Short thesis for the degree of Doctor of Philosophy (PhD)

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

By: Fatemeh Hateffard

Supervisors:

Novák Tibor József (Ph.D. Associate Professor)

Szatmári Gábor (Ph.D. Senior Research Fellow)

University of Debrecen

Doctoral School of Earth Sciences

Debrecen, 2023

1. Introduction

Digital soil mapping (DSM) is important in sustainable land management, land use planning, and precision agriculture. The process of DSM involves providing soil observations and environmental covariates, using statistical and mathematical techniques to capture their relationship, and producing spatial and/or temporal predictions. The main framework of DSM is based on the concept of fundamental soil development theory, and there has been an expansion in the development of DSM techniques at different scales. Machine learning (ML) algorithms and geostatistics are common ways of predicting soil properties and delivering spatial information on soils. The accuracy of predictive models is crucial for informing policy-making decisions, but selecting an appropriate approach and identifying the best model can be challenging. Different ML techniques, such as decision trees, random forest, artificial neural networks, and support vector machines, can be used for regression or classification purposes. To conduct DSM, soil samples are used to train the model, while environmental covariates serve as predictors to calibrate the model. These covariates are essential in explaining the soil-forming factors and other physical and chemical processes that contribute to the spatial variation of soil properties. DSM has some knowledge gaps and issues, such as data availability and quality, multivariate mapping, and uncertainty analysis. Researchers are working on innovative methods to address these gaps and improve accuracy. The

dissertation focuses on extrapolation and joint spatial modeling issues of soil properties.

Firstly, soil mapping is limited by the lack of soil observations in many areas. Spatial extrapolation, which transfers a model to a new geographic location from a donor area, can be used to predict soil properties in areas without observations (recipient), but it requires the similarity of soil-forming factors between the two areas. Secondly, soil data is complex and ever-changing, making modeling difficult. Multivariate geostatistics is a widely used approach that considers the joint spatial variability of variables and explicitly takes into account spatial interdependence. Combining geostatistics with ML algorithms can improve soil property predictions and modeling of uncertainty.

2. Aim and objectives of the study

The objectives of my doctoral research are:

1. Predict and map the spatial distribution of soil properties in two small-scale areas which are different from physiographic conditions.
2. Evaluate the potential and efficiency of different techniques in spatial predictions of soil properties. Also, select the best model with the highest accuracy and least error to extrapolate over the larger areas.
3. Assess the possibility of extrapolation by Area of Applicability (AOA) method and validate the results by samples taken from large areas.

4. Estimate the similarity between two areas by different methods and if there is an agreement between these methods.
5. Explore the possibility of predicting over an unknown area by available dataset.
6. Predicting and mapping SAS indicators by applying ensemble machine learning.
7. Jointly modeling the prediction results with multivariate geostatistical techniques.

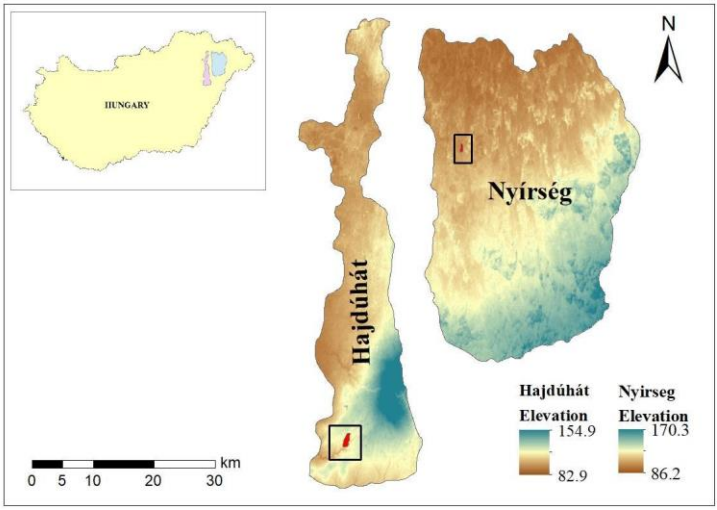
In my research, I conducted three case studies to achieve my objectives. The first case study involved comparing various ML models for predicting soil properties in small-scale areas, with extrapolation techniques applied to larger areas (Objectives 1-3). The second case study focused on determining the potential for extrapolation between areas based on similarity of soil-forming factors (Objectives 4 and 5). The third case study aimed to identify high salinity locations in arable land using joint spatial modeling of salt-affected soils (SAS) indicators (Objectives 6 and 7).

3. Materials and Methods

In case study one, the sampling design was implemented using the conditioned Latin Hypercube sampling (cLHS) method to select thirty surface samples from each of the four study areas in Hungary; Látókép, Westsik, Hajdúhát, and Nyírség (Figure 1). Environmental covariates such as Landsat images and DEM derivatives were used as inputs for cLHS to cover the spatial

variation of soil properties efficiently. Bulk density, soil organic carbon (SOC), pH, electrical conductivity (EC), and carbonate content were measured in the laboratory.

For Látókép and Westsik, three models were trained using Random Forest (RF), Artificial neural network (ANN), and Support Vector Machine (SVM), and were compared with Multiple linear regression (MLR), as a benchmark technique to compare with other algorithms. The best model was selected and fine-tuned. The trained model for Látókép was then applied to extrapolate over Hajdúhát, and the trained model for Westsik was applied to extrapolate over Nyírség. The Area of Applicability (AOA) was also applied over these areas. To validate the predictions for Hajdúhát and Nyírség, samples were taken from these areas and compared with the model predictions.



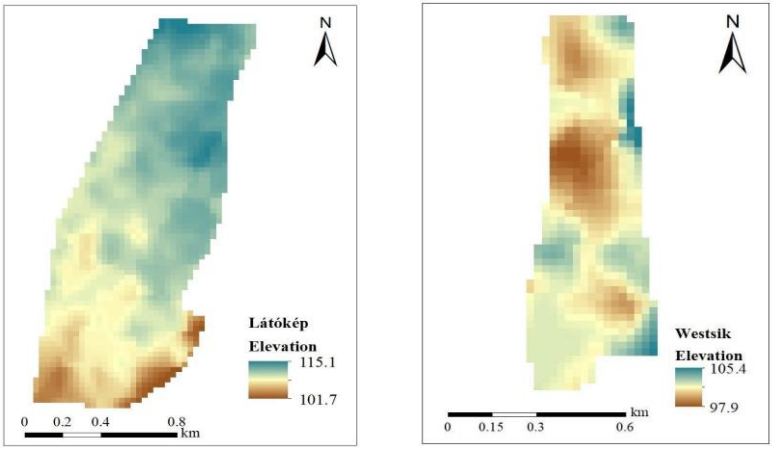


Figure 1. Study areas for case study one. The first figure is the location of microregions in Hungary which are shown on the Digital Elevation Model (DEM). Squares are the location of Látókép inside Hajdúhát microregion, and Westsik inside Nyírség microregion. The second figure is Látókép and Westsik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

In case study two, the data for four countries of Africa including Ethiopia, Kenya, Burkina Faso, and Nigeria (Figure 2) were extracted from the ISRIC Africa Soil Profiles (AfSP) database which is publicly available. The soil properties of interest for this case study were: soil organic carbon, clay content and pH. First, I 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, I had 12 scenarios. I identified the similarities between countries by using four

different methods including similarity in soil types, homosoil approach, dissimilarity index by AOA and quantile regression forest (QRF) prediction interval width. I validated the prediction points by observation in each country.
Homosoil is a method to extrapolate soil information from similar areas where soil data is scarce. The method uses Gower's similarity index to compare environmental covariates between two areas, with higher values indicating greater similarity. The index is calculated in three hierarchical steps, selecting areas with similar climate conditions, lithological classes, and topography.
AOA is a methodology that quantifies the differences in environmental covariates between donor and recipient areas and determines the extendable areas where a trained model can be employed based on the dissimilarity index. Its application helps to determine the area for which the model can be expected to make predictions with an error comparable to the model performance. QRF can calculate all quantiles of the prediction distribution, allowing for the quantification of prediction uncertainty at all prediction locations. Prediction interval width can be calculated from the difference between lower and upper quantiles of estimations. The width of prediction intervals is expected to be wider in areas with higher uncertainty due to extrapolation.

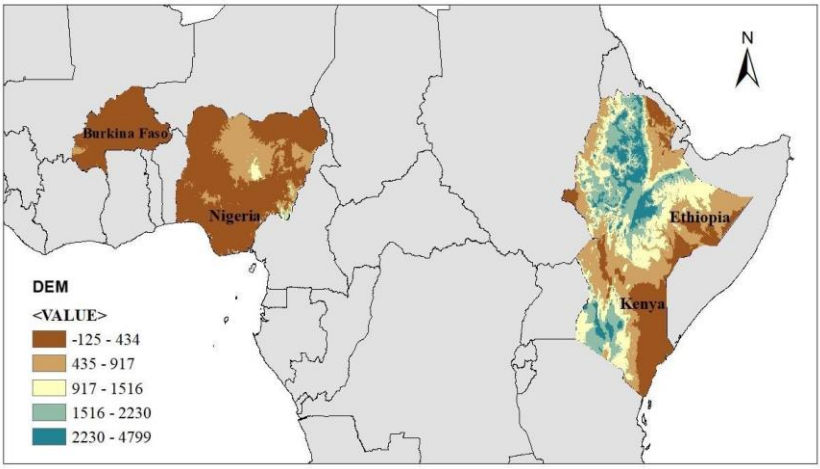


Figure 2. Study areas for case study two. Ethiopia, Kenya, Nigeria and Burkina Faso shown on DEM.

In case study three: This study collected 85 soil samples from a regular grid area near Dunavecse, Hungary (Figure 3), using soil tubes to a depth of 1m, but only analyzed the topsoil up to 30cm. Samples were taken to the laboratory to measure SAS indicators, including pH, EC, and SAR, and were carried out by colleagues of the Institute for Soil Sciences. An ensemble modeling approach was used with five individual models including RF, SVM, ANN, Extreme Gradient Boosting (XGBoost) and Generalized Linear Models with Lasso or Elastic Net Regularization (GLM). I 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.The accuracy of spatial predictions and estimation of uncertainties were evaluated using 10-fold cross-validation.

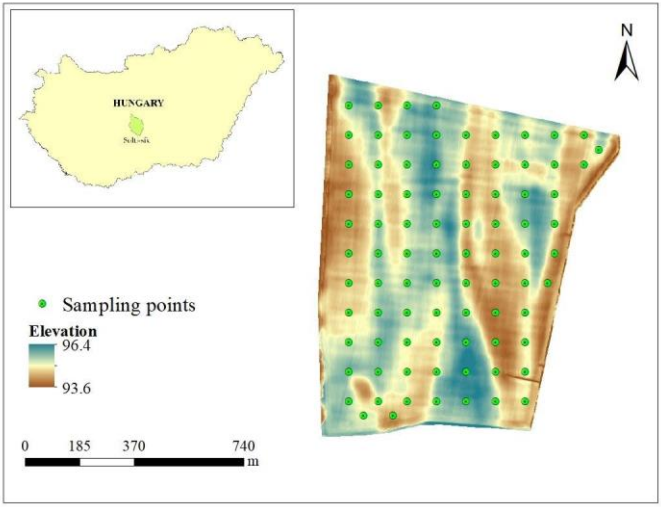


Figure 3. Study area for case study three. A plot near Dunavecse shown on Digital Elevation Model (DEM) and sampling points.

4. New Scientific Results (Theses)

1. The spatial distribution of soil properties in each area can be influenced by various environmental factors, which are related to the relationship between soil properties and environmental covariates in that specific area. In both Látókép and Westsik, geology and climate indices are constant. Therefore, topographic indices were found to be the major factor affecting the spatial distribution of soil properties

in Látókép, while NDVI was the most influential explanatory variable in Westsik, indicating that vegetation cover is the primary factor affecting soil spatial variation in the latter.

2. Each model has strengths and weaknesses in predicting soil properties, depending on soil-forming factors and local/regional conditions. My research found that despite limited observations, RF outperformed MLR, SVM, and NN in both study areas, with an R^2 coefficient of 80%. RF handles outliers, unbalanced data, and complex relationships well and is flexible in incorporating various covariates. SVM delivered acceptable results in predicting SOC and soil pH in Látókép and was more successful in the Westsik area, explaining 30-40% of spatial variation. ANN performed worse than using the spatial average of data, probably due to limited observations in this study. Also, MLR failed due to the complicated interrelationships between variables.
3. I calculated the AOA of the best predictive model, which is RF, for Látókép in order to determine the possible areas to extrapolate in Hajdúhát, and for Westsik in order to determine the possible areas to extrapolate in Nyírség, where the models could learn about relationships in Látókép and Westsik. The dissimilarity index calculated by AOA values for each soil property showed different ranges due to various relationships between the soil property of interest and predictors. Since the selection of important covariates used to train the model differed, the results of the AOA calculation also differed. The masked extrapolation maps of SOC stock and EC in Hajdúhát, as well as the BD map in Nyírség, revealed significant areas

that were outside the AOA, indicating that distinct soil-forming factors are crucial in explaining how these soil properties vary spatially between the two regions. In summary, we found that predictions inside AOA have fewer errors and the values are closer to the error measurements of predictive models, while predictions outside should be considered invalid due to larger dissimilarity. New locations with different environmental properties may lead to inaccurate predictions.

4. The degree of similarity in soil-forming factors between donor and recipient areas is important for extrapolating soil information, with more success in countries with heterogeneous conditions. Countries in the same region, such as Ethiopia with Kenya and Nigeria with Burkina Faso, have more similarities in terms of soil types and the homosoil approach. There was a correlation between spatial dissimilarity and uncertainty in predictions, with areas with significant environmental differences having higher uncertainty. Geographical proximity is also important for transferring the trained models to recipient areas, as indicated by similarities found between neighboring countries. Burkina Faso showed lower values compared to Ethiopia and Kenya in all measures of extrapolation, possibly due to significant environmental differences and geographic distance.
5. Results showed poor performance when extrapolating to recipient countries, highlighting the risks of extrapolation due to complex soil-landscape interactions and difficulty in matching soil-forming factors. The poor performance of

spatial extrapolation may also be attributed to the selected model, RF, which may perform poorly in extrapolation when there are large areas without observations and new predictors have different characteristics from what the model has learned. In general, all these four measures of extrapolation (homosoil, soil type similarity, dissimilarity index by AOA, and QRF prediction interval width) can be useful to give us information beforehand how well the extrapolation might work.

6. SuperLearner outperformed individual learners in predicting pH and SAR, while RF was the best-performing individual model among the five learners. This confirms the effectiveness of ensemble modeling in reducing noise and variance in predictions and preventing overfitting. Interestingly, RF outperformed SuperLearner in predicting EC. This could be due to factors such as limited covariates, artifacts in covariates, or insufficient sampling points, which can affect the relationship between EC and covariates. Soil EC is sensitive to modeling and mapping due to its rapid changes over time and space.
7. Joint modeling of the spatial distribution of SAS indicators using multivariate geostatistics and ML methods was found to be crucial for accurately predicting their spatial distribution since it showed clearly that SAS indicators are spatially interdependent along the study area. The study demonstrated that spatial prediction uncertainty of SAS indicators is in line with spatial cross-correlation between the indicators, providing advantages for soil quality management and

precision agriculture. This approach has potential for future soil mapping and management efforts.

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

DEBRECENI EGYETEM
Egyetemi és Nemzeti Könyvtár