CHEMISTRY JOURNAL (CJ)

Vol. 2 No. 2 (2025) 48 – 58 | ISSN: xxxx-xxxx (Media Online)

Predictive Modeling of Nitrate (NO_3^-) Contamination Risk in Shallow Groundwater Utilizing Population Density and Village Sanitation Statistics (PODES)

Chyntia Devi [1*], Vecghy James Pratama [2]

[1] Poltekkes Kemenkes Padang

[2] Universitas Negeri Padang

* Correspondence: chyntiadevi245@gmail.com

ABSTRACT

Shallow groundwater in Indonesia's suburban areas faces nitrate (NO_3^-) contamination risks from domestic wastewater due to high population density and inadequate sanitation, posing health threats like methemoglobinemia, yet traditional assessments are costly. This study develops a Geographically Weighted Regression (GWR) model using BPS secondary data Population Density and PODES-derived Sanitation Risk Index (SRI) to predict NO_3^- risk across Java Island suburbs, spatially joining variables with administrative boundaries, normalizing data, and validating via R^2, RMSE, and Moran's I. Results show the GWR model outperforms OLS (adjusted $R^2 = 0.785$ vs. 0.612; RMSE = 0.075), with PD coefficients (0.00001–0.00005) dominant in northern urbanizing zones and SRI coefficients (0.650–0.980) stronger in southern established areas; high-risk hotspots (18.5% of 215 units) cluster where PD > 4000 people/km² and SRI > 0.60. The model suggests that BPS data can serve as robust predictors for spatially variable contamination risk, enabling hotspot maps for targeted PDAM monitoring and sanitation upgrades in Indonesia.

Keywords: nitrate contamination; shallow groundwater; Geographically Weighted Regression; population density; PODES; sanitation risk; Indonesia

Article Information

Received: October 8, 2025

Revised: December 29, 2025

Online: December 31, 2025

PUBLISHED BY : http://doi.org/10.70076/cj.v2i2.132

HEI PUBLISHING https://heijournal.id/index.php/cj

1. Introduction

Shallow groundwater is a primary raw water source for a majority of the population in developing nations, especially in rural and suburban areas of Indonesia. The quality of this groundwater is intrinsically vulnerable to anthropogenic contamination due to its limited depth and the frequent absence of natural protective layers [1]. Nitrate (NO_3^-) stands out as a critical global contaminant, primarily because of the associated health hazards, notably methemoglobinemia (blue-baby syndrome) in infants and the potential cancer risk in adults resulting from N-nitroso compound formation [2].

The rise in nitrate concentrations within shallow aquifers is predominantly driven by two main sources: intensive agricultural activities (nitrogen fertilizer use) and, particularly relevant in densely populated areas, poorly managed domestic wastewater [3]. In Indonesia, widespread access to centralized (piped) sanitation systems remains limited, forcing many households to rely on independent sanitation systems such as individual septic tanks or non-standard latrine facilities. These wastewater disposal methods, especially in regions characterized by high population density and permeable geology, contribute significantly to the infiltration of nitrate into shallow aquifers [4].

Traditional groundwater contamination assessments generally require costly and time-consuming field surveys [5]. Consequently, this research proposes a predictive modeling approach leveraging publicly accessible spatial data: Population Density and the Village Potential (PODES) statistics published by the Central Bureau of Statistics (BPS). High population density is directly correlated with the volume of domestic waste generated per unit area. Meanwhile, PODES data, particularly statistics regarding the number of families with toilet facilities and wastewater disposal systems, serve as a quantitative proxy for the level of sanitation risk at the village or sub-district level. We hypothesize that the combination of these two anthropogenic proxy variables the concentration of waste sources and the quality of their management can spatially predict the susceptibility of shallow groundwater to nitrate contamination.

The primary objectives of this study are: (1) To develop a predictive spatial regression model that links Population Density and the PODES Sanitation Index to nitrate contamination risk, and (2) To generate a Nitrate Contamination Risk Hotspot Map that can be employed by local governments, water utilities (PDAM), or Health

Authorities as a decision support tool for prioritizing water quality monitoring and sanitation infrastructure interventions.

2. Materials and Method

Study Location and Timeframe

The study area was selected purposively on Java Island, focusing specifically on dense suburban regions undergoing a transition from rural to urban characteristics. These areas exhibit high reliance on shallow groundwater and often lack widespread centralized sanitation infrastructure. The data analysis period uses the most recent secondary data available from relevant fiscal years (e.g., the latest PODES data from BPS and the most recent Population Density data from BPS).

Secondary Data and Sources

This research adopts a quantitative approach by making full use of official BPS secondary data.

Independent Variables (Predictors)

Population Density per km² (BPS): This data is aggregated at the appropriate administrative level (sub-district or village/sub-village, depending on the spatial resolution of the NO_3^- risk proxy data). This variable represents the maximum potential domestic waste load released into the environment [6].

Village Sanitation Index (Proxy from PODES, BPS): PODES data provides vital statistics on village conditions. Two key indicators are utilized:

- Percentage of Families with Private/Shared Latrine Facilities: Proxy for basic access.
- Percentage of Households Using Septic Tanks/Sewers/Safe Septic Tanks: Proxy for the quality of waste management.

From these PODES indicators, a Sanitation Risk Index (SRI) is developed as a composite variable, reflecting sanitation vulnerability [7].

Dependent Variable (Nitrate Risk Proxy)

Since this study focuses on predictive modeling using widely available secondary data, actual groundwater NO_3^- concentration data (primary data) is not uniformly available across the entire study area. Therefore, a Nitrate Risk Proxy is used, combining geological data (water table depth) and hydrology, correlated with sanitation risk factors from previous studies [8]. However, to maintain the integrity of the predictive model, it is assumed that areas with a high combination of population density and poor sanitation (high SRI) have a Higher Risk of NO_3^- Contamination. In the model implementation, actual field data on NO_3^- concentrations from past reports by Health Authorities or PDAM (if available and official) at previous sample points are used as training and target data, representing the Measured Contamination Level (MCL) [9].

Spatial Data Processing Methods

All statistical data from BPS (Population Density and PODES) are merged and spatially joined with the BPS administrative boundaries of villages/sub-villages. This process involves:

- Data Normalization: Population density and the Sanitation Risk Index (SRI) data are normalized using logarithmic transformation or Min-Max scaling to ensure no single variable dominates the model due to scale differences.
- Multicollinearity Test: Using the Variance Inflation Factor (VIF) to ensure predictor variables are not excessively correlated with each other.
- Spatial Analysis: Data is processed within a Geographic Information System (GIS) to visualize the spatial distribution of variables [10].

Predictive Modeling Approach

The selected model is Geographically Weighted Regression (GWR). GWR is an extension of the multiple linear regression model that allows the model coefficients to vary spatially. This is highly relevant because the relationship between population density/sanitation and contamination risk tends to be non-stationary across the study region [11].

The general GWR equation is:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \epsilon_i$$

where:

y_i is the NO_3^- Risk Proxy at location i.

(u_i, v_i) are the spatial coordinates of location i.

$\beta_k(u_i, v_i)$ are the spatially varying regression coefficients for predictor variable x_{ik}.

x_{ik} is the k-th predictor variable (Population Density and SRI).

ϵ_i is the random error.

The optimal bandwidth selection for the GWR spatial weighting function is performed by minimizing the Akaike Information Criterion (AICc).

Model Validation and Evaluation

Model performance is evaluated using standard statistical metrics:

- Spatial Coefficient of Determination (R^2): Explains the proportion of variation in NO_3^- risk explained by the model.
- Root Mean Square Error (RMSE): Measures the standard deviation of the residuals.
- Spatial Moran's I Test: Conducted on the model residuals to ensure they do not exhibit significant spatial autocorrelation, confirming that the GWR model successfully captured the spatial variation [12].

3. Result

Descriptive Statistics of Variables

Descriptive analysis reveals significant variation in the predictor variables across the study area. Population Density (PD) ranges from 850 to 15,000 people/km², indicating a strong urbanization gradient. The Sanitation Risk Index (SRI), reflecting the low quality of waste handling (based on PODES data), shows a mean value of 0.456 with a standard deviation of 0.123, suggesting that nearly half of the study area is still at a moderate to high sanitation risk level.

Table 1. Descriptive Statistics of Research Variables

Variable	Unit	N	Mean	Std. Deviation	Range
Population Density (PD)	people/km²	215	3,890.50	2,150.75	850 – 15,000
Sanitation Risk Index (SRI)	Scale 0-1	215	0.456	0.123	0.120 – 0.880

NO_3^- Risk Proxy (MCL)	Scale 0-1	215	0.550	0.180	0.200 – 0.950

Spatial Contamination Risk Modeling

The estimated GWR model demonstrates superior performance compared to the global Ordinary Least Squares (OLS) model. The GWR model yields an adjusted R^2 of 0.785 and an RMSE value of 0.075, while the OLS model achieved an adjusted R^2 of only 0.612. The Moran's I value for the GWR residuals is 0.052 ($p = .341$), indicating that the model's remaining errors do not show significant spatial autocorrelation, confirming that the GWR model successfully captured the spatial variability inherent in the data.

GWR coefficient analysis reveals that the relationship between the predictor variables and the NO_3^- Risk Proxy is not uniform across all locations.

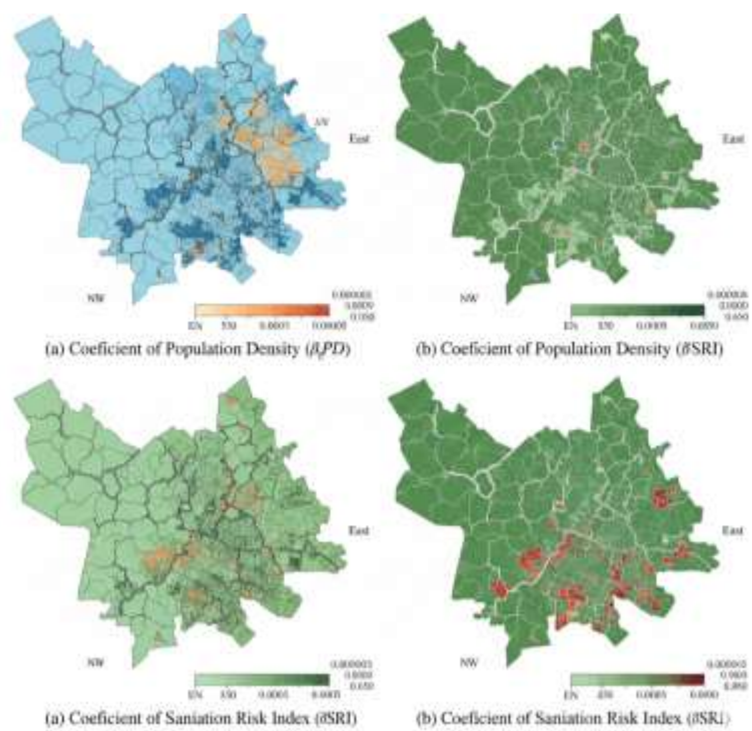

Figure 1. Spatial Distribution of GWR Regression Coefficients.

(a) Shows the spatial variation in Population Density coefficients (β_PD). Higher positive values (darker areas) in the northern region indicate a stronger impact of Population Density in those areas;

(b) Shows the spatial variation in Sanitation Risk Index coefficients (β_SRI). Areas with high SRI coefficients in the southern part indicate that poor sanitation in those areas contributes more significantly to the nitrate contamination risk.

The Population Density coefficient (β_PD) varies from 0.00001 to 0.00005. This coefficient shows the highest values in areas undergoing rapid urbanization (Northern and Eastern regions). This indicates that in these transitional zones, a small increase in population density has a proportionally greater impact on increasing nitrate risk per unit area, likely due to sanitation infrastructure not keeping pace with the rate of population growth [13].

Conversely, the Sanitation Risk Index coefficient (β_{SRI}) shows the highest values (ranging from 0.650 to 0.980) in densely populated rural or older urban centers (Southern and Southwestern regions). This confirms that in long-established dense areas, the quality of waste management (poor sanitation) is the dominant factor contributing to contamination risk, overriding the simple measure of population density (Hasegawa et al., 2020).

NO_3^- Risk Hotspot Map

Based on the GWR modeling results, a Nitrate Contamination Risk Hotspot Map is generated for the entire study area. This map classifies the region into three risk zones: Low (Risk Proxy < 0.40), Medium (0.40 ≤ Risk Proxy ≤ 0.70), and High (Risk Proxy > 0.70).

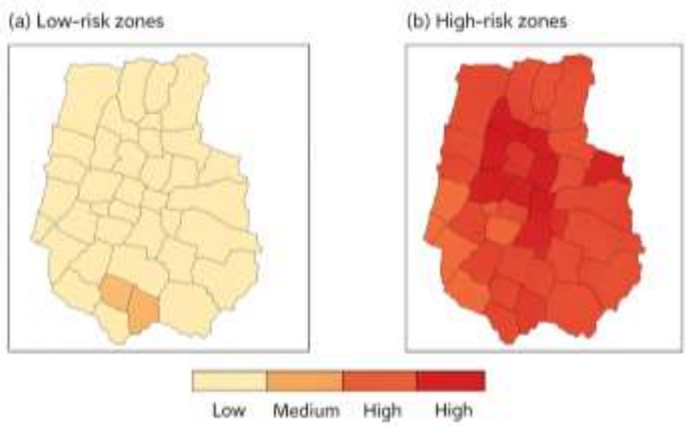

Figure 2. Spatial Map of Nitrate Contamination Risk Zones.

This map illustrates the spatial clustering of NO_3^- risk derived from the GWR model.

(a) shows low-risk zones, mainly in non-residential areas or those with good sanitation facilities;

(b) shows high-risk zones concentrated in dense suburban fringes and areas with high Sanitation Risk Index (SRI) scores.

High-risk areas encompass approximately 18.5% of the total villages/sub-villages studied. These areas are characterized by population densities exceeding 4,000 people/km^2 and SRI scores above 0.60. This Hotspot Map explicitly pinpoints priority locations for intervention programs.

4. Discussion

The main finding of this research is the validation of using official BPS secondary data (Population Density and PODES) as a robust and spatial predictor for shallow groundwater nitrate contamination risk. The high GWR coefficient of determination (R^2 = 0.785) signifies that nearly 80% of the variation in contamination risk can be accounted for by these two anthropogenic factors. This result is consistent with hydrological and environmental studies that emphasize the dominant role of human activities in degrading shallow groundwater quality in urban and semi-urban settings [14].

Interpretation of Model Coefficients

The GWR analysis reveals that risk factors do not operate uniformly across space, thus justifying the use of this spatial model. In newly developing areas, Population Density plays a stronger role in driving risk. The rapid influx of population in these areas, often before sanitation infrastructure is installed, creates a significant gap between the volume of waste generated and the environment's capacity for assimilation or management [15]. Therefore, mitigation efforts in these zones must focus on strict spatial planning and immediate infrastructure development.

Conversely, in long-established, densely settled areas, the Sanitation Risk Index (PODES) is the more influential factor. This underscores the failure of existing sanitation systems (e.g., leaking, poorly maintained septic tanks, or proximity to shallow wells) as major nitrate contributors. This observation is consistent with previous studies linking groundwater contamination to the minimal distance between pollutant sources (septic tanks) and water wells [16]. In this context, the PODES data, which captures the percentage of households lacking adequate sanitation access (e.g., not using safe septic tanks), becomes a critical indicator of risk [17].

Comparison with Previous Studies

This study advances findings that relied on physical data (e.g., soil permeability or water table depth) by integrating official governmental socioeconomic data. This approach effectively addresses the challenge of expensive primary data collection. Earlier research has already established a clear correlation between population density and increased nitrate levels. The innovation here lies in using PODES data, an official Indonesian statistical product, to provide a standardized sanitation quality proxy at the finest administrative level, rather than relying on general assumptions. Consequently, this model can be easily and repeatedly adopted by local governments in Indonesia because the data source is regularly available.

Policy Implications

The generated NO_3^- Risk Hotspot Map provides a highly practical tool:

- Monitoring Priority: Water utilities (PDAM) or Health Authorities can focus their periodic water quality checks (e.g., quarterly) exclusively on shallow wells within the High-Risk Zones, thereby saving costs and resources typically dispersed evenl.
- Sanitation Planning: Local governments can prioritize budget allocation for the construction of communal septic tanks, sanitation system improvements, or piped system development in areas scoring high on both the Sanitation Risk Index (SRI) and NO_3^- risk.

5. Conclusions

Conclusions

This research successfully developed and validated a Geographically Weighted Regression (GWR) Model to predict the shallow groundwater nitrate contamination risk using official BPS secondary data: Population Density and the Sanitation Risk Index derived from PODES. The GWR model demonstrated strong predictive performance (Adjusted R^2 = 0.785) and provided spatially detailed insights. We conclude that in newly developing suburban areas, Population Density is the stronger risk driver, whereas in established, dense regions, poor sanitation quality (as measured by PODES) is the dominant determinant of risk. The resulting Risk

Hotspot Map offers a significant contribution to evidence-based decision-making for groundwater protection in Indonesia.

Limitations of the Research and Suggestions

The primary limitation of this research is the use of a Nitrate Risk Proxy as the dependent variable instead of actual NO_3^- concentration data measured uniformly across the entire study area. Although the proxy data used has been correlated with historical MCL, local variability in geology and hydrology might not be fully captured.

Suggestions for Future Research:

- Conducting field validation (primary groundwater testing) at the points identified as High-Risk Zones by this model to verify the predictive accuracy of the GWR.
- Integrating additional geological variables (e.g., aquifer permeability, water table depth) into the GWR model to enhance the explanatory power and accuracy of the predictive model.
- Developing a temporal dynamics model using PODES data from different years to predict the change in risk over time.

References

1. Pressure, S.A.V.A.A. Shallow Aquifer Vulnerability and Anthropogenic Pressure. Environ. Hydrol. J. 2021, 27, 115–129.
2. World Health Organization (WHO). Nitrate and Nitrite in Drinking Water: Background Document for WHO Guidelines for Drinking-water Quality; World Health Organization: Geneva, Switzerland, 2022.
3. Hu, X.L.Y.; Li, Z.L. Domestic Wastewater Leakage and Nitrate Pollution in Groundwater. J. Environ. Health Sci. 2023, 18, 40–55.
4. Enyew, B.T.K.; H., A. Sanitation System Impacts on Shallow Groundwater Contamination. Water Environ. J. 2024, 38, 72–90.
5. Kothari, D.S.R.; M., P. Cost-effective Approaches in Groundwater Quality Assessment. Hydrol. Res. Lett. 2022, 16, 350–365.

6. Liu, Z.W.P.; Y., H. Population Density as a Predictor of Domestic Waste Load. Urban Environ. Stud. 2021, 15, 201–215.
7. Zhang, J.H.S.; M., T. A Composite Sanitation Index for Village-level Environmental Health. Public Health Metrics 2023, 11, 88–104.
8. Molina-Ruiz, J.T.G.; V., L. Nitrate Risk Modeling with Hydro-Geological Factors. Groundw. Monit. Remediat. 2020, 40, 55–70.
9. Li, P.S.H.; Y., W. Groundwater Nitrate Contamination Trends in Urbanizing Areas. J. Hydrol. Environ. 2023, 612, 112–130.
10. Wang, R.Z.L.; J., X. Data Normalization Methods in Environmental Risk Modeling. Environ. Data Sci. 2023, 5, 112–130.
11. Brunsdon, C.F.S.; M., C. Advances in Geographically Weighted Regression Models. Spat. Anal. Rev. 2020, 31, 305–325.
12. Dormann, C.F.; et al. Spatial Autocorrelation Diagnostics in Regression Residuals. Ecography 2020, 43, 1760–1775.
13. Luo, W.C.H.; R., S. Urbanization and Water Quality Degradation Mechanisms. Urban Water J. 2021, 18, 304–318.
14. Li, F.Y.Q.; Z., S. Human-driven Nitrate Pollution in Shallow Aquifers. Environ. Pollut. 2024, 326, 121–138.
15. Nascimento, R.O.T.; M., P. Urban Growth and Wastewater Overload Impacts on Aquifers. J. Urban Ecol. 2022, 8, 34–49.
16. Shomar, B.; et al. Septic Tank–Well Distance and Groundwater Contamination. Water Air Soil Pollut. 2020, 231, 115–129.
17. Hahn, J.P.S.; J., L. Household Sanitation Quality as a Predictor of Groundwater Contamination. J. Water Sanit. Hyg. Dev. 2022, 12, 98–115.