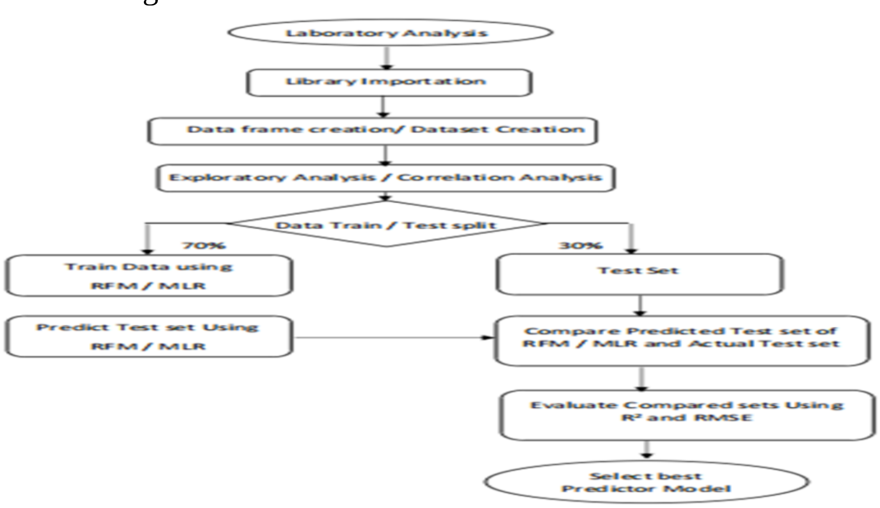


Figure 1: Flowchart of study methodology of RFM and MLR prediction of groundwater heavy metal concentrations

3.0 RESULTS AND DISCUSSIONS

3.1 Concentrations of Heavy Metals in Groundwater

The results of groundwater contamination study is presented in Table 2 for pH and the concentrations of various heavy metals under study which includes manganese (Mn), iron (Fe), zinc (Zn), arsenic (As) and lead (Pb) taken from groundwater from 14 locations across 3 Local Government Area (Uhunwhonde, Egor and Oredo) of Edo State in comparison with some standards of portable water.

Table 2: Concentration of heavy metals in groundwater taken from 14 Locations

LGA	Location	pH	Concentration (mg/l)				
			Mn	Fe	Zn	As	Pb
Oredo	GRA	6.40	0.38	0.30	1.14	0.02	0.03
	New Benin	6.21	0.36	0.31	1.25	0.02	0.03
	Ibiwe	6.31	0.35	0.28	1.22	0.02	0.03
	Ihogbe	6.40	0.32	0.27	1.19	0.02	0.03
	Ekehuan	6.51	0.33	0.27	1.19	0.01	0.03
	Oreoghene	6.55	0.30	0.26	1.18	0.01	0.03
Egor	Uselu	6.67	0.25	0.24	1.17	0.01	0.02
	Uwelu	6.63	0.27	0.24	1.15	0.01	0.02
	Okhoro	6.76	0.26	0.24	1.14	0.01	0.02
	Ugbighoko	6.87	0.24	0.23	1.12	0.01	0.02
Uhunwhonde	Obadan	6.94	0.23	0.23	1.12	0.01	0.02
	Ogheghe	7.03	0.24	0.24	0.11	0.01	0.02
	Igieduma	7.05	0.20	0.23	1.10	0.00	0.02
	Ehor	7.05	0.18	0.23	1.08	0.00	0.02
	Control	7.07	0.17	0.23	1.08	0.00	0.02
	Mean (mg/l)	6.70±0.29	0.27± 0.06	0.25± 0.03	1.08± 0.26	0.01± 0.01	0.02± 0.01
	NIS (2015)	6.5 - 8.5	0.10	0.30	5.00	0.01	0.01
	WHO (2017)	6.5 - 8.5	0.40	0.30	3.00	0.01	0.01

Nigerian Industrial Standard (NIS), World Health Organization (WHO)

From Table 2, New Benin exhibited the highest concentrations for the heavy metals except for manganese while the control exhibited the lowest concentrations across all observed heavy metals. The mean values of pH was found to be 6.70 ± 0.29 which is within the permissible limit by [15] and [14] standards for portable drinking water. Water with lower or much higher pH is unsafe for drinking and this can result from chemical or heavy metal pollution [16]. The mean concentration of the heavy metals in the ground water from the different locations was found to be 0.27 ± 0.06, 0.25 ± 0.03, 1.08 ± 0.26, 0.01 ± 0.01 and 0.02 ± 0.01 mg/l for Mn, Fe, Zn, As and Pb respectively. Zinc had the highest concentration in the groundwater while Arsenic had the least concentration. One notable observation was the inverse relationship between pH values and metal concentrations. As the concentration of heavy metal species increased across the sampling locations, the pH values decreased. This phenomenon indicates a positive correlation between acidity and increasing heavy metal concentrations. This correlation can be attributed to several factors, including increased solubility and the liberation of hydrogen ions in acidic environments. Additionally, competitive sorption processes may play a role, causing faster displacement of heavy metals from soil into groundwater. It is noteworthy that, in general, the concentrations of heavy metal species at most sampling locations was within the accepted limit of both [15] and [14] standards of portable water. However, at a few sampling points, concentrations exceeded these accepted limits. The average concentrations of Mn, Fe and Pb were observed to be lower than those reported by [17] who reported the values of these HMs to be 1.76, 0.71 and 0.04 mg/l respectively for groundwater samples collected from Okomu National Park in Ovia South West LGA of Edo State while the concentration of Zn from this study was higher than the value of 0.01 mg/l the authors reported. Also, the average concentrations of Mn, Zn and Pb in groundwater reported in this study were found to above the range of 0.0014 to 0.0904 mg/l, 0.0014 to 0.2829 mg/l and 0 to 0.0115 mg/l of Mn, Zn and Pb respectively reported by [18] for studies carried out on groundwater samples from selected communities in Edo State while the average concentration of Fe from this study was within the range of 0.0218 to 0.4570 mg/l the authors reported.

3.2 Correlation between Groundwater Variables

The correlation analysis of pH and the concentrations of heavy metals present in the groundwater under study is presented in Table 3.

Table 3: Correlation between groundwater variables (pH and the heavy metals)

	pH	Mn	Fe	Zn	As	Pb
pH	1					
Mn	-0.9464	1				
Fe	-0.8980	0.9253	1			
Zn	-0.4653	0.2807	0.2734	1		
As	-0.8766	0.9083	0.8174	0.1697	1	
Pb	-0.8598	0.8747	0.8901	0.3471	0.7206	1

From Table 3, pH showed strong negative correlations with Mn (r =-0.9464), Fe (r = -0.8980), As (r = -0.8766) and Pb (r = -0.8598) and weak negative correlation with Zn (r = -0.4653), indicating that the concentrations of the metals decreased with rise in pH. Mn showed strong positive correlations with Fe (r =0.9253), As (r = -0.9083) and Pb (r = 0.8747) and weak positive correlation with Zn (r = 0.2807), indicating that the concentrations of Mn in the groundwater increased with the concentrations other metals in the water. Fe showed weak positive correlations with Zn (r = 0.2734) and strong positive correlations with As (r = -0.8174) and Pb (r = 0.8901) indicating that the concentrations of Fe in the groundwater increased with the concentrations other metals in the water. Zn showed weak positive correlations with As (r = 0.1697) and Pb (r = 0.3471) while As showed strong positive correlations with Pb (r = 0.7206). The negative correlation exhibited by pH on the concentration of heavy metals in the groundwater and the positive correlations between the heavy metals are consistent with results reported by [19] for correlation of heavy metals in groundwater sourced from Hunan Province, China. The correlation between these variables is represented diagrammatically by the correlation map in Figure 2.

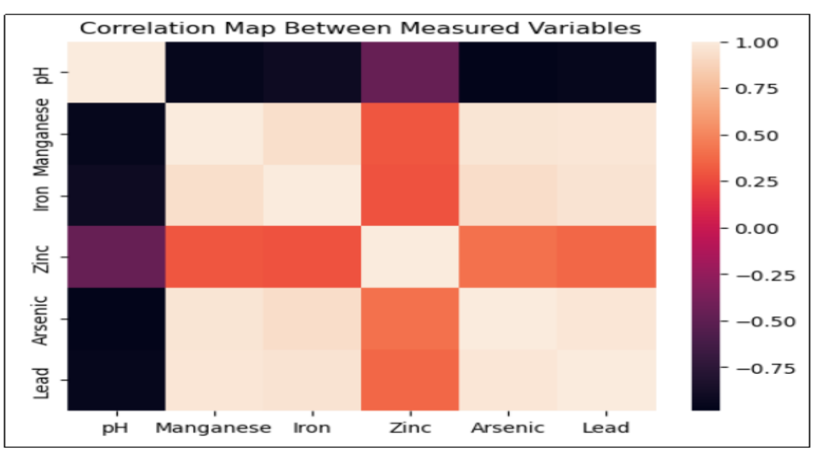
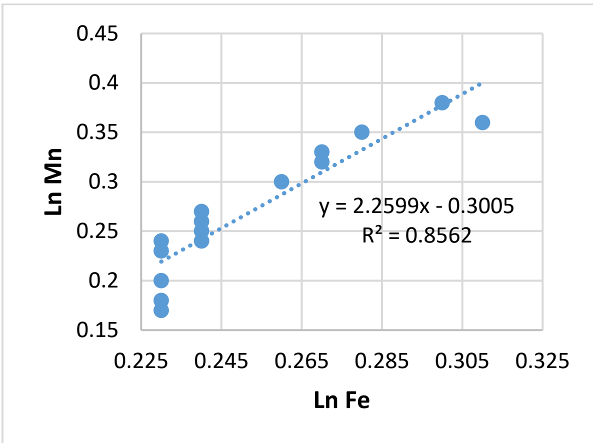


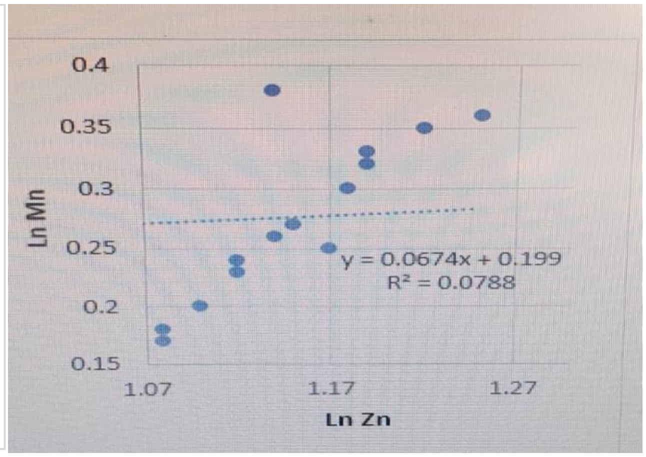
Figure 2: The correlation map between the measured variables

3.3 Linear Regression between the Heavy Metals in the Groundwater

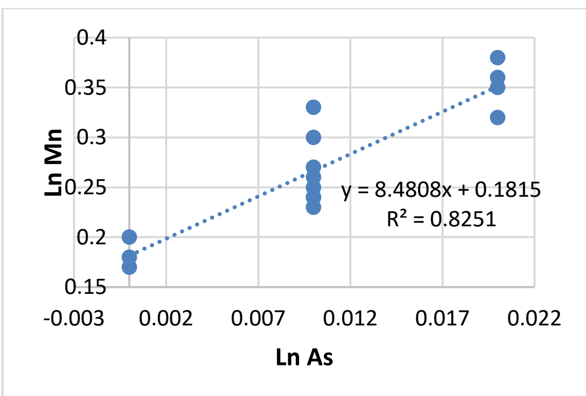
The linear regressions between the concentrations of the heavy metals in the groundwater are shown in Figure 3. From Figure 3 (a) to (i), there were positive relationships between Mn, Fe, Zn, As and Pb, with especially stronger relations of Mn with Fe, As and Pb, indicating that these heavy metals likely co-varied with each other [20]. This indirectly proved that these heavy metals could have the same source of metals and nonferrous metal minerals present in the study areas [19].



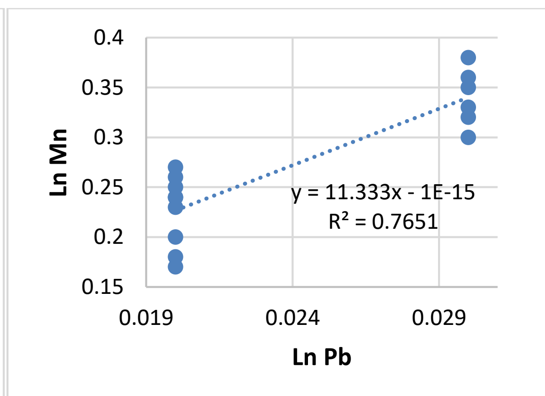
(a)



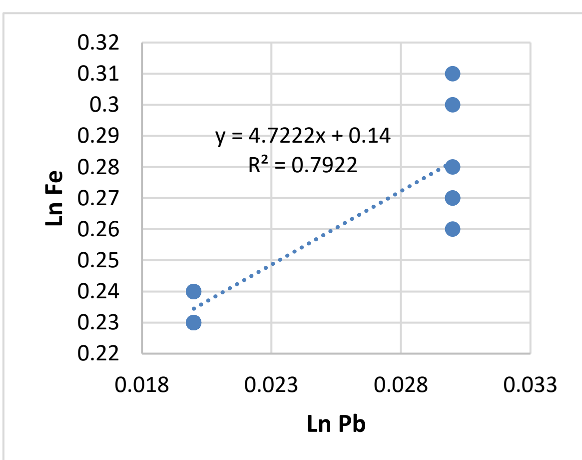
(b)



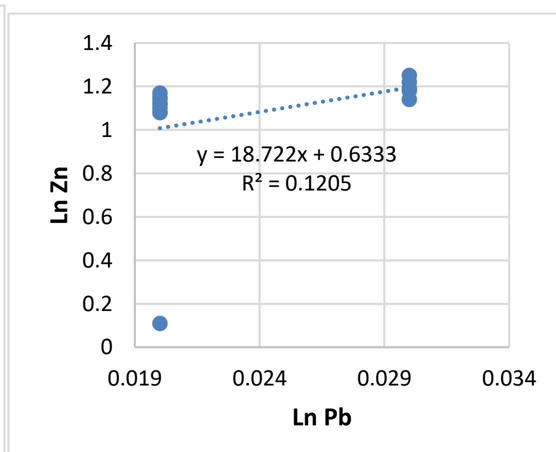
(c)



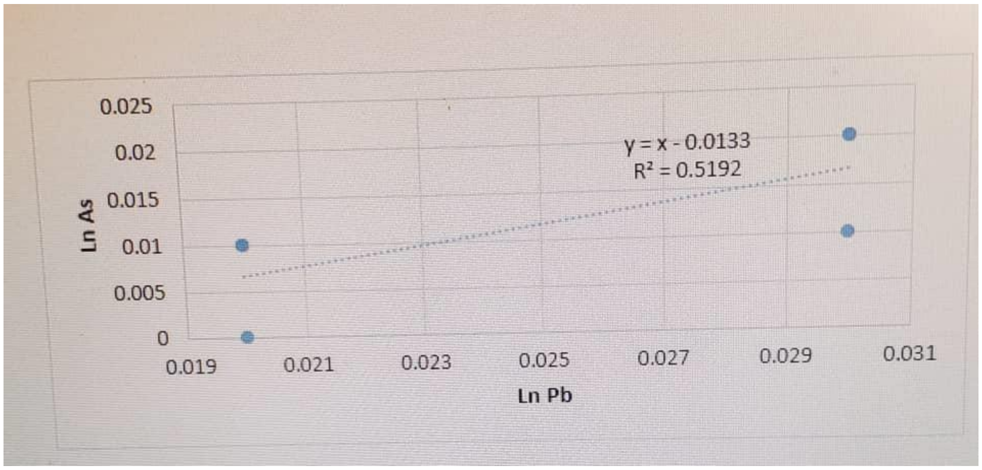
(d)



(g)



(h)

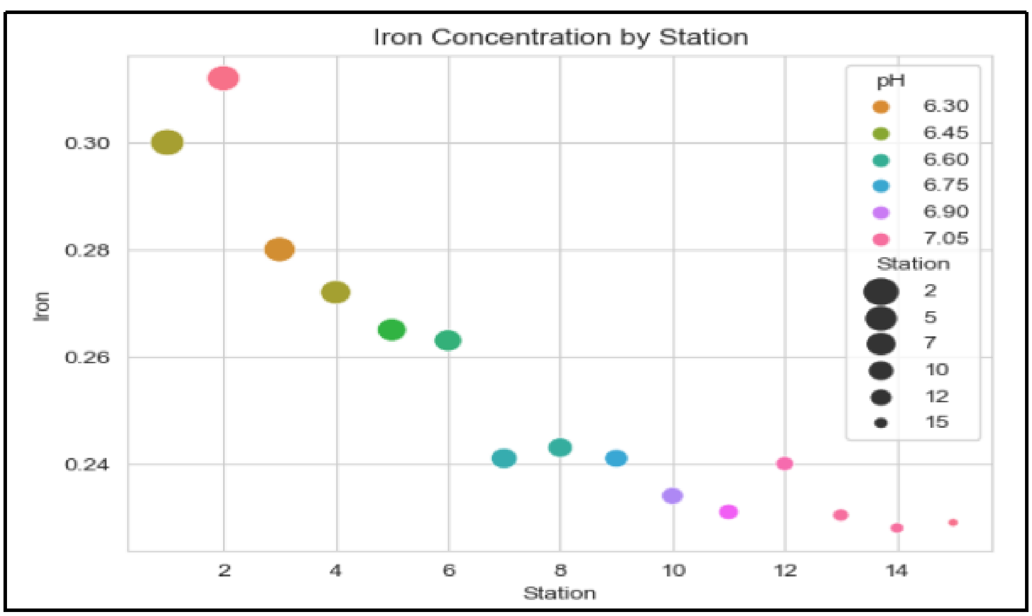


(f)

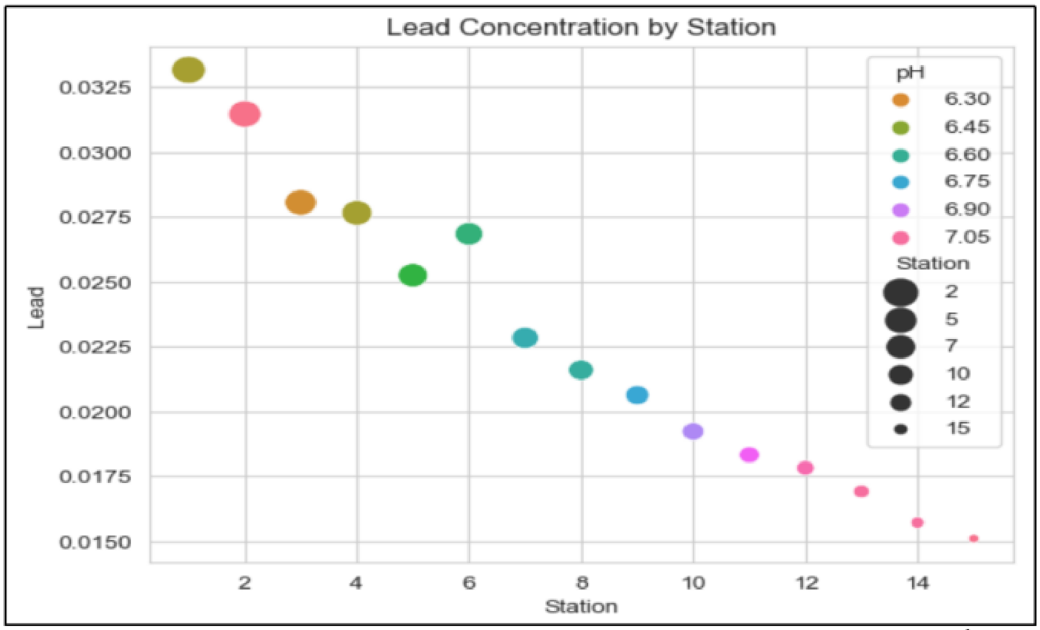
Figure 3: Linear regressions between the concentrations of the heavy metals in the groundwater (a) Mn vs Fe, (b) Mn vs Zn, © Mn vs As, (d) Mn vs Pb, (e) Fe vs Zn, (f) Fe vs As, (g) Fe vs Pb, (h) Zn vs Pb and (i) As vs Pb

3.4.1 Evaluation for Concentration of Fe, Pb and As in groundwater

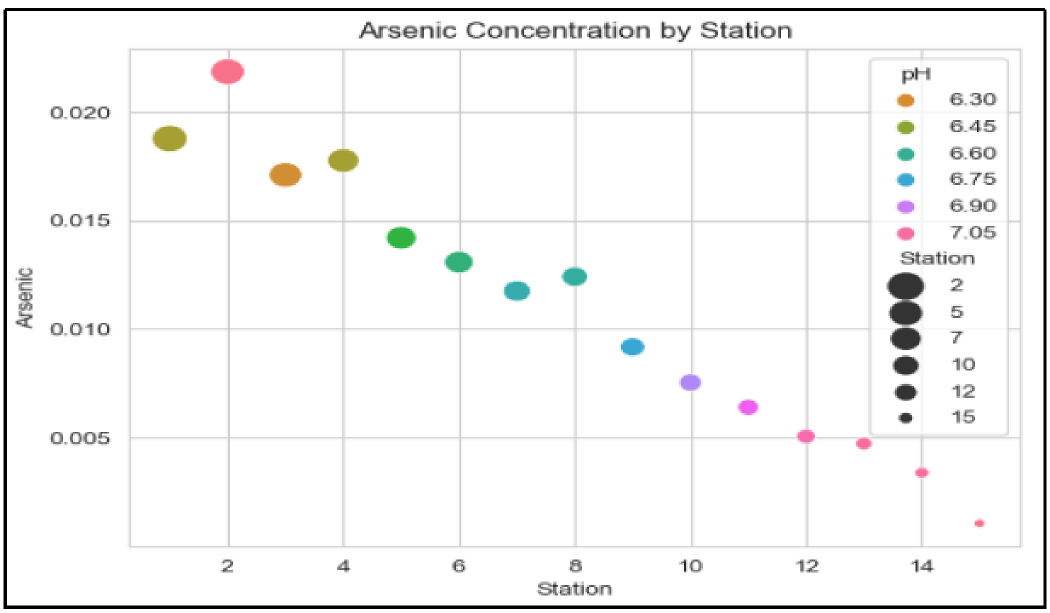
The scatter plot in Figure 4 (a), (b) and (c) visually represent Fe, Pb and As concentration across the 15 stations respectively, with an emphasis on the pH levels. The hue of the markers intensifies as the acidity increases, while the size of the markers corresponding to higher Fe concentrations. This visualization effectively illustrates the positive correlation between Fe concentration and acidity as seen in Figure 4 (a). Similar to iron, the markers' color intensifies with increasing acidity, and their size increases with increasing Pb concentrations (Figure 4b). The positive relationship between acidity and Pb concentration was also demonstrated by this visualization. The behavior observed in the relationship between As concentration and acidity remained consistent with the trends seen in the cases of Fe and Pb as seen in Figure 4 (c).



(a)



(b)

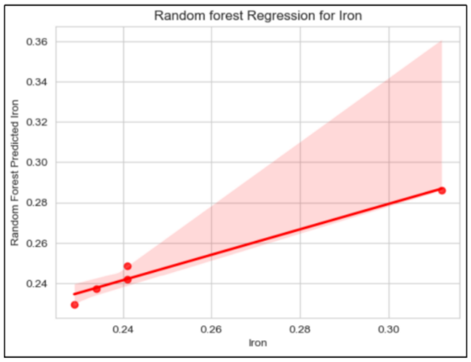


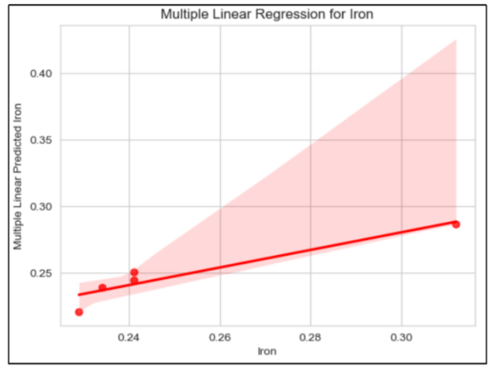
(c)

Figure 4: (a) Fe, (b) Pb and (c) As concentrations in groundwater stations

3.4.2 Evaluation for RFM and MLR for Concentration of Fe, Pb and As in groundwater

For comparative evaluation of RFM and MLR suitability for predicting Fe concentrations, RFM outperformed the MLR model in terms of R^2 and RMSE metrics. Specifically, the RFM gave a better fit with a higher R^2 value of 0.8410, which indicates over 84% explanation of the variability in Fe concentrations by the predictor variables in the model while the MLR model achieved a slightly lower R-squared value of 0.8182, suggesting that it can explain approximately 82% of the variability in Fe concentrations in the groundwater samples. Furthermore, the RFM also exhibited a lower RMSE value of 0.0122 in comparison with 0.0130 obtained by the MLR model. This indicate 1.22% deviation in RFM and 1.33 in MLR model. The regression plots for (a) RFM and (b) MLR Model for Fe concentration in groundwater is Figure 5 (a) and (b) respectively.

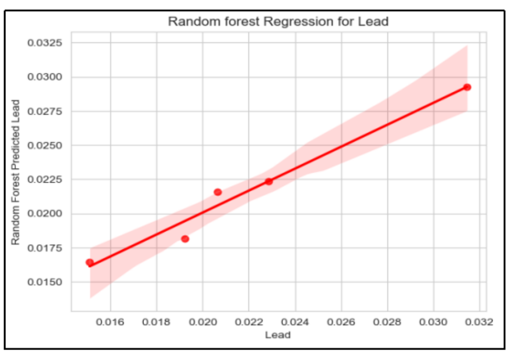


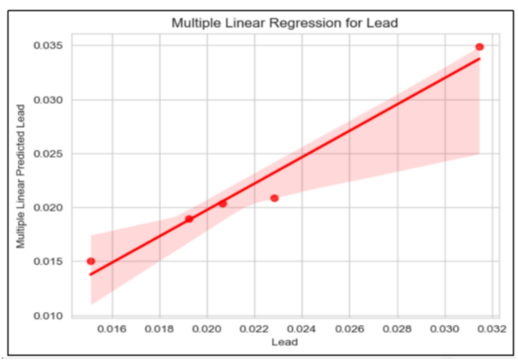


(a) (b)

Figure 5: The regression plots for (a) RF and (b) MLR models for Fe concentration in groundwater

For the prediction of Pb concentrations in groundwater by MLR and RFM, both models showed promise in predicting Pb concentrations, although the RFM model performed better. The RFM model achieved an R-squared value of 0.9389, signifying that approximately 94% of the variation in the Pb concentrations can be elucidated by the predictor variables. In contrast, the MLR model attained a marginally lower R-squared value of 0.8919, suggesting an explanation of about 89% of the variation in Pb concentrations. The superior R-squared value achieved by the RFM model denotes a more optimal fit to the data in comparison to the MLR model. Additionally, the RFM model displayed a lower RMSE value of 0.0013 than the MLR model with RMSE OF 0.0017. Specifically, RFM indicates an average deviation of approximately 0.13% of concentrations of Pb in groundwater while MLR model showed 0.17% deviation of the observed Pb concentrations in groundwater. The regression plots for (a) RFM and (b) MLR Model for Pb concentration in groundwater is Figure 6 (a) and (b) respectively.



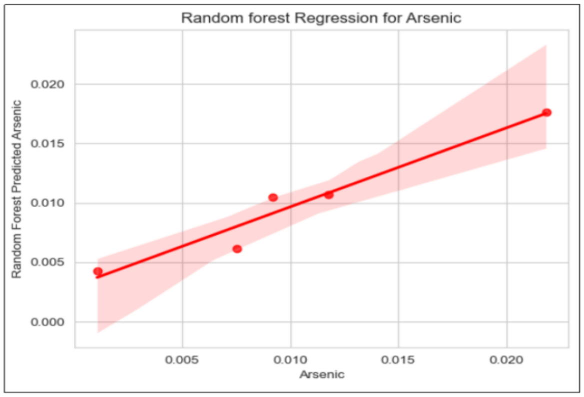


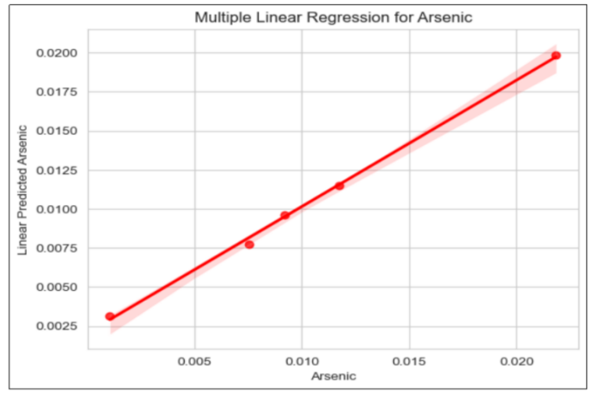
(a) (b)

Figure 6: The regression plots for (a) RFM and (b) MLR models for Pb concentration in groundwater

3.4.3 Evaluation for Concentration of Arsenic

In contrast to the Fe and Pb cases, the MLR model emerged as the superior performer in predicting arsenic concentrations in groundwater. Specifically, the MLR model exhibited an R^2 value of 0.9619. This value indicates that approximately 96% of the variance in As concentrations could be explained by the predictor variables in the model. Moreover, the RMSE associated with the MLR model was measured at 0.0013. This low RMSE value suggests that, on average, the predictions made by the MLR model deviated from the observed As concentrations by only 0.13%. In contrast, the RFM model, while still performing well, showed slightly lower results for the prediction of As concentrations. The RFM model yielded an R^2 value of 0.8577, indicating that it could explain approximately 86% of the variance in As concentrations. Furthermore, the corresponding RMSE value for the RFM model was 0.0025, which is slightly higher than that of the MLR model. These comparative metrics suggest that, for the As case, the MLR model provided more accurate and reliable predictions compared to the RFM model. The higher R-squared value and lower RMSE associated with the MLR model indicate a better fit to the data and a smaller prediction error. This outcome underscores the importance of considering the specific characteristics of each metal species and selecting the most appropriate modeling approach accordingly. The regression plots for (a) RFM and (b) MLR Model for As concentration in groundwater is Figure 7 (a) and (b) respectively.





(a)……………………………………………………………..(b)

Figure 7: The regression plots for (a) RFM and (b) MLR for As concentration in groundwater

3.4.4 Most Influential Environmental Factors Contributing to Groundwater Contamination

Tables 4, 5 and 6 gives the feature importance data of groundwater heavy metal contaminants indicating the most significant factors (heavy metals) affecting groundwater contamination using Fe, Pb and As respectively. From the tables, it can be observed that pH was the most important explanatory variable to describe the groundwater contamination by As, Fe and Pb in this study.

Table 4: Features importance of Fe contamination

Index	Features	Feature_Importance
0	pH	0.2948
1	Manganese	0.1979
2	Iron	0.2087
3	Zinc	0.1186
4	Lead	0.1799

Table 5: Features importance of Pb contamination

Index	Features	Feature_Importance
0	pH	0.3054
1	Manganese	0.2251
2	Zinc	0.1307
3	Arsenic	0.1567
4	Lead	0.1821

Table 6: Features importance of As contamination

Index	Features	Feature_Importance
0	pH	0.3066
1	Manganese	0.1773
2	Iron	0.1813
3	Zinc	0.1399
4	Arsenic	0.1950

CONCLUSION

The RFM algorithm demonstrated superior performance over the MLR algorithm in predicting concentrations of the heavy metals (iron and lead) with higher R squared value and lower RMSE except for Arsenic. The observed differences in model performance can be attributed to the intricate nature of heavy metal behavior in groundwater. From the study it can be concluded that both MLR and RFM can be used to model or predict groundwater contamination with heavy metals. RFR is therefore considered a very promising technique for large-scale modeling of groundwater heavy metal contamination.

REFERENCES

1. Basu, A., Saha, D., Saha, R., Ghosh, T. and Saha, B. (2014). A review on sources, toxicity and remediation technologies for removing arsenic from drinking water. Res. Chem. Intermediation, 40:447–485.

2. Pandey, H.K., Duggal, S.K. and Jamatia, A. (2016). Fluoride contamination of groundwater and its hydrological evolution in District Sonbhadra (U.P.) India. Proc Nat Acad Sci India Sect A Phys Sci., 86:81–93.

3. Subba Rao, N., Ravindra B, Wu J (2020) Geochemical and health risk evaluation of fuoride rich groundwater in Sattenapalle Region, Guntur district, Andhra Pradesh, India. Hum Ecol Risk Assess, 26:2316–2348

4. He, X., Li, P., Ji, Y., Wang, Y., Su, Z. and Elumalai, V. (2020). Groundwater arsenic and fuoride and associated arsenicosis and fuorosis in China: occurrence, distribution and management. Expo Health, 12:355–368.

5. Li, P., Karunanidhi, D., Subramani, T. and Srinivasamoorthy, K. (2021). Sources and Consequences of Groundwater Contamination. Archives of Environmental Contamination and Toxicology (2021) 80:1–10.

6. Salem, M., Mohamed, E.L., Mossad, M. and Mahanna, H. (2022). Random forest modeling and evaluation of the performance of a full-scale subsurface constructed wetland plant in Egypt. Ain Shams Engineering Journal, 13(6): 101778.

7. Haggerty, R., Sun, J., Yu, H. and Li, Y. (2023). Application of machine learning in groundwater quality modeling - A comprehensive review, Water Research, 233: 119745.

8. Cracknell MJ (2014) Machine learning for geological mapping: algorithms and applications. Ph.D. University of Tasmania, Tasmania.

9. Madani, A., Hagage, M. and Elbeih, S.F. (2022). Random Forest and Logistic Regression algorithms for prediction of groundwater contamination using ammonia concentration.Arabian Journal of Geosciences, 15(1619): 1 - 12.

10. Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classifcation. Pattern Recognit. Lett., 27 (4):294–300.

11. Pham BT, Prakash I (2019) A novel hybrid model of Bagging-based Naïve Bayes Trees for landslide susceptibility assessment. Bull. Eng. Geol. Environ., 78(3):1911–1925.

12. Ouedraogo, I., Defourny, P. and Vanclooster, M. (2018). Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeology Journal, 4(1), 23 - 29.

13. Cho, K.H., Sthipong, S., Pachepsky, Y.A., Kyoung-Woong, K. and Kim, J.H. (2011). Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. Water research, 45(17): 5535-5544.

14. WHO (2017). Guidelines for drinking-water quality, 4th edition, incorporating the 1st addendum. World Health Organization.

15. NIS (2015). Nigerian Standard for Drinking Water Quality. Federal Ministry of Water Resources, Nigeria.

16. Cirino, E. (2019). *What pH Should My Drinking Water Be?* Healthline Media LLC. pH of Drinking Water: Acceptable Levels and More (healthline.com).

17. Omonona, A.O., Solotan, T.V. and Adetuga, A.T. (2020). Heavy Metals Contamination in Water, Soil and Plant Samples of Okomu National Park, Edo State, Nigeria. Afr. J. Biomed. Res., 23: 29 - 34.

18. Odiana, S. and Edosomwan, E.U. (2019). Heavy metal concentrations in domestic water sources in some communities of Edo state, Nigeria. FUW Trends in Science & Technology Journal, 4(2): 539 - 544.

19. Wei, H., Yu, H., Zhang, G., Pan, H., Lvb, C. and Meng, F. (2018). Revealing the correlations between heavy metals and water quality, with insight into the potential factors and variations through canonical correlation analysis in an upstream tributary. Ecological Indicators, 90: 485–493.

20. Canli, M. and Atli, G. (2003). The relationships between heavy metal (Cd, Cr, Cu, FE, Pb, Zn) levels and the size of six Mediterranean fish species. *Environ. Pollut.* 121 (1): 129–136.