Spatially Robust Machine Learning for Groundwater Contaminant Prediction in the Paso del Norte Region

Prince Appiah[1], Mark A. Engle[2], Maria C. Mariani[1], Lin Ma[2], Naoko Lopez[2], Kwabena G. Duodu[1]

[1] *Department of Mathematical Sciences, The University of Texas at El Paso, Texas, USA*

[2] *Department of Earth, Environmental and Resource Sciences, The University of Texas at El Paso, Texas, USA*

Abstract

Groundwater contaminant prediction is challenging because concentrations are spatially heterogeneous, highly skewed, and unevenly sampled. This study evaluated a spatially robust machine-learning framework for predicting nitrate (NO3), arsenic (As), and uranium (U) concentrations in groundwater in the Paso del Norte region. The analysis included 962 nitrate, 383 arsenic, and 91 uranium observations. Concentrations were modeled using log1p transformation, and frozen five-fold spatial partitions were used to reduce optimistic performance estimates caused by geographic dependence. A spatial naive baseline was compared with Elastic Net, Random Forest, XGBoost, CatBoost, and Extra Trees across multiple predictor tiers. Preprocessing and hyperparameter tuning were performed within the training data to prevent leakage. The selected models were Elastic Net for nitrate and CatBoost for arsenic and uranium. Spatial out-of-fold MAEs were 5.21, 4.68, and 6.96, respectively, compared with baseline MAEs of 5.81, 5.24, and 7.54. Diagnostic analysis showed substantial geographic variability and systematic underprediction of the highest concentrations, particularly for arsenic and uranium. Overall, machine learning provided modest but reproducible improvements over naive spatial prediction while highlighting the importance of spatial validation and diagnostic analysis in groundwater-quality modeling.

Keywords: groundwater quality; nitrate; arsenic; uranium; spatial cross-validation; machine learning; Paso del Norte