

Gema Lingkungan Kesehatan

Vol. 24, No. 4 (2026), pp 607-617

e-ISSN 2407-8948 p-ISSN 1693-3761

doi: <https://doi.org/10.36568/gelinkes.v24i4.591>

Journal Homepage: <https://gelinkes.poltekkesdepkes-sby.ac.id/>

Sanitation Risk Prediction Model Based on G.I.S and A.I to Improve the Residential Sanitation Index in Palembang

Rio Purnama^{1*}, Rahmi Garmini¹, Ayu Dekawaty² Zalmanda Vadinna Arla¹, Arza Yulawati¹

¹ Diploma III Environmental Health Study Program, Universitas Muhammadiyah Ahmad Dahlan Palembang, Palembang, Indonesia

² Nursing Science Study Program, Universitas Muhammadiyah Ahmad Dahlan Palembang, Palembang, Indonesia

*Correspondence: rio.purnama1210@gmail.com

ABSTRACT

Residential sanitation conditions vary spatially across Palembang City, but a transparent and reproducible framework is needed to convert heterogeneous sanitation data into area-level priorities. This ecological cross-sectional study assessed all 18 subdistricts of Palembang by integrating a composite sanitation-risk index with Geographic Information System (GIS) mapping and an exploratory Random Forest analysis. The composite index was constructed from four direct sanitation and environmental indicators: access to improved latrines, access to improved water, wastewater-management access (SPAL), and illegal waste-disposal points. Indicators were min-max normalized to a common 0-1 risk direction and aggregated using equal weights. Population density, building density, diarrhea rate, and dengue rate were included as contextual variables and were not used to construct the index. Risk scores ranged from 13.09 to 73.83. Kertapati had the highest score (73.83), followed by Seberang Ulu II (54.08) and Seberang Ulu I (47.85). Tertile classification identified six high-, six medium-, and six low-priority subdistricts. Random Forest regression was evaluated using Leave-One-Out Cross-Validation because only 18 ecological units were available. The model showed MAE=14.01, RMSE=17.05, and $R^2=-0.15$, and did not outperform a mean-prediction baseline, indicating limited out-of-sample predictive performance in the present dataset. Accordingly, Random Forest results were interpreted as exploratory and not as causal evidence or as an externally validated operational prediction model. The combined GIS and composite-index framework provides a reproducible basis for relative spatial prioritization, while future predictive modeling requires independent sanitation-risk labels, finer spatial units, multi-year data, and external validation.

Keywords: Artificial Intelligence; Geographic Information System; Random Forest; Sanitation Risk

INTRODUCTION

Sanitation is a fundamental component of environmental health and sustainable urban development because inadequate sanitation can increase environmental pollution and the risk of sanitation-related diseases. Access to adequate sanitation, safe water, wastewater management, and proper waste disposal are therefore important components in reducing environmental health risks and improving the quality of residential environments. In Indonesia, sanitation challenges remain particularly complex in urban areas where population growth, settlement expansion, environmental conditions, and unequal access to basic sanitation infrastructure interact across space. Research on environmental sanitation in residential areas along the Musi River in Palembang has also demonstrated that sanitation conditions are closely associated with settlement characteristics and environmental conditions, highlighting the need for area-based sanitation

assessment (Trisnaini et al., 2022; Rahardjo et al., 2025).

Palembang City faces a particularly complex sanitation context because of its rapid population growth, continued residential development, and geographical characteristics that include low-lying and swampy areas that are vulnerable to flooding and water pollution. These conditions can increase the potential for domestic wastewater pollution and other environmental health problems, particularly in densely populated settlements and areas located along the Musi River. Previous assessments in Palembang have identified limitations in sanitation infrastructure and the practice of discharging domestic waste directly into water bodies as important factors contributing to sanitation risks in vulnerable areas of the city. Moreover, sanitation problems are multidimensional and cannot be adequately represented by a single indicator because they involve interactions

among population density, access to sanitation facilities, water supply, wastewater management, waste disposal, settlement density, and environmental health conditions (Iryanto et al., 2022; Purnama R, 2024).

Conventional approaches to sanitation risk assessment, including Environmental Health Risk Assessment, are useful for identifying environmental health risks but tend to provide relatively static assessments and require considerable time and resources. However, conventional sanitation assessment has not always been integrated with spatial analysis and data-driven analytical approaches. Integrating sanitation indicators with GIS may therefore provide a more systematic approach for identifying spatial variations in sanitation risk and supporting area-based intervention planning. Geographic Information Systems (GIS) provide an important framework for addressing this limitation because they allow environmental and sanitation variables to be integrated and visualized spatially. Recent research has highlighted the increasing application of GIS in water, sanitation, and hygiene research, particularly for understanding spatial patterns, identifying vulnerable areas, and supporting evidence-based interventions (Pradhan et al., 2025; Nilandita et al., 2022).

Machine-learning methods may complement spatial assessment by exploring nonlinear relationships among multiple variables. Random Forest is an ensemble tree method that can be used for regression or classification and can represent complex interactions without requiring a prespecified functional form (Breiman, 2001).

However, spatial mapping primarily characterizes the geographical distribution of observed conditions and does not by itself evaluate multivariable predictive relationships. Recent developments in geospatial artificial intelligence and machine learning provide opportunities to complement spatial analysis by exploring complex relationships among environmental, demographic, infrastructure, and health-related variables. Geospatial artificial intelligence has increasingly been applied in environmental epidemiology to integrate spatial information with computational modeling (Iyer et al., 2025), while machine learning approaches have also been used to predict sanitation facility access and environmental health risks (Yitageasu et al., 2025 ; van der Heijden et al., 2025; Kasaluhe et al., 2022). These developments indicate that integrating spatial analysis with machine learning can provide a more dynamic approach to identifying areas with different levels of environmental and sanitation risk.

Among machine learning techniques, Random Forest is particularly relevant for sanitation risk modeling because it can simultaneously process multiple predictor variables and capture complex and nonlinear relationships among them. As an ensemble learning method, Random Forest combines multiple decision trees and can accommodate nonlinear relationships among predictors. Its application in environmental and sanitation research can therefore

provide an exploratory multivariable framework for assessing patterns among demographic, built-environment, and health-related indicators. Recent spatial machine learning research has demonstrated that population density and building density can be important factors in explaining the spatial distribution of urban sanitation problems, while the influence of sanitation-related factors may vary according to location (Yi et al., 2024; Hardiknas Wopari et al., 2022).

Previous studies in Indonesia have demonstrated the usefulness of GIS for sanitation mapping and environmental health assessment, including sanitation-risk mapping at the local level (Ali et al., 2024), web-based sanitation mapping (Widiyatmoko et al., 2023), and GIS-based sanitation data management (Musliyana et al., 2024). Nevertheless, a methodological gap remains in integrating transparent sanitation-risk indices, GIS-based spatial prioritization, and machine-learning analysis within a single framework. Existing studies have generally focused on conventional risk assessment or descriptive GIS mapping, while fewer studies have examined whether contextual demographic, built-environment, and health indicators provide additional multivariable information related to spatial sanitation-risk patterns. Integrating GIS with exploratory machine-learning analysis may therefore extend conventional spatial assessment while allowing the predictive potential and limitations of the available data to be evaluated explicitly (Ahn et al., 2025).

Based on this problem, the present study develops a sanitation-risk assessment framework by integrating a transparent composite sanitation-risk index, GIS-based spatial analysis, and exploratory Random Forest regression across the 18 subdistricts of Palembang City. The study incorporates variables representing sanitation access, water access, wastewater management, illegal waste disposal, population and settlement characteristics, and environmental health indicators. The composite index is used to identify relative sanitation priorities, while Random Forest is used to explore whether contextual demographic, built-environment, and health indicators contain predictive information related to the spatial sanitation-risk index.

The importance of this research lies in integrating transparent index construction, spatial prioritization, and an explicit evaluation of machine-learning predictive performance within a single analytical framework. Rather than assuming that the machine-learning component provides an operational prediction model, its predictive performance is evaluated using cross-validation and interpreted according to the limitations of the available ecological dataset. The resulting framework is intended to support identification of priority areas for field verification and sanitation planning, while also providing an empirical basis for determining the data and validation requirements needed for stronger predictive modeling in future studies.

METHODS

This study used an ecological, cross-sectional spatial assessment design with the 18 subdistricts of Palembang City as the units of analysis. The analytical workflow consisted of: data compilation and harmonization; conversion of disease counts to rates; classification of waste-disposal locations; construction of a composite sanitation-risk index; assignment of relative risk categories; linkage of the index to subdistrict polygons in GIS; and exploratory Random Forest regression with Leave-One-Out Cross-Validation (LOOCV).

Eight study variables were included in the analysis. Four variables represented direct sanitation or environmental conditions and were used to construct the composite index: improved latrine access, improved water access, wastewater-management access (SPAL), and illegal waste-disposal points. Four variables were treated as contextual/exploratory variables: population density, building density, diarrhea rate, and dengue rate. Health outcomes were therefore not interpreted as causes of sanitation risk. Building density was derived from GHSL GHS-BUILT-S R2023A, epoch 2024; SPAL used the official Palembang wastewater-management dataset supplied for the study, with the city-level reference documented in the Palembang LPPD 2023; diarrhea and dengue data were obtained from health-service administrative records; illegal waste-disposal locations were compiled from the environmental-service database; and administrative boundaries were based on the 18-subdistrict polygon layer used for the GIS analysis. Other service and demographic indicators were taken from the latest district-level administrative records available to the research team during the 2025 data compilation.

For the illegal waste-disposal indicator, the waste-location database contained 483 recorded locations with heterogeneous labels. A location was counted as an illegal waste-disposal point when its type/status explicitly indicated TSL, TPSL, TPS Liar, Timbulan Liar, Liar, SL, or an equivalent label explicitly describing an illegal waste location. This rule identified 179 illegal waste-disposal points across the 18 subdistricts. Counts, rather than percentages, were used in the analysis.

The composite sanitation-risk index was constructed following the principles of transparent indicator direction, normalization, and aggregation recommended for composite indicators (OECD et al., 2008)(Sakti et al., 2024). For adverse indicators, normalized risk was calculated as $(X - X_{\min}) / (X_{\max} - X_{\min})$. For protective indicators, normalized risk was calculated as $(X_{\max} - X) / (X_{\max} - X_{\min})$. Thus, a larger normalized value consistently represented greater relative risk. Improved latrine access, water access, and SPAL access were treated as protective indicators; illegal waste-disposal points were treated as an adverse indicator.

Equal weighting was used to avoid unsupported subjective weighting. The Composite Sanitation Risk Index (CSRI) was calculated as $CSRI = 100 \times (R_{latrine} + R_{water} + R_{SPAL} + R_{waste}) / 4$.

For risk classification, the CSRI was interpreted as a relative within-city prioritization index rather than an absolute public-health threshold. The 18 scores were divided into tertiles to preserve a transparent rank-based comparison across the small set of subdistricts. The resulting cut-offs were Low ≤ 21.67 ; Medium > 21.67 to ≤ 37.68 ; and High > 37.68 , producing six subdistricts in each category.

For the GIS analysis, the CSRI score and category were linked to the administrative polygon of each subdistrict using the harmonized subdistrict name as the join key. A three-class choropleth was then produced to display low, medium, and high relative sanitation priority. The map was used as a spatial prioritization tool; no causal spatial inference was made. The GIS workflow therefore consisted of attribute harmonization, table-to-polygon join, category assignment, choropleth visualization, and comparison of geographically adjacent priority areas.

For the machine-learning component, Random Forest was applied as an exploratory regression analysis because the target was the continuous CSRI and the sample comprised only 18 ecological units. The predictors were population density, building density, diarrhea rate per 100,000 population, and dengue rate per 100,000 population. The analysis was implemented in Python 3.13 using scikit-learn 1.8.0. The final model used 500 trees, bootstrap sampling, $\text{max_depth}=2$, $\text{min_samples_leaf}=2$, $\text{max_features}=2$, and $\text{random_state}=42$. These restrictions were selected a priori to constrain tree complexity in the small dataset. Model performance was evaluated using LOOCV, in which each subdistrict was held out once, and was summarized using mean absolute error (MAE), root mean squared error (RMSE), and R^2 . A training-mean prediction was used as a simple benchmark. Out-of-bag R^2 was also reported as a secondary diagnostic. Because the data are spatially structured and small, the model was not treated as externally validated; structured cross-validation principles caution that predictive error can be underestimated when dependence is ignored (Roberts et al., 2017).

Finally, impurity-based Random Forest feature importance was calculated from the final full-data fit only as a descriptive model diagnostic. Feature importance was not interpreted as causality, effect direction, or proof of a determinant. In particular, importance rankings from a model with limited out-of-sample predictive performance should be interpreted cautiously and cannot substitute for out-of-sample performance. The definitions, units, analytical roles, data sources, and measurement periods of all study variables are summarized in Table 1.

Table 1.
Definition, unit, analytical role, source, and measurement period of study variables.

Variable	Definition	Unit	Role	Source	Period/reference
Improved latrine access	Households with improved/eligible latrine access	%	CSRI component	District administrative sanitation record	Latest record compiled in 2025
Improved water access	Population/households covered by improved/piped water data	%	CSRI component	District administrative water record	Latest record compiled in 2025
SPAL access	Wastewater-management access	%	CSRI component	Official Palembang wastewater-management dataset; city reference in LPPD 2023	2024 reference / district dataset used in study
Illegal waste-disposal points	Locations classified as illegal waste disposal after cleaning	points	CSRI component	Environmental-service waste-location database	Dataset compiled for 2025 analysis
Population density	Population per unit area	persons/km ²	Contextual RF predictor	District administrative demographic record	Latest record compiled in 2025
Building density	Built-up surface as percentage of subdistrict area	%	Contextual RF predictor	GHSL GHS-BUILT-S R2023A	Epoch 2024
Diarrhea	Reported diarrhea cases standardized by subdistrict population	cases/100,000	Contextual health indicator / RF predictor	Health administrative record	Dataset compiled for 2025 analysis
Dengue	Reported dengue cases standardized by subdistrict population	cases/100,000	Contextual health indicator / RF predictor	Health administrative record	Dataset compiled for 2025 analysis

RESULTS AND DISCUSSION

Descriptive characteristics

The dataset showed substantial heterogeneity across the 18 subdistricts. Improved water access was generally high (mean 96.59%, SD 2.47), whereas latrine access showed greater variation (mean 88.17%, SD 14.14; range 50.41-100.00%). SPAL access ranged from 76% to 94% (mean 86.94%). Building density ranged from

4.71% to 41.04%, and population density from 1,096.98 to 11,074.52 persons/km². The cleaned waste database identified 179 illegal waste-disposal locations, with 2-27 points per subdistrict. Diarrhea and dengue were standardized as rates because raw counts are strongly affected by population size. The descriptive statistics of all study variables are presented in Table 2.

Table 2.
Descriptive statistics of the study variables

Variable	Unit	n	Mean	SD	Min-Max
Improved water access	%	18	96.59	2.47	90.91-99.44
Building density	%	18	23.47	11.84	4.71-41.04
Population density	persons/km ²	18	6,259.80	3,314.55	1,096.98-11,074.52

Variable	Unit	n	Mean	SD	Min-Max
Diarrhea rate	per 100,000	18	4,039.15	222.06	3,641.20-4,796.21
Dengue rate	per 100,000	18	76.02	21.67	46.24-116.84
Improved latrine access	%	18	88.17	14.14	50.41-100.00
SPAL access	%	18	86.94	5.17	76.00-94.00
Illegal waste-disposal points	points	18	9.94	7.27	2-27

Spatial Distribution of Sanitation Risk

The CSRI ranged from 13.09 to 73.83. Kertapati had the highest relative risk score (73.83), followed by Seberang Ulu II (54.08), Seberang Ulu I (47.85), Gandus (47.13), Alang-Alang Lebar (44.12), and Jakabaring (41.77). The ranking and magnitude of the composite sanitation-risk scores across all 18 subdistricts are shown in Figure 2. These six subdistricts formed the high-priority tertile. The

medium group comprised Kalidoni, Sematang Borang, Ilir Barat II, Sukarami, Plaju, and Ilir Timur I; the low group comprised Ilir Barat I, Ilir Timur III, Kemuning, Sako, Bukit Kecil, and Ilir Timur II. The ranking reflects the composite index constructed from direct sanitation and environmental components using the explicit normalization and weighting procedure described above. The spatial distribution of these sanitation-risk categories across the 18 subdistricts is illustrated in Figure 1.

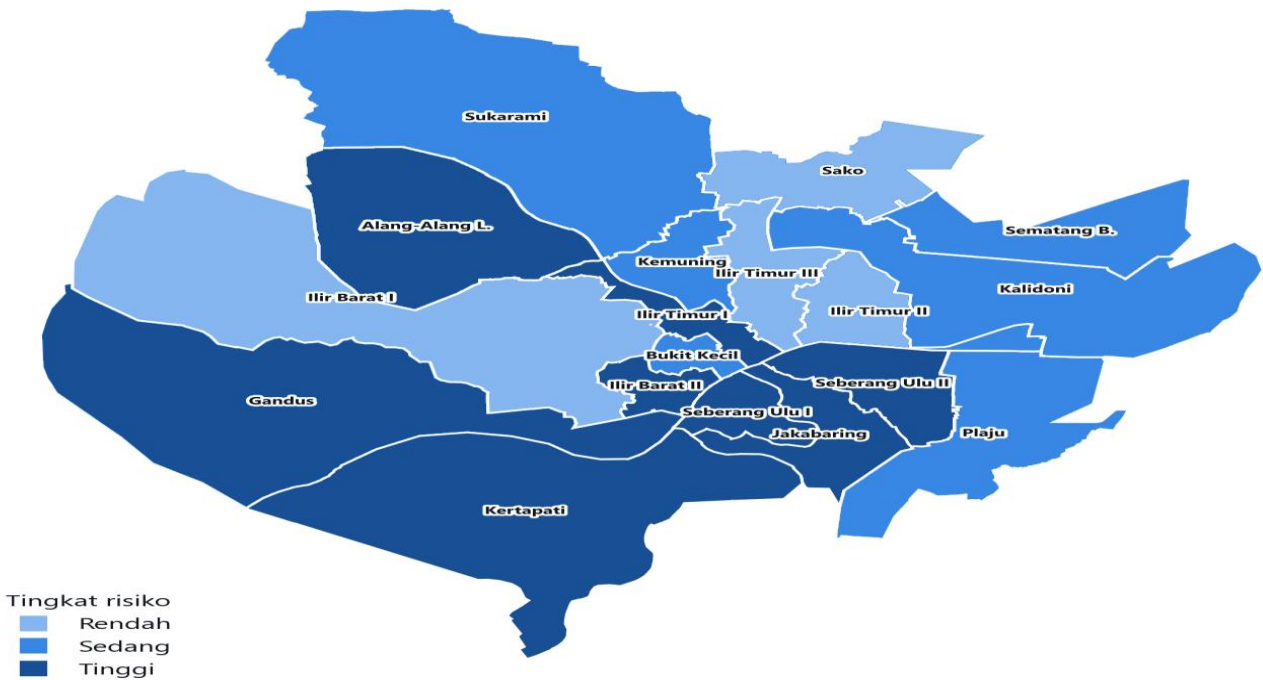


Figure 1. Spatial distribution of sanitation-risk levels across 18 subdistricts of Palembang City.

The map should be interpreted as a prioritization layer rather than a direct measurement of household-level risk. The distribution also demonstrates why sanitation planning should be spatially differentiated. Several high-priority areas are adjacent in the southern part of the city, especially Kertapati, Seberang Ulu I, Seberang Ulu II, and Jakabaring. GIS is useful in this

context because it makes service and environmental inequality visible and supports location-specific planning (Pradhan et al., 2025). The spatial clustering of several high-priority subdistricts in the southern part of Palembang also has planning implications beyond individual administrative boundaries. Sanitation problems can be connected

through drainage networks, waterways, population movement, and shared waste-service systems; therefore, neighboring high-priority areas may benefit from coordinated rather than isolated interventions (Maliga & Darmin, 2020). A systematic review by (Pradhan et al., 2025) found that GIS has been widely used in WASH research for mapping facilities, examining spatial patterns, linking WASH conditions with health outcomes, and supporting planning, while also identifying data availability and data quality as recurring challenges. (Rahardjo et al., 2025) likewise showed that GIS-based

mapping can identify spatial pollution patterns and support community sanitation planning. For Palembang, this means that the risk map is most useful when combined with field verification and service-network information, particularly in areas where wastewater, drainage, and solid-waste problems may cross subdistrict borders. The map should therefore guide where investigation and resources are directed first, rather than being treated as a substitute for site-level environmental assessment (Cuadros et al., 2024).

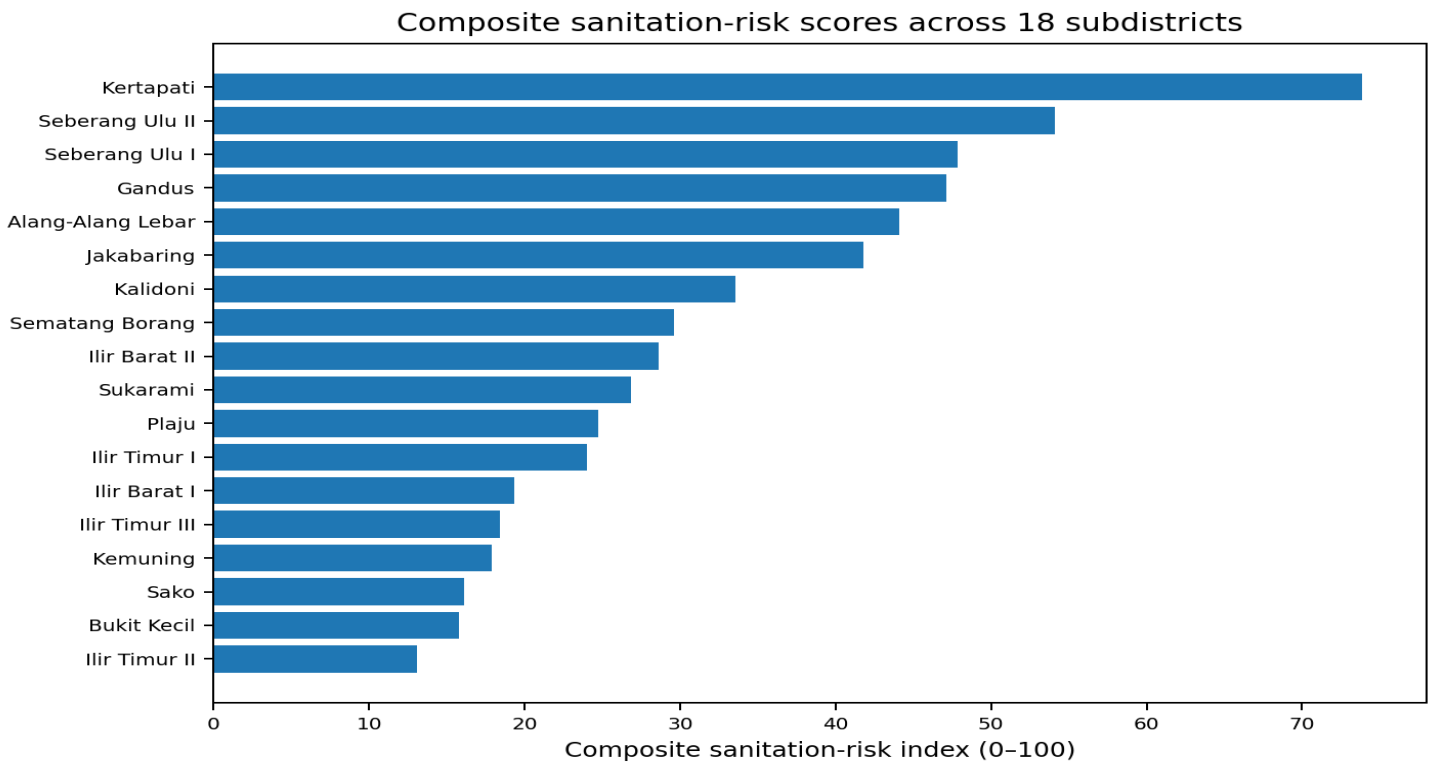


Figure 2. Composite sanitation-risk scores for the 18 subdistricts.

Table 3.

Risk score, category, and principal index contributors for all 18 subdistricts.

Rank	Subdistrict	CSRI	Category	Main index contributors
1	Kertapati	73.83	High	Low latrine access; low water access
2	Seberang Ulu II	54.08	High	Low latrine access; low water access
3	Seberang Ulu I	47.85	High	Low water access; low latrine access
4	Gandus	47.13	High	Low SPAL access; low water access
5	Alang-Alang Lebar	44.12	High	Illegal waste-disposal points; low SPAL access
6	Jakabaring	41.77	High	Low water access; illegal waste-disposal points

Rank	Subdistrict	CSRI	Category	Main index contributors
7	Kalidoni	33.59	Medium	Low SPAL access; low water access
8	Sematang Borang	29.61	Medium	Low SPAL access; illegal waste-disposal points
9	Iilir Barat II	28.62	Medium	Low SPAL access; low water access
10	Sukarami	26.86	Medium	Illegal waste-disposal points; low SPAL access
11	Plaju	24.77	Medium	Low SPAL access; low water access
12	Iilir Timur I	24.03	Medium	Illegal waste-disposal points; low water access
13	Iilir Barat I	19.32	Low	Illegal waste-disposal points; low SPAL access
14	Iilir Timur III	18.44	Low	Low SPAL access; illegal waste-disposal points
15	Kemuning	17.86	Low	Illegal waste-disposal points; low SPAL access
16	Sako	16.11	Low	Low SPAL access; low latrine access
17	Bukit Kecil	15.81	Low	Low latrine access; low SPAL access
18	Iilir Timur II	13.09	Low	Low SPAL access; low latrine access

As shown in Table 3, the area-specific implications of the CSRI can be derived from the principal index components contributing to the risk profile of each priority subdistrict. In Kertapati and Seberang Ulu II, the high-priority classification was primarily associated with relatively low improved-latrine and water-access indicators, suggesting that field verification and interventions should prioritize these basic sanitation services. Gandus showed a different profile, characterized mainly by relatively low SPAL and water access, indicating a greater need to assess domestic wastewater management and water-service conditions. In Alang-Alang Lebar, the principal contributors were illegal waste-disposal points and relatively low SPAL access, suggesting that waste-control measures and improvements in domestic wastewater management should receive particular attention. These differences demonstrate that subdistricts within the same high-risk category should not automatically receive identical interventions; rather, the CSRI should be used as an initial spatial screening tool, followed by field verification of the specific sanitation components contributing to the risk profile before intervention priorities are determined.

Exploratory Random Forest validation

The exploratory Random Forest analysis was evaluated using LOOCV, yielding MAE=14.01, RMSE=17.05, and $R^2=-0.15$, while the mean-prediction benchmark yielded MAE=13.89, RMSE=16.81, and $R^2=-0.12$. The out-of-bag R^2 was -0.097. The Random Forest validation results and comparison with the mean-prediction baseline are summarized in Table 4. These validation results indicate that predictive generalizability remains limited in the present 18-subdistrict dataset. Nevertheless, the analysis provides an initial evaluation of how contextual demographic, built-environment, and health indicators relate to the spatial sanitation-risk index within the integrated GIS–AI framework. Given the limited number of ecological units, the Random Forest findings are interpreted as exploratory rather than causal or operationally predictive, and further validation using larger, independent datasets is warranted.

The limited validation performance should be interpreted in relation to sample size and outcome construction rather than as evidence that machine learning has no role in sanitation assessment. Large datasets provide substantially more information for tree-based algorithms to learn stable patterns. (Yitageasu et

al., 2025), for example, applied machine learning to more than 500,000 household observations from Demographic and Health Surveys across Sub-Saharan Africa and reported useful performance for predicting household sanitation-facility access. In environmental epidemiology, (Iyer et al., 2025) (Saumi Anggit Musofi, 2024), similarly noted the expanding role of GeoAI for integrating geospatial and environmental information, while emphasizing methodological issues related to data quality, validation, spatial scale, and interpretability. The

contrast with the present study is important: the Palembang analysis contains only 18 ecological observations and therefore cannot provide the same learning conditions as household-level or large multi-site datasets. The principal value of Random Forest here is consequently exploratory—testing whether contextual variables contain a detectable multivariable signal—while the GIS-based index remains the more directly interpretable basis for current spatial prioritization.

Table 4.
Random Forest validation results.

Metric	Random Forest LOOCV	Mean baseline	Interpretation
MAE	14.01	13.89	RF did not improve absolute error
RMSE	17.05	16.81	RF did not improve squared-error performance
R ²	-0.15	-0.12	Negative out-of-sample R ² ; no useful predictive performance
OOB R ²	-0.097	—	Secondary diagnostic also indicates weak generalization

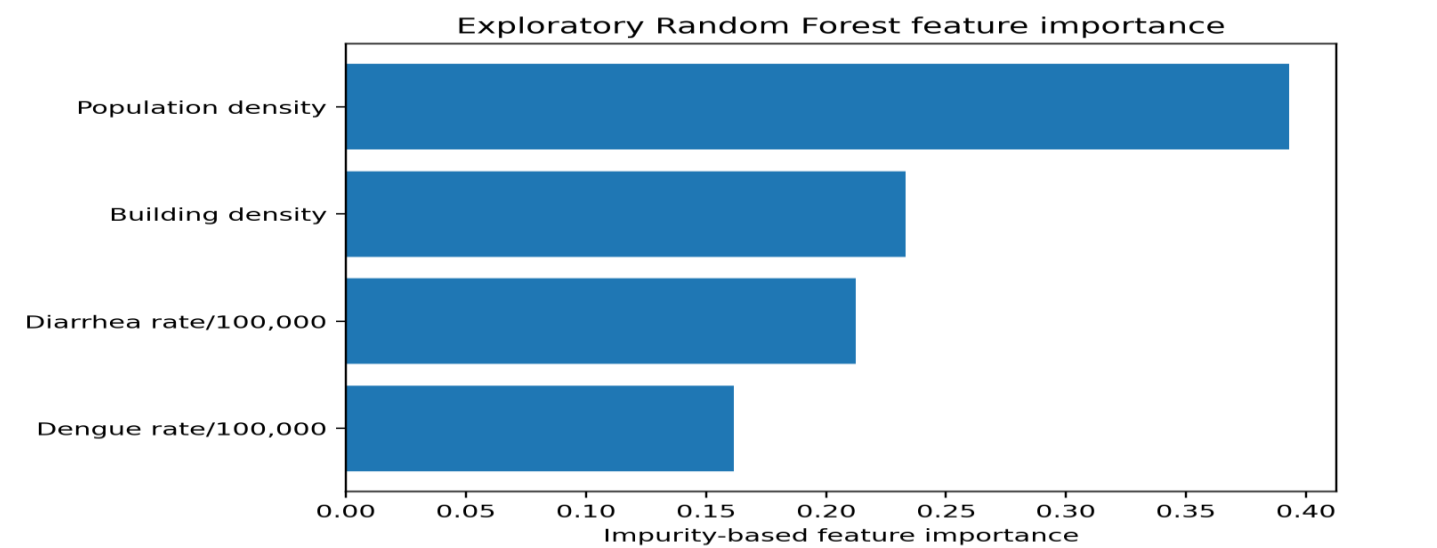


Figure 3. Exploratory impurity-based feature importance from the full-data Random Forest fit.

As shown in Figure 3, population density had the largest impurity-based importance in the full-data Random Forest fit (0.393), followed by building density (0.233), diarrhea rate (0.212), and dengue rate (0.161). Because out-of-sample predictive performance was limited, these values should not be interpreted as reliable predictive rankings, causal determinants, or effect sizes. They only describe how the fitted trees partitioned this small dataset. This distinction is important because feature importance can be misleading when model predictive performance is limited, and it does not provide the direction or causal meaning of a relationship. Recent methodological work on Random Forest interpretation similarly emphasizes caution when assigning substantive

meaning to variable-importance measures, particularly when predictors are correlated or statistically dependent (Sudasman & Gereuw, 2025 ; Maran Ata Albertus & Mauguru Mesakh Edwin, 2025 ; Rothacher & Strobl, 2024).

The prominence of population and building density in the fitted trees is nevertheless epidemiologically and environmentally plausible as a hypothesis for future investigation. Dense urban settlements can concentrate demand for toilets, wastewater conveyance, drainage, water supply, and solid-waste services within limited space. However, density should not automatically be interpreted as a sanitation hazard; its effect depends on infrastructure

capacity, service coverage, settlement form, and environmental context. (Yi et al., 2024 ; Fernández-Ortega et al., 2026) demonstrated the value of interpretable spatial machine learning for examining urban sanitation problems and showed that relationships between urban-form variables and sanitation outcomes may vary spatially. Likewise, (van der Heijden et al., 2025) found that the relative importance of environmental and socio-demographic factors for diarrheal disease differed according to the spatial scale of analysis. These studies support a cautious interpretation of the present feature-importance results: population and building density are useful signals for hypothesis generation and spatial targeting, but their substantive role in Palembang requires confirmation using finer spatial units and independent outcome data (Muhammad Zorga et al., 2025).

The contribution of this study is primarily methodological and spatial, while the AI component provides an exploratory assessment of predictive potential. First, the sanitation-risk index is fully reproducible, with explicit indicator direction, normalization, weighting, and cut-offs. Second, GIS translates the index into a spatial prioritization layer covering all 18 subdistricts. Third, the exploratory Random Forest analysis is evaluated rather than assumed to be predictive; its limited LOOCV performance is reported transparently. This contrasts with descriptive GIS-only studies by adding a formal multivariable diagnostic while also avoiding unsupported claims of prediction or causality. Large-scale sanitation machine-learning studies can achieve useful predictive performance when hundreds of thousands of observations and independent outcomes are available (Yitageasu et al., 2025 ; Debriela Salvarina Rosadianty et al., 2024), whereas the present $n=18$ ecological dataset provides more limited conditions for predictive generalization.

From a policy perspective, the combined GIS and composite-index approach is useful because sanitation investment decisions rarely depend on a single indicator. (Ali et al., 2024) (A et al., 2024), in a sanitation-risk assessment in Central Jakarta, similarly organized sanitation risk around multiple dimensions and emphasized the value of risk assessment for identifying priority locations and supporting infrastructure-budget decisions. GIS-based sanitation information systems in Indonesian settings have also been developed to improve the organization and visualization of sanitation data for local government use (Musliyana et al., 2024; Widiyatmoko et al., 2023). In Palembang, the practical implication is that the CSRI should function as a screening and prioritization layer: high-priority subdistricts can be investigated first, after which field verification can determine the specific technical intervention required. This sequencing is important because an area may obtain a high composite score through different combinations of

deficits, so identical interventions across all high-priority areas would not necessarily be efficient.

This study has several limitations. The analysis uses 18 subdistricts, which limits statistical power, model stability, and generalizability. The study is ecological and cross-sectional; therefore, associations at the subdistrict level cannot be interpreted as household-level or causal relationships. The CSRI is a relative prioritization index rather than an independently validated health-risk outcome. Equal weighting was selected for transparency and should be tested through sensitivity analysis in future studies. Tertile categories are relative to the 18 Palembang subdistricts and are not universal risk thresholds. Disease indicators may reflect healthcare access, reporting practices, seasonality, climate, mobility, and other factors beyond sanitation. The Random Forest model showed a negative out-of-sample R^2 , indicating limited predictive generalizability in the present dataset; external validation with larger and independent datasets is required before operational predictive use. Future work should use finer spatial units, multi-year data, independent field/EHRA/STBM labels, spatially appropriate validation, and external validation before any predictive decision-support system is proposed.

CONCLUSIONS

This study provides a transparent GIS-based assessment of relative residential sanitation risk across all 18 subdistricts of Palembang. The composite index, based on improved latrine access, water access, SPAL access, and illegal waste-disposal points, identified Kertapati, Seberang Ulu II, Seberang Ulu I, Gandus, Alang-Alang Lebar, and Jakabaring as the six highest-priority subdistricts for field verification and targeted sanitation planning. The GIS component demonstrates substantial spatial heterogeneity and supports differentiated area-based intervention. Random Forest analysis showed limited out-of-sample predictive performance in the present dataset and is therefore interpreted as exploratory; no causal relationship is claimed from feature importance. The principal value of the framework is reproducible spatial prioritization. Stronger predictive modeling will require independent sanitation-risk outcomes, larger and finer-scale samples, temporal data, and external validation.

SUGGESTION

Based on the findings of this study, sanitation interventions in Palembang City should prioritize areas identified as having higher sanitation risk, particularly those with unfavorable demographic, sanitation infrastructure, and environmental health conditions. Local governments are encouraged to use the sanitation risk map as supporting information for prioritizing sanitation programs, improving access to improved sanitation facilities, strengthening wastewater management, and addressing environmental health problems in high-risk areas. The integration of Geographic Information

Systems and Random Forest can also be considered as a supporting approach for periodic sanitation monitoring and evidence-based planning. Further research should expand the analysis using more detailed spatial data and additional environmental health indicators, as well as validate the prediction model with field observations and updated datasets, to improve the accuracy and applicability of sanitation

REFERENCES

- A, S., Mazroa, A. Al, Maashi, M., T.M, N., & V, P. (2024). Remote sensing and machine learning approach for zoning of wastewater drainage system. *Desalination and Water Treatment*, 319. [[Crossref](#)] , [[Publisher](#)]
- Ahn, K., Suh, Y., Han, S., Yang, J., Taubenböck, H., & Cha, M. (2025). Mapping Reduced Accessibility to WASH Facilities in Rohingya Refugee Camps With Sub-Meter Imagery. *Journal of Humanity Group*. [[Crossref](#)], [[Publisher](#)]
- Ali, F., Lestari, D. L., Putri, M. D., & Azmi, K. N. (2024). Mapping Sanitation Risk at the Hamlet Level: A Comprehensive Assessment of Sub-District in Central Jakarta. *International Journal of Technology*, 15(6), 1936–1945. [[Crossref](#)], [[Publisher](#)]
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. [[Crossref](#)], [[Publisher](#)]
- Cuadros, D. F., Chen, X., Li, J., Omori, R., & Musuka, G. (2024). Advancing Public Health Surveillance: Integrating Modeling and GIS in the Wastewater-Based Epidemiology of Viruses, a Narrative Review. In *Pathogens* (Vol. 13, Number 8). Multidisciplinary Digital Publishing Institute (MDPI). [[Crossref](#)], [[Publisher](#)]
- Debriela Salvarina Rosadianty, Dessy Lailatul Maghfiroh, Fadila Tullazulfa, Ajeng Ramadanti N, Putri, & Ibnu Rois. (2024). Pemetaan Rumah Sehat berdasarkan Faktor Kesehatan Lingkungan di Dukuh Sumuran, Palbapang, Bantul. *Jurnal Abdimas Kesehatan (JAK)*. [[Crossref](#)], [[Publisher](#)]
- Fernández-Ortega, J., Barberá, J. A., & Andreo, B. (2026). New insights into machine learning prediction techniques for real-time sanitary risk assessment in karst drinking water sources affected by faecal contamination. *Water Research*, 290. [[Crossref](#)], [[Publisher](#)]
- Hardiknas Wopari, A. E., Rumbiak, R., & Sudiro. (2022). Utilization of SIG Analysis for Groundwater Potential Mapping In Keerom District. *Jurnal Wilayah, Kota Dan Lingkungan Berkelanjutan (JWIKAL)*, 1(2). [[Crossref](#)], [[Publisher](#)]
- Iryanto, A. A., Firmansyah, Y. W., Widyantoro, W., & Zolanda, A. (2022). Spatial Patterns Of Environmental Sanitation Factors As Determinants Of Toddlers' Diarrhea In Pauh District, Padang City In 2021. *Jurnal Kesehatan Lingkungan*, 14(2), 71–81. [[Crossref](#)], [[Publisher](#)]
- Iyer, H. S., Karasaki, S., Yi, L., Hswen, Y., James, P., & VoPham, T. (2025). Harnessing Geospatial Artificial Intelligence (GeoAI) for Environmental Epidemiology: A Narrative Review. In *Current Environmental Health Reports* (Vol. 12, Number 1). Springer Science and Business Media Deutschland GmbH. [[Crossref](#)], [[Publisher](#)]
- Kasaluhe, M. D., Gansalangi, F., Sambeka, Y., & Lalombo, A. S. (2022). Sistem Informasi Geografi Untuk Pemetaan Cakupan Desa SBS (Stop Buang Air Besar Sembarangan) di Kabupaten Kepulauan Sangihe. *INSOLOGI: Jurnal Sains Dan Teknologi*, 1(1), 28–33. [[Crossref](#)], [[Publisher](#)]
- Maliga, I., & Darmin. (2020). Analisis Penilaian Risiko Kesehatan Lingkungan dengan Menggunakan Pendekatan Environmental Health Risk Assessment (EHRA) di Kecamatan Moyo Utara. *MITL Media Ilmiah Teknik Lingkungan*, 5, 16–26. [[Crossref](#)], [[Publisher](#)]
- Maran Ata Albertus, & Mauguru Mesakh Edwin. (2025). Penilaian Risiko Kesehatan Lingkungan Sanitasi Pada Daerah Terdampak Badai Seroja di Kabupaten Timor Tengah Selatan. *IJOH: Indonesian Journal of Public Health*. [[Publisher](#)]
- Muhammad Zorga, R., Susetyo, B., Eosina Hosen, P., & Sholeh Iskandar. (2025). Analisis Kinerja Sistem Sanitasi Berbasis Webgis Di Kota Bogor. *Jurnal Warunayama*. [[Crossref](#)], [[Publisher](#)]
- Musliyana, Z., Helinda, A., Wardana, P., & Khadafi, M. (2024). Sistem Informasi Pendataan Sanitasi Warga Menggunakan Sistem Informasi Geografis (Sig) Berbasis Website Pada Dinas Pupr Kota Banda Aceh Sanitation Data Collection Information System For Citizens Using A Website-Based Geographic Information System (Gis) At The Pupr Department Of Banda Aceh City. *Journal of Informatics and Computer Science*, 10(1). [[Crossref](#)], [[Publisher](#)]
- Nilandita, W., Diyanah, K. C., & Suprayogi, D. (2022). Pemetaan Kondisi Sanitasi Lingkungan Dasar serta Kejadian Penyakit pada Pondok Pesantren di Kota Surabaya. *Al-Ard: Jurnal Tehnik Lingkungan*. [[Crossref](#)], [[Publisher](#)]
- OECD, Union, E., & European Commission, J. R. C. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing. [[Crossref](#)], [[Publisher](#)]
- Pradhan, S., Pradhan, B., & Joshi, A. (2025). Leveraging geographic information systems (GIS) in water, sanitation, and hygiene (WASH) research: a systematic review of applications and challenges. In *Spatial Information Research* (Vol. 33, Number 2). Korean Spatial Information Society. [[Crossref](#)], [[Publisher](#)]
- Rahardjo, P. P., Widyastuti, D., Zairina, A., Mondiana, Y. Q., Wijaya, A. C., & Farida, S. (2025). Integrasi Sistem Informasi Geografis dalam Pemetaan Kualitas Air untuk Mendukung Program Sanitasi

Purnama R., Garmini R., Dekawaty A., Arla Z. V., Yuliawati A. (2026). Sanitation Risk Prediction Model Based on G.I.S and A.I to Improve the Residential Sanitation Index in Palembang. *Gema Lingkungan Kesehatan*, 24(4), 607–617. <https://doi.org/10.36568/gelinkes.v23i4.591>

- Masyarakat. *JURNAL GREEN HOUSE*, 4(1), 58–66. [[Crossref](#)], [[Publisher](#)]
- Rio Purnama. (2024). Pengaruh Penyuluhan Terhadap Pengetahuan Siswa Mengenai Perilaku Hidup Bersih Dan Sehat di Sekolah Dasar Negeri 43 Kota Palembang. *Masker Medika*, 12(2), 255–260. [[Crossref](#)], [[Publisher](#)]
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guilleri-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. [[Crossref](#)], [[Publisher](#)]
- Rothacher, Y., & Strobl, C. (2024). Identifying Informative Predictor Variables With Random Forests. *Journal of Educational and Behavioral Statistics*, 49 (4), 595–629. [[Crossref](#)], [[Publisher](#)]
- Sakti, A. D., Deliar, A., Hafidzah, D. R., Chintia, A. V., Anggraini, T. S., Ihsan, K. T. N., Virtriana, R., Suwardhi, D., Harto, A. B., Nurmaulia, S. L., Aritenang, A. F., Riqqi, A., Hernandi, A., Soeksmantono, B., & Wikantika, K. (2024). Machine learning based urban sprawl assessment using integrated multi-hazard and environmental-economic impact. *Scientific Reports*, 14(1). [[Crossref](#)], [[Publisher](#)]
- Saumi Anggit Musofi. (2024). Kajian Risiko Sanitasi Lingkungan Permukiman di Sempadan Sungai Code Kabupaten Sleman Daerah Istimewa Yogyakarta Saumi Anggit Musofi. *Jurnal Paradigma*. [[Crossref](#)], [[Publisher](#)]
- Sudasman, F. H., & Gereuw, G. E. (2025). Integration of AI and GIS in Clean Water Quality Monitoring in Urban and Rural Communities. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 4(2), 5338–5344. [[Crossref](#)], [[Publisher](#)]
- Trisnaini, I., Idris, H., & Purba, I. G. (2022). Kajian Sanitasi Lingkungan Pemukiman di Bantaran Sungai Musi Kota Palembang. *JURNAL KESEHATAN LINGKUNGAN INDONESIA*, 18(2), 67. [[Crossref](#)], [[Publisher](#)]
- van der Heijden, R., Doran, E. M. B., King, P., Brown, K. P., Rizzo, D. M., & Gleason, K. M. (2025). Using Spatial Technologies to Assess Risk Factors for Diarrheal Disease Under Environmental Variability in Bangladesh: A Machine Learning Study. *International Journal of Environmental Research and Public Health*, 22(11). [[Crossref](#)], [[Publisher](#)]
- Widiyatmoko, A. T., Nugroho, A., & Pasa, I. Y. (2023). Pengembangan Aplikasi Pemetaan Desa Rawan Sanitasi Berbasis Web Menggunakan Open StreatMap. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(2), 92–98. [[Crossref](#)], [[Publisher](#)]
- Yi, S., Li, X., Wang, R., Guo, Z., Dong, X., Liu, Y., & Xu, Q. (2024). Interpretable spatial machine learning insights into urban sanitation challenges: A case study of human feces distribution in San Francisco. *Sustainable Cities and Society*, 113. [[Crossref](#)], [[Publisher](#)]
- Yitageasu, G., Alemu, E. A., Worede, E. A., Tigabie, M., & Demoze, L. (2025). Machine learning-based prediction of household sanitation facility access in Sub-Saharan Africa: insights from DHS data (2012–2024). *Archives of Public Health*, 83(1). [[Crossref](#)], [[Publisher](#)]