



# The Least Limiting Water Range to Estimate Soil Water Content Using Random Forest Integrated with GIS and Geostatistical Approaches

Pelin ALABOZ<sup>a\*</sup>, Orhan DENGİZ<sup>b</sup>

<sup>a</sup>Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Isparta University of Applied Sciences, Isparta, Turkey

<sup>b</sup>Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Ondokuz Mayıs University, Samsun, Turkey

## ARTICLE INFO

### Research Article

Corresponding Author: Pelin ALABOZ, E-mail: pelinalaboz@isparta.edu.tr

Received: 29 Jun 2022 / Revised: 15 Feb 2023 / Accepted: 16 Feb 2023 / Online: 24 Oct 2023

### Cite this article

ALABOZ P, DENGİZ O (2023). The Least Limiting Water Range to Estimate Soil Water Content Using Random Forest Integrated with GIS and Geostatistical Approaches. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi)*, 29(4):933-946. DOI: 10.15832/ankutbd.1137917

## ABSTRACT

Algorithms that exist in every area today have become the center of our lives with technological developments. The uses of machine learning algorithms are being researched with the new developments in the agricultural field. The present study determined the least limiting water range (LLWR) contents of alluvial lands with different soils distributed in the Bafra Plain, where intensive agricultural activities are carried out, and revealed the compression and aeration problems in the area with distribution maps. Also, the predictability of LLWR was evaluated with the random forest (RF) algorithm, one of the machine learning algorithms, and the usability of the prediction values distribution maps was revealed. The LLWR contents of the soils varied in the range of 0.049-0.273 cm<sup>3</sup> cm<sup>-3</sup> for surface soils. There were aeration problems in 6.72%, compaction problems in 20.16%, and aeration and compaction problems in 0.8% of the surface soils examined in the study area. Furthermore, 72.32%

of the soil was under optimal conditions. For the 20-40 cm depth, an aeration problem in 5.88%, a compaction problem in 28.57%, and both an aeration and a compaction problem in 2.52% of the points were detected. In estimating LLWR with the RF algorithm, the root mean square error (RMSE) value obtained for 0-20 cm depth was determined to be 0.0218 cm<sup>3</sup> cm<sup>-3</sup>, and for 20-40 cm depth, it was 0.0247 cm<sup>3</sup> cm<sup>-3</sup>. In the distribution maps of the observed and predicted values obtained, the lowest RMSE value was determined by the SK interpolation methods for 0-20 cm depth and the OK interpolation methods for 20-40 cm. The distribution of obtained and predicted values in surface soils was similar. However, variations were found in the distribution of areas with low LLWR below the surface. As a result of the study, it was determined that LLWR can be obtained with a low error rate with the RF algorithm, and distribution maps can be created with lower error in surface soils.

Keywords: Physical properties, Moisture constants, Machine learning, Bafra delta plain

## 1. Introduction

Today, with the developing technology, evaluating soil and product-yield status has become essential for managing and using soils without losing their functions in the terrestrial ecosystem and the ecosystem services they provide. The aim of applications such as global positioning system, geographic information systems (GIS), remote sensing, yield monitoring, and estimation, and applications such as smart agriculture and precision agriculture for agricultural production is to make crop production within the scope of maximum efficiency and ecological-economic sustainability with minimum input amount. In this context, the physical properties of the soil are critical quality indicators that directly or indirectly affect plant production and yield elements (Şenol et al. 2020). Among the physical parameters of soils, properties such as soil water content, air-filled porosity, temperature, and penetration resistance directly affect plant root growth. In contrast, other properties such as bulk density, texture, aggregate stability, and pore size distribution have an indirect effect (Letey 1958). However, some of these parameters are negatively affected due to compression and compaction due to the pressure applied under heavy field traffic, and the bulk density of the soil increases. From an agricultural point of view, soil or soil layers are considered compacted when the porosity of soils, especially air-filled porosities, is low enough to limit aeration. In such a case, each or both fluids in the soil, air, and water are partially removed from the compacted soil mass, decreasing soil porosity. As a result of the increase in the proportion of small pores in the pore size distribution, aeration, root penetration, water flow, and drainage are prevented.

Aksakal (2004) states that, due to soil compaction, the penetration resistance to root growth increases while the water holding and aeration capacity decreases.

Due to compaction, the soil's changing penetration resistance or bulk density value cannot clearly explain the plant development status. The LLWR feature, a combination of various soil physical properties, is considered one of the soil structural quality indicators. Letey (1958) defined the non-LWR as the water ranges affected by the water content that the plant and the aeration and penetration resistance can take up. Da Silva et al. (1994) developed the LLWR approach by evaluating the soil bulk density in the model. LLWR is the soil water content at which limitations on plant growth in relation to water potential, aeration, and penetration resistance are minimal. The upper limit of the effect of LLWR on root growth was determined as air-filled pore volume (10%) or field capacity, the lower limit was determined as the wilting point or soil water content at which root growth was limited, and a 2 MPa soil penetration resistance occurred. (Da Silva et al. 1994). LLWR decreases with the increasing penetration resistance and bulk density with soil compaction (Haghighi Fashi et al. 2017). Some research has also reported that LLWR's wide variation range enables plants to utilize soil water more effectively and positively affects crop yield (Da Silva & Kay 1997; Chan et al. 2006). Neğiş et al. (2020) stated that soil compaction decreased, whereas LLWR increased with the organic material application. However, Alaboz et al. (2021) evaluated that LLWR showed a positive correlation with clay, organic matter, and CaCO<sub>3</sub> and a negative correlation with bulk density.

LLWR determination is a quality indicator that is a combination of laborious and time-consuming features. Therefore, studies on the predictability of the LLWR feature by pedotransfer functions have been carried out (Da Silva & Kay 1997; Leão et al. 2005; Tavanti et al. 2019). Alaboz et al. (2021) determined LLWR with deep learning with higher prediction accuracy than ANN. Akar & Güngör (2013) have stated that the random forest (RF) algorithm, one of the machine learning algorithms, is generally preferred since it shows higher accuracy than other approaches. Also, Watts & Lawrence (2008) have stated that the RF algorithm provides high accuracy in determining agricultural regions when applied to an object-oriented approach. RF algorithm, a learning-based approach, is generally used in digital soil mapping (Stum et al. 2010; Machado et al. 2019), and the studies on soil physical properties remain limited.

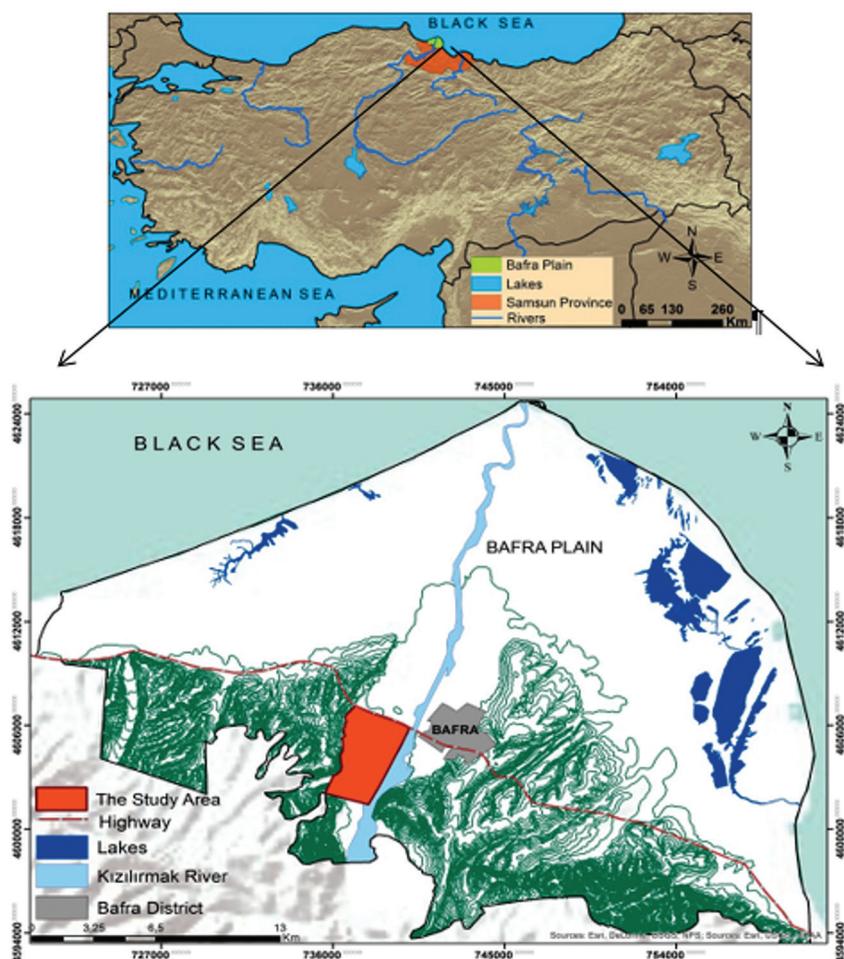
It is challenging to represent the area of the point samples of soil properties that vary depending on many factors in the field. Also, revealing spatial evaluations instead of point values in determining dynamic features contributes to sustainable management. Geostatistics, which estimates variables by interpolating between variables that do not have observations in a particular observation area and variables with observations, is frequently used to close this gap (Mihalikova et al. 2016; Tunçay et al. 2018). Developing computer, sensor technologies, and programs can easily reveal the variability of soil properties with spatial distribution maps. Alaboz et al. (2020) determined the interpolation methods showing the highest accuracy in the field capacity and wilting point distributions of soils as OK's Gaussian [root mean square error (RMSE): 4.289%] and Cokriging (RMSE: 3.187%), respectively. On the other hand, According to Tunçay et al. (2018) determined the lowest mean absolute error (MAE) and mean squarer error (MSE) values with the regression kriging method in the creation of field capacity distribution maps, while the wilting point was obtained with the Cokriging method. Furthermore, the distribution of the observed values and the values estimated from the algorithms with different methods showed a similar pattern (Alaboz et al. 2020; Şenol et al. 2020).

The present study aimed to; i) determine the least limiting water range (LLWR) contents in the alluvial lands distributed in the Bafra Plain formed on the sediments carried by the Kızılırmak river and reveal the compression and aeration problems in the area with distribution maps, ii) evaluate the predictability of LLWR with the RF algorithm and iii) determine the usability of the distribution maps of the estimated values obtained.

## 2. Material and Methods

### 2.1. General characteristics of the study area

The present study was conducted in the Samsun-Bafra delta plains of the Kızılırmak River in the Central Black Sea Region of Turkey (Figure 1).



**Figure 1- Location maps of the study area**

The study area is 30 km from the Samsun city center and covers an area of approximately 1365.4 ha. The climate in the region is semi-humid. The average temperature is 22.2 °C in July and 6.9 °C in January. The annual average temperature is 13.6 °C. Precipitation and evaporation are 764.3 mm and 726.7 mm, respectively (TSMS, 2021). According to the Soil Survey Staff (2014), soil temperature and moisture regimes are mesic and ustic, respectively. The study area is slightly sloping (0.0-2.0%) and is mainly located on the river alluvium carried by the Kızılırmak River. Also, some of the soils of the study area are distributed on colluvial clay deposits from the slopes located in the northwest parts of the area. The soils in the study area were classified as Vertisol, Inceptisol, and Entisol (Soil Survey Staff 2014), and Regosol, Fluvisol, Leptosol, Cambisol, and Vertisol, according to WRB (Figure 2). Intensive agriculture, including vegetables, fruits, and grains, is carried out on the flatlands in the study area.

## 2.2. Soil sampling and analysis

A total of 214 soil samples were taken from the study area, including surface (0-20 cm) and subsurface (20-40 cm) samples (Figure 2). Soil texture, gravimetric water content, saturated water content-field capacity, and wilting point values were determined by methods described by Gee & Bauder (1986), Blake & Hartge (1986) & Klute (1986), respectively. Penetration resistance measurements were determined by penetrometer (Eijkelkamp 1990). Bulk density was determined using undisturbed soil sampling cylinders.

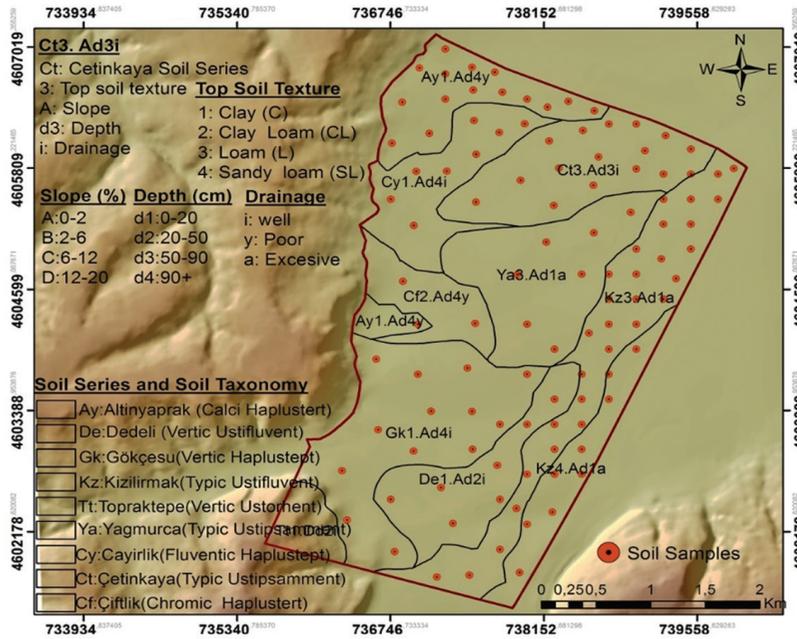


Figure 2- Soil map and soil samples pattern of the study area

The LLWR contents of the soils were determined according to Wu et al. (2003). Using the moisture content at 10% air-filled pore volume ( $\theta_{Ap}$ ), field capacity ( $\theta_{FC}$ ), wilting point ( $\theta_{WP}$ ), and moisture content at 2MPa penetration resistance ( $\theta_{PR}$ ), four possibilities were evaluated as follows, and LLWR was calculated.

- 1- If  $\theta_{Ap} \geq \theta_{FC}$  and  $\theta_{PR} \leq \theta_{WP}$ , then  $LLWR = \theta_{FC} - \theta_{WP}$  (Available water content of the soil)
- 2- If  $\theta_{Ap} \geq \theta_{FC}$  and  $\theta_{PR} \geq \theta_{WP}$ , then  $LLWR = \theta_{FC} - \theta_{PR}$  (Penetration resistance limits root development)
- 3- If  $\theta_{Ap} \leq \theta_{FC}$  and  $\theta_{PR} \leq \theta_{WP}$ , then  $LLWR = \theta_{Ap} - \theta_{WP}$  (Aeration is poor)
- 4- If  $\theta_{Ap} \leq \theta_{FC}$  and  $\theta_{PR} \geq \theta_{WP}$ , then  $LLWR = \theta_{Ap} - \theta_{PR}$  (Plant growth is limited as both aerations are poor and penetration resistance is high).

To determine the moisture content ( $\theta_{PR}$ ) at 2 MPa, the moisture obtained depending on different depths was calculated according to Busscher's (1990) Equation 1, taking into account penetration resistance measurement and bulk density values. The coefficients of the equation were found in a similar study by Alaboz et al. (2021) evaluated as stated.

$$PR = a\theta_b^d \quad \text{Equation (1)}$$

The water content in which the aeration porosity is 10% was calculated by the equation (2):

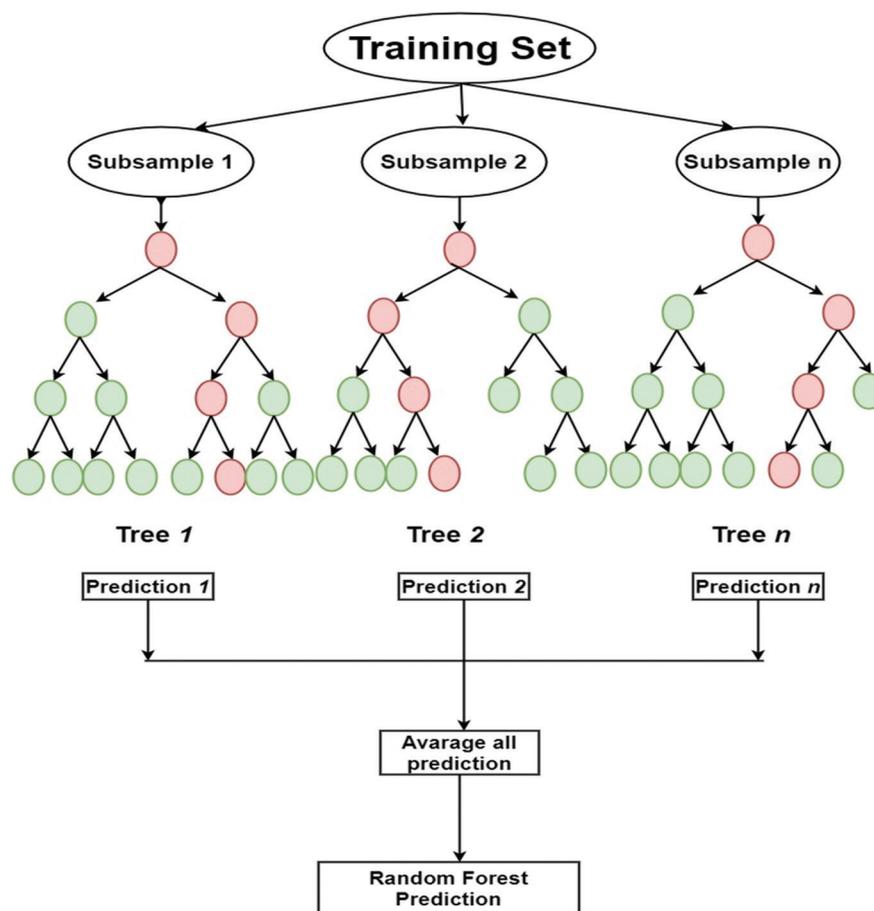
$$\theta_a = \theta_s - 0.10 \quad \text{Equation (2)}$$

Here, PR is the penetration resistance (MPa),  $\theta$  is the volumetric water content,  $D_b$  is the bulk density ( $g\ cm^{-3}$ ), and  $\theta_s$  is the saturated soil water content ( $cm^3\ cm^{-3}$ ).

### 2.3. Prediction approach using RF

RF is one of the tree-type learning algorithms. In the RF method,  $[h(x, \theta_k) \ k=1, \dots]$  tree-type classifiers are used. Here,  $x$  represents the input data, and  $\theta_k$  represents the random vector (Breiman 2001). The RF classifier parameters are the number of variables used at each node (mtry) and the number of trees to be developed (ntree) to determine the best split (Pal 2005). The user randomly selects the initial mtry value, increased or decreased, according to the next generalized errors, or the most appropriate mtry is determined by performing the tuning process. Thus, classification precision increases, whereas error decreases. According to Breiman (2001), when choosing the mtry variable value, the total number of variables equal to the square root usually gives optimum results. The RF uses the classification and regression tree (CART) algorithm to develop the largest tree without pruning (Breiman 2001). The CART algorithm divides a node by applying a certain criterion. The RF method adopts the Gini index. The cleavage position with the smallest Gini index is determined

with Gini measurements. When the Gini index reaches zero, the tree-branching process ends when one class remains at each leaf node (Watts & Lawrence 2008). The best branch is determined for each node, and many trees are produced depending on how many trees are desired to be produced (Liaw & Wiener 2002). According to the division criteria determined by using the training data, the nodes are divided into branches, and tree structures are formed (Figure 3).



**Figure 3- Structure of random forest**

The tree with the best performance among the determined trees is assigned to a class (Liaw & Wiener 2002). The RF method has no fixed model, constraint, or pattern. It functions with as many trees as the user wants and is rapid. A RF algorithm is determined using the R package software. In the R Core software, “randomforest (Liaw & Wiener, 2002)”, “caret (Kuhn et al. 2020)”, and “mice (Van Buuren & Groothuis Oudshoorn 2011)” packages are used. To make the best estimation while creating the model, the tuning process was performed, and  $mtry$  was determined as 2,  $ntree = 50$  for 0-20 cm, and  $ntree = 60$  for 20-40 cm. Furthermore, 70% of the data set was evaluated as training and 30% as the test set. Also, the distribution of soil properties was checked using the Kolmogorov-Smirnov test. The “soil texture” package is used in the texture triangle created in the R software.

#### 2.4. Interpolation models

In the present study, various interpolation [kriging, inverse distance weighting (IDW), and radial base function (RBF)] methods were applied to determine the most suitable model for the creation of spatial distribution maps of LLWR. Kriging, one of the scholastic approaches, uses a linear combination of weights at known points to estimate the value at an unknown point (Oliver & Webster 2015). A semivariogram, a measure of the spatial correlation between two points, is generated. There are several kriging interpolation methods, including simple kriging (SK), ordinary kriging (OK), and universal kriging (UK). Prior to the geostatistical estimation, a variogram was calculated for the distance classes between sample pairs.

SK is based on the logic of trying to estimate the value of a variable at any unknown point, using the values of the known points, similar to other estimation models in general (Li & Heap 2008).

OK (OK makes a calculation very similar to SK, but only OK changes the  $\mu$  parameter in the  $\mu(x_0)$  general equation  $[1 - \sum_{i=1}^n \lambda_i] = 0$  to  $\sum_{i=1}^n \lambda_i = 1$  instead of the value in the general equation while using the local average (Li & Heap 2008).

In the UK, the OK method cannot be used in case the variable values increase continuously depending on the increasing distance in a certain direction in the study area or space. In case the variable values do not increase continuously at a certain distance, the trends are removed using residual semivariograms, and estimates are made as a result of kriging (Christensen 1990; Brus & Heuvelink 2007).

IDW and RBF models, which are deterministic methods, were also applied in the study.

IDW is one of the most widely used multivariate interpolation methods. IDW is based on estimating the weighted average values of the value of the unknown point from the known point while using the inverse distance functions of the distances. There is a logic that as the distance from the point known to the target point increases, the similarities decrease (Li & Heap 2008).

RBF is currently a method used in the interpolation of multidimensional data. It is generally used for estimating limited data or hard-to-guess area points. The biggest advantage of this method is that it can be easily used in any size due to the low general restrictions (Wright 2003).

In the present study, completely regularized spline, thin plate spline, and spline with tension methods in RBF were evaluated.

ArcGIS 10.5v program was utilized in the creation and evaluation of scattering maps.

#### 2.4. Assessment of the selected models

In the present study, MAE, root means square error (RMSE), and mean absolute percentage error (MAPE) parameters were used to examine the relationships between the predicted and observed values with different interpolation techniques and RF algorithm. Estimates were determined using the following formulas (Equations 3, 4, 5).

$$MAE = \frac{1}{n} \sum_{i=1}^n |Z_i - Z| \quad (\text{Equation 3})$$

$$RMSE = \sqrt{\frac{\sum (Z_i - Z)^2}{n}} \quad (\text{Equation 4})$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Z_i - Z}{Z} \right| * 100 \quad (\text{Equation 5})$$

In the models,  $Z_i$  is the estimation value,  $Z$  is the observed value, and  $n$  is the number of observations.

Also, soil characteristics were calculated in the descriptive statistics of the analysis. The present study used the IBM SPSS 23 software to calculate the values such as minimum, maximum, mean, standard deviation, coefficient of variation, skewness, and kurtosis of the parameters as descriptive statistics. The flow chart of the study is given in Figure 4.

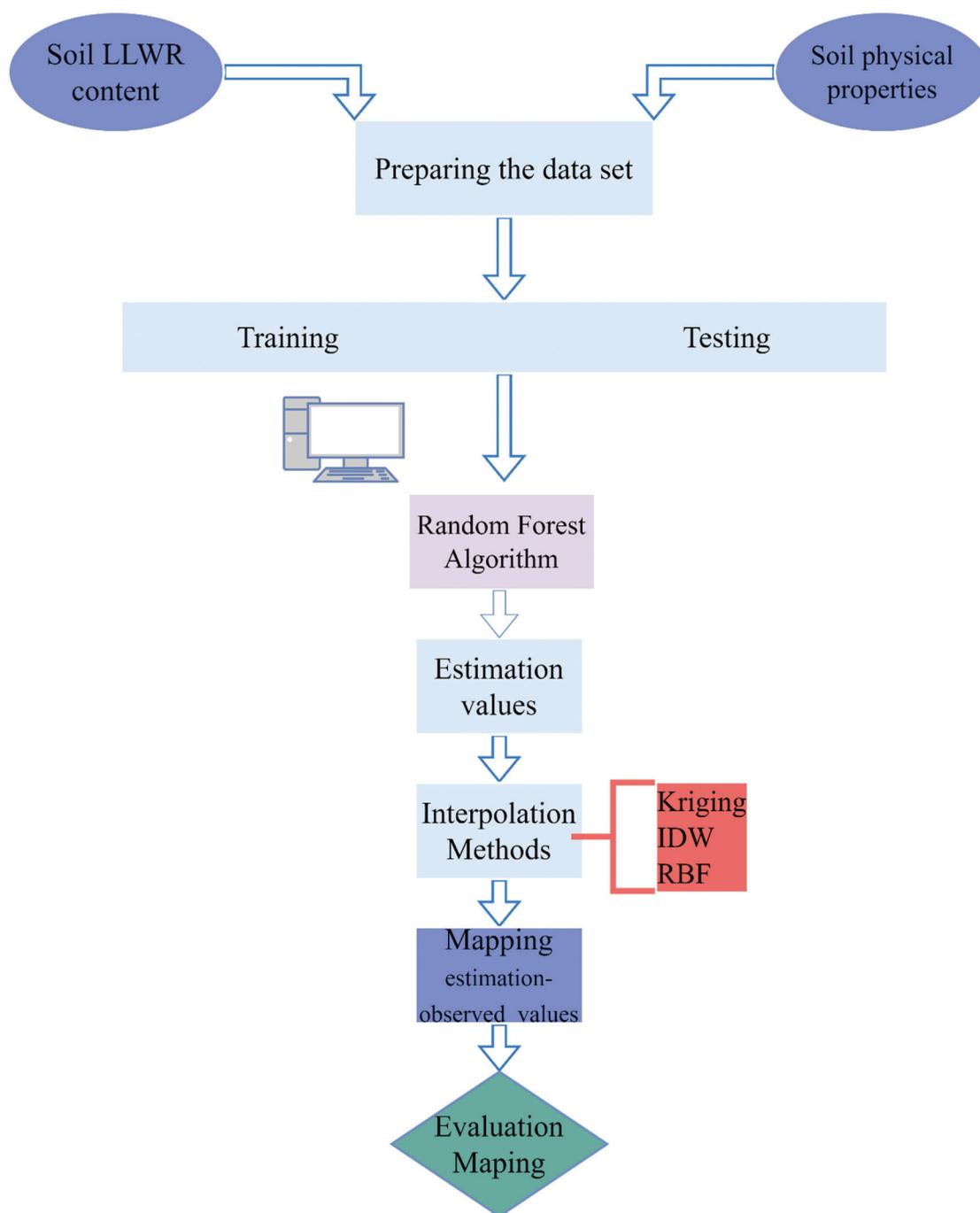


Figure 4- Flow chart

### 3. Results and Discussion

#### 3.1. Soil physical properties and LLWR

Descriptive statistics for soils of different depths (0-20 cm and 20-40 cm) are given in Table 1. Since the study area is an alluvial land, it shows a significant change, especially in sand and clay distribution rates. Dengiz (2010) has stated that there are significant changes in particle size distribution over short distances in soils formed on sediments transported by rivers. The soils' sand, silt, and clay contents are determined in the ranges of 8.86-78.46%, 11.90-55.79%, and 6.93-63.90%, respectively. On the other hand, the texture class ranged from sandy loam to clay (Figure 5). Medium (loam 31.0%), medium-fine (clay loam 38%), and fine texture (23.52%) are dominant in the study area soils.

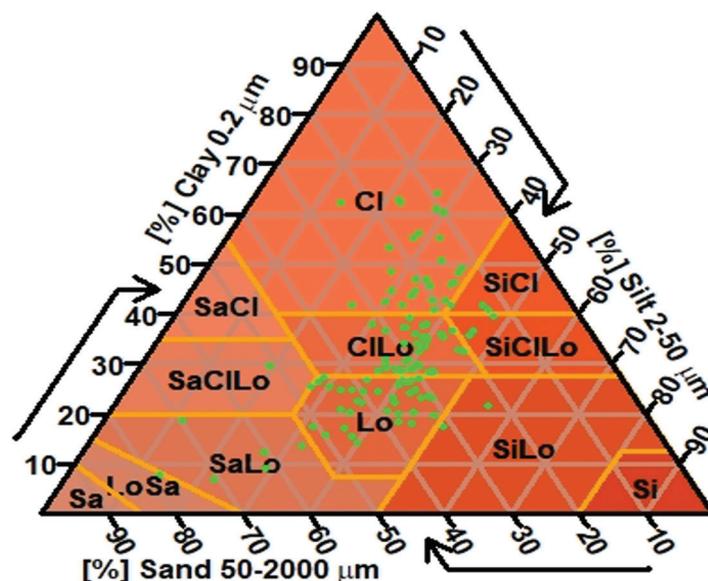


Figure 5- Selected soils in soil texture triangle  
Sa: Sand, Lo: Loam, Si: Silt, Cl: Clay

The moisture content where the air-filled pore volume is 10% ( $\theta_{Ap}$ ) was 0.330-0.444  $\text{cm}^3 \text{cm}^{-3}$  for surface soils, and it was determined in the range of 0.265-0.435  $\text{cm}^3 \text{cm}^{-3}$  at 20-40 cm depth. Depending on the depth, the increase in compaction leads to a decrease in the air-filled pore volume. The bulk density values were determined in the range of 1.20-1.56  $\text{g cm}^{-3}$ , and the penetration resistance was determined in the range of 0.26-4.64 MPa in the surface soil. These properties were 1.24-1.68  $\text{g cm}^{-3}$  and 0.51-4.40 MPa at 20-40 cm, respectively. As seen from the soil's bulk density and penetration resistance values, depth-dependent compaction was determined. This is thought to be caused by the heavy field traffic applied, especially in the area. Munsuz (1985) has reported that the compression on the surface causes an increase in penetration resistance and bulk density and states that the volume of air-filled pores decreased from 17.3% to 7.2% with the increase in bulk density. The soil's field capacity and wilting point contents were found to be 0.313 and 0.182  $\text{cm}^3 \text{cm}^{-3}$  on average. It is well known that moisture constants vary depending on texture, organic matter, and structure. Also, the moisture content in the field capacity is significantly affected by the change in the pore volume due to compression (Karahan et al. 2014).

Table 1- Descriptive statistic of soil properties

Properties	Min	Max	Mean	SD	CV	Skewness	Kurtosis
0-20 cm							
Sand %	8.86	78.46	30.60	12.83	41.98	1.09	1.93
Clay %	6.93	63.98	32.60	12.57	38.55	0.52	-0.02
Silt %	11.90	55.79	37.50	7.61	20.61	-0.80	1.27
$\theta_{Ap}$ $\text{cm}^3 \text{cm}^{-3}$	0.330	0.444	0.372	2.90	6.24	-0.07	1.79
$\theta_{FC}$ $\text{cm}^3 \text{cm}^{-3}$	0.133	0.455	0.332	6.64	18.82	-0.53	0.42
$\theta_{WP}$ $\text{cm}^3 \text{cm}^{-3}$	0.058	0.347	0.198	5.82	30.25	0.36	-0.13
$\theta_{PR}$ $\text{cm}^3 \text{cm}^{-3}$	0.037	0.361	0.124	7.01	53.96	0.98	0.46
LLWR $\text{cm}^3 \text{cm}^{-3}$	0.049	0.263	0.134	3.87	25.73	-0.35	5.12
BD $\text{g cm}^{-3}$	1.20	1.56	1.40	0.07	4.82	-0.69	0.25
PR MPa	0.26	4.64	1.27	0.88	40.35	1.51	2.46
20-40 cm							
Sand %	9.51	74.08	30.70	13.87	45.24	0.91	0.86
Clay %	6.87	72.26	31.30	13.19	42.14	0.52	-0.06
Silt %	8.34	63.18	38.20	9.68	25.47	-0.58	0.80
$\theta_{Ap}$ $\text{cm}^3 \text{cm}^{-3}$	0.265	0.435	0.344	3.77	8.14	-1.72	7.59
$\theta_{FC}$ $\text{cm}^3 \text{cm}^{-3}$	0.137	0.452	0.313	7.03	20.11	-0.52	0.10

**Table 1. Continued**

$\theta_{WP}$ cm <sup>3</sup> cm <sup>-3</sup>	0.053	0.348	0.182	6.53	34.66	0.20	-0.57
$\theta_{PR}$ cm <sup>3</sup> cm <sup>-3</sup>	0.024	0.380	0.160	6.77	41.12	0.62	0.72
LLWR cm <sup>3</sup> cm <sup>-3</sup>	0.048	0.279	0.129	4.01	28.58	-0.18	2.51
BD g cm <sup>-3</sup>	1.24	1.68	1.51	0.07	4.89	-0.62	-0.44
PR MPa	0.51	4.40	1.76	0.88	42.34	1.04	0.49

Min: Minimum, Max: Maximum, SD: Standard deviation, CV: Coefficient of variance,  $\theta_{Ap}$ : Moisture content at 10% air-filled pore volume,  $\theta_{FC}$ : Field capacity moisture content,  $\theta_{WP}$ : Wilting point moisture content,  $\theta_{PR}$ : Moisture content at 2MPa penetration resistance; LLWR: Least limiting water range, BD: Bulk density, PR: Penetration resistance

In the present study, the LLWR contents of the soils varied in the range of 0.048-0.279 cm<sup>3</sup> cm<sup>-3</sup>. There were aeration problems in 6.72%, compaction problems in 20.16%, and aeration and compaction problems in 0.8% of the surface soils in the study area, whereas 72.32% were determined under optimum conditions. In 20-40 cm depth, aeration problems were detected in 5.88%, compaction problems in 28.57%, and aeration and compaction problems in 2.52%. According to Kay & Anger (2002), LLWR values are classified in the range of “less than” and “good”. The LLWR contents are classified as “less than” if <0.1 cm<sup>3</sup> cm<sup>-3</sup>, “poor” if 0.1-0.15 cm<sup>3</sup> cm<sup>-3</sup>, “moderate” if 0.15-0.2 cm<sup>3</sup> cm<sup>-3</sup>, “good” if >0.2 cm<sup>3</sup> cm<sup>-3</sup>. Approximately 9.2% of the LLWR contents of the surface soils were <0.1 cm<sup>3</sup> cm<sup>-3</sup>, and 61% of the soils were determined in the range of 0.1-0.15 cm<sup>3</sup> cm<sup>-3</sup>. Of the subsurface soils, 16.8% were determined to be “less than” and 58% as “poor”. The LLWR contents of >0.02 cm<sup>3</sup> cm<sup>-3</sup> in surface soils constituted 6.72% while this value decreased to 3.3% at 20-40 cm subsurface depth. In light of the obtained data, one of the significant results was that the amount of water the plant could use under the surface decreased. It is also known that the range of LLWR narrows with the change in pore volume with soil compaction (Kahlon & Chawla 2017). The LLWR contents of soils can exhibit high variability due to textural fraction ratios, compaction, and aeration problems (Alaboz et al. 2021). Negiş et al. (2020), on the other hand, determined significant increases in the lower and upper limits of LLWR with the increase in the doses of organic materials. Also, Silva and Kay (1997) have stated that LLWR exhibited negative correlations with the increase in bulk density and clay content, whereas positive with the organic matter content.

The coefficient of variance (CV) is an important factor in determining the variability of soil properties in the data set. Wilding (1985) classified the CV value as ≤15%, 15-30%, and ≥30% as low, medium, and high variability, respectively. Among the properties examined in the study,  $\theta_{Ap}$  and BD properties were low, sand, clay,  $\theta_{WP}$  and PR were high, and other properties showed moderate variability. The skewness and kurtosis coefficients being close to 0 indicates a normal distribution. The feature closest to the normal distribution was LLWR, whereas the features farthest from the normal distribution were determined to be  $\theta_{Ap}$  and PR. A negative skewness coefficient indicates skewness to the left and a positive skewness to the right. In contrast, a negative kurtosis coefficient suggests that the curve is flatter than normal and the positivity is steeper.

### 3.2. Estimation of LLWR with RF

LLWR estimates were carried out using the RF algorithm with the soils' sand, silt, clay, and bulk density values. To make the most appropriate estimation in the model estimation, mtry 2 was selected due to the tuning process. According to Breiman (2001), it was determined that optimum results were obtained when the mtry variable value was chosen as the square root of the total number of variables. Figure 6 shows the ntree (number of trees used) and error rates applied depending on the depths. The error must be stable and at the lowest level to make the most accurate estimation. The lowest error rate was determined with ntree 50 for surface soils and ntree 60 for 20-40 cm depth. In RF, the number of trees (ntree) and mtry are parameters that are often modified to regulate the complexity of the models. mtry indicates the number of randomly sampled indicators as candidates in each compartment. What is sampled for the split at each node is the number of mtry estimators.

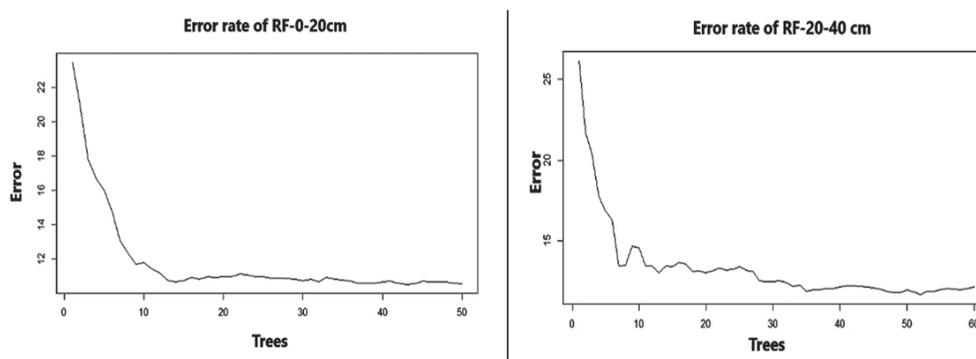


Figure 6- Error rate of trees

The RF model estimates the importance of covariates depending on how good, or bad the prediction will be when one or more variables are removed. It also reveals the errors that may occur by eliminating the good ones (Prasad et al. 2006). The variable importance of the RF algorithm is indicated in Figure 7. Sand and clay were determined as the best predictors for both depths. For 0-20 cm, the order of importance of the features is sand > clay > silt > BD. The feature with the highest error when it was removed from the model was determined as sand. If this feature is not included in the model, approximately an error of 4.5% occurs. The narrow range of variation of the bulk density feature compared to other features is also evident from the CV values. The effect of this narrow range of variation was lower among traits with high variability.

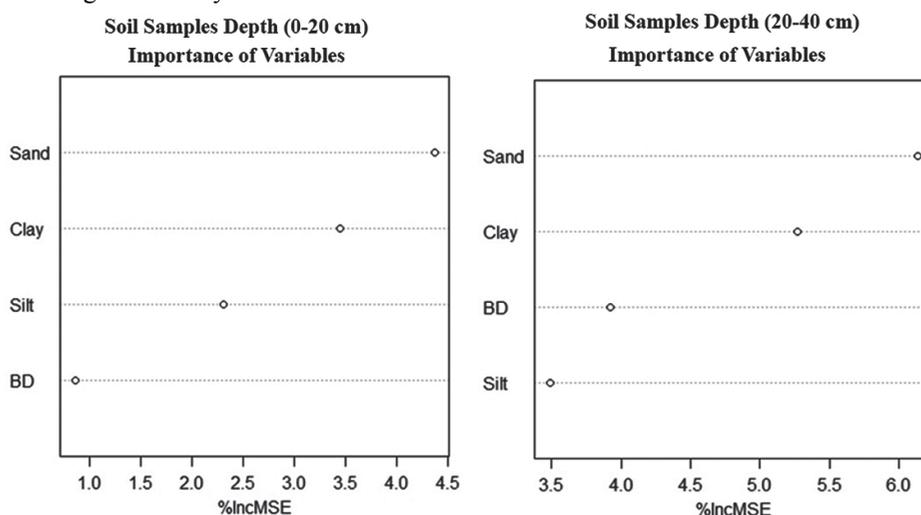


Figure 7- Importance of variables

The order of importance in estimating the LLWR for a soil depth of 20-40 cm is sand > clay > BD > silt. The property with the highest error when removed from the model was determined as sand. If this feature is not included in the model, an error of approximately 6% occurs. It was determined that the increase in compaction in subsurface soils has a defect in the estimation of LLWR. At 20-40 cm soil depth, an error of about 4% is expected as the BD moves away from the model. This value was about 1% in the surface soil, whereas the error showed a four-fold increase at 20-40 cm depth. It has been reported that there were significant negative correlations between bulk density and LLWR. Also, it has been reported that LLWR narrows with increasing bulk density (Haghighi Fashi et al. 2017; Alaboz et al. 2021). Increasing soil compaction depending on the bulk density can increase soil water retention both at field capacity and at wilting point, indicating that it provides higher water retention in the soil with the increase in medium and micro-sized pores due to the reduction of macro-sized pores (Safadoust et al. 2014). However, it is assumed that root development will be inhibited due to the existing compaction of the plant before the soil water content reaches the wilting point. Sand and clay were found to be important in the estimation of LLWR for both depths. Alaboz et al. (2021) determined a statistically significant positive relationship between the clay content of soils and LLWR ( $r=0.30$ ). The high variability in the sand content of the soils can be understood from the CV value. The highest CV value was determined in the sand among the textural fractions. Therefore, the contribution rate of this feature to the model was found to be high.

### 3.3 Assessment of models' performance for LLWR estimation

The model estimation performance values obtained in the testing and training phase of the LLWR estimation with the RF algorithm using the sand, silt, clay, and bulk density values of the soils are given in Table 2.

**Table 2- Performance assessment of RF model**

Depth	Training			Testing		
	RMSE ( $\text{cm}^3 \text{cm}^{-3}$ )	MAPE (%)	MAE ( $\text{cm}^3 \text{cm}^{-3}$ )	RMSE ( $\text{cm}^3 \text{cm}^{-3}$ )	MAPE (%)	MAE ( $\text{cm}^3 \text{cm}^{-3}$ )
0-20 cm	0.0223	13.645	0.0158	0.0218	13.28	0.0190
20-40 cm	0.0203	12.634	0.0154	0.0247	8.45	0.0167

RMSE: Root mean square error, MAPE: Mean absolute percentage error, MAE: Mean absolute error

Similar error rates were determined during the training and testing phases. This shows that the learning performance of the model is high. The RMSE values obtained in the evaluation of the predictive accuracy of the model were determined in the range of 0.0203-0.0247  $\text{cm}^3 \text{cm}^{-3}$ , and the MAPE values were found to be 8.45-12.645%. Lewis (1982) classified models with a MAPE value of less than 10% are "very good", models between 10-20% are "good", models between 20-50% are "acceptable", and models above 50% are "wrong and faulty". According to the classifications, it was determined that the model's performance was good. In the test phase, the MAE was found to be 0.0190  $\text{cm}^3 \text{cm}^{-3}$  at 0-20 cm depth and 0.0167  $\text{cm}^3 \text{cm}^{-3}$  at 20-40 cm depth. Low RMSE, MAPE, and MAE values are desirable for the model's validity in model studies. The more data trained in machine learning algorithms, the higher the probability the model predicts. Alaboz et al. (2021) achieved the best performance with the deep learning algorithm in their studies investigating the predictability of LLWR with artificial neural networks and deep learning algorithms. In another study (Tavanti et al. 2019), LLWR was evaluated with pedotransfer functions, models obtained from the literature, and artificial neural networks were examined. ANN determined the lowest RMSE value to be 0.0142  $\text{m}^3 \text{m}$ . Akar and Güngör (2013) stated that the RF algorithm exhibits higher accuracy than other approaches. In obtaining successful results, a high tree depth is considered as running the model with many trees in the background.

### 3.4. Spatial variation of LLWR

Spatial distribution maps of LLWR's observed and predicted values were created according to the most appropriate model using different interpolation models. The data conformity to the normal distribution was checked with the Kolmogorov-Smirnov test, and logarithmic transformation was applied to both data. The RMSE values of the model parameters created for the observed and predicted LLWR values are given in Table 3. In the surface soil (0-20 cm), SK's spherical model was determined as the most suitable model for the distribution of observed values. In contrast, SK's Gaussian model was determined for the distribution of predicted values. Also, it has been determined that OK's Gaussian model is the most suitable model for the spatial distributions of observed and predicted values in subsurface (20-40 cm) soils.

**Table 3- Cross-validation and their RMSE values according to different interpolation models**

Criteria	I	Inverse distance weighing			Radial basis function					
		2	3	TPS	CRS		ST			
LLWR (0-20 cm)	Observe	0.0252	0.0250	0.0249	0.0277	0.0250	0.0250			
	Estimate	0.0180	0.0181	0.0183	0.0224	0.0186	0.0184			
LLWR (20-40 cm)	Observe	0.0275	0.0280	0.0286	0.0341	0.0283	0.0281			
	Estimate	0.0244	0.0250	0.0257	0.0315	0.0255	0.0252			
Criteria	Ordinary	Kriging				Universal				
		Gau.	Exp.	Sph.	Gau.	Exp.	Sph.	Gau.	Exp.	Sph.
LLWR (0-20 cm)	Observe	0.0254	0.0253	0.0253	0.0237	0.0238	0.0236	0.0254	0.0253	0.0253
	Estimate	0.0180	0.0181	0.0181	0.0178	0.0179	0.0179	0.0180	0.0181	0.0180
LLWR (20-40 cm)	Observe	0.0271	0.0277	0.0276	0.0273	0.0272	0.0272	0.0276	0.0277	0.0276
	Estimate	0.0237	0.0240	0.0239	0.0239	0.0238	0.0238	0.0238	0.0240	0.0239

RMSE: Root mean square error, Gau.: Gaussian, Exp.: Exponential, Sph.: Spherical, TPS: Thin plate spline, CRS: Completely regularized spline, ST: Spline with tension, LLWR: Least limiting water range

The spatial distribution patterns of the values of both observed (LLWR-O) and estimated (LLWR-E) in surface soils and subsurface soils in the study area, located on alluvial land with different soil properties, showed closeness to each other (Figure 8). In general, low LLWR values in surface (0-20 cm) soils were determined in Ay1, Ad4y, Ct3, Ad3i, and Kz4, which are in the Gold leaf (Calci Hapluster), Çetinkaya (Typic Ustipssament), and Kızılırmak (Typic Ustifluent) soil series. Low LLWR values were determined in the Ad1a mapping units, while low LLWR values were determined in similar series in the distribution map of the values estimated by RF. Furthermore, examining the LLWR distribution of subsurface soils, which is very important for plant root development, it was seen that low LLWR values are concentrated in the Ct3, Ad3i, and Kz4. Ad1a mapping units located in the Çetinkaya and Kızılırmak soil series and the northeast of the study area. Although a similar case for subsurface (20-40 cm) soils was also seen in the LLWR estimated by RF, the areas with low estimated LLWR values were mostly distributed in the Kızılırmak soil series, which is classified as Typic Ustipssament.

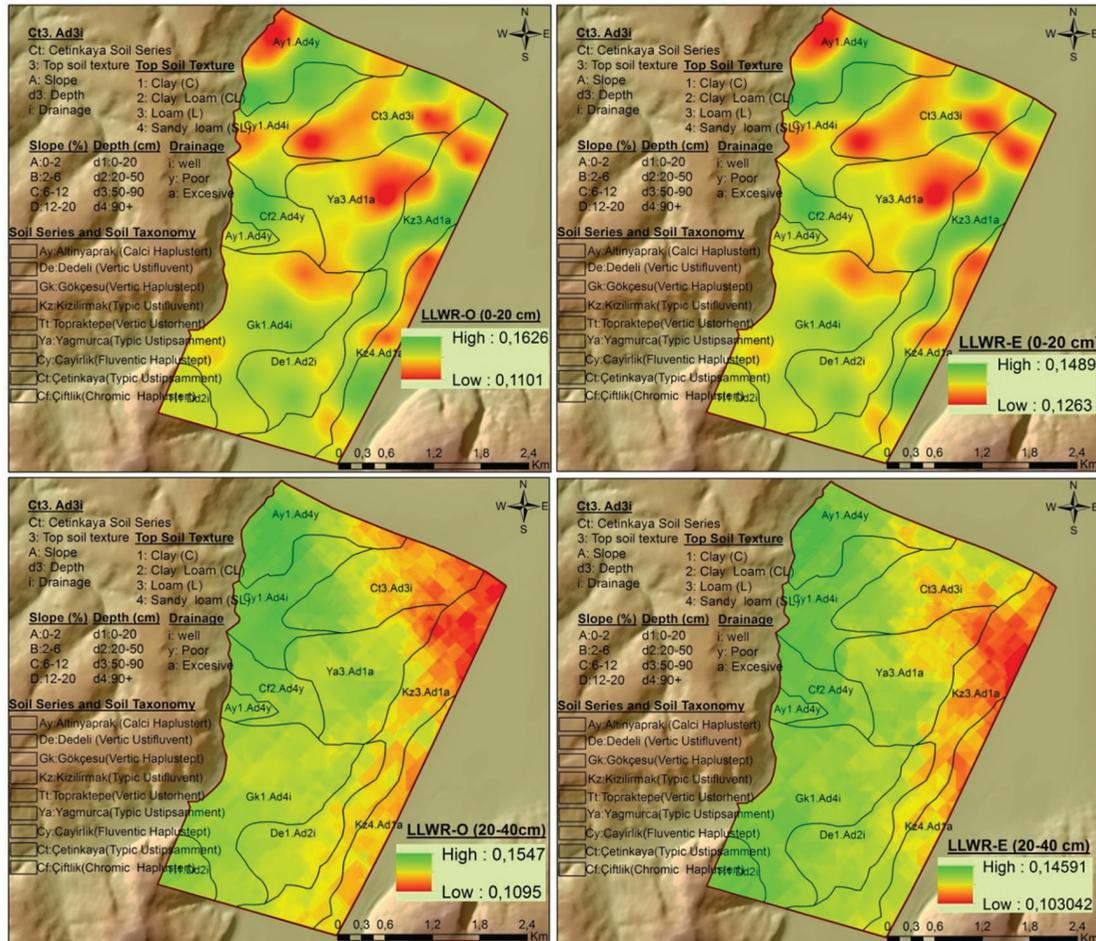


Figure 8- Spatial distribution maps of the LLWR (observed: LLWR-O and estimated: LLWR E for surface and sub-surface soil depths)

#### 4. Conclusions

In the present study, the LLWR contents of the soils distributed on the alluvial lands in the Bafra Delta Plain were evaluated for two different soil depths, and the predictability of LLWR with the RF algorithm, one of the machine learning methods, was investigated. Also, the observed and predicted values of the studied feature were evaluated with different interpolation methods.

As a result, the LLWR contents of the soils were determined in the range of  $0.049\text{-}0.279\text{ cm}^3\text{ cm}^{-3}$ . Aeration problems were determined in 6.72% of the surface soils, compaction in 20.16%, and both aeration and compaction problems in 0.8%. In 20-40 cm depth, aeration problems were detected in 5.88%, compaction problems in 28.57%, and aeration and compaction problems in 2.52%. Soil properties that are effective in estimating LLWR with RF were determined as sand and clay. The importance of BD in the model has increased with the increase of depth-dependent PR. It was shown that the estimation of LLWR from sand, silt, clay, and BD with the RF algorithm could be carried out with high accuracy by training the dataset.

Alluvial soils have unstable properties due to different geological processes. The use of the RF model, which effectively solves complex structures in estimating these properties, was successfully demonstrated. Also, spatial distribution maps were successfully created in alluvial soils using the LLWR estimated values obtained by RF. As a result, it was revealed that there are aeration and compression problems on the surface and subsurface soils in the study area. Also, it was found that spatial distribution maps can be created for the region by successfully estimating LLWR utilizing the RF algorithm. For future studies, it is recommended that spatial distribution maps be updated at regular intervals for sustainable land management, focusing on the possibilities of determining based on LLWR value using the RF algorithm, especially in areas where land traffic will be intense.

**Data availability:** Data are available on request due to privacy or other restrictions.

**Authorship Contributions:** Concept: P.A., Design: P.A., O.D., Data Collection or Processing: O.D., Analysis or Interpretation: P.A., O.D., Literature Search: P.A., O.D., Writing: P.A., O.D.

**Conflict of Interest:** No conflict of interest was declared by the authors.

**Financial Disclosure:** The authors declared that this study received no financial support.

## References

- Akar Ö & Güngör Ö (2013). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation* 1(2): 105-112. doi.org/10.9733/jgg.241212.1
- Aksakal EL (2004). Soil compaction and its importance for agriculture. *Atatürk University Journal of Agricultural Faculty* 35(3-4): 247-252.
- Alaboz P, Başkan O & Dengiz O (2021). Computational intelligence applied to the least limiting water range to estimate soil water content using GIS and geostatistical approaches in alluvial lands. *Irrigation and Drainage* 70(5): 1129-1144. doi.org/10.1002/ird.2628
- Alaboz P, Demir S & Dengiz O (2020). Determination of Spatial Distribution of Soil Moisture Constant Using Different Interpolation Model Case study, Isparta Atabey Plain. *Journal of Tekirdag Agricultural Faculty* 17(3): 432-444. doi.org/10.33462/jotaf.710411
- Blake G R & Hartge K H (1986). Bulk density. In: Klute A, Ed., *Methods of Soil Analysis, Part 1—Physical and Mineralogical Methods*, 2<sup>nd</sup> Edition, Agronomy Monograph 9, American Society of Agronomy—Soil Science Society of America, Madison, 363-382.
- Baillie I C (2001). Soil survey staff 1999, soil taxonomy: a basic system of soil classification for making and interpreting soil surveys, agricultural handbook 436, Natural Resources Conservation Service, USDA, Washington DC, USA, pp. 869.
- Breiman L (2001). Random forests. *Machine Learning* 45(1): 5-32.
- Brus D J & Heuvelink G B (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138(1-2): 86-95. doi.org/10.1016/j.geoderma.2006.10.016
- Busscher W J (1990). Adjustment of flat-tipped penetrometer resistance data to a common water content. *American Society of Agricultural and Biological Engineers* 33(2): 519-524.
- Chan K, Oates A, Swan A, Hayes R, Dear B & Peoples M (2006). Agronomic consequences of tractor wheel compaction on a clay soil. *Soil and Tillage Research* 89(1): 13-21. doi.org/10.1016/j.still.2005.06.007
- Christensen R (1990). The equivalence of predictions from universal kriging and intrinsic random-function kriging. *Mathematical Geology* 22(6): 655-664. doi.org/10.1007/bf00890514
- Da Silva A, Kay B & Perfect E (1994). Characterization of the least limiting water range of soils. *Soil Science Society of America Journal* 58(6): 1775-1781. doi.org/10.2136/sssaj1994.03615995005800060028x
- Da Silva A P & Kay B (1997). Estimating the least limiting water range of soils from properties and management. *Soil Science Society of America Journal* 61(3): 877-883. doi.org/10.2136/sssaj1997.03615995006100030023x
- Dengiz O (2010). Morphology, Physico-Chemical Properties and Classification of Soils on Terraces of the Tigris River in the South-East Anatolia Region of Turkey. *Journal of Agricultural Sciences* 16(3): 205-212. doi.org/10.1501/tarimbil\_0000001139
- Eijkelkamp (1990). Equipment for soil research. Giesbeek (The Netherlands): Eijkelkamp Corporation.
- Gee G W & Bauder J W (1986). Methods of soil analysis: physical and mineralogical analysis. Madison (WI): *American Society of Agronomy Particle-size analysis* 388-409.
- Haghighi Fashi F, Gorji M & Sharifi F (2017). Least limiting water range for different soil management practices in dryland farming in Iran. *Archives of Agronomy and Soil Science* 63(13): 1814-1822. doi.org/10.1080/03650340.2017.1308688
- Kahlon M S & Chawla K (2017). Effect of tillage practices on least limiting water range in Northwest India. *International Agrophysics* 31(2): 183-194. doi.org/10.1515/intag-2016-0051
- Karahan G, Erşahin S & Öztürk H S (2014). Field Capacity Dynamics Affected by Soil Properties. *Journal of Agricultural Faculty of Gaziosmanpaşa University* 30(1): 1-9. doi.org/10.13002/jafag177
- Kay B D & Anger D A (2002). Soil structure in soil physics companion 249-296.

- Klute A (1986). Water Retention: Laboratory Methods. Methods of soil analysis: part 1 physical and mineralogical methods. 5: 635-662. doi.org/10.2136/sssabookser5.1.2ed.c26
- Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, & Team R C (2020). Package 'caret'. *The R Journal* 23(7).
- Leão T P, Da Silva A P, Perfect E & Tormena C A (2005). An algorithm for calculating the least limiting water range of soils. *Agronomy Journal* 97(4): 1210-1215. doi.org/10.2134/agronj2004.0229
- Letey J (1958). Relationship between soil physical properties and crop production. *Advances in soil science* 1: 277-294.
- Lewis C D (1982). Industrial and Business Forecasting Methods. Butterworths Publishing.
- Li J & Heap A D (2008). A review of spatial interpolation methods for environmental scientists. *Geoscience Australia* 137-145.
- Liaw A & Wiener M (2002). Classification and Regression by Random Forest, *R News* 2-3: 18-22.
- Machado G, Vilalta C, Recamonde-Mendoza M, Corzo C, Torremorell M, Perez A & VanderWaal K (2019). Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. *Scientific Reports* 9: 1-12. doi.org/10.1038/s41598-018-36934-8
- Mihalikova M, Özyazıcı M A & Dengiz O (2016). Mapping soil water retention on agricultural lands in central and eastern parts of the Black Sea Region in Turkey. *Journal of Irrigation and Drainage Engineering* 142(12): 05016008-1. doi.org/10.1061/(asce)ir.1943-4774.0001094
- Munsuz N (1985). Soil Mechanics and Technology. *Ankara University. Faculty of Agriculture Publications*
- Negiş H, Şeker C & Çetin A. (2020). Effects of different organic materials on soil compaction and least limiting water range. *Journal of Soil Science and Plant Nutrition* 8(2):118-127. doi.org/10.33409/tbbbd.778834
- Oliver M A & Webster R (2015). Basic steps in geostatistics: the variogram and kriging. Springer International Publishing.
- Pal M (2005). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing* 26(1): 217-222. doi.org/10.1080/01431160412331269698
- Prasad A M, Iverson L R & Liaw A (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9: 181-199. doi.org/10.1007/s10021-005-0054-1
- Safadoust A, Feizee P, Mahboubi A, Gharabaghi B, Mosaddeghi M & Ahrens B (2014). Least limiting water range as affected by soil texture and cropping system. *Agricultural Water Management* 136: 34-41. doi.org/10.1016/j.agwat.2014.01.007
- Soil Survey Staff (2014). Keys to Soil Taxonomy. *USDA National Resources Conservation Services*.
- Stum A K, Boettinger J L, White MA & Ramsey R D (2010). Random forests applied as a soil spatial predictive model in arid Utah. *Digital Soil Mapping* 179-190.
- Şenol H, Alaboz P, Demir S & Dengiz O (2020). Computational intelligence applied to soil quality index using GIS and geostatistical approaches in semiarid ecosystem. *Arabian Journal of Geosciences* 13(23): 1-20. doi.org/10.1007/s12517-020-06214-9
- Tavanti R F, Freddi O D S, Tavanti T R, Rigotti A & Magalhães W D A (2019). Pedofunctions applied to the least limiting water range to estimate soil water content at specific potentials. *Engenharia Agrícola* 39(4): 444-456. doi.org/10.1590/1809-4430-Eng.Agric.v39n4p444-456/2019
- TSMS (2021). Turkish State Meteorological Service. Ankara, Turkey.
- Tunçay T, Başkan O, Bayramin İ, Dengiz O & Kılıç Ş (2018). Geostatistical approach as a tool for estimation of field capacity and permanent wilting point in semiarid terrestrial ecosystem. *Archives of Agronomy and Soil Science* 64(9): 1240-1253. doi.org/10.1080/03650340.2017.1422081
- Watts J D & Lawrence R L (2008). Merging random forest classification with an object-oriented approach for analysis of agricultural lands, *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* 37(B7): 579-582.
- Wilding L P (1985). Spatial variability: Its documentation, accommodation and implication to soil surveys. In: *Soil spatial Variability*, Las Vegas NV, 30 November-1 December 1984. 1985. p. 166-194.
- WRB (2014). World reference base for soil resources. International soil classification system for naming soils and creating legends for soil maps. *Food and Agriculture Organization of United Nations* 6: 203.
- Wright G B (2003). Radial basis function interpolation: numerical and analytical developments. University of Colorado at Boulder.
- Wu L, Feng G, Letey J, Ferguson L, Mitchell J, Mc Cullough-Sanden B & Markegard G (2003). Soil management effects on the nonlimiting water range. *Geoderma* 114(3-4): 401-414. doi.org/10.1016/S0016-7061(03)00052-1
- Van Buuren S & Groothuis-Oudshoorn K (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3): 1-67.

