**1949**

# Using Machine Learning Techniques to Solve Digital Soil Mapping Issues: Spatial Extrapolation and Joint Spatial Modelling of Soil Properties

Thesis for the Degree of Doctor of Philosophy (PhD)

**Fatemeh Hateffard**

Supervisors:

Novák Tibor József (Ph.D. Associate Professor)

Szatmári Gábor (Ph.D. Senior Research Fellow)

UNIVERSITY OF DEBRECEN

Doctoral Council of Natural Sciences and Information Technology

Doctoral School of Earth Sciences

Debrecen, 2023

*Hereby I declare that I prepared this thesis within the Doctoral Council of Natural Sciences and Information Technology, Doctoral School of Earth Sciences, University of Debrecen in order to obtain a PhD Degree in Natural Sciences at Debrecen University.*
*The results published in the thesis are not reported in any other PhD theses.*

*Debrecen, June 2023.* ……

*Signature of the candidate*

*Hereby I confirm that Fatemeh Hateffard candidate conducted her studies with my supervision within the Landscape Protection and Climate Doctoral Program of the Doctoral School of Earth Sciences, between 2019 and 2023. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.*

*I also declare that the results published in the thesis are not reported in any other theses.*
*I support the acceptance of the thesis.*
*Debrecen, June 2023* ……

*Signature of the supervisor*

*Hereby I confirm that Fatemeh Hateffard candidate conducted her studies with my supervision within the Landscape Protection and Climate Doctoral Program of the Doctoral School of Earth Sciences, between 2019 and 2023. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.*

*I also declare that the results published in the thesis are not reported in any other theses.*
*I support the acceptance of the thesis.*
*Debrecen, June 2023* ……

*Signature of the supervisor*

# Using machine learning techniques to solve digital soil mapping issues: spatial extrapolation and joint spatial modelling of soil properties

Dissertation submitted in partial fulfillment of the requirements for the doctoral (PhD) degree
In Earth Sciences (discipline name)

Written by **Fatemeh Hateffard** certified Soil Scientist

Prepared in the framework of the Earth Science Doctoral School of the University of Debrecen

Dissertation advisors: Dr. Novák Tibor József
Dr. Szatmári Gábor

The official opponents of the dissertation:

Dr.
Dr.

The evaluation committee:

Chairperson: Dr.
Members: Dr.
Dr.
Dr.
Dr.

The date of the dissertation defence: …… 2023…

# Table of Contents

## List of Abbreviation

| | |
|---|---|
| ANN | Artificial neural networks |
| AOA | Area of Applicability |
| BD | Bulk density |
| cLHS | Conditioned Latin hypercube sampling |
| DEM | Digital elevation model |
| DI | Dissimilarity index |
| DSM | Digital soil mapping |
| DT | Decision trees |
| EC | Electrical conductivity |
| EVI | Satellite based Enhanced Vegetation Index |
| LHS | Latin hypercube sampling |
| ML | Machine learning |
| MLR | Multiple linear regression |
| MRRTF | Multiresolution index of the ridge top flatness |
| MRVBF | Multiresolution index of valley bottom flatness |
| NDVI | Normalized difference vegetation index |
| PIW | Prediction interval width |
| QRF | Quantile regression forest |
| RF | Random forest |
| RK | Regression kriging |
| SAR | Sodium adsorption ratio |
| SAS | Salt-affected soils |
| SEM | Structural equation modeling |
| SI | Salinity indices |
| SOC | Soil organic carbon |
| SVM | Support vector machines |

# 1. Introduction

## 1.1. Digital soil mapping

Spatial soil information is essential for sustainable land management, land use planning, and precision agriculture (Lagacherie and McBratney, 2006). This information is important in protecting soil resources and other aspects of environmental science, such as hydrological and ecological modeling, climate change, and preserving natural resources (Gray et al. 2011). In recent years, there has been a significant increase in the use of digital soil mapping (DSM), which involves the creation of soil maps using digital technology. These maps provide valuable soil information and are increasingly being used in the field of soil science. DSM, by applying statistical and mathematical techniques, can capture the relation between soil observations and environmental variables that creates spatial and temporal soil information (McBratney et al. 2003).

The main framework of DSM refers to the concept of fundamental soil development theory by Jenny (1994). This concept consists of *CLORPT* factors, which represent climate (*CL*), organisms (*O*), relief (or topography) (*R*), parent material (*P*), and time (*T*), respectively. This concept later evolved into the *scorpan* framework (McBratney et al. 2003), in which *S* stands for soil properties and classes, and *N* for space or spatial position. This updated formula fits a spatial model to describe quantitatively the relationship between soil properties and environmental variables for a given spatial location. Over the past decades, there has been an expansion in the development of DSM techniques at different scales, from the spatially detailed soil maps for a region or catchment to global soil mapping (Poggio et al. 2021). This evolution is happening at the same time as spatial data infrastructures are becoming more and more developed, such as advanced remote sensing data, which provides us detailed information in a continuous spatial way about auxiliary variables from many exhaustive resources (Boettinger et al. 2008; Dewitte et al. 2012; Weiss et al. 2014). The other reason for this evolution is the development of computational tools, such as accessible and available software solutions to analyze this information (Minasny et al. 2008; Malone et al. 2017; Rossiter, 2018).

Digital maps are available for both quantitative and qualitative soil attributes, for instance, soil classes or soil type (Behrens et al. 2005; Hounkpatin et al. 2018) as a discrete qualitative characteristic or soil pH or organic carbon as continuous quantitative variables (Brady and Weil, 2008; Dai et al. 2014). Machine learning (ML) algorithms (Lagacherie, 2008) and geostatistics (Heuvelink and Webster, 2001), or their combination, are common ways of predicting soil properties and delivering spatial (or even spatiotemporal) information on soils. The geostatistical framework (Heuvelink and Webster, 2001) has traditionally been the foundation for spatial soil prediction. In this framework, the spatial variability of a soil property is modeled as the sum of a linear combination of environmental covariates and a spatially autocorrelated (stochastic) residual, and kriging is used to predict at unobserved locations (Krasilnikov et al. 2008; Pásztor et al. 2015; Szatmári and Pásztor, 2019).

Regression kriging (RK), widely used for spatial inference of soil properties, employs correlation with environmental variables and spatial correlation together (Hengl et al. 2004; Hengl et al. 2007). Applying geostatistical models in DSM has several benefits; first, spatial autocorrelation is taken into account in the modeling. This can be useful for environmental variables since soil properties vary between different places, while there is a correlation between them. Second, a reliable statistical model for spatial variation is presumed, which makes it possible to interpret the underlying physical processes that the model is trying to portray (Burrough, 2001; Wadoux et al. 2020). Third, the prediction is associated with the measure of uncertainty. Uncertainty maps are necessary for many situations where the prediction is not the only thing that matters, like during a decision-making process (Heuvelink and Webster, 2022). Further studies by Heuvelink et al. (1989) and Goovaerts (2001) demonstrated how this soil map uncertainty might be carried over into later analyses.

Geostatistics also has some limitations; first, it is assumed that the residuals are normally distributed, stationary, and have constant mean and variance. Second, the relationship between soil properties and covariates is not exclusively linear; therefore, it is not straightforward and might lead to additional challenges. Also, it may fail to capture detailed spatial soil variation and it may be computationally demanding with large and diverse datasets (Kanevski et al. 2004; Heuvelink and Webster, 2022).

Another common way of DSM is the application of ML algorithms. With the development of remote sensing and the increase in the number and accessibility of environmental covariates, the focus has turned to employing a flexible, non-linear model structure to enhance the quality and accuracy of modeling (Khaledian and Miller, 2020; Hateffard et al. 2019). Nowadays, ML is frequently applied for regression and classification tasks in different topics (Hengl et al. 2015; Heung et al. 2016). Despite geostatistical approaches, ML algorithms can handle a variety of variables as predictors and do not make an assumption about the distribution of the observations. Several studies have demonstrated the advantages of ML methods for DSM (Hengl et al. 2015; Chen et al. 2019). ML solely depends on algorithms, unlike geostatistics, which depends on explicit statistical models. Although ML offers many valuable advantages, it should be utilized with caution because it might be vulnerable to over-fitting and lacks transparency (Arrouays et al. 2020).

ML has much flexibility to offer when modeling the relationship between dependent and independent variables. When applied effectively with a large data set and many covariates at fine resolutions, it can significantly increase the accuracy of soil maps (Hengl et al. 2015).

The accuracy of DSM is influenced by several factors, including number of soil samples, sampling design, the number of environmental covariates, and the choice of ML algorithm. I will discuss these factors in detail later. Additionally, I will address the current challenges and gaps in the application of DSM.

### 1.2. Soil samples and sampling design

By sampling, one can estimate the characteristics of a whole population by selecting a subset of individuals from the entire population (Cochran, 1977). A key part of the DSM process involves predicting soil properties of interest at locations that have not yet been visited. Different statistical models have been widely employed for this purpose, assuming that the models are compatible with real conditions (Szatmári et al. 2015). Therefore, the sampling should be representative enough to cover the whole population, as the sample size and spatial location of the samples contribute to the accuracy of mapping. Sampling design, which has given much attention to optimization, specifies which points or locations should be visited before going to the field (Hengl et al. 2003). Depending on the availability of environmental

covariates and the shape of the study area, the sampling design can be different. For example, if the area has a regular shape and there is no information about the covariates, it is recommended to do sampling based on a regular grid such as square, triangular, etc (Vašát et al. 2010). While in the case of the availability of covariates, which is the case in most areas nowadays, and if we assume there is a correlation between soil properties and their environmental covariates, a decent technique is to make sure that the measurements are evenly distributed throughout the feature space (Wadoux et al. 2019). Spatial coverage sampling, feature space coverage sampling or conditioned Latin Hypercube sampling (cLHS) (Minasny and McBratney, 2006; Brungard et al. 2015) are the common ways for this purpose. To guarantee distribution in both geographic and feature space, spatial coordinates can be added to the list of variables.

Sampling design can also be divided into two categories; 1. non-probability sampling, such as grid-based sampling, and cLHS, which is the most used by far (Wadoux et al. 2020); 2. probability sampling, such as simple random sampling (Tziachris et al. 2019; Brus, 2022), which is less applied in studies. There are many different sampling designs, but there is no best sampling design, according to Brus (2019), and the best scheme depends on the mapping approach. However, he suggested applying cLHS technique to select the soil samples location since it can perfectly represent the variability of auxiliary variables in a feature space.

Latin hypercube sampling (LHS) is a stratified random process that effectively samples variables from their multivariate distributions. By maximally stratifying the marginal distribution, it fully covers the range of each variable. The method was initially developed to choose input variables for computer models during Monte-Carlo simulation effectively. It has been used to evaluate the uncertainty in prediction models in various environmental research and soil science (Pebesma and Heuvelink, 1999; Minasny and McBratney, 2002).

While a precondition for employing cLHS, as proposed by Minasny and McBratney (2006), is that the sampling locations must actually exist in the real world. In this method, there may be countless strata combinations based on different environmental factors, and samples are randomly fallen in each stratum. As a result, sample sets generated by cLHS can differ dramatically between runs with the same sample size (Yang et al. 2020).

Auxiliary variables can play an important role in optimizing sampling design since this information can provide insight into soil variation (Hendriks et al. 2019). The observations can be conducted at "smarter" locations, such as at locations where soil variability was anticipated to be the highest. Also, the number of soil observations can be reduced, in which field and/or laboratory measurement costs might be generally lowered, increasing the overall cost-effectiveness of the soil survey (Szatmári et al. 2019). Many studies concluded that by properly utilizing the available tools and techniques, they could have decreased their sampling effort (Kempen et al. 2009; Shirani et al. 2015). In a study in the Hunter Valley of New South Wales, Australia, Minasny and McBratney (2006) demonstrated that cLHS closely resembled the original distribution of the environmental variables with a small sample size. Scarpone et al. (2016) applied cLHS to identify the locations for field measurements of soil thickness.

### 1.3. Environmental covariates of digital soil mapping

Environmental covariates or auxiliary variables are another requirement of the DSM process. They are used as predictors to calibrate the model. They are proposed to explain the soil-forming factors and other physical and chemical processes which lead to spatial variation of soil properties.

Multiple sets of environmental variables have been utilized in DSM. Here, we can list them based on the representation of *scorpan* factors, but we have to note they might not be easily available in all case investigations;

1. climate; precipitation and temperature maps
2. organisms; Land cover type, biomass map, Normalized Difference Vegetation Index (NDVI), Satellite based Enhanced Vegetation Index (EVI), and any other indices which represent vegetation cover of the area can be applied.
3. relief and topography; Elevation, slope, aspect, plan and profile curvature, topographic wetness index, multiresolution index of valley bottom flatness (MRVBF), multiresolution index of the ridge top flatness (MRRTF), flow direction, channel network, landform classes.
4. parent material; type of parent material and bedrock, depth to bedrock and thickness, mineralogy.

5. age of the soils; this is rarely used in practice since it is difficult to determine accurately, due to the complexity of soil formation and the limited availability of appropriate methods.
6. spatial position; directions, longitude and latitude, distance to nearest ocean/mountain/hill/river.
7. anthropogenic influences; land use and landcover management, probability of erosion, application of fertilizers.

There are some resources that one can achieve the intended covariates at different resolutions. The most common resources to derive covariates are;

1. Digital elevation model (DEM) can help acquire covariates that mostly represent topography
2. Remote sensing images from platforms like Landsat 7 and 8, MODIS, or Sentinel
3. Geological and lithological map of the study area
4. Global land cover/ land use maps
5. Available information of climate stations.

Each study depends on the size and heterogeneity of the area, budget, and accessibility of these covariates can use a combination of them. Some researchers applied only a few of them (Dai et al. 2014), while others used a large set of covariates (Nussbaum et al. 2018; Hengl et al. 2017). Some publications have exclusively used parameters generated from DEM (e.g., Silva et al. 2016; Sharififar et al. 2019), climate variables (Mansuy et al. 2014), or remote sensing data (Minasny and McBratney 2016) to calibrate the model.

Temporal covariates can be applied to incorporate the effect of time or age on soil spatial variation. For instance, Heuvelink et al. (2020) mapped the temporal dynamic of organic carbon in Argentina between 1982 and 2007 using time series of MODIS products. The accuracy of spatial predictions is improved by including more relevant covariates that can better explain the distribution of soil attributes.

Recently, the application of coordination of observations or maps of distances from observation locations as an indicator of spatial position in the *scorpan* formula has been supported in the prediction process (Hengl et al. 2018; Behrens et al. 2018).

As mentioned previously, numerous covariates can account for the spatial variation in soil. However, it is important to carefully select the most influential covariates to assess the magnitude of their contribution in this particular area. Also, the goal of covariate selection is to employ fewer covariates to calibrate ML models to accelerate the model process, reduce complexity, prevent overfitting and improve prediction accuracy.

Covariate selection can be made before the calibration of the ML model, or it can be done during the calibration method as a wrapper function. For example, Hamzehpour et al. (2019), by calculating Pearson Correlation Coefficient between the variables and eliminating the highly correlated ones before calibration, chose the covariates to be utilized in ML. The other type of covariate selection relies on the inference made by a calibrated ML, which is called the wrapper method. Recursive feature elimination is an optimization procedure that is the most used "wrapper" technique (Minasny et al. 2018; Gomes et al. 2019). Most often, topography or terrain derivatives has been recognized as the most significant factor affecting soil attributes (McBratney et al. 2003; Nussbaum et al. 2018), whereas parent material, particularly in dry regions, is another important factor (Heung et al. 2014). Lamichhane et al. (2019) claimed that the most important factors for the spatial mapping of soil organic carbon are covariates representing organisms, climate, and topography, respectively. Also, they indicated that vegetation and land use in small areas were shown to be better predictive of soil organic carbon. In general, covariates reflecting soil surface and climatic factors (particularly precipitation) seem to be the most significant for predicting soil chemical characteristics, while a combination of relief, vegetation cover, and parent material are more important to predict soil classes and soil physical properties (Hengl and MacMillan, 2019). It is strongly advised to carefully prepare and prioritize soil covariates, as this process might frequently need time and resources.

### 1.4. Machine Learning algorithms

After providing soil observations and environmental covariates, we need to apply ML algorithms to capture the relation between them and produce spatial predictions. Many different algorithms have been used in DSM. The accuracy of predictive models is essential as it determines the quality of their predictions which is scientific evidence to advise policy-

making decisions. Therefore, it is crucial to increase accuracy by selecting a suitable approach, then finding and creating the most accurate prediction model. However, due to numerous aspects and factors involved in the modeling process, it is difficult to choose an appropriate approach and identify the most accurate predictive model for a given dataset (Wadoux et al, 2020).

Depending on regression or classification purposes, there are different methods that one can apply. The most popular ones include decision trees (DT) (Giasson et al. 2011), random forest (RF) (Heung et al. 2014; Hengl et al.2015; Hounkpatin et al. 2018), artificial neural networks (ANN) (Zhao et al. 2010; Dai et al. 2014), and support vector machines (SVM) (Kovačević et al. 2010;Pereira et al. 2022). Some of the literature tried to compare the accuracy of these models. So far, between different ML techniques, RF has proved its applicability in spatial predictions of soil properties in several studies, especially for regression purposes (Dharumarajan et al. 2017; Hengl et al. 2018; Vaysse and Lagacherie, 2015; Kinoshita et al. 2016). Recently, Vaysse and Lagacherie (2017) used quantile regression forest (QRF), a variant of RF. This method calculates the quantile of the predictions associated with the uncertainty map.

In a study in Brazil, Menezes et al. (2013) claimed that applying a knowledge-driven approach, such as expert systems with ML algorithms, can help effectively utilize the soil scientist's knowledge. They used this method under fuzzy logic to predict soil types and properties. To produce accurate soil organic maps, Dai et al. (2014) incorporated an ANN with the estimation of its residuals using conventional kriging. They concluded that ANN-kriging could efficiently increase the accuracy, and kriging is a good way to show the spatial variation of soil organic carbon. Zhao and Shi (2010) compared different methods, such as multiple linear regression (MLR), DT, ANN with kriging, universal kriging, and regression-kriging, to predict soil spatial distribution of organic carbon. In their study, DT showed the highest performance explaining 67 % of the total variation.

Pereira et al. (2022) compared the ML algorithms with ordinary kriging and inverse distance weighting. SVM has been selected as the ML model and performed better in interpolating soil attributes.

Heung et al. (2016) applied ten different ML methods and 20 environmental covariates to predict soil taxonomic units in DSM. The models like regression trees and random forests were favored more in terms of speed and interpretability of the results, while the k-nearest neighbor and support vector machine with radial basis function had the highest accuracy, near 72%. Also, they indicated that the choice of model and sampling design can significantly impact the outputs. Khaledian and Miller (2020) reviewed five different ML applicability in terms of the number of hyperparameters, size of samples, feature selection, training time, and interpretability of the resulting model. If training time is limited, algorithms such as SVM, MLR, and cubist should be considered. At the same time, ANN would produce better results with large datasets and without time limitations. With a small dataset, RF, SVM, and K-nearest neighbors will likely have better results than MLR and ANN. Also, RF and MLR are more appropriate algorithms in terms of interpretability since they do not operate as "black boxes," unlike ANN. One of the reasons the models such as DT and RF are increasingly becoming popular is that these algorithms can handle both linear and non-linear relationships of the data and require less preprocessing. They believed there is no best model to predict soil properties, and each model has its advantages and limitations, which is critical to understand precisely before selecting the appropriate ML model.

Recently, some studies have suggested employing ensemble modeling, which combines predictions from two or more individual models (e.g., Mishra et al. 2020; Brungard et al. 2021). In this way, it is possible to take advantage of every single model, leading to more accurate and reliable predictions (Seni et al. 2010; Zhang and Ma, 2012). Additionally, ensembles effectively address ML algorithms' challenges, such as handling missing values and enhancing confidence estimation by weighting different variables and taking into account the most crucial ones. Song et al. (2020) trained three ML models separately for the pedoclimatic zone in China, then merged their predictions with the weighted ensemble learning model. They reported that applying ensembles could increase the accuracy of soil maps by 12.6 %. Hengl et al. (2017) fitted ensemble ML methods as a combination of three different methods. They mentioned that an improvement in accuracy and variation explanation had been observed.

Mishra et al. (2021) compared RK with four different ML models to predict the spatial distribution of surface soil organic carbon stock. RK demonstrated fewer prediction errors rather than SVM, while the accuracy was comparable to gradient boosting and RF. Then, they combined the predictions of these four ML modes. The results of ensemble prediction showed higher accuracy and more detailed spatial variation. Therefore it can be a more satisfactory choice than selecting any single model in soil mapping. Brungard et al. (2021) indicated ensembles ML for regional models are approximately as accurate as global, but they have less uncertainty.

### 1.5 Issues in digital soil mapping

There are some knowledge gaps and issues in DSM, which need to be addressed. For example, data availability and quality, sampling and sampling design, consideration of spatial information, multivariate mapping, validation and uncertainty analysis, taking account of pedological knowledge, extrapolation, and interpretation of ML models (Hempel et al. 2008; Wadoux et al. 2020). To address these issues and gaps, researchers and practitioners are working on developing new and innovative methods for data collection, integration, and analysis, as well as on improving the accuracy and consistency of the data used in DSM. In my dissertation, I only describe extrapolation and joint spatial modeling issues of soil properties.

#### 1.5.1. Extrapolation of soil properties

The application of ML techniques is getting more attention in DSM, while they require sampling datasets to train the model. The sampling needs effort and cost, and it is a time-consuming process. The density of soil samples alters dramatically among different regions. Several areas of the world have significant gaps and free spaces due to very few or no soil observations. Therefore, producing soil maps at global and national levels is limited and mainly relies on the extrapolation ability of the trained model (Poggio et al. 2021). Spatial extrapolation is transferring the model to a new geographic location from which the training data has been calibrated. Extrapolation is the opposite of spatial interpolation, which uses point values in a study area to predict the vicinity (Takoutsing and Heuvelink, 2022). In other words, extrapolation can be applied from an area with observations (donor area) to predict

the soil properties in the area without observations (recipient area). The only condition is that both areas should have similar soils, which can be evaluated by the similarity of soil-forming factors. As long as the existing soil maps have captured local environmental heterogeneity and soil–landscape interactions in the area, they are helpful as predictive models to extrapolate over unmapped regions with similar environmental characteristics. Therefore, it is crucial to determine the locations where the feature space differs from the training data. Meyer and Pebesma (2021) developed the "Area of Applicability" (AOA) methodology. Based on AOA, we can only extrapolate to the regions the trained model has seen in the donor area. This concept performs based on the dissimilarity index (DI) between the covariates in the training data and new locations. It determines a delineation of the extendable areas where the trained model can be employed.

Mallavan et al. (2010) introduced the Homosoil method as a helpful way to extrapolate from other areas with similar soil forming factors where the observation is rare or there is no detailed map for the area of interest. Several studies tried to apply this method (Malone et al. 206; Silva et al. 2016; Angelini et al. 2020). Thompson et al. (2006) created a quantitative soil-landscape model to analyze the differences between locations with identical soils. Their model tends to overestimate and underestimate the soil properties in the recipient area. They attributed the errors to the difference in terrain attribute distribution. Afshar et al. (2018), examined the similarity index between two areas by Gower's similarity index and applied the multinomial logistic regression model to evaluate soil classes in the recipient area. They found out that the extrapolation was successful up to 60% prediction accuracy, while it can be significantly efficient in terms of time and costs.Angelini et al. (2020) employed structural equation modeling (SEM) from an area in Argentina to an identical soil landscape in the United States. SEM is a hybrid technique that includes pedological knowledge to analyze the ability to extrapolate rather than empirical methods. They believed that quantifying all the soil-environment interactions over time is still challenging, and we need a better understanding of different aspects. Although combining the pedological knowledge about the recipient area with soil formation indicators can be a powerful tool for extrapolation.

Nenkam et al. (2022) questioned whether it was possible to extrapolate in areas considered similar based on the Homosoil technique and compared the results with existing soil global

maps. They discovered that extrapolation in geographic space is possible while adding local data to the training dataset can increase accuracy.

There are some problems in spatial extrapolation, which might affect the predictions for the recipient area. In addition, a comparison of the ability to extrapolate between different ML models has not yet been fully investigated. In the study by Takoutsing et al. (2022), when RK extrapolation was replaced by RK interpolation, there was a decrease in prediction performance. It is unknown if RF has a similar outcome or if RF performs geographical extrapolation better or worse than RK. Later, Takousting and Heuvelink (2022) studied the ability of extrapolation between RF and RK and concluded that extrapolation had much better results for RK than RF. Therefore, avoiding extrapolation in feature space with RF is important, which can be done by calculating AOA (Meyer and Pebesma, 2021).

Neyestani et al. (2021) evaluated the potential of extrapolation for eleven ML algorithms to predict soil classes. Also, they looked into the problem of imbalanced soil class observations through oversampling techniques, which could improve the accuracy. DT and RF performed the highest accuracy regarding the overall accuracy and Kappa value.

The other extrapolation issues include; the different history of land cover and land use, the inability of the model to cover the feature space well, the nonlinearity of soil processes (Taghizadeh-Mehrjardi et al. 2022), differences in the past and present of environmental variables, and difficulties in quantifying soil-landscape relation (Rossiter, 2021).

Regarding these issues, if we do extrapolation, it is crucial to quantify the measurements error and uncertainty of the predictions in DSM. Quantile regression forests (QRF) (Meinshausen and Ridgeway, 2006), as a modification of the RF model, compute all quantiles of prediction distribution. Therefore, it can quantify the prediction uncertainty at all prediction locations. Also, prediction interval width (PIW) can be calculated from the difference between lower and upper quantiles of estimations for any point in the predictor space (Zhang et al. 2019). It is expected to get wider intervals in areas where we extrapolate with high uncertainty with a certain probability. Some studies successfully applied RF and QRF in the spatial prediction of soil properties (Forkuor et al. 2017; Hengl et al. 2015; Vaysse and Lagacherie, 2017).

### 1.5.2. Joint spatial modeling

The soil is a complex system since it changes from one point to another, and many of its properties alter over time (Heuvelink and Webster, 2001). Also, the soil has many interactions with itself and the environment from a broader perspective. Soil-landscape interactions exhibit spatial and temporal heterogeneity across the different scales (Haygarth and Ritz, 2009). Researchers are investigating relationships between spatial soil information and environmental covariates to fully utilize the technical capabilities of mapping (Hartemink et al. 2008; Arrouays et al. 2014). Some soil properties strongly correlate with each other, or they might show spatial autocorrelation. For example, the relative combination of sand, silt, and clay in the soil determines the soil texture, and the sum of their percentages should always add up to 100%. In many cases, when modeling and predicting particle size fractions of the soil separately, there is no guarantee that the sum of the maps will consistently add up to 100% at each pixel. Similarly, when dealing with elevation differences that cannot be negative (such as the difference between the surface and the depth of shallow groundwater level), mapping these properties can result in incoherent outcomes. Therefore, when spatial modeling of more than one variable is intended, it is better to consider their spatial interdependence and jointly model their spatial distribution (Szatmári et al. 2020). The application of multivariate geostatistics can help not only make use of this interdependency's benefits in spatial modeling but also to produce consistent findings that are highly valuable. Laborczi et al. (2019) applied the spatial inference on the original sand, silt and clay variables by regression kriging, which employs environmental correlation and geostatistical interpolation. This approach uses multiple linear regression analysis to model the trend of the target variable on environmental covariates and interpolates the regression residuals using standard kriging. Therefore, instead of mapping the particle size fractions independently, they applied composite regression kriging based on Additive Log-Ratio (alr) on the original values and achieved the sum of the particle size fractions 100% together at the end. Furthermore, ML algorithms are incapable of addressing these problems due to their inability to handle the spatial aspect of the data. Similarly, the complexity of ML models, often described as 'black

boxes,' poses challenges in their interpretation, particularly when it comes to incorporating spatial perspectives, as highlighted by Brenning (2022).

Some ML models try to deal with multiple responses, which could show correlation, but these algorithms are not accepted yet. Geostatistics comes to help to take the spatial aspect of the data into account. Geostatistical approaches are widely employed in soil science, which has been completed by ML models in the past decade (Oliver, 1987; Heuvelink and Webster, 2001; Szatmári et al. 2021; Steinbuch et al. 2022). Multivariate geostatistics is a proper approach to explicitly take the joint spatial variability by considering the spatial modeling of the variables. Furthermore, it exploits spatial interdependence to give coherent and even more precise spatial predictions for the soil property of interest (Goovaerts, 1997; Wackernagel, 2003; Webster and Oliver, 2007). Multivariate geostatistics is also capable of modeling and quantifying the uncertainty related to a spatial prediction, which has become a prevalent demand in environmental modeling and mapping (Pásztor et al. 2016; Szatmári et al.2019). Applying both ML algorithms and multivariate geostatistics can take the merits of both approaches; 1.Modeling complex, non-linear relationship between the soil properties of interest and the environmental covariates, 2. Jointly modeling the stochastic part of the spatial variability of the soil properties of interest.

## 2. Aims and structure of the research

My thesis aims to explore the extrapolation potential of soil property modeling and joint spatial modeling issues. In order to achieve this, the following detailed objectives will be pursued:

1. Predict and map the spatial distribution of soil properties in two small-scale areas that have different physiographic conditions.
2. Evaluate the potential and efficiency of different techniques in spatial predictions of soil properties.
3. Select the best model with the highest accuracy and least error to extrapolate over the larger areas.
4. Assess the possibility of extrapolation by AOA method and validate the results by samples taken from large areas.
5. Estimate the similarity between two areas by different methods and evaluate if there is an agreement between these methods.
6. Explore the possibility of predicting over an unknown area by available dataset.
7. Predicting and mapping salt-affected soils (SAS) indicators by applying ensemble machine learning.
8. Jointly modeling the prediction results with multivariate geostatistical techniques.
9. Evaluation of SAS indicators at the field scale based on final maps.

As part of my PhD research, I conducted three case studies to achieve my research objectives. The first case study focused on achieving objectives 1 to 4. This study comprised two phases, with the first phase involving the comparison of various ML models for the spatial prediction of soil properties. In the second phase, I evaluated the feasibility of extrapolating the trained model to larger areas. The study was conducted in small-scale areas of Látókép and Westsik, with extrapolation techniques applied to larger areas of Hajdúhát and Nyírség. The soil properties studied included soil organic carbon (SOC) content, bulk density (BD), SOC stock, soil pH, electrical conductivity (EC), and soil carbonates, with MLR, RF, ANN, and SVM used as the applied techniques.

The spatial variability of soil properties in Látókép and Westsik was attributed to topographic variability. It is important to note that the extrapolation was only applied to the arable lands of Hajdúhát and Nyírség, as the land use for smaller areas was limited to agricultural fields.

In my second case study, I focused on achieving research objectives 5 and 6. The objective was to determine the potential for extrapolation in geographic space between two areas, based on the similarity of soil-forming factors. I hypothesized that the transferability of a model from a donor area to a recipient area depends on the similarity of soil-forming factors between the two areas. To accomplish these objectives, I employed four different methods to determine the potential for extrapolation, including similarity in soil types, homosoil approach, dissimilarity index by AOA, and QRF prediction interval width. The study was conducted in four African countries: Ethiopia, Kenya, Burkina Faso, and Nigeria, with soil properties of interest being SOC content, clay content, and soil pH. This project has been done with the help of colleagues in Wageningen University and Research and in Netherland. To achieve these objectives, I first measured the similarities by soil types and homosoil approach, and then trained a random forest (RF) model and extrapolated it to other areas. I also calculated the dissimilarity index by the AOA method and QRF prediction interval width, and finally validated the extrapolation results. The aim was to check for any agreement between different measures of extrapolation and actual cross-validation results.

My third case study aimed to achieve research objectives 7 to 9, which addressed joint spatial modeling issues. Specifically, I focused on the spatial prediction of salt-affected soils (SAS) in an arable land located in Hungary. The goal was to identify locations with high salinity and generate a high-resolution map with acceptable accuracy for policy-making decisions. SAS indicators such as pH, electrical conductivity (EC), and sodium adsorption ratio (SAR) were used for the study. I hypothesized that there might be spatial interdependency and cross-correlation between the SAS indicators. As such, the spatial distribution maps of the SAS indicators would be more accurate if we jointly modeled them.

In the context of these three case studies, it is beneficial to mention relevant publications up to the date of defense. Here is a brief description of the publication:

1. Case Study One:

- **Hateffard Fatemeh**, Szatmári Gábor, Novák Tibor József (2023): Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. Soil Science Annual (IF= 1.74, SJR = Q2) - Accepted manuscript.
- **Hateffard Fatemeh**, Novák Tibor József (2021): Soil sampling design optimization by using conditioned Latin Hypercube sampling, Advances in Modeling Soil Systems, ISMC2021-35.
- **Hateffard Fatemeh**, László Márta, Novák Tibor József (2022): Anthrosequence of soils on Aeolian Sand Dunes in Westsik's experimental field, Nyíregyháza, Hungary. Book: Soil Squences Atlas V, Publisher: Nicolaus Copernicus University Torun

2. Case Study Two:
    - **Hateffard Fatemeh**, Luc Steinbuch, Gerard B.M. Heuvelink (2023): Evaluating the extrapolation potential of random forest digital soil mapping. Geoderma (IF = 7.422, SJR = Q1) - Under review.
3. Case Study Three:
    - **Hateffard Fatemeh**, Balog Kitti, Tóth Tibor, Mészáros János, Árvai Mátyás, Kovács Zsófia Adrienn, Szűcs-Vásárhelyi Nóra, Koós Sándor, László Péter, Novák Tibor József, Pásztor László, Szatmári Gábor (2022): High-Resolution Mapping and Assessment of Salt-Affectedness on Arable Lands by the Combination of Ensemble Learning and Multivariate Geostatistics. Agronomy 12. 1858. (IF = 3.949, SJR = Q1).
    - **Hateffard Fatemeh,** Novák Tibor József, Szatmári Gábor (2021): Spatial prediction of soil pH using machine learning models in Látókép, Hungary, webGeoMATES.

## 3. Materials and Methods

### 3.1. Study area

#### 3.1.1. Case study one

Hajdúhát and Nyírség are two microregions in Hungary (Figure 1) which are located on the Great Hungarian Plain. Microregions are lower levels of the landscape system, which are more homogenous in terms of geomorphology and landform. Based on Hungary's landform element and geomorphic landscape maps, both areas have a flat landform type located on Pleistocene alluvial plains (Józsa and Fábián, 2016). Later, during the late Pleistocene, they were partially or totally reshaped by aeolian processes, including sand and dust redeposition and loess deposition. They belong to the warm temperate, fully humid climate with warm summers (Kottek et al. 2006).

The surface of Hajdúhát is covered by fertile loess and silt with chernozem soils, making it highly suitable for farming. The altitude differs from 83 to 155 m. The loess cover is 1 to 2.5 meters deep. The southern part has the largest thickness values, which reaches a depth of 10-15 meters (Kertész and Křeček, 2019).

The surface of Nyírség was covered by wide flood plains and alluvial fans of the interfluves before, while later, it was mainly formed by wind-blown sand. In the late Pleniglacial period, there was a significant sand movement, and different sand forms developed in Nyírség. Sandy loess, loess-sand, and loess, with a depth of 4 m in some places, can all be found on the dune cover. These wind-blown sands are at high risk for wind erosion that is stabilized by planting acacias, fruit trees and tobacco (Kertész and Křeček, 2019). The altitude ranges from 86 to 170 m in this area.

Látókép is a long-term experimental station that was founded in 1983. It is located between 47◦32′ 30″ – 47◦33′ 44″ N; 21◦26′ 15″ – 21◦27′ 11″ E in Hajdúhát microregion on a plain area (Figure 2). The main application of the 160 ha large area is for fertilization, cultivation, and irrigation experiments. The area has a flat topography with only small differences in elevation, which is between 111-114 meters (above sea level), and 0.5-1% is the average slope. The soil type is calcareous chernozem due to the mollic horizon in topsoil, with secondary carbonate accumulations. The soil is deep and well-aggregated.

Westsik Vilmos is a crop rotation experimental station covering 47 ha, established in 1929. It is located between 47◦59′ 21″ – 47◦58′ 35″ N; 21◦41′ 57″ – 21◦42′ 18″ E, in the Nyírség microregion (Figure 2). The station is known as a remarkable example of successive production in Hungary. It is utilized to investigate the effects of organic manure on soil properties and crop yields based on various cropping systems. Therefore, it can provide data for farmers regarding applications of green manure and fertilizers. Also, it provides useful plant and soil information regarding scientific research, which can be applied in soil quality management and sustainable production. The elevation is between 101-105 meters, on a slightly undulating sandy landscape, with a 1.5-5% average slope. The station intended to improve soil organic carbon and fertility due to sandy texture and poor aggregated soils.

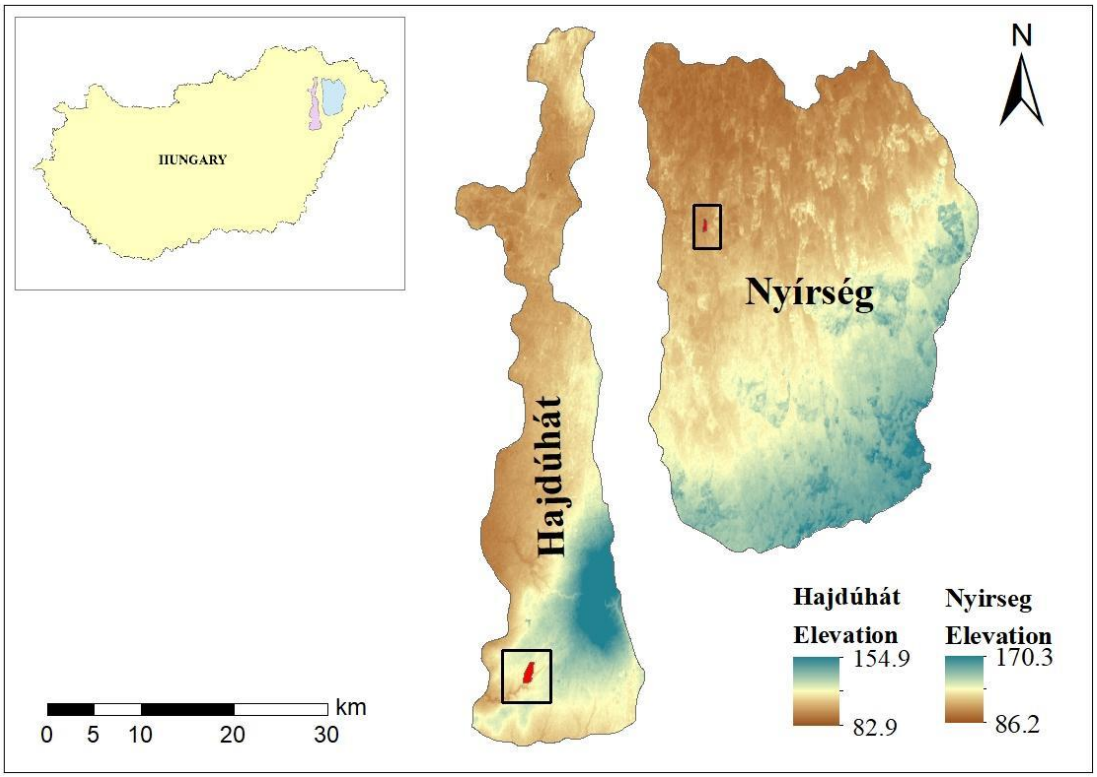



*Figure 1. Study areas for case study one. The location of microregions in Hungary shown on Digital Elevation Model (DEM). Squares are the location of Látókép inside Hajdúhát microregion, and Westsik inside Nyírség microregion. The unit of measurement is in meters.*

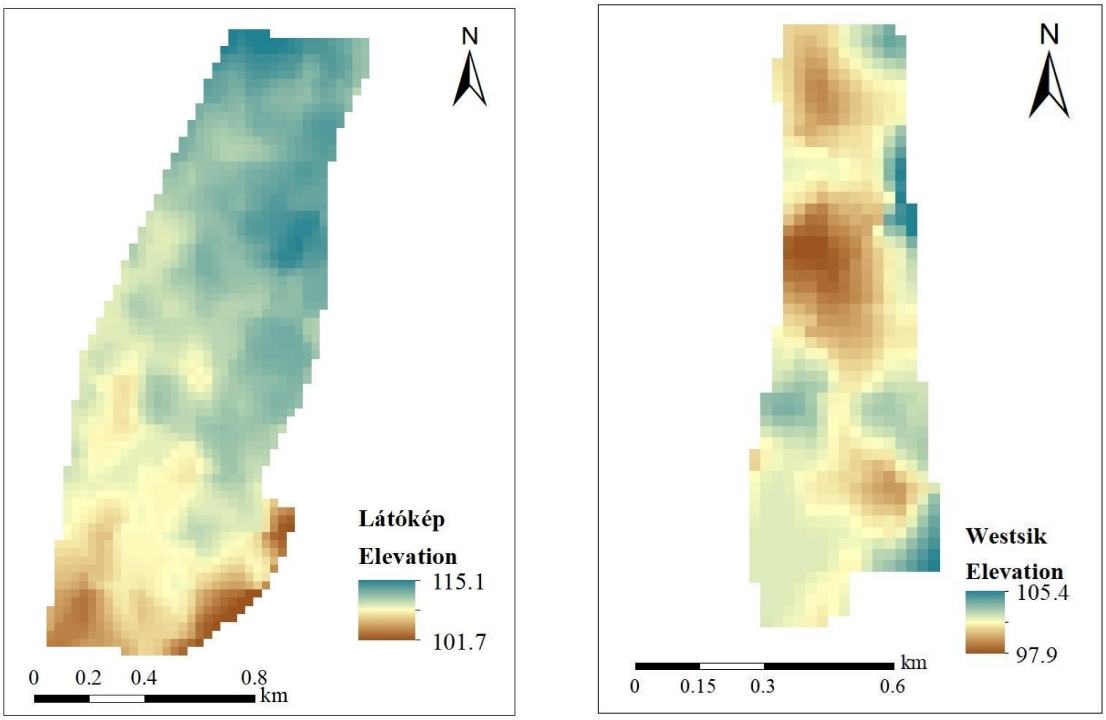



*Figure 2. Study areas for case study one. Látókép and Westsik shown on Digital Elevation Model (DEM). The unit of measurement is in meters.*

### 3.1.2. Case study two (Selected African countries)

Ethiopia, Kenya, Burkina Faso, and Nigeria were selected from African countries as our study area (Figure 3). The reasons for selecting these countries are: first, we wanted two countries with different similarity index; second, we required that each country has sufficient soil samples in the public database and has fairly good and uniform coverage of the whole country. Ethiopia and Kenya are in the eastern region, while Nigeria and Burkina Faso are in West Africa.

Kenya has a varied climate, with dry conditions in the east and north and significant rainfall and humidity in the west. Ethiopia's climate also has a considerable regional climatic variation, but it is mainly tropical. The southeast and northeast, particularly the lowlands, are known for their warm desert environment; the highlands have humid subtropical and tropical savanna climates, mainly in the central and western parts of Ethiopia. The climate in Burkina Faso is mainly dry tropical with two distinctive seasons: short rainy and long dry seasons.

Nigeria has three distinct climate zones: tropical savanna in the north, arid and semi-arid desert in the center, and tropical rainforest and monsoon season in the south (https://climateknowledgeportal.worldbank.org).

The highest mountains on the African continent are observed in Kenya and Ethiopia. Also, one of the lowest points in the African continent which is the Danakil Depression (125m below sea level) is located in Ethiopia. Therefore, the difference in elevation between highest and lowest point is vast. In addition, Nigeria consists of highland and lowland in which the difference is about 3500 m. At the same time, in Burkina Faso the difference rarely reaches 700 m and it's relatively a flat country. The average elevation in Kenya, Ethiopia, Nigeria and Burkina Faso are 735, 626, 327 and 282 meters, respectively (https://en-gb.topographic-map.com/maps). Regarding soil types based on WRB (World Reference Base for Soil Resources) classification (Fao. 2009; Anjos et al. 2015), Kenya has the most variety with 28 different soil types, dominating weakly developed mineral soils with the accumulation of clay and sodium in some parts. Also, some small parts showed well-developed soil structures with high nutrient-holding capacity. In Ethiopia, around one-third of the soils are shallow over hard bedrock, especially in mountainous parts; in other areas, deep and well-developed soil structures with the accumulation of rich clay and iron are recognised. In northern Nigeria, easily erodible sandy soils are visible, where the low weathering rate limits soil formation, and in south, soils mainly developed from basic igneous parent rock, forming deep red soils with a well-developed structure and high productivity. Other parts of Nigeria have shallow, weakly developed soils and have low water and nutrient-holding capacity. There are only 12 different soil types in Burkina Faso. Almost half of the country has soils with the accumulation of iron and manganese that develops mainly under the influence of groundwater. The other dominant soil types are slightly acid soils with clay-enriched subsoil. Depending on the type of clay, they have a different capacity to hold nutrients and water (Panagos et al. 2012; Jones et al. 2013).

Ethiopia and Kenya have been recognized for the same soil moisture regime, Aridic, Ustic, and Udic, while Nigeria and Burkina Faso have the same moisture regime, Ustic. In terms of soil temperature regimes, Ethiopia and Kenya have isohyperthermic, isomegathermic, and

isothermic conditions. Burkina Faso is isomegathermic, and Nigeria is isohyperthermic, mainly a thermic regime with the above soil temperatures of 22 °C (Jones et al. 2013).

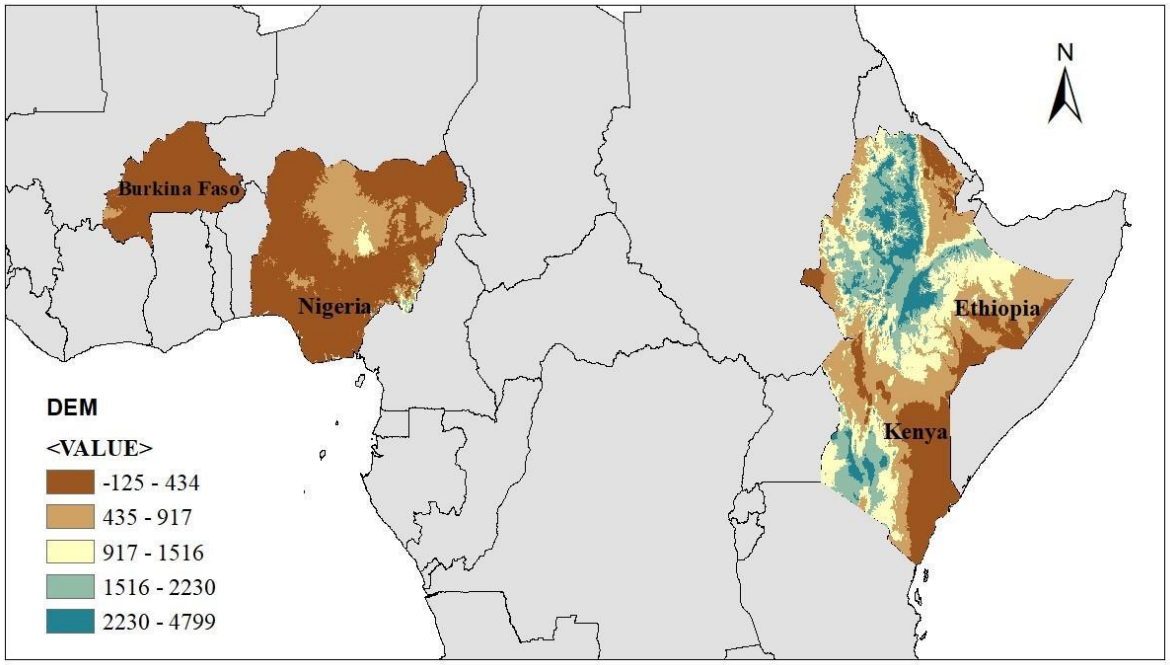



*Figure 3. Study areas for case study two. Ethiopia, Kenya, Nigeria and Burkina Faso shown on Digital Elevation Model (DEM). The unit of measurement is in meters.*

### 3.1.3. Case study three (Dunavecse)

The study area is located between the corner coordinates of 46°55′16″ N, 19°01′37″ E, 46° 55′17″ N, 19°02′12″ E, 46°55′55″ N, 19°01′41″ E, 46°55′49″ N, 19°02′12″ E, over 0.85 $km^2$ (Figure 4). The location for this plot is near Dunavecse, the Central part of Hungary and inside the microregion of "Solti-sík" (Dövényi et al. 2008). The study area, which has a quasi-rectangular shape, is impacted by soil salinity and is currently being utilized for agricultural purposes. The mean annual temperature is 10.4–10.5 °C. The average annual precipitation in this microregion is between 530 and 550 mm, with significant spatial variability. Rainfall occurs especially in spring and summer, enhancing the movement of salts (Dövényi et al. 2008). The crop rotation includes maize, sunflower, and barley, which produce reasonably good yields. This plot was ideal for the study since it has been consistently managed for the past 50 years, making gathering information about management and remote sensing data easier.

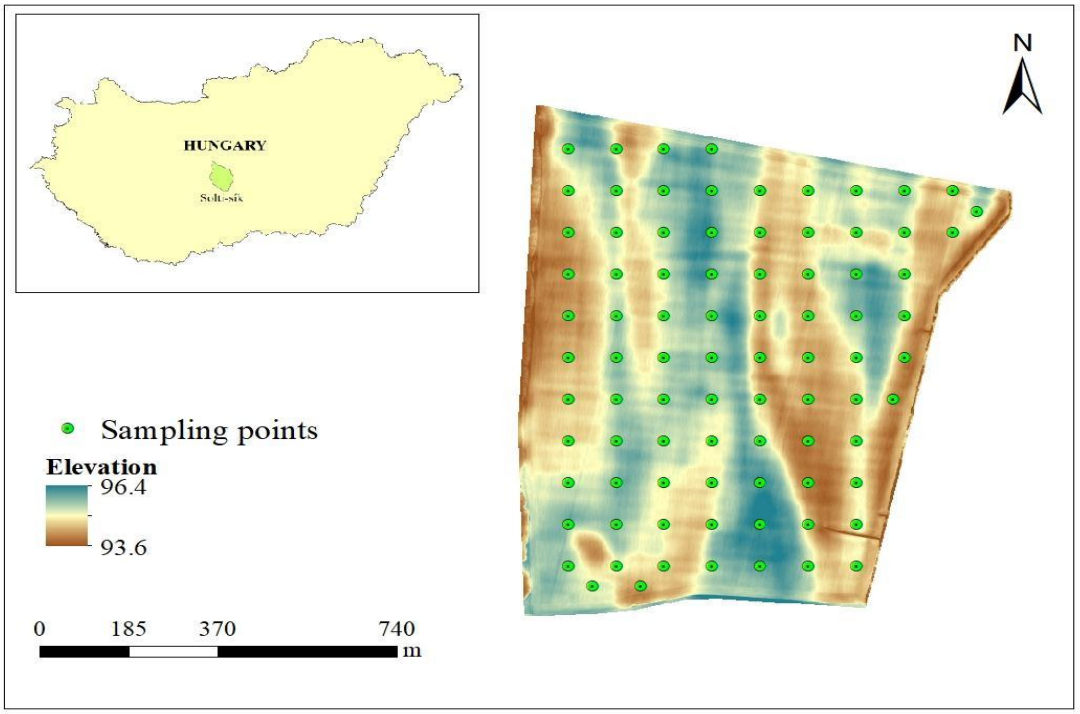



*Figure 4. Study area for case study three. A plot near Dunavecse shown on Digital Elevation Model (DEM) and sampling points. The unit of measurement is in meters.*

### 3.2. Sampling design, field survey and sample analysis

#### 3.2.1. Case study one

Since it is expected that there are relationships between environmental variables and soil properties, the appropriate sampling points are those that resemble the cumulative probability distribution of environmental variables reasonably. Therefore, before going to the field and collecting the samples, we designed the sampling points by the cLHS method for all four study areas. Based on the variability of environmental covariates in feature space, this method can efficiently cover the spatial variation of soil property of interest in study areas. The cLHS is a stratified random sampling design which attempts to ensure that the selected points are as uncorrelated and fill the space as good as possible (Brungard & Boettinger, 2010).

The inputs of cLHS were the Landsat images (https://earthexplorer.usgs.gov/) and DEM derivatives as environmental covariates, including elevation, slope, topographic wetness index, relative slope position, LS factor (topographical factor, from the USLE (Universal Soil Loss Equation) soil erosion model), valley depth, the normalized difference vegetation index

(NDVI) and Landsat 8 images Bands. Considering the size of areas, budget and time, thirty samples were selected for each study area. The sampling design for both areas were applied by the "clhs" package in R, which are presented in Figure 5 and 6. After selecting the intended locations for each area, a field sampling campaign was conducted. Each point was sampled from the uppermost 10 cm of the surface in triplicates with metal cylinders (100 $cm^3$) to measure bulk density (BD) and taken to the laboratory for further analysis.

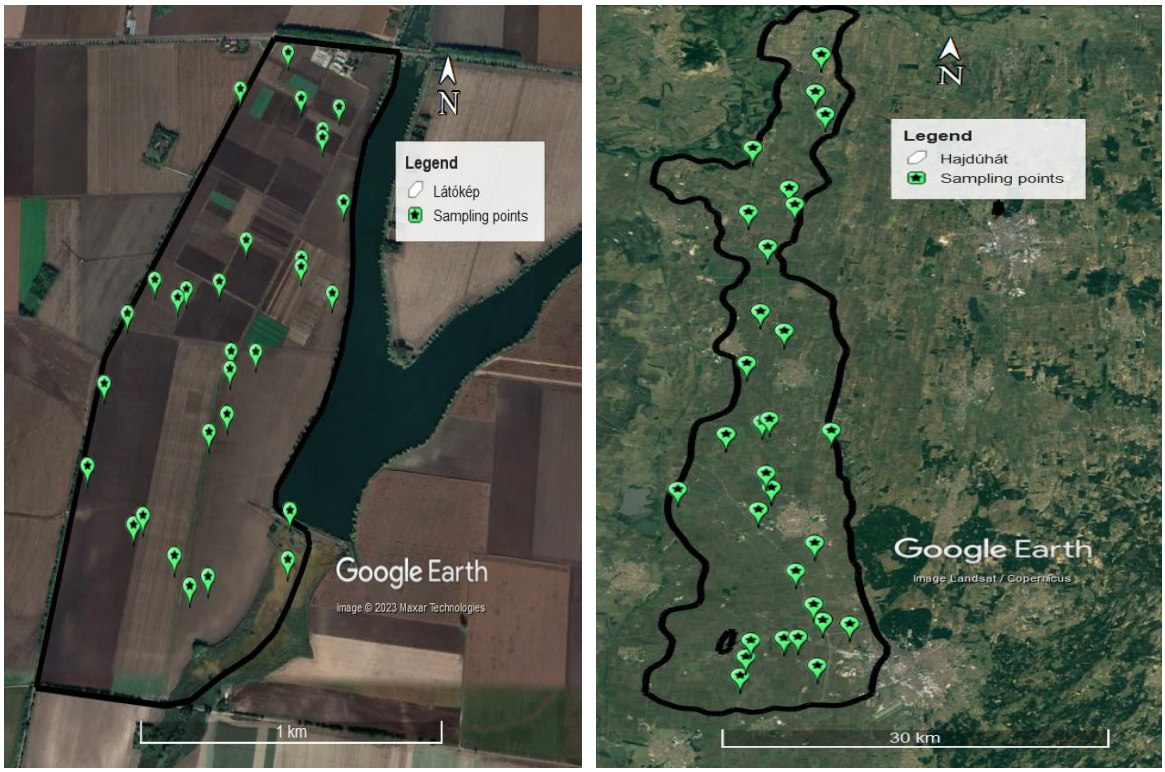



*Figure 5. Determined sampling location for Látókép (left) and Hajdúhát (right) by cLHS*

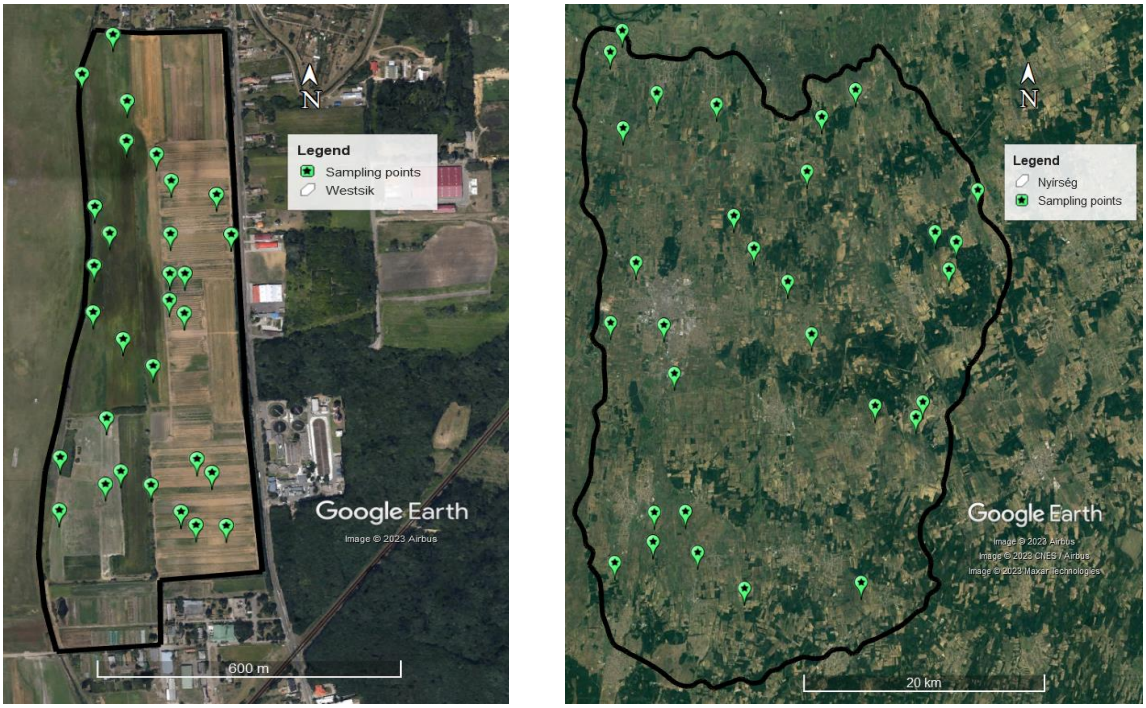



*Figure 6. Determined sampling location for Westsik (left) and Nyírség (right) by cLHS*

In the laboratory, all the soil samples were oven-dried for at least 48 hours, crushed, sieved to a size particle smaller than 200 µm, filled in plastic bags (100 g) and kept at room temperature (Figure 7). Bulk density was calculated by dividing the weight of samples after drying in the oven by the bulk volume of the cylinder at field moisture. SOC content was measured by the wet oxidation method (Davies, 1974). Then, SOC stock was obtained by multiplying the SOC and bulk density and soil thickness which was 10 cm in our case. A standard glass electrode in the 1:2.5 soil-water suspension measured electrical conductivity and pH. Soil carbonate content was determined by volumetric Scheibler calcimeter method. Each measurement for each property was repeated three times.

*Figure 7. Fieldwork, taking samples and laboratory measurements for soil organic carbon*

### 3.2.2. Case study two

In this study, we used a publicly available dataset from the ISRIC Africa Soil Profiles (AfSP) database (https://www.isric.org/projects/africa-soil-profiles-database-afsp). Nearly 18,000 soil profiles from multiple digital and analogue data sources, representing the major parts of Africa, are included in this database. $pH_{(H2O)}$, SOC and clay-content were selected as the target soil properties for modeling and mapping since these properties have a large enough sample size for all four countries. The chosen depth interval was 0 to 20 cm because we mainly focus on topsoil characteristics.

### 3.2.3. Case study three

In this study, the sampling design followed a regular grid, selecting 85 soil locations at a distances of 100 m from each other. Samples were taken by soil tubes containing undisturbed soil from the surface to a depth of 1 m. However, in this study I only used data for the topsoil up to 30 cm in depth (0-30 cm).

Samples were taken to the laboratory to measure SAS indicators, including pH, EC, and sodium adsorption ratio (SAR). pH and EC were measured in 1:2.5 soil-water suspension with a standard glass electrode. The concentrations of $Na^{+}$ and $Ca^{2+}$ +$Mg^{2+}$ in the saturation extract of the soil is required to calculate SAR expressed in millie equivalents per liter measured. It should be noted that I did not take part in the fieldwork and laboratory measurements for this study. These were carried out by the colleagues of the Institute for Soil Sciences, Centre for Agricultural Research, in the framework of a project funded by NKFIH[1].

## 3.3. Environmental covariates

A list of environmental covariates was applied for each of these studies; some of these covariates were calculated for three case studies, while others were specific to a particular case study, as presented in Table 1. All these covariates were selected based on the representation of soil-forming factors in related areas.

In study one, since the goal is to extrapolate to arable lands, we extracted the arable lands of Hajdúhát and Nyírség using the CORINE land cover map of 2018 (https://land.copernicus.eu/pan-european/corine-land-cover).

In case studies one and three, climate and land cover are homogeneous over the study areas. Therefore, only factors representing topography, organisms and soil surface were taken as the main soil forming factors.

For the first project, the source of DEM was EU-DEM in 30m resolution. EU-DEM is a hybrid product from the Shuttle Radar Topography Mission (SRTM) and ASTER DEM using a weighted averaging approach. Further details about EU-DEM can be found at the following

[1] More details about the project: Title: Optimization of the large-scale mapping of salt-affected soils under different land uses, Grant number: K124290.

link (https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem). For the second project, the source of the DEM was the SRTM digital elevation dataset 4 in 90m resolution (https://srtm.csi.cgiar.org/).

DEM applied in case study three, was surveyed and generated using an unpiloted aerial vehicle (UAV) in a fully automatic flight mode above the study area. The survey was conducted in March 2019. The field had no vegetation cover, as it was prepared for corn planting, and there had been no previous cover crop. Therefore, the UAV survey was conducted over a bare soil field. In terms of geo-transformation accuracy, both the horizontal and vertical accuracies of the photogrammetric image block were found to be below 5 cm. This calculation was based on 4 independent checkpoints using a 5-fold permutation approach, utilizing the 14 original ground control points.

The satellite images utilized for case study one were downloaded from the Landsat 8 satellite imagery source (https://earthexplorer.usgs.gov) in October 2019. This specific time frame was selected as it represented the most recent update accessible at that time and ensured the absence of vegetation cover.

For the first case study, we extracted the Normalized Difference Vegetation Index (NDVI), Carbonate Normalized Ratio, and Clay Normalized Ratio from the Landsat 8 band images (John et al. 2020). These indices were derived using the following formulas:

NDVI:

$$\frac{\text{NIR} - \text{Red}\ (5-4)}{\text{NIR} + \text{Red}\ (5+4)}$$

Carbonate Normalized Ratio:

$$\frac{\text{Red} - \text{Green}\ (4-3)}{\text{Red} + \text{Green}\ (4+3)}$$

Clay (hydroxyls) Normalized Ratio:

$$\frac{\text{SWIR(a)} - \text{SWIR(b)}\ (6-7)}{\text{SWIR(a)} + \text{SWIR(b)}\ (6+7)}$$

For the third case study, various spectral salinity indices (SI) were extracted from Sentinel-2 bands. One of these indices is SI-INDP, which is a salinity index derived from the Indo-Dutch Network Project (IDNP, 2002). Another index is SI-Aster, which is a salinity index derived

from ASTER (Bannari et al., 2008). Additionally, two intensity indices (Int 1, Int 2) were extracted to detect soil salinity levels.

Various topographic factors were extracted from the DEM to characterize the landscape for all case studies. The slope represents the degree of steepness at a particular location, while the aspect indicates the direction in which the slope faces. Plan curvature and profile curvature describe the horizontal and vertical curvature of the land surface, respectively. The LS factor combines slope steepness and length to estimate soil erosion rates. Valley depth quantifies the depth of low-lying areas, while positive and negative openness assess the visibility and openness above and below a certain elevation. The topographic wetness index evaluates the potential wetness of an area based on slope and contributing area. The terrain ruggedness index measures the roughness or variability in elevation within a defined area. The multiresolution index of valley bottom flatness and multiresolution ridge top flatness analyze the flatness of valley bottoms and ridge tops at different resolutions. Relative slope position determines the location of a point relative to its surroundings on a slope. Roughness characterizes the irregularity in elevation within a specific area. Flow direction identifies the direction of water or fluid flow across the land surface, while catchment area defines the land area draining into a specific point or stream. Lastly, modified catchment area considers adjusted criteria for contributing land (Mulder et al. 2011).

In the first and second case studies, the environmental covariates were calculated by me. However, in the third case study, colleagues at the Institute of Soil Science in Budapest calculated the environmental covariates.

All covariates had a different resolution but were resampled and harmonized to the final spatial resolution, which was 30 m, 1 km and 2 m for studies 1,2 and 3, respectively.

The DEM derivatives detailed in Table 1 were extracted using ArcGIS (http://www.esri.com) and SAGA GIS (Conard et al. 2015). All other covariate layers were introduced in R using the "*raster*" and "*sp*" package.

*Table 1. List of environmental covariates applied for each case study with their source and soil forming factor.*

| Factor | Environmental Covariate | Case study | Source |
|---|---|---|---|
| Climate | Annual mean temperature | 2 | [illegible] |

| | | | |
|---|---|---|---|
| | Annual precipitation | | |
| | Annual mean cloud cover frequency | | |
| | Annual mean water vapor pressure | | |
| | Annual mean solar radiation | | |
| | Global Reference Evapotranspiration | | |
| | Temperature seasonality (Standard deviation * 100) | | |
| | Precipitation seasonality | | |
| | Standard deviation of monthly water vapor pressure | | |
| | Standard deviation of monthly solar radiation | | |
| | Mean temperature of driest quarter | | |
| | Mean temperature of warmest quarter | | |
| | Precipitation of wettest month | | |
| | Precipitation of driest month | | |
| | Global Aridity Index | | |
| Vegetation | Normalized difference vegetation index (NDVI) | 1 | Landsat 8 |
| | Shannon ( Enhanced Vegetation Index ) | 2 | MODIS |
| | Soil adjusted vegetation index | 3 | Sentinel-2 |
| | Normalized difference vegetation index | | |
| | Vegetation soil salinity index | | |
| Topography | Elevation | 1,2,3 | DEM |
| | Slope | | |
| | Aspect | | |
| | Plan curvature | | |
| | Profile curvature | | |
| | LS factor | | |
| | Valley depth | | |

| | | | |
|---|---|---|---|
| | Positive openness | | |
| | Negative openness | | |
| | Topographic wetness index | | |
| | Terrain ruggedness Index | | |
| | Multiresolution index of valley bottom flatness | | |
| | Multiresolution ridge top flatness | | |
| | Relative slope position | | |
| | Roughness | 3 | |
| | Flow direction | | |
| | Catchment area | | |
| | Modified catchment area | | |
| | Mass balance index | | |
| Soil surface | Clay normalized ratio | 1 | Landsat 8 |
| | Carbonate normalized ratio | | |
| | MODIS RED long-term mean | 2 | Modis |
| | MODIS NIR long-term mean | | |
| | MODIS Green long-term mean | | |
| | MODIS SWIR 1 long-term mean | | |
| | MODIS SWIR 2 long-term mean | | |
| | Brightness index | 3 | Sentinel-2 |
| | Normalized difference salinity index | | |
| | Salinity index (SI) 1-5 | | |
| | Salinity ratio | | |
| | Visible infrared salinity index | | |
| | Green band | | |
| | Red band | | |
| | Near-infrared band | | |
| | Short-wave infrared-1 | | |
| | Short-wave infrared-2 | | |
| | SI1_IDNP | | |

| | | | |
|---|---|---|---|
| | SI_ASTER | | |
| | INT 1 | | |
| | INT 2 | | |

### 3.4. Machine learning models and geostatistics

Multiple linear regression (MLR) is one of the simplest regression techniques, which represents the linear relationship between several predictor variables and a continuous target variable (Andrews, 1974; James et al. 2013). Assumptions for MLR include 1) a linear relationship between the predictor and target variables, 2) normally distributed regression residuals, and 3) no correlation between predictor variables. There are no hyperparameters for MLR to adjust during training. MLR is often applied in spatial modeling and predictive mapping as a benchmark technique to compare with ML algorithms, which is expected to be the least accurate technique (Chagas et al. 2016). In reality, there is a correlation between different soil landscape elements, and this relation is not always linear (Forkuor et al. 2017). We compared MLR as a statistical approach with other ML models for the first case study in R software.

Random Forest (RF) is a tree-based algorithm which employs many decision or regression trees during the training process. RF is the most popular ML algorithm in DSM for both classification and regression tasks (Breiman, 2001; Hengl et al. 2018).

To reduce variance and improve the model's performance, RF employs bootstrap aggregation, also known as "bagging." The bootstrapping technique randomly chooses a collection of observations from the training dataset and then creates a decision tree linked to those selected observations. Instead of applying one decision tree, many trees in various classifiers for the given dataset are utilized by RF. Therefore, RF picks the prediction from each tree and, based on the majority votes of predictions, predicts the outcome (Ho, 1995). In most cases, two hyperparameters need to be fine-tuned to produce accurate results with RF; the number of decision trees "*ntree*" and number of covariates to split in each node for each decision tree "*mtry*". There are several advantages to applying RF, including; being capable of working with high dimensionality and large datasets, preventing overfitting, handling missing values

in the dataset, requiring less training time, and finally, performing with high accuracy. We applied this model in R using the "*randomForest*" and "*caret*" packages for the first and third case studies. In the first case study, we utilized *ntree* = 50 and *mtry* = 6, as they demonstrated the best performance at these specific values. In the second case study, we set the hyperparameters to their default values. Specifically, *ntree* was set to 100, and *mtry* was determined by default as one of the following: 2, 19, or 35. These values corresponded to the minimum, mean, and maximum number of covariates, respectively. Each model at each scenario was assigned one of these values based on the best results it achieved.

Artificial neural networks (ANNs) are another common algorithm in DSM that simulate biological neural networks (Behrens et al. 2005; Were et al. 2015). Information is transferred between neurons through multiple layers of the network. The network architecture is made up of many artificial neurons or processing elements connected by weights, which arrange layers and change parameters to learn from the data. Several ANN algorithms have been proposed; however, for this study, a feed-forward multilayer perceptron was chosen since it is the most frequently used type. The structure of this model consists of three layers: an input layer, one or more hidden layers, and an output layer. Each layer contains a collection of connected nodes (neurons) that function in parallel to convert the input data into output values (Lee and Evangelista, 2006; Ghaderi, 2019). The residual is calculated during the training stage by comparing the output with the input. After making adjustments to the connection weights and recalculating the residual, the algorithm goes back through the iterative propagation of errors. This procedure is repeated until the minimum residue is attained.

Changing the number of neurons, the number of hidden layers, the number of iterations, and the training type are parts of hyperparameters that can optimize the model's performance (Khaledian and Miller, 2020). We applied this model in R using the "*neuralnet*" package for the first case study. In this study, a feedforward neural network with two hidden layers (with 5 and 3 neurons, respectively) and a linear output layer was used.

Support Vector Machine (SVM) is a supervised learning model that has recently become more widely used in DSM (Kovačević et al. 2010; Were et al. 2015). For both linear and non-linear patterns, SVM categorizes the data using the best separating hyperplane. SVM keeps track of all covariates in order to separate or fit data linearly while defining a maximum

margin using support vectors, which are the observations. The margin is the separation between the decision surface and any observation, which is as far away as possible. Additionally, to address non-linear issues, this method uses kernel functions to map the non-linear correlation between features and high dimensional space (Gunn, 1998).

Cost (C) and gamma (γ) are two parameters in this model that have a significant impact on model performance (Tang et al., 2020). We utilized the "*e1071*" package in R to implement this model for the first case study. This package provides a function called "*tune*" that enables the exploration of various combinations of hyperparameters and evaluation of their performance. In the first case study, we defined ranges of possible values for these hyperparameters as cost = c(0.1, 1, 10) and gamma = c(0.5, 1, 2) for each property. The best values of cost and gamma for each property were determined and used to train the model.

We used the ensemble method for the third case study to model each SAS indicator. The ensemble consisted of five base learners; RF, SVM, ANN (which was explained above), and two more learners, including Extreme Gradient Boosting (XGBoost) and Generalized Linear Models with Lasso or Elastic Net Regularization (GLM).

XGBoost is a tree-based ML algorithm that performs based on the prediction error of the previous trees. This model can create more reliable and higher predictions by modifying and introducing a more precise approximation to the gradient boosting framework (Chen and Guestrin, 2016).

GLM technique is a generalized version of the linear regression model, the simplest and easiest predictive approach, as we described earlier. The difference between this model with MLR is computing the lasso or elastic-net penalty at a grid of values that can efficiently handle the cross-correlation between covariates and non-linearity (Friedman et al. 2010).

We applied the SuperLearner method using the "*Landmap*" package to stack all single learners. SuperLearner is an algorithm that applies cross-validation to evaluate how well one model, several models, or even the same model with various settings will perform. The model has been found to enhance prediction performance compared to using a single base learner when applied to various problems (Van der Laan et al. 2007).

To further clarify how the stacking of individual learners was accomplished, we utilized the comprehensive capabilities provided by the "*Landmap*" package. This package serves as a

fully automated benchmarking tool for predictive mapping tasks, offering a series of integrated steps from data preprocessing to model fitting, hyperparameter optimization, and validation with uncertainty assessment (Hengl et al., 2022). Within this package, the SuperLearner method leverages these functionalities to stack the individual learners effectively. It overlays the observations and covariates, performs feature selection, and fits each single learner model. It then combines the predictions from these models using an optimized approach to minimize the mean squared error. By stacking all the individual learners, it is ensure the integration of the strengths of each learner, resulting in enhanced prediction accuracy.

Multivariate geostatistics, namely regression cokriging (Stein and Corsten, 1991; Szatmári et al, 2020) combined with SuperLearner method as an ensemble of ML algorithms to predict the spatial pattern of SAS indicators in a salt-affected area. Since the indicators can be spatially interdependent or cross-correlated, it is preferable to model their spatial distribution jointly.

The following model describes the spatial variations in the SAS indicators, which are frequently applied in both DSM (Mcbratney et al. 2003) and multivariate geostatistical modeling (Wackernagel et al. 2003):

$$Z(u) = m(u) + \varepsilon(u) \quad (1)$$

where Z(u)= vector of the SAS indicators,

m(u)= vector of the deterministic component representing the spatial variations in the SAS indicators that can be described from the environmental covariates,

ε(u)= vector of the stochastic residuals that can be spatially correlated and cross-correlated,

u = vector of the geographical coordinates.

The variables must follow a normal distribution in applying kriging algorithms. Therefore, we normalized the variables in this study by normal score transformation for those that had non-normal distribution (Goovaerts, 1997).

Ensemble predictions for SAS indicators (pH, EC, and SAR) were taken to be the "*m*" in equation (1). Then, we computed the residuals by subtracting the ensemble predictions from the observed values for each SAS indicator at each sampling point in order to perform multivariate geostatistical modeling of "ε" in Equation (1). Afterwards, the variograms and

cross-variograms from the residuals were calculated, and a linear model of coregionalization (LMC) was fitted to make sure that the model was statistically valid (Goovaerts, 1997).
We also quantified the prediction uncertainty by compiling a 90% prediction interval for each SAS indicator. This can be achieved by adding and subtracting 1.64 times the kriging standard deviation from the prediction provided by regression cokriging. In the end, we transformed back the results for each SAS indicator that were normalized previously.

### 3.5. Similarity between study areas

The four selected countries of Africa (case study two) should be checked regarding similarities of soil forming factors. In our study, we used the homosoil approach to assess the similarity between two locations. This methodology, developed by Mallavan et al. (2010), utilizes Gower's similarity index to determine the similarity between environmental covariates in the donor and recipient areas. Gower's similarity index is a statistical measure that determines the similarity between two objects based on their attributes. The index ranges between 0 and 1, with higher values indicating greater similarity (Gower, 1971). In the homosoil approach, calculation of Gower's similarity index involves three hierarchical steps; first, by choosing areas with similar climate conditions that have a similarity index larger than 0.85 (homoclimes); second, by selecting through all homoclimes the areas that have similar lithological classes (homoliths); and third, selecting areas with same topography condition from all homoliths areas (homotop). This can be done for any location in the world by the global-scale spatial database for climate, topography, and lithology covariates prepared by Mallavan et al. (2010). We assumed that if two locations are similar regarding these three factors, then they probably have similar soils.
Thus, we used this method to identify similarities in soil-forming factors between the donor and recipient countries. Then, we defined and computed the "homosoil fraction", which is the fraction of the recipient's surface, which is homosoil to at least one location in the donor country.
At the same time, taking into account the pedological knowledge about the soil types also can be an effective tool in detecting similar soils since a certain set of features form a unique soil type (Blum et al. 2017). Therefore, the main soil types in each country based on WRB

classification were extracted from Soil Atlas of Africa (Panagos et al. 2012; Jones et al., 2013). Then, to quantify the similarity between the different soil types, we used the Jaccard measure of similarity, a statistical measure commonly used to assess the similarity of two or more sets of data (Awad and Khanna, 2015). The Jaccard measure was calculated based on the coverage of each soil type in each country, allowing us to identify regions with similar soil types.

Another approach to see how different a new location in the recipient area is from the locations in the donor area is the dissimilarity index (DI) of the covariates. DI can be calculated by the AOA function implemented in the CAST package (CAST: Caret Applications for Spatio-Temporal models) in R. This method needs two datasets; one contains training data (soil samples and environmental covariates) in the donor area, and the second contains new locations in the recipient area (environmental covariates). DI is a unitless index that estimates the minimum Euclidean distance of the closest training data point to the average of the distances in the training data.

DI values range from 0 to infinity, where 0 indicates that the new location has the same covariates as the training data point in the model, and an increase in DI corresponds to an increased distance from the nearest training data point. The other byproduct of this function is the AOA layer which derives based on identifying a threshold. AOA has only two values; 0 and 1, where 0 represents a new location outside of AOA, and one represents the point inside the AOA.

The training samples are represented in a covariate space with multiple dimensions. Prior to analysis, the covariates are appropriately scaled and weighted. Initially, the average distance between all training data points is computed. Subsequently, for each training data point the distance to the nearest training data point outside of the same cross-validation fold is evaluated (assuming a threefold cross-validation in the visualization). This distance is then divided by the average distance between all training data points to obtain the DI. The DI is calculated for each training data point, and the threshold for the AOA is established using the upper whisker of the DI values presented in a boxplot. When dealing with a new data point, the DI is computed accordingly, and if DI is lower than the threshold, the point will fall inside

AOA, and if it is larger than the threshold, the point will be outside of AOA (Meyer and Pebesma, 2021).

Uncertainty of prediction in the second case study was done by quantile regression forest (QRF). A 90% prediction interval width (PIW) was computed from the difference between 0.95 (upper) and 0.05 (lower) quantiles. By applying this method, we can check if it agrees with the locations determined by AOA.

### 3.6. Validation

For the first phase of the first case study, the goal is to compare the capability of ML algorithms in the spatial prediction of soil properties. The validation strategy was based on splitting the dataset into train and test, in which 70% of the data were used for training and 30% for testing the model. The validation metrics for both study areas (Látókép and Westsik) were the determination coefficient (R-square), and root mean square error (RMSE). After selecting the best model with the highest accuracy and least error, validation for the second phase was done based on cross-validation to take all observations in the training dataset. Also, to ensure that model did not over fit and improve model performance, we trained the model with feature forward selection by *"cast"* package in R.

Also, cross-validation was employed for the second and third case studies to evaluate and compare the spatial prediction performance using 10-fold. In this method, the dataset was randomly split into ten folds of the same size. Each time nine-fold was used for calibration and one-fold for validation, and this process was iterated ten times until all these ten-fold were taken in the validation set.

Mean error (ME), RMSE, Lin's concordance correlation coefficient (CCC) (Lin, 1989), and model efficiency coefficient (MEC) (Nash, J. E., & Sutcliffe, 1970) are the four validation metrics that were calculated between observations and predictions as follows;

$$ME = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (3)$$

$$MEC = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (4)$$

$$CCC = \frac{2\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sum_{i=1}^{n}(O_i - \bar{O})^2 + \sum_{i=1}^{n}(P_i - \bar{P})^2 + n(\bar{O} - \bar{P})^2} \quad (5)$$

where Oi and Pi are the observations and predictions for the location i, respectively, and $\bar{O}$ and $\bar{P}$ are the mean of the predictions and observations.

To assess the effectiveness of uncertainty quantifications, accuracy plots and G statistics were utilized. An accuracy plot, or prediction interval coverage probability plot, is based on the idea that if an uncertainty quantification provides a prediction interval with a specified width (e.g. 90%), then it is expected that 90% of the observations from the validation dataset should fall within this interval. This concept can be applied to symmetric prediction intervals of any width. As a result, an accuracy plot can visually display the proportion of observations from the validation dataset that falls within symmetric prediction intervals of different sizes. A perfectly accurate plot would ideally align with the y=x line, and the G statistics can be used to determine how closely the accuracy plot fits this line:

$$G = 1 - \int_0^1 |\xi(p) - p| dp \quad (6)$$

where the proportion of observations (represented by $\xi$(p)) and the width of the prediction interval (represented by p). The ideal outcome is a G statistics value of 1.

### 3.7. Workflow of case studies

We summarized the methodology taken for three case studies as a workflow in Figure 8. For case study one, sampling design was done by cLHS. Thirty samples were collected from the surface of the Látókép, Westsik, Hajdúhát and Nyírség study areas. Samples were taken to the laboratory and soil properties including bulk density, soil organic carbon, pH, EC and carbonate were measured. First, we trained four models (MLR, RF, ANN, SVM) for Látókép and Westsik and the best model was chosen. We fine-tuned the model and applied the trained model of Látókép to extrapolate over Hajdúhát and the trained model of Westsik to extrapolate over Nyírség. Also the AOA was applied over these areas. We validated the points in Hajdúhát and Nyírség with samples taken from there.

For case study two, the data for four African countries including Ethiopia, Kenya, Burkina Faso, and Nigeria were extracted from publicly available dataset. First, we trained the RF model for each country and each property with default hyperparameters values, and predicted that country and the other three countries each time. Therefore, we had 12 scenarios. We checked how similar they are in terms of soil types, soil forming factors by homosoil approach, and dissimilarity index by AOA. Quantile regression forest (QRF) was applied to check which locations have high uncertainty. We validated the prediction points by observation in each country.

For the third study: an ensemble modeling approach was used with five individual models (RF, XGboost, SVM, ANN, and GLM) on three indicators of salt-affected soils. We applied the SuperLearner method to stack all single learners. Regression co-kriging was performed on the stochastic residuals obtained from the ML model. Afterwards, the variograms and cross-variograms from the residuals were calculated, and a linear model of coregionalization (LMC) was fitted. The prediction uncertainty was quantified by compiling a 90% prediction interval for each SAS indicator.

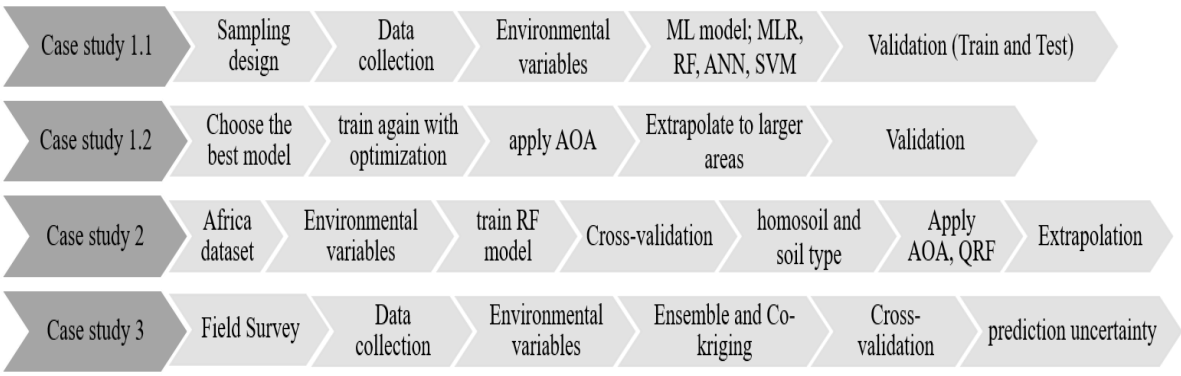



*Figure 8. Workflow of the methods and approaches for the three case studies.*

## 4. Results

### 4.1. Case study one

Soil surface characteristics and taxonomic position of soil show remarkable differences between Látókép and Westsik. The summary statistics of soil properties for both areas are

presented in Table 2 and 3. The average SOC concentration of surface samples was almost two times higher in Látókép, with a remarkably smaller standard deviation than in Westsik. The bulk density of surface soil samples is slightly higher in Westsik due to the sandy texture. However, standard deviations are pretty similar in both sites, thanks to the similar techniques applied in plowing. Both areas have normal EC and pH values.

The carbonate content of the surface in Westsik was almost triple that in Látókép. The texture of surface soil samples in Látókép was silt loam in all samples. In contrast, the texture of the Westsik site has wider variability from loamy sand to sandy loam at the surface, but also to sandy clay loam and clay loam in deeper soil horizons. Additional information on soil classification from our previous study is also presented in Table 4 (Hateffard et al. 2022). Soil profile classification showed higher soil variability in the Westsik station. The soil profiles in Látókép belonged to Chernozems and Kastanozems, and soils in the Westsik were classified into Arenosols, Phaeozems and Gleysols.

In Látókép, soil spatial variability was in the aggregation type and topsoil horizon, distinguishing the Kastanozems with coarse aggregates (medium subangular blocky) from the Chernozems with fine subangular blocky structure. At the level of the qualifiers, the vertical position and development of secondary carbonate accumulation (Endocalcic-Amphicalcic, or just protocalcic properties, therefore Haplic), and the thickness of the humus horizon (Pachic, >50 cm) showed differences. In Westsik station, based on the higher organic content of topsoil, darker color and well-developed aggregates sites were classified as Phaeozems, contrary to poor-developed surface horizons with paler color at elevated locations classified as Brunic Arenosols. The darkest surface colors showed the Chernic Gleysols connected with the deepest landforms.

*Table 2. Summary statistics of soil properties in Látókép*

| Property | Unit | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| SOC | % | 1.29 | 2.33 | 1.76 | 0.24 |
| BD | g/cm3 | 1.12 | 1.46 | 1.28 | 0.10 |
| SOC stock | ton/ha | 16.71 | 31.72 | 22.60 | 3.11 |
| pH | - | 5.36 | 8.13 | 6.52 | 0.66 |
| EC | dS/m | 0.05 | 0.63 | 0.21 | 0.16 |
| Carbonate | % | 2.32 | 7.35 | 3.25 | 1.06 |

*Table 3. Summary statistics of soil properties in Westsik*

| Property | Unit | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| SOC | % | 0.29 | 4.15 | 0.91 | 0.40 |
| BD | g/cm3 | 1.15 | 1.56 | 1.38 | 0.09 |
| SOC stock | ton/ha | 4.56 | 22.9 | 11.07 | 5.32 |
| pH | - | 4.59 | 8.26 | 6.66 | 1.36 |
| EC | dS/m | 0.02 | 0.33 | 0.15 | 0.10 |
| Carbonate | % | 2.57 | 30.2 | 8.34 | 6.46 |

*Table 4. Soil classification according to WRB (2015) of representative soil profiles in Látókép and Westsik stations (Hateffard et al. 2022)*

| **Study site** | **Principal Qualifiers** | **Reference Groups** | **Supplementary Qualifiers** |
|---|---|---|---|
| **Látókép** | Endocalcic | Chernozems | Aric, Pachic, Siltic |
| | Amphicalcic | | Aric, Siltic |
| | Haplic | | Aric, Siltic |
| | Endocalcic | Kastanozems | Aric, Siltic |
| **Westsik** | Lamellic – Brunic | Arenosols | Aric |
| | Calcaric | Phaeozems | (Anthromollic, Pantoarenic, Areninovic) |
| | Calcaric | | (Anthromollic, Anoarenic, Endoloamic, Aric) |
| | Pantocalcaric – Chernic | Gleysols | (Pantoloamic, Aric) |

### 4.1.2. Variable importance in Látókép and Westsik

After training the RF model, the variable importance for each soil property in Látókép and Westsik are presented in Figures 9 and 10. In Látókép, plan curvature and valley depth were the most important variables in the spatial variation of SOC and SOC stock. Multiresolution index of ridge top flatness, elevation and NDVI in determining BD, profile curvature and NDVI in determining carbonates play more important roles. Negative openness and clay index are the most influential in the spatial variation of pH, while plan curvature has a minor effect. Also, for EC prediction, multiresolution index of ridge top flatness and slope are the most important ones. The topographic wetness index has less power in this area to predict the spatial variation of bulk density, EC and carbonates, despite pH and SOC, which ranked

fourth and fifth, respectively. In Westsik, the most influential variable in determining the spatial variation of soil properties is NDVI (Figure 10). Other parameters have poor effects on the explanation of soil properties. Only for prediction of bulk density, elevation and carbonate index showed a significant contribution. Some predictors such as multiresolution index of ridge top flatness, aspect, valley depth, and deviation from mean value have fruitless effect.

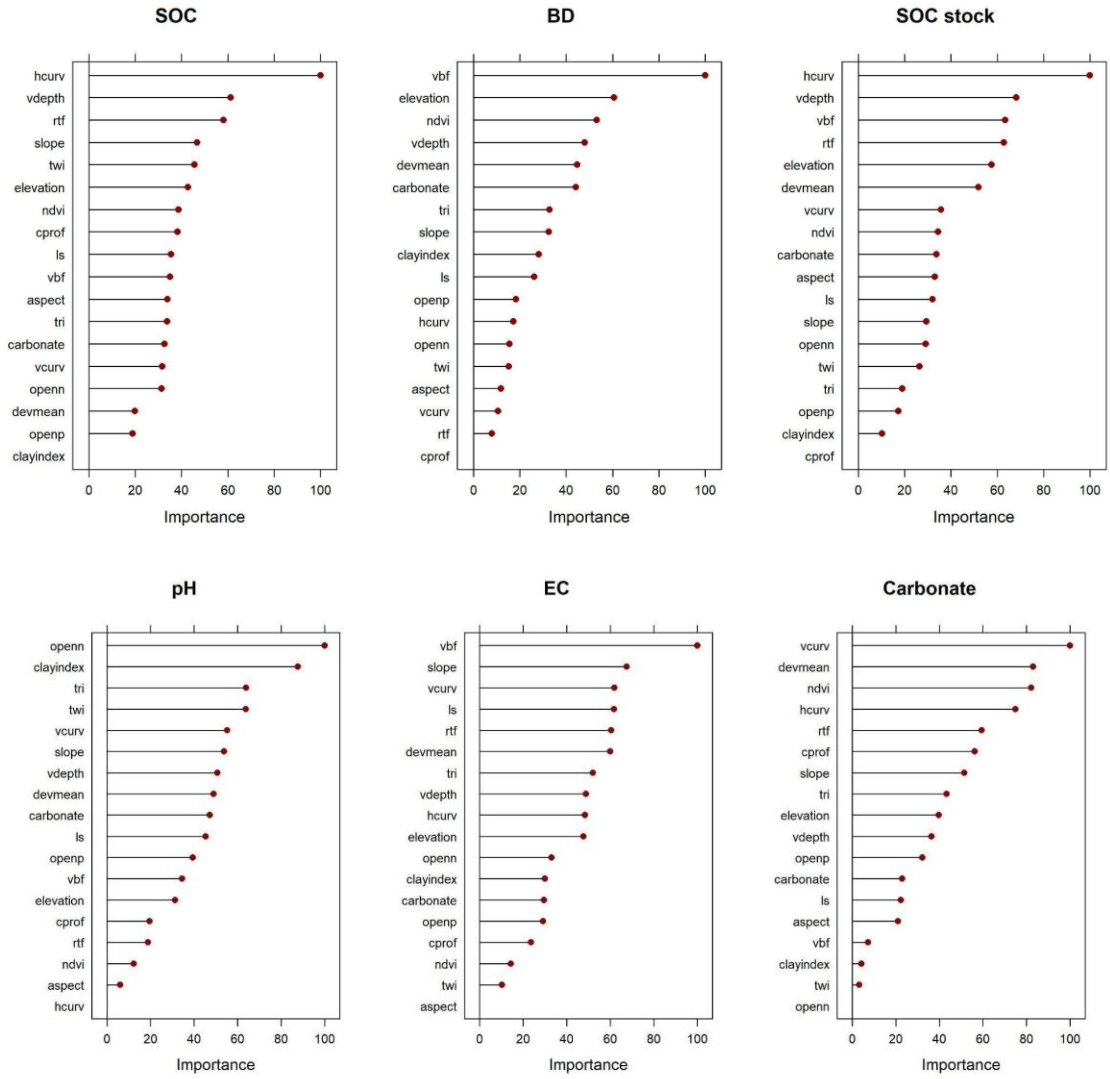



*Figure 9. Variable importance based on RF for each properties in Látókép. Abbreviations: hcurv and vcurve: plan and profile curvature, ls: ls-factor, opennessp and opennessn: positive and negative openness, tri: terrain ruggedness index, vbf: multiresolution index of valley bottom flatness, rtf: multiresolution index of ridge top flatness, twi: topographic*

*wetness index, vdepth: valley depth, devmean: deviation from mean value, carbonate: carbonate normalized ratio, NDVI: normalized difference vegetation index*

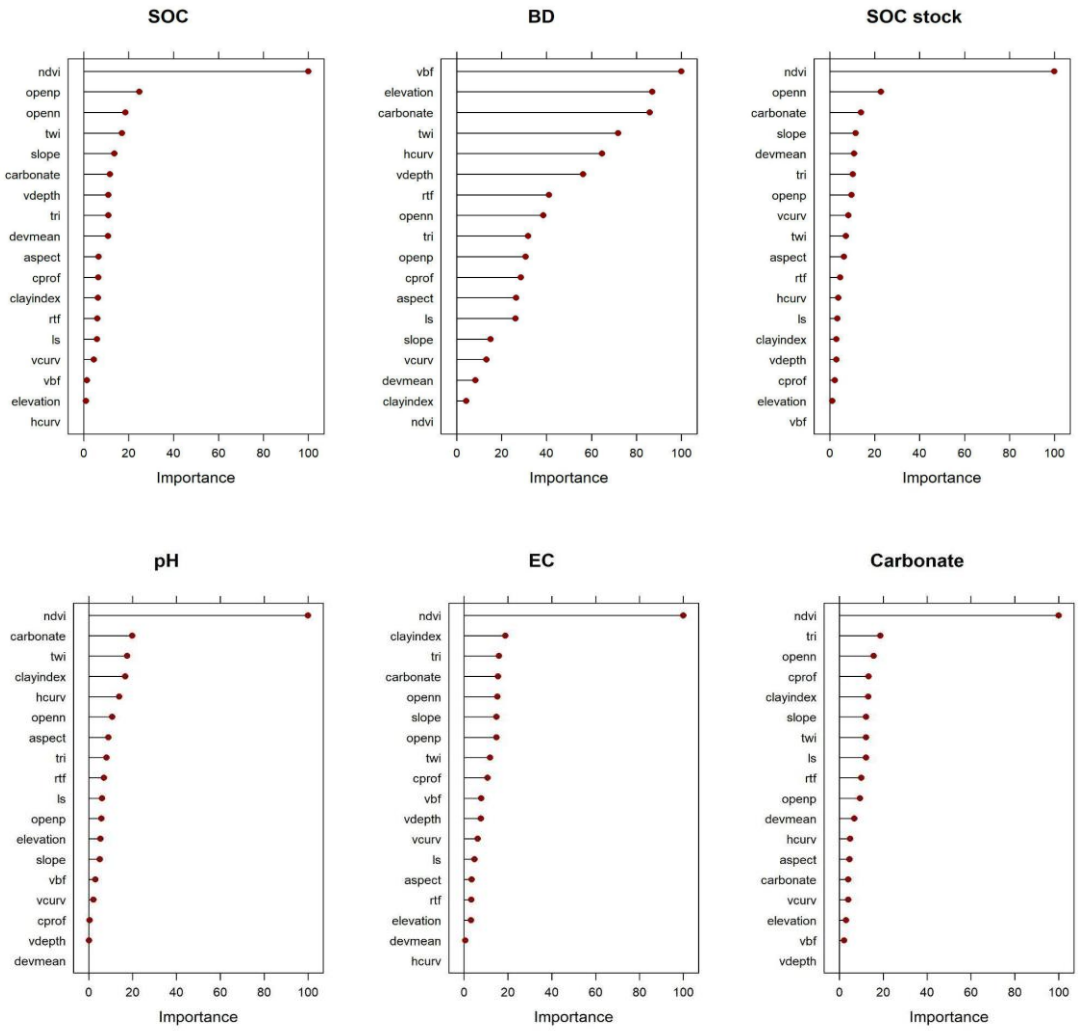



*Figure 10. Variable importance based on RF for each property in Westsik, same abbreviation in Figure 9.*

### 4.1.3. Comparison of machine learning models

In the first phase of Case study 1, the data was split into training and test sets. The score on the validation metrics for each ML model are presented in Table 5 and 6. The proposed ML algorithms showed different capabilities to predict soil properties at unsampled locations. MLR, applied in this study as a benchmarking approach, showed the highest $R^2$ with the least RMSE in the training dataset and unsatisfactory performance in the test dataset.

In Látókép (Table 5), the second-highest accuracy for training was ANN for these soil properties; BD ($R^2$ = 0.94), pH ($R^2$ =0.91), EC ($R^2$ =0.97) and carbonate ($R^2$ =0.97), except SOC ($R^2$ =0.51) and SOC stock ($R^2$ =0.14). RF, followed by SVM for SOC ($R^2$ =0.81, 0.61) and SOC stock ($R^2$ =0.90, 0.77), respectively, performed better in training the model than ANN. Also, considering RMSE, ANN performs the worst in SOC and SOC stock, while MLR, RF, and SVM values are in the same range. For other properties, the training results for RF were around $R^2$ = 0.7 ~ 0.9, and for SVM were between $R^2$ = 0.3 ~ 0.7.

Assessing how well these models functioned with the test dataset, RF outperformed other models by delivering around 80 % of all soil properties variability except soil carbonate, which was around 55 %. Appealingly, the $R^2$ values for ANN were negative in most cases except in EC. It means that ANN performed even worse than if we only took the average of the target values as our predictions. In the case of EC, ANN has $R^2$ and RMSE 0.51 and 0.12, which is acceptable. In addition, the best performance of SVM is related to soil pH, which explains 33 % spatial variability with RMSE 0.34, while the results showed negative values for EC predicted by SVM. In the prediction of SOC stock by SVM, the $R^2$ is almost equal to zero, which means the same results when considering only the average of target values.

*Table 5. Summary of ML algorithms on the train and test datasets in Látókép.*

| Train Dataset | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | BD | | SOC | | SOC stock | | pH | | EC | | Carbonate | |
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| MLR | 0.99 | 0.01 | 0.95 | 0.05 | 0.96 | 0.69 | 0.94 | 0.16 | 0.93 | 0.04 | 0.94 | 0.28 |
| RF | 0.80 | 0.05 | 0.81 | 0.10 | 0.90 | 1.04 | 0.87 | 0.25 | 0.77 | 0.07 | 0.80 | 0.53 |
| ANN | 0.94 | 0.03 | 0.51 | 0.30 | 0.14 | 3.10 | 0.91 | 0.20 | 0.97 | 0.02 | 0.96 | 0.21 |
| SVM | 0.61 | 0.07 | 0.68 | 0.14 | 0.77 | 1.58 | 0.88 | 0.24 | 0.35 | 0.12 | 0.66 | 0.69 |
| Test Dataset | | | | | | | | | | | | |
| MLR | 0.06 | 0.86 | 0.05 | 0.56 | 0.04 | 4.36 | 0.01 | 1.27 | 0.01 | 0.38 | 0.03 | 1.50 |
| RF | 0.89 | 0.04 | 0.82 | 0.10 | 0.88 | 0.74 | 0.81 | 0.12 | 0.81 | 0.07 | 0.55 | 0.16 |
| ANN | -0.75 | 0.14 | -0.46 | 0.30 | -0.27 | 2.43 | -0.33 | 0.35 | 0.51 | 0.12 | -0.7 | 0.31 |
| SVM | 0.19 | 0.11 | 0.27 | 0.26 | 0.09 | 2.05 | 0.33 | 0.34 | -0.15 | 0.18 | 0.11 | 0.37 |

The results for Westsik (Table 6) in the training dataset showed that MLR has the highest $R^2$, followed by RF for some properties and ANN for others. Similar results were obtained between ANN, RF and SVM for training soil carbonates and pH, in which $R^2$ were more than

90% in all cases. The $R^2$ values indicate that the RF, SVM and ANN models deliver 85%, 74% and 45% of SOC stock variability, respectively. In calibrating the model for BD and EC, the results for RF and ANN showed a similar ability to predict ($R^2$ ~ 0.7 and RMSE ~ 0.04 for BD, $R^2$ ~ 0.9 and RMSE ~ 0.02 for EC).

Likewise, the RF technique performed more effectively and achieved the best R2 (0.8) and the least RMSE for all soil properties in the test dataset. The SVM model for some properties, such as BD and SOC stock, could not predict better than the mean dataset. At the same time, it performed sufficiently for other properties as the R2 for SOC, pH, EC and carbonate were 0.30, 0.42, 0.35 and 0.4, respectively. Also, ANN showed a negative R square and high RMSE for BD, SOC stock, EC and carbonates while well-performed for spatial prediction of soil pH delivering around 40 %.

*Table 6. Summary of the ML algorithms on the train and test datasets in Westsik.*

| Train Dataset | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | BD | | SOC | | SOC stock | | pH | | EC | | Carbonate | |
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| MLR | 0.97 | 0.02 | 0.99 | 0.03 | 0.93 | 1.23 | 0.99 | 0.11 | 0.98 | 0.01 | 0.98 | 0.46 |
| RF | 0.79 | 0.04 | 0.88 | 0.29 | 0.85 | 1.83 | 0.95 | 0.27 | 0.92 | 0.02 | 0.94 | 0.96 |
| ANN | 0.76 | 0.04 | 0.99 | 0.04 | 0.45 | 5.78 | 0.98 | 0.13 | 0.94 | 0.02 | 0.97 | 0.68 |
| SVM | 0.59 | 0.06 | 0.52 | 0.58 | 0.74 | 2.45 | 0.93 | 0.34 | 0.85 | 0.03 | 0.94 | 1.58 |
| Test Dataset | | | | | | | | | | | | |
| MLR | 0.24 | 0.52 | 0.02 | 1.86 | 0.14 | 20.96 | 0.05 | 1.74 | 0.04 | 0.16 | 0.06 | 10.57 |
| RF | 0.82 | 0.05 | 0.80 | 0.17 | 0.86 | 2.12 | 0.86 | 0.47 | 0.93 | 0.02 | 0.89 | 1.82 |
| ANN | -0.17 | 0.12 | 0.21 | 0.33 | -0.74 | 7.36 | 0.39 | 1.0 | -0.31 | 0.1 | -0.24 | 6.21 |
| SVM | 0.01 | 0.11 | 0.30 | 0.31 | 0.02 | 5.53 | 0.42 | 0.97 | 0.35 | 0.07 | 0.4 | 5.46 |

The spatial prediction maps for each property in Látókép and Westsik are presented in Figures 11 and 12. In both areas, RF demonstrated more detailed information on the spatial distribution of soil properties. As we can see, the prediction range for MLR in all soil properties was too extensive, and even some properties delivered negative values, which is unrealistic. For example, SOC and BD were negative values, which is impossible. ANN and SVM could explain the general pattern for each study area and have worthy information. Based on RF maps, the southern part of Látókép depicted higher organic carbon and bulk density and small patches of high carbonates with pH values over 7.5. In Westsik, the eastern

parts of the station revealed lower SOC and SOC stock with less pH, EC and Carbonate values.

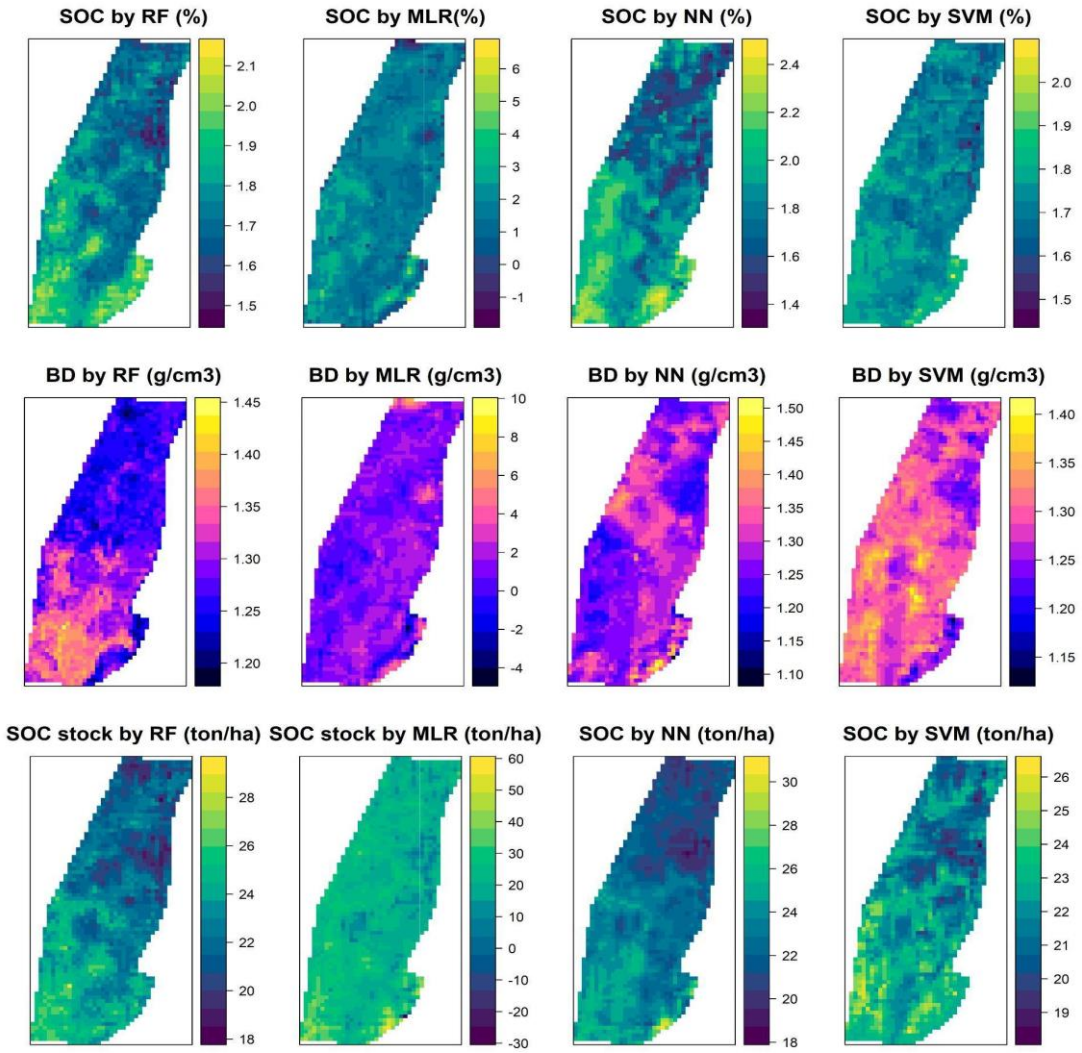



*Figure 11. Final maps predicted by each model in Látókép.*

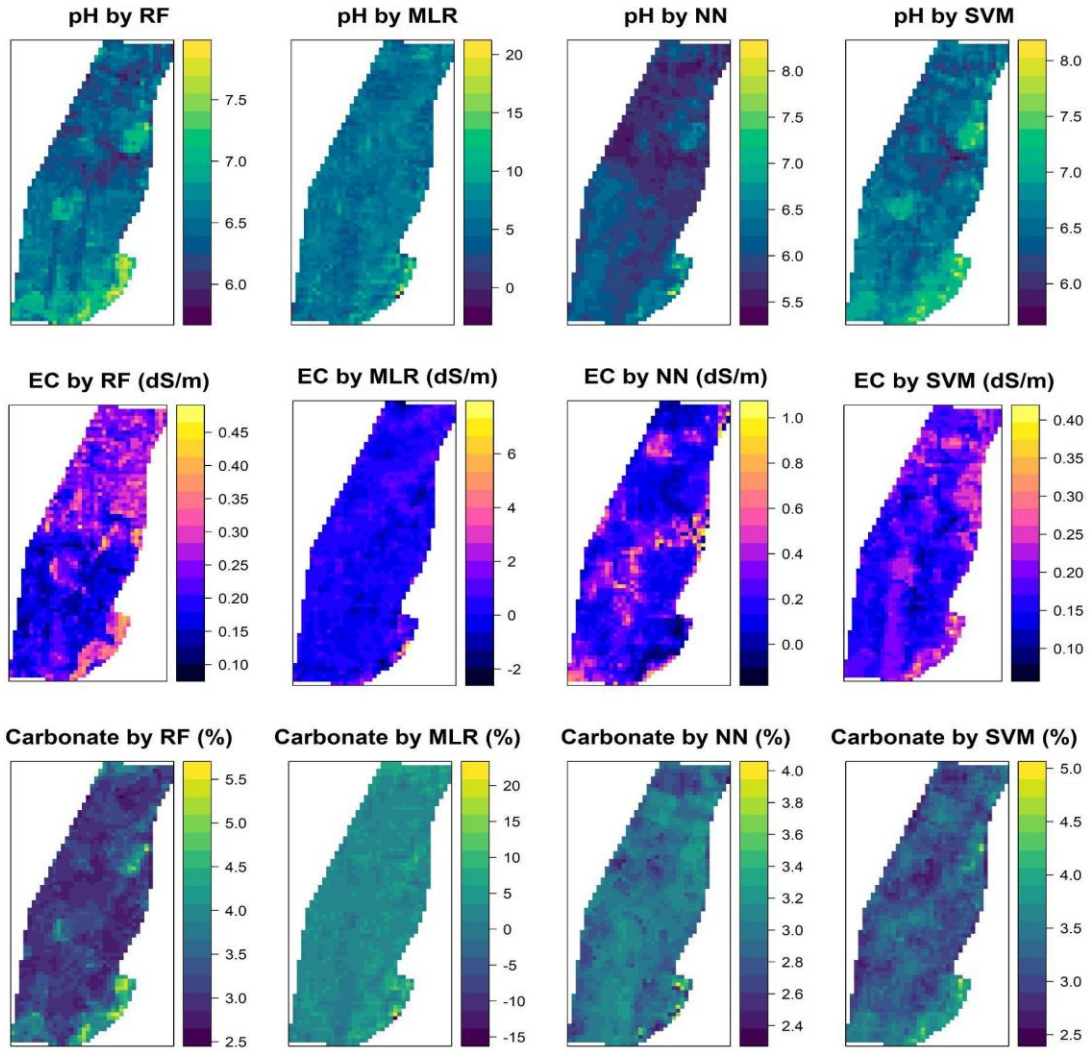



*Figure 11 (continue). Final maps predicted by each model in Látókép.*

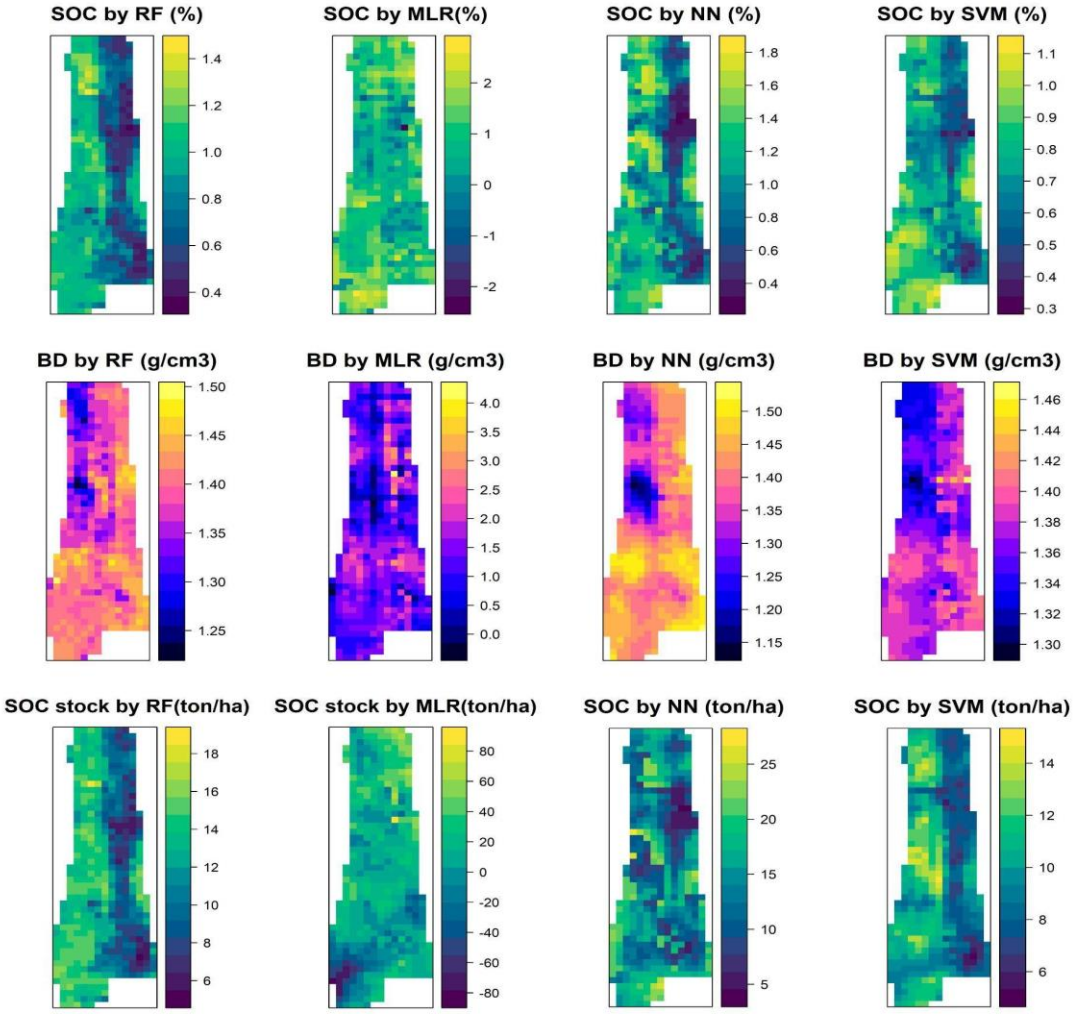



*Figure 12. Final maps predicted by each model in Westsik.*

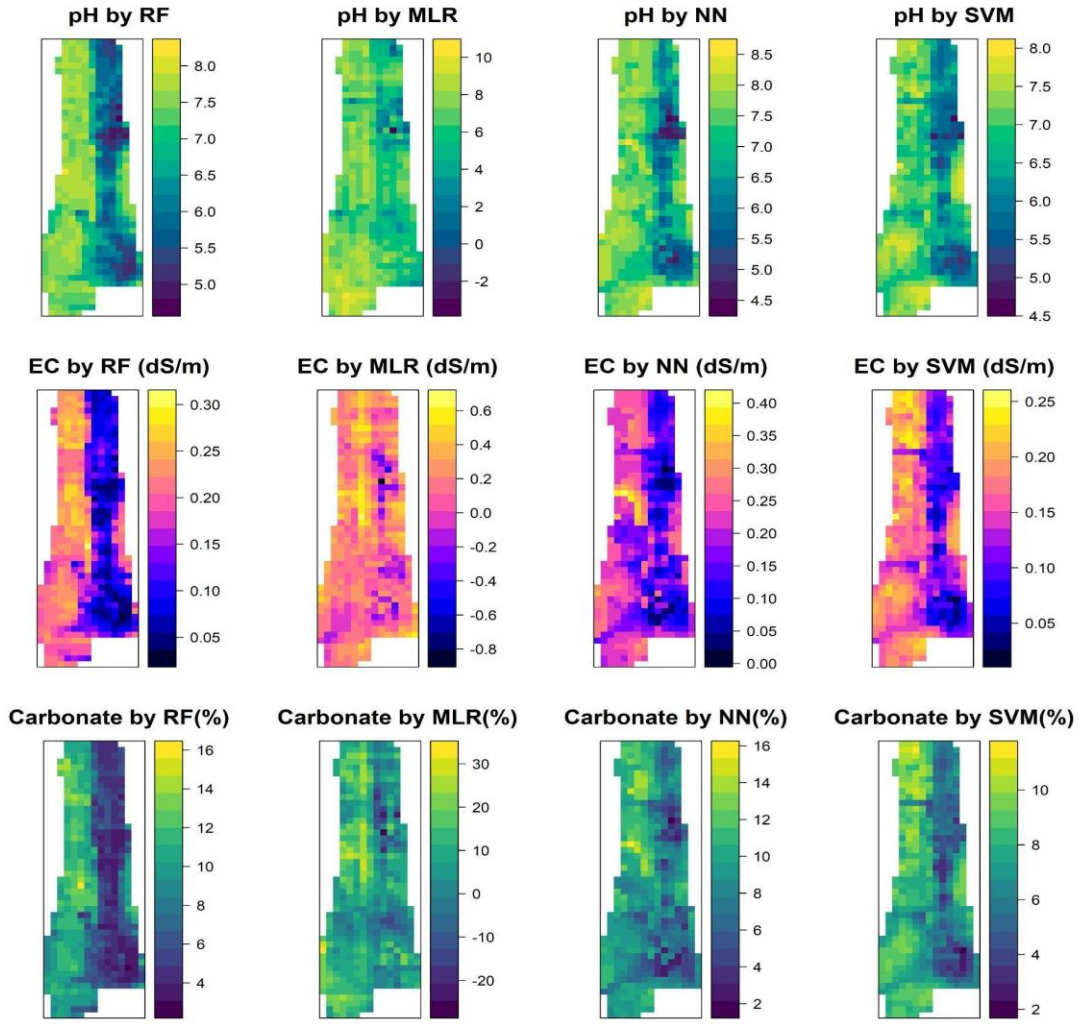



*Figure 12 (continue). Final maps predicted by each model in Westsik.*

### 4.1.4. Extrapolation and Area of Applicability

The prediction results for both areas, Látókép and Westsik, confirmed that RF outperformed other ML algorithms. Therefore, RF with feature selection is employed to train the model in Látókép for extrapolation purposes over Hajdúhát, and in Westsik for extrapolation over Nyírség. In addition, the AOA method was also applied for each area based on the trained model. A summary of dissimilarity index values for each property is presented in Tables 7 and 8. DI can take values from 0 to infinity. Hence, the minimum for all DI layers is zero. Increasing the DI values means the diversity between environmental covariates in donor and recipient areas increases. The DI values are over threshold (threshold is calculated as the

upper whisker of DI values, as explained in section 3.5) placed outside of the applicability area. The prediction maps for each property masked by the deriving area of applicability are presented in Figures 13 and 14. The gray areas in each map show the areas that are outside of AOA. As we can see in Tables 7 and 8, each property, depending on their relation with environmental covariates in the area, has different values of DI, showing different possible areas to extrapolate.

In Hajdúhát, the maximum DI between soil properties has been recognized for soil carbonates (DI = 102.5), while the least is related to SOC (1.04). Other properties such as BD, SOC stock and EC have maximum DI of around 11, while soil pH has a higher value (16.62). Extrapolating the ML model to another area will function well when the DI values get close to zero. The spatial average of DI for BD (0.22) and carbonate (0.2) are lower than other properties, while SOC stock is highest (0.48). The threshold calculated in soil EC is larger than others (2.43). The prediction map (Figure 13) for SOC showed that some areas of the southern part are outside of AOA. For BD, pH and Carbonate, these values are small patches the size of two or more pixels, which might not be visible in these figures. However, there are large areas in the northern part of Hajdúhát, which are too different from what the trained model for SOC stock and EC in Látókép has seen.

*Table 7. Summary of spatial dissimilarity index values for Hajdúhát*

| Property | Mean | Max | Threshold |
|---|---|---|---|
| SOC | 0.28 | 1.04 | 0.67 |
| BD | 0.22 | 11.44 | 0.51 |
| SOC stock | 0.48 | 11.74 | 1.48 |
| pH | 0.40 | 16.62 | 0.78 |
| EC | 0.45 | 10.79 | 2.43 |
| Carbonate | 0.2 | 102.5 | 0.43 |

Assessing the dissimilarity between environmental covariates in Westsik and Nyírség (Table 8), the spatial average of BD (19.52) is significantly higher than others, around 0.1 ~ 0.2. The highest dissimilarity also is discovered in BD (Max= 46.4), while the highest identicality to the donor area is related to soil EC (Max= 2.43).

The spatial prediction map for BD demonstrated that large parts of the area are outside of AOA and only a few parts of the southern parts are within AOA. The spatial predictions for

other properties are generally located inside AOA, meaning it is possible to extrapolate over these areas except for a few small patches (Figure 14).

*Table 8.Summary of spatial dissimilarity index values Nyírség*

| Property | Mean | Max | Threshold |
|---|---|---|---|
| SOC | 0.12 | 9.11 | 0.28 |
| BD | 19.52 | 46.4 | 26.4 |
| SOC stock | 0.11 | 6.99 | 0.25 |
| pH | 0.22 | 7.50 | 0.53 |
| EC | 0.14 | 2.43 | 0.34 |
| Carbonate | 0.16 | 3.77 | 0.29 |

After extrapolating over Hajdúhát and Nyírség, the predictions were validated based on sampling taken from these areas. The sampling collection was done before applying any model. Therefore, those points that fell outside of AOA were excluded from validation. For each area, 30 points were collected. The results of validation of the points inside and outside of AOA are briefed in Tables 9 and 10. We do not expect a large error for points outside of AOA since the number of samples to validate might be few. The ME values should generally be near zero; otherwise, the measurements would be biased. In Hajdúhát (Table 9), the ME values for the prediction of SOC, BD and EC inside the AOA are unbiased. Their prediction outside of AOA mainly remained the same. For soil carbonate, the ME is =0.82, which is biased, and considering the RMSE (1.17), the contribution of systematic error in predictions is considerable. Therefore, the predictions might have high errors, although all the points fell inside the AOA by chance.

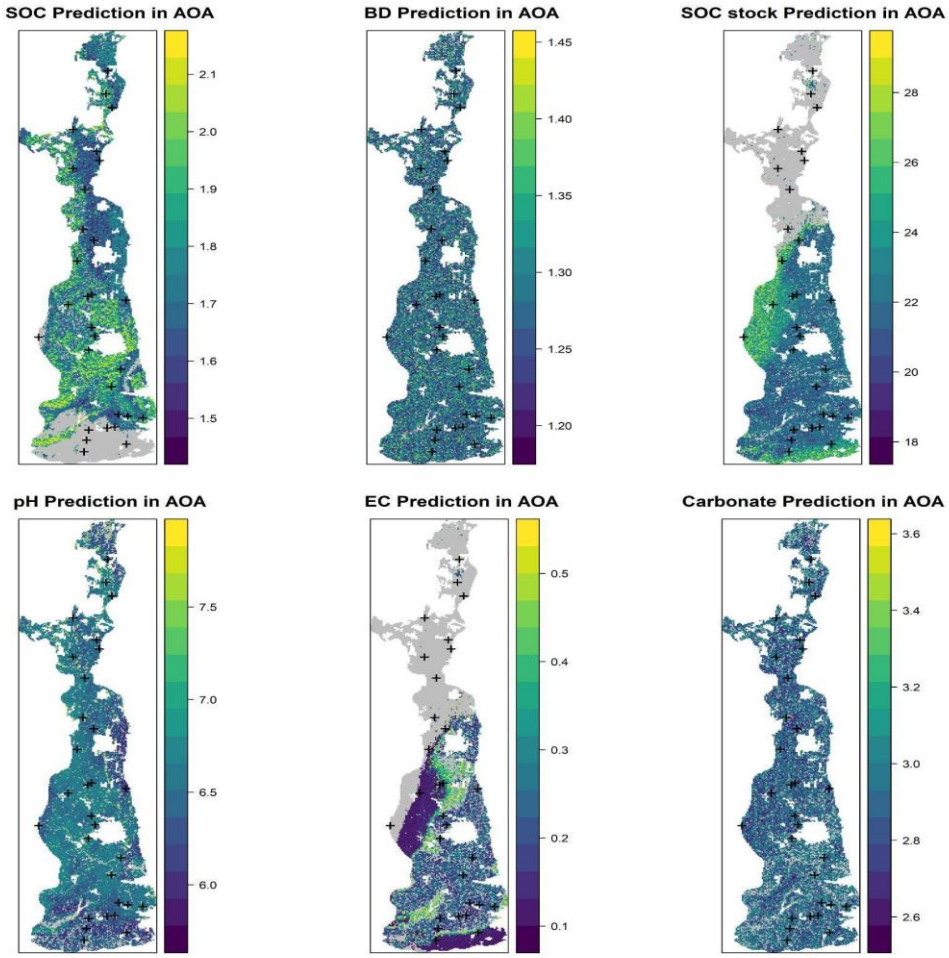



*Figure 13. The predictions masked by the derived area of applicability for Hajdúhát, where the areas outside the area of applicability are shown in gray. Also, the location of sampling points for validation are presented. The unit for each soil property is the same with Figure 11.*

When AOA applied for SOC stock, one-third of the sampling points fell outside applicable areas. By checking the validation inside and outside of AOA for SOC stock, it is visible that the ME (-4.83) and RMSE (6.59) significantly increase outside of AOA compared to the values inside AOA (ME = -0.23 and RMSE = 2.41). The soil pH validation metric showed that the points inside AOA, which are 26, are overestimated (ME = 0.40), while those outside AOA are underestimated (ME = -0.40).

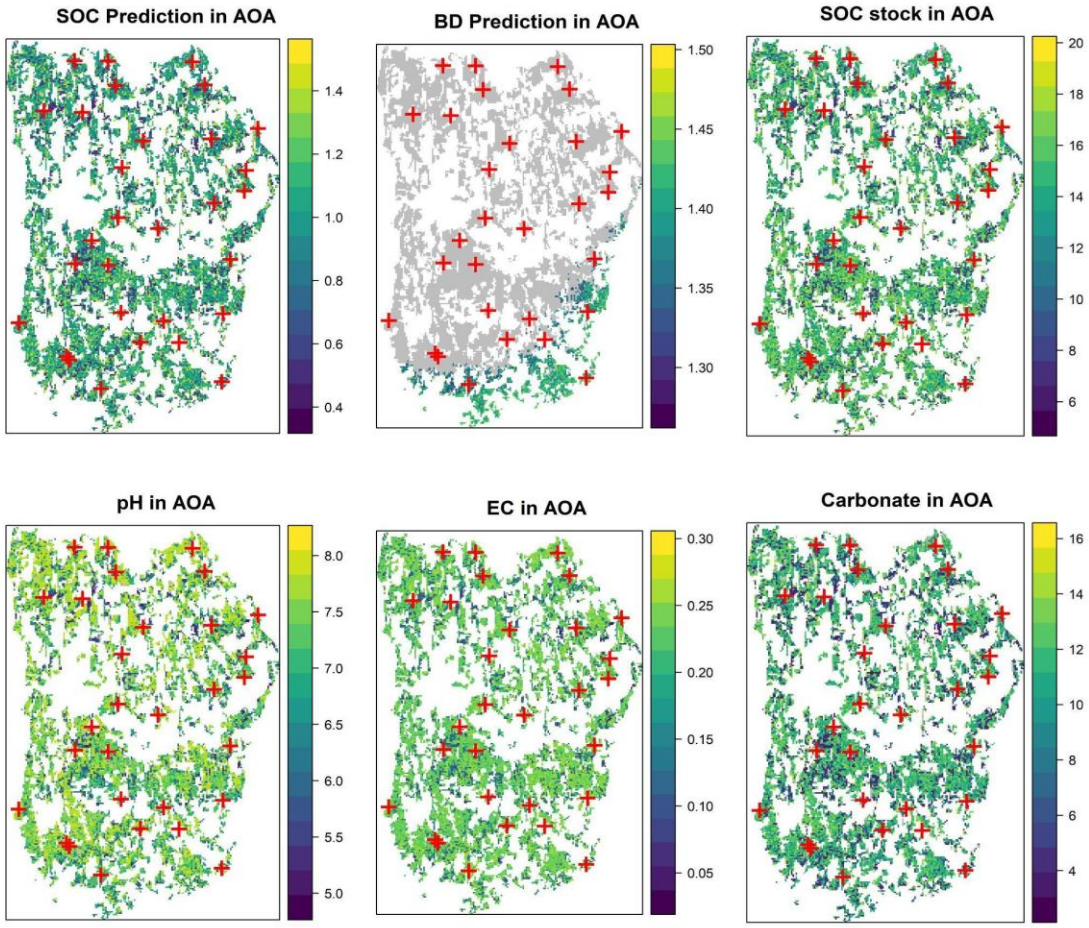



*Figure 14. The predictions masked by the derived area of applicability for Nyírség, where the areas outside the area of applicability are shown in gray. Also, the location of sampling points for validation are presented. The unit for each soil property is the same with Figure 12.*

*Table 9. Validate the results with samples taken from Hajdúhát; N means the number of points fell within AOA.*

| Property | Inside AOA | | | Outside AOA | | | N |
|---|---|---|---|---|---|---|---|
| | ME | MAE | RMSE | ME | MAE | RMSE | |
| SOC | -0.05 | 0.32 | 0.24 | -0.08 | 0.22 | 0.46 | 20 |
| BD | -0.02 | 0.07 | 0.08 | -0.05 | 0.07 | 0.09 | 25 |
| SOC stock | -0.23 | 2.08 | 2.41 | -4.83 | 5.63 | 6.59 | 21 |
| pH | 0.40 | 0.74 | 0.89 | -0.40 | 0.88 | 0.98 | 26 |
| EC | 0.008 | 0.11 | 0.12 | 0.09 | 0.02 | 0.15 | 16 |
| Carbonate | 0.82 | 0.83 | 1.17 | - | - | - | 30 |

In validating extrapolation results with Nyírség sampling (Table 10), the ME values for SOC and points inside BD are close to zero. The results for BD are only for 3 points because most of the parts in this area are outside AOA. The results for SOC stock showed a considerable error when extrapolating the points outside of AOA, in which RMSE increased from 9.14(inside) to 13.19(outside). Soil pH and EC sampling points fell inside AOA; therefore, there is no information to see how it changes the error outside of AOA. Still, the ME and RMSE values for soil pH are -1.0 and 1.62, indicating a larger systematic error. By comparing the validation for soil carbonate, it can be shown that the outside values for the ME (-6.18) and RMSE (16.23) are much higher than the inside values.

*Table 10. Validate the results with samples taken from Nyírség; N means the number of points fell within AOA.*

| Property | Inside AOA | | | Outside AOA | | | N |
|---|---|---|---|---|---|---|---|
| | ME | MAE | RMSE | ME | MAE | RMSE | |
| SOC | 0.03 | 0.6 | 0.78 | - | - | - | 30 |
| BD | 0.005 | 0.09 | 0.1 | -0.02 | 0.09 | 0.1 | 3 |
| SOC stock | 0.13 | 7.05 | 9.14 | -0.1 | 12.67 | 13.19 | 27 |
| pH | -1.0 | 1.30 | 1.62 | - | - | - | 30 |
| EC | -0.11 | 0.12 | 0.14 | - | - | - | 30 |
| Carbonate | -1.2 | 6.89 | 10.91 | -6.18 | 6.18 | 16.23 | 28 |

**4.2. Case study two**

The observations from the AfSP database were extracted for the three soil properties of Kenya, Ethiopia, Nigeria and Burkina Faso for the depth interval 0-20 cm from the surface. Their summary statistics are presented in Table 11.

*Table 11. Summary statistics of selected countries observations of three soil properties. SD is standard deviation and n is number of observations. The unit for Clay is g/100g and for SOC is g/kg.*

| Country | Variable | Min | Mean | Max | SD | N |
|---|---|---|---|---|---|---|
| Kenya | Clay | 0 | 37.7 | 88 | 19.6 | 400 |
| | SOC | 0.3 | 14.4 | 360 | 19.2 | 848 |
| | pH | 4 | 7 | 11 | 1.2 | 845 |
| Ethiopia | Clay | 2 | 35.4 | 90 | 17.3 | 1082 |
| | SOC | 0.6 | 24.4 | 251 | 23.6 | 1661 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | pH | 4 | 7 | 9.9 | 1.1 | 1710 |
| Nigeria | Clay | 0 | 20.2 | 84 | 18.9 | 1074 |
| | SOC | 0.2 | 9.7 | 102.4 | 8.3 | 1667 |
| | pH | 3 | 6 | 9.3 | 0.8 | 1753 |
| Burkina Faso | Clay | 1 | 21.5 | 64 | 14.8 | 616 |
| | SOC | 0.9 | 9.4 | 43.1 | 6.4 | 613 |
| | pH | 4.6 | 6.4 | 8.8 | 0.6 | 595 |

### 4.2.1. Similarity in soil types and homosoil

The Jaccard index was calculated to identify the similarity of soil types between two countries with the same proportion (Figure 15). A 100% in similarity would mean that two countries have the same soil type in the same proportions. The two countries with the most similar soil types are Burkina Faso and Nigeria (46.5%), followed by Ethiopia and Burkina Faso (43.4%), whereas Burkina Faso and Kenya have the lowest similarity (26%).

Ethiopia and Kenya, located in the eastern parts of Africa, share a 38% similarity. Nigeria and Ethiopia, Nigeria and Kenya have 35.3 % and 41.3 % identical soil types, respectively.

Based on the similarity of soil-forming factors between the donor and recipient countries, the homosoil technique identifies similar soils. We quantified the homosoil approach as the fraction of similar locations so that each time one country is the donor, others are the recipients (Table 12).

The highest similarity between soils was observed when Kenya was a donor and Ethiopia, and Nigeria were recipients, 41 % and 36%, respectively. At the same, only 6.6% similarity was found between Kenya (donor) and Burkina Faso (recipient) and almost nothing in the reverse situation. When Burkina Faso played as the donor country, the resemblance in soil forming factors with Nigeria was only around 14%. Ethiopia as a donor and Kenya as a recipient might have about one-third similarities in soils, followed by Burkina Faso (20.9%) and Nigeria (14.6%).

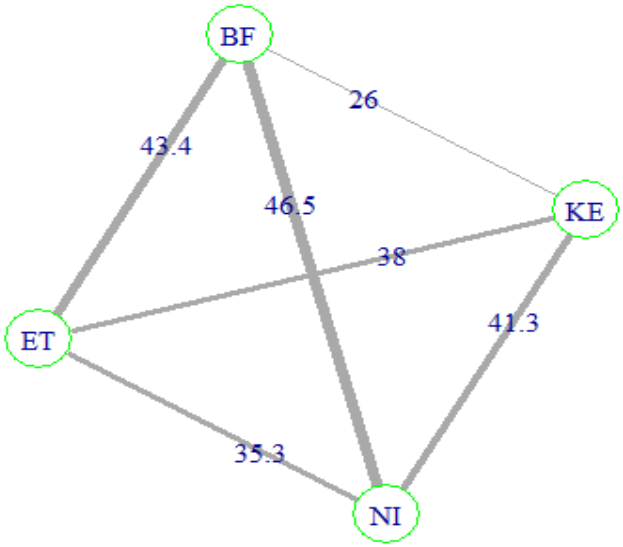



*Figure 15. Similarity in soil types between selected countries (%), Abbreviation; BF: Burkina Faso, KE: Kenya, ET: Ethiopia, NI: Nigeria.*

*Table 12. Homosoil fractions (%). Horizontal the donors, vertical the recipients.*

| | Kenya | Ethiopia | Nigeria | Burkina Faso |
|---|---|---|---|---|
| Kenya | - | 29.5 | 33.2 | 0.5 |
| Ethiopia | 41.4 | - | 14.2 | 1.1 |
| Nigeria | 36.5 | 14.6 | - | 14 |
| Burkina Faso | 6.6 | 20.9 | 14.3 | - |

### 4.2.2. RF model and dissimilarity index by AOA

It is required to train the model before calculating the dissimilarity index by AOA. Random Forest model results by applying 10-fold cross-validation are shown in Table 13. According to the modeling efficiency (MEC), the model could deliver between 30 and 59 % of the spatial variation of Clay and SOC in selected countries. However, it explained greater MEC values for spatial prediction of soil pH, around 50 to 70%. In the case of Burkina Faso, the MEC indicated poor predictions for clay and pH, about only 18% and 15%, respectively. The ME values for soil pH for all four countries were near zero; therefore, the predictions were unbiased. But the ME values for SOC in Nigeria and clay in Kenya were the lowest. The best model with higher MEC and lower error measurements in the spatial prediction of clay was observed in Nigeria, followed by SOC in Burkina Faso and pH in Kenya.

*Table 13. Results of trained model for each country. The unit for Clay is g/100g and for SOC is g/kg.*

| Property | Clay | | | SOC | | | pH | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | ME | RMSE | MEC | ME | RMSE | MEC | ME | RMSE | MEC |
| Kenya | -0.04 | 16.2 | 0.31 | 0.18 | 15.29 | 0.37 | 0.01 | 0.64 | 0.72 |
| Nigeria | 0.26 | 12.22 | 0.57 | 0.06 | 6.0 | 0.43 | 0 | 0.55 | 0.53 |
| Ethiopia | 0.1 | 14.3 | 0.32 | 0.17 | 16.26 | 0.53 | 0.01 | 0.64 | 0.67 |
| Burkina Faso | 0.4 | 13.44 | 0.18 | 0.16 | 4.8 | 0.44 | 0 | 0.58 | 0.15 |

The trained models for each country and each property were employed to obtain the dissimilarity index maps by calculating AOA. From now on, we will only present figures for Kenya as a donor country (Figure 16); other figures are provided in Appendix.

Generally speaking, the eastern countries (Ethiopia with Kenya) exhibit more similarity across all characteristics, than the western countries (Nigeria with Burkina Faso). For example, in Figure 16, the trained model in Kenya was applied to calculate DI in other countries. We can see that most of Ethiopia has dark blue colors, indicating less dissimilarity, while Nigeria and Burkina Faso have lighter colors for clay and SOC, meaning that dissimilarity increases. However, the DI in soil pH for Nigeria and Burkina Faso are closer to the middle colors. When Burkina Faso is a donor country, the DI increases significantly especially in the central parts of Ethiopia and Kenya, which in case of soil pH is getting worse and DI goes over 20, showing yellow colors (Figure SM1, supplementary material).

Figure 17 displays the distributions of the prediction DI of the trained model for the donor country versus the prediction DI of the recipients. There is some overlap between Kenya (Donor) and Ethiopia (Recipient) in their DI distribution. At the same time, there is little overlap between Kenya and Nigeria (recipient), and almost no coverage between Ethiopia and Burkina Faso has been recognized (recipient). The range of dissimilarity in the case of soil pH decreases compared to Clay and SOC. Interestingly, when Burkina Faso plays as a donor country, the DI density plot for itself is getting narrower, meaning that a smaller range for dissimilarity and more points are concentrated in this range. In contrast, the recipient countries are shaped entirely flat due to a more extensive range of dissimilarity (Figure SM4,

supplementary material). This means that covariates in recipient countries are different from covariates in Burkina Faso.

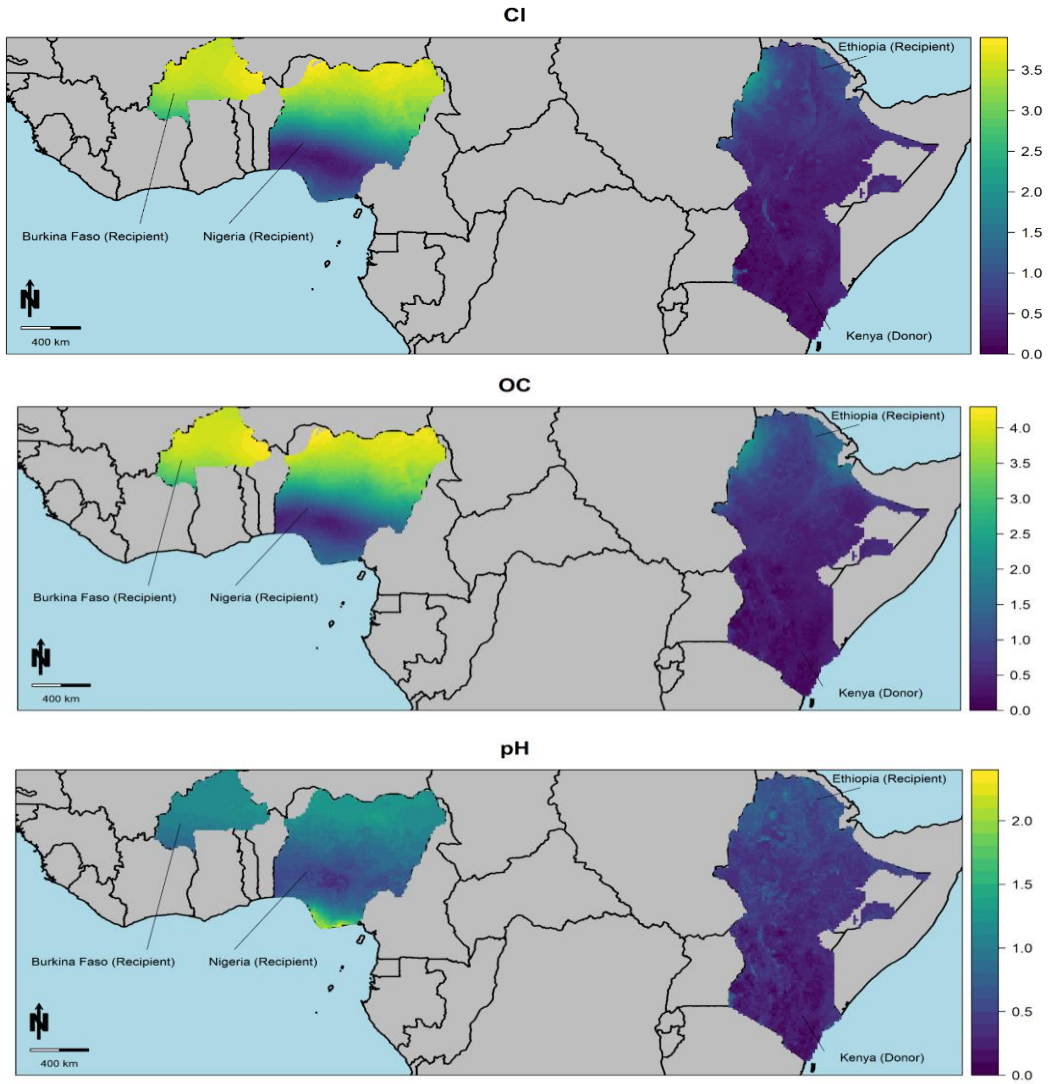



*Figure 16. Dissimilarity index map when Kenya is a donor country and other countries are recipients.*

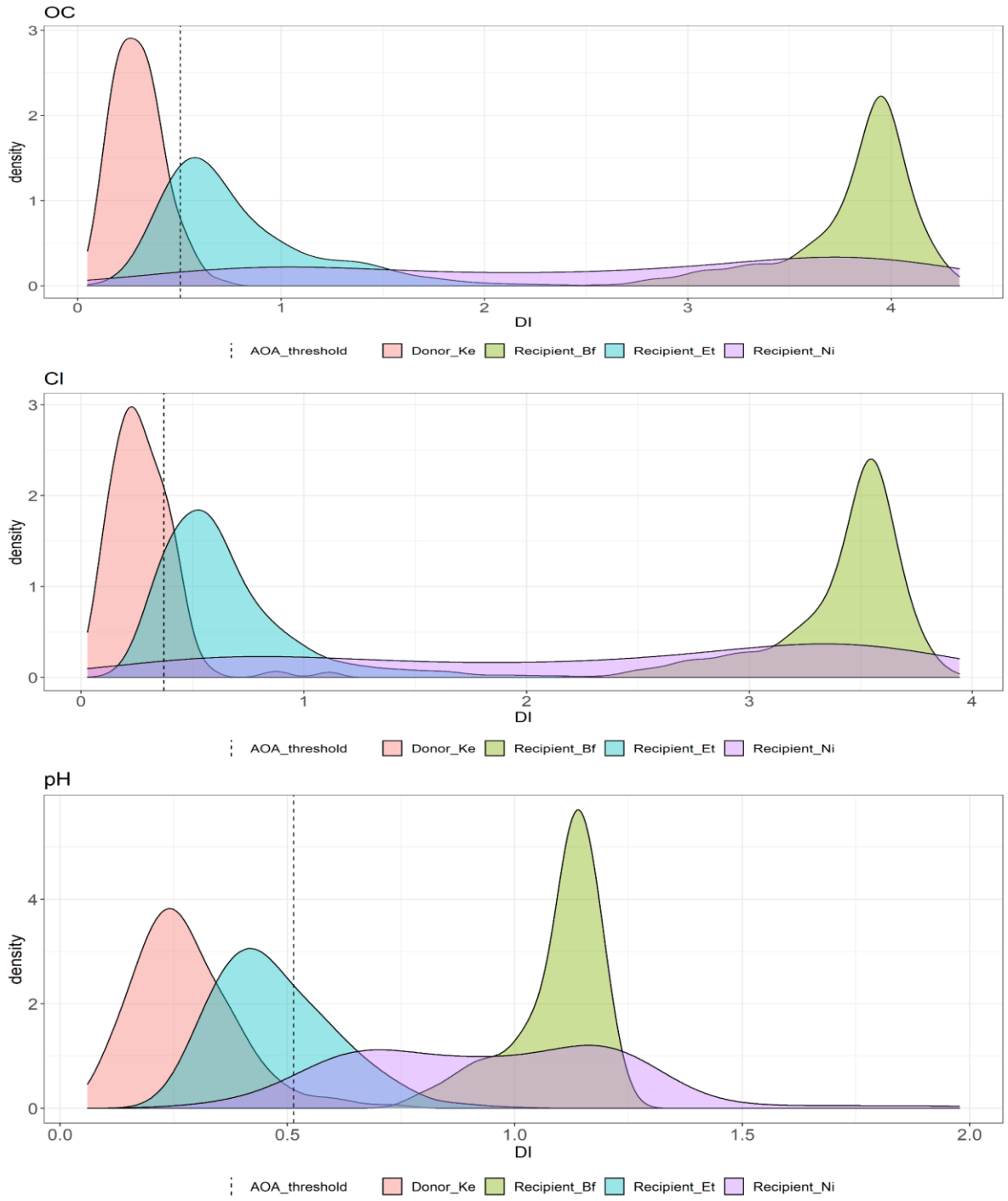



*Figure 17. Distribution of the prediction DI, Kenya is a donor country and other countries are recipients. Abbreviation; BF: Burkina Faso, KE: Kenya, ET: Ethiopia, NI: Nigeria.*

### 4.2.3. Uncertainty and comparison

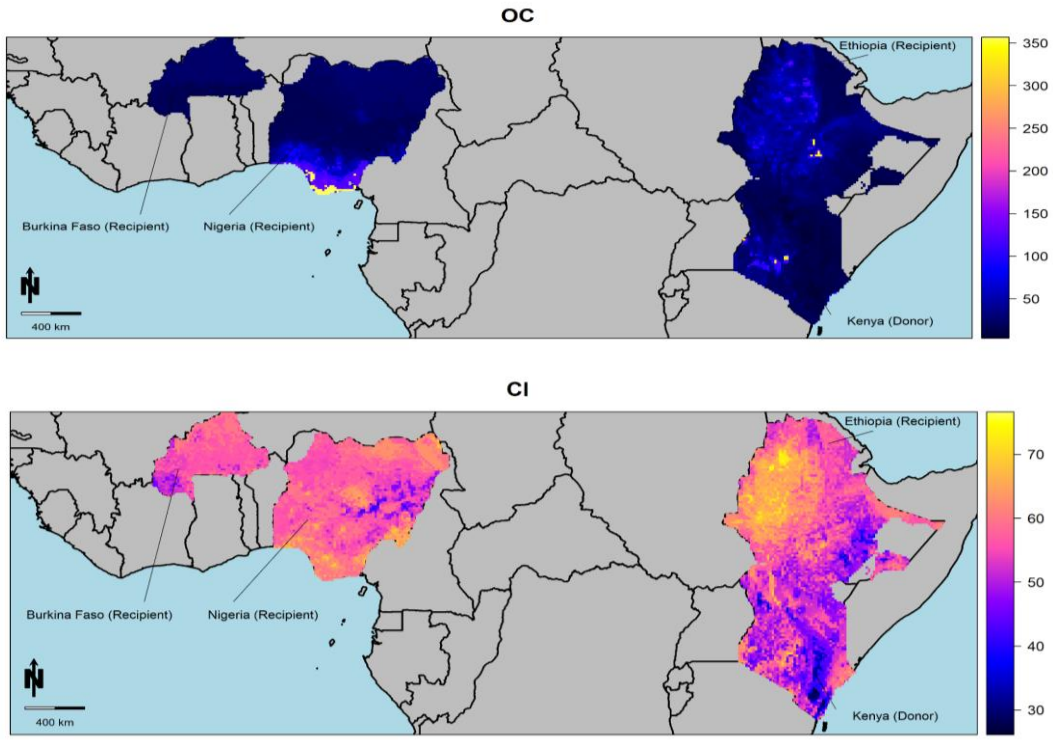
QRF estimation of uncertainty by deriving a 90% prediction interval width (PIW) is shown in Figure 18. Same as in the previous section, we only presented the maps for Kenya as a donor country for all properties, and other maps are in Appendix. Although there were differences in the magnitude of uncertainty, all figures generally delivered a comparable spatial pattern of uncertainty between countries in the same region. In Kenya, the PIW map for clay displayed varying patterns, with certain regions exhibiting a smaller PIW and others demonstrating an expanded PIW. However, the width of the PIW increases further when making predictions for recipient countries. Extrapolation of soil pH based on the trained model for Kenya can be less uncertain in Ethiopia compared to northern parts of Nigeria and Burkina Faso. The PIW for Burkina Faso when acting as donor produced a smaller range than when other countries act as donors. For example, this range is at most 60% for clay in Burkina Faso, while for the other three countries, this value is over 70%. Therefore, extrapolation to other countries by the trained model in Burkina Faso is too uncertain, and all maps demonstrated wider PIW (Figure SM7, supplementary material).

The difference between the 0.05- and 0.95-quantile in Nigeria (donor) maps showed a wider width for clay and pH in extrapolation to other countries. Although the PIW for Nigeria's SOC map showed the pattern's middle colors, it produced out-of-range values in extrapolating over Ethiopia and some tiny parts of Kenya (Figure SM9, supplementary material).

It is demanded to validate the extrapolation results over the recipient countries by their available datasets in each country, as presented in Table 14. Overall, the statistics in this table show that the capability to extrapolate is much worse compared to what was seen in the cross-validation results of the trained model in Table 13. Yet, it would be more straightforward to predict for the countries in the same region. Compared to RMSE numbers, the ME values, which represent the systematic error of the predictions, can occasionally be rather high.

Since the trained model was unbiased (Table 13), the ME values should be very small. But we see that when extrapolation happens, the systematic error increases, and it sometimes gets too high or too low. This indicates that the RMSE is biased regarding systematic and random error contributions. For instance, extrapolating SOC over Nigeria with the trained model in

Ethiopia, the ME and RMSE values are 11.49 and 13.63, respectively, indicating the random error contribution is too low. In some cases, severe under and overestimation of prediction has been observed. For example, predictions over Ethiopia and Kenya by Burkina Faso's trained model for clay were underestimated, and ME values were -12.37 and -15.54, respectively. Spatial prediction of soil pH had less error compared to Clay and SOC. Ethiopia and Kenya could deliver the spatial prediction of pH to each other by 23 % (When Ethiopia was the donor) and 22% (When Kenya was the donor). Also, extrapolating over Nigeria by the trained model in Kenya had the MEC around 24 %, which slightly performed better in extrapolation. In situations where the MEC were close to zero or negative, the predictions are equal or worse if we only apply the average of the measurements.



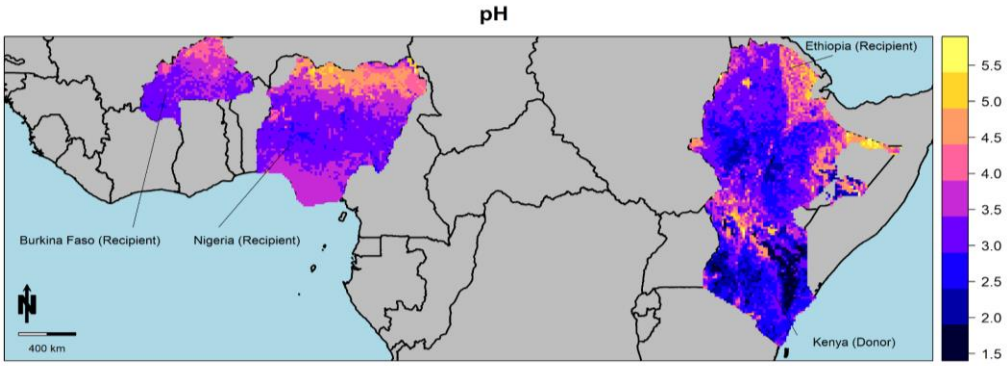



*Figure 18. Prediction interval width when Kenya is a donor country and other countries are recipient*

*Table 14. Final validation metrics; comparison between all predictions. Abbreviation; BF: Burkina Faso, KE: Kenya, ET: Ethiopia, NI: Nigeria.*

| | | Clay | | | OC | | | pH | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Donor | Recipient | ME | RMSE | MEC | ME | RMSE | MEC | ME | RMSE | MEC |
| KE | Ni | 8.08 | 18.51 | 0.02 | 1.63 | 6.96 | 0.24 | 0.71 | 1.01 | -0.61 |
| ET | Ni | 17.87 | 27.52 | -1.16 | 11.49 | 13.63 | -1.92 | 0.45 | 0.89 | -0.26 |
| BF | Ni | 3.0 | 18.08 | 0.07 | 1.61 | 7.6 | 0.09 | 0.42 | 0.89 | -0.24 |
| ET | Ke | -2.2 | 18.24 | 0.13 | 6.77 | 18.19 | 0.11 | -0.36 | 1.06 | 0.23 |
| BF | Ke | -15.54 | 24.82 | -0.61 | -2.94 | 18.23 | 0.1 | -0.61 | 1.33 | -0.22 |
| NI | Ke | -13.92 | 23.39 | -0.43 | -3.66 | 18.66 | 0.06 | -1.08 | 1.5 | -0.54 |
| KE | Et | 2.09 | 16.66 | 0.07 | -7.55 | 25.18 | -0.14 | -0.38 | 0.98 | 0.22 |
| BF | Et | -12.37 | 21.43 | -0.54 | -11.33 | 26.31 | -0.24 | -0.71 | 1.31 | -0.4 |
| NI | Et | -9.17 | 19.56 | -0.28 | -14.01 | 27.44 | -0.35 | -1.11 | 1.51 | -0.86 |
| KE | Bf | 4.55 | 15.4 | -0.08 | -1.52 | 6.07 | 0.11 | 1.01 | 1.23 | -2.79 |
| ET | Bf | 12.34 | 19.5 | -0.74 | 5.95 | 8.34 | -0.69 | 0.95 | 1.16 | -2.39 |
| NI | Bf | 1.34 | 14.25 | 0.07 | -1.68 | 6.23 | 0.06 | 0.05 | 0.63 | 0 |

### 4.3. Case study three

A summary statistics of the SAS indicators from topsoil, including EC, pH and SAR, are presented in Table 15. Soil pH in the study area is slightly alkaline, about 7.9 ~ 8.2. The difference between the minimum and maximum values observed for EC and SAR indicated that it might be saline and sodic at some points. The average of the estimated SAR values is 15.8, which refers to dominating Na+ in soil samples. In soil science, a SAR value above 13 is considered sodic with high levels of exchangeable sodium.

Higher organic matter, salt, silt content, and lower carbonate concentrations can all be found in the soil of formerly heavily vegetated local depressions. Most measured soil properties

correlated with elevation (94.6–96.2 m a.s.l). According to the IUSS Working Group's World Reference Base for Soil Resources (ANJOS et al. 2015), seven major soil types could be distinguished on the study site: Chernozem (63.53%), Phaeozem (15.29%), Kastanozem (7.06%), Calcisol (5.88%), Gleysols (4.71%), Cambisols (2.35%), and Regosols (1.18%) (Tóth et al. 2022). These soil groups describe the variation in SOC, the humus layer's thickness and the carbonate content's spatial variation.

There are no major saline-sodic soil groups among the soil profiles due to agricultural use and management; however, checking the soil qualifiers indicates the presence of sodium and former salt accumulation. In most cases, ex-situ measurements of the soil columns reveal salt maxes in the C horizon, as the soil qualifiers support this remark.

*Table 15. Summary statistics of the point observations (n = 85).*

| SAS indicators | Unit | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| EC | µS $cm^{-1}$ | 136.4 | 428.0 | 214.0 | 59.94 |
| pH | - | 7.90 | 8.79 | 8.201 | 0.15 |
| SAR | - | 0.13 | 181.0 | 15.79 | 37.32 |

In this study, the histogram of pH values followed a normal distribution, while EC and SAR did not show the expected behavior. Therefore, we applied the normal score transformation for EC and SAR values. The performance of the SuperLearner model and each independent ML model are shown in Table 16. In comparison to the performance of each model separately, the SuperLearner model predicts pH values and SAR with the highest accuracy (R2) and the least error (RMSE). Evaluating the performance of each base learner revealed that RF was the best model by delivering 36 % of spatial prediction of pH followed by SVM. Also, RF, SVM and GLM succeeded in explaining SAR spatial variation by 90 % and RMSE 0.2 ~ 0.3. In contrast, the worst performance was related to XGBoost for the spatial prediction of pH and NN for the spatial prediction of SAR (Table 16).

However, in the case of EC, only RF could deliver the spatial variation in acceptable results, while SuperLearner and other individual models were unsuccessful. The map of EC by SuperLearner showed artifacts and produced irrational values in the case of NN and GLM.

Therefore, we only used and provided the RF results to overcome this issue. Table 16 showed that R2, RMSE and MAE for EC by RF model were 0.39, 0.97 and 0.79, respectively.

*Table 16. Summary of performance of ensemble modeling and each ML algorithm which was included.*

| ML algorithms | $R^2$ | | | RMSE | | | MAE | | |
|---|---|---|---|---|---|---|---|---|---|
| | pH | EC | SAR | pH | EC | SAR | pH | EC | SAR |
| RF | 0.36 | 0.39 | 0.96 | 0.12 | 0.97 | 0.21 | 0.10 | 0.79 | 0.09 |
| XGBoost | 0.09 | - | 0.91 | 5.42 | - | 0.83 | 5.42 | - | 0.65 |
| NN | 0.09 | - | 0.10 | 0.15 | - | 1.00 | 0.11 | - | 0.81 |
| SVM | 0.22 | - | 0.90 | 0.13 | - | 0.33 | 0.10 | - | 0.21 |
| GLM | 0.12 | - | 0.95 | 0.14 | - | 0.27 | 0.11 | - | 0.14 |
| SuperLearner | 0.43 | - | 0.96 | 0.11 | - | 0.20 | 0.09 | - | 0.11 |

Afterwards, the residuals for pH and SAR were computed by calculating the difference between observations and prediction by SuperLearner. In the case of EC, the residuals were derived only from the RF prediction results. Then direct and cross-variograms of the residuals were calculated and are given in Figure 19. Between the residuals, an explicit spatial dependency and interdependency have been recognized. This indicates that involving the residuals (multivariate geostatistics) in the spatial variability of these indicators is reasonable. Also, we fitted a linear model of coregionalization (Goovaerts, 1997) using a spherical model type and range value of 350 m to ensure that a statistically sound model is employed in multivariate geostatistical modeling.

Also, as we mentioned earlier, uncertainty measurements enable us to assemble probability maps, which are quite useful in practice. Thus, we compiled the probability map of SAR values which shows the likelihood that SAR will exceed 13 (threshold). This map (Figure 21) indicates the high probability of finding soils with high sodium content in the northern part of the study plot, which is in general agreement with our observations. Figure 21 provides valuable insights that can be utilized to offer guidance and advice to stakeholders.

As mentioned, we validated the spatial prediction results by using 10-fold cross-validation. Accordingly, the accuracy measurements, including ME, RMSE, CCC, and MEC are reported in Table 17. The most promising unbiased results can be interpreted from the least ME and

RMSE with the highest accuracy. According to the results, the spatial predictions for all three indicators were accurate and acceptable. The ME values in all indicators were close to zero. The CCC values changed from 0.39 for EC, 0.59 for pH and 0.97 for SAR. The modeling efficiency showed the highest value for SAR (0.95) and the lowest value for EC (0.24).

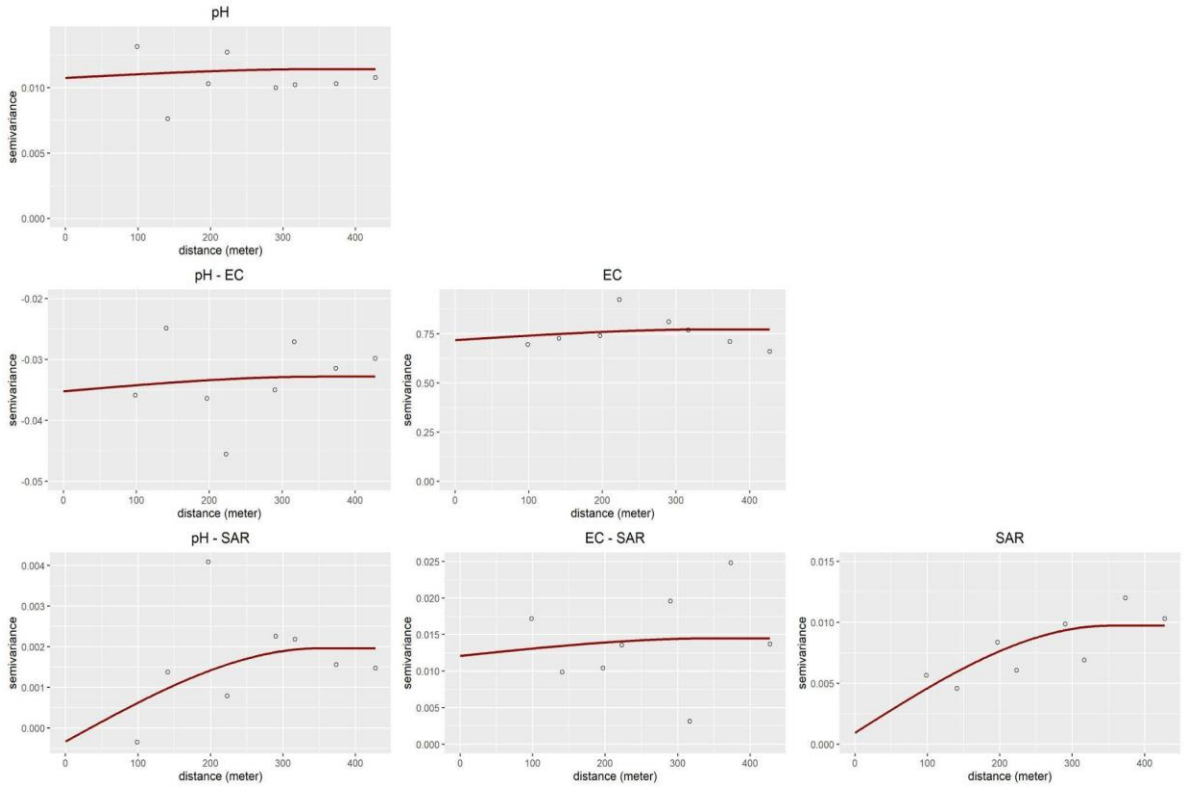



*Figure 19. The computed direct and cross-variograms (open circles) and fitted linear model of coregionalization (solid line).*

In the end, the accuracy plots with computation of G statistics for pH, EC and SAR are displayed in Figure 22. The accuracy plots confirmed that the uncertainty quantifications are valid for each indicator since it follows the y=x line. Also, the calculated G statistics, expected to be close to their expected value (i.e., 1), further support this.

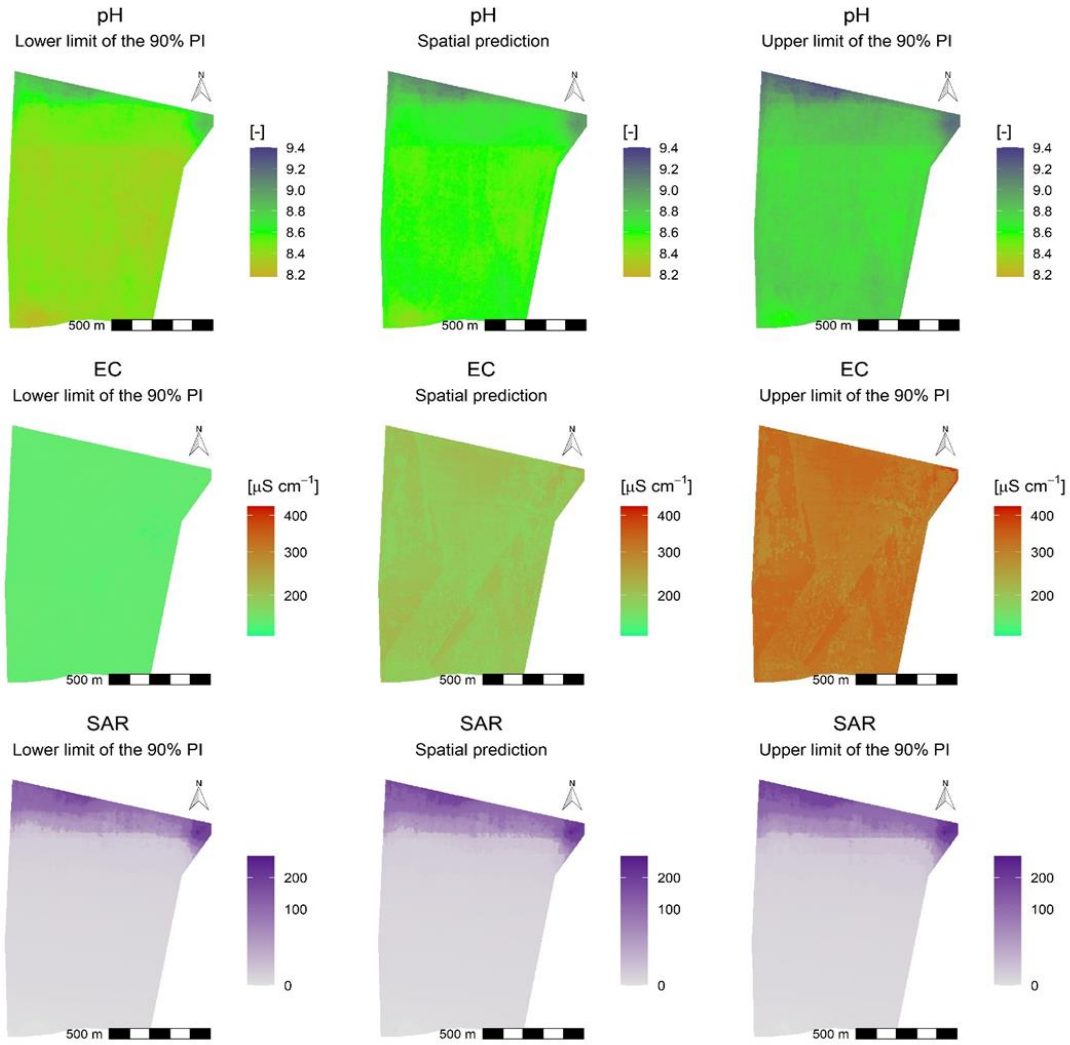



*Figure 20. Spatial predictions of the salt-affected soils indicators with their associated prediction uncertainty expressed by lower and upper limit of the 90% prediction interval (PI).*

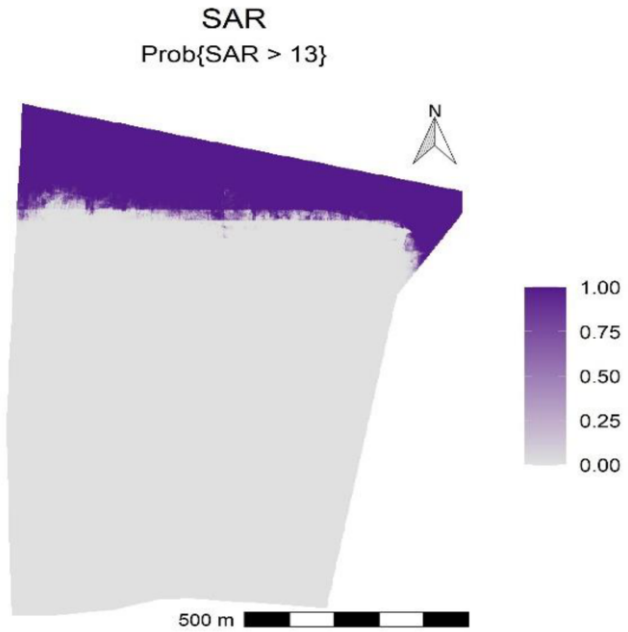



*Figure 21. Probability map of sodium adsorption ratio (SAR) for the SAR values greater than the threshold of 13.*

*Table 17. The performance of spatial predictions of SAS indicators by 10-fold cross-validation.*

| SAS indicators | ME | RMSE | CCC | MEC |
|---|---|---|---|---|
| pH | <0.001 | 0.11 | 0.59 | 0.41 |
| EC | <0.001 | 0.86 | 0.39 | 0.24 |
| SAR | 0.007 | 0.22 | 0.97 | 0.95 |

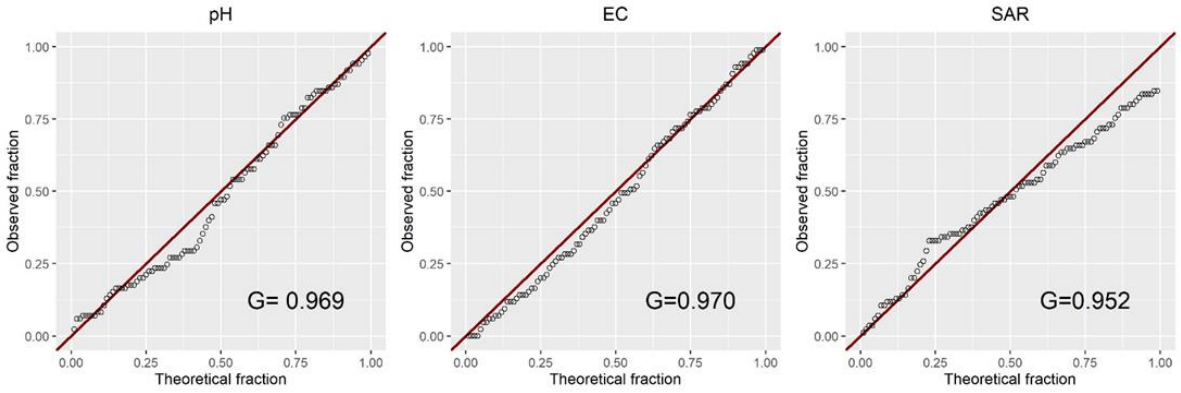



*Figure 22. Accuracy plots with the computed G statistics.*

# 5. Discussion

## 5.1. Case study one

### 5.1.1. Soil properties and environmental covariates relationship in Látókép and Westsik

The most important environmental variables based on the random forest model for predicting soil properties in Látókép were mainly those derived from DEM, such as elevation, plan and profile curvature, slope and multiresolution index of ridge top flatness. Therefore, it can be concluded that topographic indices are the major factor affecting the spatial distribution of soil properties in this study area. Some parameters, such as geology and climate, are the same over the area and were considered constant. Hence, the only parameters that make a difference in the spatial variation of soil properties are terrain attributes which express landscape morphometry. Previous studies have shown that soil properties strongly correlate with topography indices extracted from DEM (McBratney et al. 2003; Heung et al. 2016; Tziachris et al. 2019; Zhou et al. 2020). Forkuor et al. 2017 showed elevation as the most significant variable for the spatial prediction of SOC and nitrogen. Zhou et al. 2020, indicated that topographic variables, especially elevation, are the most explanatory variables for soil health indicators.

The only Landsat indices that had a markable influence in explaining the soil spatial variation in Látókép were NDVI in determining BD and Clay index in soil pH. Salehi Hikouei et al. 2021, investigated the importance of vegetation indices in estimating soil bulk density. NDVI was assigned the second most crucial variable by the models, which verifies that soil bulk density is highly affected by vegetation variation and structure.

Natarajan et al. 2022 indicated that satellite images could be a powerful tool in modeling soil pH. Tajik et al. 2020, demonstrated that terrain attributes and vegetation indices have a significant relationship with soil microorganisms communities which can be used for soil quality assessment. However, Ghazali et al. 2020, found a weak correlation between satellite images from Landsat 8 and soil pH values, but they believed it might help identify soil pH.

At the same time, in Westsik, the most influential explanatory variable for all soil properties was NDVI, while the parameters extracted from DEM showed almost no significant effect. The soil-forming factors in this area are also homogenous, and even elevation and its

derivatives showed few differences. It seems the only variable affecting the soil spatial variation is vegetation cover over this area which is reasonable since it's an experimental research field and every year, the rotation of vegetation is changed. Asgari et al. 2020, showed that NDVI and soil adjusted vegetation index were among the most important predictors in the spatial prediction of soil great groups. In addition, Bhunia et al. 2019 concluded a powerful relationship between NDVI and SOC stock in estimating spatial SOC stock by a multivariate regression model. Song et al. 2017, showed remote sensing variables, especially NDVI, have a great impact in explaining soil organic matter variation since both vegetation and organic matter are large sources of carbon which could have a strong correlation and act as a driving force.

### 5.1.2. Comparison of the performance of machine learning models

Each model has advantages and disadvantages in terms of predictability, depending on different soil-forming elements and local/regional circumstances. Despite having few observations, our research convincingly demonstrates that RF outperforms MLR, SVM, and NN in both areas. The results revealed a high correlation between soil data and environmental variables, so the final map by RF could explain soil variation for these properties with an $R^2$ coefficient of 0.8. Outliers, unbalanced data, and non-linear and complex relationships may all be handled very well with RF (Ao et al. 2019). Additionally, it was demonstrated by Hengl et al. (2018) that the RF model is the most popular and attractive model for spatial predictions. Also, they noted that the most significant features of this model are flexibility in combining, incorporating, and extending covariates of various types and the capability to provide more informative and detailed maps. Similar to the results of this study, many researchers have reported the acceptable accuracy of the RF model in mapping different soil properties (Pahlavan, 2015; Heung et al. 2016; Camera et al. 2017; Fathizad et al. 2020). To predict SOC stock, John et al. (2020) used various machine learning techniques, including NN, SVM, cubist, RF, and MLR. They select MLR as the least effective model and RF as the one that performs the best. Vaysse and Lagacherie (2015) utilized RF in the spatial prediction of soil pH for different depths and acquired a high performance (R2 ~ 0.7 to 0.8).

In predicting SOC and soil pH in Látókép, SVM performed slightly better compared to MLR and NN, with $R^2$ of about 0.27 and 0.33, respectively. In addition, this model was more successful in the Westsik study area, as it could explain between 30% to 40 % of the spatial variation of SOC, pH, EC and carbonate. Similar to these results, Campbell et al. (2019) predicted SOC and clay for topsoil using spectral data and achieved $R^2$ values of 0.45 and 0.25, respectively. However, Dotto et al. (2018) applied SVM on SOC stock values and validated the results with an accuracy of 0.75 on an independent dataset. Were et al. (2015) revealed that SVM performed better in the spatial prediction of SOC stocks than RF and NN. In this study, the non-linear relationship between soil attributes and environmental variables may have contributed to MLR's failure; however, the interrelationships between the variables are complicated. According to Forkuor et al. (2017), alternative machine learning algorithms perform better than MLR when handling non-linear and complex relations between variables. Negative $R^2$ was observed in the performance of NN, although this is not a computational or mathematical problem. The results indicated the model performed even worse than if we had merely employed the spatial average of the data. The inability of NN could be due to limited observation (Moody, 1994; Tu, 1996; Hernández-Lobato and Adams, 2015). Khaledian and Miller (2020) reported that the major drawback of employing ANN is its sensitivity to sample size. There are some ML algorithms that are not sensitive to the number of observations, such as Cubist and RF. Although some studies accomplished strong results with ANN, a large sample size is required to produce stable results, which DSM studies typically lacked. Nevertheless, some research has supported the great predictability of ANN, even with a smaller sample size. For instance, SOC (0-100cm) with 595 points (Dotto et al. 2018) and soil moisture content (topsoil) with 137 points (Mahmoudabadi et al. 2017) both achieved $R^2$ values for independent validation of about 0.79 and 0.85, respectively. An acceptable outcome was obtained by Zhao et al. (2009) in their study on soil texture prediction using an ANN, which had a relative overall accuracy of 80% for clay and sand content. Furthermore, Li et al. (2013) observed that the radial basis function NN model demonstrated a more realistic spatial pattern and outperformed MLR and regression kriging in predicting SOC.

The other limitation of algorithms such as ANN is that the data should be transformed, and their assumption relies on normal data distribution. However, in other algorithms, the data

does not necessarily need to be normalized to some scales in order to function properly (Kuhn and Johnson, 2013; Hengl et al. 2015).

### 5.1.3. Extrapolation and AOA

We calculated the AOA of predictive models, by which we represented the area where the models were enabled to learn about relationships in Látókép and Westsik. Then we predicted the model over Hajdúhát and Nyírség and estimated the prediction error inside and outside of AOA by the samples taken from these areas. The application of AOA is quantifying the differences in the environmental covariates between donor and recipient areas and determining the area for which the model can be expected to make predictions with an error comparable to the model performance.

The dissimilarity index values for each of these soil properties showed different ranges due to various relationships between the soil property of interest and predictors. For example, the correlation between SOC and elevation in one area might differ with soil pH. As we can see also in Figure 9 and 10, the selection of important covariates applied for weighting in training the model is different, leading to different results in the calculation of the AOA. The most important covariate for the predictive model in Westik was NDVI for all soil properties except BD. Therefore, the possible areas to extrapolate over Nyírség for BD considerably differ from other properties.

The masked spatial extrapolation maps of SOC stock and EC in Hajdúhát and the map of BD in Nyírség showed large areas which are outside of AOA. These findings imply that distinct soil-forming factors are crucial in explaining how these soil properties vary spatially between the two regions. In other words, the environmental covariates applied for training the model in donor areas were not enough to cover the heterogeneity of larger areas. This result is consistent with Taghizadeh-Mehrjardi et al. (2022), who concluded that different scorpan factors are responsible for the explanation of soil class variations in two areas. Also, they showed that in places where the covariates in the donor area are too different from the recipient area, the DI increases and probably the predictions in those areas contain higher uncertainty.

By assessing the validation results inside and outside of AOA, we found that the predictions inside have considerably fewer errors, and the values are closer to the error measurements of predictive models. This is similar to the results of Meyer and Pebesma (2021), who discovered that prediction errors within the AOA are similar to the cross-validation error based on a database of 972 simulations. The predictions outside of AOA should not be considered valid since the dissimilarity is larger than the DI of the trained model. In general, where the DI is near 0, ML models perform well. Comparability in terms of hydrology, pedology, and geomorphological processes in both the donor and recipient area is crucial for practical extrapolation (Lagacherie et al. 2001). According to Nenkam et al. (2022), the homosoils concept, which defines the similarity of soil forming factors, aids in extrapolating soil data between two areas using the DSM model.

Furthermore, the significant error in larger areas (Hajdúhát and Nyírség) in terms of some properties can be due to the small sample size in both areas. In contrast, better performance accuracy may be caused by more observations being represented in the sample data (Debella-Gilo and Etzelmüller, 2009). Jafari et al. (2012) noted that a critical element affecting the purity of the map is the size of the sampling units in relation to the entire research area. In other words, uncertainty increases with a smaller sample size.

Grinand et al. (2008) examined the extent to which the model produces a reliable prediction in France using a supervised learning technique for extrapolating soil types. It was discovered that the predictions for the donor area were more accurate than those for the recipient area. Due to the complexity of soil spatial variation and the difficulty of matching soil-forming components between two areas, the low accuracy of the spatial extrapolation was expected.

Knowledge of the AOA is helpful when there is limited observation (e.g. field data), and predictions are accomplished for heterogeneous areas, or in other cases when there is uncertainty about applying the model to a new environment.

The results indicated that it is demanded to account for insufficient coverage of environmental covariates and to limit predictions to areas that are comparable in their predictors compared to the training data and are therefore inside the AOA. In other words, some new spaces (locations in the recipient area) might be different in terms of environmental properties, and the trained algorithm has never seen such properties.

In this study, similarity between study areas was only based on expert knowledge. The two areas (Látókép vs Hajdúhát; Westsik vs Nyírség) are formed during the same periodic time (late Pleistocene), and have the same parent material and climatic zones. The land use for all four areas was arable lands. They are categorized as flat landform types on alluvial plains. Although few differences in elevation have been observed in each area, it seems topography is the main factor affecting soil formation. Therefore, an important factor that should be considered for a successful extrapolation is the similarity between the donor and the recipient areas.

### 5.2. Case study two

#### 5.2.1. Similarities by different methods

From case study one, I have learned a degree of similarity between the two areas in terms of all soil-forming factors plays a key role in the possibility of extrapolating soil information from the donor to the recipient areas. Therefore, it is important to quantify the similarity between two areas in terms of different aspects (SCORPAN factors), with more observation. It is the reason we defined the second case study to use ISRIC Africa Soil Profiles (AfSP) which has many observations in all African countries with sufficient coverage of samples over some countries. In this case study, I quantified the similarities in terms of soil type and soil forming factors.

Different SCORPAN factors play important roles in explaining the spatial variability of soil types in the two areas. In this study, countries in the same region, Ethiopia with Kenya, Nigeria with Burkina Faso, have more similarities in terms of soil types and homosoil approach.

Considering factors such as climate, variety of soil types and difference in elevation with results of the homosoil approach indicates that Kenya and Ethiopia have the largest diversity, followed by Nigeria. Therefore, the probability of finding more similar soils from a country with a more heterogeneous condition to other countries increases. At the same time, Burkina Faso is recognized as having the lowest variation in soil-forming factors among the countries included in the research.

When predictions are made for heterogeneous countries, the trained model has seen more spatial variability of soil information such as in Ethiopia and Kenya. Therefore, models trained in these countries can be more successfully transfered to other countries, as indicated in the DI maps and plots, in which the DI range decreases (Figure 16. DI maps in Kenya). Furthermore, the range of DI increases in more homogenous areas like Burkina Faso specially in case of soil pH, as indicated in Figure SM1, supplementary material. This means that when Burkina Faso is a donor, the new geographic space (other countries as recipients) might differ considerably in its environmental covariates from what has been seen in the training data. In other words, the environmental covariates presented in recipient countries are not sufficiently covered by the training data in Burkina Faso.

Comparing the DI and PIW maps, indicated a subtle correlation between spatial dissimilarity to the donor area and related uncertainty. In those areas with significant differences in environmental covariates, the predictions probably contain higher uncertainty compared to the areas with lower DI. The results are in general agreement with findings of Malone et al. (2016) which showed similar areas tend to have less uncertainty compared to when dissimilarity increases. Taghizadeh-Mehrjardi et al. (2022) indicated ML models work better in areas where the DI is closer to zero. Additionally, Jafari et al. (2012) showed that predictions with high levels of uncertainty probably result from an insufficient conceptual model.

Also, countries in the same region revealed closer or similar spatial patterns to each other based on DI and PIW maps. Similarities found between neighboring countries considering all four methods (soil type, homosoil, DI and PIW) indicated that geographical proximity is regarded to be crucial for transferring the trained models to the recipient areas, which was confirmed by Nenkam et al. (2022) and Angelini et al. (2020). This might explain the low predictive power of Burkina Faso versus Kenya/Ethiopia. The other reason might be the high differences in soil spatial variation between these countries which affected the relationship between soil and environmental covariates, and finally impacted the ability of the trained model to predict over other countries(Nenkam et al. 2022).

### 5.2.2. Extrapolation results

The results for cross-validation of the trained model were acceptable (Table 13. Results of trained model for each country) but when extrapolation occurs (Table 14. Final validation metrics), the results were quite poor. In general, the validation results are unsatisfactory, which emphasizes the possible dangers of extrapolation between two areas.

Due to the intricacy of soil-landscape interaction and the difficulties of fully matching soil-forming factors, the low performance of spatial extrapolation was anticipated (Malone et al. 2016). This finding aligns with Grinand et al. (2008), who discovered that the predictive accuracy was quite low when their trained model for one area was extrapolated to another area. Nenkam et al. (2022) also found that transferring the model within homosoil areas performed weakly although homosoils can be an effective tool in transferring soil data between two areas. Taghizadeh-Mehrjardi et al. (2022) found that spatial interpolation always has a higher accuracy and lower uncertainty compared to spatial extrapolation.

The other reason for the incompatibility of extrapolation might be related to the selected model. RF has been recognized as the most promising model in many DSM studies for handling complex and non-linear relationships between covariates and observation (Hengl et al. 2015). In addition, it has been proven that RF works well in predictions of soil properties where we have enough coverage of training data. Nevertheless, in addition to this powerful ability, RF has a drawback; the model would perform weakly in extrapolation in feature space. Therefore, the extrapolation by RF would be problematic when there are large spaces with no observations, and the new predictors have different characteristics from what the trained model has learned (Meyer and Pebesma 2021).

Malone et al. (2016) quantified the similarity between the donor and recipient areas under homosoil approach with available covariates and extrapolated the model between two areas. The similarity to the donor area was only about 47%, and found that there is limited ability for extrapolation, and similarity between two areas would affect the predictive power of models.

All four methods (homosoil, soil type similarity, dissimilarity index by AOA, and QRF prediction interval width) can be useful to give us information beforehand on how well the

extrapolation might work. There is slight correlation between these methods, e.g. all of them showed neighboring countries are more similar. However, when it comes to extrapolation, it seems DI plots by AOA have the strongest agreements with statistical validation metrics computed from data in the prediction area and can be used as a preliminary document in case of having no/few observations

### 5.2.3. Limitations

There are some limitations to case study two that need to be considered for further studies. One limitation is the quality of the training dataset, from collecting samples in the field to measurements in laboratories, which might cause some errors.

A second limitation is the inaccuracy of the global environmental covariates applied in the homosoil approach, which may not fully capture the soil spatial variation of basic soil forming factors. In addition, some factors related to soil formation are neglected in homosoil, for example, anthropogenic effect, biological factors and age.

Different studies in DSM have pointed out that human activities significantly impact soil formation (Richter et al. 2007; Kuzyakov and Zamanian, 2019), as well as biological factors (Ladd et al. 1996; Meurer et al. 2020; Sothe et al. 2022).

Third, the inability of the trained model to capture all soil spatial variation, as observed in the training model for Burkina Faso in the case of pH and clay, the trained model could only explain less than 20% of spatial variability. The failure of RF in extrapolating in feature space has been discussed by Hengl et al. (2021). Takousting and Heuvelink (2022) showed regression kriging performed better in extrapolation than RF. Selection of the most important features might help to increase the accuracy of predictive models (Meyer et al. 2019; Karasu and Altan, 2019).

## 5.3. Case study three

### 5.3.1. Ensemble machine learning model and multivariate geostatistics

We discovered that the SuperLearner significantly outperformed each single learner in the spatial prediction of pH and SAR (Table 16). This finding confirmed the high performance of ensemble modeling, which reduces noise and variance in predictions by combining the

merits of individual models. At the same time, ensembles avoid overfitting and can produce reliable and acceptable performance rather than any other single model. This finding is consistent with other studies which confirmed the effectiveness of ensemble modeling over using one single model (Cressie, 1993; Geiger et al. 2012; Hengl et al. 2022). Mishra et al.(2020) employed regression kriging in conjunction with several ML techniques for SOC stock mapping. They discovered that an ensemble prediction obtained from all four ML strategies performed superior to any individual model in terms of giving more spatial information and accuracy for estimating the spatial variation in the soil attribute of interest.

Similarly, Taghizadeh-Mehrjardi et al. (2021) stated that SuperLearner can be a favorable approach for spatial prediction of soil properties, as it can produce more reliable and accurate predictions than any single other model. Also, it was noteworthy to discover that even the poorest model contributed to the super learner's creation.

However, it is important to highlight that RF was the top-performing model among the five individual learners, providing remarkably comparable results and nearly equivalent to the SuperLearner. This may be related to RF's capacity to manage data nonlinearity and outliers (Hengl et al. 2015).

RF outperformed the SuperLearner in the predictive mapping of EC; hence RF was employed in multivariate geostatistical modeling instead of the SuperLearner, even though the results of this investigation validated the SuperLearner's outstanding performance for pH and SAR.

The failure of SuperLearner for EC mapping can be due to the relationship between EC and covariates. For example, lack of covariates to explain the heterogeneity of spatial EC, artifacts in covariates, or even a few sampling points (Lück et al. 2009; Jafari et al. 2012). Soil EC is an attribute which changes quickly over time and space (Li et al. 2013; Paz et al. 2020). This might be a reason for the sensitivity of the EC when it comes to modeling and mapping.

Several factors need to be considered for the interpretability of any spatial prediction model, including the accuracy of collecting samples and laboratory measurements, type of covariates and their resolution (Wadoux et al. 2020), soil-landscape interactions (Rossiter, 2018; Hateffard et al. 2019), and type of selected ML model (Khaledian and Miller, 2020). It is challenging to compare the impact of different environmental covariates since the covariates

affect the coefficients both directly (by collinearity) and indirectly (via the nature of the data) (Khaledian and Miller, 2020).

Accordingly, the applicability of each model for each soil attribute varies in each area. As a result, there is no best model which could be used and advised in every circumstance. Multiple individual models, and ensembles of these models, should be evaluated because the ensemble will use each model's potential and, generally speaking, can improve prediction and accuracy. The calculated direct and cross-variograms illustrated in Figure 19 verified our hypothesis that it is preferable to jointly model their spatial distribution utilizing multivariate geostatistics since it showed clearly that SAS indicators are spatially interdependent along the study area. Multivariate geostatistics is widely used in DSM (Odeh et al. 1995; Lark et al. 2014), and Szatmári et al. (2020) has thoroughly reviewed its benefits and drawbacks in SAS mapping. Tziachris et al. (2019) applied hybrid methods, including different ML methods with kriging of their residuals and concluded that the application of joint modeling in spatial prediction of soil organic matter could significantly increase the accuracy.

The regression kriging method incorporates environmental correlation and spatial autocorrelation to predict soil property of interest. Regression kriging, in contrast to other methods, typically results in fewer prediction errors (Hengl et al. 2007; Mishra et al. 2012; Minasny et al. 2013).

The most important finding of this study is that the spatial prediction uncertainty of SAS indicators is in line with spatial cross-correlation between the indicators. This has many advantages, especially when the complex evaluation of the indicators is intended, for example, soil quality management or precision agriculture.

It's important to note that a moderately large nugget variance has been recognized in some of the computed variograms (Figure 19), which is common in DSM (Vaysse et al. 2015). In our study, this concern might be due to the applied sampling strategy.

### 5.3.2. Assessment of predicted map of salt-affected soils

In Hungarian lowlands, salt accumulation and related processes such as sodification and alkalinization are common characteristics (Tóth et al. 2001). In order to increase agricultural production, it is necessary to produce spatial and temporal maps to detect soil properties

variation. Also, in salt-affected soils, which are less suitable for agriculture, it is required to map salinity in high resolution and accuracy to have more productive lands. Besides this, it is important to make probability maps which express the possibility of the presence of limiting factors in the area of interest. This information is crucial for designing agrotechnical activities and giving the proper advice considering the size of the area. For example, the threshold values based on bio-physical criteria in Europe (Van Orshoven et al. 2012), for soil salinity and sodicity are ECe > 4 dS/m and ESP > 6; SAR > 13, respectively. Therefore, the criteria state that 16.9% (0.16 km2) of the study plot would meet the requirement for being subsidized based on a probability value of 0.9.

The exchangeable sodium (or soil sodicity), due to insufficient data for lower salt levels, was not completely delineated, which is a limitation for productive agriculture. As an outcome of agro technical procedures, soil salinity declined to a medium or low level on the surface (plowed horizon). The maximum level of salinity can be found in the C horizon. The map of the spatial distribution of soil EC is followed by topography, and salinity showed correlation with elevation. Likewise, Nabiollahi et al. (2021) evaluated SAS indicators employing DSM and hybridized RF and found that the most important covariates were elevation, groundwater table, categorical maps, salinity index in the predictive mapping of pH, EC, and SAR. This demonstrates that the groundwater table and topography are essential for properly evaluating SAS indicators.

Our information about groundwater levels in this study plot is limited. From a few observations (around 10 locations), we can see that it is about the critical level (according to Kovda et al. 1973), meaning that the risk of accumulation of salts in the fluctuation zone might exist. Nonetheless, it was not possible to consider these values in our spatial prediction. We should note that salinity is an ever-changing parameter that can be easily solved in water or accumulate on the surface depending on water conditions. Still, a detailed spatial and temporal map of soil salinity is required, especially in areas with a probability of secondary salt accumulations.

The spatial map of soil alkalinity shows pH values of more than 8.5 and higher SAR values in the northern part of the study plot (Figure 20). The soil pH values can affect the availability of soil nutrients for absorption in plants and crops. Therefore, depending on plant types, the

soil pH should be preserved at a required level; otherwise can reduce productivity and yield. Therefore, identifying these alkaline parts on high-resolution maps can be helpful for stakeholders to develop cost-effective solutions.

## 6. Conclusion

The first phase of the first case study aimed to explore the applicability of selected ML models in the spatial prediction of soil properties in two different geographical conditions. Our research has brought us to the conclusion that the RF approach, especially when using a small number of observations, produced more reliable results than the other models in both areas. ANN, SVM, and MLR did not produce satisfactory results in terms of accuracy and detailed spatial pattern. Furthermore, the importance of DEM derivatives as representative of relief characteristics in scorpan factor in Látókép was highlighted, while in Westsik the most important predictor for all soil properties was NDVI. Overall, considering the actual situation in both fields and expert knowledge, these predictions and the pattern produced in the final maps by RF seem reliable. It is essential to comprehend the strengths and weaknesses of various ML algorithms before choosing the best ML technique according to specific DSM problems and mapping conditions. Considering the limitations and purpose of the research also can help to select the most proper ML technique. For example, ANN is sensitive to small datasets, as we also noticed in this study.

Látókép and Westsik are both experimental research stations; hence, a detailed and accurate spatial distribution of soil properties across the areas would help stakeholders design sustainable management practices.

The second phase of the first case study was related to extrapolating the trained model by the most promising model to their related microregion by assessing the area of applicability method. From the first phase, we found that RF is the most acceptable model in both areas. Therefore, we predicted soil properties in Hajdúhát by the trained model in Látókép and Nyírség by the trained model in Westsik. At the same time, AOA was applied to select the possible areas to predict based on the similarity between donor and recipient area covariates. Validation was applied inside and outside the AOA, and we concluded that areas inside the AOA were closer to the actual observations. This method, AOA, can be considered a powerful

tool to detect similar areas to the donor area and consequently apply the trained model for predicting soil properties in similar areas in which we have few/no observations. Also, we conclude that different scorpan factors are important in the spatial distribution of soil properties. When prediction over an unknown area is intended, two important factors need to be considered; first, the model should be able to fully capture the soil-landscape interaction, and second, the extrapolation is only possible in geographic space with the assumption of the similarities in scorpan factors between two areas. In other words, similarities in feature space, where the covariate might be different from the donor area, the predictions will expose the risk and should be avoided or used with caution.

From the first study, we understood similarities in soil forming factors are important in transferring the model between two areas. Therefore, the objective of the second case study was to quantify the similarities in soil type, homosoil approach, train the RF model, calculate dissimilarity index by AOA and their uncertainty by QRF prediction interval width. We intended to check the possibility of extrapolating in geographic space. The results showed that all these four methods somehow are in general agreement, for instance, they revealed more similarities in countries from the same region or locations with high dissimilarity by AOA were identified also by high uncertainty. However, when the model trained in one country was applied to predict in the other three countries, the extrapolation showed poor results compared to actual observations in those countries. The trained model might not adequately cover some environmental covariates to deliver reliable predictions.

The trained model generated for soil pH had higher accuracy and performed a bit better in extrapolation. We can conclude that achieving higher accuracy in training the dataset for the donor country can increase the potential of extrapolation to the recipient areas.

We did not separate similar areas and validate the observation within and outside of similar areas since it was out of the scope of this study, but it can be used as further studies.

All these four methods can be helpful to give us information beforehand to discover the potential of transferring the model between two areas and can be used as a preliminary document in case of having no observation and should not be used as final maps. Application of these methods offer a cheap and fast way to generate digital soil maps for areas with scarce soil data, since they can be implemented faster than new soil surveys.

The third case study aimed to estimate the spatial distribution of SAS indicators in the salt-affectedness arable plot in Hungary by jointly modeling the ensemble machine learning and multivariate geostatistics. Our ensemble modeling consisted of five base learners: Ranger, XGBoost, SVM, NN, and GLM. Our results showed that ensemble machine learning integrated with multivariate geostatistics could be a favorable method for delivering a detailed spatial distribution of selected SAS indicators at high spatial resolution and evaluating salt-affected areas on arable lands. Nevertheless, the outcomes demonstrated that ensemble machine learning does not consistently outperform the single models and even in some cases, it is preferable to employ the best base learner alone rather than ensemble modeling. Additionally, the application of multivariate geostatistics was found to be a crucial factor in the success of the approach. Overall, the third case study highlights the importance of jointly modeling the spatial distribution of soil properties and demonstrates the potential of this approach for future soil mapping and management efforts.

## 7. Summary

**Introduction**: DSM has been utilized successfully for several applications since the early 2000s, including precision agriculture, environmental monitoring, and land use planning. DSM is the process of creating maps that represents the spatial distribution of soil properties and characteristics and involves the integration of various types of data, including field-based soil surveys, remote sensing data, and geospatial data. The most common strategies for predicting soil properties and delivering soil spatial and temporal maps are geostatistics and machine learning algorithms. It is true that DSM has the potential to greatly improve our understanding of soil properties, but it also faces several challenges and gaps that limit its accuracy and effectiveness. In this study, we reviewed two of them. One issue is that many parts of the world have few/no observations, and the DSM process requires costly and time-consuming sampling efforts; therefore, the production of accurate soil maps is limited. One of the potential solutions for this problem is the extrapolation of soil properties from areas with observations to those without, which relies on the similarity of soil-forming factors between the two areas.

The second issue is that the spatial aspect of soil data and the interdependence between variables can make modeling difficult since soil is a complex and ever-changing system that interacts with itself and the environment. Multivariate geostatistics is a widely used approach in soil science that considers the joint spatial variability of variables and explicitly takes into account spatial interdependence. Combining multivariate geostatistics with ML algorithms can leverage the strengths of both approaches for more precise soil property predictions and modeling of uncertainty.

Two case studies were defined for extrapolation issues and one for joint spatial modeling.

**Aims**: The first study aimed to compare ML models in predicting and mapping soil properties in small-scale areas and to evaluate the potential and efficiency of extrapolating the best model to larger areas. The second study explored the possibility of extrapolating a ML model for predicting soil properties in one area to another area based on their similarity.

The third study aimed to predict and map salt-affected soils in Hungary using ensemble machine learning and joint modeling with multivariate geostatistical techniques.

**Material and methods:** First case study; Four machine learning models (MLR, RF, ANN, SVM) were trained on the Látókép and Westsik study areas, with the best model selected, fine-tuned and applied to predict soil properties in Hajdúhát and Nyírség. The results were validated by sampling in these areas and applying AOA. Second case study: similarities were identified by using four different methods including similarity in soil types, homosoil approach, dissimilarity index by AOA and QRF prediction interval width, and validates the results using cross-validation, in four countries in Africa.

Third case study: an ensemble modeling approach was used with five individual models (RF, XGboost, SVM, NN, and GLM) on three indicators of salt-affected soils. A multivariate geostatistics analysis was performed on the stochastic residuals obtained from the machine learning modeling. The accuracy of spatial predictions and estimation of uncertainties were evaluated using 10-fold cross-validation.

**Results and discussion:** First case study; In Látókép, the most important environmental variables for predicting soil properties were topographic indices, while in Westsik it was vegetation cover as expressed by NDVI. The results showed that RF outperformed the other models. The results showed that the predictions inside the AOA in Hajdúhát and Nyírség had fewer errors, and it is crucial to limit predictions to areas that are comparable to the training data and are inside the AOA. Second case study; All four methods used in the study somehow were in general agreement. The extrapolation results were not satisfactory, with low performance due to the complexity of soil-landscape interaction and the difficulties of fully matching soil-forming factors. Third case study; The study found that ensemble modeling was effective in mapping and assessing salt-affected soil, producing better results than base learners for two indicators (pH and SAR) with high $R^2$ values. The random forest prediction was found to be acceptable for EC. The methodology used in the study, which included 10-fold cross-validation for performance and uncertainty quantification, was found to be efficient for mapping salt-affected soil with high spatial resolution.

**Conclusion:** First case study; we found that the AOA method was effective in predicting soil properties in areas with similar scorpan factors. The study highlights the importance of considering the strengths and weaknesses of ML algorithms and the similarities in scorpan factors between donor and recipient areas before applying the trained model for predictions

over unknown areas. Second case study; The four methods used in the study provide useful information for discovering the potential of transferring models between areas and can be used as a preliminary step in generating digital soil maps for areas with limited data, faster than new soil surveys. Third case study; Ensemble machine learning and multivariate geostatistics can deliver a detailed spatial distribution of the indicators but may not consistently outperform single models, with the best base learner being used alone in some cases. High-resolution mapping of SAS indicators is crucial for precision agriculture.

## 8. Acknowledgments

I am deeply grateful for the invaluable support and guidance provided by several people without whom this project would not have been possible. Firstly, I would like to express my sincere appreciation to my supervisors, Dr. Tibor Novák and Dr. Gábor Szatmári for their patience, sound advice, and constant support throughout this project. Dr. Novák's depth of expertise and diligent editing have been extremely helpful and valuable to me. I am grateful for his guidance and belief in me as a student, which has continued over the years. I would also like to thank Dr. Gábor Szatmári for showing me the way to conduct research and for believing in my abilities. Dr. Szatmári has also created opportunities for me to collaborate with the Institute of Soil Science in Budapest and the Wageningen University and Research.

Also, I would like to express my sincere gratitude to Wageningen University and Research for their invaluable assistance with the second case study of my dissertation. In particular, I would like to extend my heartfelt thanks to Prof. Dr. Gerard Heuvelink for his insightful comments, extensive knowledge, and excellent leadership throughout the collaboration.

I am also deeply grateful to Dr. Luc Steinbuch for his exceptional support in the technical aspects of the research, as well as his guidance and advice in writing and all other aspects of the project.

I am also grateful to the University of Debrecen and the Department of Landscape Protection and Environmental Geography for their guidance and mentorship, and to the Stipendium Hungaricum Scholarship for providing financial support. I thank my committee members for their insightful comments during my defense.

Finally, I want to express my heartfelt thanks to my husband, family, and numerous friends both inside and outside my home country of Iran for their continuous support and understanding throughout my research and writing of this project. Your unwavering encouragement and belief in me have been a source of inspiration and motivation, and I am deeply grateful for your love and support.

## 9. References


Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics, 16(4), 523-531.

Angelini, M. E., Kempen, B., Heuvelink, G. B. M., Temme, A. J., & Ransom, M. D. (2020). Extrapolation of a structural equation model for digital soil mapping. Geoderma, 367, 114226.

Anjos, L., Gaistardo, C. C., Deckers, J., Dondeyne, S., Eberhardt, E., Gerasimova, M., ... & Zhang, G. L. (2015). World reference base for soil resources 2014 international soil classification system for naming soils and creating legends for soil maps.

Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z, (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174, 776-789.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A. C., Morgan, C. L., ... & Mulder, V. L. (2020). Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Regional, 20, e00255.

Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B., (2014). GlobalSoilMap: Basis of the global spatial soil information system. CRC Press.

Asgari, N., Ayoubi, S., Jafari, A., Demattê, J. A, (2020). Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. International Journal of Remote Sensing, 41(19), 7624-7648.Henderson, P., 1982. Inorganic geochemistry. Pergamon Press, New York.

Awad, M., & Khanna, R. (2015). Efficient learning machines: theories, concepts, and applications for engineers and system designers (p. 268). Springer nature.

Bannari A, Guedon AM, El-Harti A, Cherkaoui F, El-Ghmari A (2008) Characterization of slightly and moderately saline and sodic soils in irrigated agriculture land using simulated data of advanced land imaging (E0-1) sensor. Commun Soil Sci Plant Anal 39:2795– 2281.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D., & Goldschmitt, M. (2005). Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science, 168(1), 21-33.

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. European journal of soil science, 69(5), 757-770.

Bhunia, G. S., Kumar Shit, P., Pourghasemi, H. R, (2019). Soil organic carbon mapping using remote sensing techniques and multivariate regression model. Geocarto International, 34(2), 215-226.

Blum, W. E., Schad, P., & Nortcliff, S. (2017). Essentials of Soil Science: soil formation, functions, use and classification (World Reference Base, WRB). Gebr. Borntraeger Science Publishers.

Boettinger, J. L., Ramsey, R. D., Bodily, J. M., Cole, N. J., Kienast-Brown, S., Nield, S. J., ... & Stum, A. K. (2008). Landsat spectral data for digital soil mapping. In Digital soil mapping with limited data (pp. 193-202). Springer, Dordrecht.

Brady, N. C., Weil, R. R. (2008). The nature and properties of soils (Vol. 15, pp. 989-992). Upper Saddle River, NJ: Prentice Hall.

Brenning, A. (2022). Spatial machine-learning model diagnostics: a model-agnostic distance-based approach. International Journal of Geographical Information Science, 1-23.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239, 68-83.

Brungard, C., Nauman, T., Duniway, M., Veblen, K., Nehring, K., White, D., ... & Anchang, J. (2021). Regional ensemble modeling reduces uncertainty for digital soil mapping. Geoderma, 397, 114998.

Brus, D. J. (2019). Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma, 338, 464-480.

Brus, D. J. (2022). Spatial Sampling with R. CRC Press.

Burrough, P. A. (2001). GIS and geostatistics: Essential partners for spatial analysis. Environmental and ecological statistics, 8(4), 361-377.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., & Bruggeman, A. (2017). A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma, 285, 35-49.

Campbell, P. M. D. M., Fernandes Filho, E. I., Francelino, M. R., Demattê, J. A. M., Pereira, M. G., Guimarães, C. C. B., & Pinto, L. A. D. S. R. (2019). Digital soil mapping of soil properties in the "Mar de Morros" environment using spectral data. Revista Brasileira de Ciência do Solo, 42.

Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., & Li, L. (2019). A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. ISPRS International Journal of Geo-Information, 8(4), 174.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd international conference on knowledge discovery and data mining (pp. 785-794).

Cressie, N.A.C. (1993). Statistics for Spatial Data; Wiley: Hoboken, NJ, USA.

Cochran, W. G. (1977). Sampling techniques. John Wiley & Sons.

da Silva Chagas, C., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena, 139, 232-240.

Dai, F., Zhou, Q., Lv, Z., Wang, X., & Liu, G. (2014). Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators, 45, 184-194.

Davies, B. E. (1974). Loss-on-ignition as an estimate of soil organic matter. Soil Science Society of America Journal, 38(1), 150-151.

Debella-Gilo, M., & Etzelmüller, B. (2009). Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. Catena, 77(1), 8-18.

Dewitte, O., Jones, A., Elbelrhiti, H., Horion, S., & Montanarella, L. (2012). Satellite remote sensing for soil mapping in Africa: An overview. Progress in physical geography, 36(4), 514-538.

Dharumarajan, S., Hegde, R., & Singh, S. K. (2017). Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. Geoderma Regional, 10, 154-162.

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., & Grunwald, S. (2018). A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma, 314, 262-274.

Dövényi, Z., Ambrózy, P., Juhász, Á., Marosi, S., Mezősi, G., Michalkó, G., ... & Tiner, T. (2008). Inventory of microregions in Hungary. OTKA Research Reports| OTKA Research Reports .

FAO/IIASA/ISRIC/ISSCAS/JRC, (2009). Harmonized World Soil Database (version 1.1). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Fathizad, H., Ardakani, M. A. H., Sodaiezadeh, H., Kerry, R., & Taghizadeh-Mehrjardi, R. (2020). Investigation of the spatial and temporal variation of soil salinity using random forests in the central desert of Iran. Geoderma, 365, 114233.

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PloS one, 12(1), e0170478.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

Geiger, J. Some Thoughts on the Pre- and Post-Processing in Sequential Gaussian Simulation and Their Effects on Reservoir Characterization. In New Horizons in Central European Geomathematics, Geostatistics and Geoinformatics; Geiger, J., Pál-Molnár, E., Malvic, T., Eds.; GeoLitera: Szeged, Hungary,2012; pp. 17–34.

Ghaderi, A., Abbaszadeh Shahri, A., Larsson, S, (2019). An artificial neural network based model to predict spatial soil type distribution using piezocone penetration test data (CPTu). Bulletin of Engineering Geology and the Environment, 78(6), 4579-4588.

Ghazali, M. F., Wikantika, K., Harto, A. B., & Kondoh, A. (2020). Generating soil salinity, soil moisture, soil pH from satellite imagery and its analysis. Information Processing in Agriculture, 7(2), 294-306.

Giasson, E., Sarmento, E. C., Weber, E., Flores, C. A., & Hasenack, H. (2011). Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agricola, 68, 167-174.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., & Fernandes Filho, E. I. (2019). Modelling and mapping soil organic carbon stocks in Brazil. Geoderma, 340, 337-350.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press on Demand.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. Geoderma, 103(1-2), 3-26.

Goovaerts, P. Geostatistics for Natural Resources Evaluation; Oxford University Press: New York, NY, USA; Oxford, UK, 1997; ISBN 9780195115383.

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Gray, J. M., Humphreys, G. S., & Deckers, J. A. (2011). Distribution patterns of World Reference Base soil groups relative to soil forming factors. Geoderma, 160(3-4), 373-383.

Grinand, C., Arrouays, D., Laroche, B., & Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma, 143(1-2), 180-190.

Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16.

Hamzehpour, N., Shafizadeh-Moghadam, H., & Valavi, R. (2019). Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. Catena, 182, 104141.

Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M. de L., (2008). Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Hateffard, F., Márta, L., Novák, T.. (2022). Anthrosequence of soils on Aeolian Sand Dunes in Westsik's experimental field, Nyíregyháza, Hungary. Soil Squences Atlas V. Nicolaus Copernicus University Torun. Chapter 11.

Hateffard, F., & Novák, T. J. (2021). Soil sampling design optimization by using conditioned Latin Hypercube sampling (No. ISMC2021-35). Copernicus Meetings.

Hateffard, F., Dolati, P., Heidari, A., & Zolfaghari, A. A. (2019). Assessing the performance of decision tree and neural network models in mapping soil properties. Journal of Mountain Science, 16(8), 1833-1847.

Haygarth, P. M., & Ritz, K. (2009). The future of soils and land use in the UK: soil systems for the provision of land-based ecosystem services. Land use policy, 26, S187-S197.

Hempel, J. W., Hammer, R. D., Moore, A. C., Bell, J. C., Thompson, J. A., & Golden, M. L. (2008). Challenges to digital soil mapping. Digital soil mapping with limited data, 81-90.

Hendriks, C. M. J., Stoorvogel, J. J., Lutz, F., & Claessens, L. (2019). When can legacy soil data be used, and when should new data be collected instead?. Geoderma, 348, 181-188.

Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1-2), 75-93.

Hengl, T., Heuvelink, G. B., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. Computers & geosciences, 33(10), 1301-1315.
Hengl, T. (2021). Extrapolation is tough for trees (tree-based learners), combining learners of different types makes it less tough [online]. www.medium.com.

Hengl, T., & MacMillan, R. A. (2019). Predictive soil mapping with R. Chapter 4.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., ... & Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PloS one, 10(6), e0125814.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

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

Hengl, T., Rossiter, D. G., & Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Research, 41(8), 1403-1422.

Hengl, T.; Parente, L.; Bonannella, C. Spatial and Spatiotemporal Interpolation/Prediction Using Ensemble Machine Learning. 2022. Available online: https://opengeohub.github.io/spatial-prediction-eml/ (accessed on 1 June 2022).

Hernández-Lobato, J. M., & Adams, R. (2015, June). Probabilistic backpropagation for scalable learning of bayesian neural networks. In the International conference on machine learning (pp. 1861-1869). PMLR.

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma, 214, 141-154.

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.

Heuvelink, G. B. M., & Webster, R. (2001). Modelling soil variation: past, present, and future. Geoderma, 100(3-4), 269-301.

Heuvelink, G. B., & Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial Statistics, 100639.

Heuvelink, G. B., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., ... & Sanderman, J. (2021). Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science, 72(4), 1607-1623.

Heuvelink, G. B., Burrough, P. A., & Stein, A. (1989). Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information System, 3(4), 303-322.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Hounkpatin, K. O., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten, T., ... & Welp, G. (2018). Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). Scientific reports, 8(1), 1-16.

IDNP. Indo-Dutch Network Project (2002) A methodology for identification of waterlogging and soil salinity conditions using remote sensing. Central Soil Salinity Research Institute, Karnal, India.

Jafari, A., Finke, P. A., Vande Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. European Journal of Soil Science, 63(2), 284-298.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jenny, H. (1994). Factors of soil formation—a system of quantitative pedology. McGraw-Hill.

John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., Marcus Afu, S, 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. Land, 9(12), 487

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Thiombiano, L., Van Ranst, E., Yemefack, M. , Zougmore R., (eds.), 2013, Soil Atlas of Africa. European Commission, Publications Office of the European Union, Luxembourg. 176 pp.

Józsa, E., & Fábián, S. Á. (2016). Mapping landforms and geomorphological landscapes of Hungary using GIS techniques. Studia Geomorphologica Carpatho-Balcanica, 50, 19.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., & Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environmental Modelling & Software, 19(9), 845-855.

Karasu, S., & Altan, A. (2019, November). Recognition model for solar radiation time series based on random forest with feature selection approach. In 2019 11th international conference on electrical and electronics engineering (ELECO) (pp. 8-11). IEEE.

Kempen, B., Brus, D. J., Heuvelink, G. B., & Stoorvogel, J. J. (2009). Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma, 151(3-4), 311-326.

Kertész, Á., & Křeček, J. (2019). Landscape degradation in the world and in Hungary. Hungarian Geographical Bulletin, 68(3), 201-221.

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81, 401-418.

Kinoshita, R., Roupsard, O., Chevallier, T., Albrecht, A., Taugourdeau, S., Ahmed, Z., & van Es, H. M. (2016). Large topsoil organic carbon variability is controlled by Andisol properties

and effectively assessed by VNIR spectroscopy in a coffee agroforestry system of Costa Rica. Geoderma, 262, 254-265.

Kottek M., Grieser J., Beck C., Rudolf B., Rubel F. (2006): World map of the Köppen-Geiger climate classification up-dated. Meteorol, 15: 259–263.

Kovačević, M., Bajat, B., & Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. Geoderma, 154(3-4), 340-347.

Krasilnikov, P., Carre, F., & Montanarella, L. (2008). Soil geography and geostatistics. Concepts and Applications. JRC Scientific and Technical Reports, 204.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

Kuzyakov, Y., & Zamanian, K. (2019). Reviews and syntheses: Agropedogenesis–humankind as the sixth soil-forming factor and attractors of agricultural soil degradation. Biogeosciences, 16(24), 4783-4803.

Kvoda, V.A.; van den Berg, C.; Hagan, R.M. Irrigation, Drainage and Salinity; Hutchinson/FAO/UNESCO, 1973.

Laborczi, A., Szatmári, G., Kaposi, A. D., & Pásztor, L. (2019). Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma, 352, 360-372.

Ladd, J. N., Foster, R. C., Nannipieri, P., & Oades, J. M. (1996). Soil structure and biological activity. Soil biochemistry, 9, 23-78.

Lagacherie, P. (2008). Digital soil mapping: a state of the art. Digital soil mapping with limited data, 3-14.

Lagacherie, P., & McBratney, A. B. (2006). Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Developments in soil science, 31, 3-22.

Lagacherie, P., Robbez-Masson, J. M., Nguyen-The, N., & Barthès, J. P. (2001). Mapping of reference area representativity using a mathematical soilscape distance. Geoderma, 101(3-4), 105-118.

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.

Lark, R.M.; Ander, E.L.; Cave, M.R.; Knights, K.V.; Glennon, M.M.; Scanlon, R.P. Mapping Trace Element Deficiency by Cokriging from Regional Geochemical Soil Data: A Case Study on Cobalt for Grazing Sheep in Ireland. Geoderma **2014**, 226–227, 64–78.

Lee, S., & Evangelista, D. G. (2006). Earthquake-induced landslide-susceptibility mapping using an artificial neural network. Natural Hazards and Earth System Sciences, 6(5), 687-695.

Li, H.Y.; Shi, Z.; Webster, R.; Triantafilis, J. Mapping the Three-Dimensional Variation of Soil Salinity in a Rice-Paddy Soil. Geoderma 2013, 195, 31–41.

Li, Q. Q., Yue, T. X., Wang, C. Q., Zhang, W. J., Yu, Y., Li, B., Bai, G. C, (2013). Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. Catena, 104, 210-218.

Lin, L. I. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometric, 45, 255-268.

Lück, E., Gebbers, R., Ruehlmann, J., & Spangenberg, U. (2009). Electrical conductivity mapping for precision farming. Near Surface Geophysics, 7(1), 15-26.

Mahmoudabadi, E., Karimi, A., Haghnia, G. H., & Sepehr, A. (2017). Digital soil mapping using remote sensing indices, terrain attributes, and vegetation features in the rangelands of northeastern Iran. Environmental monitoring and assessment, 189(10), 1-20.

Mallavan, B. P., Minasny, B., & McBratney, A. B. (2010). Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In Digital Soil Mapping (pp. 137-150). Springer, Dordrecht.

Malone, B. P., Jha, S. K., Minasny, B., & McBratney, A. B. (2016). Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma, 262, 243-253.

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). Using R for digital soil mapping (Vol. 35). Cham, Switzerland: Springer International Publishing.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., ... & Beaudoin, A. (2014). Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. Geoderma, 235, 59-73.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2), 3-52.

Menezes, M. D. D., Silva, S. H. G., Owens, P. R., & Curi, N. (2013). Digital soil mapping approach based on fuzzy logic and field expert knowledge. Ciência e Agrotecnologia, 37, 287-298.

Meurer, K., Barron, J., Chenu, C., Coucheney, E., Fielding, M., Hallett, P., ... & Jarvis, N. (2020). A framework for modelling soil structure dynamics induced by biological activity. Global change biology, 26(10), 5382-5403.

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12(9), 1620-1633.

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. Ecological Modelling, 411, 108815.

Minasny, B., & Mc Bratney, A. B. (2002). Uncertainty analysis for pedotransfer functions. European Journal of Soil Science, 53(3), 417-429.

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & geosciences, 32(9), 1378-1388.

Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I. (2013). Chapter One -Digital mapping of soil carbon. Adv. Agron. 118, 1–47.

Minasny, B., McBratney, A., & Lark, R. M. (2008). Digital soil mapping technologies for countries with sparse data infrastructures. Digital soil mapping with limited data, 15-30.

Minasny, B., Setiawan, B. I., Saptomo, S. K., & McBratney, A. B. (2018). Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma, 313, 25-40.

Mishra, U., Gautam, S., Riley, W. J., & Hoffman, F. M. (2020). Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. Frontiers in big Data, 3, 528441.

Mishra, U., Torn, M. S., Ogle, S., and Masanet, E. (2012). Improving regional soil carbon inventories: combining IPCC carbon inventory method with regression kriging. Geoderma, 189–190, 288–295.

Mishra, U.; Gautam, S.; Riley, W.J.; Hoffman, F.M. Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region. Front. Big Data 2020, 3, 528441.

Moody, J, 1994. Prediction risk and architecture selection for neural networks. In From statistics to neural networks, 147-165. Springer, Berlin, Heidelberg.

Mulder, V. L., De Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping—A review. Geoderma, 162(1-2), 1-19.

Nabiollahi, K.; Taghizadeh-Mehrjardi, R.; Shahabi, A.; Heung, B.; Amirian-Chakan, A.; Davari, M.; Scholten, T. Assessing Agricultural Salt-Affected Land Using Digital Soil Mapping and Hybridized Random Forests. Geoderma **2021**, 385, 114858.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3), 282-290.

Natarajan, V. A., Kumar, M. S., Tamizhazhagan, V., & Chevdumoi, R. M. (2022). Prediction Of Soil Ph From Remote Sensing Data Using Gradient Boosted Regression Analysis. Journal of Pharmaceutical Negative Results, 29-36.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., ... & Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil, 4(1), 1-22.

Odeh, I.O.A.; McBratney, A.B.; Chittleborough, D.J. Further Results on Prediction of Soil Properties from Terrain Attributes: Heterotopic Cokriging and Regression-Kriging. Geoderma **1995**, 67, 215–226.

Oliver, M. A. (1987). Geostatistics and its application to soil science. Soil use and management, 3(1), 8-20.

Pahlavan, M. R. (2015). Digital soil mapping using Random Forest model in Golestan province. Journal of Water and Soil Conservation, 21(6), 73-93.

Panagos, P., Van Liedekerke, M., Jones, A., & Montanarella, L. (2012). European Soil Data Centre: Response to European policy support and public data requirements. Land use policy, 29(2), 329-338.

Pásztor, L., Laborczi, A., Takács, K., Szatmári, G., Dobos, E., Illés, G., ... & Szabó, J. (2015). Compilation of novel and renewed, goal oriented digital soil maps using geostatistical and data mining tools. Hungarian Geographical Bulletin, 64(1), 49-64.

Pásztor, L., Szabó, K. Z., Szatmári, G., Laborczi, A., & Horváth, Á. (2016). Mapping geogenic radon potential by regression kriging. Science of the total environment, 544, 883-891.

Paz, A.M.; Castanheira, N.; Farzamian, M.; Paz, M.C.; Gonçalves, M.C.; Monteiro Santos, F.A.; Triantafilis, J. Prediction of Soil Salinity and Sodicity Using Electromagnetic Conductivity Imaging. Geoderma 2020, 361, 114086.

Pebesma, E. J., & Heuvelink, G. B. (1999). Latin hypercube sampling of Gaussian random fields. Technometrics, 41(4), 303-312.

Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Santos, N. T., & Fernandes-Filho, E. I. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. Precision Agriculture, 1-16.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil, 7(1), 217-240.

Richter Jr, D. D. (2007). Humanity's transformation of Earth's soil: Pedology's new frontier. Soil science, 172(12), 957-967.

Rossiter, D. G (2021). Soil mapping today: computer-generated predictive soil maps–their role in soil survey and land evaluation. Guidelines for Authors Agriculture for Development, 26.

Rossiter, D. G. (2018). Past, present & future of information technology in pedometrics. Geoderma, 324, 131-137.

Salehi Hikouei, I., Kim, S. S., & Mishra, D. R. (2021). Machine-learning classification of soil bulk density in salt marsh environments. Sensors, 21(13), 4408.

Scarpone, C., Schmidt, M. G., Bulmer, C. E., & Knudby, A. (2016). Modelling soil thickness in the critical zone for Southern British Columbia. Geoderma, 282, 59-69.

Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis lectures on data mining and knowledge discovery, 2(1), 1-126.

Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019). Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350, 84-92.

Shirani, H., Habibi, M., Besalatpour, A. A., & Esfandiarpour, I. (2015). Determining the features influencing physical quality of calcareous soils in a semiarid region of Iran using a hybrid PSO-DT algorithm. Geoderma, 259, 1-11.

Silva, S. H. G., de Menezes, M. D., Owens, P. R., & Curi, N. (2016). Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma, 267, 65-77.

Song, X. D., Wu, H. Y., Ju, B., Liu, F., Yang, F., Li, D. C., ... & Zhang, G. L. (2020). Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China. Geoderma, 363, 114145.

Song, Y. Q., Yang, L. A., Li, B., Hu, Y. M., Wang, A. L., Zhou, W., ... & Liu, Y. L. (2017). Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. Sustainability, 9(5), 754.

Sothe, C., Gonsamo, A., Arabian, J., & Snider, J. (2022). Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma, 405, 115402.

Stein, A.; Corsten, L.C.A. Universal Kriging and Cokriging as a Regression Procedure. Biometrics **1991**, 47, 575–587

Steinbuch, L., Brus, D. J., & Heuvelink, G. B. (2022). Mapping depth to Pleistocene sand with Bayesian generalized linear geostatistical models. European Journal of Soil Science, 73(1), e13140.

Szatmári, G., Barta, K., & Pásztor, L. (2015). An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. Hungarian Geographical Bulletin, 64(1), 35-48.
Szatmári, G., & Pásztor, L. (2019). Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma, 337, 1329-1340.

Szatmári, G., László, P., Takács, K., Szabó, J., Bakacsi, Z., Koós, S., & Pásztor, L. (2019). Optimization of second-phase sampling for multivariate soil mapping purposes: Case study from a wine region, Hungary. Geoderma, 352, 373-384.

Szatmári, G.; Pásztor, L. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma **2019**, 337, 1329–1340.

Szatmári, G., Bakacsi, Z., Laborczi, A., Petrik, O., Pataki, R., Tóth, T., & Pásztor, L. (2020). Elaborating Hungarian segment of the global map of salt-affected soils (GSSmap): national contribution to an international initiative. Remote Sensing, 12(24), 4073.

Szatmári, G., Pásztor, L., & Heuvelink, G. B. (2021). Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. Geoderma, 403, 115356.

Taghizadeh-Mehrjardi, R., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., & Scholten, T. (2022). Semi-supervised learning for the spatial extrapolation of soil information. Geoderma, 426, 116094.

Tajik, S., Ayoubi, S., & Lorenz, N. (2020). Soil microbial communities affected by vegetation, topography and soil properties in a forest ecosystem. Applied Soil Ecology, 149, 103514.

Takoutsing, B., & Heuvelink, G. B. (2022). Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma, 428, 116192.

Takoutsing, B., Heuvelink, G. B., Stoorvogel, J. J., Shepherd, K. D., & Aynekulu, E. (2022). Accounting for analytical and proximal soil sensing errors in digital soil mapping. European Journal of Soil Science, 73(2), e13226.

Tang, W., Li, Y., Yu, Y., Wang, Z., Xu, T., Chen, J., ... & Li, X. (2020). Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms. Chemosphere, 253, 126666.

Thompson, J. A., Pena-Yewtukhiw, E. M., & Grove, J. H. (2006). Soil–landscape modeling across a physiographic region: Topographic patterns and model transportability. Geoderma, 133(1-2), 57-70.

Tóth, T.; Gallai, B.; Novák, T.; Czigány, S.; Makó, A.; Kocsis, M.; Árvai, M.; Mészáros, J.; László, P.; Koós, S.; et al. Practical Evaluation of Four Classification Levels of Soil Taxonomy, Hungarian Classification and WRB in Terms of Biomass Production in a Salt-Affected Alluvial Plot. Geoderma **2022**, 410, 115666.

Tóth, T.; Várallyay, G. Past, Present and Future of the Hungarian Classification of Salt-Affected Soils. Soil Classif. **2001**, 125–135.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231.

Tziachris, P., Aschonitis, V., Chatzistathis, T., & Papadopoulou, M. (2019). Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. Catena, 174, 206-216.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).

Van Orshoven, J.; Terres, J.-M.; Tóth, T. Updated Common Bio-Physical Criteria to Define Natural Constraints for Agriculture in Europe; Office for Official Publications of the European Communities: Luxembourg, 2012.

Vašát, R., Heuvelink, G. B. M., & Borůvka, L. (2010). Sampling design optimization for multivariate soil mapping. Geoderma, 155(3-4), 147-153.

Vaysse, K., & Lagacherie, P. (2015). Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional, 4, 20-30.

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, 291, 55-64.

Vaysse, K.; Lagacherie, P. Evaluating Digital Soil Mapping Approaches for Mapping GlobalSoilMap Soil Properties from Legacy Data in Languedoc-Roussillon (France). Geoderma Reg. **2015**, 4, 20–30.

Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.

Wadoux, A. M. C., Brus, D. J., & Heuvelink, G. B. (2019). Sampling design optimization for soil mapping with random forest. Geoderma, 355, 113913.

Wadoux, A. M. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists. John Wiley & Sons.

Weiss, D. J., Atkinson, P. M., Bhatt, S., Mappin, B., Hay, S. I., & Gething, P. W. (2014). An effective approach for gap-filling continental scale remotely sensed time-series. ISPRS Journal of Photogrammetry and Remote Sensing, 98, 106-118.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52, 394-403.

Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B., ... & Zhou, C. (2020). Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. Geoderma, 369, 114337.

Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. Springer Science & Business Media.

Zhao, Y. C., & Shi, X. Z. (2010). Spatial prediction and uncertainty assessment of soil organic carbon in Hebei Province, China. In Digital Soil Mapping (pp. 227-239). Springer, Dordrecht.

Zhao, Z., Chow, T. L., Rees, H. W., Yang, Q., Xing, Z., Meng, F. R, 2009. Predict soil texture distributions using an artificial neural network model. Computers and electronics in agriculture, 65(1), 36-48.

Zhao, Z., Yang, Q., Benoy, G., Chow, T. L., Xing, Z., Rees, H. W., & Meng, F. R. (2010). Using artificial neural network models to produce soil organic carbon content distribution maps across landscapes. Canadian Journal of Soil Science, 90(1), 75-87.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Science of The Total Environment, 729, 138244.

Zhu, A. X., Liu, J., Du, F., Zhang, S. J., Qin, C. Z., Burt, J., ... & Scholten, T. (2015). Predictive soil mapping with limited sample data. European Journal of Soil Science, 66(3), 535-547.

## 10. Supplementary Material; case study two

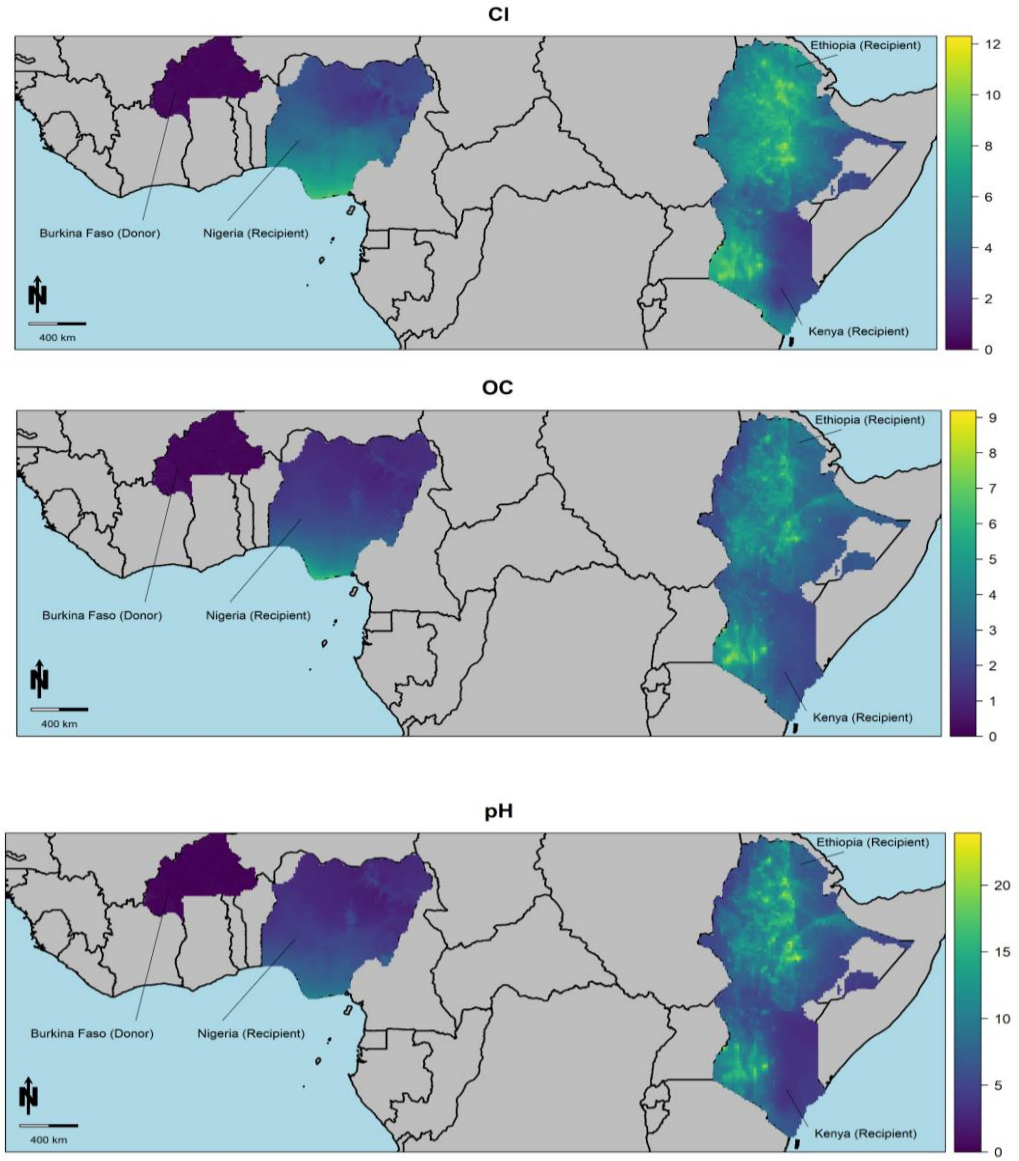



*Figure SM1. Dissimilarity index (DI) maps; BurkinaFaso is a donor country and other countries are recipients.*

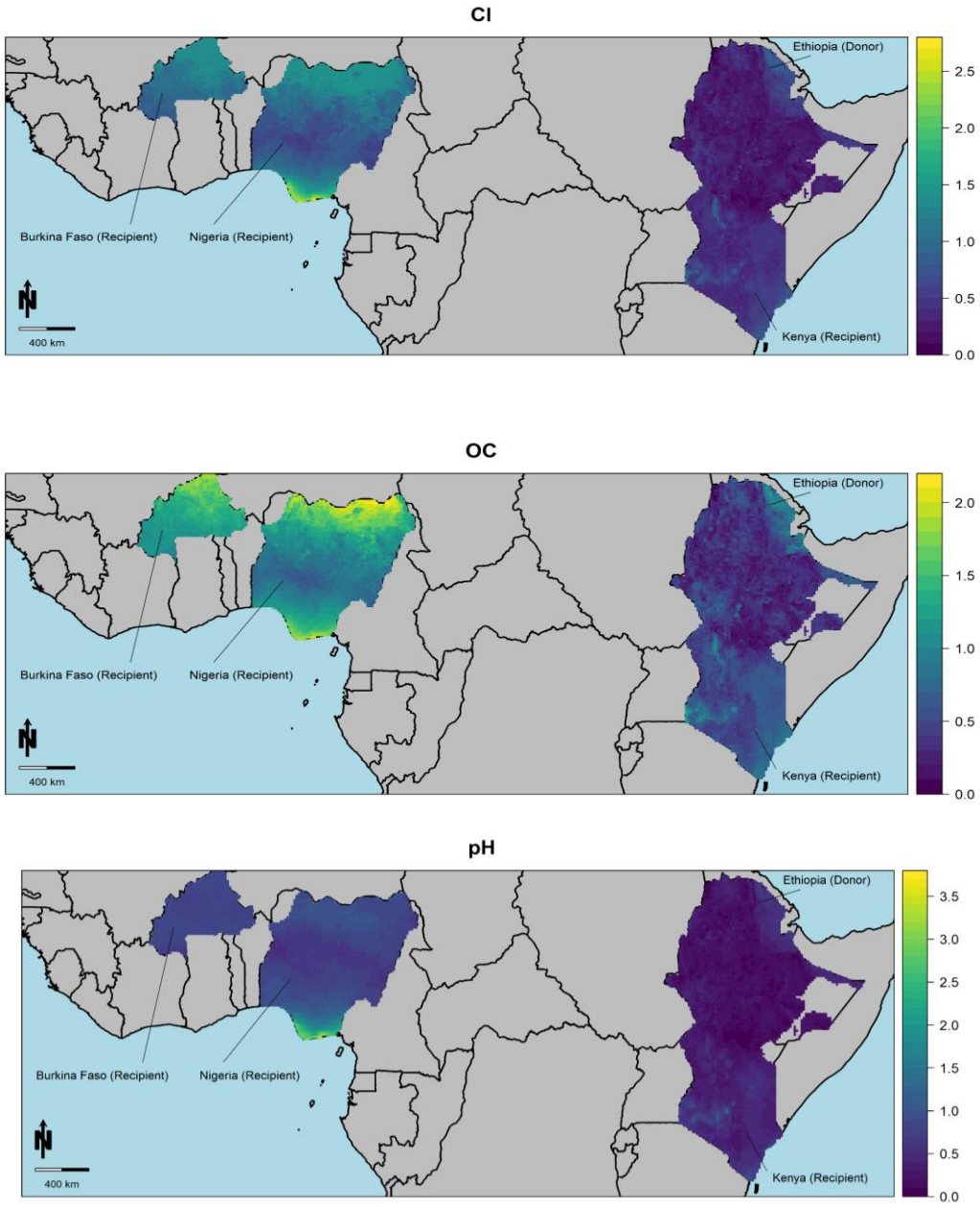



*Figure SM2. Dissimilarity index (DI) maps; Ethiopia is a donor country and other countries are recipients.*

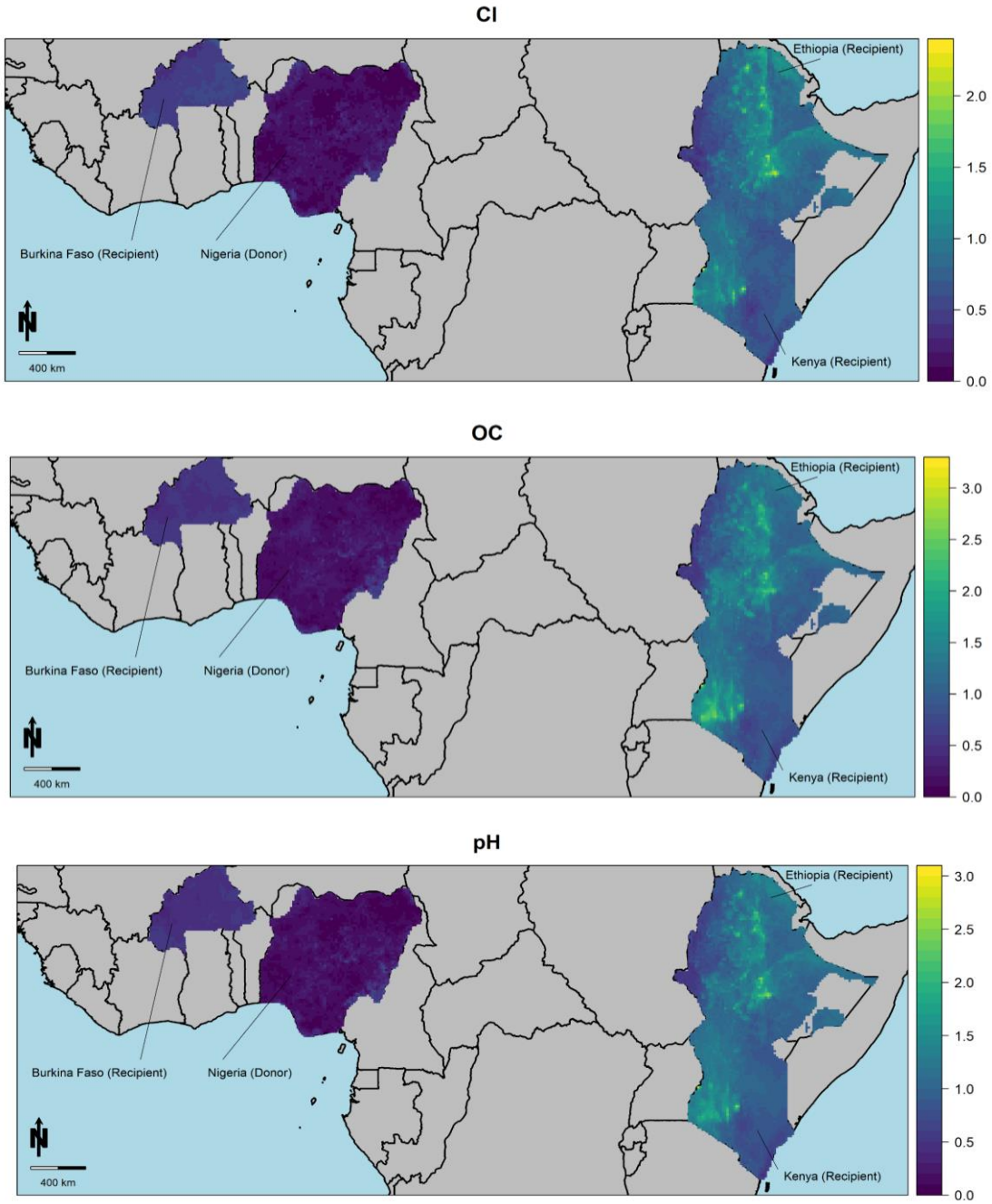



*Figure SM3. Dissimilarity index (DI) maps; Nigeria is a donor country and other countries are recipients.*

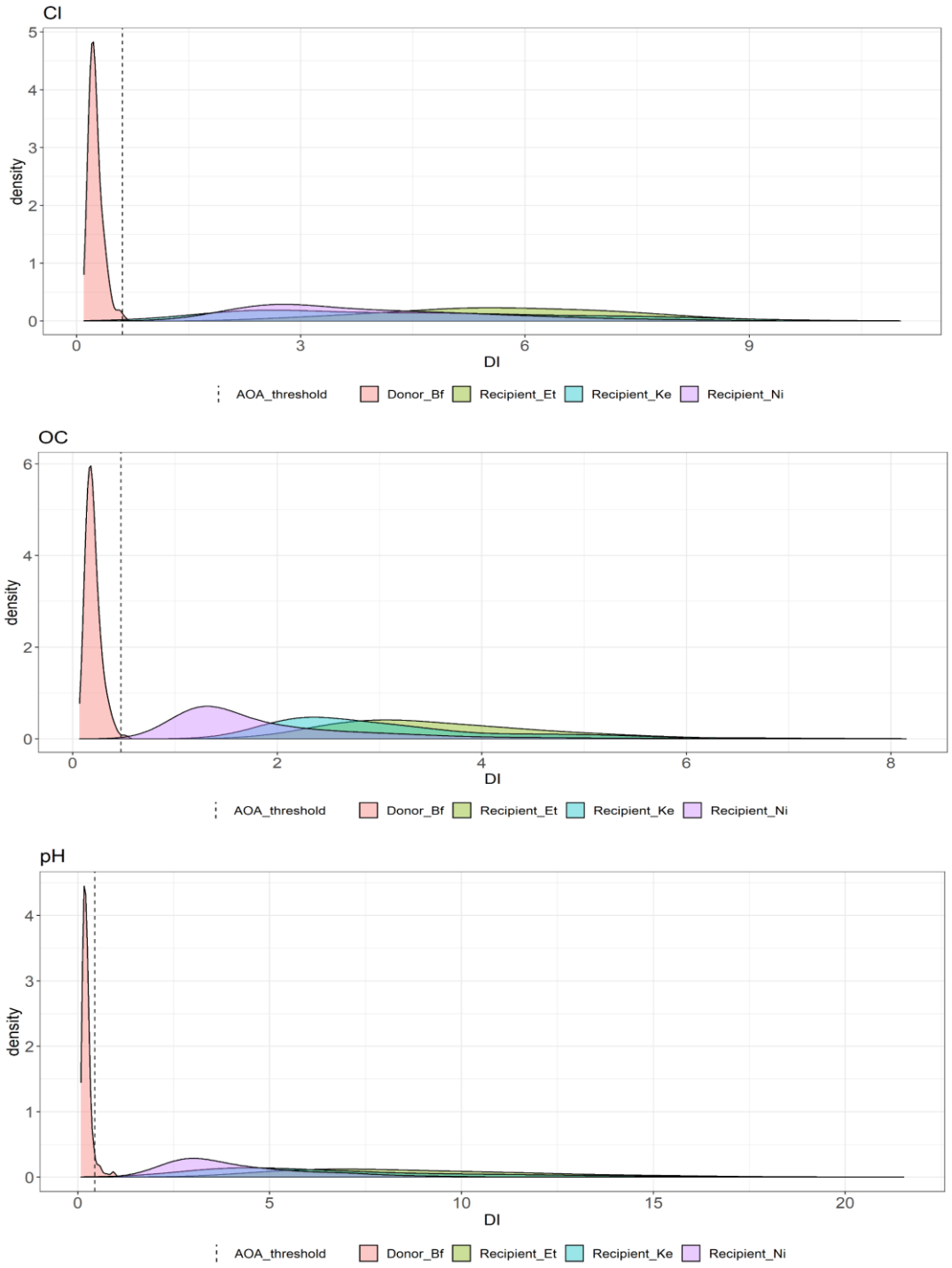



*Figure SM4. Distribution of the prediction DI, BurkinaFaso is a donor country and other countries are recipients.*

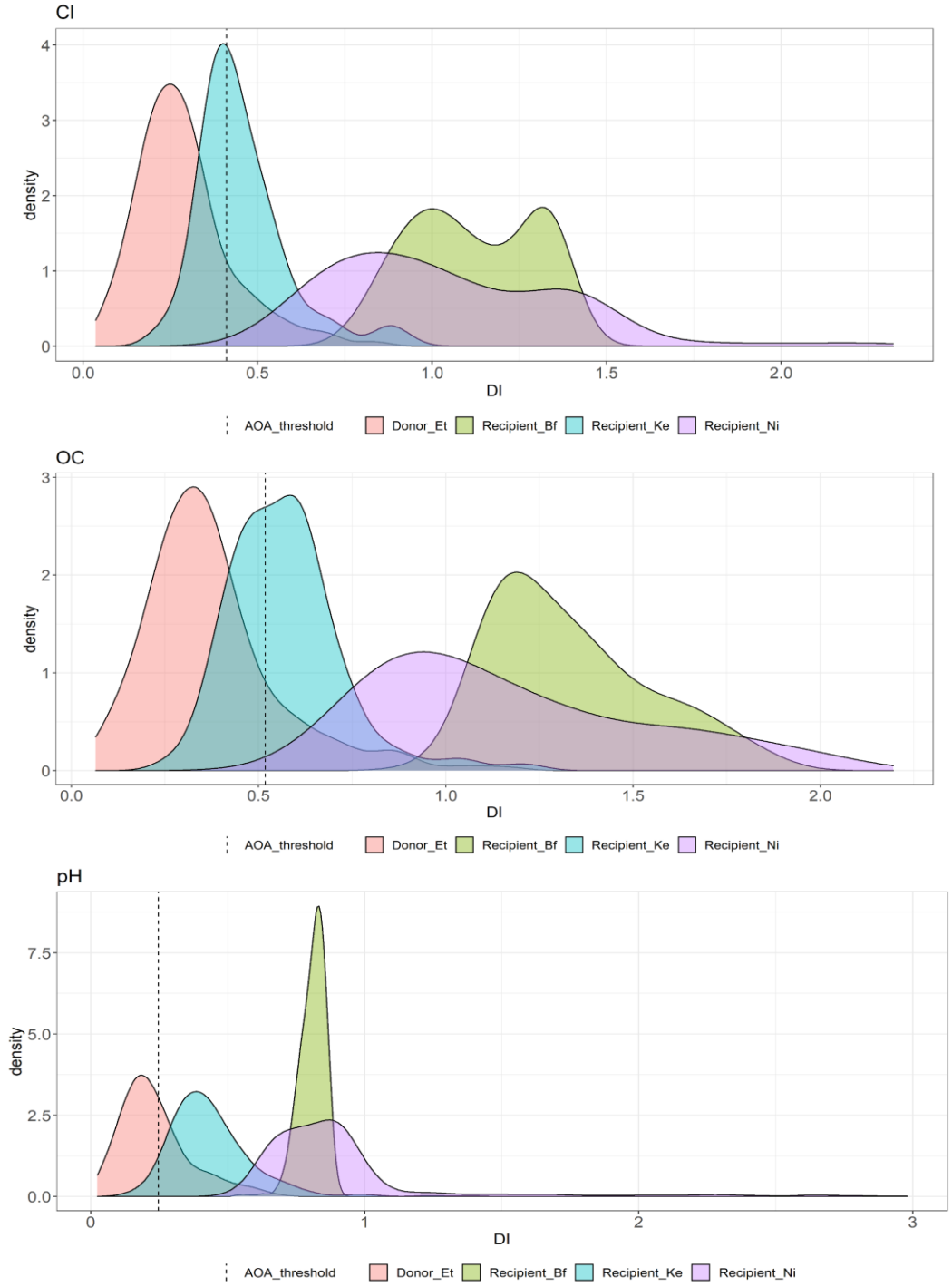



*Figure SM5. Distribution of the prediction DI, Ethiopia is a donor country and other countries are recipients.*

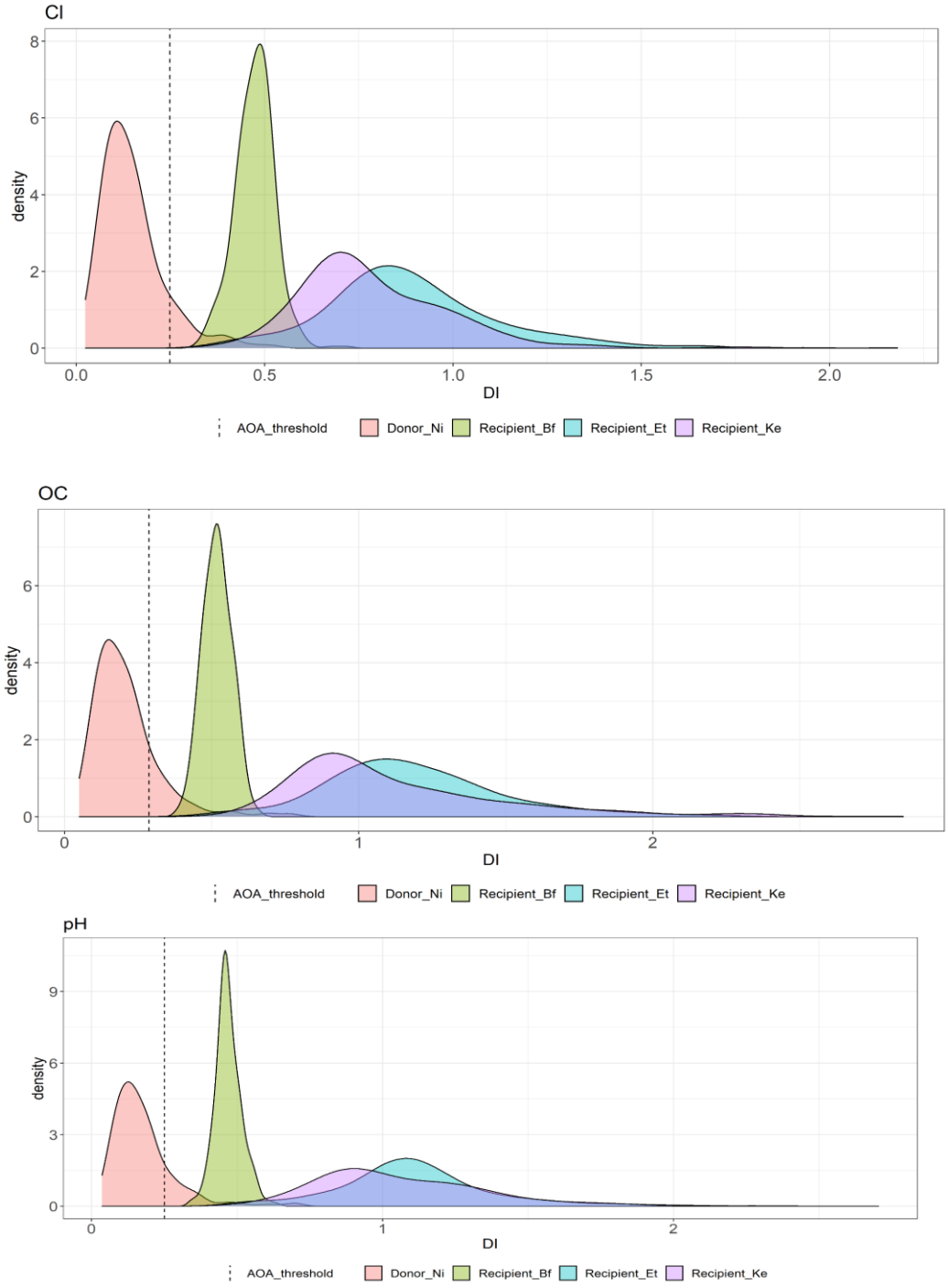



*Figure SM6. Distribution of the prediction DI, Nigeria is a donor country and other countries are recipients.*

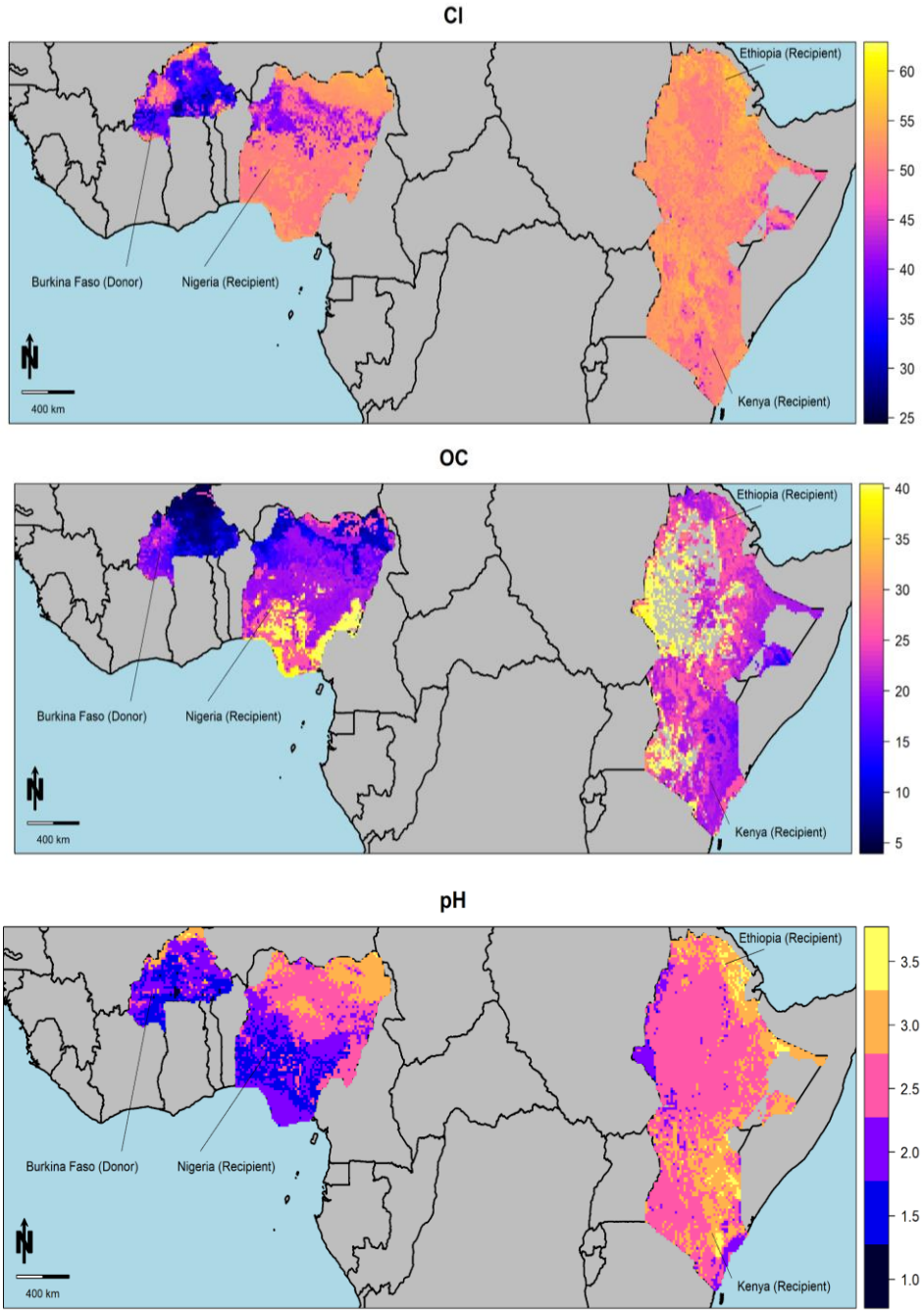



*Figure SM7. Prediction interval width maps, when Burkina Faso is a donor country and other countries are recipients.*

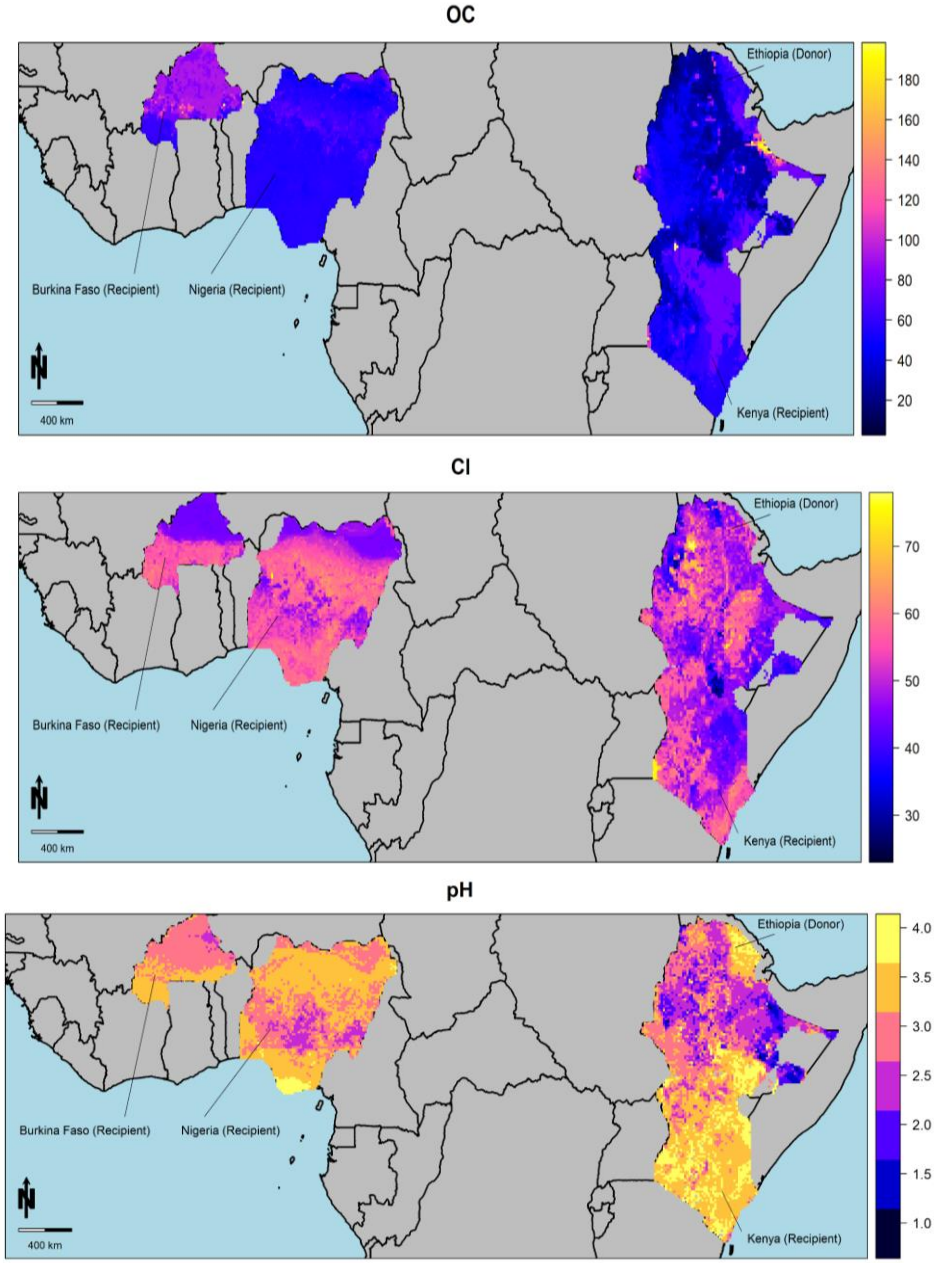



*Figure SM8. Prediction interval width maps, when Ethiopia is a donor country and other countries are recipients.*

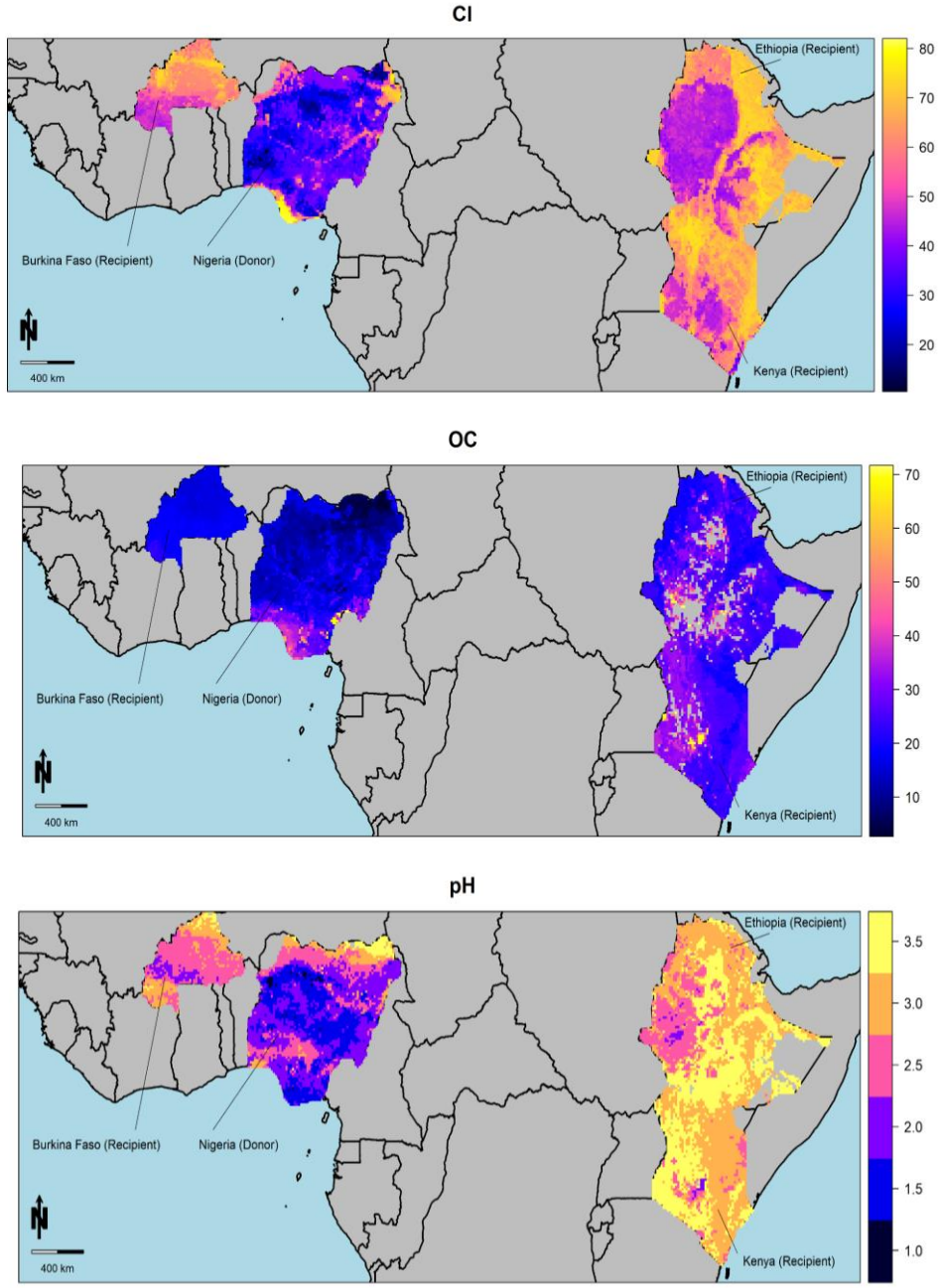



*Figure SM9. Prediction interval width maps, when Nigeria is a donor country and other countries are recipients.*

## 11. List of Publication

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/114/2023.PL
Subject: PhD Publication List

Candidate: Fatemeh Hateffard
Doctoral School: Doctoral School of Earth Sciences

**List of publications related to the dissertation**

Foreign language international book chapters (1)

1. **Hateffard, F.**, Márta, L., Novák, T.: Anthrosequence of soils on Aeolian Sand Dunes in Westsik`s experimental field, Nyíregyháza, Hungary.
   In: Soil sequences Atlas V.. Ed.:Marcin Świtoniak, Przemysław Charzyński, Nicolaus Copernicus University, Torun, 167-180, 2022. ISBN: 9788323149606

Foreign language scientific articles in international journals (2)

2. **Hateffard, F.**, Balog, K., Tóth, T., Mészáros, J., Árvai, M., Kovács, Z. A., Szűcs, V. N., Koós, S., László, P., Novák, T., Pásztor, L., Szatmári, G.: High-Resolution Mapping and Assessment of Salt-Affectedness on Arable Lands by the Combination of Ensemble Learning and Multivariate Geostatistics.
   *Agronomy-Basel. 12* (8), 1-19, 2022. EISSN: 2073-4395.
   DOI: http://dx.doi.org/10.3390/agronomy12081858
   IF: 3.949 (2021)
3. **Hateffard, F.**, Mohammed, S., Alsafadi, K., Enaruvbe, G. O., Heidari, A., Abdo, H. G., Rodrigo-Comino, J.: CMIP5 climate projections and RUSLE-based soil erosion assessment in the central part of Iran.
   *Sci. Rep. 11* (1), 1-17, 2021. EISSN: 2045-2322.
   DOI: http://dx.doi.org/10.1038/s41598-021-86618-z
   IF: 4.996

DEBRECENI EGYETEM Egyetemi és Nemzeti Könyvtár

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Foreign language abstracts (1)

4. **Hateffard, F.**, Novák, T.: Soil sampling design optimization by using conditioned Latin Hypercube sampling.
In: 3rd ISMC Conference - Advances in Modeling Soil Systems, [s.n.], [s.l.], 85, 2021.

**Total IF of journals (all publications): 8,945**
**Total IF of journals (publications related to the dissertation): 8,945**

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

18 April, 2023