Machine Learning-Based Prediction of Heavy Metal Risks in Urban Stormwater Sediments

Prédiction des Risques Liés Aux Métaux Lourds Dans les Sédiments des Eaux Pluviales Urbaines par Apprentissage Automatique

María Alejandra Pimiento[a], Jose Anta[a], Andrés Torres[b]

[a]Universidade da Coruña, Water and Environmental Engineering Group, Center for Technological Innovation in Construction and Civil Engineering (CITEEC), m.a.pimiento@udc.es, jose.anta@udc.es
[b] Pontificia Universidad Javeriana. andres.torres@javeriana.edu.co

RÉSUMÉ

Les sédiments des eaux pluviales urbaines accumulent souvent des métaux lourds (ML), ce qui représente des risques environnementaux importants. Cette étude propose une méthodologie fondée sur l'apprentissage automatique (AA) pour prédire les indices de risque du facteur d'enrichissement (FE) des ML dans un bassin urbain de Bogotá, en Colombie. Des échantillons de sédiments ont été prélevés sur deux ans et analysés pour la distribution granulométrique (DG) et les concentrations totales de Cd, Pb, Cu, Zn, Cr et Ni. Les données de précipitations ont été caractérisées par la hauteur, l'intensité et la période sèche antérieure (PSA). Les tests statistiques ont identifié le FE du Pb comme variable cible la plus adaptée, montrant des corrélations significatives avec les paramètres de la DG (D20–D70, $p < 0,05$). Un modèle à machine à vecteurs de support (SVM) a été élaboré à l'aide d'une approche de Monte-Carlo avec 1 000 itérations de calibration-validation. Le noyau ANOVA a offert les meilleures performances, obtenant un accord substantiel (kappa de Cohen = 0,71, $p = 0,037$) et prédisant correctement les niveaux de risque pour sept des huit échantillons de validation. Les résultats démontrent la faisabilité de prédire les indices de risque FE à partir de données pluviométriques facilement disponibles et d'informations sédimentaires limitées, réduisant les coûts de suivi et soutenant des stratégies proactives de gestion des sédiments.

ABSTRACT

Urban stormwater sediments often accumulate heavy metals (HMs), posing significant environmental risks. This study proposes a machine learning (ML)-based methodology to predict enrichment factor (EF) risk indices for HMs in an urban catchment in Bogotá, Colombia. Sediment samples were collected over two years and analyzed for particle size distribution (PSD) and total concentrations of Cd, Pb, Cu, Zn, Cr, and Ni. Rainfall data were characterized by precipitation height, intensity, and antecedent dry weather period (ADP). Statistical tests identified Pb_EF as the most suitable target variable, showing significant correlations with PSD parameters (D20–D70, $p < 0.05$). A Support Vector Machine (SVM) model was developed using a Monte Carlo approach with 1,000 calibration-validation iterations. The ANOVA kernel achieved the best performance, yielding substantial agreement (Cohen's Kappa = 0.71, $p = 0.037$) and correctly predicting risk levels for seven out of eight validation samples. Results demonstrate the feasibility of predicting EF risk indices using readily available rainfall data and limited sediment information, reducing monitoring costs and supporting proactive sediment management strategies.

KEYWORDS

Enrichment Factor, Heavy metals, Prediction, Sediments, Stormwater

1 INTRODUCTION

Stormwater is a valuable natural resource, yet in urban environments it is often degraded by pollution. Preventing its contamination is therefore essential (Rentachintala et al., 2022). Among the primary contributors to this pollution are sediments transported by runoff, which can accumulate toxic substances. Monitoring these sediments is critical to evaluate their toxicity and calculate pollution indices that indicate potential environmental risks. However, contamination levels vary significantly depending on local factors such as land use and hydrological conditions (Nawrot et al., 2021; Pimiento et al., 2023).

Heavy metal pollution in urban stormwater has been extensively studied due to its harmful effects and the urgent need to mitigate environmental impacts. While assessing pollution risk is an important step, predicting this risk is equally crucial for developing effective strategies to reduce negative outcomes (Ahmed et al., 2024; Fatmi et al., 2024; Pimiento et al., 2023). Machine Learning (ML) has proven highly effective for predicting environmental variables, offering accuracy and reliability in solving complex, non-linear problems and supporting better decision-making. When combined with multivariate analysis, ML provides deeper insights into pollution sources and patterns (Fatmi et al., 2024; Bi et al., 2024).

This study proposes a methodology to predict heavy metal risk levels in urban stormwater sediments using ML techniques. The approach focuses on identifying relationships between local conditions and environmental risk indices, enabling more proactive and informed management of urban water quality.

2 MATERIAL AND METHODS

2.1 Study Site

This study was conducted in an urban catchment in Bogotá, Colombia, which primarily receives stormwater runoff from residential áreas (*Fig. 1*). This land use implies that sediments originate from both natural materials and impermeable surfaces, reflecting typical urban conditions.

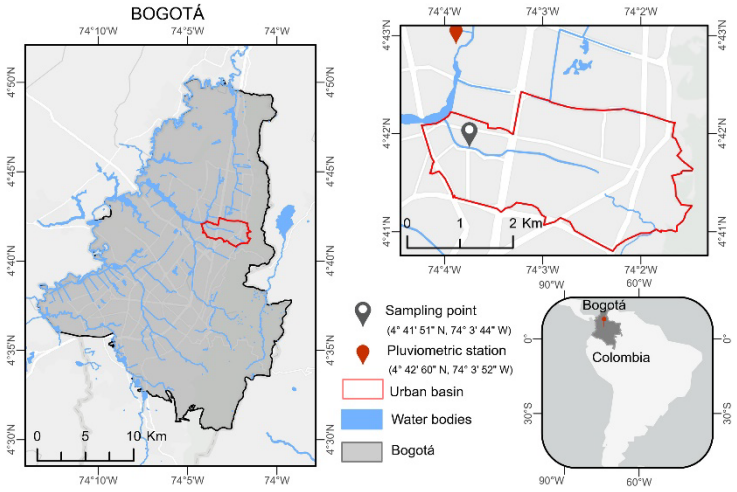

Fig. 1 Study site location

2.2 Sediment and hydrology characterization

Sediment characterization focused on particle size distribution (PSD) and total concentrations of the most common urban heavy metals (HM): Cadmium (Cd), Lead (Pb), Cooper (Cu), Zinc (Zn), Chromium (Cr), and Nikel (Ni). Samples were collected over a two-year period from the sampling point within the urban catchment to capture temporal variability. PSD was determined using laser diffraction Mastersizer 3000E Particle Size Analyzer, and heavy metal concentrations were measured following following EPA Method 3051A method, ensuring compliance with standard protocols. The enrichment factor (EF) was selected for this study because it is commonly used to determine whether contaminants adhered to sediments originate from natural or

anthropogenic sources (Gopal et al., 2023; Rumuri et al., 2023).

Hydrological characterization was based on rainfall data obtained from District Risk Management Institute of Bogotá. Precipitation records were analyzed to identify precipitation height (mm), rainfal intensity (mm/h) and antecedes dry weather period (ADP).

2.3 Modelling methodology

Fig. 2 presents the methodology applied for modelling EF risk indices, which relies on statistical tests to examine relationships among variables. To predict each risk index (Y variable), the corresponding explanatory variables (X variables) were first identified. Cohen's kappa coefficient was applied to determine associations between risk index categories and PSD and rainfall, both converted into categorical variables. Variable pairs with a p-value < 0.05 were retained for further analysis. Because risk index classification depends on defined thresholds, correlation tests were then used to numerically validate the strength of these relationships.

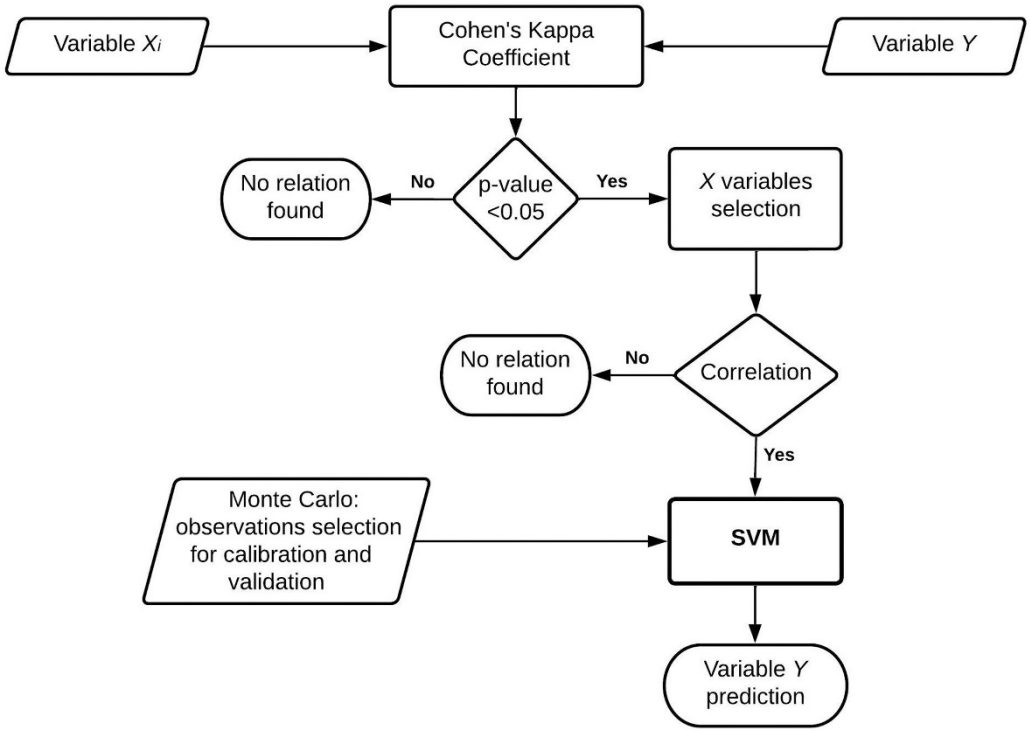

Fig. 2 Modeling methodology

A Monte Carlo approach was applied to enhance model robustness by generating 1,000 prediction models through random selection of calibration (two-thirds of observations) and validation (one-third) datasets. Model performance was assessed using Cohen's Kappa coefficient to compare predicted and observed categories in the validation set. All kernels available in the ksvm function were tested, and the most effective kernel was selected based on Kappa results, retaining models with p-values < 0.05 during calibration. Validation relied solely on rainfall variables, supplemented by PSD characteristic diameters from the first calibration observation to ensure replicability without additional sediment sampling, reducing costs and improving decision-making. The final model was chosen based on the highest agreement between predicted and observed classes and the lowest p-value (< 0.05).

3 RESULTS

Statistical analysis identified significant relationships only for Ni_EF and Pb_EF, with Pb_EF showing strong correlations with PSD parameters (D20–D70, $p < 0.05$), positioning it as the primary target for predictive modeling. Ni_EF displayed a positive but non-significant correlation with antecedent dry weather period (ADP).

During model development, the Monte Carlo approach generated 1,000 calibration-validation iterations, and the ANOVA kernel was selected as the optimal SVM classifier for Pb_EF prediction. The model achieved substantial

agreement (Cohen's Kappa = 0.71, p = 0.037) and correctly predicted risk levels for seven out of eight validation simples (*Fig. 3*). While prediction accuracy varied across enrichment classes, the model consistently identified significant enrichment, supporting its applicability for sediment management. Misclassification in one observation was linked to rainfall variability rather than PSD, highlighting the influence of hydrological conditions on EF prediction.

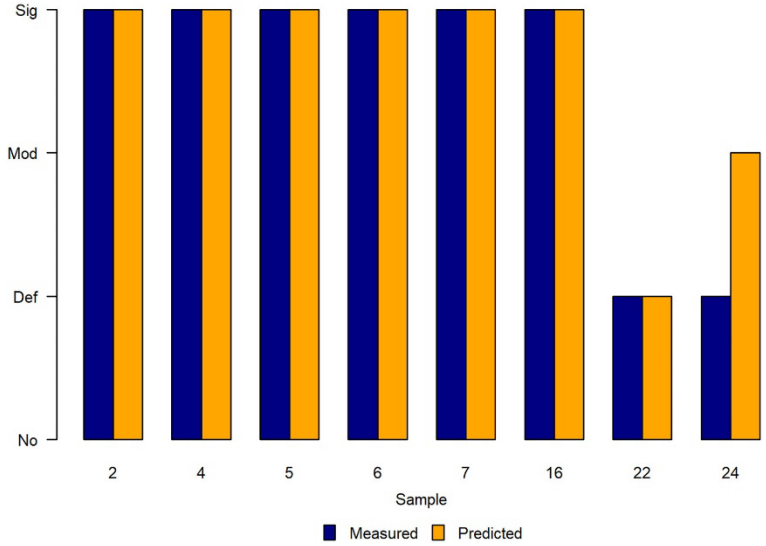

Fig. 3 Model results

4 CONCLUSIONS

The SVM model using the ANOVA kernel achieved substantial agreement (Kappa = 0.71), accurately predicting enrichment classes in most validation cases.

The proposed methodology demonstrates that EF risk indices can be predicted using rainfall data supplemented by minimal sediment information, reducing field sampling requirements and improving decision-making for urban stormwater management.

Future research should expand the dataset and explore additional ML algorithms to enhance prediction accuracy across all enrichment classes.

LIST OF REFERENCES *(only for scientific papers)*

Ahmed, B., Islam, S., Quraishi, S. B., Alam, M. N. E., Ahsan, M. S., & Kabir, A. (2024). *A probabilistic risk assessment of heavy metal in water and sediment: An industrially affected urban river in Bangladesh*. Water Environ. Res., 96(8), e11097.

Bi, Z., Sun, J., Xie, Y., Gu, Y., Zhang, H., Zheng, B., ... & Wei, N. (2024). *Machine learning-driven source identification and ecological risk prediction of heavy metal pollution in cultivated soils*. J. Hazard. Mater., 476, 135109.

Fatmi, B., Hazzab, A., Rahmani, A., & Ghenaim, A. (2024). *Examining temporal trends in heavy metal levels to analyze sediment pollution dynamics in the Saida urban watershed (N-W Algeria)*. Water Environ. Res., 96(8), e11084.

Gopal, V., Krishnamurthy, R. R., Vignesh, R., Nathan, C. S., Anshu, R., Kalaivanan, R., & Abioui, M. (2023). *Assessment of heavy metal contamination in the surface sediments of the Vedaranyam coast, Southern India*. Regional Studies Marine Sci., 65, 103081.

Nawrot, N., Wojciechowska, E., Mohsin, M., Kuittinen, S., Pappinen, A., & Rezania, S. (2021). *Trace metal contamination of bottom sediments: a review of assessment measures and geochemical background determination methods*. Minerals, 11(8), 872.

Pimiento, M. A., Duque, V., & Torres, A. (2023). *Urban stormwater sediment risk assessment from drainage structures in Bogotá, Colombia.* Environ. Sci.: Water Res. Technol., 9(12), 3269-3280.

Rentachintala, L. R. N. P., Reddy, M. M., & Mohapatra, P. K. (2022). *Urban stormwater management for sustainable and resilient measures and practices: a review*. Water Sci. Technol., 85(4), 1120-1140.

Rumuri, R., Ramkumar, T., Vasudevan, S., & Gnanachandrasamy, G. (2023). *Enrichment of heavy metals as function of salinity and pH of estuarine sediments, South East Coast of India*. Geology, Ecol. and Landscapes, 7(3), 212-220.