Mathematical Geosciences
https://doi.org/10.1007/s11004-024-10163-4

Incorporating Spatial Information for Regionalization of Environmental Parameters in Machine Learning Models

Marc Ohmer[1] · Fabienne Doll[1] · Tanja Liesch[1]

Received: 29 September 2023 / Accepted: 30 September 2024

Abstract

Machine learning models have gained popularity for environmental variable predictions due to their capacity to capture complex relationships and automate learning. However, incorporating spatial information as covariates into these models remains a challenge, as they may struggle to recognize spatial structures or autocorrelation without explicit training. In this study, we address this challenge by integrating spatial information into a random forest model, enhancing nitrate concentration predictions in groundwater. Using a dataset from 1,550 well locations in Baden-Wuerttemberg, Germany, spanning 2016 through 2019, we consider various environmental covariates including climate data, topography, land cover, soil properties, and hydrology. To incorporate spatial information, we employ eight techniques leveraging spatial coordinates (geographic coordinates, polynomial geographic coordinates, oblique geographic coordinates) or distances (Wendland transformed coordinates, Euclidean distance fields, Euclidean distance matrix, principal component analysis, eigenvector spatial filtering). Results are compared with a baseline model and a univariate ordinary kriging benchmark, evaluated through leave-one-out cross validation, various error metrics, and Moran's I of residuals. Our findings highlight that integrating spatial information significantly enhances random forest model accuracy in predicting groundwater nitrate concentrations. Distance-based methods, like the Euclidean distance matrix, outperform coordinate-based approaches, albeit with higher computational requirements. Employing a dimension-reduced matrix strikes a balance between performance and accuracy. This study advances groundwater management and demonstrates the effectiveness of machine learning models in environmental studies.

Keywords Spatial machine learning · Random forest modeling · Spatial autocorrelation · Spatial covariates

✉ Marc Ohmer
marc.ohmer@kit.com

[1] Division of Hydrogeology, Institute of Applied Geosciences, Karlsruhe Institute of Technology , Adenauerring 20b, 76131 Karlsruhe, Germany

Published online: 18 December 2024

Abbreviations

10-fold CV	10-fold cross-validation
B	Bias
LUBW	Baden-Wuerttemberg State Institute for the Environment, Survey and Nature Conservation
ESF	Eigenvector spatial filtering
EDF	Euclidean distance fields
EDM	Euclidean distance matrix
GC	Geographic coordinates
GWR	Geographically weighted regression
GW	Groundwater
kNN	k-nearest neighbors
LISA	Local indicators of spatial association cluster map
LOO-CV	Leave-one-out cross validation
IDW	Inverse distance weighting
MAE	Mean absolute error
ML	Machine learning
OK	Ordinary kriging
OGC	Oblique geographic coordinates
PCA	Principal component analysis
PGC	Polynomial geographic coordinates
RBF	Radial basis functions
R^2	Coefficient of determination
RMSE	Root mean squared error
RF	Random forest
noinfo	Random forest model without spatial information
SA	Spatial autocorrelation
SVD	Singular value decomposition
WTC	Wendland transformed coordinates

1 Introduction

In numerous environmental disciplines, reliable and accurate spatial predictions of continuous data are crucial for informed decision-making. These predictions often depend on point measurements, making appropriate regionalization methods essential for estimating continuous data. For instance, in hydrogeology, measurements of groundwater (GW) level and GW quality parameters can be taken only at monitoring wells, but spatially continuous maps are imperative to derive important information, such as groundwater flow direction.

Regionalization challenges have traditionally been tackled using interpolation methods such as deterministic or geostatistical techniques such as kriging. These methods use the values of neighboring sample locations and the spatial structure of the data to estimate values at unsampled locations, assuming a level of spatial autocorrelation where close values are more similar than distant ones. This concept is often referred to as the first law of geography, as stated by Tobler (1970). However, common spatial interpolation methods like inverse distance weighting, spline interpolation, and kriging (except for co-kriging and universal kriging) have limited capacity to account for additional spatially correlated variables, as is the case for topography or land use, for example, which may affect GW level or GW quality (Ohmer et al. 2017).

In geographically weighted regression (GWR), covariates can be considered, allowing for the modeling of relationships between variables on a local level while accounting for spatial variation. GWR is widely used for assessing spatial variations in data relationships, particularly in spatial non-stationarity research like environmental studies (Brunsdon et al. 1996). Extensions such as multi-scale GWR (Fotheringham et al. 2017) enable modeling of spatial processes at different scales using a bandwidth vector, while GWR assumes a uniform bandwidth parameter. Further extensions include user-specified kernel functions based on spatial proximity (e.g., Liang et al. 2023), spatial weight kernels (e.g., Du et al. 2020), or the nonlinear relationship between spatial proximity and nonstationary weights (e.g., Wang et al. 2022). However, compared to machine learning (ML), traditional GWR may not be as flexible in capturing complex nonlinear relationships or handling large datasets. ML models such as neural networks can often capture a wider range of data structures and patterns, especially when relationships between variables are nontrivial or interactions between variables are complex. ML models automatically extract features and identify interactions, whereas in GWR, users usually need to specify features and their interactions, limiting its applicability when understanding of data and relationships is limited. In contrast to classical interpolation methods, ML models utilize spatially correlated input features (covariates), which exhibit spatial continuity and are consistently available across the entire study area. By learning complex relationships between these covariates and the target variable from the training data, these models can make predictions at unsampled locations. However, without explicitly incorporating spatial information, ML models may not fully account for the spatial autocorrelation of target values.

Most recent studies that apply ML models for the regionalization of hydrogeological or environmental parameters do not analyze or discuss the consideration of spatial autocorrelation. Some studies do not incorporate spatial information at all (e.g., Ransom et al. 2022; Knoll et al. 2020), while others use only one method, such as including XY-coordinates as predictors (e.g., Chowdhury et al. 2010; Kirkwood et al. 2022; Tsangaratos et al. 2014; Wadoux 2019; Walsh et al. 2017; Zanella et al. 2017), without examining other potential alternatives or evaluating the plausibility of the results.

Incorporating spatial dependencies in ML models is a subject of growing interest. Different approaches have been developed to capture spatial relationships more accurately than traditional geographic coordinates. These spatial covariates provide predictive utility by encompassing both environmental correlations and spatial dependence in the prediction process (Behrens et al. 2018).

The incorporation of spatial lag features is another common approach for capturing spatial relationships. These features aggregate target values from neighboring observations using methods such as k-nearest neighbors (kNN) or inverse distance weighting (IDW) (e.g., Credit 2022; Kiely and Bastian 2020; Leirvik and Yuan 2021; Liu et al. 2022; Sekulić et al. 2020) to combine the aspects of spatial interpolation and ML regression. Including spatial lag features from neighboring observations risks data leakage and bias in the model. Direct data leakage occurs when the target variable is used to derive features for the same target, leading to overfitting and inflated performance. Indirect data leakage happens when the target variable influences features for related variables, resulting in biased estimates. This can lead to overfitting and poor generalization. Hence, we excluded the spatial lag features approaches from our analysis.

After reviewing literature from various disciplines, we have identified several methods for incorporating spatial information into ML models:

1. Geographic coordinates (GC)
2. Transformed coordinate-based input features, including (i) higher polynomial geographic coordinates (PGC) (e.g., Borcard and Legendre 2002; Li et al. 2011) and (ii) oblique geographic coordinates (OGC) (e.g., Møller et al. 2020)
3. The use of radial basis function (RBF) methods such as Wendland transformed coordinates (WTC) (e.g., Nychka et al. 2015; Chen et al. 2022)
4. Euclidean distance metrics including (i) Euclidean distance fields (EDF) (e.g., Behrens et al. 2018; Hengl et al. 2018) and (ii) pairwise Euclidean distance matrix (EDM) (Ahn et al. 2020)
5. Dimensionally reduced representations of EDM including (i) principal component analysis (PCA) (e.g., Ahn et al. 2020) and (ii) eigenvector spatial filtering (ESF) (e.g., Borcard and Legendre 2002; Diniz-Filho and Bini 2005; Islam et al. 2022)

In this study, we methodically examined these different strategies for integrating spatial data into an ML model, utilizing a random forest (RF) regression model and a GW quality dataset focused on nitrate levels in Baden-Wuerttemberg, Germany, as published in Karimanzira et al. (2023). We compared the proposed approaches to a baseline RF model that lacks spatial information as well as a univariate ordinary kriging (OK) interpolation through a comparative analysis. Although OK does not incorporate secondary information, we selected it as a benchmark because it remains one of the most widely utilized interpolation techniques, thereby providing a straightforward and effective basis for comparison.

The evaluation process involved assessing selected cross-validation error measures. Additionally, we assessed the plausibility of the regionalization results, including the identification of artifacts. To our knowledge, this is the first systematic comparison of these approaches. While our study focused on hydrogeological data, we believe that our findings can apply to other environmental disciplines as long as the data exhibit spatial autocorrelation.

2 Theory and Background

2.1 Random Forest (RF)

The RF algorithm, developed by Breiman (2001), is a supervised learning method for classification or regression problems that builds an ensemble of decision trees based on a training dataset. The algorithm randomly selects a subset of features and observations for each tree, ensuring diversity in the decision trees. Each tree is constructed by repeatedly splitting the data based on the selected features until a stopping criterion is met. The final prediction from the RF is then calculated as the average of the predictions from all the individual decision trees. Mathematically, the RF model can be described as

$$\hat{\theta} B(\mathbf{x}) = \frac{1}{B} \sum b = 1^{B} t_b(\mathbf{x}; \mathbf{T}_b). \tag{1}$$

In this equation, $\hat{\theta} B(\mathbf{x})$ represents the predicted output for a given input vector $\mathbf{x}$. The prediction is obtained by averaging the predictions of B individual decision trees, where t_b^* represents the bth decision tree, and the average is taken over all the trees in the forest. Each decision tree is constructed using a bootstrap sample $\mathbf{T}_b$ from the original training data. RF is robust to dataset probability distributions and variable correlations. Through construction of decision trees from random data subsets, it effectively mitigates the impact of collinearity.

2.2 Ordinary Kriging (OK)

OK is a widely used geostatistical interpolation method for estimating unknown values at unsampled locations from a set of sampled points. It is based on a semivariogram model that describes the spatial autocorrelation of the variable of interest. The semivariogram model is then used to estimate the covariance between any two points as a function of their spatial separation distance. The kriging estimate at an unsampled location is a weighted average of the neighboring sampled values, where the weights are determined by the covariance between each neighboring point and the unsampled location. This estimate can be written as

$$\hat{z}(u) = \sum_{i=1}^{n} \lambda_i z(x_i), \tag{2}$$

where $\hat{z}(u)$ is the estimated value at location u, n is the number of sampled locations, λ_i is the weight assigned to the ith sampled location, and $z(x_i)$ is the known value at the ith sampled location. The weights are determined by minimizing the kriging variance, which is a measure of the uncertainty of the estimation

$$\mathrm{Var}(\hat{z}(u)) = \gamma(0) - \sum_{i=1}^{n} \lambda_i \gamma(h_i), \tag{3}$$

where $\gamma(0)$ is the variance of the target variable, $\gamma(h_i)$ is the spatial autocovariance between the unsampled location u and the sampled location xi, and h_i is the distance between u and xi.

2.3 Spatial Autocorrelation

Spatial autocorrelation refers to the tendency of nearby observations to exhibit similarities, leading to a correlation pattern based on their spatial proximity. Positive spatial autocorrelation indicates clustering of similar values, while negative spatial autocorrelation suggests an inverse relationship (dispersion) between nearby values. Conversely, the absence of spatial autocorrelation implies a random distribution of values across space (Griffith and Peres-Neto 2006).

Spatial autocorrelation, described by Tobler's first law of geography (Tobler 1970), states that neighboring locations tend to have similar values in geographic data such as temperature, groundwater level, or soil properties. Ignoring spatial autocorrelation in statistical analysis can lead to biased estimates, inflated standard errors, and erroneous conclusions. Accounting for spatial autocorrelation is therefore crucial in modeling to ensure accurate and reliable results (Dormann et al. 2007).

Moran's I is a widely used statistic in spatial statistics for measuring spatial autocorrelation, indicating the level of clustering or dispersion in a dataset. It quantifies the similarity between neighboring observations on a scale from -1 (perfect dispersion) to $+1$ (perfect clustering), with 0 denoting no spatial autocorrelation. Moran's I index is computed using the formula

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{4}$$

where N represents the total number of units in the analysis, S_0 is the sum of weights (w_{ij}) for all possible unit combinations, x_i denotes the value of the variable for unit i, $\bar{x}$ is the average value of the variable across all units, and w_{ij} represents the weight between units i and j signifying their spatial relationship.

The p value associated with Moran's I statistic indicates the probability of observing a Moran's I value that is as extreme as or more extreme than the computed value, assuming no spatial autocorrelation. The p value is calculated using the formula

$$p = \frac{n(\text{Permuted Moran's I values} \geq \text{Observed Moran's I value} + 1)}{\Sigma n(\text{permutations}) + 1}. \tag{5}$$

In this formula, the numerator represents the number of permuted Moran's I values greater than or equal to the observed Moran's I value, and the denominator refers to the total number of permutations conducted during the permutation test. A smaller p value indicates stronger evidence against the null hypothesis of no spatial autocorrelation (SA), indicating higher statistical significance (Rey and Anselin 2010; Rey et al. 2023).

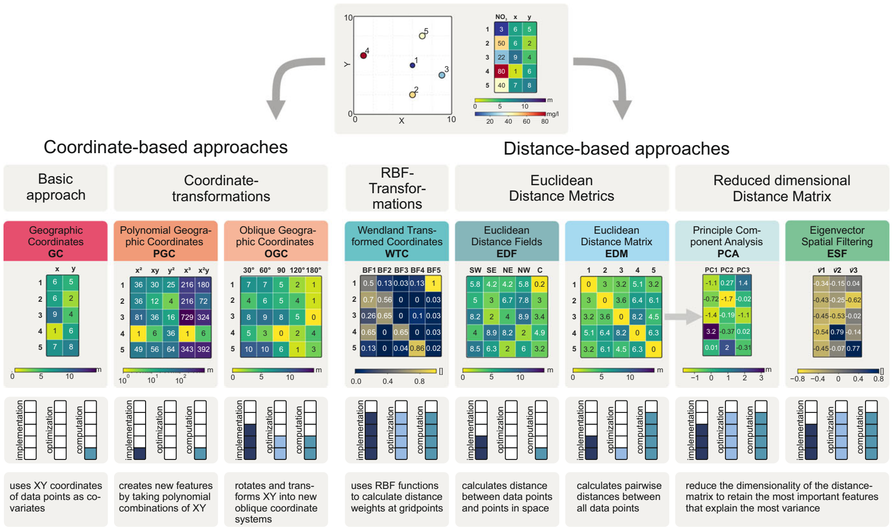
Coordinate-based approaches
Distance-based approaches
Basic approach
Coordinate-transformations
RBF-Transformations
Euclidean Distance Metrics
Reduced dimensional Distance Matrix
Geographic Coordinates GC
Polynomial Geographic Coordinates PGC
Oblique Geographic Coordinates OGC
Wendland Transformed Coordinates WTC
Euclidean Distance Fields EDF
Euclidean Distance Matrix EDM
Principle Component Analysis PCA
Eigenvector Spatial Filtering ESF
implementation
optimization
computation
uses XY coordinates of data points as covariates
creates new features by taking polynomial combinations of XY
rotates and transforms XY into new oblique coordinate systems
uses RBF functions to calculate distance weights at gridpoints
calculates distance between data points and points in space
calculates pairwise distances between all data points
reduce the dimensionality of the distance-matrix to retain the most important features that explain the most variance
X
Y
NO3
mg/l
m

Fig. 1 Overview of methods investigated to incorporate spatial information into a random forest (RF) model. These methods aim to capture spatial dependencies and enhance the performance of the predictive model. The matrices display the feature values generated from the *X* and *Y* coordinates of the sample dataset for each approach

3 Methodology

3.1 Spatial Covariates

To introduce spatial autocorrelation into ML models and emulate established interpolation techniques, several approaches for generating spatial covariates from geographic coordinates or distance matrices can be used. These covariates play a crucial role in enhancing the model's capacity to capture spatial relationships. The study conducted tests on three coordinate-based approaches and five distance-based approaches, which are detailed below. All approaches are summarized in Fig. 1.

3.1.1 Coordinate-Based Approaches

Geographic Coordinates (GC)

Geographic coordinates provide a simple means of integrating the spatial reference of the data into regression models, as shown by their widespread use in practice (Chowdhury et al. 2010; Gilardi and Bengio 2003; Kirkwood et al. 2022; Langella et al. 2010; Tsangaratos et al. 2014; Wadoux 2019; Walsh et al. 2017; Zanella et al. 2017). However, relying solely on geographic coordinates could restrict the model to linear relationships and may not be sufficient to capture the full complexity of spatial patterns and relationships in the data, leading to lower model accuracy and sharp blocky artifacts (e.g., Hengl et al. 2018; Behrens et al. 2018).

Polynomial Geographic Coordinates (PGC)

Polynomial geographic coordinates increase the complexity of regression models by including higher-order terms of geographic coordinates. By using PGC, relationships between the target variable and GC can be modeled as nonlinear, curvilinear, or quadratic. This approach has been shown to improve the ability of the model to capture complex spatial structures (e.g., patterns, gradients, hot spots) that cannot be accounted for by simple linear models (Borcard and Legendre 2002; Li et al. 2011). In PGC, geographical coordinates are converted into higher-order terms (e.g., x^2, y^2, x^3, y^3, etc., and their combinations), which are then included as covariates in regression models. However, a commonly cited drawback of this method is the high correlation between the spatial covariates generated, leading to multicollinearity and a loss of the model's ability to identify independent structures (Borcard and Legendre 2002). While this may result in high accuracy in the training data, the model may not be able to accurately predict outcomes at unsampled locations (Meyer et al. 2019).

Oblique Geographic Coordinates (OGC)

Oblique geographic coordinates are synthetic coordinates calculated from known geographic coordinates at different oblique angles relative to the geographic x-axis. Incorporation of OGC increases the spatial complexity in the model, improving its adaptation to more complex spatial structures. The calculation of OGC for a point with geographic coordinates (a_1, b_1) is performed using the formula

$$\mathrm{OGC} = b_2 = \sqrt{a_1^2 + b_1^2} \cdot \cos(\theta - \arctan(a_1/b_1)), \tag{6}$$

where b_2 is the oblique geographic coordinate of the point, a_1 is the y-coordinate, b_1 is the x-coordinate, and θ is the oblique angle between b_2 and the x-axis. Møller et al. (2020) calculated OGC along n axes at angles between zero and $\pi((n-1)/n)$, with a distance of π/n between them. This method is relatively new and has only been tested by the developers themselves.

3.1.2 Distance-Based Approaches

Wendland Transformed Coordinates (WTC)

Wendland transformed coordinates model spatial dependence using basis functions (e.g., Nychka et al. 2015; Chen et al. 2022). This involves expressing the spatial process $Y(s)$ as a linear combination of known covariates $x(s)$ and a random process $v(s)$ with a general nonstationary covariance function

$$\mathrm{Cov}(v(s), v(s')) = C(s, s'). \tag{7}$$

To incorporate $v(s)$ into ML models, the use of nonlinear basis functions is common. One widely used option for spatial data is the Wendland compactly supported correlation function. In a rectangular grid with grid points u_j, the basis functions are given by

$$\Phi_j^*(s) = \Phi(||s - u_j||/\theta), \tag{8}$$

where θ is a scale parameter that controls the support of the correlation function. Nychka et al. (2015) suggested a rectangular grid model with varying resolutions to achieve a large number of uniform basis functions. The number of basis functions per grid is determined by $K_h = (9 \times 2^h - 1 + 1)d$, where h is the hth grid and d is the spatial dimension. For instance, a two-dimensional model with four grids necessitates $K = 1,830$ basis functions. However, the use of the Wendland compactly supported correlation function with a large dataset can result in computational difficulties due to the significant number of basis functions that must be calculated for each geographic coordinate.

Euclidean Distance Fields (EDF)

Euclidean distance fields provide a distance-based method of representing spatial properties. In this approach, fixed points, usually the four corners (northwest, northeast, southwest, and southeast) of the spatial range of the data points and the center, are used to calculate the Euclidean distance from the sampled or unsampled locations. This technique is independent of individual sample locations and provides information about the location and spatial relationships in the study area. EDF can be used alone or in combination with GC as covariates. Compared to a Euclidean distance matrix (EDM) based on pairwise distances between the points, the resulting distance matrix has fewer columns corresponding to the number of fixed points used. This leads to significantly shorter computation times (Behrens et al. 2018).

Euclidean Distance Matrix (EDM)

Another way of using distance-based covariates to incorporate spatial information into a model is to construct a Euclidean distance matrix $\mathbf{D}$, which represents the pairwise Euclidean distance of the data points

$$\mathbf{D}_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{9}$$

where $\mathbf{D}_{ij}$ represents the distance between data points i and j, and (x_i, y_i) and (x_j, y_j) are the spatial coordinates of the respective data points.

To populate the entire distance matrix, this equation can be applied for all pairs of data points, resulting in a matrix $\mathbf{D}$ that contains the pairwise distances between all data points in the dataset. Distance matrices are commonly used in geostatistical models such as kriging. However, it is important to consider that the size of the distance matrix scales linearly with the number of measurement points, which can result in longer computation times for large datasets (Ahn et al. 2020; Hengl et al. 2018).

Principal Component Analysis of EDM (PCA)

Principal component analysis can be applied to the distance matrix $\mathbf{D}$ to reduce its dimensionality while retaining the essential spatial information. PCA identifies the directions (principal components) in the data that explain the maximum amount of variance. By using the top k principal components, we capture the significant spatial patterns while reducing the dimensionality of the distance matrix.

Eigenvector Spatial Filtering of EDM (ESF)

Eigenvector spatial filtering is another approach for capturing spatial patterns from the distance matrix. ESF involves performing a singular value decomposition (SVD) on the distance matrix $\mathbf{D}$, resulting in the decomposition equation

$$\mathbf{D} = \mathbf{USV}^T. \tag{10}$$

Here, $\mathbf{U}$ represents the matrix of eigenvectors, which capture distinct modes of spatial variation, while $\mathbf{S}$ denotes the diagonal matrix of singular values, and $\mathbf{V}^\mathsf{T}$ is the transpose of the matrix of eigenvectors. The resulting eigenvectors capture different modes of spatial variation. Each eigenvector corresponds to a distinct pattern within the spatial distribution of the data, so we retain the top k eigenvectors to capture the significant spatial structures. In this study, following the methodology introduced by Borcard and Legendre (2002), we tested different maximum neighborhood distance thresholds for the dataset. To accommodate points with a higher neighborhood distance, distances exceeding the threshold were multiplied by a factor of 4. This multiplication aimed to downweight the influence of distant points and mitigate their impact on the spatial relationships under investigation. By amplifying the distances beyond the threshold, the study sought to address potential long-range spatial dependencies while preserving the overall spatial structure of the data.

4 Experimental Setup

4.1 Nitrate Observations in Baden-Wuerttemberg

This study uses mean nitrate concentrations measured from 2016 to 2019 at 1,550 well locations in Baden-Wuerttemberg, Germany (Fig. 2a). The data are available through the Environment, Survey and Nature Conservation (LUBW) groundwater data catalog (LUBW 2023). The dataset includes only wells in the uppermost and unconfined aquifers. To eliminate significant outliers, the 1.5 IQR rule was applied. Any measurements within the wells that exceeded a threshold of 1.5 times the interquartile range (IQR) were removed, while the remaining measurements within the same wells were retained. The nitrate concentrations exhibit strong positive skewness (Fig. 2b), with a range of 0.28 to 78.99 mg/L and a median of 18.04 mg/L. The local Moran scatterplot (Fig. 2c) is used to analyze clustering patterns of a variable, divided into four quadrants: High-High (HH) for high values surrounded by high values, Low-Low (LL) for low values surrounded by low values, and so on. The p value determines the significance of the observed spatial patterns. In the local indicators of spatial association cluster map (LISA), hot spots (high-value clusters surrounded by high-value neighbors) are predominantly found in regions known for their wine production, while cold spots (low-value clusters surrounded by low-value neighbors) are primarily located in the Upper Rhine Graben and the Southwest German Basement (Black Forest) (Fig. 2d). With a positive Moran's I of 0.359 and a highly significant p value of 0.001, we can conclude with strong confidence that there is a nonrandom spatial clustering of the data. There is a statistically significant tendency for similar values to be grouped together in geographic space.

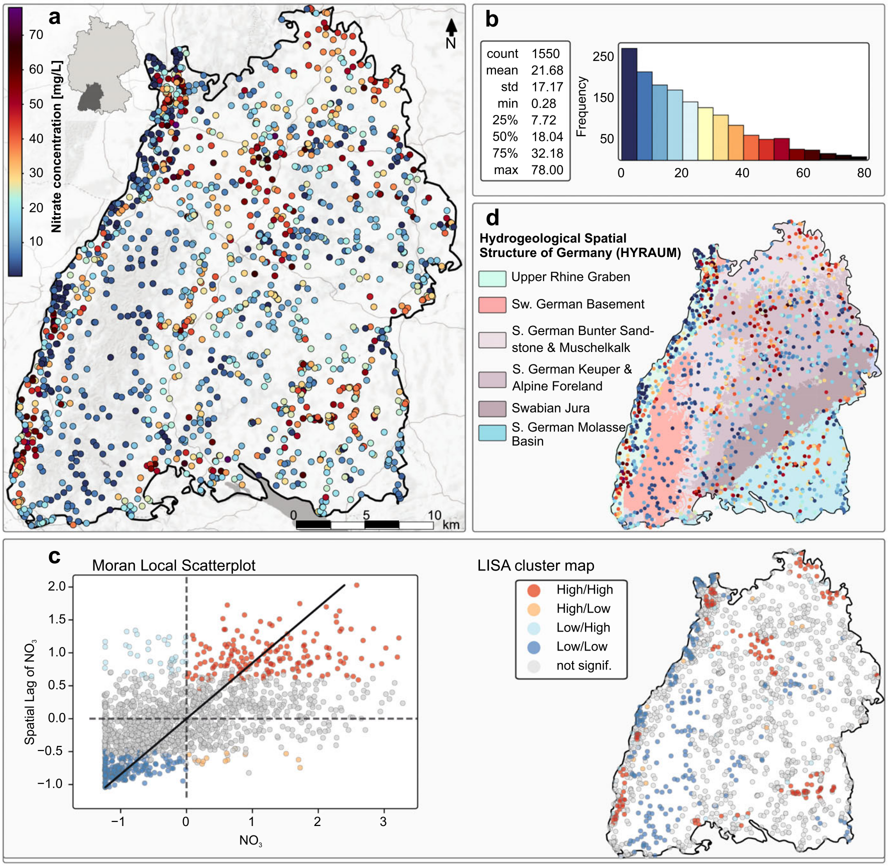

Fig. 2 **a** Spatial distribution of Nitrate measurements in Baden-Wuerttemberg, Germany; **b** histogram illustrating the distribution of nitrate measurements; **c** Moran local scatterplot, and LISA highlighting hot spots (red) and cold spots (blue), as well as outliers (orange and pale blue) at a significance level of $p = 0.05$. **d** Hydrogeological spatial structure of Baden-Wuerttemberg (HYRAUM), as referenced in BGR and SGD (2015)

4.2 Environmental Covariates

Environmental covariates, also known as predictors, encompass a range of variables that characterize the physical, chemical, and biological attributes of an environment. They serve as essential inputs for models used to comprehend and forecast environmental phenomena. These covariates encompass climate data, topographical information, land cover classification, soil properties, hydrological characteristics, remote sensing data, and human-related factors. By integrating these covariates into models, researchers gain valuable insights into environmental processes and make informed decisions for sustainable management practices.

The environmental covariates listed in Table 1 were selected, taking into consideration their conceptual understanding of their influence on nitrate in groundwater.

Table 1 Selected environmental covariates with Bayesian optimization for the prediction of nitrate concentration

Name	Parameter	Description	Type	Source
HÜK250	Hydrolog. units	Hydrogeological Map of Germany 1:250,000 (HÜK250). Hydrogeological characteristics of the upper aquifers	C	BGR and SGD (2019)
BÜK200	Soil units	Soil Map of Germany 1:200,000 (BÜK200). Information on soil type, soil type source rock in Germany	C	BGR and SGD (2020)
CLC5_18	Land cover	CORINE Land Cover 2018, min mapping unit: 5 ha (CLC5), Germany	C	BKG and SGD (2021)
GK1000	Geology	Geological Map of Germany 1:1,000,000 (GK1000)	C	BGR and SGD (2002)
HYRAUM	Hydrolog. regions	Hydrogeological spatial structure of Germany (HYRAUM), regions with similar hydrogeological characteristics	C	(BGR and SGD 2015)
HUMUS1000	Org. matter contents	Organic matter contents in topsoil of Germany 1:1,000,000 (HUMUS1000OB)	C	BGR and SGD (2007)
MUNDIALIS	Land cover	Germany 2019—Land cover classification based on Sentinel-2 data	C	Riembauer et al. (2021)
NDVI	NDVI index	MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid (250 m 16 days NDVI)	N	Didan (2021)
THÜNEN19	Crop types	National-scale crop type maps for Germany from combined time series of Sentinel-1, -2 and Landsat 8 data (2017–2019)	C	Blickensdörfer et al. (2021)

Furthermore, these covariates underwent initial Bayesian input parameter optimization tests, leading to their final selection.

Though multicollinearity poses challenges in regression, our approach tackled it effectively with RF modeling. The ensemble learning of RF inherently handles multicollinearity by constructing multiple decision trees on random data subsets. This property eliminates the need for explicit mitigation measures (Dormann et al. 2013; Lindner et al. 2022). Nonetheless, verifying the method's robustness against collinearity and multicollinearity is vital, with corrective measures taken if necessary.

4.3 Cross-validation Strategies

In this study, we implemented two cross-validation strategies to evaluate the performance of the tested approaches. Specifically, we employed a leave-one-out cross validation (LOO-CV) technique to ensure comparability with the benchmark method of OK, which commonly utilizes this form of cross-validation. This method involves iteratively training the model on all data points except one, and then evaluating its performance on the omitted data point. For the methods involving multiple hyperparameter options, namely OGC, WTC, PCA, and ESF, a preliminary step of 10-fold cross-validation (10-fold CV) was performed to identify the optimal hyperparameters. This two-step approach allowed us to effectively assess the performance of each method while addressing computational challenges that would otherwise arise.

4.4 Model Performance Criteria

We employed the following error metrics to assess the performance of the models. The mean absolute error (MAE) measures the average absolute difference between the predicted and observed values, providing a measure of overall prediction accuracy

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{11}$$

where y_i represents the observed values and $\hat{y}_i$ represents the predicted values.

The root mean squared error (RMSE) calculates the square root of the mean of squared differences between predicted and observed values, representing the magnitude of prediction errors

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}. \tag{12}$$

Additionally, we utilized the R^2 score, which measures the proportion of variance explained by the model relative to the total variance in the data

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \tag{13}$$

The bias (B) measures any systematic deviation between the predicted and observed values and provides insights into the model's tendency to consistently overestimate or underestimate the target variable. It can be calculated as

$$B = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right), \tag{14}$$

where y_i represents the observed values and $\hat{y}_i$ represents the predicted values. A B close to zero indicates that the model predictions are on average unbiased.

In addition, we assessed spatial autocorrelation in the prediction errors using the Moran's I statistic for the residuals (Eqs. 4 and 5). This statistic provides insights into the spatial structure and clustering of errors, with positive or negative values indicating spatial autocorrelation and revealing regions of similarity or dissimilarity in errors. Significant Moran's I values highlight consistent over- or under-prediction areas, helping assess model performance, identify spatial patterns, and understand the relationships among prediction errors.

5 Results

The results section is structured as follows: Sect. 5.1 presents the results of the 10-fold CV for the spatial information approaches, considering various hyperparameter setting options described in Sect. 4.3. Subsequently, Sect. 5.2 provides an overview of the LOO-CV results for all approaches, including the optimized variants discussed in Sect. 5.1. Finally, Sect. 5.3 presents and discusses the results of the spatial predictions.

5.1 10-Fold Cross-Validation Results for Hyperparameter Optimization

Figure 3 illustrates the results of the 10-fold CV for the approaches with various hyperparameter setting options used in parameter optimization. The evaluation metrics used were MAE, RMSE, and R^2. To expedite the process, a 10-fold CV was conducted prior to the more time-consuming LOO-CV. The purpose of this preliminary step was to determine the optimal parameterization for the given dataset.

Step Size of Oblique Angle Rotation, OGC

The step size parameter (Fig. 3, top row) is crucial for calculating the OGC and significantly impacts the model's predictive accuracy for the dataset at hand. It designates the angular increments (in degrees), thereby defining the range and diversity of angles employed in OGC data projections. For instance, employing a step size of 20° generates nine distinct angular projections, spanning from 20° to 180°. With the goal of minimizing RMSE and maximizing R^2, a step size of 3° was chosen for subsequent computations. This decision suggests an optimal trade-off between capturing data complexity and avoiding overfitting at this increment. However, the observation that the error tends to increase inconsistently as the step size extends beyond this point hints at the possibility that larger step sizes, which correspond to fewer oblique projections, may not sufficiently encapsulate the data's inherent complexity.

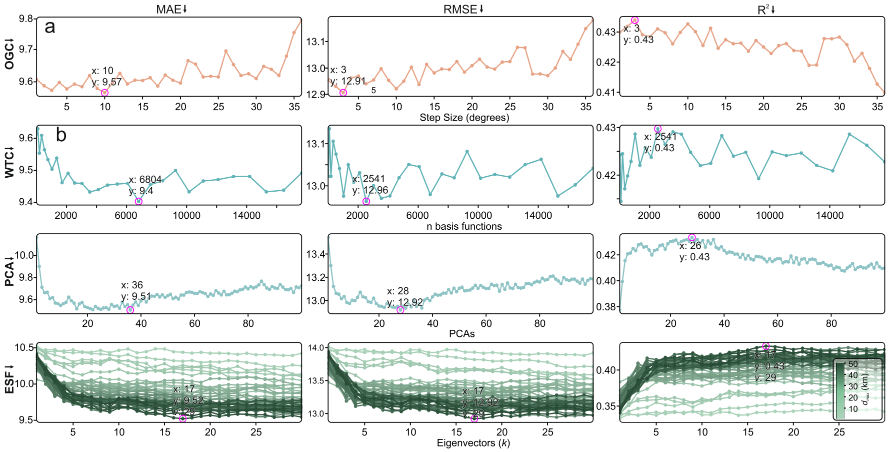

Fig. 3 Overview of 10-fold CV hyperparameter optimization for methods with multiple setting options (OGC, WTC, PCA, and ESF). Rows: OGC errors versus rotation step size, WTC errors versus basis functions, PCA errors versus principal components, and ESF errors versus eigenvectors and neighbor distance

Number of Wendland Basis Functions, WTC

The relationship between the number of basis functions and the performance of the WTC model is intricate, as depicted in Fig. 3, second row. Here, we conducted experiments with a range of 3 to 16,000 functions. Generally, increasing the number of basis functions leads to improved predictive accuracy. However, we found that the highest accuracy was achieved with 2,541 basis functions. Beyond this point, further additions may not consistently enhance accuracy, indicating the potential for overfitting. To optimize the model, we carefully considered the trade-off between complexity and accuracy. By selecting 2,541 basis functions, we aimed to strike a balance where the model captured the underlying data structure without being overly sensitive to noise. This decision was informed by the observation of an elbow point in the accuracy trend, where the addition of more basis functions resulted in diminishing returns.

Number of Principal Components

Increasing the number of principal components generally reduces errors (Fig, 3, third row), with the best performance within the tested range between 20 and 30 principal components. We have selected 28 principal components for subsequent tasks. However, further increasing the number of components does not consistently improve accuracy or capture the underlying data structure, potentially fitting the model to noise.

Number of Eigenvectors and Maximum Neighboring Distance, ESF

The ESF algorithm incorporates two important parameters: the maximum neighboring distance ($d_{\max}$) and the number of eigenvectors (k) derived from the reduced distance matrix (Fig. 3, last row). We conducted tests using a range of 1 to 30 eigenvectors and maximum neighboring distances ($d_{\max}$) between 1 and 50 km to evaluate their influence on the model's performance. Increasing the value of ($d_{\max}$) improves the predictive capabilities of the model by expanding the spatial neighborhood and encompassing a wider array of spatial patterns and relationships. Similarly, including

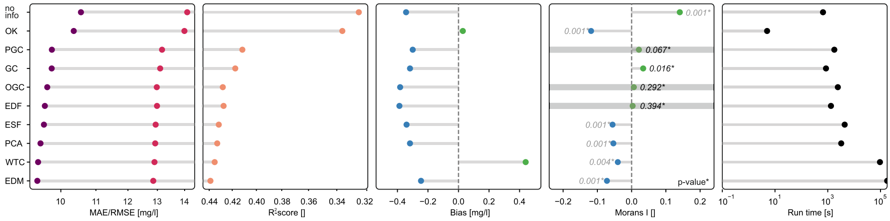

Fig. 4 Performance criteria (MAE, RMSE, R^2-score, Moran's I and calculation time) for the leave-one-out cross-validation of the proposed spatial integration methods, the RF model without spatial information, and the benchmark OK. The significance of the Moran's I results was assessed using p values (*). Residual Moran's I coefficients with p values above 0.05 are considered statistically insignificant and are represented by a gray box in the diagram

a higher number of eigenvectors (k) up to 30 improves the model metrics, such as decreasing MAE and RMSE, and increasing R^2. This improvement can be attributed to the additional spatial information captured by the eigenvectors, representing distinct spatial patterns within the dataset. By exploring different combinations of eigenvectors and assessing their effects on model performance, we gain insights into the optimal selection of these parameters for the ESF algorithm in capturing the spatial variability of nitrate concentrations. We achieved the best results with 17 eigenvectors and a maximum neighboring distance of 29 km. These parameter settings resulted in the lowest errors, including reduced MAE and RMSE, indicating improved model performance.

5.2 LOO Cross-validation Results

Figure 4 presents the detailed results of the leave-one-out cross validation for the selected methods aiming to incorporate spatial information into an RF model. The evaluated metrics include mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), B, Moran's I, Moran's p value, and computation time.

In terms of accuracy, the nonspatial RF model (referred to as "noinfo") and the geostatistical OK approach perform worst among all tested methods, with relatively similar performance (MAE 10.56 mg/L and 10.35 mg/L). Considering the slightly better accuracy and significantly shorter computation time of the OK approach, the additional effort required for the RF model may not be justified, if only these two approaches are taken into account.

As Fig. 4 shows, the inclusion of spatial information significantly enhances the accuracy of the results. While all methods that include spatial information yield results within a relatively narrow range (MAE ranging from 9.38 to 9.75 mg/L, and RMSE ranging from 12.86 to 13.17 mg/L), a clear trend emerges, indicating the superiority of distance-based methods over coordinate-based ones. Put differently, methods that generate more spatial information (EDM and WTC) features enable the model to better capture the underlying structure, albeit at the expense of increased computational time. Notably, the PCA and ESF approaches, which utilize the first principal components or eigenvectors of EDM, strike a balance between accuracy and computational efficiency. Here, PCA slightly outperforms ESF (MAE 9.40 vs. 9.46 mg/L) in our dataset.

OK exhibits the lowest bias among the methods evaluated. This can be attributed to its utilization of a best linear unbiased predictor (BLUP) estimator. Among the evaluated approaches, only WTC exhibits a positive bias, suggesting that the forecasts tend to be, on average, overestimated. In contrast, the other methods demonstrate a consistent, slightly negative bias ranging between −0.25 and −0.38 mg/L.

Analyzing the spatial autocorrelation of residuals, as measured by Moran's I, provides insights into the performance of different modeling approaches. However, the Moran's I values for all methods range from −0.12 to 0.1, indicating a weak or negligible spatial autocorrelation in the residuals. This challenges the ability to draw strong conclusions based solely on Moran's I values. Furthermore, the p values for PGC, OGC, and EDF are above 0.05, indicating that their Moran's I values are statistically insignificant. This suggests that these methods do not exhibit significant spatial patterns in the residuals. The geostatistical OK results exhibit the highest negative Moran's I values, suggesting their effectiveness in capturing spatial patterns accurately. In contrast, the nonspatial RF show the highest positive Moran's I values, indicating limitations in capturing spatial variability. Coordinate-based methods demonstrate slight positive autocorrelation near zero, indicating some spatial similarity among neighboring observations. Meanwhile, distance-based methods display a weak or slightly negative autocorrelation, suggesting a less pronounced spatial structure in the residuals.

5.3 Regionalization Results

In addition to evaluating performance, we also analyzed the regionalization results to assess their plausibility in relation to our conceptual understanding and to identify any potential spatial artifacts. Figure 5 depicts the spatial predictions generated by different approaches, and while they may initially appear similar, closer examination reveals distinct variations in finer details.

The benchmark method, OK, displays distinct bullseye artifacts characterized by concentric rings surrounding the data points. These artifacts indicate an overfitting of the spatial autocorrelation, potentially caused by the limited number of data points and insufficient consideration of the spatial structure. Furthermore, the nitrate distribution demonstrates smooth transitions over considerable distances, which contradicts hydrogeological principles. In reality, groundwater quality typically undergoes rapid changes within short distances due to factors such as land use, hydrogeological processes like transport or degradation, and impermeable boundaries between different aquifers.

In contrast, the spatial predictions from the RF models demonstrate consistent and plausible patterns across all methods, revealing greater spatial variability. The predictions heavily rely on spatial covariates, particularly land use, soil and aquifer properties, while the direct influence of spatial information appears more localized. This observation aligns with the conceptual understanding that groundwater quality is primarily influenced by local factors. However, it is crucial to complement the analysis with domain knowledge regarding nitrate contamination in specific locations for a comprehensive evaluation.

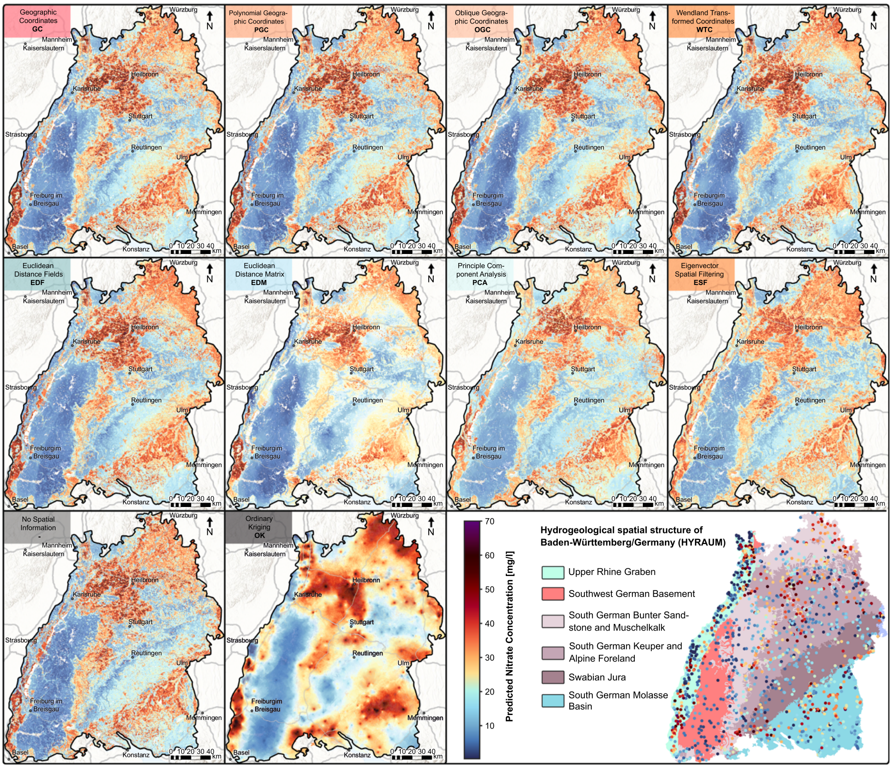
Geographic Coordinates GC
Polynomial Geographic Coordinates PGC
Oblique Geographic Coordinates OGC
Wendland Transformed Coordinates WTC
Euclidean Distance Fields EDF
Euclidean Distance Matrix EDM
Principle Component Analysis PCA
Eigenvector Spatial Filtering ESF
No Spatial Information -
Ordinary Kriging OK
Würzburg
N
Mannheim
Kaiserslautern
Heilbronn
Karlsruhe
Stuttgart
Strasbourg
Reutlingen
Ulm
Freiburg im Breisgau
Memmingen
Basel
Konstanz
0 10 20 30 40 km
Predicted Nitrate Concentration [mg/l]
70
60
50
40
30
20
10
Hydrogeological spatial structure of Baden-Württemberg/Germany (HYRAUM)
Upper Rhine Graben
Southwest German Basement
South German Bunter Sandstone and Muschelkalk
South German Keuper and Alpine Foreland
Swabian Jura
South German Molasse Basin

Fig. 5 Spatial prediction of nitrate distribution in groundwater in Baden-Wuerttemberg using the spatial random forest model without spatial information, eight investigated spatial information approaches, and interpolation using univariate ordinary kriging

Among the approaches utilizing linear geographic coordinates as spatial information, the GC method displays some pronounced linear artifacts, primarily observed in the region of South German Bunter Sandstone and Muschelkalk, located approximately halfway between Strasbourg and Stuttgart. This phenomenon likely arises from an exaggerated influence of the X-coordinate based on dominant trends along the east–west axis. In approaches like PGC and OGC, incorporating higher-order polynomial terms or employing new features from a rotated coordinate system, these artifacts are observed to a lesser extent, although they may still be visible in certain areas.

Significant variations in predictions between the different approaches are observed in areas where nitrate hot spots border nitrate low spots, particularly in the southern region of South German Bunter Sandstone and Muschelkalk, east of Freiburg, and the northern region, northeast of Heilbronn. In these localized areas, the application of EDM with only important PCA or ESF components as input features, or the use of WTC, leads to relatively lower nitrate predictions compared to other methods.

While all results from the distance-based approaches are generally plausible, it is difficult to determine which regionalized pattern is more likely without further investigations on-site or more detailed knowledge of the local situations. Additional information or field studies are necessary to gain a better understanding of the specific factors influencing the nitrate distribution in these areas.

6 Conclusion

In this study, we explored different approaches to incorporate spatial information into a random forest model for predicting nitrate concentrations in groundwater. We compared the performance of these approaches using cross-validation techniques and evaluated the accuracy and spatial patterns of the predictions among each other and to a nonspatial RF and geostatistical univariate ordinary kriging approach. The results of the cross-validation analysis demonstrated that the inclusion of spatial information significantly improved the predictive accuracy of the model compared to the nonspatial RF and geostatistical ordinary kriging approaches.

Among the spatial approaches investigated in this study, methods that incorporated distance-based features, namely the EDM, WTC, PCA, and ESF, demonstrated superior overall performance in terms of accuracy metrics and plausibility of spatial predictions. Notably, the EDM and WTC methods emerged as the top performers. However, it is important to consider that these methods can become computationally intensive, particularly when dealing with large datasets, due to the creation of numerous additional features. In contrast, the PCA and ESF approaches, which utilized the principal components or eigenvectors of the EDM, strike a balance between accuracy and computational efficiency, making them a favorable choice for practical implementation.

The spatial predictions of nitrate concentrations displayed variations across the different approaches examined in this study. The RF models that incorporated spatial information exhibited consistent and plausible patterns overall. However, in certain regions, particularly at the boundaries between nitrate hot spots and low spots, artifacts and deviations were observed, specifically with the coordinate-based approaches. These artifacts may be attributed to the influence of the coordinates and their associated spatial trends in these regions.

In conclusion, integrating spatial information distinctly improved nitrate concentration predictions in groundwater using the RF model. Distance-based methods, such as EDM, WTC, PCA, and ESF, performed well in capturing spatial patterns and enhancing accuracy, without producing artifacts in regionalization.

While integrating spatial information holds promise for improving regionalization of environmental parameters, our study encounters some challenges. The dataset may not fully capture regional nitrate variability due to both high spatial and temporal dynamics, as well as the inherently sparse data available from costly groundwater monitoring wells. Therefore, spatial predictions of groundwater parameters are generally always subject to great uncertainty. Additional spatial predictors, particularly those related to potential nitrate input pathways, are likely to enhance predictive accuracy even further. Additionally, other data with different spatial resolutions and autocorrelation levels could yield varied results. Using distance-based methods introduces computational challenges, especially with large datasets. Addressing artifacts in spatial predictions demands deeper spatial process understanding.

While our analysis focused on a specific dataset, we hypothesize that the findings can be extrapolated to other environmental data and regionalization problems employing ML methods. Although the results may vary among different approaches depending on the level of autocorrelation in the data, it is essential to consistently consider the incorporation of spatial information into ML models. Furthermore, it is advisable to employ cross-validation techniques and plausibility checks when testing and comparing various approaches. By doing so, researchers can ensure robust and reliable predictions in regionalization tasks related to diverse environmental datasets.

Acknowledgements This study is a contribution to the project: Nitrate Monitoring 4.0 - Intelligent Systems for Sustainable Reduction of Nitrate in Groundwater (NiMo 4.0), funded by the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) on the basis of a resolution of the German Bundestag.

Author Contributions MO: conceptualization, methodology, software, formal analysis, validation, investigation, visualization, writing—original draft preparation. FD: conceptualization, methodology, software, formal analysis, validation, investigation. TL: conceptualization, methodology, writing—review and editing, supervision.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement The well data are publicly available at the web service of the Baden-Wuerttemberg State Office for Environment (LUBW 2021).

Declarations

Code availability The code to perform the calculations shown in this manuscript can be found on: Ohmer (2023) or on https://github.com/marcohmer/Spatial_Information_RF.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahn S, Ryu DW, Lee S (2020) A machine learning-based approach for spatial estimation using the spatial features of coordinate information. ISPRS Int J Geo-Inf 9(10):587

Behrens T, Schmidt K, Viscarra Rossel RA, Gries P, Scholten T, MacMillan RA (2018) Spatial modelling with Euclidean distance fields and machine learning: spatial modelling with Euclidean distance fields. Eur J Soil Sci 69(5):757–770

BGR and SGD (2002) Geological Map of Germany 1:1,000,000 (GK1000): Federal Institute for Geosciences and Natural Resources (BGR). Hannover. Digital map data. Available online at: https://services.bgr.de/geologie/gk1000. Accessed 22 Oct 2024

BGR and SGD (2007) Organic matter contents in top soils of Germany 1:1,000,000 (HUMUS1000OB), Hannover, 2007. Digital map data. Available online at: https://services.bgr.de/boden/humus1000ob. Accessed 22 Oct 2024

BGR, SGD (2015) Hydrogeological spatial structure of Germany (HYRAUM). Digital map data

BGR and SGD (2019) Hydrogeological Map of Germany 1:250,000 (HÜK250). Federal Institute for Geosciences and Natural Resources (BGR) and German State Geological Surveys (SGD), Hannover. Digital map data. Avialable online at: https://www.bgr.bund.de/huek200. Accessed 22 Oct 2024

BGR and SGD (2020) Soil Map of Germany 1:200,000 (BÜK200). Federal Institute for Geosciences and Natural Resources (BGR) and German State Geological Surveys (SGD), Hannover. Digital map data. Available online at: https://www.bgr.bund.de/buek200. Accessed 22 Oct 2024

BKG and SGD (2021) WMS CORINE LAND COVER 5 HA - Status 2018. The Federal Agency for Cartography and Geodesy (BKG), Frankfurt am Main. Digital map data. Available online at: https://gdz.bkg.bund.de/index.php/default/corine-landcover-5-ha-stand-2018-clc5-2018.html. Accessed 22 Oct 2024

Blickensdörfer L, Schwieder M, Pflugmacher D, Nendel C, Erasmi S, Hostert P (2021) National-scale crop type maps for Germany from Combined Time Series of Sentinel-1, Sentinel-2 and Landsat 8 data (2017, 2018 and 2019)

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol Model 153(1–2):51–68

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr Anal 28(4):281–298

Chen W, Li Y, Reich BJ, Sun Y (2022) DeepKriging: spatially dependent deep neural networks for spatial prediction. ArXiv:2007.11972 [cs, stat]

Chowdhury M, Alouani A, Hossain F (2010) Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater. Stoch Environ Res Risk Assess 24(1):1–7

Credit K (2022) Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. Geogr Anal 54(1):58–83

Didan K (2021) MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061. NASA EOSDIS Land Processes DAAC

Diniz-Filho JAF, Bini LM (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters: spatial filtering of richness data. Glob Ecol Biogeogr 14(2):177–185

Dormann FC, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Daniel Kissling W, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30(5):609–628

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1):27–46

Du Z, Wang Z, Wu S, Zhang F, Liu R (2020) Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. Int J Geogr Inf Sci 34:1–25

Fotheringham S, Yang W, Kang W (2017) Multiscale geographically weighted regression (MGWR). Ann Am Assoc Geogr 107:1–19

Gilardi N, Bengio S (2003) Comparison of four machine learning algorithms for spatial data analysis, p 16

Griffith DA, Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87(10):2603–2613

Hengl T, Nussbaum M, Wright MN, Heuvelink GB, Gräler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518

Islam MD, Li B, Lee C, Wang X (2022) Incorporating spatial information in machine learning: the Moran eigenvector spatial filter approach. Trans GIS 26(2):902–922

Karimanzira D, Weis J, Wunsch A, Ritzau L, Liesch T, Ohmer M (2023) Application of machine learning and deep neural networks for spatial prediction of groundwater nitrate concentration to improve land use management practices. Front Water Sec Water Artif Intell 5:1193142

Kiely TJ, Bastian ND (2020) The spatially conscious machine learning model. Stat Anal Data Min ASA Data Sci J 13(1):31–49

Kirkwood C, Economou T, Pugeault N, Odbert H (2022) Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Math Geosci 54(3):507–531

Knoll L, Häußermann U, Breuer L, Bach M (2020) Spatial distribution of integrated nitrate reduction across the unsaturated zone and the groundwater body in Germany. Water 12(9):2456

Langella G, Basile A, Bonfante A, Terribile F (2010) High-resolution space-time rainfall analysis using integrated ANN inference systems. J Hydrol 387(3–4):328–342

Leirvik T, Yuan M (2021) A machine learning technique for spatial interpolation of solar radiation observations. Earth Space Sci 8(4).

Li J, Heap AD, Potter A, Daniell JJ (2011) Application of machine learning methods to spatial interpolation of environmental variables. Environ Model Softw 26(12):1647–1659

Liang M, Zhang L, Wu S, Zhu Y, Dai Z, Wang Y, Qi J, Chen Y, Du Z (2023) A high-resolution land surface temperature downscaling method based on geographically weighted neural network regression. Remote Sens 15(7):1740

Lindner T, Puck J, Verbeke A (2022) Beyond addressing multicollinearity: robust quantitative analysis and machine learning in international business research. J Int Bus Stud 53:1307–1314

Liu X, Kounadi O, Zurita-Milla R (2022) Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. ISPRS Int J Geo-Inf 11(4):242

LUBW (2021) Umwelt-Daten und -Karten Online (UDO). The State Institute for Environment Baden-Württemberg (LUBW), Karlsruhe. Available online at: https://udo.lubw.baden-wuerttemberg.de/public/. Accessed 23 Oct 2024

LUBW (2023) Groundwater Monitoring Program -Annual Data Catalog Groundwater. The State Institute for Environment Baden-Württemberg (LUBW), Karlsruhe. Digital data. Available online at: https://umweltdaten.lubw.baden-wuerttemberg.de/. Accessed 23 Oct 2024

Meyer H, Reudenbach C, Wöllauer S, Nauss T (2019) Importance of spatial predictor variable selection in machine learning applications—moving from data reproduction to spatial prediction. Ecol Model 411:108815

Møller AB, Beucher AM, Pouladi N, Greve MH (2020) Oblique geographic coordinates as covariates for digital soil mapping. SOIL 6(2):269–289

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution Gaussian process model for the analysis of large spatial datasets. J Comput Graph Stat 24(2):579–599

Ohmer M (2023) Code to incorporating spatial information for regionalization of environmental parameters in machine learning models. marcohmer/Spatial_information_rf. https://doi.org/10.5281/zenodo.8108637

Ohmer M, Liesch T, Goeppert N, Goldscheider N (2017) On the optimal selection of interpolation methods for groundwater contouring: an example of propagation of uncertainty regarding inter-aquifer exchange. Adv Water Resour 109:121–132

Ransom K, Nolan B, Stackelberg P, Belitz K, Fram M (2022) Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. Sci Total Environ 807:151065

Rey SJ, Anselin L (2010) PySAL: a Python library of spatial analytical methods. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis: software tools, methods and applications. Springer, Berlin

Rey SJ, Arribas-Bel D, Wolf LJ (2023) Geographic data science with Python. Chapman & Hall/CRC texts in statistical science. CRC Press, Boca Raton

Riembauer G, Weinmann A, Xu S, Eichfuss S, Eberz C, Neteler M (2021) Germany-wide Sentinel-2 based land cover classification and change detection for settlement and infrastructure monitoring. In: Proceedings of the 2021 Conference on Big Data from Space (BiDS'2021). Publications Office of the European Union, Luxembourg

Sekulić A, Kilibarda M, Heuvelink GB, Nikolić M, Bajat B (2020) Random forest spatial interpolation. Remote Sens 12(10):1687

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234

Tsangaratos P, Rozos D, Benardos A (2014) Use of artificial neural network for spatial rainfall analysis. J Earth Syst Sci 123(3):457–465

Wadoux AMC (2019) Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351:59–70

Walsh ES, Kreakie BJ, Cantwell MG, Nacci D (2017) A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. PLOS ONE 12(7):e0179473

Wang H, Huang Z, Yin G, Bao Y, Zhou X, Gao Y (2022) Gwrboost: a geographically weighted gradient boosting method for explainable quantification of spatially-varying relationships

Zanella L, Folkard AM, Blackburn GA, Carvalho LMT (2017) How well does random forest analysis model deforestation and forest fragmentation in the Brazilian Atlantic forest? Environ Ecol Stat 24(4):529–549