М.А. Гендельманның 110 жылдығына арналған «Сейфуллин оқулары – 19» халықаралық ғылыми - практикалық конференциясының материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 19», посвященной 110-летию М.А. Гендельмана». - 2023.- Т. I, Ч. IV. – Р. 258-263.

**UDC№63. 631.4.41**

# OPPORTUNITIES OF USING THE RANDOM FOREST AND SUPPORT VECTOR REGRESSION MODEL IN ESTIMATING SOIL ORGANIC CARBON CONTENT

*Sümeyye GÜLER*[*]
*Kastamonu University, Institute of Science, Sustainable Forestry Doct. Programme, Kastamonu, TURKEY*
*Bülent. TURGUT*
*Karadeniz Technical University, Faculty of Forestry, Dept. of Forest Soil and Ecology, Trabzon, TURKEY*
*Sezgin. AYAN*
*Kastamonu University, Faculty of Forestry, Silviculture Department, Kastamonu, Turkey*
*Corresponding author:* sumeyyeglr01@gmail.com

**Abstract:** Soils are the largest terrestrial organic carbon pool and contain different amounts of soil organic carbon depending on the geological structure, climatic conditions, soil characteristics, stand age, stand density, land use, and management. A forest ecosystem is any piece of forest that is homogeneous in terms of the composition, characters, and interrelationships of its constituent elements in a particular location. In forest ecosystems, carbon is stored both in the vegetative mass and in the soil. Accurate estimation of soil organic carbon is critical to support the estimation of changes in the earth's carbon balance. Random Forest is a Supervised Machine Learning Algorithm widely used in classification and Regression problems. It creates decision trees on different samples and gets majority votes for classification and mean in case of regression. Support Vector Regression, on the other hand, can be defined as a vector space-based machine learning method that finds a decision boundary between the two classes that are furthest from any point in the training data. These two machine learning methods have been preferred because they reach the target faster, with less cost, and with less manpower compared to conventional methods. This study discusses the usability of Random Forest and Support Vector Regression modeling methods in predicting soil carbon content using NDVI (Normalized Difference Vegetation Index) values indicating vegetation density, as well as landforms, and climate data.

**Keywords:** Organic carbon, remote sensing, machine learning

## 1. INTRODUCTION

Soil is the largest carbon reserve in the terrestrial biosphere. Carbon is the main element of the cells and the biological system. Carbon is in the atmosphere, in the structure of living things, in organic wastes, fossil fuels, rocks, and oceans. Plants convert inorganic carbon into organic carbon with photosynthesis. Decomposed above and underground parts of plants form a significant part of soil organic carbon. Soils have variability according to their organic carbon content.

Soil organic carbon plays an essential role in the global carbon cycle and controls soil quality and productivity, reduces, and eliminates the negative effects of climate change, and sustains the use of farmland (Juricova et al., 2022; Ferre et al., 2014). The main source of soil organic carbon is animal residues, dead and living microorganisms, and above and below-ground plant residues (Sitch et al., 2003). Soil organic carbon is a measure of soil organic matter and the main energy source for soil microorganisms. Soil organic carbon is stored with soil and vegetation biomass in the forest.

Methods for spatial prediction of SOC have been developed by researchers in recent years. A common approach is an interpolation based on sample data (such as Kriging) (Dou et al., 2010, Elbasiouny et al., 2014). However, a high interpolation accuracy requires a large sample size that is labor-intensive and costly to collect. The SOC at a location is related to climate, terrain, and vegetation. environmental variables are increasingly used in SOC estimation to improve prediction accuracy with limited samples due to their being regulated by interactions (Grinand et al., 2017, Lamichhane et al., 2019). Therefore, it is necessary to develop effective environmental variables and effective estimation methods to predict accurate SOC. The use of neural network (NN) models for soil organic carbon estimation based on remote sensing data has become popular in the last decade.

This study aims to review advances in machine learning (ML) for the estimation of SOC using ecosystem components such as soil properties, climate, landforms, and vegetation.

## 1.1 Variables Affecting Soil Organic Carbon Storage

The organic carbon storage capacity of the soil is affected by land use/land cover, climate, net primary production (NPP), soil class, parent material, slope, aspect, and elevation. The amount of organic carbon storage in the soil is a function of the dynamic balance between the soil properties and the organic material entering and leaving the soil.

### 1.1.1 The Effects of Climate on Soil Organic Carbon

Precipitation and temperature are the most important factors affecting the amount of organic matter in the soil in a regional sense. In humid regions with high rainfall, the organic matter and organic carbon content of the soils are generally higher. The amount of vegetable waste mixed with the soil is also high due to the net crop production is high in the lands of this region. This ensures that the amount of soil organic matter, and therefore the amount of organic carbon, is high. For the decomposition of organic matter in the soil, the living conditions of microorganisms must be at an optimum level. For this reason, in climates where precipitation is high, and temperature is low, organic matter cannot decompose and organic carbon value is high. In climatic conditions where precipitation and temperature are high, the humification of organic matter is high, as optimum conditions for soil organisms are formed. Organic carbon content is low in soils with this climate, as organic matter is constantly broken down and used as an energy source. In regions with low precipitation and high temperature, the amount of organic carbon is also low due to the net crop production is low.

### 1.1.2 The Effects of Land Use on Soil Organic Carbon

Organic carbon is stored and retained in the ecosystem by soil and plants. Plants convert the $CO_2$ taken from the atmosphere into organic compounds through photosynthesis. These organic compounds mix with the soil after the death of the plants and can be stored in the soil for many years. Levels of SOC depends on inputs of plant litter and rhizodeposition (Li et al., 2010; Park et al., 2012; Six et al., 2004). Disruption of an area's natural vegetation or inappropriate soil management practices decrease soil organic matter content (Park et al., 2012; Murty et al., 2002). Researchers have reported that changes in field use also cause changes in SOC content (Murty et al., 2002; Post and Kwon, 2000).

### 1.1.3 The Effects of Topography on Soil Organic Carbon

Another factor affecting the organic carbon storage capacity of the soil is topographic factors. The organic carbon content of sloping soils is higher than that of flat and lowland lands (Birkeland, 1984). The vegetation cover is relatively weak in sloping soils and the annual organic matter input to the soil is lower due to surface runoff. On the other hand, the organic matter content of the soils on the north-facing is higher than the south-facing soils under the same climate and slope conditions (Fanning and Fanning, 1989). Studies show that erosion is slower on north-facing slopes as the snow melts more slowly, as water seeps into the soil before it runs off, and as a result, the soils on the northern slopes are deeper. Since the dense vegetation supported by the deep soils allows more organic matter to be added to the soil annually, the organic carbon content of these soils is higher than the soils on the south-facing slopes (Birkeland, 1984). In addition, wetter north-facing slopes results in slower decomposition of organic matter added to the soil and, consequently, more organic carbon accumulation in the soil.

## 1.2 NDVI Values by Stand Coverage

The Normalized Vegetation Index (NDVI) is used to obtain information about the biophysical properties of vegetation, such as density, canopy, and plant density. NDVI is often considered one of the most critical factors influencing changes in soil organic carbon and plant productivity. Therefore, it is expected to show a significant correlation with SOC variation. NDVI is calculated as:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)}$$

NIR; the near-infrared wavelength of the light spectrum

RED; red zone wavelength

NDVI; represents the vegetation index value.

## 1.3 The Relationship between tree canopy cover and Soil Organic Carbon

It is stated that there is a relationship between the amount of organic matter accumulated on the tree canopy cover and the stand base (Kara et al., 2008). It is known that the cover of the tree canopy cover affects the penetration of light into the stand and thus plays an important role in the decomposition of the dead material. In addition, the higher the leaf surface index, the higher the amount of organic material that is shed. In the light of this information, it was assumed that it would be possible to determine the void ratio and to reveal the relationship between it and the amount of organic matter with photographs taken perpendicular to the stand roof.

The characteristics of the establishment of the forest, such as age, closure, density, mixing ratio, also affect the soil organic matter. In closed forests, it can be expected that the amount of carbon stored is high, due to the low amount of light and temperature reaching the stand, and the greater amount of leaf shedding from the trees. For the same reasons, the amount of carbon stored in the soil may increase depending on the age of the forest. In mixed forests, litter decomposition is faster because of more microbiological activity and more favorable ecological conditions.

There are many articles and models used on the modeling of SOC stocks. To run these models, inputs such as organic matter input to soils, decomposition rates of these organic substances and C, N, C/N, lignin contents are required. However, it has been revealed during the literature research that the studies on the mentioned inputs are quite limited in our country's conditions. Thereupon, the focus is on models that allow estimations for areas without data based on the relationships between soil, climate, bedrock/material, land use, vegetation and topographic factors and SOC. For this, kriging and machine learning algorithms were examined, and it was decided to use Random Forest and Support Vector Regression models.

### 1.4 Artificial Intelligence Techniques Used for Soil Organic Carbon Estimation

#### 1.4.1 Random Forest (RF)

Random Forest is a flexible, easy-to-use machine learning algorithm that produces great results most of the time, even without hyperparameter tuning (Figure 1). It is also one of the most used algorithms because it can be used for both simplicity and classification and regression tasks. The development and popularity of machine learning methods in recent years has revealed many methods that can be beneficial in modeling.

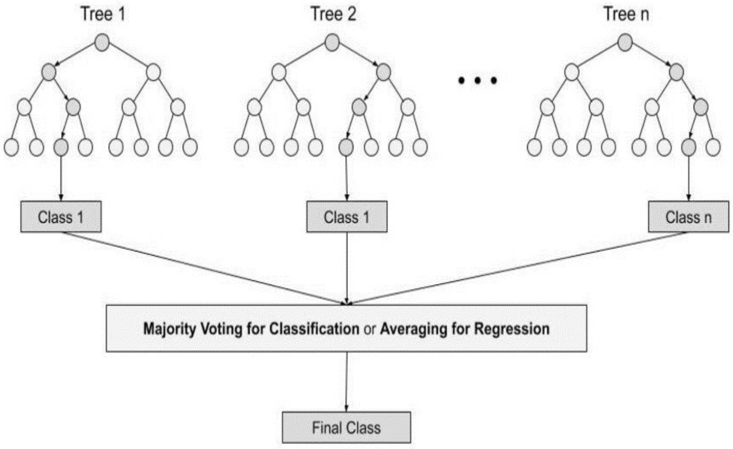
Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

**Figure 1.** Random Forest algorithm mechanism

In addition to being a method developed from RF, classification, and regression trees, it is shown as one of the most successful decision trees. In recent years, this method has been used in many disciplines and there are remarkable studies in areas such as ecology, climate change, and remote sensing (Breiman, 2001; Liaw and Wiener, 2002; Pal, 2005; Gislason et al., 2006; Evans and Cushman, 2009; Evans et al., 2011). Unlike traditional classification and regression trees, the RF method creates many decision trees and provides the opportunity to evaluate through the combination of these trees. The structure in which the decision trees are formed in the RF method is called the forest. Each decision tree in the forest is created by selecting samples from the data set with the bootstrap technique and determining several random variables among all variables at each node.

### 1.4.2 Support Vektor Regresyon

Support Vector Regression, as the name suggests, is a regression algorithm that supports both linear and non-linear regressions. This method works on the principle of Support Vector Machine. SVR differs from SVM in that SVM is a classifier used to predict discrete categorical labels while SVR is a regressor used to predict continuous ordinal variables.

In simple regression, the idea is to minimize the error rate, whereas in SVR the idea is to fit the error within a certain threshold, so the job of the SVR is to approximate the best value within a certain margin called ε-tube (Figure 2).

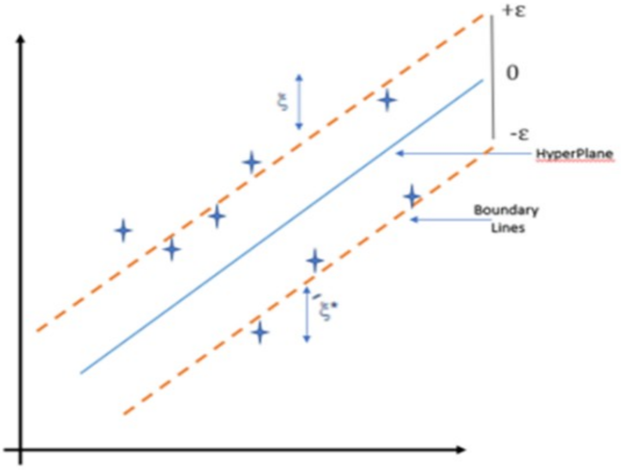

**Figure 2.** Support Vector Regression model representation

## 2. CONCLUSION

Soil organic carbon (SOC) is a reliable indicator of soil fertility and land management. Remote sensing data is widely used to estimated SOC variation. In this study, random forest (RF) and support vector regression (SVR) models used to estimate soil organic carbon content with normalized difference vegetation index (NDVI) values using satellite images (Sentinel 1- Sentinel 2) were revived. The varying importance of RF and SVR use compared to other studies has shown remarkable consistency, although some differences are evident. It has been found that multi-temporal data is more predictive than single-date data, and SVR outperforms the RF algorithm. The fact that the data in the dataset is sparse and easy to classify allows SVM to run faster and give better results. However, it gives good results in Random Forest but does not match SVM for the dataset. The choice of algorithm depends on the desired result. While both models are consistent, the performance of the algorithm is heavily dependent on data quality.

## List of used literature

1. Birkeland, P.W. (1984). Soils and geomorphology. Oxford University Press. New York, NY.
2. Breiman, L. (2001). Random Forest Mach. Learn., 45, pp. 5-32 https://doi.org/10.1023/A:1010933404324.
3. Dou, F.G., Yu, X., Ping, C.L., Michaelson G., Guo L.D., & Jorgenson T. (2010). Spatial variation of tundra soil organic carbon along the coastline of northern Alaska. *Geoderma* 154 (3–4), pp. 328-335 https://doi.org/10.1016/j.geoderma.2009.10.020
4. Elbasiouny, H., Abowaly, M., Abubaker, A., & Gad, A. (2014) Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta Egypt. *Catena* 113: 70-78 https://doi.org/10.1016/j.catena.2013.09.008
5. Evans, J.S., Cushman, S.A. (2009). Gradient modeling of conifer species using random forests. *Landscape Ecology* 24(5): 673-683.
6. Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A. (2011). Modeling Species Distribution and Change Using Random Forest. In Predictive Species and Habitat Modeling in Landscape Ecology. Springer New York, pp. 139-159.
7. Fanning, D.S., ve Fanning, M.C.B. (1989). Soil Morphology, Genesis, and classification. John Wiley & Sons. New York, USA.
8. Ferré, C., Comolli, R., Leip, A., Seufert, G. (2014). Forest conversion to poplar plantation in a Lombardy floodplain (Italy): effects on soil organic carbon stock. *Biogeosciences* 11: 6483–6493.
9. Grinand, C., Le Maire,G., Vieilledent, G., Razakamanarivo, H., Razafimbelo, T., & Bernoux, M. (2017). Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing *Int. J. Appl. Earth Obs. Geoinf.*, 54, pp. 1-14, 10. https://doi.org/10.1016/j.jag.2016.09.002.
10. Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R. (2006). Random forests for land cover classification. *Pattern Recognition Letters* 27(4): 294-300.
11. Juřicová, A., Chuman, T., & Žížala, D. (2022). Soil organic carbon content and stock change after half a century of intensive cultivation in a chernozem area. *Catena* 211(May 2021). https://doi.org/10.1016/j.catena.2021.105950
12. Kara, Ö., Bolat, İ., Çakıroğlu, K., Öztürk, M. (2008). Plant canopy effects on litter accumulation and soil microbial biomass in two temperate forests. *Biology Fertility of Soils* 45: 193–198.
13. Lamichhane, S., Kumar, L., & Wilson B. (2019) Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A Review Geoderma, 352, pp. 395-413, https://doi.org/10.1016/j.geoderma.2019.05.031
14. Liaw, A., Wiener, M. (2002). Classification and regression by random forest. *R News*, 2(3): 18-22.
15. Li, Y., Xu, M., M. Zou, M., & Xia, Y. (2010). Soil CO2 efflux and fungal and bacterial biomass in a plantation and a secondary forest in wet tropics in Puerto Rico. *Plant Soil* 268: 151-160.
16. Murty, D., Kirschbaum, M.U.F., McMurtrie, R.E., & McGilvray H. (2002). Does conversion of forest to agricultural land change soil carbon and nitrogen? A review of the literature Glob. *Change Biol.* 8: 105-123.

17. Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26(1): 217-222.

18. Park, C.W., Ko, S., Yoon, T.Y, Han, S., Yi, K., Jo, W., Jin, L., Lee, S.J., Noh, N.J. & Chung, H.Y. (2012). Son Differences in soil aggregate, microbial biomass carbon concentration, and soil carbon between *Pinus rigida* and *Larix kaempferi* plantations in Yangpyeong, Central Korea. *For. Sci. Technol*. 8: 38-46.

19. Post, W.M., Kwon, K.C. (2000) Soil carbon sequestration and land use change: processes and potential. *Glob. Chang. Biol.* 6 (3) 317–327.

20.Six, J., Bossuyt, H., Degryze, S., & Denef K. (2004). A history of research on the link between (micro) aggregates, soil biota, and soil organic matter dynamics. *Soil Tillage Res*. 79: 7-31.

21. Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J.O., Levis, S., Lucht, W., Sykes, M.T., Thonicke, K., & Venevsky S. (2003). Evaluation of ecosystem dynamics, plant geography, and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Glob. Change Biol*. 9 pp. 161-185, https://doi.org/10.1046/j.13652486.2003.00569.x.