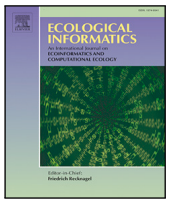
Ecological Informatics 86 (2025) 103057

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Spatial autocorrelation in machine learning for modelling soil organic carbon

Alexander Kmoch *, Clay Taylor Harrison, Jeonghwan Choi, Evelyn Uuemaa

Landscape Geoinformatics Lab, Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, Tartu, 51003, Tartu, Estonia

ARTICLE INFO

Dataset link: 10.5281/zenodo.14236579, 10.5281/zenodo.14236923

Keywords:
Soil mapping
Random forest
Environmental covariates
Spatial modelling

ABSTRACT

Spatial autocorrelation, the relationship between nearby samples of a spatial random variable, is often overlooked in machine learning models, leading to biased results. This study compares various methods to account for spatial autocorrelation when predicting soil organic carbon (SOC) using random forest models. This kind of systematic comparison has not been done previously. Five models incorporating spatial structure were compared against baseline models with no added spatial components. Cross-validation showed slight improvements in accuracy for models considering spatial autocorrelation, while Shapley Additive Explanations confirmed the importance of spatial variables. However, no decrease in spatial autocorrelation of residuals was observed. Random Forest Spatial Interpolation emerged as the top performer in capturing spatial structure and improving model accuracy. Raster-based models exhibited enhanced prediction detail. The findings emphasize the value of incorporating spatial autocorrelation for better prediction of SOC with machine learning. Considerations such as the spatial distribution of predictions and computational complexity should help guide the selection of suitable approaches for specific spatial modelling tasks.

1. Introduction

Soil is essential to all life, as it supports the plants that generate oxygen and supports the base of the food web. However, despite the rapid growth of readily available remotely sensed data, soil properties are difficult to map from space. At the same time, it is impossible to study large areas by mapping in the field due to the high time requirements and cost (Wadoux et al., 2019). Consequently, digital soil mapping (predictive spatial modelling) is becoming an irreplaceable tool for capturing the spatial variability of soil properties (Brungard et al., 2015; Lamichhane et al., 2019).

Machine learning (ML), particularly Random forest (RF) models, are the most widely used type in soil predictive modelling because they are interpretable and fast (Duan et al., 2024; Wadoux et al., 2020a; Heung et al., 2016). Unfortunately, most RF frameworks do not account for certain distinctive properties of spatial data that set it apart from other data types, especially the phenomenon of spatial dependence, i.e., spatial autocorrelation (Sekulić et al., 2020; dos Santos et al., 2023). Failure of the RF model to appropriately account for spatial autocorrelation can undermine the modelling results (Nikparvar and Thill, 2021). However, spatial autocorrelation also represents an opportunity in spatial RF modelling, especially if the modelled phenomena exhibit strong environmental gradients; the RF algorithm can implicitly learn from changes that occur along the geographical extent. However, this also creates a problem: autocorrelation violates the assumption of sample independence (Dormann et al., 2007). Moreover, autocorrelation can cause some covariates (features) to completely overpower the predictive performance of other covariates and may even lead to false conclusions about the relationships among environmental variables (Ploton et al., 2020). Therefore, it is extremely important to explicitly consider spatial autocorrelation in spatial predictive RF models while also being able to separate the predictive power created by autocorrelation from the predictive power of other covariates.

Several methods have been developed to account for spatial patterns/spatial autocorrelation in supervised machine learning models. A common approach in the literature is to incorporate covariates that capture spatial structure, like distances (Behrens et al., 2018), spatial coordinates (Hengl et al., 2018), properties of neighbours (Sekulić et al., 2020), etc., into the model data. Another approach is to add another step to the prediction process: examples include creating ensembles of local and global models (Georganos et al., 2021; Brunsdon et al., 2010), or using ordinary kriging interpolation on ML training data residuals to estimate an error term that can be incorporated into testing predictions (Fox et al., 2020). However, Jemeljanova et al. (2024) concluded in their recent review on adapting machine learning for environmental spatial data that there is no single systematic approach for addressing spatial autocorrelation in machine learning models in general. Applying several methods of incorporating spatial autocorrelation to the same model and comparing their performance

* Corresponding author.
E-mail address: alexander.kmoch@ut.ee (A. Kmoch).

https://doi.org/10.1016/j.ecoinf.2025.103057
Received 8 September 2024; Received in revised form 27 January 2025; Accepted 27 January 2025
Available online 5 February 2025

could provide valuable insights towards a better understanding of which method to choose in a particular situation. While existing literature suggests various methods for incorporating spatial autocorrelation, there is no consensus on a single, comprehensive method that uniformly improves model performance across different scenarios Jemeljanova et al. (2024) and Sarkar et al. (2024). Furthermore, the potential benefits of comparing multiple methods of addressing spatial autocorrelation within the same model framework remain unexplored. Consequently, research is needed to systematically evaluate different approaches to incorporate spatial autocorrelation and analyse their effectiveness in enhancing the predictive accuracy of soil organic carbon models.

We aim to test different accounting methods for spatial autocorrelation in machine learning to identify which method would give the best results for spatial modelling of soil organic carbon (SOC). To our knowledge, this is the first study systematically evaluating the main approaches to including autocorrelation in machine learning for SOC mapping. With this study, we provide critical insights into selecting appropriate methodologies for SOC spatial modelling tasks, ultimately improving the reliability of soil predictive models.

2. Data and methodology

We employ a comprehensive approach to modelling soil organic carbon (SOC) across Estonia, utilizing the combination of soil, elevation, land use, drainage, and normalized vegetation index data to capture the spatial heterogeneity of SOC. For baseline models, we implement two primary strategies: (1) a vector-based method that utilized the Estonian soil map EstSoil-EH (Kmoch et al., 2021) polygons for training and predicting SOC, and (2) a raster-based method where all original raster covariates (including DEM derivatives, land use, and NDVI) were maintained in raster format, while the covariates originating from EstSoil-EH (i.e. digitized fine-earth and rock fractions) from the vector polygons were converted to a 10-meter raster. We then augment the raster baseline model with spatially aware techniques by incorporating (1) sample coordinates, (2) k-nearest neighbours, (3) buffer distances, (4) kriging of prediction residuals, and (5) geographically weighted regression using localized random forest models for each sampling point.

2.1. Data

2.1.1. Data and case study area

We used soil organic carbon field samples as the target variable and soil texture, elevation derivatives, land use, normalized vegetation index, and drainage information as predictors. *Soil Organic Carbon*

We used soil organic carbon (SOC) measurements for Estonia (study area) from three sources:

1. 472 samples from EstSoil-EH (Kmoch et al., 2021) that have been collected from forest areas, open and overgrown grasslands and alvars, peatlands and arable soil transects;
2. 194 samples from European Union LUCAS Land Use/Cover Area frame Survey (LUCAS) (European Commission Joint Research Centre, 2022) data for topsoil SOC. This topsoil survey represents the first attempt to build a consistent spatial database of the soil cover across the EU based on standard sampling and analytical procedures, with all soil samples being analysed in a single laboratory;
3. 303 samples from a project *Role of grasslands in mitigating climate change* (Helm, 2023)

Altogether, we used 969 samples to train and validate the machine learning models (Fig. 1) which covered different main land use types A.4.

Soil Data

For soil data, we used EstSoil-EH dataset (Kmoch et al., 2021), which is an *an eco-hydrological modelling parameters dataset derived from the Soil Map of Estonia* (Estonian Land Board, 2017). EstSoil-EH consists of 750,000+ soil units in vector format. The following covariates were used from the soil data: clay fraction (%), silt fraction (%), sand fraction (%), rock fraction (%).

Elevation Data

We used LiDAR-based 5 m DEM from the Estonian Land Board (Estonian Land Board, 2022). We created several derivatives from DEM: terrain wetness index (TWI), terrain roughness index (TRI), LS-Factor (LSF), and slope.

Land use

We derived land use from the land use basemap created under project *Mapping and Assessment of Ecosystems and their Services in Estonia (ELME)* (Helm et al., 2020) where land use types were aggregated from various databases.

Drainage

The drainage regime considered both underground tile drainage and ditch-based drainage systems. The drainage ditches were derived from the Estonian Topographic Database (ETAK) (Estonian Land Board, 2023a). A 100 m buffer was created around ditches as the drainage influence zone. The tile drainage information as polygons was obtained from the official register of drainage systems by the Agricultural Board of the Ministry of Rural Affairs of Estonia. The two layers were merged, and the resulting layer consisted of binary information with a value of 1 for drained areas and 0 otherwise.

NDVI

The Normalized Difference Vegetation Index (NDVI) measures the amount of green vegetation in an area. We derived NDVI for June-August 2022 composite image of Sentinel 2 using Google Earth Engine.

Table 1 summarizes all datasets and their provenance.

2.2. Methods

First, we constructed one baseline raster and one baseline vector model. Next, we implemented five spatially-aware RF models by augmenting the baseline raster model. Finally, we selected model hyperparameters and evaluated model performance using 5-fold cross-validation.

2.2.1. Baseline models

To provide a baseline for comparison, we trained two RF models with no added spatial features, only with the described covariates: (1) a vector-based approach where the raster-based elevation-derived covariates and NDVI were aggregated to the original vector-based soil units of EstSoil-EH (Kmoch et al., 2021) by using zonal statistics and calculating mean, median and standard deviation. The drainage was represented as a percentage of the soil unit's area, and the land use as the majority land use type within each soil unit; (2) a raster-based approach where all original raster covariates (DEM derivatives, land use, NDVI, drainage) were kept as rasters and original soil data from vector polygons was converted into the same 10 m raster format. These models can be compared to understand the effects of transitioning from vector-based to raster-based modelling. The baseline raster model will also serve as a reference for evaluating the impact of adding spatial features to the modelling process.

For both models, we converted `landuse` from a single categorical feature into five binary "dummy" features (`landuse_type_arable`, `landuse_type_grassland`, `landuse_type_forest`, `landuse_type_artificial`, and `landuse_type_wetland`).

All raster datasets were co-registered using GDAL Warp (GDAL/OGR contributors, 2024) to 10 m resolution, aligning pixels and matching extents. A resolution of 10 m was chosen because it is the coarsest resolution among the available rasters, and upsampling to a finer resolution would not be appropriate for model-training purposes. For sources that existed as vectors, we used GDAL vector-to-raster, using

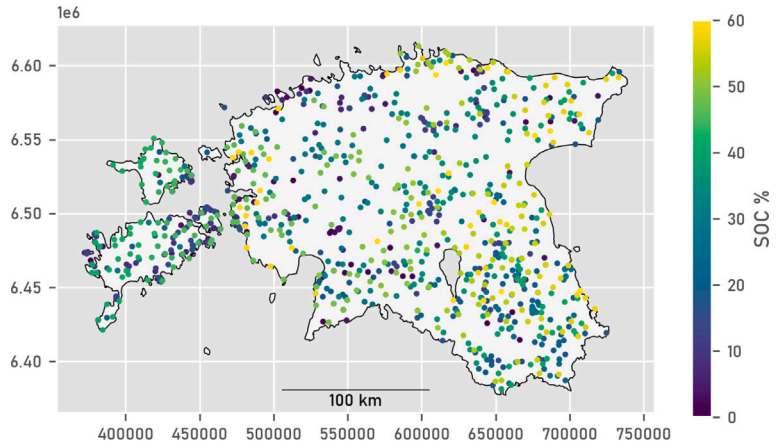


Fig. 1. Spatial distribution of soil samples in dataset across Estonia.

Table 1
Summary of data features and their origin.

Feature name	Description	Provenance	Original resolution	Data type
clay	Clay fraction (top layer)	EstSoil	Vector	Numeric
silt	Silt fraction (top layer)	EstSoil	Vector	Numeric
sand	Sand fraction (top layer)	EstSoil	Vector	Numeric
rock	Rock fraction (top layer)	EstSoil	Vector	Numeric
twi	Terrain wetness index	LiDAR DEM	5 m	Numeric
tri	Terrain roughness index	LiDAR DEM	5 m	Numeric
slope	Slope	LiDAR DEM	5 m	Numeric
lsf	LS-factor	LiDAR DEM	5 m	Numeric
landuse	Land use classification	ELME	10 m	Dummy
ndvi	Mean NDVI, July 2022	Sentinel-2	10 m	Numeric
drained	Boolean value for drainage	ETAK	Vector	Dummy
soc	Soil organic carbon (% of mass)	LUCAS +other fieldwork	Point data	Numeric

the previously generated rasters as the extent to co-register with the others. We used Python package `rasterio` to sample values from the rasters onto the SOC data points. A summary of descriptive statistics for all the continuous variables as sampled at SOC observation locations is shown in Table A.5.

2.2.2. Adding spatial awareness

Coordinates as covariates (XY)

The simplest way to incorporate spatial information into a model is to include the coordinates of the samples as covariates. This can potentially be useful if the target variable has a linear (or otherwise) relationship to easting or/and northing, or as a proxy variable for some other related covariate such as distance-to-coastline. Still, it does not inherently provide a way for a random forest to model autocorrelation between near points and may lead to blocky artefacts in predicted surfaces (Hengl et al., 2018; Behrens et al., 2018).

We included the X and Y coordinates of the SOC data points in the model as additional covariates (Fig. 2).

Random Forest Spatial Interpolation (RFSI)

Random forest spatial interpolation (RFSI) is a framework that considers the nearest observations and their distances to the prediction location as covariates in a random forest model. Sekulić et al. (2020) developed RFSI and compared it to other interpolation techniques like kriging, regression kriging, random forest, and random forest for spatial prediction (RFsp; Hengl et al., 2018), with three case studies involving synthetic data, daily precipitation data in Catalonia, Spain and mean daily temperature data in Croatia. Results showed that RFSI performed better than most deterministic interpolation techniques and performed similarly to inverse distance weighting and RFsp.

We augmented the training data with $2 * K$ covariates: K features representing the values of the target variable of the K nearest neighbours of the available sampling points, and K features representing the distances to those neighbours. When predicting, these covariates were also derived from neighbours in the training dataset. We selected $K = 7$ neighbours based on a grid-search test across the 5-15 neighbour range as recommended by Sekulić et al. (2020), eventually adding 14 additional covariates to the model (Fig. 2). The structure of the training data is shown in Fig. 2. In this case, n1 is the nearest neighbour in the training dataset that is not the current sampling point itself.

Buffer distances (BD)

The RFsp prediction framework also uses buffer distances between observation points to account for spatial autocorrelation. It can provide predictions that are as accurate and unbiased as kriging but with several advantages. RFsp is described to not rely on strict statistical assumptions about the target variable's distribution or stationarity, to be flexible in incorporating and extending different covariates, and to be able to yield more informative maps showing prediction error. Alone without additional environmental covariates, RFsp can be considered to produce results similar to ordinary kriging. But it is particularly designed to create multivariate spatial prediction models for geoscience applications (Hengl et al., 2018).

We calculated the distance from each sample to every point in the training data, and then grouped those distances into N ordered bins (per sample) of approximately the same size. A covariate was added to the model for every point in the training data, with values representing the bin that point fell into for each sample. We used $N = 20$ bins.

Fig. 2 shows the model's training data structure. Note that the first bin is bin 1, and each point is in its own bin 1. Also note that if point m is in point n's bin i, point n is not necessarily in point m's bin i as well.

Random Forest Regression Kriging (RFRK)

Kriging is an interpolation technique which fits a mathematical function, typically a Gaussian process, to a set of sample points. The weights assigned to the neighbouring points depend on their distance

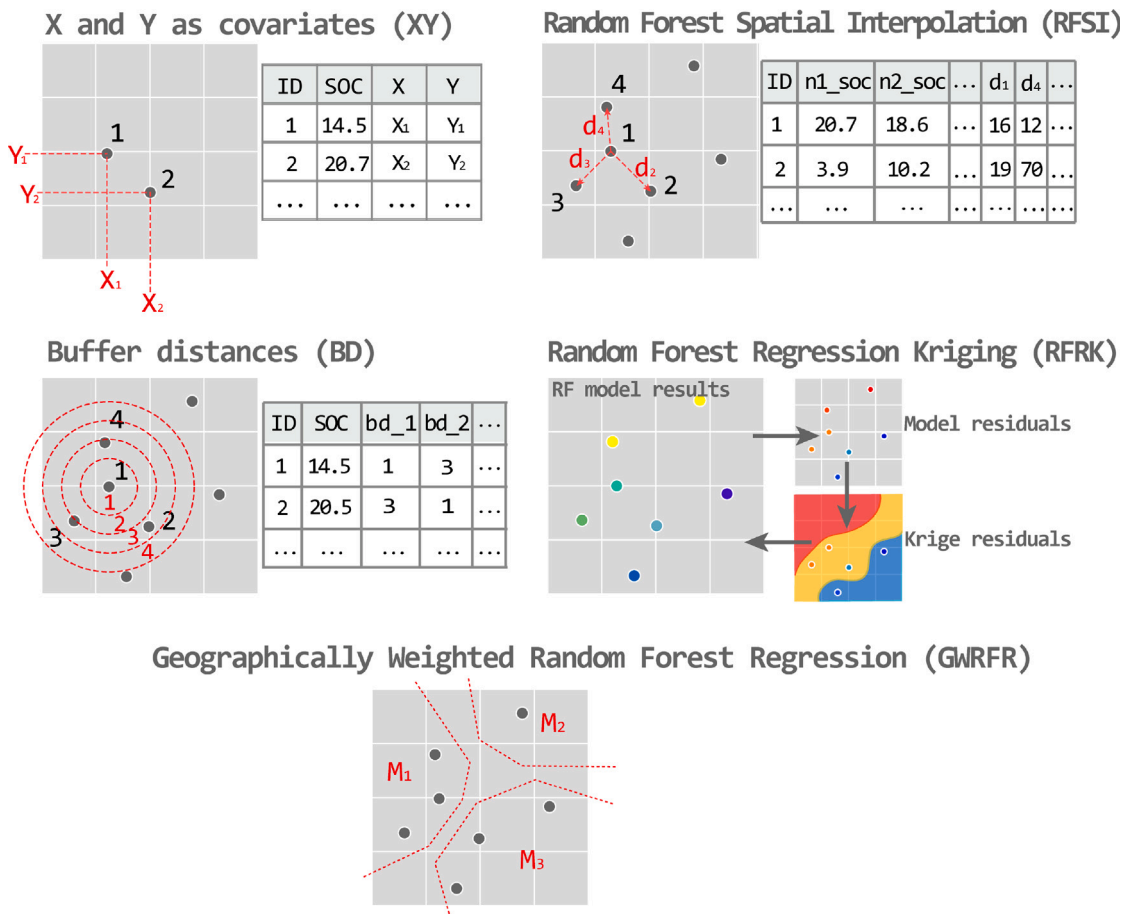


Fig. 2. Methods for incorporating spatial structure into random forest models.

and degree of clustering. Nearby points are assigned greater weights (compared to farther points), while dispersed points are assigned smaller weights (compared to clustered ones). In this way, the technique uses spatial autocorrelation in the data to predict values at unsampled locations.

Kriging and random forests are often used together in ensemble methods to improve the random forests' predictions. The typical approach is Random Forest Regression Kriging, in which the residuals of the random forest's predictions are kriged. Then, the kriging surface's predictions are added to the random forest results which potentially increases the accuracy (Fox et al., 2020; Hengl et al., 2017).

To implement random forest regression kriging, we first trained the baseline raster model, and then used it to make predictions on its training data. The residuals of these predictions were used to fit a semivariogram for ordinary kriging (a spherical model, with a nugget of 25, a sill of 31, and a range of 100,000 m). These residuals were interpolated across the study area using ordinary kriging with the calculated semivariogram. When predictions are made, the value of the kriging surface at that location is added to the baseline model's prediction, acting as an error correction (Fig. 2).

Geographically Weighted Random Forest Regression (GWRFR)

Geographically Weighted Regression (GWR) accounts for spatial nonstationarity in regression models by fitting a regression to each feature in the dataset that is trained only on the points in that feature's neighbourhood, as well as a global model including all training features (Brunsdon et al., 2010). GWRFR applies the principles of GWR using random forests as the regression model. This has the advantage of accounting for spatial nonstationarity in the data while also handling non-linear relationships between features and the target variable. Georganos et al. (2021) found that proper neighbourhood size selection and local model weight can lead to a model with improved performance over regular RF.

For each sampling point we trained a local random forest model on the K nearest neighbours of each training sample, in addition to a global RF model (Fig. 2). Predictions were a weighted average of the global model's prediction and the prediction of the nearest local model to each input sample. We used $K = 100$ neighbours, following recommendations in Georganos et al. (2021).

2.2.3. Model validation

Each model was evaluated using 5-fold cross-validation with a 80-20 train-test split ratio. We measured the models' performance with three statistical metrics: the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE). We found the hyperparameters shown in Table A.6 using the grid-search algorithm. These tuned values where found to be optimal for all models in this study. We also calculated residuals for all the models. Residuals of predictions on known data points can reveal if spatial autocorrelation affects a model's predictions; a spatial trend in the residuals indicates the model is systematically better in some areas, implying an unmodelled spatial relationship (Wadoux et al., 2020a; Kaveh et al., 2023). Kim (2021) found that the residual spatial autocorrelation of a model's predictions correlates with the spatial autocorrelation of the input and target features, suggesting that incorporating spatial aspects into a model can improve performance by reducing residual autocorrelation.

2.2.4. Feature importance and interpretability

An advantage of random forest models is that they are explainable AI (XAI) and there are several methods that can me used to see what

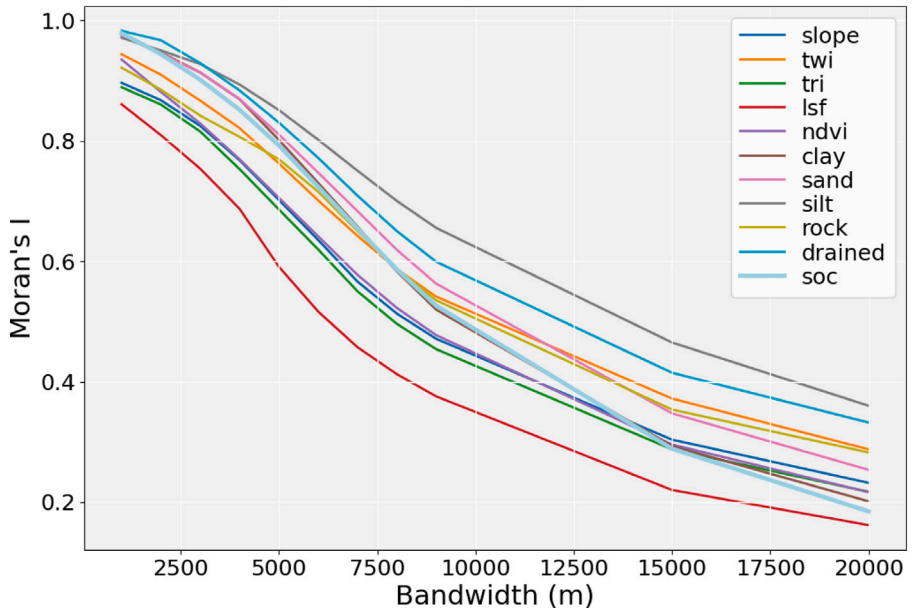


Fig. 3. Moran's I vs. kernel bandwidth by distance in meters for all continuous model features. Kernel bandwith means how far points are away from each other when Moran's I is calculated. In general, the further away points are from each other, the lower the similarity between the points which can be also seen on this figure.

are the main contributors to the model and how are they related to each other. We used Shapley Additive exPlanations (SHAP values) which are a recent advance in the interpretability of ML models. SHAP values quantify the contribution of each feature to the model's prediction for any given point in a game-theoretic manner. To calculate SHAP values, one must have a background dataset with F number of features (usually the training dataset, or a subset of it), and a dataset consisting of one or more samples to be explained. From the background data, a tree of 2^F models will be trained, one for each possible combination of features, beginning from the empty set. Then, for each sample, a prediction from every model can be made, and the marginal contribution of each feature to the model's prediction can be calculated over the entire tree (Lundberg and Lee, 2017). This marginal contribution of a feature to a single prediction is its SHAP value for that prediction.

We used beeswarm plots to visualize all the SHAP values for each prediction, which were grouped by the models' features and coloured according to the value of the feature for each prediction.

2.2.5. Estimating the autocorrelation

To estimate the spatial autocorrelation of each input feature, we calculated the Global Moran's I for several spatial weights matrices, each based on a kernel density function (KDF) with a different bandwidth based on regular distance intervals from each observation (Fig. 3). Plotting Moran's I over these distances allows us to visualize the scale of each input feature's spatial autocorrelation.

Values of Moran's I between 0.6 and 1 indicate strong spatial autocorrelation within distances of 5000 to 10,000 meters from sample points for different features (Fig. 3). The feature with the strongest spatial autocorrelation across most bandwidths was silt. Soil organic carbon percentage was one of the more strongly autocorrelated features at shorter distances (5000 m and below) but quickly dropped below other features at longer distances.

2.2.6. Computational performance

Despite increased computational resources, the growing complexity of models poses challenges for spatial modelling. We assessed computational time for model predictions. Training spatial models are typically fast, with only hundreds or thousands of points. However, prediction tasks are computationally intensive, often involving millions of points/pixels. Therefore, it is relevant to know the cost-effectiveness of adding spatial aspects to the models and evaluate their computational time.

Table 2
Model evaluation metrics (5-fold cross-validation).

Model	R^2	RMSE	MAE
Baseline vector	0.61	7.5	4.46
Baseline raster	0.6	7.5	4.39
XY	0.61	7.4	4.23
RFSI	0.63	7.29	4.31
BD	0.62	7.37	4.27
RFRK	0.61	7.47	4.39
GWRFR	0.6	7.49	4.57

For all data processing, modelling, and analysis, we used Python 3.9, with the following packages:

- numpy: for numerical operations (Harris et al., 2020).
- pandas: for data manipulation (Reback et al., 2021).
- geopandas: for spatial data manipulation (Jordahl et al., 2020).
- scikit-learn: for machine learning and data analysis (Pedregosa et al., 2011). Hunter (2007).
- seaborn: for plotting (Waskom, 2021).
- shap: for SHAP value analysis (Lundberg and Lee, 2017) .
- pykrige: for kriging (Murphy et al., 2022).

For mapmaking, we used QGIS 3.28 (QGIS Development Team, 2023).

3. Results

3.1. General model performance

The results from the cross-validation show minor improvements from the raster baseline random forest model for methods using spatial covariates (XY, RFSI, BD), while the models that incorporate spatial data in an extra step (GWRFR, RFRK) make only negligible improvements at best (Table 2). The best-performing model was RFSI, which showed a 0.02 improvement in R^2 and a 0.21 decrease in RMSE over baseline.

Each model was used to predict SOC for the whole of Estonia at 10 m resolution (Fig. 4), except GWRFR, which did not complete

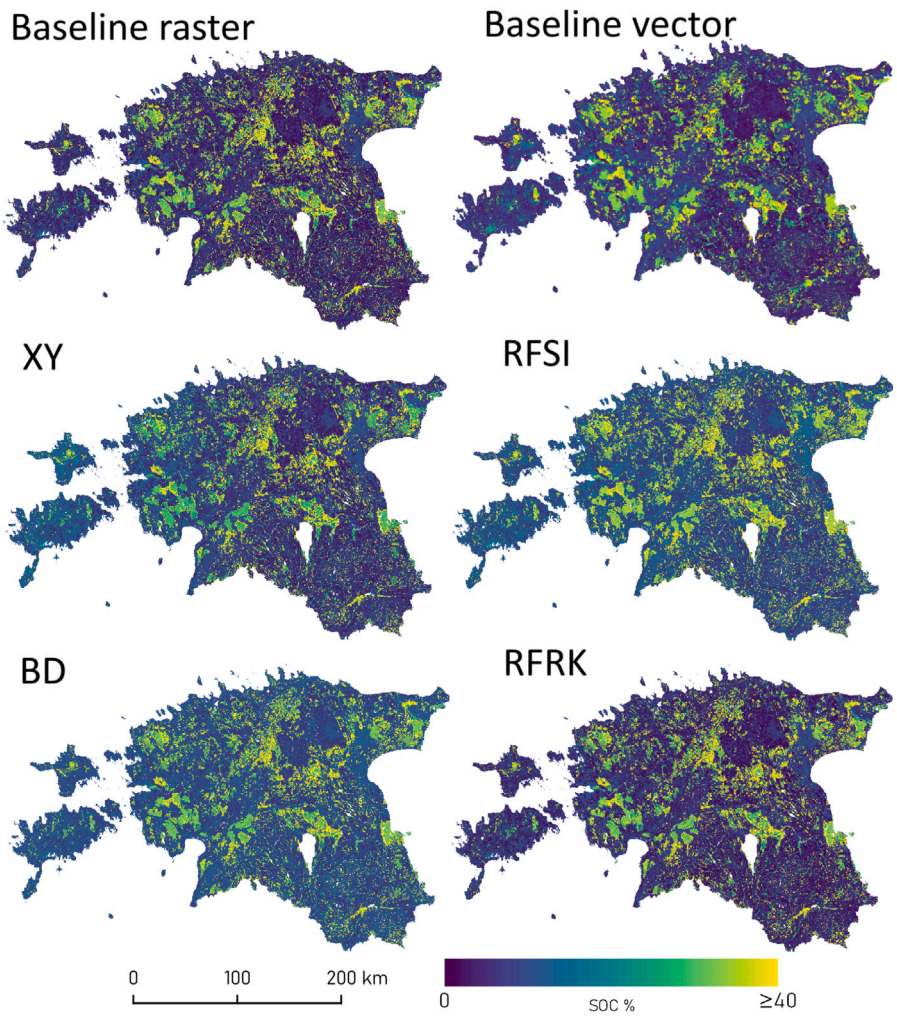


Fig. 4. Comparison of predictions by spatial and non-spatial machine learning methods across Estonia. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

successfully. A visual inspection does not enable any clear assessment, as the modelling results are pretty similar, except that RFSI and BD seem to have higher values for agricultural regions and forested areas than the baseline model (showing dark blue rather than lighter blue). This can be confirmed by consulting the histograms of the models' prediction distributions in Fig. 5. In the BD and RFSI histograms, the lower parts of the distributions have shifted towards higher values, and there are higher peaks between 5% to 20% SOC, which do not exist in the baseline models nor other spatial models (XY and RFRK). BD and RFSI also have slightly more values above 20% SOC than baseline models.

The baseline raster and RFRK models have the most similar distributions to the observed data (though they show a third peak above 40% SOC). The other models, however, show jagged, sharp peaks and troughs (Fig. 5).

However, predictions made on only validation data do not show any of these irregularities (Fig. 6). Any differences between model distributions are minimal.

3.2. Residuals

Plotting of prediction residuals on the testing dataset does not reveal any notable spatial pattern (Fig. 7). The models appear to be consistent in which samples are over/underestimated, both in direction and degree.

After calculating Moran's I for several lag distances, we can see that the autocorrelation of the baseline raster model's prediction residuals is strong at distances below 10,000 m, then quickly drops (Fig. 8). In other words, there is a spatial pattern to the model residuals. This indicates that there indeed may be some spatial autocorrelation in the physical phenomenon that the baseline raster model has not captured. However, this autocorrelation is only marginally reduced in the spatial models' residuals.

3.3. Feature importance

To better understand how each covariate (incl. spatial covariates) contributes to the models, we calculated SHAP values for the baseline raster model and the models with spatial information as covariates (XY, RFSI, BD). The other models (GWRFR, RFRK) were left out of this calculation because they do not have any explicit spatial covariates for which to calculate importance.

According to SHAP values, the most important feature for all models is clay content. The higher clay values result in significantly higher SOC values up to over 40% clay content (Fig. 9). This can be explained by the fact that clay was used as a stand-in for peat in the EstSoil dataset for hydrological modelling purposes (Kmoch et al., 2021). The order of the rest of the feature importances differed slightly across different models but among the top five features were always the land use type *arable land* and slope. Arable land use is associated with lower values of SOC, while forest and grassland land use lead the model to predict

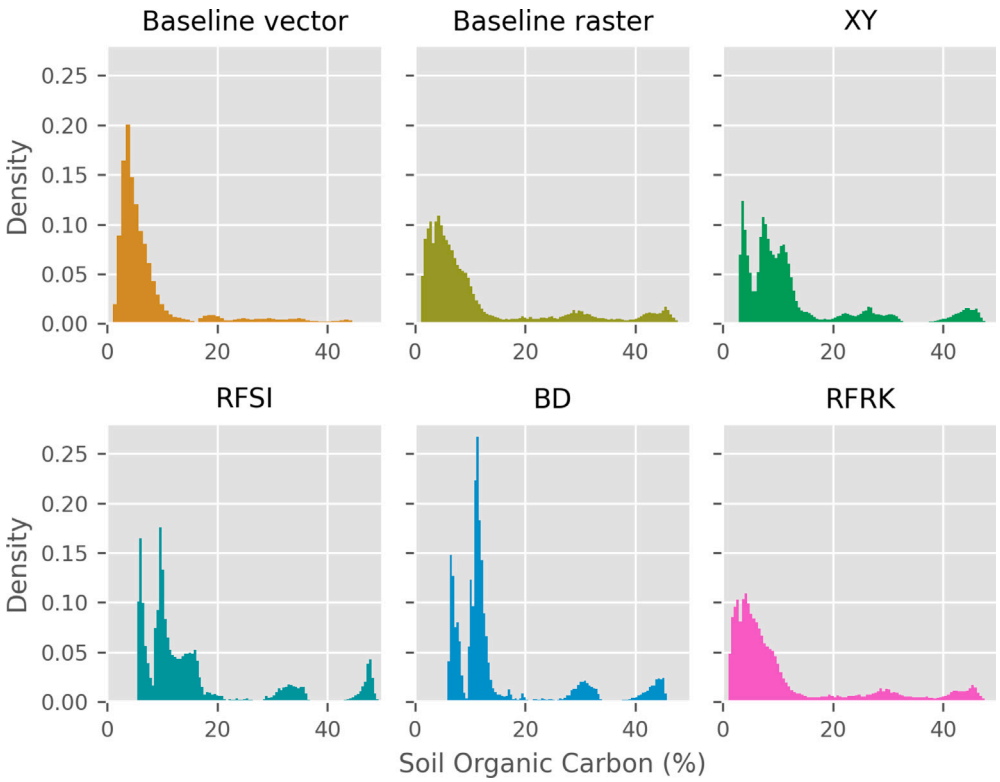


Fig. 5. Distribution of SOC predictions across all of Estonia for each model.

higher SOC. Higher slope values lead to lower SOC values. In all the spatial models (XY, RFSI, BD), the spatial component also made it to the top 5 features. For the XY model, the X coordinate is the second most important covariate, and it is negatively correlated with SOC content. The Y coordinate is not particularly important to the model, and the SHAP plot does not show any clear directional relationship between the Y coordinate value and SOC. In the RFSI model, the value of the nearest neighbour (neighbor_1_val) is the fifth-most impactful covariate and positively correlates with SOC.

Plotting SHAP values for each of the BD model's spatial features does not provide any insight, as there are hundreds of extra covariates to consider, and, if buffer distances are important to the model, each observation location will have a different set of covariates providing the most important information. However, because SHAP values are additive, they can be summed across all the buffer covariates to determine the contribution of the features as a group to the model's predictions. In this case, colour becomes a meaningless quality of the plot, because it just represents the sum of the buffer distance column values as intensity, but not the direction of correlation. However, the importance of buffer distances in general can be identified which in this case is the second most important feature.

To better understand and explore the details of the spatial variation in the prediction, we selected four sites from different regions of Estonia with different land use:

- Nõva, northwest Estonia (mainly forested with wetlands and flat)
- Rakvere, northeast Estonia (mainly arable and flat)
- Lavassaare, southwest Estonia (mainly wetland and flat)
- Valgjärve, shoutheast Estonia (mixed land use and hilly)

Comparisons between the baseline vector model, baseline raster model, and all spatial models for Nõva site are shown in Fig. 10 and the most influential model variables for Nõva site are shown in Fig. 11. Rakvere, Lavassaare and Valgjärve are shown in appendices A.12–A.15. All detailed study areas show greater spatial variation in SOC values in the raster models (including spatial models) compared to the baseline vector model. The higher level of spatial variation in raster models is likely due to the topographic parameters (TWI, TRI) that capture the spatial variation of elevation in more detail. However, the most important feature determining the dominant spatial pattern is clearly clay content that originates from the vector map.

3.4. Computation time

Model training times on a consumer notebook PC were negligible (on the order of several seconds) for all models except BD and GWRFR. BD adds hundreds of extra covariates to the model and takes approximately 40 s to train. GWRFR trains 100 sub-models. This takes about 60 s. Prediction times, however, are much greater, as predictions must be made for approximately 500 million raster pixels for all of Estonia.

After the vector baseline model (approx 800,000 polygons), the baseline raster model is the fastest, taking 1 h and 21 min to predict (Table 3). XY, RFSI, and RFRK are similar to the baseline raster, running 20-30 min longer due to extra computational steps, and in the case of XY and RFSI, extra model features. BD, containing hundreds of additional features, takes significantly longer to make predictions, at 4 h and 13 min. GWRFR was not calculated across the entire study area, as it was estimated to take up to 300 h to predict on the available hardware. However, given enough resources, the task could potentially be parallelized to make prediction times manageable.

4. Discussion

We implemented five spatially-aware random forest models to enhance SOC predictions and account for spatial autocorrelation. Our results demonstrated that incorporating spatial covariates into predictive models of soil organic carbon resulted in only minor improvements over the baseline random forest model, with the RFSI model performing best, achieving a 0.02 increase in R^2 and a 0.21 decrease in RMSE. SHAP values highlighted clay content as the most important factor

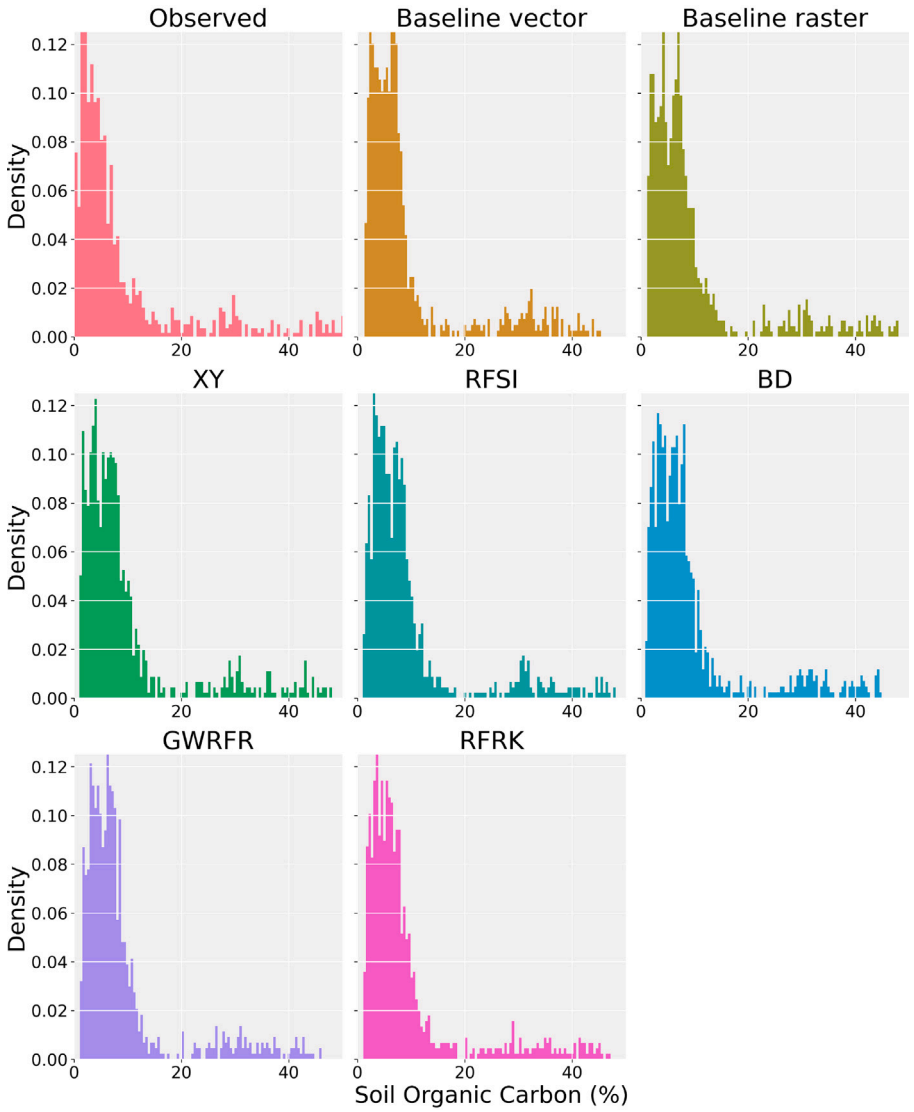


Fig. 6. Distribution of SOC predictions for each model on observed/testing data.

Table 3
Time for each model to make predictions across the entire study area, and associated hardware.

Model	Time	CPU count	RAM (GB)
Baseline vector	22 s	8	64
Baseline raster	1 h 21 m	16	168
XY	1 h 45 m	16	148
RFSI	1 h 43 m	16	160
BD	4 h 13 m	16	192
RFRK	1 h 51 m	16	260

influencing SOC, followed by land use types and slope. Additionally, site-specific analyses demonstrated that raster models captured greater spatial variation in SOC than the baseline vector model, largely due to topographic factors. While the baseline raster model was the fastest to predict, BD and GWRFR models required significantly more computation time, underscoring the balance between model complexity and practicality for large-area studies.

4.1. Challenges in adding spatial structure

RFSI emerged as the top performer in capturing spatial structure and improving model accuracy on validation data, followed by the BD and XY models. However, the RFSI and BD models' usefulness beyond validation data is questionable, as they appear to overfit when applied on a broader scale, resulting in unnatural SOC prediction distributions characterized by higher double peaks (Fig. 5). Meyer et al. (2019) warn that highly autocorrelated spatial proxies can lead to significant overfitting in this context and create artefacts in the prediction. In our case, although there was a geographical overlap between the sampling and prediction area, the sampling points distribution may still have been too dispersed to prevent some extrapolation entirely. This could have been the reason for slightly too high values in wetlands and forests (double peaks).

The GWRFR model, despite being computationally intensive, does not significantly outperform other models and struggles with capturing local spatial effects, possibly due to its large bandwidth. Meanwhile, the RFRK model shows minimal improvement over the baseline, with its kriging surface generated from training data residuals producing predictions that closely resemble the baseline model, offering little added value.

Therefore, we can say that overall improvements that came with spatial machine learning strategies in this study are small and show only a small decrease in residual autocorrelation. Several previous studies have shown that adding a spatial aspect to random forest models improves cross-validation results (Hengl et al., 2018; Sekulić

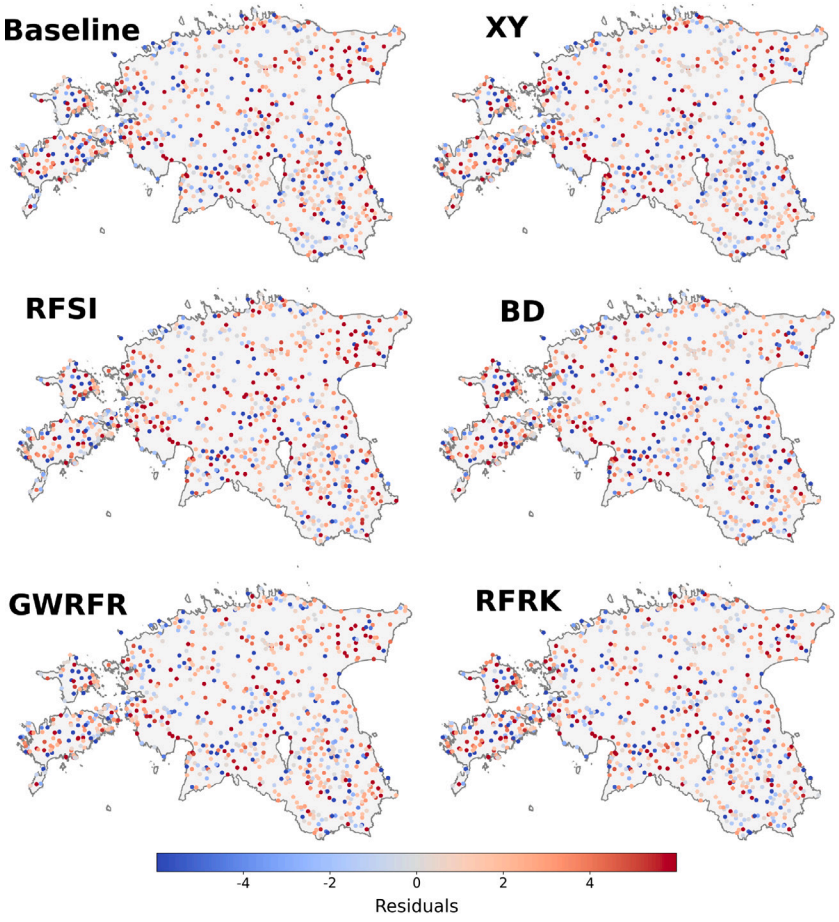


Fig. 7. Spatial distribution of residuals of predictions on testing data (all folds merged) for each model.

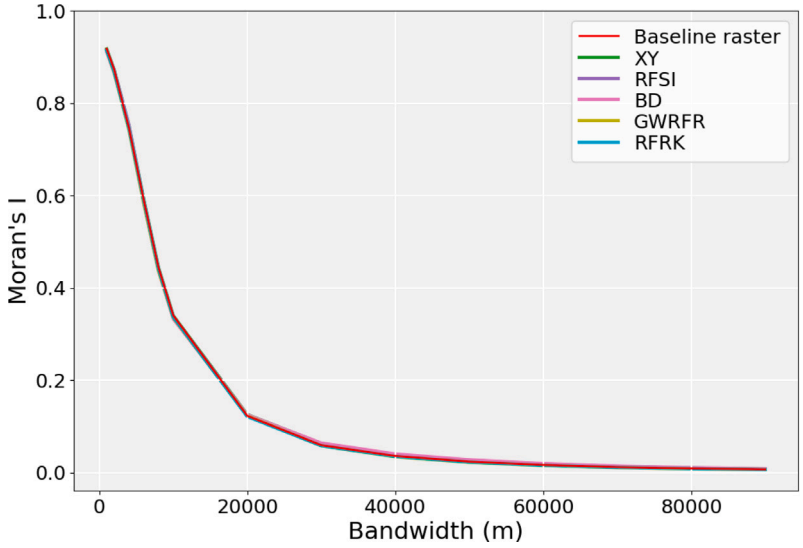


Fig. 8. Moran's I vs. distance plot for raster model residuals.

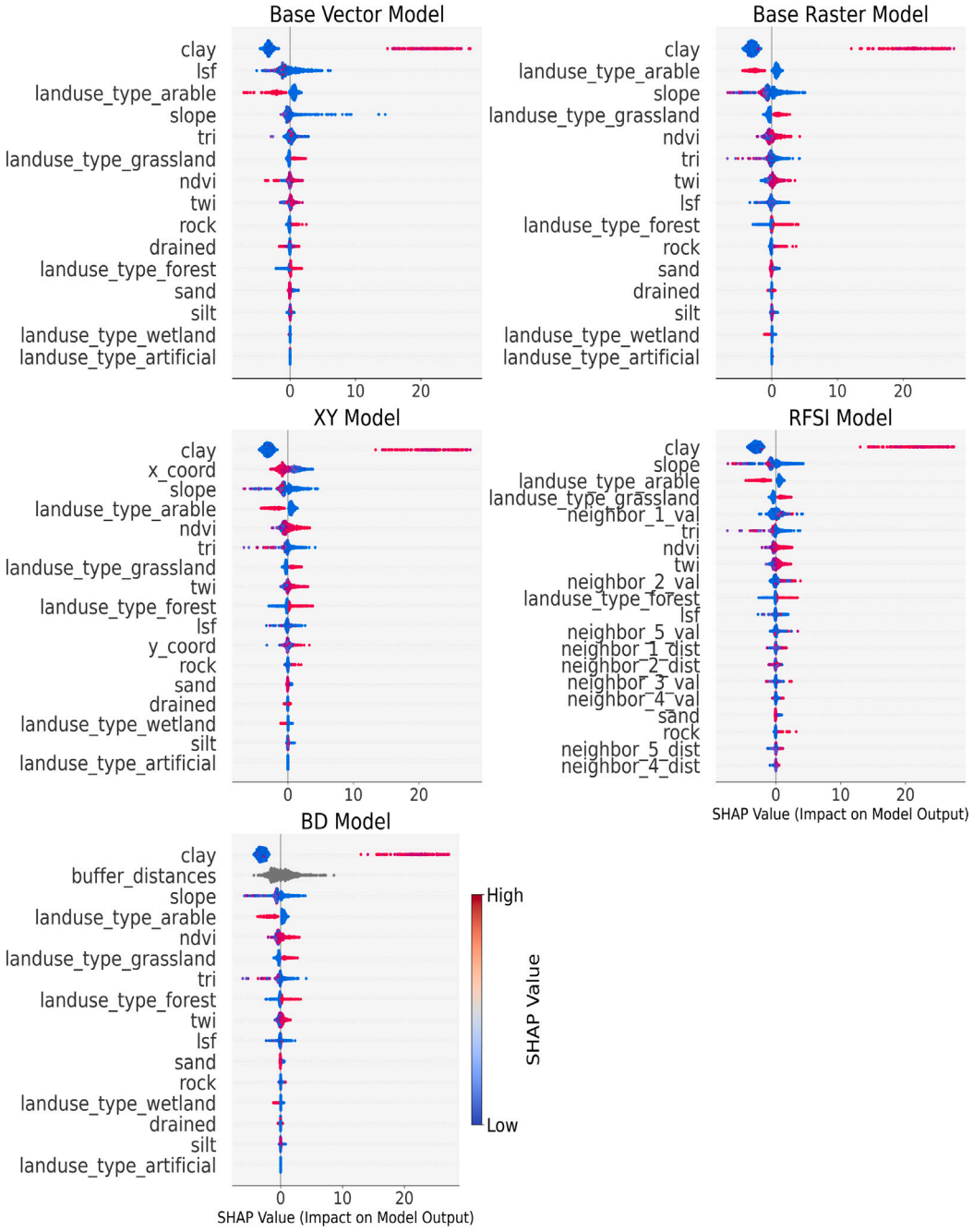


Fig. 9. SHAP values for all models.

et al., 2020; Georganos et al., 2021; Fox et al., 2020), and can reduce autocorrelation in prediction residuals (Beale et al., 2010; Kim, 2021) which is somewhat surprising that both effects are only minor here. However, Beale et al. (2010) observed that even well-fitted spatial models may have residual autocorrelation and it does not necessarily indicate a problem.

One potential explanation for the limited effect observed here is the spatial scale of the autocorrelation phenomenon. Although there is some autocorrelation among model inputs and SOC at short distances (less than 5 km), it generally involves only a few very near neighbours (Fig. 3). Georganos et al. (2021) mention that, for geographically weighted regression, the size of the neighbourhood used for training local models "can be described as the operational scale of the relationship which includes just enough data points to capture the inherent localities [of the modelled process] while at the same time rejecting/reducing unnecessary training data that come from locations afar, that might be considered as noise to the model". This makes sense in relation to the nature of machine learning, which generally requires a large number of samples to be effective. It follows, naturally, that if there are not enough data points to capture inherent localities while rejecting noise, then effective modelling will be hard. Moreover, Milà et al. (2024) emphasize that Random Forest models incorporating spatial proxies are not well-suited for extrapolating to new areas, especially when there is a lack of geographical overlap between the sampling and prediction areas.

Neighbourhood size is critical to the effectiveness of spatial machine learning and may be important beyond just its implications for geographically weighted random forests; however, it can also be applied more broadly to other types of spatial ML models. In this case, the scale of the autocorrelation in soil organic carbon content is smaller than

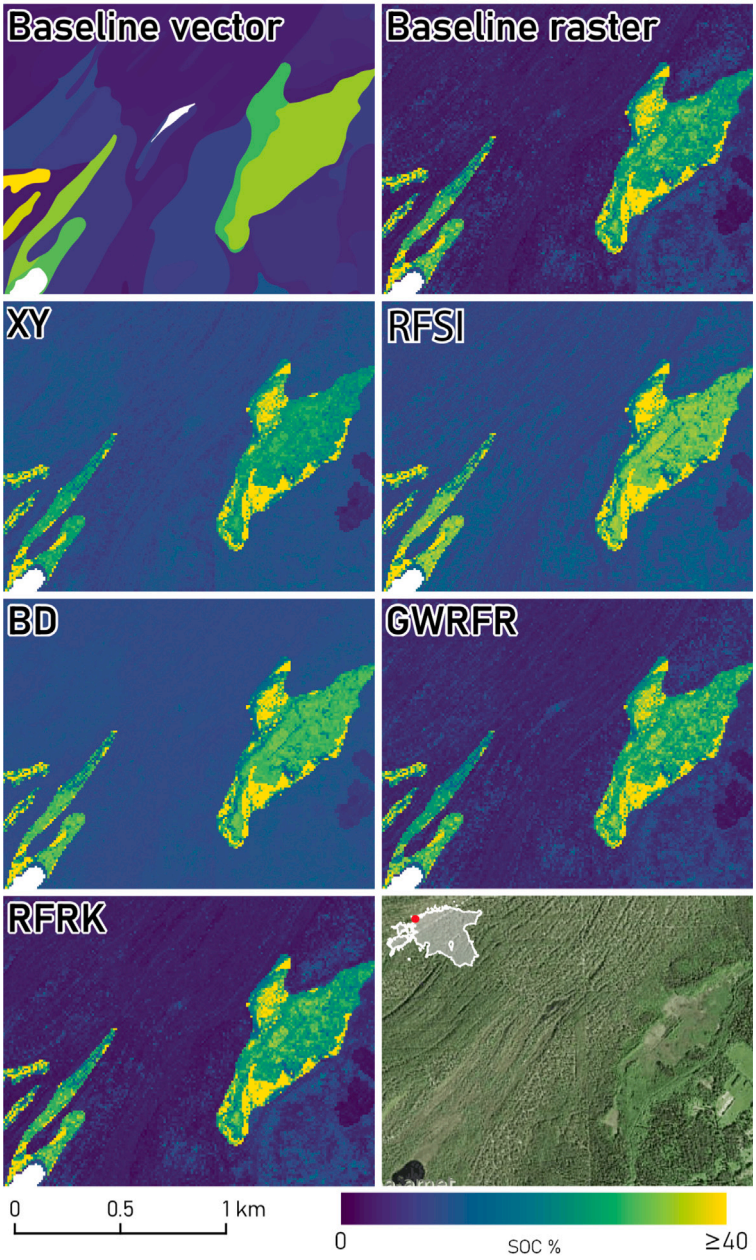


Fig. 10. Comparison of predictions by spatial and non-spatial machine learning methods (northwest Estonia, Nõva), with orthophoto (Estonian Land Board, 2022).

what can be captured given the density of the available samples. This would apply particularly to the buffer distances model (as only ~0.05% of the additional covariates are autocorrelated to any given sample) and geographically weighted regression (because there are not many autocorrelated samples to build local regression models from).

Another possibility is that the relatively small amount of spatial autocorrelation of SOC could have been largely captured in the baseline model by some combination of the other covariates. This idea is supported by the fact that in the spatial models, the spatial covariates showed some importance in SHAP value analysis above other useful environmental covariates. In contrast, the model performance did not improve significantly. This could imply that the non-spatial covariates had already captured some spatial autocorrelation, and when actual spatial covariates were introduced to the model, they depressed the influence of non-spatial covariates. On the other hand, this should also raise cautiousness because spatial covariates can act as pseudo-covariates which are meaningless and not related to soil-forming factors and processes Wadoux et al. (2020b). Further studies to confirm this idea might involve systematically removing (combinations of) environmental covariates from the spatial models and the baseline model to see if any are necessary to baseline model performance but not spatial model performance.

In this study, the scale of autocorrelation in SOC was smaller than what the available sample density could capture, which could explain the double-peak effects observed in the RFSI and BD models. These models attempt to exploit spatial autocorrelation by fitting random forest models to spatial structures alongside environmental covariates. However, when prediction points are too distant from training data points, the models may rely too heavily on spatial variables, leading to less accurate predictions. To further explore this issue, the concept of Areas of Applicability (AOA, Meyer and Pebesma (2021)) could be used, allowing for the comparison of prediction distributions within and outside the AOA. In addition, the availability of field observations for soils and properties remains a challenge in modelling (Safaee et al., 2024), but the performance of the final predictive model largely hinges

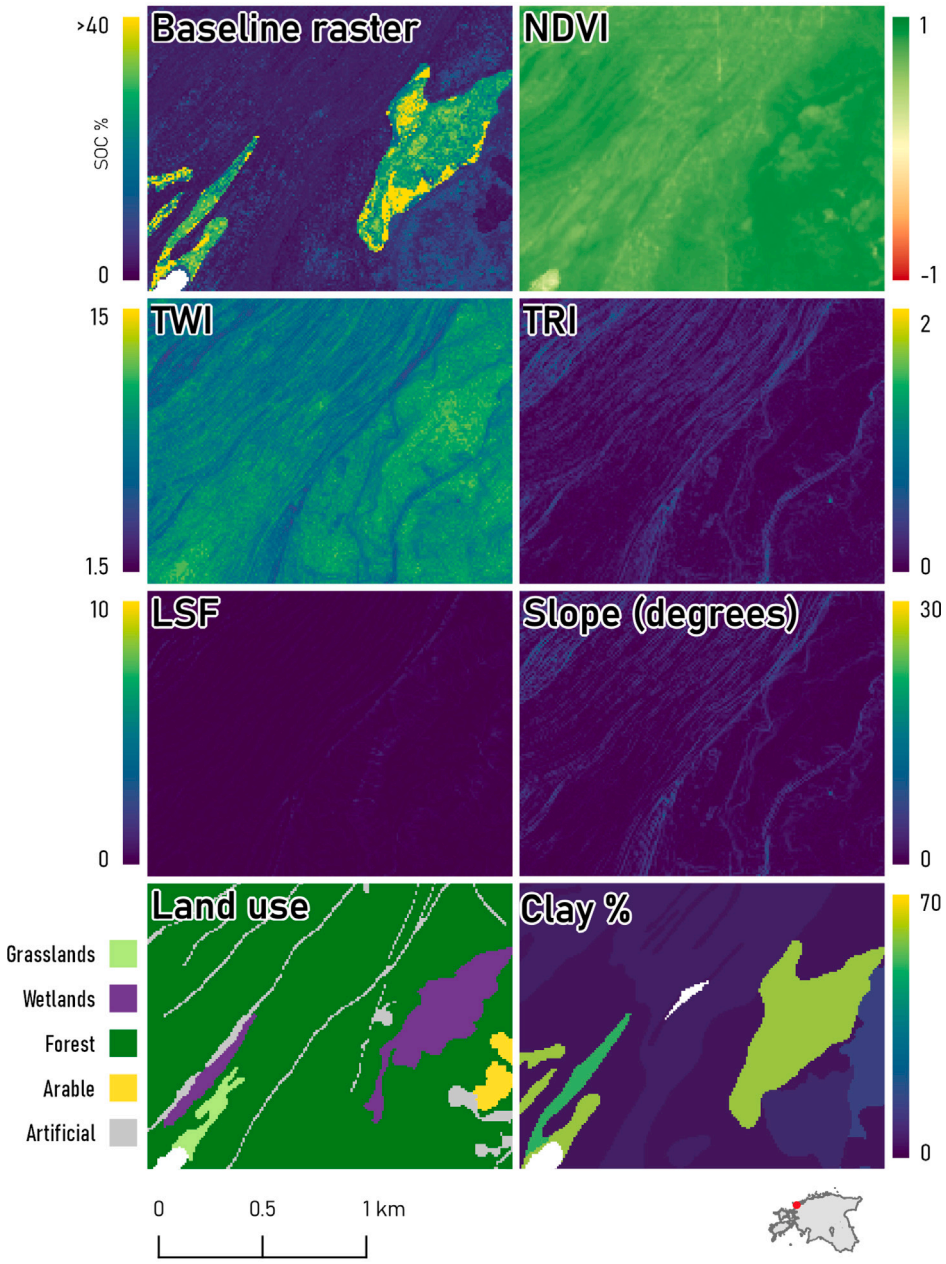


Fig. 11. Maps of baseline raster predictions and influential model variables (northwest Estonia, Nõva).

on the quality of the samples, specifically their size and representativeness, to accurately reflect real-world variability (Bouasria et al., 2023). Therefore, it is crucial to increase the density of SOC measurements, especially in areas experiencing high spatial variability, but at the same time, consider sampling strategy.

4.2. Computational considerations and validation challenges

The raster models showed much more detail than can be achieved by predicting the original soil units using vector model (Fig. 10). Predictions made on data aggregated to polygonal vector units are limited to the discrete shape and size of those units and it is not possible to account for smooth transition of environmental variables. The outlines of the original soil units are still easily discernible in many places because data for clay, sand, silt, and rock fractions have been rasterized from vector polygons and thus are represented as uniform within soil units. While the results of the raster-based models provide a higher level of spatial detail that is not available in vector-based models, there are some necessary considerations.

Firstly, in the current study, it was not possible to validate the detail of the output maps separately from the model itself, because there was no validation SOC data available in such spatial detail. In the vector-based model, environmental covariates aggregated over entire soil units are used to make predictions of what SOC is like in those 750,000 soil units generally. In the raster-based models, the values of these covariates are known at specific 10 m x 10 m locations, and predictions are made at hundreds of millions of them. For better validation, more SOC measurements are needed preferably on carefully selected study sites to capture the local SOC variation.

Secondly, the scale of computation necessary to produce raster predictions at this level of detail is considerable. Calculating predictions for the entire country required several hours of parallel computing time on a large cluster of CPUs. This kind of computation is becoming increasingly accessible and could be accelerated by switching to GPU-oriented computation, so it is not an insurmountable obstacle in general. However, computational complexity is still an important consideration when working with spatial data in machine learning, especially because data size increases quadratically with finer resolution.

Ensemble-based methods for incorporating spatial data into ML models, such as GWRFR, are particularly compute-hungry, and GWRFR did not show any clear advantage over other models during cross-validation. Using it to generate a map of all of Estonia would have required an estimated 300 h of computing time on available resources, so this step of analysis was skipped. XY and RFSI showed good performance, and they were computationally significantly less demanding.

5. Conclusions

In this study, we investigated whether SOC predictions could be improved by incorporating spatial autocorrelation into random forest models. Our findings indicate that spatial autocorrelation minimally enhances prediction accuracy. Among the five models tested, the RFSI method emerged as the most effective, which suggests that leveraging the value and proximity of neighbouring observations can effectively capture spatial dependencies, offering a robust approach for enhancing predictive performance in SOC modelling. We also examined whether a continuous raster approach delivers more accurate spatial distributions of SOC predictions compared to traditional vector methods. The results demonstrated that raster-based models provided finer spatial detail and greater variability in SOC predictions across Estonia, as they were able to utilize covariate values at precise 10 m x 10 m locations. This enhanced spatial resolution allows for a more nuanced understanding of SOC distribution, which is less feasible with aggregated vector-based predictions. However, the limitation in validation data at such spatial scales underscores the need for more detailed SOC measurements to confirm these results.

This study not only contributes to the specific field of SOC modelling but also provides valuable insights about spatial aspects in modelling that can improve environmental mapping practices in general. Most of the environmental variables exhibit more or less spatial autocorrelation; therefore, it is relevant to consider this when making spatial predictions of these variables. Our study increases the awareness of spatial dependencies in the environmental data which can increase the prediction accuracy but there is also a danger of overfitting. This also underscores the importance of high-resolution validation data which in turn highlights a broader necessity in environmental science to invest in comprehensive data collection efforts to reinforce the validity of modelling, especially machine learning approaches.

In summary, this study supports the integration of spatial autocorrelation into random forest models as a viable means of improving SOC predictions. However, the validity of the detailed spatial predictions remains uncertain due to insufficient validation data. This underscores the need for ongoing refinement of modelling techniques and collecting high-resolution validation data to advance geospatial SOC analysis. Future research should focus on evaluating these spatially aware models using high-density validation data from field observations coupled with strategic spatial sampling. Additionally, comparative analyses of model predictive performance within and outside specific Areas of Applicability (AOAs) will provide valuable insights into their reliable application contexts.

CRediT authorship contribution statement

Alexander Kmoch: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Clay Taylor Harrison:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Jeonghwan Choi:** Writing – review & editing, Visualization, Software, Investigation, Formal analysis, Data curation. **Evelyn Uuemaa:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Acknowledgements

This work was funded by the Estonian Research Agency (grant number PRG1764, PSG841), Estonian Ministry of Education and Research (Centre of Excellence for Sustainable Land Use (TK232)), Estonian Environment Agency (grant MULD2) and by the European Union (ERC, WaterSmartLand, 101125476). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The authors are thankful for the technical support from the High Performance Computing Center of the University of Tartu. The authors also like to extend their gratitude to Dr. Alexandre Wadoux for his insightful feedback on critical aspects that greatly contributed to improving this manuscript.

Appendix

See Figs. A.12–A.17 and Tables A.4–A.6.

Table A.4
Land use classes, and distribution of SOC samples across the land use classes.

Land use type	Numerical code	Num. of SOC samples
grasslands	1	253
wetlands	2	36
forests	3	415
arable	4	214
artificial	5	51

Table A.5
Descriptive statistics for all variables.

Feature name	Mean	Std. Dev.	Min.	Median	Max.
soc	9.214	12.012	0	4.633	60
slope	1.573	2.528	0.017	0.804	26.49
twi	9.489	1.639	3.621	9.791	13.307
tri	0.119	0.169	0.003	0.066	1.806
lsf	0.241	0.683	0	0.084	9.76
ndvi	0.777	0.124	0.011	0.81	0.936
clay	17.553	18.342	0	15	70
sand	67.489	23.737	15	65	100
silt	14.959	9.778	0	15	50
rock	7.757	15.699	0	0	85
drained	0.258	0.438	0	0	1

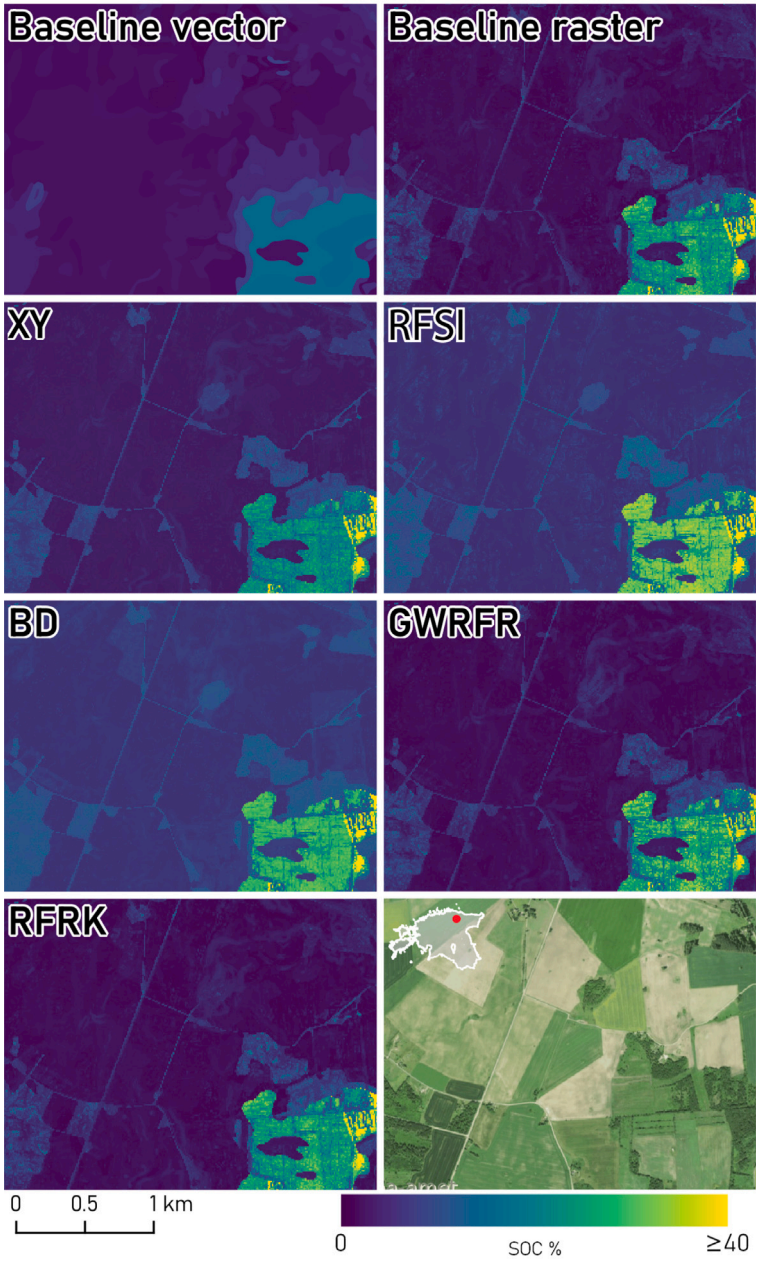


Fig. A.12. Comparison of predictions by spatial and non-spatial machine learning methods (detail, northeast Estonia, Rakvere), with orthophoto (Estonian Land Board, 2023b).

Table A.6
Hyperparameters for all models in this study.

Hyperparameter	Value
n_estimators	766
max_features	1.0
max_depth	20
min_samples_split	2
min_samples_leaf	4
bootstrap	True

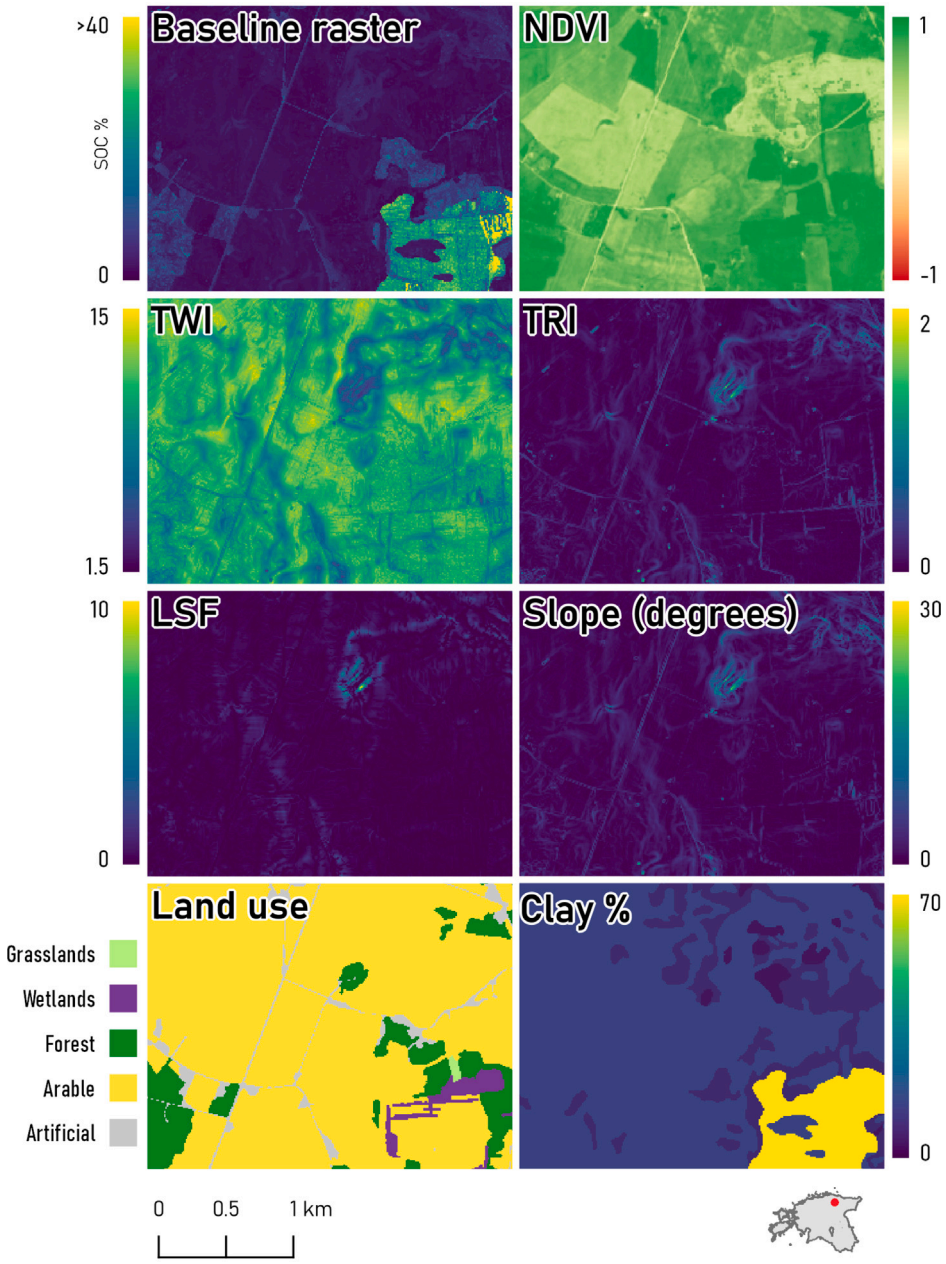


Fig. A.13. Maps of baseline raster prediction and influential model variables (detail, northeast Estonia, Rakvere).

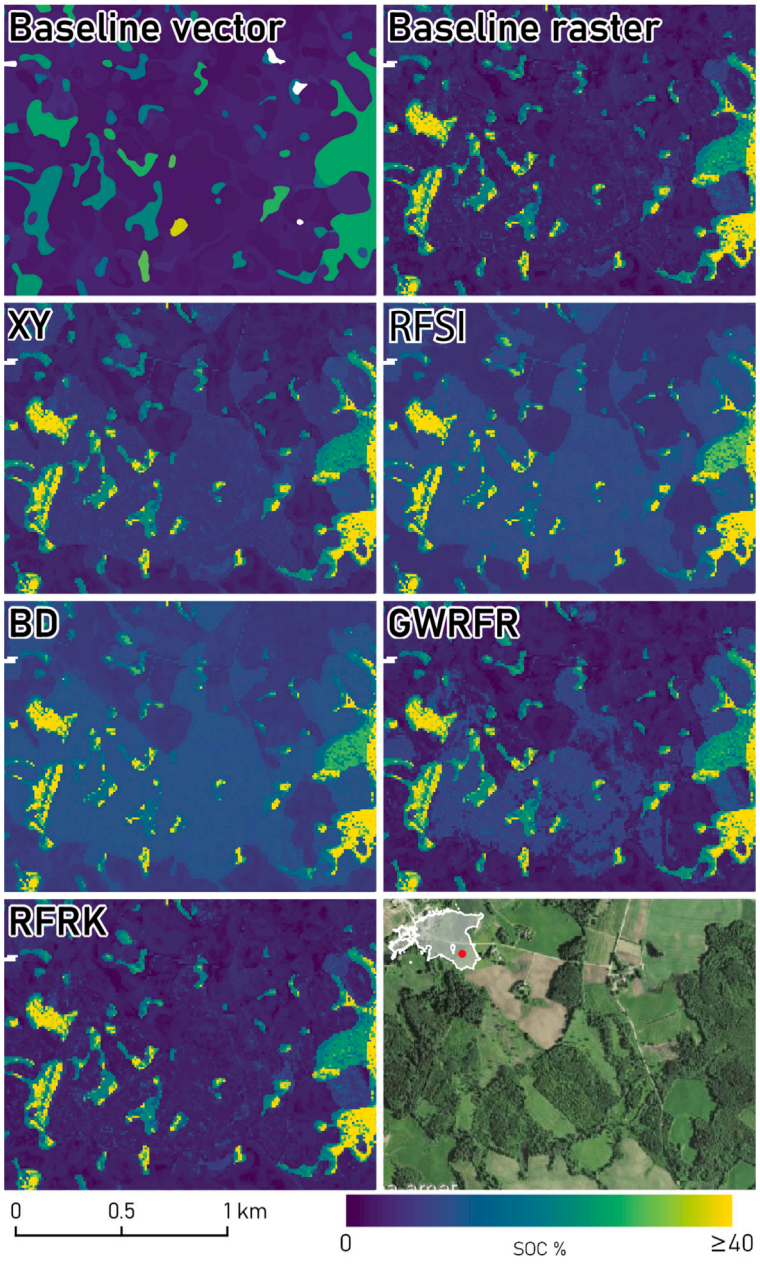


Fig. A.14. Comparison of predictions by spatial and non-spatial machine learning methods (detail, southeast Estonia, Valgjärve), with orthophoto (Estonian Land Board, 2023b).

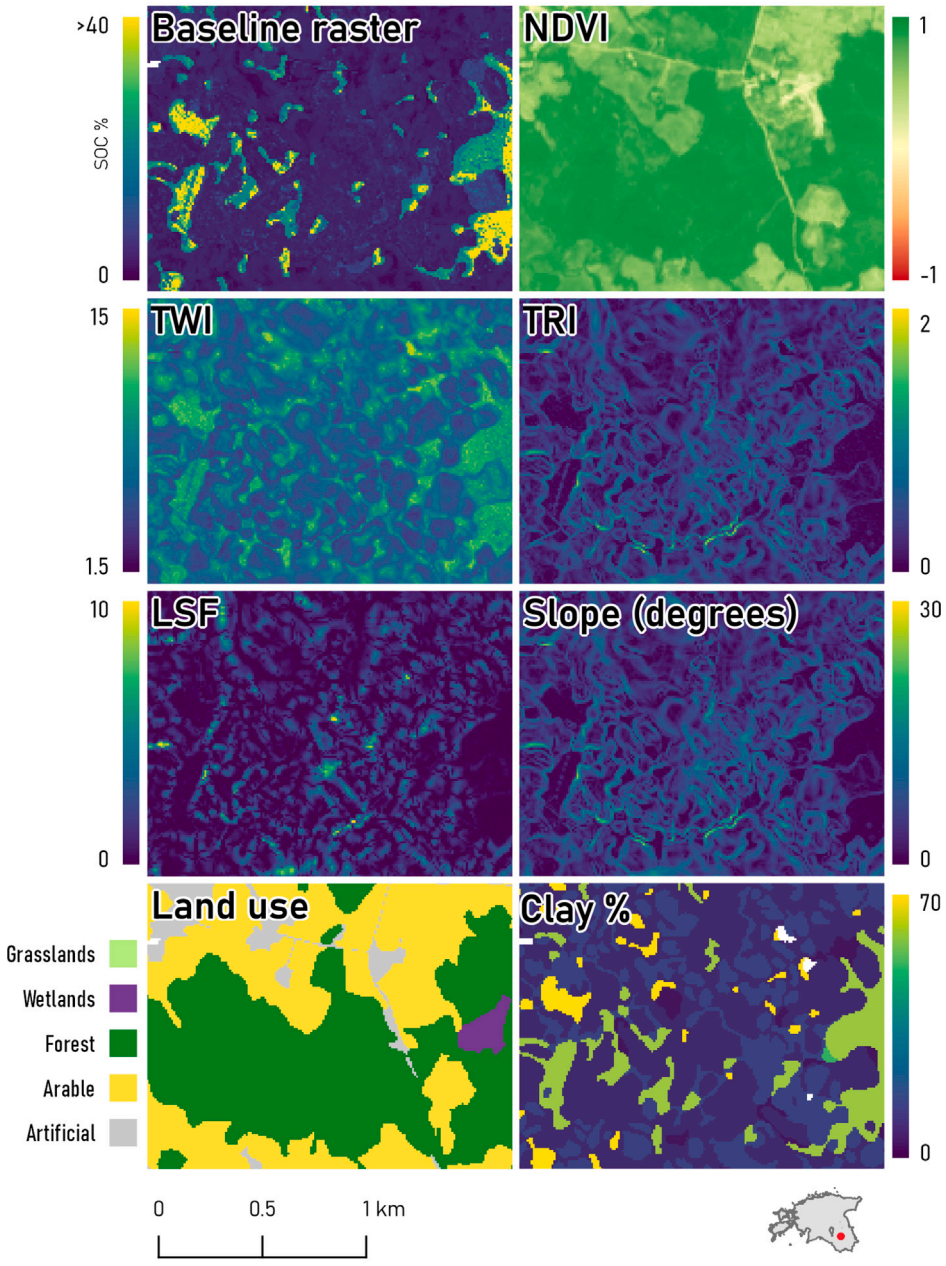


Fig. A.15. Maps of baseline raster prediction and influential model variables (detail, southeast Estonia, Valgjärve).

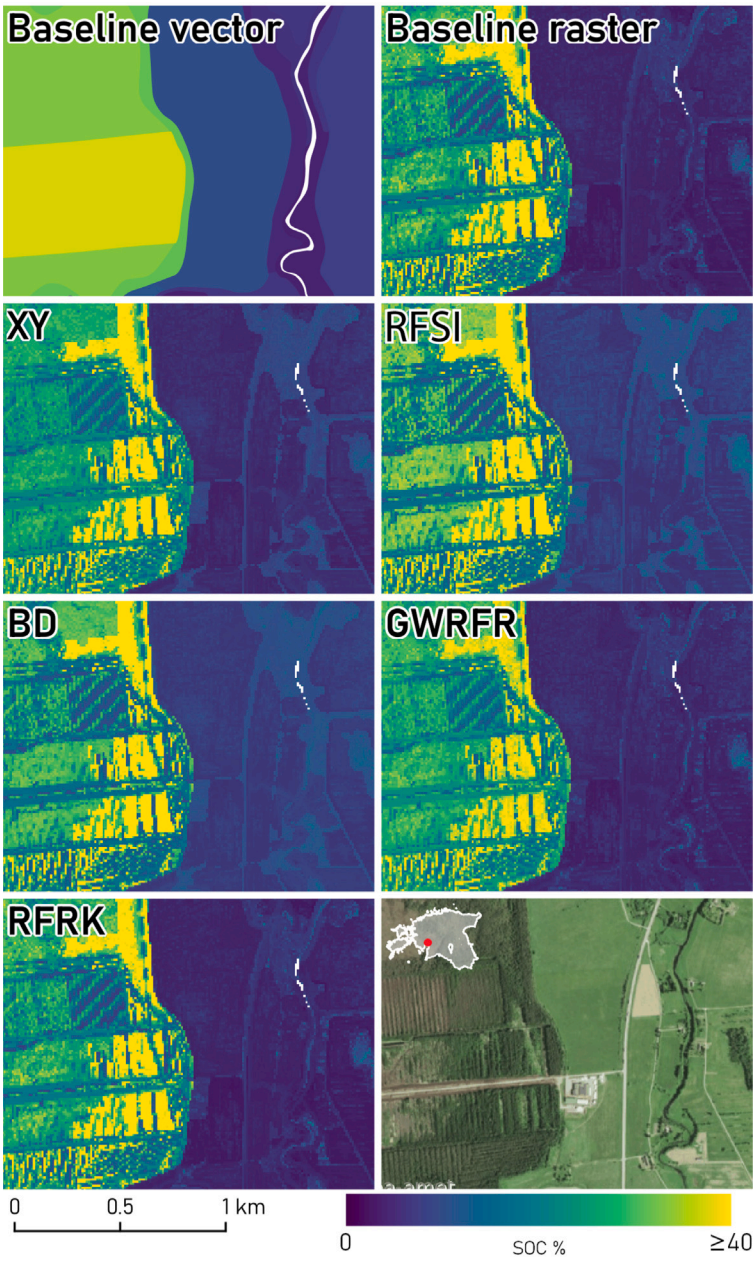


Fig. A.16. Comparison of predictions by spatial and non-spatial machine learning methods (detail, southwest Estonia, Lavassaare), with orthophoto (Estonian Land Board, 2023b).

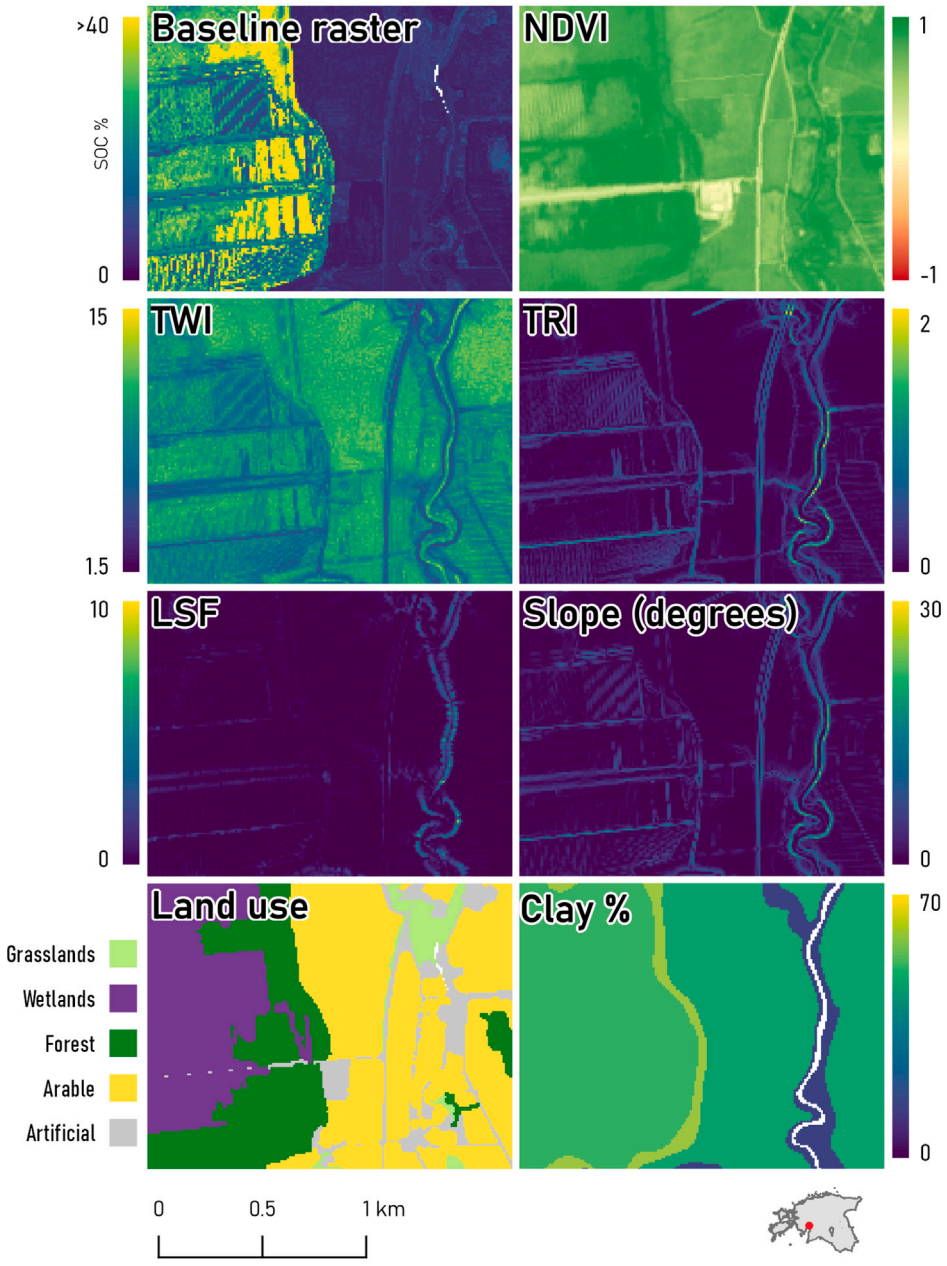


Fig. A.17. Maps of baseline raster prediction and influential model variables (detail, southwest Estonia, Lavassaare).

Data and code availability statement

The data (Uuemaa and Kmoch, 2024) and scripts/codes (Kmoch et al., 2024) that support the findings of this study are openly available on Zenodo. Data DOI: 10.5281/zenodo.14236579, code DOI: 10.5281/zenodo.14236923 with an associated GitHub repository).

References

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13 (2), 246–264. http://dx.doi.org/10.1111/j.1461-0248.2009.01422.x, Publisher: Wiley.

Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., MacMillan, R.A., 2018. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci. 69 (5), 757–770. http://dx.doi.org/10.1111/ejss.12687.

Bouasria, A., Bouslihim, Y., Gupta, S., Taghizadeh-Mehrjardi, R., Hengl, T., 2023. Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. Ecol. Inform. 78, 102294. http://dx.doi.org/10.1016/j.ecoinf.2023.102294.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83. http://dx.doi.org/10.1016/j.geoderma.2014.09.019.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 2010. Geographically weighted regression: A method for exploring spatial nonstationarity. Geogr. Anal. 28 (4), 281–298. http://dx.doi.org/10.1111/j.1538-4632.1996.tb00936.x, Publisher: Wiley.

Dormann, C., M. McPherson, J., B. Araujo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30 (5), 609–628. http://dx.doi.org/10.1111/j.2007.0906-7590.05171.x.

dos Santos, E.P., Moreira, M.C., Fernandes-Filho, E.I., Demattê, J.A.M., dos Santos, U.J., da Silva, D.D., Cruz, R.R.P., Moura-Bueno, J.M., Santos, I.C., de Sá Barreto Sampaio, E.V., 2023. Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. Ecol. Inform. 77, 102240. http://dx.doi.org/10.1016/j.ecoinf.2023.102240.

Duan, M., Song, X., Li, Z., Zhang, X., Ding, X., Cui, D., 2024. Identifying soil groups and selecting a high-accuracy classification method based on multi-textural features with optimal window sizes using remote sensing images. Ecol. Inform. 81, 102563. http://dx.doi.org/10.1016/j.ecoinf.2024.102563.

Estonian Land Board, 2017. Soilmap of Estonia - Mullastiku kaart. Estonian Land Board, http://dx.doi.org/10.15155/RE-72.

Estonian Land Board, 2022. Estonian elevation data. URL: https://geoportaal.maaamet.ee/eng/Maps-and-Data/Elevation-data/Download-Elevation-Data-p664.html.

Estonian Land Board, 2023a. Estonian topographic database. URL: https://geoportaal.maaamet.ee/eng/spatial-data/estonian-topographic-database-p305.html.

Estonian Land Board, 2023b. Estonian orthophotos. URL: https://geoportaal.maaamet.ee/eng/spatial-data/orthophotos-p309.html.

European Commission Joint Research Centre, 2022. LUCAS 2018 Soil Module: Presentation of Dataset and Results. Publications Office, http://dx.doi.org/10.2760/215013.

Fox, E.W., Ver Hoef, J.M., Olsen, A.R., 2020. Comparing spatial regression to random forests for large environmental data sets. Plos One 15 (3), http://dx.doi.org/10.1371/journal.pone.0229509.

GDAL/OGR contributors, 2024. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation, http://dx.doi.org/10.5281/zenodo.5884351.

Georganos, S., Grippa, T., Gadiaga, A.N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 36 (2), 121–136. http://dx.doi.org/10.1080/10106049.2019.1595177, Publisher: Taylor & Francis.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2, Publisher: Springer Science and Business Media LLC.

Helm, A., 2023. Role of Grasslands in Mitigating Climate Change, 2022–2023. University of Tartu, Published: EIC Environmental program project.

Helm, A., Kull, A., Veromann, E., Remm, L., Villoslada, M., Kikas, T., Aosaar, J., Tullus, T., Prangel, E., Linder, M., Otsus, M., Külm, S., Sepp, K., 2020. Country-Wide Assessment and Mapping of the Economic Value of Ecosystem Services Provided by Estonian Terrestrial Ecosystems. ELME projekt, Estonian Envionmental Agency, URL: https://keskkonnaagentuur.ee/media/1482/download.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. Plos One 12 (2), http://dx.doi.org/10.1371/journal.pone.0169748.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6 (nil), e5518. http://dx.doi.org/10.7717/peerj.5518.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. http://dx.doi.org/10.1016/j.geoderma.2015.11.014.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95. http://dx.doi.org/10.1109/MCSE.2007.55, Publisher: IEEE COMPUTER SOC.

Jemeljanova, M., Kmoch, A., Uuemaa, E., 2024. Adapting machine learning for environmental spatial data - A review. Ecol. Inform. 81, 102634. http://dx.doi.org/10.1016/j.ecoinf.2024.102634.

Jordahl, K., Bossche, J.V.d., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A.G., Farmer, C., Hjelle, G.A., Snow, A.D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L.J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, Leblanc, F., 2020. geopandas/geopandas: v0.8.1. Zenodo, http://dx.doi.org/10.5281/zenodo.3946761.

Kaveh, N., Ebrahimi, A., Asadi, E., 2023. Comparative analysis of random forest, exploratory regression, and structural equation modeling for screening key environmental variables in evaluating rangeland above-ground biomass. Ecol. Inform. 77, 102251. http://dx.doi.org/10.1016/j.ecoinf.2023.102251.

Kim, D., 2021. Predicting the magnitude of residual spatial autocorrelation in geographical ecology. Ecography 44 (7), 1121–1130. http://dx.doi.org/10.1111/ecog.05403, Publisher: Wiley.

Kmoch, A., Harrison, C.T., Uuemaa, E., Choi, J., 2024. Code Supplement: Spatial Autocorrelation in Machine Learning for Modelling Soil Organic Carbon. Zenodo, http://dx.doi.org/10.5281/zenodo.14236923.

Kmoch, A., Kanal, A., Astover, A., Kull, A., Virro, H., Helm, A., Partel, M., Ostonen, I., Uuemaa, E., 2021. EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia. Earth Syst. Sci. Data 13 (1), 83–97. http://dx.doi.org/10.5194/essd-13-83-2021.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395–413. http://dx.doi.org/10.1016/j.geoderma.2019.05.031.

Lundberg, S.M., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30. NIPS '17, Curran Associates, Inc., Red Hook, NY, USA, pp. 4768–4777. http://dx.doi.org/10.5555/3295222.3295230.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. http://dx.doi.org/10.1111/2041-210X.13650.

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. http://dx.doi.org/10.1016/j.ecolmodel.2019.108815.

Milà, C., Ludwig, M., Pebesma, E., Tonne, C., Meyer, H., 2024. Random forests with spatial proxies for environmental modelling: opportunities and pitfalls. Geosci. Model. Dev. 17 (15), 6007–6033. http://dx.doi.org/10.5194/gmd-17-6007-2024.

Murphy, B., Yurchak, R., Müller, S., 2022. GeoStat-Framework/PyKrige: v1.7.0. Zenodo, http://dx.doi.org/10.5281/zenodo.7008206.

Nikparvar, B., Thill, J.C., 2021. Machine Learning of Spatial Data. ISPRS Int. J. Geo-Inf. 10 (9), http://dx.doi.org/10.3390/ijgi10090600.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), http://dx.doi.org/10.1038/s41467-020-18321-y, Publisher: Springer Science and Business Media LLC.

QGIS Development Team, 2023. QGIS Geographic Information System. QGIS Association, URL: https://www.qgis.org.

Reback, J., jbrockmendel, McKinney, W., Bossche, J.V.d., Augspurger, T., Cloud, P., Hawkins, S., gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., Hayden, A., Saxton, D., Gorelli, M.E., Shadrach, R., Jancauskas, V., McMaster, A., Li, F., Battiston, P., Seabold, S., Dong, K., chris-b1, 2021. pandas-dev/pandas: Pandas 1.2.5. Zenodo, http://dx.doi.org/10.5281/zenodo.5013202.

Safaee, S., Libohova, Z., Kladivko, E.J., Brown, A., Winzeler, E., Read, Q., Rahmani, S., Adhikari, K., 2024. Influence of sample size, model selection, and land use on prediction accuracy of soil properties. Geoderma Reg. 36, e00766. http://dx.doi.org/10.1016/j.geodrs.2024.e00766.

Sarkar, M.S., Majhi, B.K., Pathak, B., Biswas, T., Mahapatra, S., Kumar, D., Bhatt, I.D., Kuniyal, J.C., Nautiyal, S., 2024. Ensembling machine learning models to identify forest fire-susceptible zones in Northeast India. Ecol. Inform. 81, 102598. http://dx.doi.org/10.1016/j.ecoinf.2024.102598.

Sekulić, A., Kilibarda, M., Heuvelink, G.B., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote. Sens. 12 (10), http://dx.doi.org/10.3390/rs12101687.

Uuemaa, E., Kmoch, A., 2024. Data Supplement: Spatial Autocorrelation in Machine Learning for Modelling Soil Organic Carbon. Zenodo, http://dx.doi.org/10.5281/zenodo.14236579.

Wadoux, A.M.-C., Brus, D.J., Heuvelink, G.B., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913. http://dx.doi.org/10.1016/j.geoderma.2019.113913.

Wadoux, A.M.-C., Minasny, B., McBratney, A.B., 2020a. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Sci. Rev. 210, 103359. http://dx.doi.org/10.1016/j.earscirev.2020.103359, Publisher: Elsevier BV.

Wadoux, A.M.J.-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020b. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. 71 (2), 133–136. http://dx.doi.org/10.1111/ejss.12909.

Waskom, M.L., 2021. seaborn: statistical data visualization. J. Open Source Softw. 6 (60), 3021. http://dx.doi.org/10.21105/joss.03021, Publisher: The Open Journal.