Levantine Journal of Applied Sciences, Volume 1 (2021) 30-37.

Levantine Journal of Applied Sciences

journal homepage: https://www.ljoas.com

Rapid Spatial Estimation of Soil pH using Machine Learning under Limited Covariate Conditions

Hikmet Günal[1], Miraç Kılıç[2*], Mesut Altındal[3], Recep Gündoğan[1]

[1] Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Harran University, Şanlıurfa, Turkey.
[2] Kahta Vocational School, Adıyaman University, Kahta, Adıyaman, Turkey
[3] Fruit Research Institute Directorate, General Directorate of Agricultural Research and Policies of Turkey, Eğirdir, Isparta, Turkey

(Orcid: 0000-0002-4648-2645, 0000-0001-8026-5540, 0000-0002-0332-6677, 0000-0001-8877-1130)

Keywords

MLP-ANN
SVR
Geostatistic
Ordinary Kriging
Soil pH
Machine learning
Spatial analysis

ABSTRACT

Conventional soil mapping approaches require to spend long time in the field and laboratory, and are most of the time expensive; therefore, soil scientists continue to study producing reliable digital soil maps in a short time and at a less cost. The main aim of this study was to map the spatial distribution of soil pH at a field scale with fine resolution, and to assess the ability of two commonly used machine learning approaches to estimate soil pH at a scale. The machine learning models applied in this study were Multi-Layer Perception Artificial Neural Network (MLP-ANN) and Support Vector Regression (SVR). The study area covers an approximately 107.1 ha land, and is located in the orchards of fruit research station in Egirdir, Turkey. One hundred and three surface soil samples (0-30 cm) were collected from the corners of 50 x 50 m grid in the study area. The pH value ranged between 7.52 and 8.33 with a mean value of 7.95. The number of hidden node in the MLP-ANN architecture was 16 where the RMSE values in the validation (0.08), test (0.12) and training datasets (0.06) were the lowest. The RMSE, MAE and R^2 values of SVR algorithm in the training and test datasets were 0.054, 0.043, 0.759, and 0.075, 0.060, 0.483, respectively. The accuracy of estimated soil pH map produced using MLP-ANN and SVR algorithms were 55.3 and 24.22% higher than the prediction map obtained by conventional ordinary kriging. The finding of the study revealed that machine learning algorithms can be used to produce spatial estimation maps of soil properties which are costly and require intensive time and labor.

1. Introduction

Soil reaction (pH) is one of the most important indicator of soil quality due to the influence on availability of plant nutrients and biochemical reactions; therefore, reliable spatial distribution of soil pH provides valuable information on fertilizer and environmental management (Shen et al., 2013). In addition, pH controls the decomposition of soil organic matter, which increases the availability of mineral nutrients in soils, thus also affects the physical properties of soils (Shukla et al., 2006). Solubility of aluminum increases when pH is less than 5.0, and high concentration of Al^{3+} ions inhibits root growth and is toxic to most crops (Brady and Weil, 2008). In contrast to the low pH, a pH value over 8.5 indicates the high sodium ion concentration to cause dispersion of clay particles, leading to destruction of soil structure, preventing infiltration of water, causing surface runoff and water erosion, and an inappropriate environment for plant growth (Slessarev et al., 2016).

* Corresponding author.
Email address: mirackilic@adiyaman.edu.tr (Kilic M.)
http://dx.doi.org/10.56917/ljoas.7

Soil pH in a landscape may vary in a very short distance due to the changes in factors affecting the soil pH. Therefore, management of soil pH at field scale is important for crop production, and is also crucial to control water quality and land degradation at the broader scale (Merry and Sabljic, 2009). Estimation and mapping the spatial distribution of soil pH (Slessarev et al., 2016), which has a significant impact on the storage and supply of nutrients in soil, is extremely important, especially for the management of plant nutrients in agricultural production (Shen et al., 2013).

The use of machine learning approaches in digital soil mapping have recently been increased with the developments in computer and geographic information systems technologies (Wadoux et al., 2020). Digital soil mapping techniques provide reliable quantitative estimation and are fast and cost-effective approaches. Therefore, the DSM techniques have been used for the spatial estimation of many soil properties, such as soil organic matter (Rivera and Bonilla, 2020), particle size distribution (Rivera and Bonilla, 2020; Taghizadeh-Mehrjardi et al., 2020), water stable aggregates (Rivera and Bonilla, 2020), calcium carbonate equivalent (Zeraatpisheh et al., 2019), plant nutrients (Hengl et al., 2021) and salinity (Nabiollahi et al., 2021). Chen et al. (2019) developed a bootstrapping hybrid framework for estimating soil salinity with a limited number of samples. However, the DSM techniques have not been adequately applied for the spatial analysis and estimation of soil pH. Thus, this study was aimed to compare two machine learning approaches for digital mapping of soil pH in the orchards of fruit research station in Egirdir, Turkey.

2. Material and Method

2.1. Study Area and Soil Sampling

The study was carried out in orchards of Fruit Research Institute in Egirdir, Isparta, Turkey. The study area with an area of approximately 107.1 ha is located between 30° 52' 10"E - 30° 52' 45"E longitudes and 37° 48' 50"N and 37° 49' 20"N latitudes. The study area was divided into 50x50 m square grids, and the coordinates of sampling points were recorded using a GPS to analyze spatial distribution of soil pH. Ninety-one soil samples were collected from 0-30 cm depth in the corners of square grids. Two fine transects with 1, 3, 7, 12, 20 and 30 m intervals were sampled to determine the spatial variability within shorter distances than 50 m (Figure 1). Total of 103 samples were collected from the corners of grids and the fine intersects. Soil samples were dried at room temperature and sieved through 2 mm sieves and made ready for analysis. pH values of soils were determined according to U.S. Salinity Laboratory Staff (1954).

2.2. Spatial Distribution Model Approaches

2.2.1. Preparation of Dataset

Spatial distribution model approaches used in the study were Multi-Layer Perceptron Artificial Neural Network (MLP-ANN), Support Vector Regression (SVR), and Ordinary Kriging (OK). The geographical coordinates of soil sampling points were used as a covariate in both models to evaluate the estimation success of the models. The entire dataset was divided into three subsets as 70% training, 15% validation and 15% testing for the training and accuracy assessment of MLP-ANN and OK models. The dataset was randomly divided in MATLAB 2020a using the 'dividerand' command. The validation dataset was used to select the best performing MLP-ANNs produced in different architectures, the test dataset was used to estimate the accuracy of the selected MLP-ANN. Thus, the model was trained with an iterative method to avoid overfitting during the training phase, and the validation dataset was used to operate the early stop procedure (Ciaburro, 2018). The train dataset of ANN was used for SVR and OC training (70%), and the remaining dataset (30%) was used for testing.

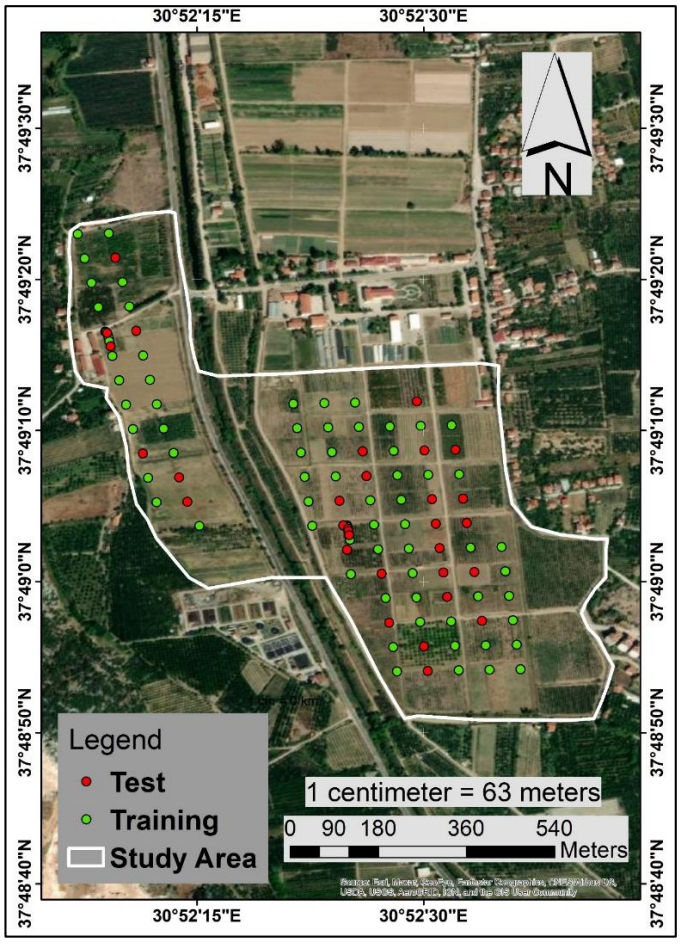


Figure 1. Locations of training and test points in the study area

2.2.2. Multi-layer Perceptron Artificial Neural Network (MLP-ANN)

The MLP-ANN neural network architecture was designed using Matlab 2020a (Figure 2). The hyperbolic tangent sigmoid (tansig) activation function was used between the input layer and the hidden layers, and the linear (pureline) transfer function was used between the hidden layers and the output layer in the MLP-ANN architecture. The number of hidden layers was determined as 2 considering the estimation accuracy by trial and error method. The number of hidden layer node was determined

by changing the number of nodes between 1 and 30, taking the prediction accuracy of the network in the validation and test datasets into account (Sergeev et al., 2019). Data standardization was achieved by mapping the minimum and maximum values of the matrix row to the range [-1 1] with the 'mapminmax' command (Ciaburro, 2018).

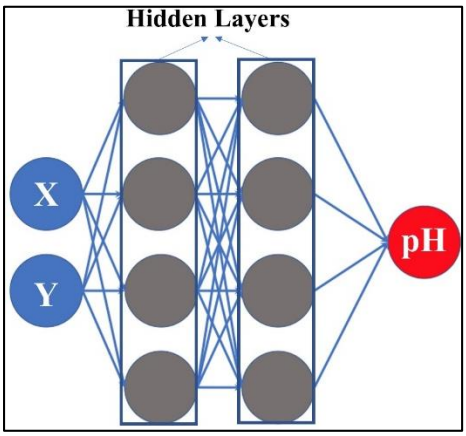


Figure 2. MLP-ANN spatial distribution model architecture of soil pH

2.2.3. Support vector regression

The second approach used in spatial modeling of soil pH was Support Vector Regression. Bayesian Optimization algorithm for the optimization of hyper parameters in Support Vector Regression was carried out using Matlab 2020a. Thus, the Minimum Objective Function, which aims to minimize the difference between the prediction and observation values, was determined (Yang et al., 2020). The acquisition functions evaluate the expected improvement in the optimization function, ignoring the values that cause the deviation in the estimation to achieve the expected improvement.

$$EI(x,Q) = E_Q\left[\max(0, \mu_Q(x_{best}) - f(x))\right] \quad (1)$$

In the equation, EI is the expected improvement, x_{best} is the lowest mean location in the previous iteration and $\mu_Q(x_{best})$ is the lowest object value in the previous iteration (Frazier, 2018).

The epsilon (ε) parameter of Support Vector Regression (SVR) determines the width of a tube around the predicted function (hyperplane) (Figure 3). The points remaining inside the tube are considered correct and are not penalized by the algorithm. Slack (ζ) measures the distance to points outside the tube, and the box constrain parameter determines the importance of this distance. The algorithm aims to minimize the error by defining a function that can place as many points as possible inside the tube and reduce the ζ. The structure focuses on linear examples. Nonlinear structures are transformed into a linear function in a multidimensional space With the kernel trick (Dobilas, 2020). In this study, the optimal values or variables were determined for SVR's Box Constrain, Kernel Scale, Epsilon and Kernel Function hyperparameters with Bayesian algorithm.

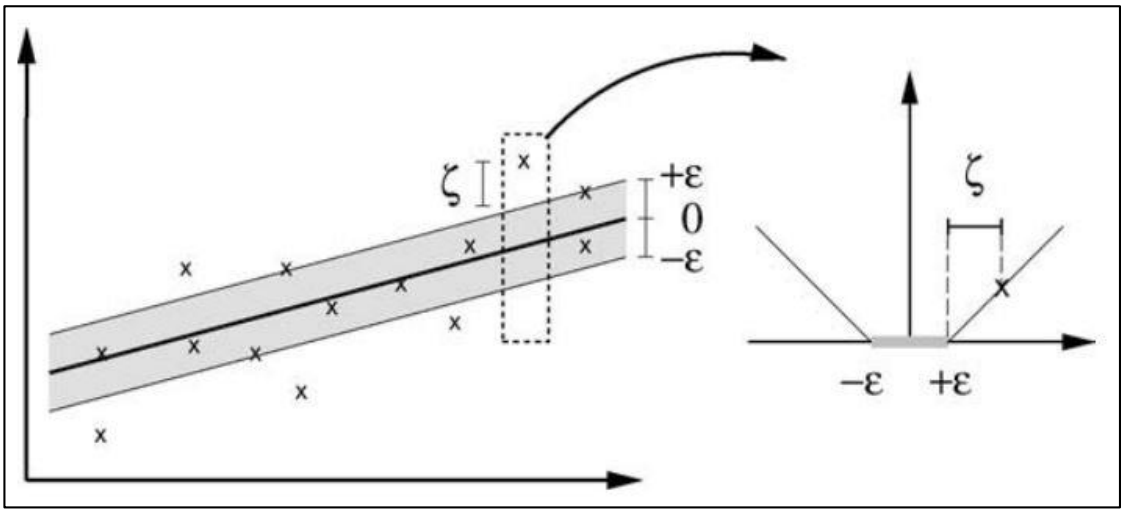


Figure 3. Hyperplane line with boundary lines defined by Support Vector Regression (SVR)-+epsilon (ε) (Schölkopf and Alexander, 2002)

2.3. Accuracy assessment

Statistical error metrics such as Root Mean Square Error (RMSE) (Equation 2), mean absolute error (MAE) (Equation 3) and coefficient of determination (R^2) (Equation 4) were used to evaluate model success.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(M_i - E_i)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_i - E_i| \quad (3)$$

$$R^2 = \frac{\sum_{i=1}^{N}(M_i - M_{mean}) \times (E_i - E_{mean})}{\sqrt{\sum_{i=1}^{N}(M_i - M_{mean})^2 \sum_{i=1}^{N}(E_i - E_{mean})^2}} \quad (4)$$

In Equations 2, 3, and 4; Mi, Ei, and n are the estimated and measured soil pH values and n the number of samples, respectively (Somaratne et al., 2005).

3. Results and Discussion

3.1. Descriptive Statistics

Parameters of descriptive statistics for training, testing and all datasets were presented in Table 1. The highest and lowest pH values of surface soils in the entire study area were 7.52 and 8.33 with a mean value of 7.95. Most field and horticultural crops grow well in moderately acidic to neutral (i.e. an optimal range of 5.5-7.0) soil conditions, while grow well in soil pH of 7–8 (Brady and Weil, 2008).

Table 1. Descriptive Statistics

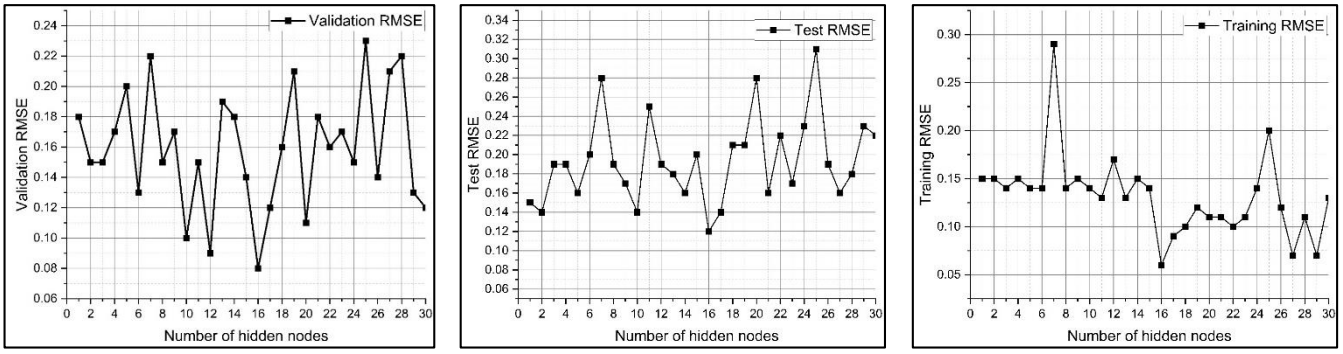
	Min.	Max.	Mean	Median	Std. Dev	CV (%)
All dataset	7.52	8.33	7.95	7.94	0.19	2.39
Training Dataset	7.52	8.33	7.94	7.94	0.18	2.33
Test Dataset	7.59	8.32	7.97	7.94	0.20	2.50

The range of pH values in study area shows that some areas in the study area may have slight problems in nutrient availability due to alkaline nature. The pH range recorded in the study area is similar to the soil pH values in semi-arid climate regions. Chen et al. (2019), who prepared a map of soil pH for the whole of China using Random Forest and XGBoost machine learning techniques, stated that the median pH in arid and semi-arid region of entire Chine was 8 or greater, and more than 90 % of soils had a pH value over 7.0. The coefficients of variation (CV%) of all, training, and test datasets were 2.39, 2.33 and 2.50%, respectively.

The descriptive statistics of the training and test dataset were quite similar with the entire dataset. The similarity in CV values revealed that the training and test dataset represent the entire dataset, and the similarity of datasets is convenient for model training and accuracy assessment (Aitkenhead and Coull, 2020).

3.2. MLP Architecture and Accuracy Assessment

The number of hidden nodes for MLP-ANN model was determined by evaluating the error statistics of the validation and test datasets. The RMSE was one of the error statistics used in this assessment process. The response of the RMSE value was determined by changing the number of hidden nodes between 1 and 30, and the number of hidden nodes with the lowest RMSE value was selected for the MLP-ANN model. The lowest RMSE values in the validation, test and training datasets were obtained with the 16th hidden node as 0.08, 0.12 and 0.06, respectively (Figure 4). Khanal et al. (2018), spatially predicted some soil properties in Iran using a set of machine learning algorithms. Similar to our study area, the soils in their study area were formed over sedimentary rocks such as limestone, sandstone, conglomerate and shale, and the RMSE value of the pH value in the ANN model was reported as 0.65. Khanal et al. (2018), used remote sensing data and many soil properties as the covariates. However, the coordinates of the sampling points were used in the estimation of pH values in the current study. Considering the limited-data conditions, the estimation success for pH values is compatible with the literature. The resolution of open source remote sensing data is insufficient; therefore, is not suitable for small areas as in our study area. In this context, the prediction success of the MLP-ANN test dataset can be considered sufficient with the literature.



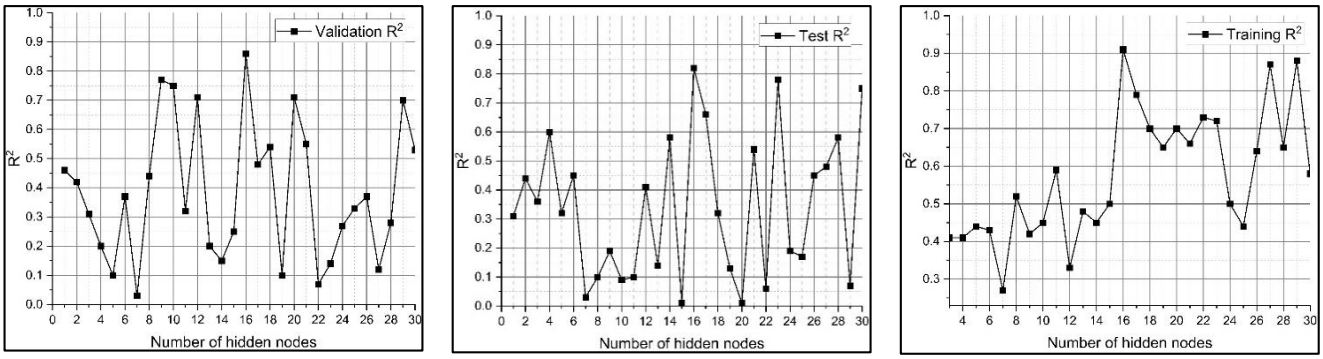
(a) (b) (c)

Figure 4. The response of RMSE values for validation (a), test (b), and training (c) datasets to the change in the number of hidden nodes

The error statistics used to determine the number of hidden nodes, which is one of the important components of the MLP-ANN architecture, is the coefficient of determination (R^2). The R^2 values between the prediction and actual values of the validation, test and training sets obtained with the change in the number of hidden nodes are shown in Figure 5. The highest R^2 values for the validation, test and training datasets were obtained in the 16th hidden node number as 0.86, 0.82 and 0.91, respectively. Tziachris et al. (2020) investigated the effect of optimized ANN models on spatial estimation success of soil pH values in the Grevena region, northern Greece. The researchers reported that the R^2 value increased from 0.278 to 0.760 with the optimization of the model hyperparameters. Similarly, the R^2 value varied considerably depending on the variation of the model parameter. The highest R^2 values in all datasets were recorded when the optimal number of neurons was 16. The result shows that the prediction success does not always increase with the increase in the number of hidden nodes.

Because, the lowest R^2 value in the test dataset was obtained at the 20th hidden node, while the R^2 value of the training and validation sets was 0.7 at this hidden node. Therefore, the optimal number of hidden nodes should be determined according to the R^2 values agreement of all three datasets to overcome the overfitting problem (Ciaburro, 2018).

Chen et al. (2019) who produced a high-resolution soil pH map of the entire China by two machine learning algorithms (Random Forest and XGBoost) assessed the quality of predictions by cross-validation. The researchers stated that the RMSE and Lin's Concordance Correlation Coefficient of predictions were acceptable (0.71 and 0.84 pH units per point, respectively).



(a) (b) (c)

Figure 5. The response of validation (a), test (b) and training (c) dataset R^2 values to the change in the number of hidden nodes

3.3. Support Vector Regression

The minimum objective function value calculated following Bayesian optimization was 0.024768 and obtained in the 18th iteration. The upper confidence limit of the objective function value using the Bayesian optimization algorithm Gaussian process model was 0.024406 for all possible hyperparameter sets (Figure 6).

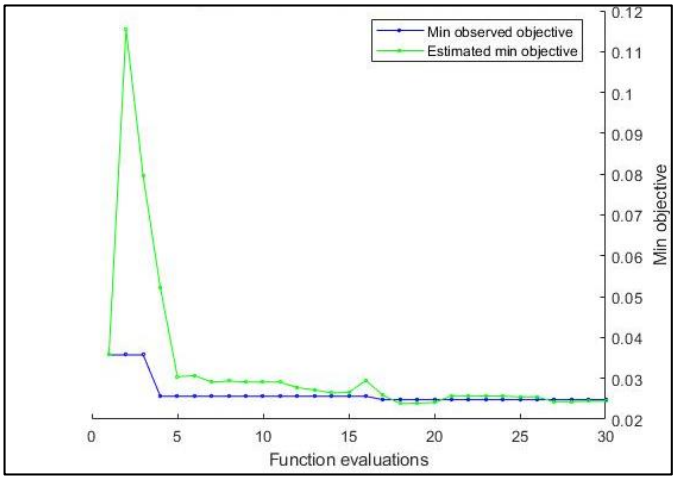


Figure 6. Minimum objective vs. Number of function evaluations

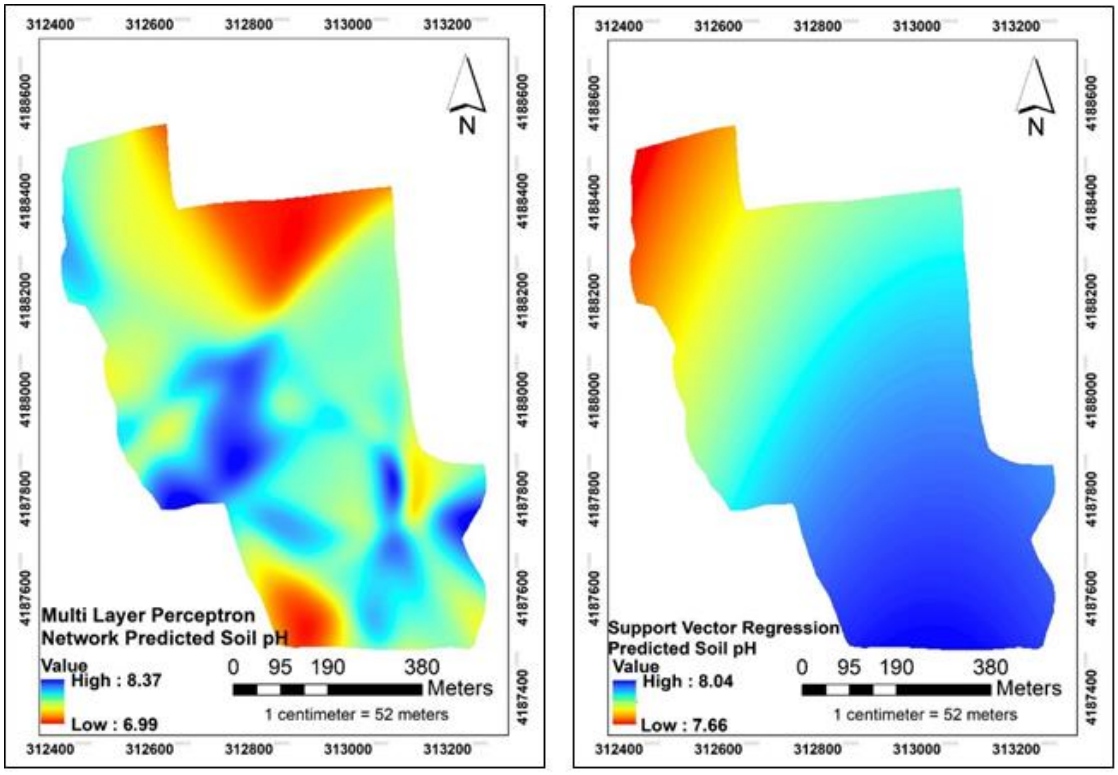
The kernel function of the SVR model under optimal hyperparameter conditions was polynomial. The degrees of box constraint, epsilon, and polynomial were 0.0031515, 0.00018779 and 2, respectively (Table 2). The RMSE, MAE and R^2 values of SVR algorithm in the training dataset were 0.054, 0.043 and 0.759, respectively. The RMSE, MAE and R^2, the error statistics, in the test dataset were 0.075, 0.060 and 0.483, respectively (Table 2). Kovačević et al. (2010) estimated soil pH values using SVR algorithm based on known values of some chemical and physical properties in soil profiles located at the Soil and Reclamation Institute, Faculty of Agriculture, University of Belgrade in eastern Serbia. The researchers reported the R^2 value as 0.90 for the linear SVR algorithm and 0.94 for the Gaussian SVR algorithm. Kovačević et al. (2010) used covariates such as soil organic matter, total nitrogen, sulfur and cation exchange capacity, which have a significant correlation with soil pH, in their SVR models. The aim of our study was to evaluate the prediction success of the machine learning algorithm with the easily obtained data. Therefore, the geographical coordinates of the sampling points were the only covariate used in the estimation of pH values. The compromise of our SVR model is acceptable compared to Kovačević et al. (2010), and the results are in good agreement with the findings of the MLP-ANN model (validation R^2 0.86, test R^2 0.82 and training R^2 0.91). Yang et al. (2019) compared soil pH prediction capabilities of SVR and ANN algorithms using near-infrared spectroscopy data of 523 soil samples collected from paddy fields in the Chinese Yangtze Plain. The RMSE value in the SVR algorithm was reported as 0.36 and the R^2 value was 0.74. In the ANN algorithm, the RMSE value was 0.33 and the R^2 value was 0.76. The prediction success of the ANN and SVR models reported by Yang et al. (2019) were quite similar, while this similarity could not be obtained in our study. The results are related to the number of samples and the attributes of the covariates used in the studies. The success of machine learning algorithms used in soil properties estimation is highly affected by the statistical (eg correlation) relationship between covariates and target variables (Aitkenhead and Coull, 2020). The spatial dependence and the cross-validation R^2 value for pH values in the same study area using the same dataset was reported as 23.27% and 0.366 by Altındal (2011) who used ordinary kriging

in estimation model. The R^2 value in this study was improved by 55.36% using the MLP-ANN and 24.22% using the SVR model compared to ordinary kriging used by Altındal (2011). The result reveals that artificial intelligence algorithms can be a reliable alternative to conventional geostatistics methods.

Table 2. The results of hyperparameters optimization and error statistics of the SVR algorithm

Box Constraint	Epsilon	Kernel Function	Polynomial Order	Training			Test		
				RMSE	MAE	R^2	RMSE	MAE	R^2
0.0031515	0.000188	Polynomial	2	0.054	0.043	0.759	0.075	0.060	0.483

Most reports on spatial distribution of soil properties indicate that variability of soil pH is less than other soil properties investigated. However, this conclusion may mislead the producers in nutrient management, because the pH value is a logarithmic transformation of hydrogen ion concentration, and the actual variability of hydrogen ions is similar to the other soil properties (Merry and Sbaljic, 2009). The maps of pH spatial prediction models are given in Figure 7. Histogram statistics of the estimated pH map produced using the MLP-ANN model were: minimum 6.99, maximum 8.37, mean 7.86 and standard deviation 0.36 (Figure 7a). The histogram statistics of the estimated pH map produced using the SVR model were; minimum 7.66, maximum 8.04, mean 7.92 and standard deviation 0.09 (Figure 7b). The coefficients of variation (CV) for the estimated spatial distribution obtained using the MLP-ANN and SVR models were 4.58 and 1.13%, respectively. The soil pH maps created with two separate machine learning models are highly similar to each other. The most significant difference was observed in the pH values of the lands located in the south of the study area. The pH values of all lands in the south were above 8.0 in the SVR model, while the pH values in this land varied between 6.99 and 8.37 in the soil pH map produced by the MLP-ANN model (Figure 7a and 7b). The result demonstrates that the MLP-ANN model is capable of generating predictive soil pH maps that are capable of representing spatial heterogeneity. Higher MLP-ANNCV value compared to the entire dataset can be attributed to the fact that the spatial estimation of pH values was only based on the geographical locations.



(a) (b)

Figure 7. Spatial distribution of soil pH maps predicted using (a) MLP-ANN, and (b) SVR models

4. Conclusion

The present study aimed to evaluate the success of nonlinear machine learning algorithms in estimating soil pH in a field with both aforementioned conditions. The results revealed that MLP-ANN and SVR algorithms provided 55.3 and 24.22% higher prediction accuracy, respectively, compared to the conventional

geostatistics kriging model. This finding indicates that machine learning algorithms can be reliably used for spatial estimation of laborious and expensive soil properties. The results have revealed that machine learning algorithms can produce predictive soil maps with high spatial resolution and provide more reliable estimation at field scale compared to the conventional geostatistics methods.

The application and success of nonlinear machine learning algorithms in estimation of soil properties using the mathematical relationship between known chemical and physical soil properties are quite abundant in the literature. In particular, the spatial estimation ability of machine learning algorithms with limited covariates is very important when access to soil analysis is not easy or open source remote sensing data cannot provide sufficient spatial detail in small areas due to poor resolution.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge the Tokat Gaziosmanpasa University of Turkey for funding this research.

Funding

This study was supported by the Scientific Research Projects Commission of Tokat Gaziosmanpasa University (Grant no; 2009/06).

Cite this article: Günal, H., Kılıç, M., Altındal, M., Gündoğan, R., 2021. Rapid Spatial Estimation of Soil pH using Machine Learning under Limited Covariate Conditions. Levantine Journal of Applied Sciences, Volume 1,30-37. http://dx.doi.org/10.56917/ljoas.7

References

Aitkenhead, M., Coull, M. 2020. Mapping soil profile depth, bulk density and carbon stock in Scotland using remote sensing and spatial covariates. European Journal of Soil Science, 71(4), 553-567.

Altındal, M., 2011. Soil survey of the Egirdir Horticultural Research Institute lands and mapping of the spatial distribution of plant available micronutrient contents. Master's thesis, Gaziosmanpaşa University, Institute of Science and Technology, Tokat, Turkey (in Turkish).

Brady, N. C., Weil, R. R., Weil, R. R. 2008. The nature and properties of soils (Vol. 13, pp. 662-710). Upper Saddle River, NJ: Prentice Hall.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., … Shi, Z. (2019). A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. Science of The Total Environment, 655, 273–283. https://doi.org/10.1016/j.scitotenv.2018.11.230

Ciaburro, G., 2018. MATLAB for Machine Learning, Journal of Materials Processing Technology.

Dobilas, S., 2020. Support Vector Regression (SVR) — One of the Most Flexible Yet Robust Prediction Algorithms [www Document]. Towards Data Science. URL https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60

Frazier, P.I., 2018. A Tutorial on Bayesian Optimization. arXiv preprint arXiv:1807.02811. https://doi.org/https://doi.org/10.48550/arXiv.1807.02811

Hengl, T., Miller, M. A., Križan, J., …. & Crouch, J. 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Scientific Reports, 11(1), 1-18. https://doi.org/10.1038/s41598-021-85639-y

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Computers and Electronics in Agriculture, 153, 213–225. https://doi.org/10.1016/j.compag.2018.07.016

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154, 340–347. https://doi.org/10.1016/j.geoderma.2009.11.005

McGeorge, W.T., 1954. Diagnosis and Improvement of Saline and Alkaline Soils. Soil Science Society of America Journal 18, 348.

Merry, R. H., Sabljic, A., 2009. Acidity and alkalinity of soils. Environmental and ecological chemistry, 2, 115-131.

Nabiollahi, K., Taghizadeh-Mehrjardi, R., Shahabi, A., Heung, B., Amirian-Chakan, A., Davari, M., & Scholten, T. (2021). Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. Geoderma, 385, 114858. https://doi.org/10.1016/j.geoderma.2020.114858

Rivera, J. I., Bonilla, C.A., 2020. Predicting soil aggregate stability using readily available soil properties and machine learning techniques. Catena, 187, 104408. https://doi.org/10.1016/j.catena.2019.104408

Schölkopf, B., Alexander, J.S., 2002. Learning with kernels: support vector machines, Regularization, 1.2. ed. MIT press.

Sergeev, A. P., Buevich, A. G., Baglaeva, E. M., & Shichkin, A. V. (2019). Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. CATENA, 174, 425–435. https://doi.org/10.1016/j.catena.2018.11.037

Shen, C., Xiong, J., Zhang, H., Feng, Y., Lin, X., Li, X., ... & Chu, H. (2013). Soil pH drives the spatial distribution of bacterial communities along elevation on Changbai Mountain. Soil Biology and Biochemistry, 57, 204-211. https://doi.org/10.1016/j.soilbio.2012.07.013

Shukla, M. K., Lal, R., Ebinger, M. 2006. Determining soil quality indicators by factor analysis. Soil and Tillage Research, 87(2), 194-204. https://doi.org/10.1016/j.still.2005.03.011

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y., Schimel, J. P., Chadwick, O.A., 2016. Water balance creates a threshold in soil pH at the global scale. Nature, 540(7634), 567-569. http://dx.doi.org/10.5281/zenodo.61996

Somaratne, S., Seneviratne, G., Coomaraswamy, U., 2005. Prediction of Soil Organic Carbon across Different Land-use Patterns. Soil Science Society of America Journal 69, 1580–1589. https://doi.org/10.2136/sssaj2003.0293

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., & Schmidt, K. (2020). Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma, 376, 114552. https://doi.org/10.1016/j.geoderma.2020.114552

Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., Doukas, I.D., 2020. Comparing Machine Learning Models and Hybrid Geostatistical Methods Using Environmental and Soil Covariates for Soil pH Prediction. ISPRS International Journal of Geo-Information 9, 276. https://doi.org/10.3390/ijgi9040276

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

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. https://doi.org/10.1016/j.geoderma.2020.114337

Yang, M., Xu, D., Chen, S., Li, H., Shi, Z., 2019. Evaluation of Machine Learning Approaches to Predict Soil Organic Matter and pH Using vis-NIR Spectra. Sensors 19, 263. https://doi.org/10.3390/s19020263

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.2018.09.00