Machine Learning for Predicting Groundwater Quality and Understanding Controls on Contaminants

Prince Appiah, Mark A. Engle, Maria C. Mariani, Lin Ma, Naoko Lopez

Abstract

Reliable prediction of trace constituent concentrations in groundwater requires methods that respect spatial dependence and leverage geochemical processes and landscape context. Despite their low concentration, trace constituents make up the majority of inorganic elements in drinking water standards and impact water quality for millions of people. However, data on trace elements in groundwater are sparse, making it difficult for stake holders and decisions makers to understand and evaluate regional water quality. Efforts to expand these data through mathematical prediction of concentrations of redox sensitive constituents in groundwater, such as arsenic (As), nitrate (NO_3^-), and uranium (U), have proved highly challenging, partially as a result of sparse data, censored results, and complex processes that are difficult to tease out in models. Here, we present a reproducible pipeline that (i) imputes predictors with MissForest outside crossvalidation while locking coordinates, (ii) performs strictly spatial validation via KMeansLeave-One-Group-Out with buffer sweeps (010 km), (iii) compares a broad model roster, and (iv) quantifies drivers using fold-wise permutation importance. Applied to regional groundwater in the Paso del Norte region, As achieved its best skill with tuned CatBoost on a Short+External feature set ($R^2 \approx 0.32$, improving over a no-externals baseline $R^2 \approx 0.22$). U performed best with a tuned multilayer perceptron at a 10 km buffer ($R^2 \approx 0.17$; baseline random forest $R^2 \approx 0.13$). NO_3^- remained challenging; a median-loss hist-gradient booster yielded modest skill ($R^2 \approx 0.006$). Dominant predictors for As included sulfate, fluoride, chloride, calcium, sodium, pH, elevation, and settlement proximity; for U, screen depth, volcanic/fault proximity, fluoride, alkalinity, sodium, and soil-U indices; for NO_3^-, farmland share, wastewater proximity, building density, and Mg/Cl proxies. Results highlight the necessity of spatial CV with buffers and the value of targeted external covariates. We outline next stepsadding pressure/transport indicators (septic, fertilizer, irrigation, wastewater networks) and uncertainty mappingto advance deployable groundwater-quality prediction.