SOIL SCIENCE ANNUAL
2023, 74(1), 165879

https://doi.org/10.37501/soilsa/165879

Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions

Fatemeh Hateffard*[1], Gábor Szatmári[2,3], Tibor József Novák[1]

[1] University of Debrecen, Department of Landscape Protection and Environmental Geography, H-4032 Debrecen, Hungary
[2] Institute for Soil Sciences, Centre for Agricultural Research, Department of Soil Mapping and Environmental Informatics, H-1022 Budapest, Hungary
[3] University of Debrecen, Department of Physical Geography and Geoinformatics, H-4032 Debrecen, Hungary

* PhD candidate, Fatemeh Hateffard, Hateffard.fatemeh@science.unideb.hu, ORCID ID: https://orcid.org/0000-0003-0265-4991

Received: 2022-10-18
Accepted: 2023-05-04
Published online: 2023-05-04
Associated editor: Ł. Mendyk

Keywords:

Digital soil mapping
Soil variations
Machine learning
Soil properties
Random forest

Abstract

A reliable overview of the spatial distribution of soil properties is a straightforward approach in soil policies and decision-making. Soil organic carbon (SOC) content, SOC stock and bulk density (BD) directly affect soil quality and fertility. Therefore, an accurate assessment of these crucial soil parameters is required. To do this, we used machine learning algorithms (MLAs) including, multiple linear regression (MLR), random forest (RF), artificial neural network (ANN), and support vector machine (SVM) with the help of environmental covariates to predict SOC content, BD, and SOC stock. The study was conducted in two different areas, Látókép and Westsik (East Hungary), both experimental research fields but different from physio geographic points of view. Thirty topsoils (0–10 cm) samples were collected for each study area using conditioned Latin Hypercube Sampling strategy. Environmental covariates were extracted from a digital elevation model (DEM) and satellite images based on the representation of soil forming factors. We validated the results by randomly splitting the dataset into a train (two-third) and test (one-third) and calculated the root mean square error and R^2. Our results showed that RF provided the most accurate spatial prediction with R^2 of about 80% for each soil property in both study areas. This study highlighted the importance of terrain attributes (including plan and profile curvature, elevation and valley depth) and NDVI derived from satellite images in presenting a spatial distribution of selected soil properties in two different areas. We conclude that comparing these methods can help to determine the most accurate maps under diverse geographical conditions and heterogeneities at different scales, which can be used in precision soil quality management.

1. Introduction

Application of digital soil mapping (DSM) and prediction models for diverse soil characteristics, like soil organic carbon, macronutrient, or moisture content of soils, are widely popular and useful tools for supporting agricultural production (Ma et al., 2019; Owens et al., 2020). Nowadays, thanks to the development of technology and remote sensing, continuous modeling of the soil surface is possible due to the availability of ancillary information, which was the main limitation in conventional soil survey. Therefore, in DSM, this auxiliary information representing the main soil forming factors, along with soil observations, can predict the spatial distribution of soil properties by capturing the relation between these variables using machine learning methods (McBratney et al., 2003). The main concept behind DSM is based on the soil development theory proposed by Jenny (1994), which outlines the five key soil forming factors known as *clorpt* (climate, organisms, relief or topography, parent material, and time). This theory was later refined to the *scorpan* framework (McBratney et al., 2003), where "*s*" represents the soil properties of interest and "*n*" refers to the spatial position. The updated framework describes the quantifiable relationship between soil properties and environmental factors at a specific location in space, providing a spatial model for DSM. In this way, even with limited number of soil observations, it is feasible to map the entire area in which remote sensing data exists, and the soil variation can be effectively captured by the relation between them (Mora-Vallejo et al., 2008). The development of DSM techniques has expanded in recent years, resulting in the availability of digital soil maps for both qualitative and quantitative soil attributes (Behrens et al., 2005; Dai et al., 2014), from detailed regional maps (Hateffard et al., 2022) to global soil maps (Poggio et al., 2021). Machine learning algorithms (MLAs) are frequently used for DSM

and have demonstrated many advantages, such as the ability to handle a variety of variables and the flexibility to model the relationship between dependent and independent variables (Hengl et al., 2015; Chen et al., 2019). However, it is important to use caution when applying MLAs, as it may be vulnerable to over-fitting and lack transparency (Arrouays et al., 2020). The accuracy of predictive models in DSM is crucial as it influences the quality of the predictions, which are used to support policy-making decisions. The selection of the right approach and the identification of the most accurate predictive model, however, is challenging due to the many factors involved in the modeling process. This makes it difficult to find the appropriate approach and the most accurate model for a given dataset (Wadoux, 2020).

There are various MLAs that can be used for regression or classification tasks in DSM, including decision trees (DT) (Hateffard et al., 2019), random forest (RF) (Hengl et al., 2015; Hounkpatin et al., 2018), artificial neural networks (ANN) (Dai et al., 2014), and support vector machines (SVM) (Kovačević et al., 2010). Different studies have compared the accuracy of these models in prediction of soil properties (e.g. Zhao and Shi, 2010; Hengl et al., 2018). Zeraatpisheh et al. (2019) compared different methods including Cubist, RF, DT and multiple linear regression (MLR) to estimate the spatial distribution of soil organic carbon, clay content and calcium carbonate equivalent in a semi-arid region in Iran. The best models were found to be Cubist and RF for predicting soil properties in this area.

In their study, Zhao and Shi (2010) evaluated various techniques, including MLR, DT, ANN with kriging and universal kriging, for predicting the spatial distribution of organic carbon in soil. The results showed that decision trees had the best performance, delivering 67% of the total variation. Heung et al. (2016) used ten different MLAs and 20 environmental covariates in a DSM approach to predict soil great groups and orders. The models, such as regression trees and RF, were preferred for their speed and interpretable results, while the k-nearest neighbor and SVM showed the highest accuracy at 72%. The authors also pointed out that the choice of model and sampling design can greatly affect the outcomes.

Since these techniques are standardized and applied independently from landscape types, they have to function evenly reliable under diverse geographical conditions showing up heterogeneities at different scales. Comparison of these methods can be a powerful tool for finding the most accurate maps, which later can be used as inputs in precision soil quality management (Stoorvogel et al., 2015; Balducci et al., 2018). In some cases, within cultivation parcels with small sizes, it is urgent to have an overview of the spatial variability of soils to cut down the expenses related to sampling and increase productivity. In this study, we want to point out these objectives in two areas which differ from topography, parent materials, and soil texture; i) predict the spatial distribution of soil properties, including soil organic carbon (SOC) content, bulk density (BD), and SOC stock using machine learning techniques; ii) assess the efficiency and potential of different models for both study areas; iii) dependency of model construction and applicability of these methods for different physiographic conditions. We supposed that within these areas, the spatial variability of soil properties is largely influenced by on the variability of topography. Therefore, this study not only provides a comprehensive comparison of different machine learning techniques but also provides insight into how these methods perform in different physio geographic conditions, making it a valuable contribution to the field of soil mapping and management.

2. Materials and methods

2.1. Study sites

Two study sites (Fig. 1) with a comparable extent on the Great Hungarian Plain were selected for mapping of soil characteristics in detail. Both are managed in the framework of long-term crop rotation and plant nutrition for the University of Debrecen. The climate of the study areas belongs to the warm temperate, fully humid climate with warm summers (Kottek et al., 2006). Both are located on an early Pleistocene alluvial plain, reshaped partially or totally by aeolian processes (sand and dust re-deposition, and loess deposition) during the late Pleistocene.

The Westsik Vilmos crop rotation experimental station (47°59'21'' – 47°58'35'' N; 21°41'57'' – 21°42'18'' E) was founded in 1929 with a area of 47 ha in the direct vicinity of Nyíregyháza, on a slightly undulating sandy landscape, at 101–105 m a.s.l, with 1.5–5% average slope. It aimed to improve sand textured, poorly aggregated soils, poor in SOC. It is one of the oldest in Europe, and the management involves the regular application of standardized amounts of fertilizers and plowing green manure crops into the soil.

The Látókép long-term field experiment (47°32'30'' – 47°33'44'' N; 21°26'15'' – 21°27'11'' E) is located on a plain area with flat topography, at about 111–114 m a.s.l.. It has few elevation differences and a 0.5–1% average slope, with deep, well-aggregated, mollic topsoil, with secondary carbonate accumulations. Long-term experiments were founded here in 1983 and recently include fertilization, cultivation, and irrigation experiments over 160 ha.

2.2. Soil sampling and analysis

Thirty sampling points were determined based on the conditioned Latin Hypercube Sampling (cLHS) method to reasonably cover the spatial variation of soil properties in both study areas. The cLHS is a kind of stratified random sampling design efficient locations based on the variability of environmental covariates in feature space (Brungard and Boettinger, 2010). The covariates mainly extracted from DEM and Landsat images (see Section 2.3), and then principal component analysis was used to select the most important features. Therefore, the input data of the cLHS were elevation, slope, topographic wetness index, the normalized difference vegetation index (NDVI), Band 3, 7, and 9 of Landsat 8 images. The R package "*clhs*" was used for this sampling design (Hateffard and Novák, 2021). Each sampling point was sampled in triplicates with 100 cm^3 metal cylinders

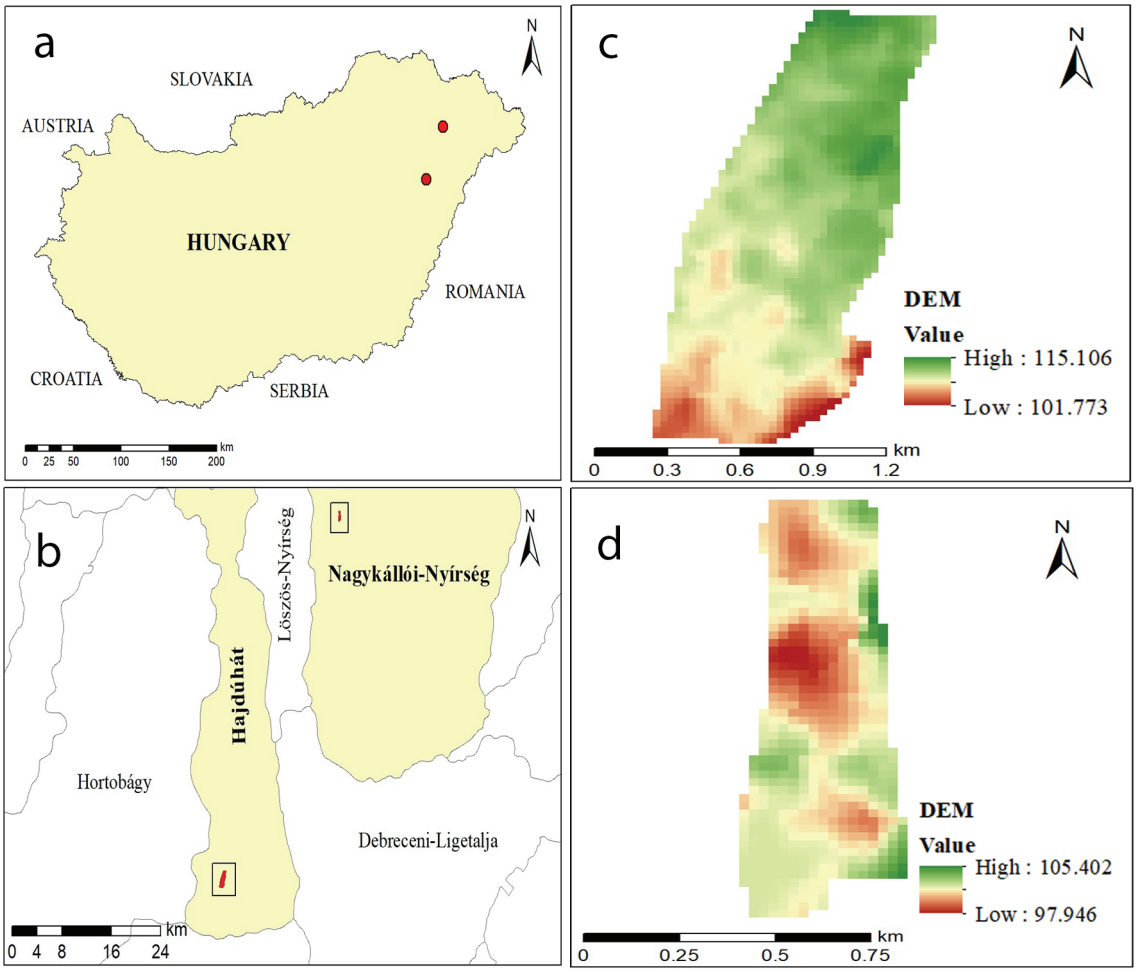

Fig. 1. Study area; a: locations of study area in Hungary, b: locations of study areas in related microregions, Látókép located in Hajdúhát and Westsik located inNagykállói-Nyírség, c: Látókép, d: Westsik.

to determine bulk density. Soil samples representing the uppermost 10 cm of the surface were collected for basic laboratory analysis.

The SOC content was determined by the wet oxidation method (Davies, 1974). Also, we calculated SOC stock obtained from SOC and bulk density based on the following equation:

$$SOC_{stock} = SOC \cdot BD \cdot D,$$

Where SOC_{stock} is the soil organic carbon stock [unit: g cm^{-2}], *SOC* is the soil organic carbon content [unit: percentage of soil dry mass], *BD* is the bulk density [unit: g cm^{-3}], and *D* stands for soil thickness [unit: cm] which was 10 cm in our case. Due to using a more convenient way of expressing SOC stock, we transformed the unit to [tons·ha^{-1}], as it is more common and easy to interpret.

2.3. Environmental covariates

Covariates representing the soil forming factors were employed to map the selected soil properties (Dai et al., 2014). As we described (Section 2.1), the climate, land cover, and parent material can be considered homogeneous over the study area. Therefore, a digital elevation model (Bashfield and Keim, 2011) with a spatial resolution of 26 m was used for characterizing topography and relief conditions. Various terrain attributes were derived including elevation, slope, aspect, plan and profile curvature, LS-factor, positive and negative openness, terrain ruggedness index, multiresolution index of valley bottom flatness, multiresolution index of ridge top flatness, topographic wetness index, valley depth and deviation from mean value (Tool Focal Statistics) using ArcGIS (http://www.esri.com) and SAGA GIS (Conrad et al., 2015). Moreover, additional information about the topsoil based on spectral data was obtained via Landsat 8 satellite imagery (https://earthexplorer.usgs.gov) which had been collected in February 2019 with a spatial resolution of 30 m. The NDVI (Band 5/4), Clay (hydroxyls) normalized ratio (Band 6/7), and Carbonate normalized ratio (Band 4/3) was calculated from Landsat 8 (John et al., 2020). Therefore, we assumed that this approach would provide certain soil features from remotely sensed information. All the auxiliary variables are projected in the same spatial dimension into a Universal Transverse Mercator (UTM) Zone 34N coordinate system. Then, resampling was applied to harmonize the resolution of 30 m for input layers to further process; accordingly, the final maps are also in the same resolution. This information was assigned corresponding to every geo-referenced soil sample.

2.4. Machine learning algorithms

In soil mapping, MLAs can significantly explain and model complex non-linear relationships between soil properties and the available environmental variables related to the soil-forming factors (Hengl et al., 2018). In this study, the following MLAs were applied: multiple linear regression (MLR), random forest (RF), artificial neural networks (ANN), and support vector machine (SVM). Note that only a brief introduction on the MLAs used in this study is provided as they are frequently used in DSM; for more details, see the cited papers and textbooks.

2.4.1. Multiple linear regression (MLR)

Multiple linear regression is a statistical model representing the relation between various independent variables and the response variable (Piekutowska et al., 2021). The premise of this algorithm is the existence of a linear relationship between independent and dependent variables that lead to applying a straight line between the points. Hence, this assumption could be a limitation since linearity rarely stands in soil science (Forkuor et al., 2017).

2.4.2. Random Forest (RF)

The Random Forest (RF) algorithm is based on the classification and regression trees that runs by building many decision trees at training time (Breiman, 2001; Hengl et al., 2018). In this model, the number of trees and covariables to split in each node, namely mtry, are the most powerful items which can be fine-tuned and generate more accurate results (Ghafouri Kesbi et al., 2016). In the end, the weighted average of individual trees will be the concluding predictions. RF is available to implement through several packages in R, and we applied the R package "*randomForest*" based on Breiman's (2001) original RF algorithm.

2.4.3. Artificial Neural Network (ANN)

The artificial neural network (ANN) is another machine learning algorithm that simulates biological neurons' structure and attempts to allow computers to learn similarly to humans' brains. Multi-layer perceptron is the most frequently used ANN type (Ghaderi et al., 2019), which follows the feed-forward model, meaning inputs are sent into the neuron, processed, and result in an output. Therefore, it consists of three layers, i.e., input layer, hidden layer, and linear output layer (Liu et al., 2020). The best results arise from trial and error in changing the number of neurons in the hidden layer during the training stage. In this study, we adopted neural networks backpropagation using the "*neuralnet*" package in R.

2.4.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) was developed in 1990 as a supervised learning model that classifies the data based on optimal separating hyperplane using kernel function for linear and non-linear patterns. The hyperplane is fixed by maximizing the margins of class boundaries (Were et al., 2015). In this model, certain parameters are effective in model performance, including cost (C) and gamma (γ) (Tang et al., 2020). Cost is the parameter for the soft margin, which controls the influence of each support vector, and the process involves trading error penalty for stability. Cost and width of margin have reverse relationships. Gamma is the free parameter of the Gaussian radial basis function which works with non-linear kernel functions. A small gamma means that the class of this support vector will have a significant influence; consequently, if the gamma is large, then the variance is slight (James et al., 2013). We used „*e1071*" package in R to utilize this model.

2.5. Validation

The datasets were randomly split into two sets. Two-thirds of the data went to the train set, which were used for training, whereas one-third of the data went to the test set, which was used for evaluating the models' performance. To compare and evaluate the performance of the four MLAs for each soil property in both study areas, we used the "*goof*" function of the R package "*ithir*", which compute a number of error measures. In this study, we evaluated the performance of each model based on their R^2 and root mean square error (RMSE), prioritizing models with higher R^2 values and smaller RMSE values. We summarized the methodology used in this study as a flowchart in Fig. 2.

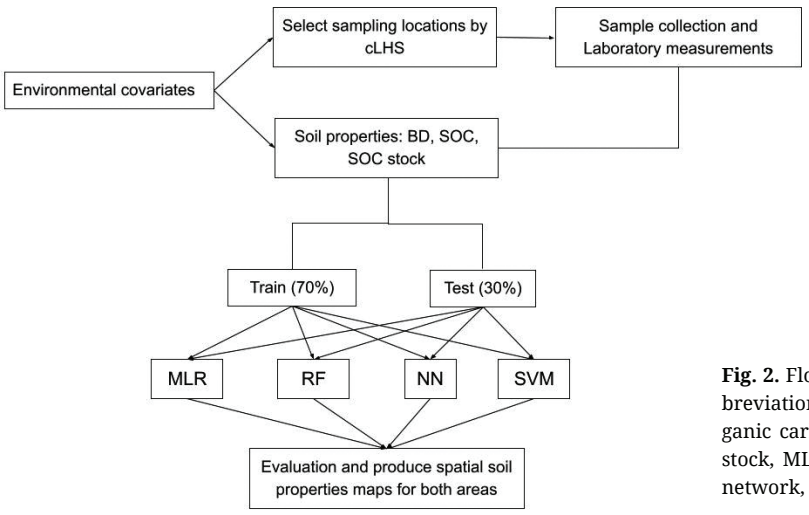

Fig. 2. Flowchart of the methods and approaches used in this study. Abbreviations: cLHS: conditioned Latin Hypercube Sampling, SOC: soil organic carbon content, BD: bulk density, SOC stock: soil organic carbon stock, MLR: multiple linear regression, RF: random forest, NN: neural network, and SVM: support vector machine.

3. Results

3.1. Soil properties

The Table 1 presents the statistical summary of selected soil properties in both study areas. In te rms of SOC content, Látókép has a significantly higher mean value of 1.76 g·kg^{-1} compared to Westsik's mean value of 0.40 g·kg^{-1}. The SOC content in Látókép ranged from 1.29 to 2.33 g·kg^{-1}, while the range in Westsik was wider, varying from 0.91 to 4.15 g·kg^{-1}. BD also shows some differences between the two areas, with Látókép having a slightly lower mean value of 1.28 g·cm^{-3} compared to Westsik's mean value of 1.38 g·cm^{-3}. Also, the minimum and maximum values of SOC stock in Látókép were also higher than those in Westsik.

Table 1
Summary statistics of soil properties in both study areas

Study area		Látókép				Westsik			
Property	Unit	Min	Max	Mean	SD	Min	Max	Mean	SD
SOC content	g·kg^{-1}	1.29	2.33	1.76	0.24	0.29	4.15	0.91	0.40
BD	g·cm^{-3}	1.12	1.46	1.28	0.10	1.15	1.56	1.38	0.09
SOC stock	t·ha–^{1}	16.71	31.72	22.60	3.11	4.56	22.9	11.07	5.32

3.2. Comparison of MLAs

As it was mentioned, we fine-tuned the hyperparameters of each MLAs to achieve higher performance. For this purpose, the number of trees selected in RF was 50, with *mtry* between 6 and 8. Changes in the number of hidden neurons can lead to tuning the ANN, for which we employed a combination of 3 and 5 layers. "*Tune*" function in the "*e1071*" package gives the possibility to specify the best cost and gamma combination for SVM in each dataset. Here, the cost = 10 and gamma = 2 achieved the best model performance. Afterward, we calculated the values of RMSE and R^2 to check model accuracy. Table 2–5 summarize the performance of the fine-tuned MLAs in Látókép and Westsik, respectively, which were compared to the results of validation carried out on the test dataset.

First of all, in the calibration dataset related to the Látókép (Table 2), MLR has shown the highest R^2, which was around 95% in these properties, following RF (R^2 = 80% ~ 90%), then ANN, and SVM, which illustrate nearly similar capability to predict. However, considering RMSE, ANN performs the worst values in SOC content and SOC stock while the MLR, RF, and SVM are quite the same, but the least RMSE related to MLR, which is 0.01 for bulk density, 0.05 SOC content, and 0.69 SOC stock. In the second place, assessing the results of the test dataset (Table 3), as evident from the tables, RF achieved the highest accuracy regarding R square and RMSE ($R^2 \approx 0.8$). At the same time, MLR and ANN outcomes have the most unfavorable results.

The outcome in Westsik dataset is the same as in the Látókép. MLR operated more effectively in calibration than other models (R^2 = 0.9), followed by RF, ANN, and SVM, that the R-squared values of the fitted model are around 0.8, 0.7, and 0.5, respectively (Table 4).

Table 2
Summary of the fine-tuned machine learning algorithms on the train dataset in Látókép.

Model	Bulk density		SOC content		SOC stock	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
MLR	0.990	0.010	0.951	0.053	0.957	0.691
RF	0.797	0.047	0.812	0.104	0.902	1.039
ANN	0.936	0.026	0.509	0.296	0.136	3.100
SVM	0.609	0.066	0.678	0.136	0.774	1.582

Abbreviations: MLR: multiple linear regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine, SOC: soil organic carbon, and RMSE: root means square error.

Table 3
The predictive performance of the fine-tuned machine learning algorithms on the test dataset in Látókép.

Model	Bulk density		SOC content		SOC stock	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
MLR	0.064	0.863	0.046	0.558	0.039	4.358
RF	0.885	0.035	0.821	0.104	0.882	0.739
ANN	–0.746	0.138	–0.456	0.299	–0.274	2.432
SVM	0.188	0.114	0.270	0.258	0.090	2.054

Abbreviations: MLR: multiple linear regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine, SOC: soil organic carbon, and RMSE: root means square error.

Table 4
Summary of the machine learning algorithms on the train dataset in Westsik.

Model	Bulk density		SOC		SOC stock	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
MLR	0.969	0.015	0.989	0.029	0.934	1.226
RF	0.791	0.039	0.878	0.292	0.853	1.830
ANN	0.761	0.042	0.989	0.041	0.453	5.775
SVM	0.590	0.055	0.516	0.584	0.739	2.445

Abbreviations: MLR: multiple linear regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine, SOC: soil organic carbon, and RMSE: root means square error.

Likewise, the RF technique performed quite favorably and achieved the best R^2 (0.8) and the least RMSE in the test dataset. Afterward, the other three models misbehaved in the test dataset, indicating raised RMSE and deficient R^2 (Table 5).

When it comes to the final maps produced by each model, it is also visible that the RF model succeeded in showing the most detailed information related to the distribution of each property across the study areas (Fig. 3 and 4). Of course, the other maps also contain worthy information; also, they were capable of explaining the whole pattern of the sites. However, we can observe that the prediction range is unrealistic for some properties.

Table 5
The predictive performance of the fine-tuned machine learning algorithms on the test dataset in Westsik.

Model	Bulk density		Organic Carbon		Soil carbon stock	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
MLR	0.240	0.524	0.023	1.857	0.137	20.956
RF	0.823	0.047	0.80	0.165	0.856	2.117
ANN	–0.169	0.123	0.206	0.330	–0.737	7.363
SVM	0.010	0.114	0.295	0.311	0.018	5.534

Abbreviations: MLR: multiple linear regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine, SOC: soil organic carbon, and RMSE: root means square error.

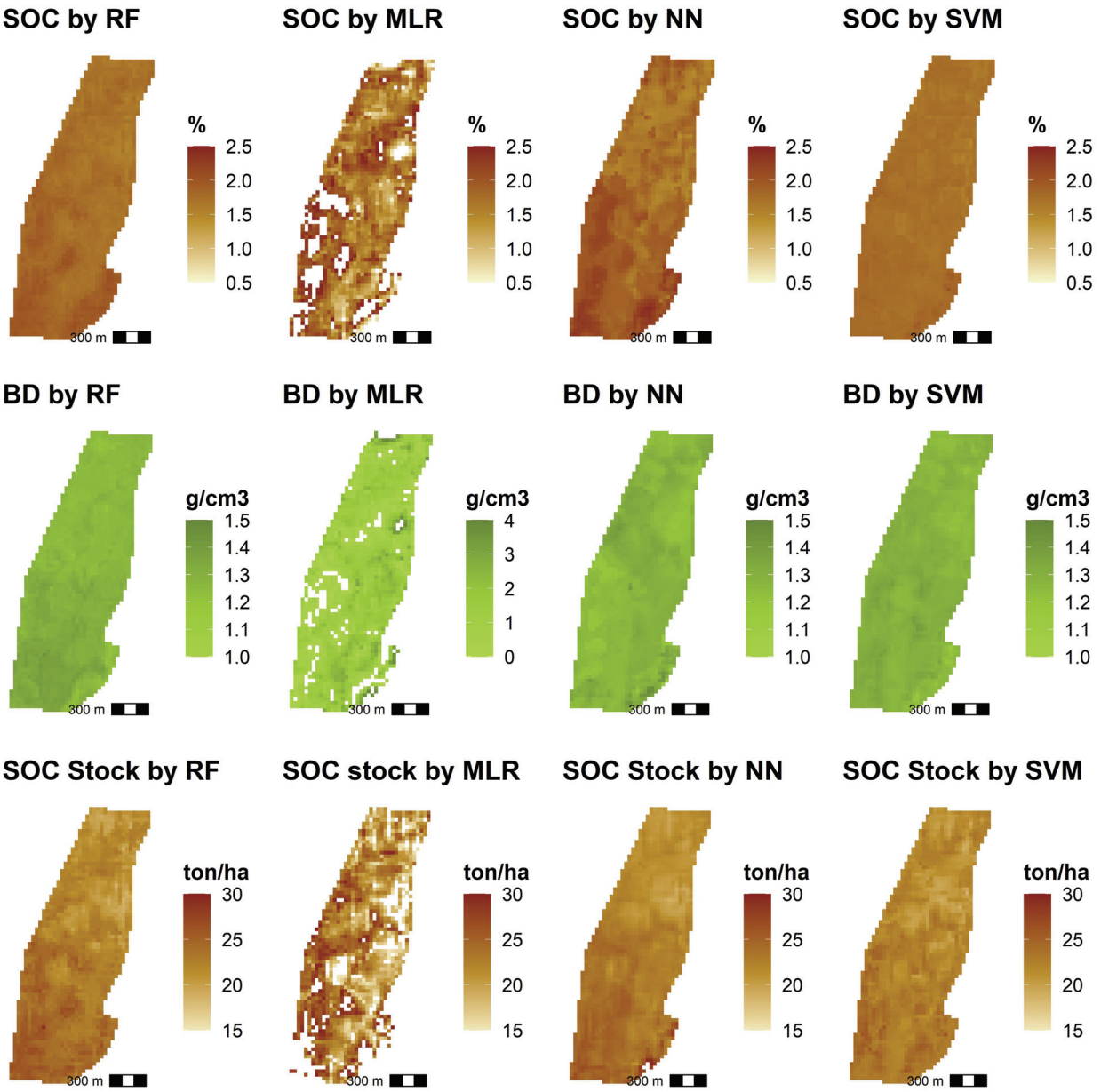

Fig. 3. Final maps predicted by each model in Látókép. Abbreviation: SOC: soil organic carbon content, BD: bulk density, SOC stock: soil organic carbon stock, RF: random forest, MLR: multiple linear regression, NN: neural network, and SVM: support vector machine.

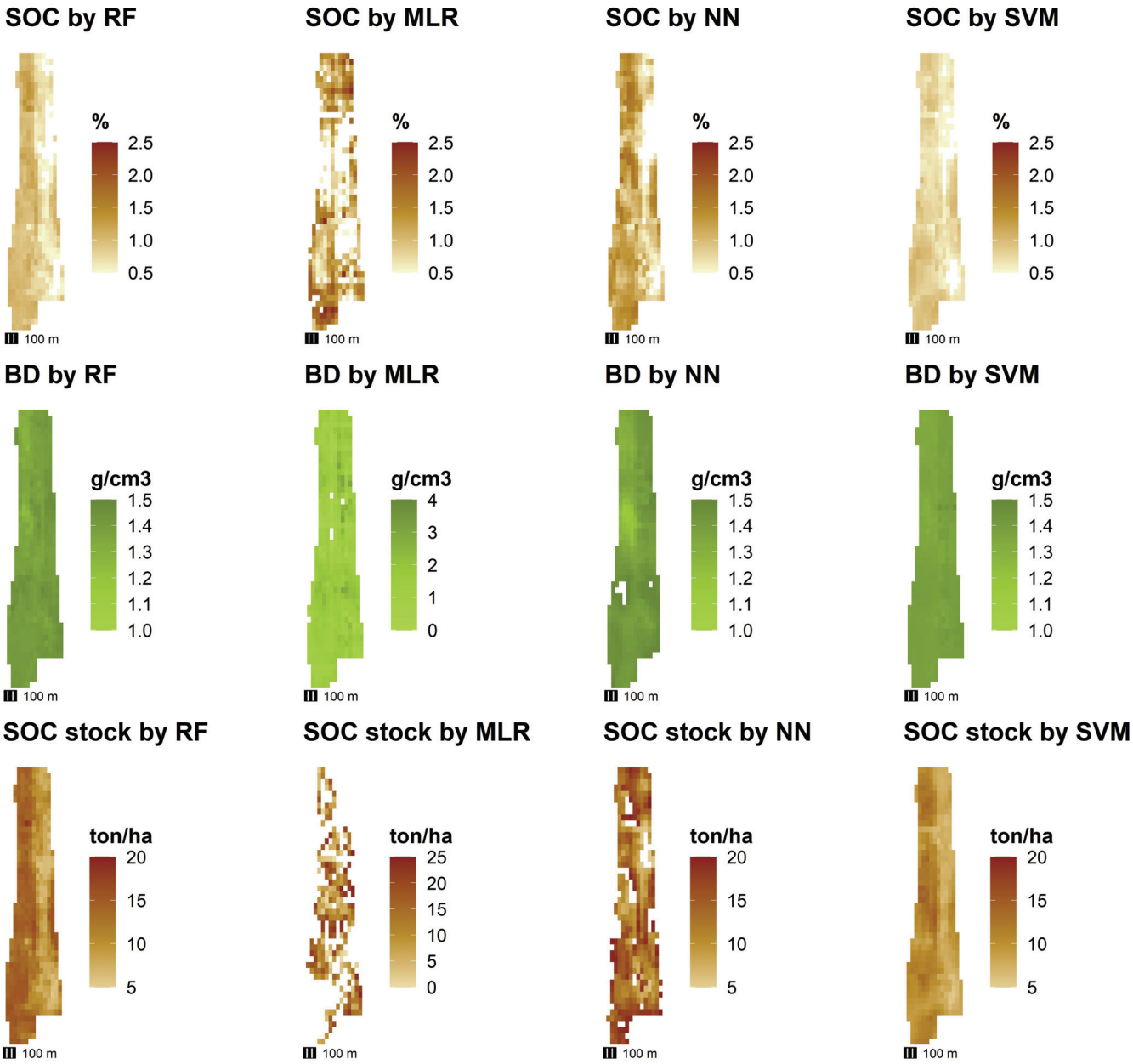

Fig. 4. Final maps predicted by each model in Westsik. Abbreviation: SOC: soil organic carbon content, BD: bulk density, SOC stock: soil organic carbon stock, RF: random forest, MLR: multiple linear regression, NN: neural network, and SVM: support vector machine.

3.3. Relationship and importance of environmental covariates in mapping soil properties

The relationship between SOC content and predictor variables in Látókép is presented in Fig. 5. SOC has a relatively strong correlation with elevation (–0.541), plan curvature (–0.586), multiresolution index of the ridge top flatness (–0.529), valley depth (0.538) which were extracted from DEM and carbonate index (0.476) obtained by Landsat images. On the other hand, a poor but meaningful relationship has been seen between SOC content and slope (0.231), NDVI (–0.247), and terrain ruggedness index (0.245). Similarly, valley depth (0.414), elevation (–0.437), clay index (–0.582), and NDVI (–0.592) are the predictors which explain the potent relationship with SOC content in the Westsik area. In contrast, the relationship between SOC content and slope, terrain roughness index, and openness is relatively weak in Westsik (Fig. 6).

Variable importance measures, one of the byproducts of the RF model, confirm which predictors are the most influential in the modeling and predicting process. Here, the results of most significant variables for SOC content, bulk density, and SOC stock based on the RF model are presented in Fig. 7 and 8 for Látókép and Westsik, respectively. The study in Látókép indicated that the plan curvature and valley depth were the dominant factors in explain-

ing the spatial variability of SOC content and SOC stock. Additionally, the multiresolution index of ridge top flatness, elevation, and NDVI were found to play a significant role in determining the BD in Látókép (Fig. 7). On the other hand, in Westsik, the most crucial variable in explaining the spatial variation of SOC content and SOC stock was NDVI. Furthermore, the variables that contributed significantly to predicting BD were the multiresolution index of valley bottom flatness, elevation, and carbonate index (Fig. 8).

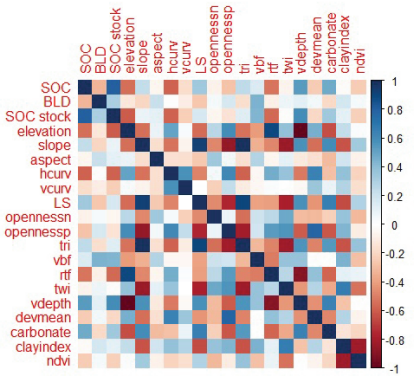

Fig. 5. Correlation between predictors and soil organic carbon content (SOC) in Látókép. Abbreviations: hcurv and vcurv: plan and profile curvature, ls: ls-factor, opennessp and opennessn: positive and negative openness, tri: terrain ruggedness index, vbf: multiresolution index of valley bottom flatness, 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, SOC stock: soil organic carbon stock, and BD: bulk density.

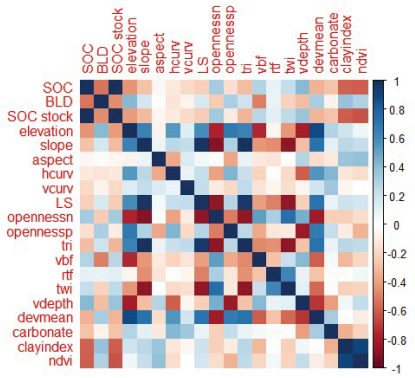

Fig. 6. Correlation between predictors and soil organic carbon content (SOC) in Westsik. Abbreviations: hcurv and vcurve: plan and profile curvature, ls: ls-factor, opennessp and opennessn: positive and negative openness, tri: terrain ruggedness index, vbf: multiresolution index of valley bottom flatness, 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, SOC stock: soil organic carbon stock, and BD: bulk density.

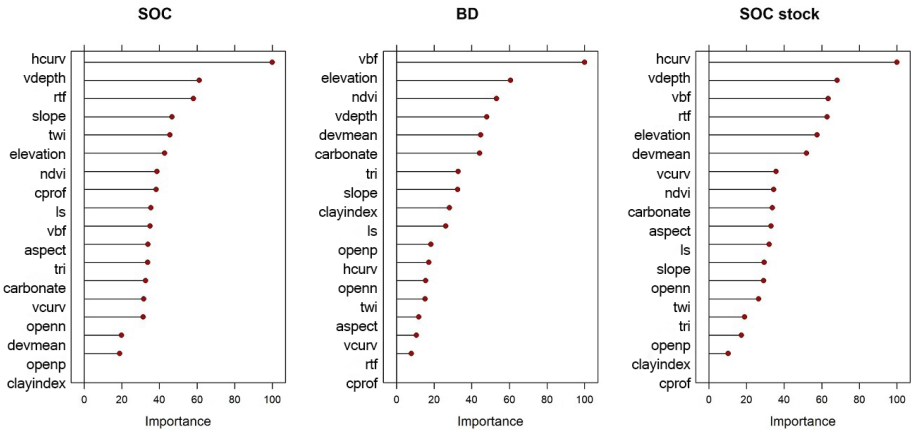

Fig. 7. Variable importance based on RF for each properties in Látókép. Abbreviations: hcurv and vcurve: plan and profile curvature, ls: ls-factor, opennessp and opennessn: positive and negative openness, tri: terrain ruggedness index, vbf: multiresolution index of valley bottom flatness, rtf: multiresolution 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, SOC stock: soil organic carbon stock, and BD: bulk density.

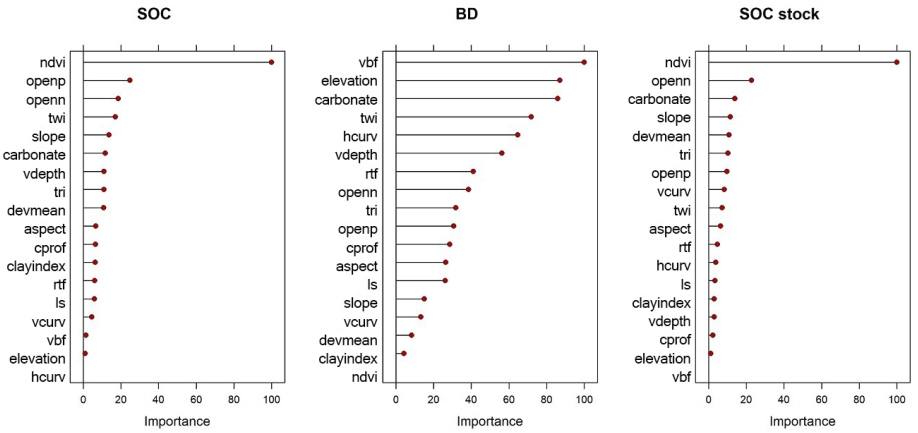

Fig. 8. Variable importance based on RF for each properties in Westsik. Abbreviations: hcurv and vcurve: plan and profile curvature, ls: ls-factor, opennessp and opennessn: positive and negative openness, tri: terrain ruggedness index, vbf: multiresolution index of valley bottom flatness, rtf: multiresolution 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, SOC stock: soil organic carbon stock, and BD: bulk density.

4. Discussion

4.1. Interpretation of the relationship between soil properties and environmental covariates

The Fig. 7 revealed that plan curvature, valley depth, multiresolution index of valley bottom flatness, and elevation were the four most relevant predictors for mapping SOC content, bulk density, and SOC stock in the Látókép. While in Westsik, NDVI and carbonate index, multiresolution index of valley bottom flatness, elevation, and openness were the strongest ones (Fig. 8). This result proves the importance of topographic indices in representing landscape morphometry, which (Heung et al., 2016) support. Similarly, Forkour et al. (2017) recorded elevation as the most significant covariate for SOC and nitrogen spatial distribution. (Piccini et al., 2014) found that parameters driven by DEM, including topographic wetness index, plan and profile curvature, and soil type, were among the most significant factors in estimating soil organic matter.

In Látókép, Landsat indices have less power to explain the variability of soil properties, as the markable influence of NDVI was observed only in bulk density. Similar conclusions were reported by Piccini et al., 2014). On the other hand, the most influential predictors in Westsik were the NDVI and clay index extracted from remote sensing data, which was marked by (Asgari et al., 2020). In addition, Bhunia et al. (2019) estimated spatial SOC stock with satellite data-derived indices and a multivariate regression model, which approved a significant relation between NDVI and SOC stock. Contrary to the report by Dai et al.(2014), the correlation between environmental variables and organic matter content should not be necessarily high. These relations were shallow in their case study, especially for elevation, precipitation, and NDVI. Also, in the other studies of (Mosleh et al., 2016; Song et al., 2017; John et al., 2020), the inability of environmental variables to predict soil properties, especially in low relief conditions, were reported.

4.2. Advances and limits of the applied MLAs

Each model has its pros and cons in predictability, which depends on various soil forming factors and local/regional conditions. But our research clearly shows that RF has an advantage over MLR, SVM, and ANN in both areas, though with limited number of observations. The benefits of RF are doing great with handling outliers, unbalanced data, non-linear data, and complex relationships (Ao et al., 2019). Furthermore, Hengl et al. (2018) presented that the RF model is an attractive and popular model for spatial predictions. Besides, they described the most important characteristics of this model, which include flexibility in combining, incorporating, and extending covariates of different types and the ability to make more informative and detailed maps. John et al. (2020) utilized different machine learning algorithms to predict SOC stock, including ANN, SVM, cubist, RF, and MLR. They choose RF as the best-performing model and MLR as the poorest one. Despite this, Were et al. (2015) reported better performability of SVM rather than RF and ANN in predicting SOC stocks. Moreover, Tang et al. (2020) used MLR and SVM to predict the biodegradation rate of organic chemicals. They claimed that SVM models have satisfactory goodness-of-fit, robustness, and external predictive abilities.

The failure of MLR in this research could be the non-linearity relationship between soil properties and environmental variables, but here the relationships between variables are complex. Forkuor et al. (2017) stated the inability of MLR to manage non-linear relationships between variables and better performance of other machine learning algorithms compared to MLR. In the case of ANN, noticing negative R^2 is not a mathematical or computational issue. Still, such an outcome means the model behaved even worse than if we only used the spatial average of the data for prediction, and the observed mean is a better predictor than the model by itself. One of the main reasons for ANN's inability to predict can be the limited observation (Moody, 1994). However, the high predictability of ANN was confirmed by many studies. For example Zhao et al. (2008), in their research on soil texture prediction with an artificial neural network, achieved a relative overall accuracy of 80% for clay and sand content which is an acceptable result. Furthermore, Li et al. (2013) reported that the radial basis function neural network model performed much better than multiple linear regression and regression kriging in predicting SOC, moreover showed a more realistic spatial pattern.

4.3. Spatial distribution of selected soil properties

In these two study areas, a clear spatial distribution pattern of soil properties is evident in RF maps (Fig. 3 and 4). Therefore, we will describe the broad overview of each property mainly based on RF predictions. In Látókép, the southeast and southwest have higher SOC content; if we move along the western border, some narrow parts still have high SOC content until we reach northern parts with brighter spots, which means lower SOC content. In the central part of the area, we can observe different values of SOC content as small patches. In MLR maps, there are many white patches that show out-of-range values. While in the Westsik area, the western part has higher values in SOC content, BD, and SOC stock, which NDVI maps and elevation differences might influence. Lower values of SOC content can be found on the eastern side of the area.

Soil surface characteristics and soil profile horizonation show remarkable differences between the two study sites. The soil texture in Látókép is silt loam, and the soil texture in Westsik has wider variability from loamy sand to sandy loam at the surface. Despite textural differences between the two areas, the bulk density of surface soil samples are slightly higher in the Westsik due to the sandy texture. Still, standard deviations are pretty similar in both sites, thanks to the similar techniques applied in plowing.

The average SOC content proved to be almost twice higher in the Látókép, with a remarkably smaller standard deviation, than in the Westsik (Table 1). Generally, the spatial variability of selected properties in the Westsik is more expansive, while in the Látókép, it appears more homogenous than the other. These differences can be reasonable, considering the elevation and related attributes heterogeneity which also can be recognized in

the importance plot of the covariates (Fig. 7 and 8). These factors, like the influence of elevation, affect many soil processes; for example, the local depression causes the deposition of SOC while higher positions are exposed to erosion, precipitation, and transportation.

5. Conclusions

This study aimed to investigate the applicability of different MLAs in predicting SOC content, bulk density, and SOC stock in two geographical conditions to display the spatial variability of these properties and the dependency of model building. We examined and compared four statistical models: RF, MLR, ANN, and SVM. Our work has led us to conclude that the RF technique achieved more reliable results than the other models, especially with limited number of observations. In contrast, SVM, MLR, and ANN did not deliver satisfactory results for both provided accuracy and spatial pattern. In addition, we indicated the importance of relief characteristics and spectral information, including elevation, plan and profile curvature, valley depth, and NDVI, in predicting soil properties. Our results highlighted the importance of differences in topography compared to each other, leading to observing different spatial variations of selected soil properties across the areas. Overall, these prediction accuracies and the provided pattern in the final maps by RF seem trustworthy, considering the actual situation in both field and expert knowledge. These two study areas, Látókép and Westsik, are both research experiments fields; therefore, detailed information about the spatial distribution of soil properties would support sustainable management practices and precision agriculture.

References

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 modelling. Journal of Petroleum Science and Engineering 174, 776–789. https://doi.org/10.1016/j.petrol.2018.11.067

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. https://doi.org/10.1016/j.geodrs.2020.e00255

Asgari, N., Ayoubi, S., Jafari, A., Demattę, 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. https://doi.org/10.1080/01431161.2020.1763506

Balducci, F., Impedovo, D., Pirlo, G., 2018. Machine learning applications on agricultural datasets for smart farm enhancement. Machines 6(3), 38. https://doi.org/10.3390/machines6030038

Bashfield, A., Keim, A., 2011. Continent-wide DEM creation for the European Union. In 34th International Symposium on Remote Sensing of Environment. The GEOSS Era: Towards Operational Environmental Monitoring. Sydney, Australia (pp. 10–15).

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. https://doi.org/10.1002/jpln.200421414

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. https://doi.org/10.1080/10106049.2017.1381179

Breiman, L., 2001. Random forests. Machine learning 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brungard, C.W., Boettinger, J.L., 2010. Conditioned latin hypercube sampling: Optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In Digital soil mapping 67–75. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-8863-5_6

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. https://doi.org/10.3390/ijgi8040174

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1. 4. Geoscientific Model Development 8(7), 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

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. https://doi.org/10.1016/j.ecolind.2014.04.003

Forkuor, G., Hounkpatin, O.K., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in southwestern Burkina Faso: a comparison of machine learning and multiple linear regression models. PloS one 12(1), e0170478. https://doi.org/10.1371/journal.pone.0170478

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. https://doi.org/10.1007/s10064-018-1400-9

Ghafouri Kesbi, F., Rahimi Mianji, G., Honarvar, M., Nejati Javaremi, A., 2016. Tuning and application of random forest algorithm in genomic evaluation. Research On Animal Production (Scientific and Research) 7(13), 185–178. https://doi.org/10.18869/acadpub.rap.7.13.185

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. https://doi.org/10.1016/j.geoderma.2019.01.007

Hateffard, F., Balog, K., Tóth, T., Mészáros, J., Árvai, M., Kovács, Z.A., Szatmári, G., 2022. High-Resolution Mapping and Assessment of Salt-Affectedness on Arable Lands by the Combination of Ensemble Learning and Multivariate Geostatistics. Agronomy 12(8), 1858. https://doi.org/10.3390/agronomy12081858

Hateffard, F., Dolati, P., Heidari, A., Zolfaghari, A.A., 2019. Assessing the performance of decisiontree and neural network models in mapping soil properties. Journal of Mountain Science 16(8). https://doi.org/10.1007/s11629-019-5409-8

Hateffard, F., Novák, T.J., 2021. Soil sampling design optimization by using conditioned Latin Hypercube sampling (No. ISMC2021-35). Copernicus Meetings. https://doi.org/10.5194/ismc2021-35

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. https://doi.org/10.1371/journal.pone.0125814

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. https://doi.org/10.7717/peerj.5518

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques

for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

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). https://doi.org/10.1038/s41598-018-28244-w

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer. https://doi.org/10.1007/978-1-0716-1418-1

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. https://doi.org/10.3390/land9120487

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. https://doi.org/10.1127/0941-2948/2006/0130

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. https://doi.org/10.1016/j.geoderma.2009.11.005

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. https://doi.org/10.1016/j.catena.2012.11.012

Liu, E., Liu, J., Yu, K., Wang, Y., He, P., 2020. A hybrid model for predicting spatial distribution of soil organic matter in a bamboo forest based on general regression neural network and interative algorithm. Journal of Forestry Research 31(5), 1673–1680. https://doi.org/10.1007/s11676-019-00980-3

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). European Journal of Soil Science 70(2), 216–235. https://doi.org/10.1111/ejss.12790

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Moody, J., 1994. Prediction risk and architecture selection for neural networks. In From statistics to neural networks, 147–165. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-79119-2_7

Mora-Vallejo, A., Claessens, L., Stoorvogel, J., Heuvelink, G.B., 2008. Small scale digital soil mapping in Southeastern Kenya. Catena 76(1), 44–53. https://doi.org/10.1016/j.catena.2008.09.008

Mosleh, Z., Salehi, M.H., Jafari, A., Borujeni, I.E., Mehnatkesh, A., 2016. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental monitoring and assessment 188(3), 1–13. https://doi.org/10.1007/s10661-016-5204-8

Owens, P.R., Dorantes, M.J., Fuentes, B.A., Libohova, Z., Schmidt, A., 2020. Taking digital soil mapping to the field: Lessons learned from the Water Smart Agriculture soil mapping project in Central America. Geoderma Regional 22, e00285. https://doi.org/10.1016/j.geodrs.2020.e00285

Piccini, C., Marchetti, A., Francaviglia, R., 2014. Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. Ecological Indicators 36, 301–314. https://doi.org/10.1016/j.ecolind.2013.08.009

Piekutowska, M., Niedbała, G., Piskier, T., Lenartowicz, T., Pilarski, K., Wojciechowski, T., Czechowska-Kosacka, A., 2021. The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest. Agronomy 11(5), 885. https://doi.org/10.3390/agronomy11050885

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., and Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL 7, 217–240. https://doi.org/10.5194/soil-7-217-2021, 2021

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. https://doi.org/10.3390/su9050754

Stoorvogel, J.J., Kooistra, L., Bouma, J., 2015. Managing Soil Variability. Soil-specific farming: precision agriculture, 22, 37.

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. https://doi.org/10.1016/j.chemosphere.2020.126666

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. https://doi.org/10.1016/j.earscirev.2020.103359

Were, K., Bui, D.T., Dick, Ř.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. https://doi.org/10.1016/j.ecolind.2014.12.028

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006

Zhao, Y.C., Shi, X.Z., 2010. Spatial prediction and uncertainty assessment of soil organic carbon in Hebei Province, China. Digital soil mapping: bridging research, environmental application, and operation 227–239. https://doi.org/10.1007/978-90-481-8863-5_19

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. https://doi.org/10.1016/j.compag.2008.07.008