EGU23-12842, updated on 08 Mar 2024
https://doi.org/10.5194/egusphere-egu23-12842
EGU General Assembly 2023

Incorporating spatial information for regionalization of hydrogeological parameters in machine learning models

Marc Ohmer, Fabienne Doll, and Tanja Liesch
KIT - Karlsruhe Institute of Technology, Institute for Applied Geosciences, Department of Hydrogeology, Karlsruhe, Germany (marc.ohmer@kit.edu)

Modeling spatially continuous variables from point measurements is integral to environmental scientific research in many fields. This is especially the case for groundwater, which is accessible only at boreholes and springs. Often, further management decisions are dependent on spatially continuous values of groundwater level or quality parameters.

For this task, deterministic or geostatistical interpolation methods are traditionally used, involving the spatial structure of the point locations as a set of XY-coordinates. In this case, the spatial model is usually created based solely on the geographic location of the measurement and the spatial autocorrelation of the target values. With few exceptions (e.g. co-kriging), classical interpolation techniques do not support the incorporation of covariates that are spatially correlated to improve spatial prediction accuracy.

Spatial predictions using machine learning (ML) models are an attractive and increasingly prevalent alternative, which use correlated, spatially continuous covariates (e.g. meteorological data, land-use, or geological maps) as predictors for groundwater level or quality parameters. They are trained on the nonlinear relationship between these predictors and the target values with the available point measurement data. However, spatial autocorrelations of the target values are usually not considered, as the points are treated independently of their location. Therefore, the machine learning models cannot exploit and represent the spatial dependence structures in the target data, without any further information about the geographic locations. The incorporation of locational information into ML models is, however, not trivial. From other fields, diverse approaches (e.g. using coordinates directly, distances to certain locations in space, Euclidean distance matrices, transformed coordinates) exist, and also different assessments of their suitability. For example, it is often stated that XY-coordinates are very well suited for the use with decision trees, but not for neural networks.

We systematically investigate the impact of the most commonly applied methods for the integration of spatial information, such as XY-coordinates, transformed coordinate-based input features, Euclidean distance matrix or distances to corners or center, Wendland transformed coordinates, and combinations of the aforementioned, on the interpolation results for selected hydrogeological parameters and different test sites in two ML models (Random Forest and Multi-Layer-Perceptron). We compare the results by cross-validation and with kriging reference models

as well as visual assessment for plausibility.

The results show that the incorporation of spatial covariates can significantly improve model performance, especially when the data have high spatial autocorrelation (and the data set is sufficient to capture this). In particular, the Euclidean distance matrix and Euclidean distances to defined locations proved to be efficient approaches to provide the spatial data structure for the model, while the application of XY-coordinates often resulted in significant artifacts in the resulting prediction surfaces.