Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43) 3** : 1–6.
ISSN-0037-8712
doi:10.5269/bspm.78472

Statistical Modeling of Groundwater Pollution Using Logistic Regression

Aseel A. Jaaze, Mohanad N. Abdul Sayed* and Rana H. Shamkhi

ABSTRACT: Binary logistic regression is used to examine groundwater pollution and determine the main contributing elements. Groundwater level, Northing coordinates, and Total Dissolved Solids (TDS) were the hydrogeological and geographical variables that were examined using SPSS. The study showed the significant impact of variables on the state of pollution. The results provide important information for methods to mitigate pollution and manage groundwater resources.

Key Words: Groundwater, logistic Regression, pollution.

Contents

1. Introduction

Groundwater is a major source of fresh water in arid and surface water-poor areas. One of the most important reasons that make it vulnerable to pollution is human waste, represented by urban expansion, liquid industrial waste, and agricultural runoff. To evaluate the impact of hydrogeological factors and model groundwater pollution statistically, we will use binary logistic regression. Combating pollution requires developing sustainable water management methods. There are significant concerns about the safety of groundwater in Al-Ghareeb-Marawah area of Iraq due to poor monitoring and increasing land use. Groundwater pollution means the deterioration of groundwater quality due to the introduction of hazardous chemical, biological or physical substances into the aquifer. The main causes of pollution include improper disposal of waste, excessive use of chemical fertilizers, water leakage from irregular waste dumps, and leakage of industrial waste. These pollutants cause a change in the natural composition of groundwater, making it unfit for human and agricultural use for studies on groundwater pollution, see [9,5]. Given the difficulty of observing groundwater pollution, unlike surface water pollution, which is clearly visible, long periods may pass without groundwater pollution being noticed. For this reason, predictive models must be developed to assess the potential for contamination.

* Corresponding author.
2010 *Mathematics Subject Classification:* 62XX, 6207.
Submitted August 18, 2025. Published October 07, 2025

2. Logistic Regression

Logistic regression is one of the most common statistical methods for modeling and estimating the probability of a binary outcome see [2]. For groundwater contamination, it is either contaminated (1) or uncontaminated (0), the dependent variable in our study. The independent variables are: total dissolved solids (TDS_ppm), groundwater level (Water_Level_m), and north coordinates.

The logistic regression model follows the form:

$$P(Y=1) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

Here, $P(Y=1)$ represents the probability that a groundwater sample is polluted. The variables x_i are the predictor variables Represented by (Northing, Water Level, and TDS), respectively, while β_i are the regression coefficients determined through data analysis.

Logistic regression is very good for environmental studies because of its ability to model the relationship between multiple environmental conditions and the probability of contamination.

3. Data

The data used in our study are from hydrogeological assessments of the Upper Cretaceous aquifer located in the Gharib-Marwa area. Specifically, data were collected and processed from 21 existing groundwater wells. The data included spatial coordinates (east and north), elevation, water level, aquifer characteristics (crest and thickness), and total dissolved solids (TDS). The contamination status of each well was determined based on a specific TDS threshold (e.g., TDS $>$ 1000 ppm indicates contamination). These data were taken from [1], who provided comprehensive GIS-based hydrogeological maps and an assessment of the aquifer system in this area. Binary logistic regression analysis was applied to examine the relationship between groundwater contamination (binary dependent variable: polluted or uncontaminated) and independent predictors: north, water level (m), and total dissolved solids (TDS) per million. The aim of this study was to identify pollution indicators and evaluate the accuracy of model predictions.

3.1. Table of data

The data is shown in the Table 1:

The following variables represent:

The dependent variable is the pollution status (polluted = 1, unpolluted = 0). The independent variables are as follows:

- North: Represents the geographical location.
- Groundwater level (Water Level) Water salinity (TDS): It is a key variable in determining groundwater contamination.

A new variable for contamination status has been added based on a specific salinity threshold (e.g., TDS $>$ 1000 ppm is considered a contaminant).

Table 1: Summary of Hydrogeological and chemical characteristics for 21 groundwater wells

Well_No	Easting	Northing	Elevation_m	Water_Level_m	Aquifer_Top_m	Aquifer_Thickness_m	TDS_ppm	Pollution_Status
3590-3-B14	508500	3602300	500	390	450	50	416	0
3589-4-H4	508300	3593750	474	452	451	45	450	0
3590-3-C7	519950	3598500	390	365	367	110	470	0
3590-3-C17	519985	3603300	360	330	395	80	432	0
3590-3-C19	519873.1	3602209	358	338	455	95	432	0
3590-3-C20	519397	3600862	356	343	420	116	495	0
3590-3-C5	517950	3604050	354	322	340	118	441	0
3590-3-B15	514881.3	3598488	420	406	430	70	467	0
3590-3-III-C18	514376.7	3602523	451	427	450	50	439	0
3590-3-C23	521941.1	3602180	357	350	550		494	0
3590-2-A9	523906.6	3602878	370	316			788	0
3589-4-D-2	507700	3585400	495	285	370	20	1276	1
3589-4-H11	513150	3586775	480	268	380	120	972	1
3589-3-H-MEW2	539710.1	3593443	503	293	291	30	453	0
3589-I-G-WMEW2	530466	3588210	444	254	255	229	1030	1
3590-III-C22	520990.8	3600391	375	401	500	50	500	0
3590-III-C14	520159.1	3602731	341	321	380	60	460	0
3589-I-MEW1	540873.9	3599252	505	267	290	196	445	0
3589-I-G-WMEW1	529243.6	3592956	438	279	275	135	260	0
3590-II-F-ZEW1	543169.1	3609433	425	392	392	193	383	0
3590-II-F30	541425	3605675	522	470	470		500	0

4. Data Analysis and Results

We will now use SPSS to examine the data and use logistic regression to model groundwater pollution, see [7].

4.1. Data processing

Data quality was checked before analysis to determine if there were any missing values.

None of the 21 groundwater well records have any missing information, guaranteeing the dataset's completeness and integrity. As a result, the binary logistic regression model may be applied continuously without requiring data exclusion or imputation, as shown in the Table 2:

Table 2: Summary of data completeness and quality assessment

Category	Number (N)	Percentage (%)
Values used in the analysis	21	100%
missing values	0	0%

4.2. Evaluate the basic model without the independent variables

Initially, no predictor variables were included in the baseline model, also known as the null model. The model achieved an overall classification accuracy of 85.7% by classifying all cases as "unpolluted" (status = 0). This model's inability to detect any of the real contaminated instances (status = 1), however, underscores both its prediction limitations and the need to include pertinent independent variables, as shown in the Table 3:

Table 3: Summary of data completeness and quality assessment

Reality	Expectation (Unpolluted)	Expectation (Polluted)	Percentage (%)
Unpolluted(0)	18	0	100%
Polluted(1)	3	0	0%
Total accuracy	85.70		

4.3. The importance of variables before entering them into the model

All three variables: Northing, Water_Level_m, and TDS_ppm have a significant effect on pollution ($P < 0.05$).

TDS_ppm (water salinity) is the most influential variable, followed by geographical location (Northing), then groundwater level (Water_Level_m).

An overall P value of 0.000 means that the independent variables have a significant effect on improving the model, as shown in the Table 4.

Table 4: Significance of predictor variables based on Chi-Square Test

variable	Chi-Square	df	Sig. (P-Value)
Northing (Geographical Location - North)	12.952	1	0.000
Water_Level_m (Groundwater Level)	5.574	1	0.018
TDS_ppm (Groundwater Salinity)	17.335	1	0.000
Total Statistics	18.551	3	0.000

4.4. Model significance after introducing variables

The whole model showed a statistically significant improvement over the null model when the independent variables were added, with a chi-square value of 17.225 ($df = 3$, $p = 0.001$). This shows that all of the predictors together significantly increase the explanatory power of the model, as shown in the Table 5:

Table 5: Overall model significance after including predictors

Chi-Square	df	Sig. (P-Value)
17.225	3	0.001

4.5. Model quality

The performance of the model was further assessed through the use of pseudo R-squared statistics. An extraordinarily high level of explanatory power is suggested by the Nagelkerke R^2 value of 1.000, which shows that the model fully explains the variance in the dependent variable. This can indicate a very good model fit. The strength was further supported by the Cox and Snell R^2 of 0.56 and the -2 Log Likelihood value of 0, as shown in the Table 6:

Table 6: Model quality indicators using pseudo R-squared Measures

-2 Log Likelihood	Cox & Snell R^2	Nagelkerke R^2
0	0.56	1.000

5. Discussion

Many recent studies [3,4,6], Recent work highlights arsenic contamination [8,10], have applied logistic regression to evaluate groundwater contamination and validate its effectiveness in environmental modeling. For instance: Iqbal et al in 2023 [3], conducted a comparative analysis between logistic regression, artificial neural networks, and random forest models to predict arsenic contamination in Pakistan. Their findings highlighted the reliability and interpretability of logistic regression despite its simplicity. Perez et al. [9], combined logistic regression with the Analytical Hierarchy Process (AHP) to identify the main factors contributing to E. coli contamination in shallow groundwater. Ahmed et al. [5], used logistic regression alongside random forest models to assess groundwater contamination related to ammonia levels.

6. Conclusions

Based on logistic regression analysis, total dissolved solids (TDS) are identified as the primary factor driving groundwater pollution in the study area. Elevated salinity levels are closely associated with contaminated wells. Geographic location, represented by the northing coordinate, also plays a notable role, with higher contamination rates observed in the southern part of the region. Groundwater level has an inverse relationship with pollution but contributes less significantly compared to the other factors. The final model displayed excellent predictive ability, correctly classifying all wells with 100

7. Recommendations

Based on the study's outcomes, several recommendations are proposed:

1. Regularly monitor TDS concentrations to assess water quality trends.

2. Focus pollution control efforts on the southern section of the aquifer where contamination is more prevalent.

3. Implement regulatory measures to limit industrial, agricultural, and domestic activities near groundwater sources.

4. Support future research integrating additional hydrochemical indicators and spatial models for more comprehensive pollution prediction.

References

1. S. Hamad. Gis based evaluation of upper cretaceous aquifer in al ghareeb marawah area. In *Proceedings of the GI_Forum 2008*, pages 1–9, Austria, 2008.
2. D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, Hoboken, NJ, USA, 2013.
3. J. Iqbal, C. Su, M. Ahmad, M.Y.J. Baloch, A. Rashid, Z. Ullah, H. Abbas, A. Nigar, A. Ali, and A. Ullah. Hydrogeochemistry and prediction of arsenic contamination in groundwater of vehari, pakistan: comparison of artificial neural network, random forest and logistic regression models. *Environmental Geochemistry and Health*, 46(1):14, 2024.
4. A. Madani, M. Hagage, and S.F. Elbeih. Random forest and logistic regression algorithms for prediction of groundwater contamination using ammonia concentration. *Arabian Journal of Geosciences*, 15(20):1619, 2022.
5. Ahmed Madani, Mohammed Hagage, and Salwa F Elbeih. Random forest and logistic regression algorithms for prediction of groundwater contamination using ammonia concentration. *Arabian Journal of Geosciences*, 15(20):1619, 2022.
6. A. Mohammaddost, Z. Mohammadi, M. Rezaei, H.R. Pourghasemi, and A. Farahmand. Assessment of groundwater vulnerability in an urban area: a comparative study based on drastic, ebf, and lr models. *Environmental Science and Pollution Research*, 29(48):72908–72928, 2022.
7. J. Pallant. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*. Routledge, London, UK, 2020.
8. E.K. Putri, S. Notodarmojo, and R.R. Utami. Investigating dominant factors of coliform contamination in shallow groundwater: A logistic regression and ahp approach. *Groundwater for Sustainable Development*, 27:101384, 2024.
9. Enda Kalyana Putri, Suprihanto Notodarmojo, and Rosetyati Retno Utami. Investigating dominant factors of coliform contamination in shallow groundwater: A logistic regression and ahp approach. *Groundwater for Sustainable Development*, 27:101384, 2024.
10. Z. Zhao, A. Kumar, and H. Wang. Predicting arsenic contamination in groundwater: A comparative analysis of machine learning models in coastal floodplains and inland basins. *Water*, 16(16):2291, 2024.

Aseel A. Jaaze,
Department of Pharmacognosy and Medicinal Plants,
College of Pharmacy, University of Basrah,
Basrah, Iraq.
E-mail address: aseel.jaaze@uobasrah.edu.iq

and

Mohanad N. Abdul Sayed,
Department of Computer Engineering and Artificial Intelligence Techniques,
Qurna Polytechnic College, Southern Technical University,
Basrah, Iraq.
E-mail address: Mohanad87@stu.edu.iq

and

Rana H. Shamkhi,
Department of Pharmacognosy and Medicinal Plants,
College of Pharmacy, University of Basrah,
Basrah, Iraq.
E-mail address: rana413427@gmail.com