ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Ferhat ALBAYRAK

SHORTENING THE TEST DURATION IN LGB BASED SOIL TEXTURE ANALYSIS

DEPARTMENT OF COMPUTER ENGINEERING

ADANA-2021

ABSTRACT

MASTER THESIS

SHORTENING THE TEST DURATION IN LGB BASED SOIL TEXTURE ANALYSIS

Ferhat ALBAYRAK

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor	: Assoc. Prof. Dr. Umut ORHAN
	Year: 2021, Pages: 59
Juries	: Assoc. Prof. Dr. Umut ORHAN
	: Assoc. Prof. Dr. Ali İNAN
	: Asst. Prof. Dr. Serkan KARTAL

In this thesis, an approach based on curve fitting, support vector regression, multilayer perceptron and long short-term memory architecture is proposed to shorten the experiment time in experiments with the Laser Guided Bouyoucos device developed for soil texture analysis. For this purpose, texture analysis signals obtained from 52 soil samples, each with 14400 samples (2 hours), were used. In the traditionally used curve fitting method, the shortest signal segment is found with an acceptable absolute error by shortening the soil texture signals from the end, while in machine learning methods, the shortest signal segment is found starting from the beginning. In the curve fitting method, the most suitable curve for the soil signals was selected as the 2nd degree exponential equation with the R-squared method. In the time shortening study with SVR, the model was trained and tested in the sample range of 1000-7000. In order to determine the entrance segment length in the MLP method, tests were carried out with a segment size in the range of 50-950 and the entrance segment size was selected as 200. In the MLP method, 3 layers are used: 1 input layer with 200 inputs, 1 hidden layer with 100 neurons, and 1 output layer with output. In the LSTM method, a 3-layer architecture is used, including the 1-input input layer, the 200-neuron LSTM layer, and the single-output output layer. As a result of the time shortening studies, the soil components were estimated with the SVR.

Keywords: Time Series, Curve Fitting, Support Vector Regression, Multi Layer Perceptron, Long Short-Term Memory

ÖZ

YÜKSEK LİSANS TEZİ

LGB TABANLI TOPRAK TEKSTÜR ANALİZLERİNDE TEST SÜRESİNİN KISALTILMASI

Ferhat ALBAYRAK

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Doç. Dr. Umut ORHAN
Yıl: 2021, Sayfa: 59
Jüri : Doç. Dr. Umut ORHAN
: Doç. Dr. Ali İNAN
: Dr. Öğr. Üyesi Serkan KARTAL

Bu tezde, toprak tekstür analizi için geliştirilmiş olan Lazer Güdümlü Bouyoucos cihazı ile yapılan deneylerde deney süresi kısaltması için eğri uydurma, destek vector regresyonu, çok katmanlı algılayıcı ve uzun kısa süreli bellek mimarisine dayalı bir yaklaşım önerilmiştir. Bu amaçla, her biri 14400 örnekli (2 saat) 52 toprak örneğinden elde edilen tekstür analizi sinyalleri kullanılmıştır. Geleneksel olarak kullanılan eğri uydurma yönteminde toprak tekstür sinyalleri sondan kısaltarak en kısa sinyal parçasını kabul edilebilir bir mutlak hata ile bulunurken, makine öğrenmesi yöntemlerinde en kısa sinyal parçası başlangıçtan başlayarak bulunmuştur. Eğri uydurma yönteminde toprak sinyallerine en uygun eğri R-kare metodu ile 2.dereceden eksponansiyel denklem olarak seçilmiştir. DVR ile yapılan süre kısaltması çalışmasında model 1000-7000 örneklem aralığında eğiterek testleri yapılmıştır. ÇKA yönteminde giriş segment uzunluğunun belirlenmesi için segment boyutu 50-950 aralığında olmak üzere testler yapılmış ve giriş segment boyutu 200 olarak seçilmiştir. ÇKA yönteminde 200 girişli 1 giriş katmanı, 100 nöronlu 1 gizli katman ve 1 çıkışlı çıkış katmanı olmak üzere 3 katman kullanılmıştır. UKSH yönteminde 1 girişli giriş katmanı, 200 nöronlu UKSH katmanı ve tek çıkışlı çıkış katmanı olmak üzere 3 katmanlı mimari kullanılmıştır. Yapılan süre kısaltma çalışmalarının sonucunda DVR ile toprak bileşenlerinin tahminleri yapılmıştır.

Anahtar kelimeler: Zaman Serileri, Eğri Uydurma, Destek Vektör Regresyonu, Çok Katmanlı Algılayıcı, Uzun Kısa-Süreli Hafıza

EXTENDED ABSTRACT

In this thesis, in the soil texture analysis experiment carried out to increase soil fertility, the experiment time shortening study for the experiments to be carried out with the Laser Guided Bouyoucos device and in the light of the data obtained as a result of these studies, the soil component ratios were estimated with a shortened test period. For this purpose, the traditional method of curve fitting was applied on the soil signals obtained with the LGB device. Due to the success of the curve fitting method, an approach based on machine learning methods has been proposed in order to achieve higher success. Since the soil signals obtained with the LGB device are time series, machine learning methods such as Support Vector Regression (SVR), Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) architectures have been studied. For all methods, 52 different soils, which were previously measured with both the hydrometer method and the LGB device, were used. For each soil, 2-hour soil texture analysis signals obtained from the LGB device were studied.

Due to the fact that the hydrometer method is cumbersome and lacking in technology, the expert opinion and lack of technology of the hydrometer method have been prevented thanks to the LGB device developed. With this thesis, which is made with more than one approach, it has been seen that the LGB device can provide an improvement in the experimental time, which is one of the missing aspects. Since the soil texture analysis results obtained in the experiments with the LGB device are time series, tests were carried out with the methods that gave the most successful results in time series estimation.

The curve fitting method, which is traditionally used in time series forecasting, was applied to the 2-hour data of 52 different soil signals obtained with LGB. In this method, experiments were made with many functions such as logarithmic, exponential, polynomial, equipotential, gaussian, linear. As a result of the experiments, the calculated R-squared values were calculated and the 2nd

degree exponential equation was determined as the most suitable function. By using the Levenberg-Marquardt algorithm as the training algorithm, the R-squared value was higher. In the curve fitting method, the equation of the curve is derived by reducing 200 (approximately 1.5 minutes) samples from the last sample (14400) for each of the soil signals. For each soil, 50 different tests and estimations were performed. The absolute error values were chosen according to the sensitivity of the LGB device, with an average of 0.2V and a maximum of 0.4V. According to the selected error values, the shortest curve segment with which an acceptable estimation can be made has been found. As a result of the tests carried out with this method, it has been seen that the 2-hour experiment period can be shortened to 1 hour and 10 minutes, approximately 42%.

In the studies carried out with the SVR method, a single input and output model was used. All soil signals were divided as training set and the remaining samples as test set by increasing 500 (approximately 4 minutes) samples in the range of 1000 to 7000 samples. Samples allocated as training set were subjected to model training with 1 time interval and estimation was performed with the trained model. By calculating the absolute error values between the predicted values and the soil signals, the shortest signal segment suitable for the acceptable error was found. As a result of the tests made with the SVR, it was seen that the soil texture analysis experiment could be shortened to 25 minutes, approximately 79%.

In the MLP method, 1 input layer, 1 hidden layer and 1 output layer are architecturally selected. In the MLP architecture, the soil signals are divided into training and test data in itself. In the training of MLP, the signal part reserved for training was divided into certain segment lengths and given as input, and the first value after this segment was given as the output value. In this way, the training of the MLP architecture was completed with one sequence input and one output. The tests were carried out by increasing the input layer by 50 samples between 50-950 sizes on randomly selected soils and the input layer length was determined as 200 since it was observed that the least incorrect result was obtained when the input layer length was 200. Tests were carried out by increasing the number of samples by 500 from the first 1000 samples of the soil signals to 7000 samples with the MLP. And as a result of the tests, it has been observed that the test time can be shortened to 25 minutes, approximately 79%.

The LSTM method, which is another artificial neural network model with high success in time series forecasting, has been used to obtain more successful results. Thanks to the forget gate in the LSTM architecture, the number of input values in the input layer is used as 200. The LSTM model used in the study is a 3-layer architecture as 1-input input layer, 200-neuron LSTM layer and 1-output output layer. The inter-layer link is used as a fully-connected. In this method, as in the MLP method, the tests were carried out by increasing the sample number by 500 from the first 1000 samples of the soil signals to 7000 samples. Since the LSTM method is sensitive to the standardization of the collected data, all collected data were standardized before using this method. As a result of the tests made with this method, it was observed that the soil texture analysis experiment time could be shortened from 2 hours to 16.5 minutes, approximately 86%.

According to traditional or machine learning methods used in time series forecasting, certain parts of the soil signal are studied. Since the error values of the curve fitting method are high on ground signals shorter than 4500 samples, curve fitting method studies were not performed on ground signals shorter than 4500 samples. In machine learning methods, since successful results were obtained on signals shorter than 7000 samples, these methods were not used on soil signals longer than 7000 samples.

Soil texture analysis was carried out using the SVR method in the light of the test time shortening tests. The ratio of soil components (sand, silt and clay) was estimated using abbreviated soil texture signals. The absolute error values between the estimated rates and the rates calculated by the hydrometer method are given. As a result of the study carried out with this method, it was observed that the estimation of the sand and silt ratios could be made in a short time, and a 2-hour test period was needed only for a more accurate determination of the clay. It has been seen that the test time can be shortened to 3600 samples by taking into account a small margin of error in the clay determination and soil components can be detected with an acceptable margin of error.

As a result of the work done in this thesis, 52 different soils were studied with the LGB device. In the soil texture analysis time shortening experiments performed on the soil signals obtained with the LGB device, the test duration; can be shortened to 8400 samples (approximately 42%) with the curve fitting method, 3000 samples (approximately 79%) with the SVR and MLP method, and 2000 samples (approximately 86%) with the LSTM method. It has been observed that the soil component ratios could be estimated with the SVR method with a 30-minute experiment period.

GENİŞLETİLMİŞ ÖZET

Bu tezde, toprak verimliliğini arttırmak amacıyla yapılan toprak tekstür analizi deneyinde Lazer Güdümlü Bouyoucos cihazı ile yapılacak olan deneyler için deney süresi kısaltma çalışması ve bu çalışmalar sonucunda elde edilen veriler ışığında kısa deney süresi ile toprak bileşen oranlarının kestirimi yapılmıştır. Bu amaçla LGB cihazı ile elde edilen toprak sinyalleri üzerinde öncelikle geleneksel yöntem olan eğri uydurma yöntemi uygulanmıştır. Eğri uydurma yönteminin başarı göstermesi sebebi ile daha yüksek başarı elde edebilmek için makine öğrenmesi yöntemlerine dayalı bir yaklaşım önerilmiştir. LGB cihazı ile elde edilen toprak sinyalleri zaman serisi olduğu için, makine öğrenmesi yöntemlerinden Destek Vektör Regresyonu (DVM), Çok Katmanlı Algılayıcı (ÇKA) ve Uzun Kısa-Zamanlı Hafıza (UKSH) mimarileri ile çalışılmıştır. Tüm yöntemler için önceden hem hidrometre yöntemi hem de LGB cihazı ile ölçümleri yapılan 52 farklı toprak kullanılmıştır. Her bir toprak için LGB cihazından elde edilen 2 saatlik toprak tekstür analiz sinyalleri ile çalışılmıştır.

Hidrometre yönteminin hantal ve teknoloji yoksunu bir yöntem olması sebebi ile geliştirilen LGB cihazı sayesinde hidrometre yönteminin uzman görüşü ve teknoloji yoksunluğunun önüne geçilmiştir. Birden fazla yaklaşım ile yapılan bu tez ile LGB cihazının eksik kalan yönlerinden biri olan deney süresi konusunda gelişme sağlayabileceği görülmüştür. LGB cihazı ile yapılan deneylerde elde edilen toprak tekstür analizi sonuçları zaman serisi olduğu için zaman serisi tahminleme de en başarılı sonuçlar veren yöntemler ile testler yapılmıştır.

Zaman serisi tahminleme de geleneksel olarak kullanılan eğri uydurma yöntemi LGB ile elde edilen 52 farklı toprak sinyallerinin 2 saatlik verilerine uygulanmıştır. Bu yöntemde logaritmik, üssel, polinomal, ekponansiyel, gaussian, doğrusal olmak üzere birçok fonksiyon ile deneme yapılmıştır. Yapılan denemeler sonucunda hesaplanan R-kare değerleri hesaplanmış ve 2.dereceden eksponansiyel denklem en uygun fonksiyon olarak belirlenmiştir. Eğitim algoritması olarak

Levenberg-Marquardt algoritması kullanılarak R-kare değerinin daha yüksek olması sağlanmıştır. Eğri uydurma yönteminde toprak sinyallerinin her biri için son örneklemden (14400) itibaren 200 örneklem azaltılarak eğrinin denklemi çıkarılmıştır. Her toprak için 50 farklı test ve tahminleme işlemi yapılmıştır. Mutlak hata değerleri LGB cihazının hassasiyetine göre ortalama 0.2V, maksimum olarak ise 0.4V seçilmiştir. Seçilen hata değerlerine göre kabuledilebilir tahminleme yapılabilen en kısa eğri parçası bulunmuştur. Bu yöntem ile yapılan testler sonucunda 2 saatlik deney süresinin 1 saat 10 dakikaya kadar, yaklaşık %42 oranında kısaltılabileceğini görülmüştür.

DVR yöntemi ile yapılan çalışmalarda tek giriş ve çıkışlı model ile çalışılmıştır. Tüm toprak sinyalleri 1000 ile 7000 örneklem aralığında 500 (yaklaşık 4 dakika) örneklem arttırılarak eğitim seti olarak ve kalan örneklemler test seti olarak bölünmüştür. Eğitim seti olarak ayrılan örneklemler 1 zaman aralığı ile model eğitimine tabi tutulmuştur ve eğitilen model ile tahminleme işlemi yapılmıştır. Tahmnilenen değerler ile toprak sinyalleri arasındaki mutlak hata değerleri hesaplanarak kabul edilebilir hataya uygun olan en kısa sinyal parçası bulunmuştur. DVR ile yapılan testler sonucunda toprak tekstür analizi deneyinin 25 dakikaya, yaklaşık %79 oranında kısaltılabileceği görülmüştür.

ÇKA yönteminde 1 giriş katmanı, 1 gizli katman ve 1 çıkışlı çıkış katmanını mimari olarak seçilmiştir. ÇKA mimarisinde toprak sinyallerini kendi içerisinde eğitim ve test verisi olarak ayrılmıştır. ÇKA'nın eğitiminde eğitim için ayrılmış olan sinyal parçasını kendi içerisinde belirli segment uzunluklarına ayrılarak giriş olarak verilip, bu segmentten sonraki gelen ilk değer çıkış değeri olarak verilmiştir. Bu sayede ÇKA mimarisinin eğitimi bir sekans girişli ve tek çıkışlı olarak tamamlanmıştır. Rastgele seçilen topraklar üzerinden giriş katmanını 50-950 büyüklük arasında 50'şer örneklem arttırarak testleri yapılmış ve giriş katman uzunluğu 200 olduğunda en az hatalı sonucun elde edildiği gözlemlendiğinden dolayı giriş katman uzunluğu 200 olarak belirlenmiştir. ÇKA ile toprak sinyallerinin ilk 1000 örnekleminden itibaren 7000 örneklemine kadar örneklem sayısı 500 arttırarak testler yapılmıştır. Ve yapılan testler sonucunda deney süresinin 25 dakikaya, yaklaşık %79 oranında kısaltabileceği gözlemlenmiştir.

Zaman serisi tahminleme de yüksek başarı gösteren bir başka yapay sinir ağı modeli olan UKSH yöntemi daha başarılı sonuçlar elde etmek için kullanılmıştır. UKSH mimarisinin kendi içerisinde bulunan unut kapısı sayesinde giriş katmanındaki giriş değerlerinin sayısı 200 olarak kullanılmıştır. Çalışmada kullanılan UKSH modeli 1 girişli giriş katmanı, 200 nöronlu UKSH katmanı ve 1 çıkışlı çıkış katmanı olarak 3 katmanlı mimaridir. Katmanlar arası bağlantıyı tam bağlantı olarak kullanılmıştır. Bu yöntemde de ÇKA yönteminde olduğu gibi toprak sinyallerinin ilk 1000 örnekleminden itibaren 7000 örnekleme kadar örneklem sayısını 500 arttırarak testler gerçeklenmiştir. UKSH yöntemi toplanmış verilerin stadardizasyonuna duyarlı olduğu için bu yöntem kullanılmadan önce toplanan tüm veriler standardize edilmiştir. Bu yöntem ile yapılan testler sonucunda toprak tekstür analizi deney süresinin 2 saatten 16,5 dakikaya, yaklaşık %86 oranında kısaltılabileceğini gözlemlenmiştir.

Zaman serisi tahminlemelerinde kullanılan geleneksel veya makine öğrenmesi yöntemlerine göre toprak sinyalinin belirli parçaları üzerinde çalışılmıştır. Eğri uydurma yönteminin 4500 örneklemden daha kısa olan toprak sinyalleri üzerinde hata değerlerinin yüksek olması sebebi ile 4500 örneklemden daha kısa toprak sinyalleri üzerinde eğri uydurma yöntemi çalışması yapılmamıştır. Makine öğrenmesi yöntemlerinde ise 7000 örneklemden daha kısa sinyaller üzerinden başarılı sonuçlar elde edildiği için 7000 örneklemden daha uzun toprak sinyalleri üzerinde bu yöntemler kullanılmamıştır.

Yapılan deney süresi kısaltması testleri ışığında DVR yöntemi kullanılarak toprak tekstür analizi çalışması yapılmıştır. Kısaltılmış toprak tekstür sinyalleri üzerinden toprak bileşenlerinin (kum, silt ve kil) oran tahmini yapılmıştır. Tahmin edilen oranlar ile hidrometre yöntemi ile hesaplanmış oranlar arasındaki mutlak hata değerleri verilmiştir. Bu yöntem ile yapılan çalışma sonucunda kum ve silt oranlarının tahminin kısa sürede yapılabileceği, sadece kilin daha doğru tespiti için 2 saatlik deney süresine ihtiyaç olduğu gözlemlenmiştir. Kil tespitinde ufak bir hata payı göze alınarak deney süresinin 3600 örnekleme kısaltılabileceği ve kabul edilebilir bir hata payı ile toprak bileşenlerinin tepit edilebileceği görülmüştür.

Bu tezde yapılan çalışma sonucunda LGB cihazı ile 52 farklı toprak üzerinde çalışılmıştır. LGB cihazı ile elde edilen toprak sinyalleri üzerinde yapılan toprak tekstür analizi deneylerinde, deney süresini eğri uydurma yöntemi ile 8400 örneklem yaklaşık %42 oranında, DVR ve ÇKA yöntemi ile 3000 örneklem yaklaşık %79 oranında ve UKSH yöntemi ile 2000 örneklem yaklaşık %86 oranında kısaltılabileceği görülmüştür. DVR yöntemi ile toprak bileşen oranlarının 30 dakikalık deney süresi ile tahminlenebileceği incelenmiştir.

ACKNOWLEDGEMENT

The study presented in this thesis was carried out under the supervision of Asst. Prof. Dr. Umut ORHAN. I would like to express my sincere gratitude to him for his supervision guidance, patience, motivation, useful suggestions, and his valuable time for this work.

I would like to thank members of MSc thesis jury, Assoc. Prof. Dr. Ali İNAN and Asst. Prof. Dr. Serkan KARTAL for their suggestions and corrections.

I would like to thank to TÜBİTAK for supporting this study. This study is part of 118O162 project, supported by TÜBİTAK.

CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF ABBREVIATIONS

Laser	: Light Amplification by Stimulated Emission of Radiation
LGB	: Laser Guided Bouyoucos
SVM	: Support Vector Machine
SVR	: Support Vector Regression
MLP	: Multilayer Perceptron
XOR	: Exclusive Or Gate
ANN	: Artificial Neural Network
LSTM	: Long Short-Term Memory
RNN	: Recursive Neural Network
pH	: Power of Hydrogen
$NaPO_3$	: Sodium Hexametaphosphate
LDR	: Light-Dependent Resistor
USDA	: United States Department of Agriculture
t	: Time Stamp
y	: Output Value
$\hat{y}$	: Predicted Value of y
$\overline{y}$	: Mean Value of y
IR-UWB	: Impulse Radio Ultra-Wideband
TH PPM	: Time-Hopping Pulse Position Modulation
BPSK	: Binary Phase Shifting Key
LSCFM	: Least Square Curve Fitting Method
URR	: Ultimate Recoverable Resources
GO	: Graphene Oxide
TiO_2	: Titanium Oxide
f	: Function
W	: Weight
b	: Bias

C	: Candidate Vector
GRU	: Gradient Recurrent Unit
SRN	: Simple Recursive Neural Network
ANOVA	: Analysis of Variance
CW-RNN	: Clockwork Recursive Neural Network
CPU	: Central Process Unit
MAE	: Mean Absolute Error
MaxAE	: Maximum Absolute Error
Seq2one	: Sequence-to-one
AE	: Absolute Error
μ	: Mean
x	: Input Value
BiLSTM	: Bidirectional Long Short-Term Memory

1. INTRODUCTION

Soil is the outermost layer of the earth, where millions of living species live and which is indispensable for life. It is the most basic resource necessary for life since the beginning of human history. It is formed as a result of the disintegration and crumbling of soil rocks by physical, chemical and biological effects and changing their composition. The structure of the soil, its formation, composition, and the events taking place in the soil are important for the continuation of life on earth. Soil texture is important for areas that greatly affect human life, from the plant to be grown to the structure of the construction to be made. For this reason, soil analysis is necessary in order to learn the soil texture, and with these analyzes, the soil can be classified. In order to classify the soil, it is necessary to know its texture.

Soil texture is determined by the proportions of sand, silt and clay particles in the soil. These inorganic particles are separated according to their size. Those with a diameter between 2 and 0.02 mm are called sand, those between 0.02 and 0.002 mm are called silt and those with a diameter of less than 0.002 mm are called clay. There are many methods in the literature for the determination of sand, silt, and clay ratios. Among these methods, the pipette method is the most reliable and gives detailed results. However, the laboratory equipment required for the pipette method is excessive and the analysis time is long. The Bouyoucos-hydrometer method, which is another analysis method, is the most preferred and traditionally accepted method because it gives quicker results compared to other methods and requires less test equipment. Despite these advantages of the hydrometer method, it is tiring, lacking in technology and has a high margin of error.

There are many studies conducted to eliminate the disadvantages of the hydrometer method. In the light of these studies, the Light Amlification by Stimulated Emission of Radiation (Laser) Guided Bouyoucos (LGB) device was produced. The LGB device works with the principle of recording and analyzing the change in the intensity of the laser passed through the measuring tape with the sedimentation of the soil in the suspension formed with the soil to be analyzed, depending on time.

Time series are the values obtained by measurements made on one or more variables with certain time intervals of events that can be expressed numerically. In time series, each data has a timestamp. In short, it is a series of observations taken sequentially over time. The annual flow of a river, the monthly production of a factory, seconds of energy spent in a living body are examples of time series. Below is the change between the monthly Turkish lira and US dollar exchange rates between January 2006 and December 2010.

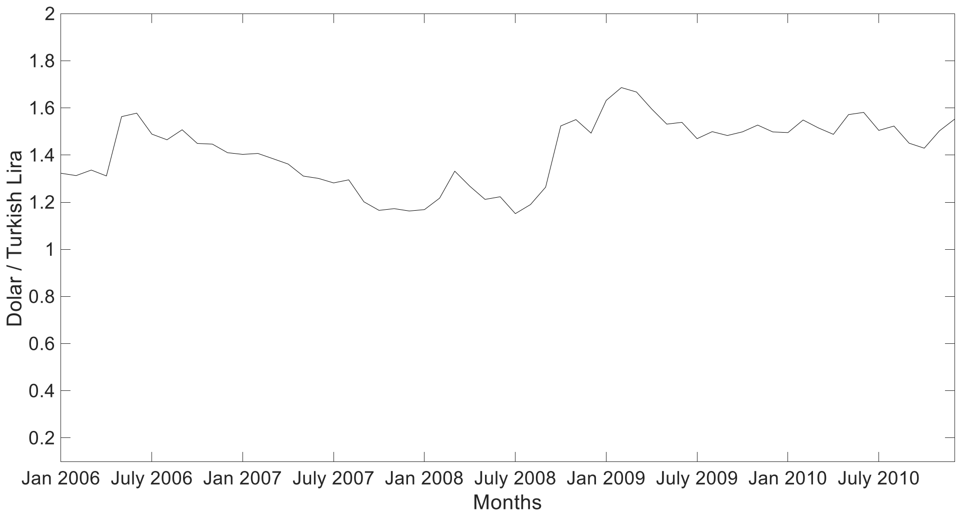

Figure 1.1. 5-year change between Turkish lira nad US dollar Exchange rate.

As seen in Figure 1.1, the ratio between the Turkish lira and the US dollar was recorded daily between January 2006 and December 2010, and a time series was obtained by ordering the recorded data according to time. Time series are used to detect and solve problems in fields such as agriculture, finance, production, energy, health, security, weather, and astronomy. For this, it is necessary to analyze the obtained time series. With the other definition of time series analysis methods, with signal processing methods, successful analyzes can be made on time series such as pattern recognition, information extraction, future prediction. The methods that analyze the signal depending on the frequency are called frequency domain methods, and the methods that analyze the signal depending on the time are called time domain methods.

The most widely used and traditionally accepted method in signal processing is the curve fitting method. The curve fitting method is based on the chronological ordering of the timestamped data and finding the most suitable curve for this ordering. The mathematical function of the curve found can be exponential, trigonometric, exponential, linear or polynomial. Below is an example of the curve method applied to the monthly average temperature data of Istanbul in 2012.

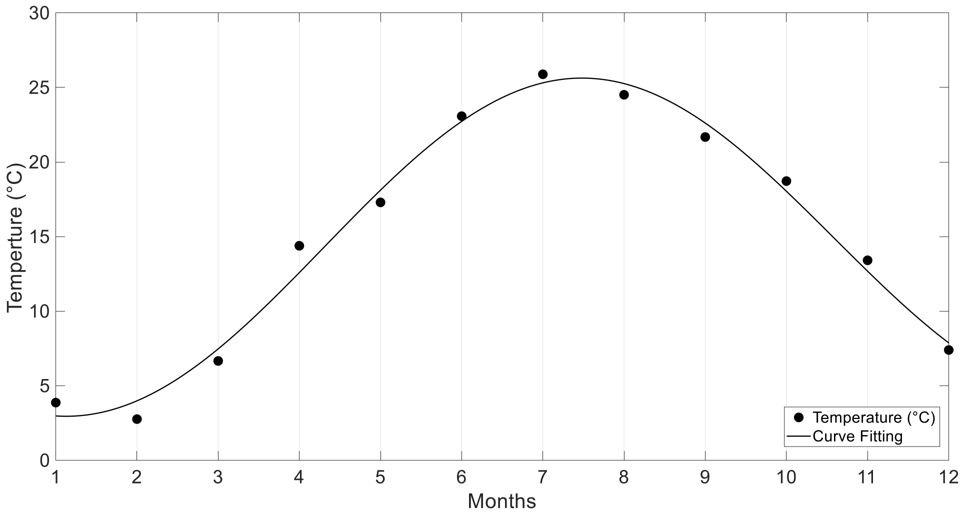

Figure 1.2. Sinusoiodal curve fitting method in accordance with the temperature data calculated in Istanbul in 2012.

As shown in Figure 1.2, depending on the data obtained, the curve closest to the sign with a certain margin of error can be found with the curve fitting method and the mathematical function of this curve can be reached.

The support vector machine (SVM) is one of the most widely used and highly successful artificial neural network models for classification since the 1960s. As a result of the studies carried out due to its high success for classification, the concept of support vector regression (SVR) emerged in the 2000s. The SVR method is a simple ANN architecture that can be used for both regression analysis and classification and can produce successful results in both subjects.

Another method frequently used in signal processing is the Multilayer Perceptron (MLP) method, which is one of the deep learning methods. MLP is a feed forward artificial neural network (ANN) model developed to solve the Exclusive Or Gate (XOR) problem without linear separability.

The Long Short-Term Memory (LSTM) model used for signal processing is a Recursive Neural Network (RNN) architecture. It was developed to solve the vanishing gradient problem that occurs in traditional RNNs. Unlike a feed forward neural network, LSTM also has feedback connections. Thanks to the feedback, the effect of important events that occurred randomly or with a delay in the time series can be calculated. For this reason, it is a very suitable model for classification, processing and forecasting in time series. A standard LSTM cell consists of three separate doors, the entrance door, the exit door, and the forget door.

In this thesis, it was studied on shortening the soil analysis in time series collected with the LGB device, which was developed based on the standards of the hydrometer method for soil analysis. The LGB device is a promising method for learning the soil texture, but it has not been able to innovate in the experiment time required for soil analysis. For this reason, the soil analysis experiment time needs to be developed technologically. In the time series obtained with the LGB device, estimation was made by curve fitting, SVR, MLP and LSTM methods and the results were presented comparatively. In addition, using the shortened test results, tests were made on the determination of soil components with the SVR method, and the results were presented.

2. SOIL TEXTURE ANALYSIS

Soil analysis is the determination of the proportions of sand, silt and clay particles in the soil. Soil classification can be done with the determined ratios. Soil class has a great effect on the work to be done on the soil. Especially in plant production, the fertilizer to be used according to the soil class causes the product to be obtained more and to be of high quality. The beginning of soil classification dates back to the work of Attenberg (1916). Attenberg was influenced by Ritten von Rittenger's work on sieves and Oden's (1915) application of Stoke's law to soil science. Soil is named according to the particles with a higher percentage when classifying. This nomenclature is made according to the soil texture triangle. The soil texture triangle is given in Figure 2.1.

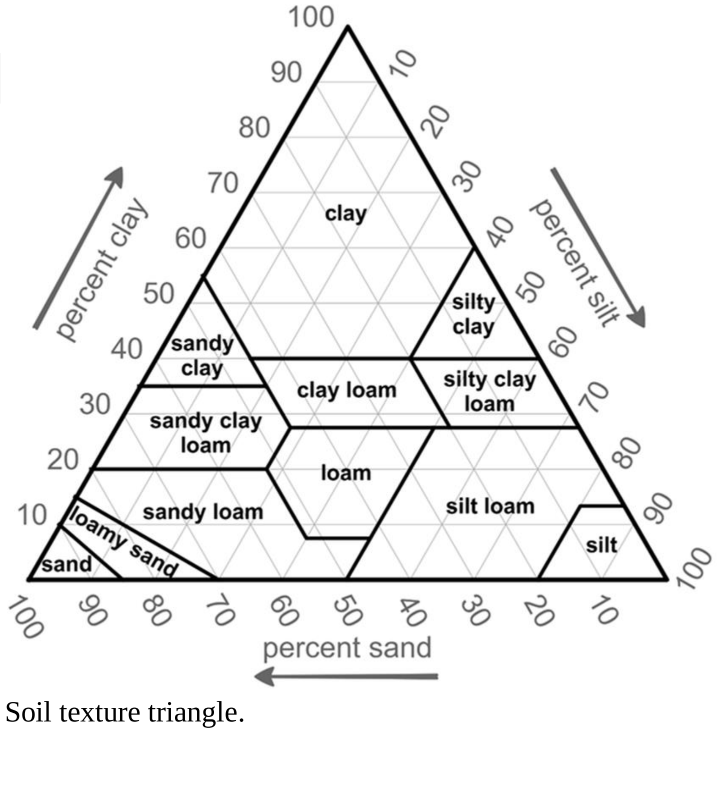

Figure 2.1. Soil texture triangle.

As shown in Figure 2.1, the soil names are clay, silt, sand, loam, silty clay, sandy clay, clay loam, silt loam, sandy loam, loamy sand, silty clay loam and sandy clay loam.

Today, there are more than 400 methods for soil texture analysis (Barth and Sun, 1985; Syvitski, 1991). However, most of these methods require equipment for analysis. For this reason, methods such as sedimentation, pipette, hydrometer, which are acceptable methods and require less equipment than other methods, are used more frequently. In mechanical analysis, which is another method, particles passing through a 2mm sieve are counted. Mechanical analysis was standardized in 1928 by the Soil Science Society.

Particles in the soil can take many different forms. For this reason, there is no common decision on analyzes and analysis results. In addition, external factors such as equipment used for analysis, environment and temperature also affect the sensitivity of the analysis. For this reason, the sensitivity of the analysis varies according to the environment in which the analysis is carried out and factors such as the error possibilities of the equipment used. In addition, even within the same region, soil samples taken from more than one place may differ from each other. For this reason, the soil sample to be taken from the region to be analyzed must have the characteristics of the region. The prerequisites for the analysis are that the soil sample is taken from the soil suitable for sampling and that it is sufficient in terms of the particles it contains. Samples taken in accordance with the preconditions are called “repeatable” and these samples are suitable for post-analysis experiments. According to the experiments to be carried out, more than 2,000 tests can be performed with 1 kg soil sample.

Particle shape and size are essential for particle classification required in soil texure analysis. But since it can only be represented by a single parameter, experts classify particles according to their spherical state. Expert experience, high error rate and laboratory environment required in traditional soil texture analysis methods make soil texture analysis difficult. For this reason, there are many technological studies in the literature (Ringrose-Voase and Bullock, 1984; Tyler and Wheatcraft, 1992; Grout et. al., 1998, Stabinger et. Al., 1967). However, the equipment used in these studies is expensive and scarce. In the recent past, it has been seen that with the help of cheap and easily available sensors, with the right method and hard work, smooth, realistic and reliable data can be produced, and soil analysis can be made by processing these data.

The LGB made in the light of the traditional methods used for soil analysis and the bouyoucos-hydrometer method are given in detail below.

2.1. Traditional Methods

The characterization process is done by looking at the amount of large particles. Particles between 10 cm and 1 mm are considered large for soil texture analysis methods. Although there are sieving methods for calculating the amounts of these particles, the most preferred method is the direct measurement method. The direct measurement method represents the measurement made with an equipment such as a ruler, compass, or scale without any intervention. Coroni and Maraga (1983) made estimations with the help of a compass, Laxton (1980) with a photographic technique, Shiozawa and Campbell (1991) with the help of particle radius and geometric standard deviation, and Nadeau (1985) with the help of a microscope. Allen et al. They compiled the most used microscope studies in 1996. Tovey and Smart used electron microscopy in 1982 to examine particles smaller than 20µm.

With the direct measurement method, full analysis of the particles can be made. However, the preparation time for analysis is long (Griffiths, 1967). And during the preliminary preparation period, the organic substances in the soil dissolve and the soil content changes. The unequal particle size is a problem for direct measurement. While particles of equal size can be easily examined with a microscope, the measurement differences between the axes will make a big difference for particles smaller than 5µm (Nadeau et. al., 1984). Also, the error rate depends on the vertical size of the particle. As the vertical dimension increases, the difference between the sections to be taken will increase and this will increase the error rate (Allen et. al., 1996).

The elimination method, which is one of the traditional methods, is a primitive but at the same time one of the most effective tools. The sieves used to determine the type of soil have square or circular openings between 125 mm and 5 μm. Circular perforated sieves measure the particle by radius, while square perforated sieves measure the particles according to the distance between its two parallel faces and the diagonal distance in the particle. The sieving method can be done on wet or dry soil. While there are tools for wet sieving, there is a method called air jet for dry sieving (AFNOR, 1982). Square hole sieves are used for particles below 2 mm. Electronically shaped sieves are used for particles below 37 μm (ISO, 1998). Sieves made with metal plates are used to separate large particles because they have larger holes. It is still a method used for soil analysis today, as they can produce good results with the right method and sufficient care. However, the sieving method is not recommended for particles below 30 μm, as it is a tiring and a method for smaller particles. The majority of the mistakes made in the sieving method are due to the use of the sieves by shaking them in only one direction. Putting a large amount of soil samples on the sieve and sieving using sieves with faulty holes are other errors in the sieving method (Metz, 1985; Head, 1992). In addition, the difficulty of cleaning and maintaining the sieves is another disadvantage of the sieving method.

The sedimentation method is the most widely used traditional soil analysis method today. The sedimentation method is made by the movement of the suspension made with the soil sample to be analyzed towards the bottom of the container over time with the effect of gravity. The first sieving method followed by the sedimentation method is the most commonly used method for particle size analysis. Before the sedimentation method, the soil sample to be analyzed is set up and sieved. Afterwards, estimation is made on the particles sinking to the bottom of the container by being thrown into a solution. The solution used is sodium hexametaphosphate (NaPO3), which has a pH value of 9.5, and it may rarely differ in some solutions (Klute, 1986). The solution used for soils containing volcanic ash should be acidic (Maeda et. al., 1977).

Estimation is made by assuming that each particle in the created solution behaves according to Stokes' law. Allen et. al., the results of their study in 1996 are below.

- Thin and flat particles sink to the bottom by zigzagging, and therefore their settling speed is slower.
- The sedimentation method is unreliable for particles smaller than 1µm.
- Since the electrical interaction between the diluted electrolytes and the soil particles does not affect the sedimentation rate of the particles, it can be ignored.

In the pipette method, density method and centrifuge method developed based on the sedimentation method, it is assumed that the particles behave according to Stokes' law.

For the pipette method, also called macro pipette, the soil to be analyzed is placed in a container and a suspension is created by adding water to it. The mixture is taken from the upper part of this suspension at certain time intervals and dried. Particle ratios are estimated according to the differences of the particles measured from the dried parts. The pipette method is generally used after sieving particles smaller than 63µm (ISO, 1998; U.K. BSI, 1998; Germany DIN, 1983; France AFNOR, 1983c). The United States Department of Agriculture (USDA, 1996) and Agriculture Canada (Carter, 1993) prefer the pipette method for soil texture analysis. The pipette method is very laborious for particles smaller than 2µm. Burt et. al. who wants to avoid this disadvantage get conducted a micropipette study in 1993 and presented the results of this study in comparison with the macropipette.

Density method is a method developed with the logic of sedimentation method. Their density is lower when the soil contains excess air or organic matter. On the contrary, in the case of excess inorganic substances (such as iron, manganese, titanium), their densities are high. Bouyoucos-hydrometer method is the most used method among the density method. As a preliminary preparation in the hydrometer method, the soil sample is kept in the oven at approximately 100ºC for 1 day, and then it is purified from large particles by passing through a 2mm sieve. 40gr sodium hexametaphosphate is mixed with 1 liter of distilled water and left for 1 day. 50-100 gr soil sample is mixed with 125ml sodium hexametaphosphate and left for 1 day. After mixing for 5-10 minutes in a light mixer, it is placed in a 1000 ml measuring cup and distilled water is added up to the 1000 ml line and mixed. 40 seconds after the suspension is prepared and mixed, the hydrometer is immersed, and the reading is recorded. This value represents the sum of the clay and silt values in the soil sample. The second reading is done after 60 seconds, the third reading is done after 1 hour and the last reading is done after 2 hours. The temperature value of the suspension is also recorded at each reading operation. The value obtained with the last reading gives the clay ratio. The silt ratio is calculated by the difference between the first read value and the last read value. The sand ratio is also obtained by subtracting the calculated values from 100. The hydrometer method has been standardized by ISO (ISO, 1977; ISO, 1998).

The reason why the Bouyoucos-hydrometer method is the most commonly used method in soil analysis is that it is simple and requires less equipment. However, the hydrometer method has many disadvantages. The most common mistake made in the hydrometer method is the wrong reading of the hydrometer value (Head, 1992). Excess organic matter in the suspension is the main reason for this error. Even foaming formed by a suitable oxidation method affects the hydrometer value. Another reason is the presence of soluble salt in the soil sample. Salts dissolved in the suspension affect the results of the hydrometer method as they change the density. In addition, the valve part of the hydrometer can cause errors (Allen et. al., 1996). Gee and Bauder in 1986 showed that 40gr soil sample for 1 liter suspension would give good results in their study. There are studies to accelerate the analysis time in the hydrometer method (Sur and Kukal, 1992). In order to improve the method and eliminate the need for expert control, Stabinger et al. get. (1967) used an ultrasonic technique, Orhan and Kılınç (2020) used laser and photosensitive resistor. In these two studies, the applicability of the hydrometer method by digitizing was defended. In another study, a significant relationship was established between the density of the suspension and the amount of particles smaller than 2 μm (Smith, 2001).

Another method created in the light of the sedimentation method is the centrifugation method. Centrifuge is the artificial realization of gravity with tubes connected to a high-speed motor. The centrifuge method is generally used for particles smaller than 1μm. The centrifugation time and the sedimentation of particles by Tanner and Jackson (1947) were accepted as the standard for small particles by Avery and Bascomb (1982) and USDA (1996).

2.2. Laser Guided Bouyoucos

The hydrometer method is cumbersome and has a high margin of error. In order to overcome these problems, Orhan and Kılınç (2020) designed a new electronic device called LGB, adhering to the preliminary preparations and measurement standards of the hydrometer method used for soil texture analysis, and it has been demonstrated how close the hydrometer analysis results can be achieved with computerized estimation in this device. The soil to be analyzed in this study is first dried as in the hydrometer experiment. Then it is sieved and mixed with $NaPO_3$ and left for 1 day and poured into a measuring tape. The LGB test setup is shown in Figure 2.2.

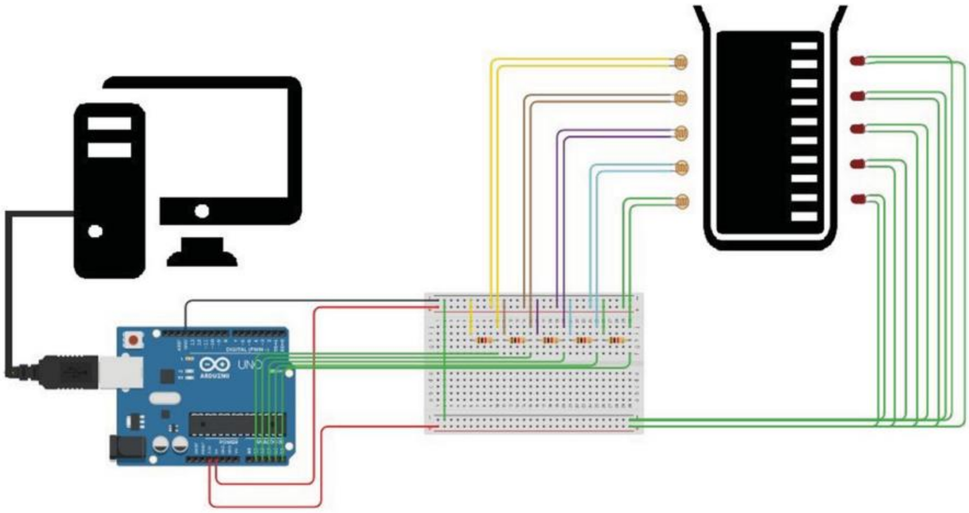

Figure 2.2. LGB test setup (Orhan ve Kılınç, 2020).

As seen in Figure 2.2, a laser light source and LDR sensor were placed at 5 different heights on two opposite surfaces of the 1000 ml measuring tape, passing through its center on the circular axis. The light intensities coming to the LDR sensor were recorded 2 times per second with the help of a microcomputer. In the study, the most efficient and consistent results were obtained with measurements made a few centimeters below the top of the soil-water mixture. The basic logic of the study is based on the correlation of the changes in light intensity with the sedimentation over time with the laser and LDR sensor assembly placed opposite each other, taking advantage of the difference in the sedimentation rates of sand, silt, and clay in the soil-water mixture. With the analysis of the obtained 2-hour time series, estimation was made on the proportions of sand, silt and clay in the soil. The signal obtained from the tested soil is shown in Figure 2.3.

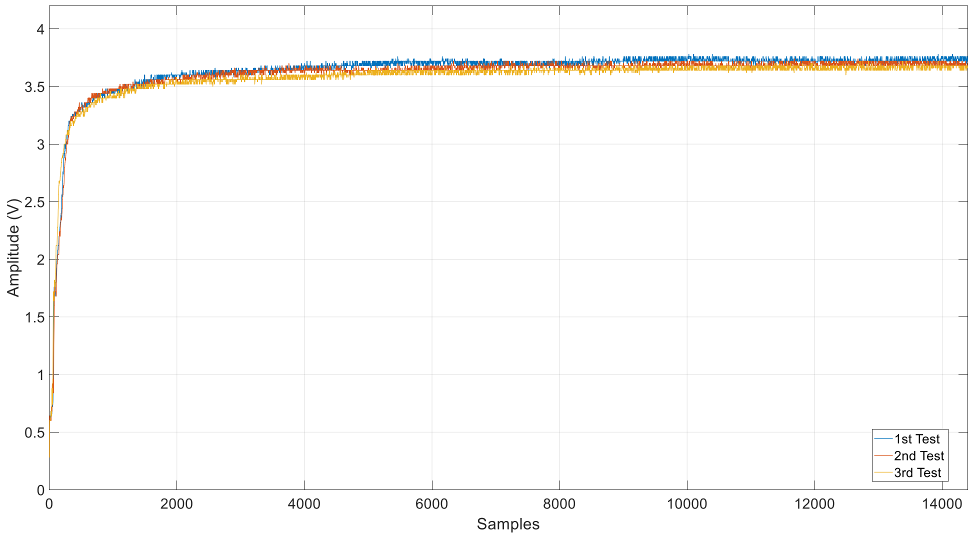

Figure 2.3. Obtained soil signal.

When Figure 2.3 was examined, it was observed that as the sand, silt and clay particles in the soil precipitated, the intensity of light coming to the sensor increased and the slope of the curve changed at certain intervals. The resulting sign traditionally appears to have the appearance of an exponential or exponential curve.

LGB is a promising method for digitalizing the hydrometer method and solving the most common error, the disadvantage of misreading the hydrometer. In addition, unlike the hydrometer method, LGB does not need a laboratory environment. As in the hydrometer method, excess organic matter in the suspension, foaming that may occur with the oxidation method and soluble salt in the soil sample are also factors that may affect the results for LGB.

3. TIME SERIES FORECASTING METHODS

A time series is a set of data calculated at certain time intervals. Thanks to the analysis to be made in these time series, the future can be predicted from the calculated data. Time series forecasting is an important aspect for deep learning methods apart from the goal of predicting the future (Masum et. al. 2018). The estimation to be made over the soil signals obtained with the LGB device is a univariate estimation. In the ANN models to be made, the input data of the model will be timestamps, and the output data will be values with that time stamp. Each input data acts as a window that moves through the time series (Shen el. al., 2015).

The values recorded at moments t_0, t_1, t_2, … will be the output data, while moments t_0, t_1, t_2 will be the input data. In this way, model training is done through the windows given as input into the model, and the model, whose training is completed through these windows, has the capacity to make predictions for other time series. The purpose of training the ANN model is to make the model capable of making accurate predictions with an acceptable margin of error for the time series it may encounter.

The traditional method used in time series estimation is curve fitting. Curve fitting method shows high performance for many time series analysis. Other methods are SVR, MLP and LSTM methods, which are deep learning methods. The aforementioned methods are explained in detail below.

3.1. Curve-Fitting Method

The curve fitting method used for time series estimation is the method of chronologically ordering the data collected at certain time intervals and finding the most suitable curve for this series over the obtained series. With the finding of this curve, the function of the curve is obtained. On this curve, operations such as estimation, pattern recognition, and information extraction can be performed on the time series. In order to qualify the fitted curve as the most suitable curve, the R-squared value is checked. The R-squared value is calculated according to equation (1).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{1}$$

The variable y_i mentioned in equation (1) represents the actual values, the variable $\hat{y}_i$ represents the predicted values, and the variable y_i represents the mean of the y_i values. The R-squared value calculated by this equation is between 0 and 1.

The curve functions obtained by the curve fitting method can be linear as well as exponential, trigonometric, polynomial or exponential. Table 3.1 gives the names and representative functions of some curve functions.

Table 3.1. Some cure fitting function names and representation functions.

Curve Fitting Function	General Model of Function
Linear	$y = a * x + b$
Polinomial (2nd degree)	$y = a * x2 + bx + c$
Sum of Sine (1 term)	$y = a * sin(bx + c)$
Exponential (1 term)	$y = a * eb * x$
Power (2 terms)	$y = a * xb + c$

In addition to the functions given in Table 3.1, there are many functions such as interpolant, Fourrier, Gauss, and Weibull. The curve fitting method can be used for time series estimation as well as in many other fields.

Shabbir at. al., (2020) Shabbir investigated the structural integrity of the propellant grain by performing numerical simulations of the rocket engine for various loading conditions such as pressure and temperature. They said that the repellent material behaves like a rubber material and that materials such as rubber can be characterized using hyperelasticity. Using the strain energy function and the Mooney Rivlin material model to describe the material model, they estimated the material coefficients using the test results by curve fitting. They then used these coefficients to examine the structural integrity of the pusher grain.

Huyen et. al., in 2020, shifted time pulses with radio impulse ultra wide band (IR-UWB), time-hopping pulse position modulation (TH PPM), binary phase shift keying (BPSK) systems together with the least square curve fitting method (LSCFM). They have worked to increase the accuracy of finding buried objects in heterogeneous environments by combining with them. In their study, they validated the analytical statements by simulation and evaluation of the performance of the systems. And as a result of their study, they showed that the method they proposed was more successful than the traditional methods.

Working on the production of fossil fuels, Wang and Feng (2016) examined the reasons why curve fitting models show inconsistent and different results. The effects of ultimate recoverable resources (URR), curve shape, number of cycles and maximum consumption rate on curve fitting models, which are the 4 main factors affecting the results, were analyzed theoretically. In addition, in the same study, they modeled between studies on fossil fuel production in China and 4 basic factors. As a result of the studies, it has been shown that the curve fitting method is not suitable for URR estimation and the curve shape is the most effective feature in a well-founded URR estimation. They argued that increasing the number of cycles in the model should be used with caution, as it adapts to the past data, but creates high production levels and decreases in production after these levels. They showed that when the maximum consumption rate, which is the last basic feature, is limited, it leads to mathematically optimal and practically impossible results. They argued that more consistent results could be obtained in future studies with a comprehensive evaluation of the importance of these four basic factors.

Moghada et. al., (2020) studied the effects of hybrid use of graphene oxide (GO) and titanium oxide (TiO_2) nanomaterials on water thermal conductivity. They carried out their experiments at 0.05, 0.1, 0.2, 0.4, 0.6, 0.8 and 1% concentrations and temperatures between 20 and 50 degrees Celsius (ºC). In their study, they reached the conclusion that temperature and nano-material concentrations resulted in higher thermal conductivity and the maximum thermal conductivity increase was 32.8% by curve fitting method. They also presented a new correlation for thermal conductivity based on experimental data obtained in the laboratory in terms of temperature and nanomaterial volume ratio.

3.2. Support Vector Machine and Regression

Support Vector Machine (SVM) is one of the most well-known methods, which still provides the best classification success in many applications since its foundation in the 1960s. The algorithm, which was first developed by Vapnik and Lerner (1963), took its final form known today with the studies in the AT&T Bell laboratory. Thanks to its extraordinary success in many different problems, it has become one of the standard methods in machine learning (Boser et. al., 1992; Guyon et. al., 1993; Cortes and Vapnik, 1995; Schölkopf et. al., 1995; Schölkopf et. al., 1996; Vapnik et. al., 1997). Schölkopf and Smola (2002) studied the use of SVM for regression and introduced the concept of Support Vector Regression (SVR). The SVR is defined based on a linear or nonlinear "kernel" function and is solved by the quadratic programming approach given by equation (2) – (4).

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} ||w||^2 + C(e^T\xi + e^T\xi^*) \tag{2}$$

$$s.t. \;\; (\phi(\mathrm{A})\mathrm{w} + \mathrm{eb}) - Y \le e\varepsilon + \xi \,, \;\; \xi \ge 0e \tag{3}$$

$$Y - (\phi(\mathrm{A})\mathrm{w} + \mathrm{eb}) \le e\varepsilon + \xi^*, \xi^* \ge 0e \tag{4}$$

The variable indicated by C in equation (2) is a predefined value and provides a balance between errors and regression. A represents the training examples and Y represents the regression outputs. On the other hand, the expressions ξ and ξ^* are artificial examples, while the expression e is the unit vector. Equation (2) – (4) turns into the form in equation (5) using the Lagrangian multipliers a ve a^*.

$$\max_{a,\ a^*} - \frac{1}{2}\ (a^* - a)^T K(A,\ \ A^T)(a^* - a) +\ Y^T(a^* - a) +\ \varepsilon e^T(a^* + a) \qquad (5)$$

$$s.t.\ \ e^t\ (a^* + a) = 0 \qquad (6)$$

$$0e\ \leq a\,, a^* \leq Ce \qquad (7)$$

In order to solve the problem that is the subject of the SVR, the goal becomes to find the b threshold and a^* by solving equations (5) – (7). This solution is also represented by equation (8).

$$f(x) =\ \sum_{i=1}^{n}(\ a^* - a\ )\ K(x_i, X) + b \qquad (8)$$

Here, the kernel function denoted by $K(x_i, x)$ is $\phi(x_i)$ in hyperspace. $\phi(x)$ represents the cross product. Here, the ϕ function can be selected from functions such as linear, polynomial, exponential, diametric, depending on the problem.

3.3. Multilayer Perceptron

Artificial neural networks (ANN) are a model based on living nature. It is basically based on imitating the nervous system of a living thing, allowing the brain to react according to the situation encountered and learning new information from these situations. Just as the nervous system needs an effect to function, the

ANN needs an activation function to function. This function is generally chosen as a nonlinear function. It is used to produce solutions in many areas such as classification, information extraction, future prediction, which many fields are interested in with ANN. ANN stores the previously learned information in their memory for later use. In this way, they can respond more successfully to the situation they may encounter in the future. So much so that if a high-successful ANN encounters missing or erroneous information, it can correct this situation and reach the correct result. In addition to the many advantages of ANN, it has disadvantages such as having a complex structure, difficult modeling, being dependent on hardware, and long learning processes. Figure 3.1 shows a neuron and an ANN cell.

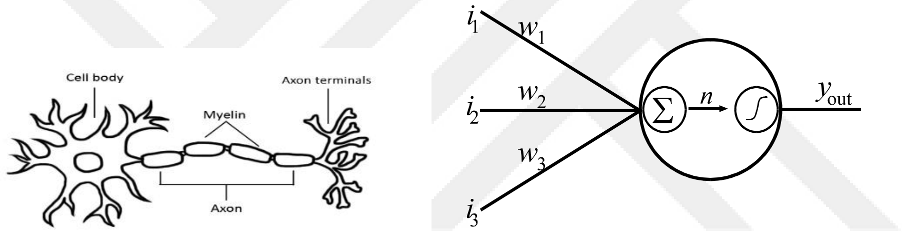

Figure 3.1. Comparison of biological neuron cell (left) and ANN cell (right).

As seen in Figure 3.1, a neuron cell, and an ANN cell, which took its design from a neuron cell, are largely similar to each other. Just as neurons communicate with each other, ANN cells also communicate with each other. Just as neurons transmit the information they receive from the outside to the next neuron; ANN cells transform the information they receive into an output with the help of the functions they contain.

Multi-Layer Perceptron (MLP), which is a type of ANN, has a hidden layer/layers in addition to the input and output layer, unlike a classical ANN. MLP is a feedforward ANN model developed to solve the XOR problem without linear separability. Below is an example of an ANN with the MLP model.

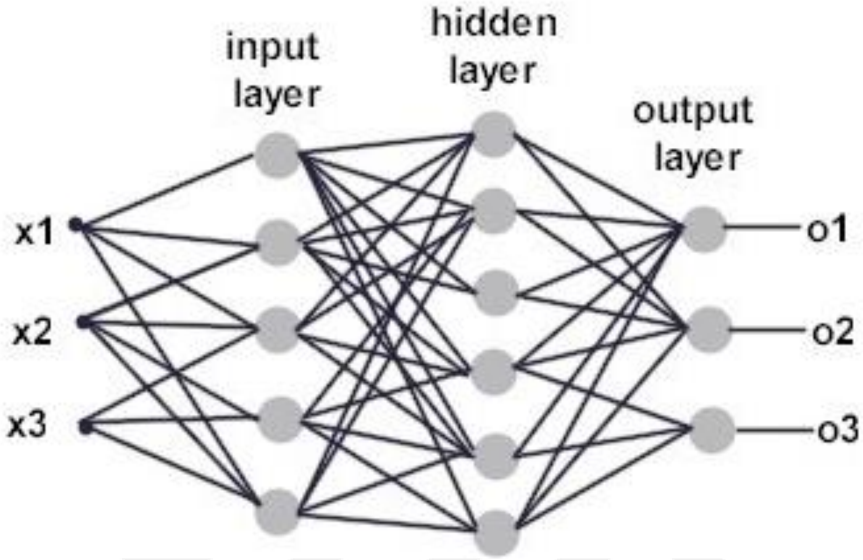

Figure 3.2. An ANN with a MLP model.

As shown in Figure 3.2, artificial neural networks with MLP model consist of three layers: input layer, hidden layer and output layer. While the number of hidden layers may be more than one, the number of inputs (x_1, x_2, x_3, …, x_n) and the number of outputs (o_1, o_2, o_3, …, o_m) may differ.

As in non-multilayer ANN, the input layer is the layer of data coming into the MLP cell. Incoming data is transmitted to the middle layer, also known as the hidden layer, without any interference. While the increase in hidden layers and cells in the hidden layer causes the ANN to learn more accurately, the increase in the number of calculations causes a waste of time. The output layer produces the output of the ANN model by processing the information coming from the middle layer. The model given in Figure 3.2 is a feed forward model. There are also ANN models with feedback. A feedback ANN recalculates the weights in the network using the values obtained from the output layer as back propagation. ANNs are divided into three according to their learning algorithms.

- **Supervised Learning**

Output values are specified for input values transmitted to the network. It updates its weights according to the error between expected outputs and self-produced outputs.

- **Unsupervised Learning**

Only login information is given to the network, with no output information. The network creates its rules to classify each instance among itself. Updates the link weights so that it can group instances of the same property.

- **Reinforced Learning**

In each iteration, the result of the network is rated as good or bad, allowing the network to update the weights again. In this way, it continues operations by both learning and producing results with inputs.

3.4. Long Short-Term Memory

LSTM, an RNN model, is a highly successful ANN architecture for future prediction. It was developed to solve the vanishing gradient problem that occurs in traditional RNN. Unlike a feed forward ANN, LSTM also has feedback links. Thanks to the feedback, the effect of important events that occurred randomly or with a delay in the time series can be calculated. For this reason, it is a very suitable model for classification, processing and forecasting in time series. A standard LSTM is shown in Figure 3.3.

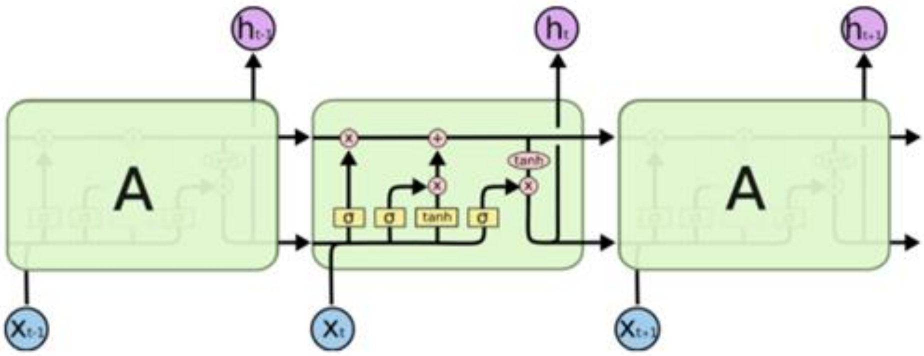

Figure 3.3. A Standard LSTM structure.

As can be seen in Figure 3.3, a standard LSTM cell structure consists of three gates, namely the input gate, the exit gate and the forget gate. Forget gates in LSTM have the features of forgetting or remembering the information obtained from their previous training. Thanks to these gates, it allows the important information in the input data to affect the results more and less the unimportant data.

The first step of LSTM is to decide what information to forget. The sigmoid layer, also called the forget gate, makes this decision according to equation (9).

$$f_t = \sigma(W_f \,.[h_{t-1}, x_t] + b_f), \tag{9}$$

According to h_{t-1} and x_t values, the result produced by equation (9) is between 0 and 1. 1 means completely remember the incoming information, 0 means completely forget the incoming information. Then, the second sigmoid layer updates the weight and bias values according to the information obtained from the first sigmoid layer, according to equation (10). The candidate value vector produced by the equation (11) of the $tanh$ layer and the values of the second sigmoid layer are combined.

$$i_t = \sigma\,(W_i \,.\, [h_{t-1}, x_t] + b_i), \tag{10}$$

$$C_t = \tanh\,(W_c \,.\, [h_{t-1}, x_t] + b_c), \tag{11}$$

Finally, the data to be sent to the output by the LSTM cell is decided. For this decision, a sigmoid layer works again according to equation (12) and this information is sent to the $tanh$ layer in equation (13).

$$o_t = \sigma(W_o \,.\, [h_{t-1}, x_t] + b_o), \tag{12}$$

$$h_t = o_t * \tanh(C_t), \tag{13}$$

The value that comes to a standard LSTM cell is sent to the next LSTM cell and affects the operations to be performed there. Although the general logic of LSTM is as described, there are also studies in the literature where it is used in different ways.

Gers and Schmidhuber, in 2000 revealed that the size of the time intervals between events conveyed the necessary information for many sequential tasks such as motor control and rhythm detection, but Hidden Markov models ignored this information, while RNN made use of this information. They showed that LSTM is the model that makes the most use of this information and performs better within the RNN architecture. They showed that the peephole powered LSTM can learn the fine distinction between spikes sequences separated by 50 or 49 discrete time steps, without the aid of a short training, and that the new LSTM variant they developed can generate very stable sequences of highly nonlinear, precisely timed spikes. With the new LSTM they have developed, they have made it a promising approach for real-world problems that require time and numbers.

Cho et. al. proposed a new neural network model called RNN Encoder-Decoder, which consists of two RNNs in 2014. One of the RNNs in the RNN Encoder-Decoder represents a symbol sequence as a fixed-length vector, while the other decodes the vector created with the first RNN into a symbol sequence. They jointly trained the encoder and decoder of the proposed model. In this way, they maximized the conditional probability of the target sequence. They showed that the performance of the statistical machine translation system empirically gave better results using the conditional probabilities of the sentence pairs calculated by the RNN Encoder-Decoder as an additional feature in the existing log-linear model. They showed that their proposed model could qualitatively learn a semantically and syntactically meaningful representation of linguistic expressions.

Koutnik et. al., (2014) taking a new approach to complex interdependencies between temporal long inputs in rank estimation and classification, developed the Clockwork RNN (CW-RNN) model. RNNs are theoretically capable of dealing with these temporary dependencies thanks to the short-term memory implemented by the feedback links. But in practice, when long-term memory is required, they are difficult to train successfully. It offers a simple but powerful change to the Simple RNN (SRN) architecture, where the hidden layer is divided into separate modules, each processing inputs at its own temporal granularity and performs calculations only. Rather than complicate standard RNN models, CW-RNN reduces the number of SRN parameters, significantly improving performance on the tasks under test, and speeding up network evaluation. They demonstrated three tasks in their proposed new architecture in preliminary experiments:

- Audio signal generation
- TIMIT spoken word classification where it outperforms both SRN and LSTM networks

- Online handwriting recognition where it outperforms SRNs

Since the development of the LSTM architecture, many variants have been derived. LSTM, which started to become more popular in the 2010s, has been used for various machine learning problems. Greff et. al., in 2016 comparing between these variants developed tested 8 different LSTM variants in three representative tasks. They preferred these three representational tasks as speech recognition, handwriting recognition and polyphonic music modelling. From their work, they optimized the parameters of all LSTM models for each task separately using random search and performed the evaluations with the Analysis of Variance (ANOVA) framework. In their study, they summarized the results of a total of 5400 experimental studies, approximately 15 years of CPU time. The results show that none of the variants can significantly improve the standard LSTM architecture and that the forget gate and output enable function are the most critical components of LSTM.

Jozefowicz et al., (2015), who tested on LSTM architecture and RNN architectures, one of which is to facilitate the training of RNN architecture, which is difficult to train, showed that some RNN architectures work with higher performance than LSTM. They evaluated over ten thousand different RNN architectures and conducted extensive architectural research. And they found an architecture that outperformed both the LSTM and the recently developed Gated Recurrent Unit (GRU) in some, if not all tasks. They showed that adding 1 bias to the forget gate of LSTM bridges the gap between LSTM and GRU.

4. PREDICTION RESULTS

In this section, the application of the methods described in the Time Series Estimation Methods section to the study and the results obtained are explained. Thus, it has been investigated that the experimental period can be shortened to a certain extent with LGB. Mean absolute error (MAE) sensitivity of LGB device (0.2V) and maximum absolute error (MaxAE) were chosen to be twice the sensitivity of LGB device (0.4V). As a result of the tests, it was examined that each tested method could shorten the LGB experiment time with a certain margin of error. However, in case of a measurement under the shortened test times, it is predicted that the error may increase in the sand, silt and clay estimations, which is the main purpose of the LGB device, as well as the margin of error determined for the signal.

4.1. Prediction with Curve-Fitting

As explained in the curve fitting method section, the amplitude value of a certain sample time t can be found by formulating the curves based on the graphs of the measurements in the dataset. In this study, equations such as exponential (1^{st} and 2^{nd} order) and polynomial (many different degrees) with various variations were tried and it was decided that the second-order exponential equation was most suitable for the curves measured by LGB, according to the R-square values of these equations. Figure 4.1 shows a second-order exponential curve superimposed on the measurement curve of a sample soil.

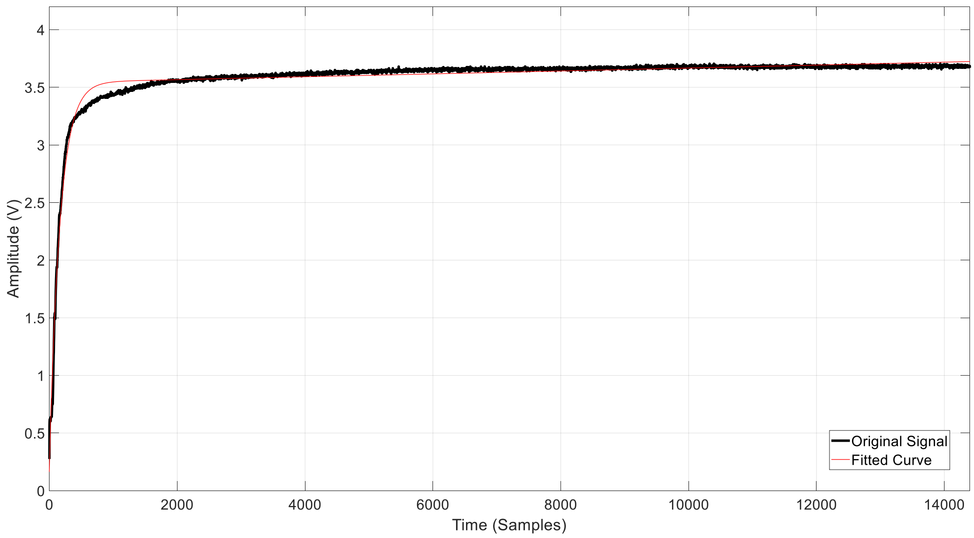

Figure 4.1. Fitted curve on the measurement curve of a sample soil.

The sample soil measurement curve shown in Figure 4.1 and the curve fitted using the 2nd order exponential equation overlap 97.78% according to the R-squared value. Since this curve can be formulated as $f(x) = a * e^{b*x} + c * e^{d*x}$ after this point the coefficients of the curve fitted instead of the actual measurement value of the soil $(a = 3.54), (b = 3.47e - 06), (c = -3.34), (d = -0.01)$ the amplitude value of any sample time t can be found with a small margin of error. For example, when 14.400 (last sampling moment) is given instead of x value in the equation of the curve fitted in the Figure 4.1, the obtained amplitude value is 3.72V. Considering that the 14.400th value of the actual measurement data is 3.76V, the actual measurement data can be captured with a very small margin of error such as 0.04V with the curve fitting method.

Some experiments were carried out on the success of estimating what value the real measurement data of the soil would show after 2 hours (in the 14.400th sample) over the curve to be fitted, without using all of the measurement data of the soil with the above-described processes.

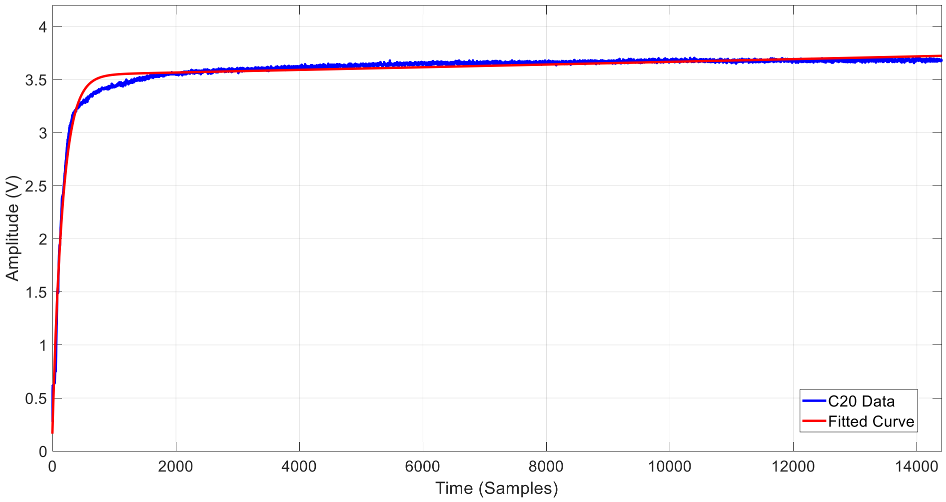

Figure 4.2. C20 labeled soil signal and fitted curve with 14400 samples.

Figure 4.2 shows the curve obtained using the entire C20 labeled soil signal. The R-squared value of the fitted curve with 14400 samples in this soil was calculated as 97.28%. While the average absolute error value between the fitted curve and the soil signal was calculated as 0.028V, the maximum absolute error was calculated as 0.359V. It was observed that the reason for the maximum error to be higher than the average error was due to the fact that the sand particles in the soil sedimented rapidly at the beginning of the experiment, causing a rapidly changing slope on the curve between 0-1000 samples. Since the average and maximum absolute error values are acceptable, a new test was performed by subtracting a certain amount from the end of the C20 soil signal and the result is given in Figure 4.3.

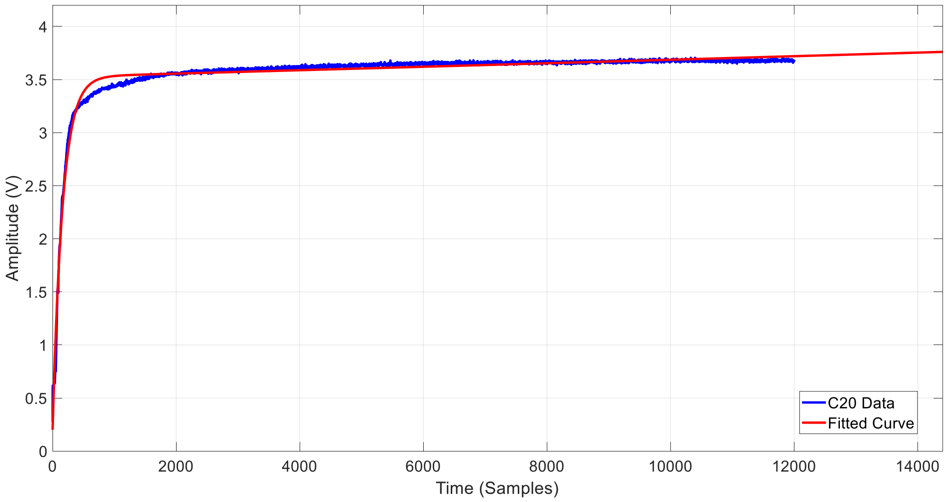

Figure 4.3. C20 labeled soil signal and fitted curve with 12000 samples.

As can be seen in Figure 4.3, 2400 samples from the end of the soil signal labeled C20 are subtracted and a new curve fitting process is performed with the first 12000 samples. When Figure 4.2 and Figure 4.3 are compared, it is seen that the fitted curve is more erroneous. It has been determined that the length of the soil signal has a direct effect on the error values when estimating with the curve fitting method.

In the curve fitting process with 12000 samples, the R-squared value was calculated as 98.13%, the average error value was 0.037V, and the maximum error value was 0.376V. Although the R-square value increased, that is, the fitted curve and the soil signal were more compatible, the reason for the increase in error values was that the conformity of the fitted curve between 12000 samples and 14400 samples with the soil signal decreased. As in the first test, the maximum error value in the second test is between 0-1000 samples. Since the error values are acceptable values, it was foreseen that estimation could be made in a shorter time, and a third test was made by subtracting from the end, and the result is shown in Figure 4.4.

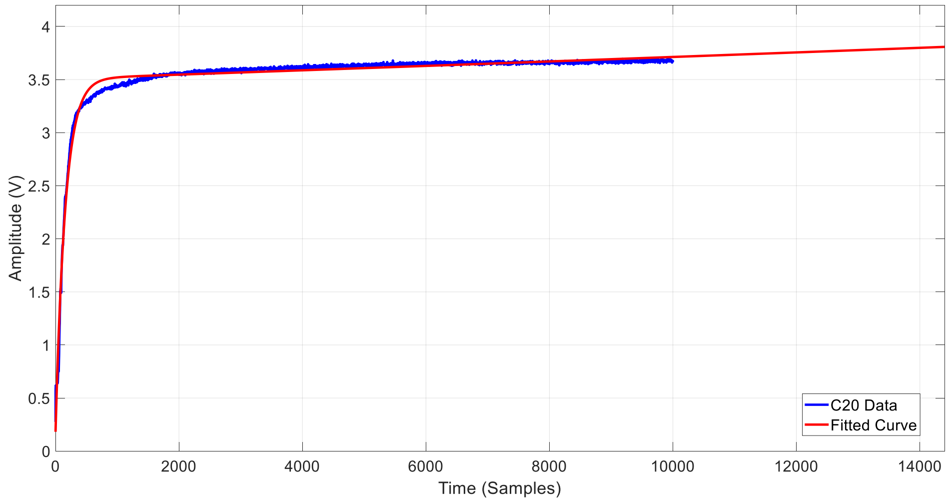

Figure 4.4. C20 labeled soil signal and fitted curve with 10000 samples.

The test result with 10000 samples on the soil signal labeled C20 is given in Figure 4.4. When Figure 4.4. is examined, it is understood that the error values have increased compared to the 12000 sample test performed with the soil signal labeled C20. As a result of the third test, the R-squared value was calculated as 98.39%, while the average absolute error was calculated as 0.04V and the maximum absolute error value as 0.37V. Compared to the previous tests, the sample range with the maximum error in the third test is between 0-1000. It has been observed that the mean absolute error value increases as the length of the soil signal labeled C20 is shortened, but the maximum absolute error remains constant or close to negligible values. As a result of the test with 10000 samples, since the absolute error values were less than the threshold values determined according to the sensitivity of the LGB device, the soil signal was shortened again and the another test was performed. The graph obtained as a result of the fourth test is given in below.

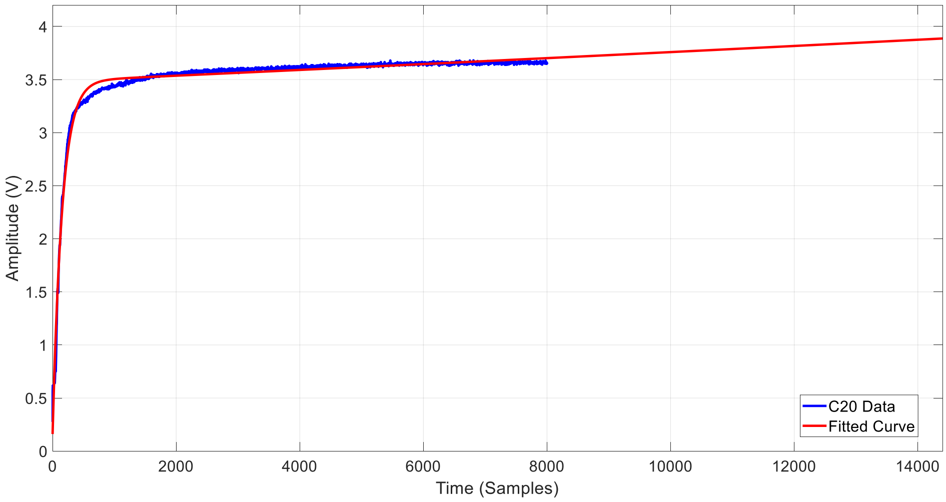

Figure 4.5. C20 labeled soil signal and fitted curve with 8000 samples.

When Figure 4.5 is examined, it is seen that the error between the curve fitted with 8000 samples and the soil signal increased in the fourth test for the soil signal labeled C20. In this test, the R-squared value was calculated as 98.66%, the mean absolute error was 0.07V, and the maximum absolute error was 0.39V. As a result of the tests, it can be said that the R-squared value and the error values increase as the sample size decreases. In other words, as the time shortening process continues, the similarity between the fitted curve and the soil signal decreases and more erroneous estimation is made.

The purpose of the tests was to determine that the soil signal (2-hour recording) could be calculated with an acceptable error rate and to see if it could produce consistent results in shorter times with the LGB device. In this context, the actual measurement data of the experiments carried out with 52 soils were used in the time shortening calculations. At each step, 200 samples were deleted from all signs (approximately 1.5 min) and the variation of absolute errors obtained using curves fitted accordingly is given in Figure 4.6.

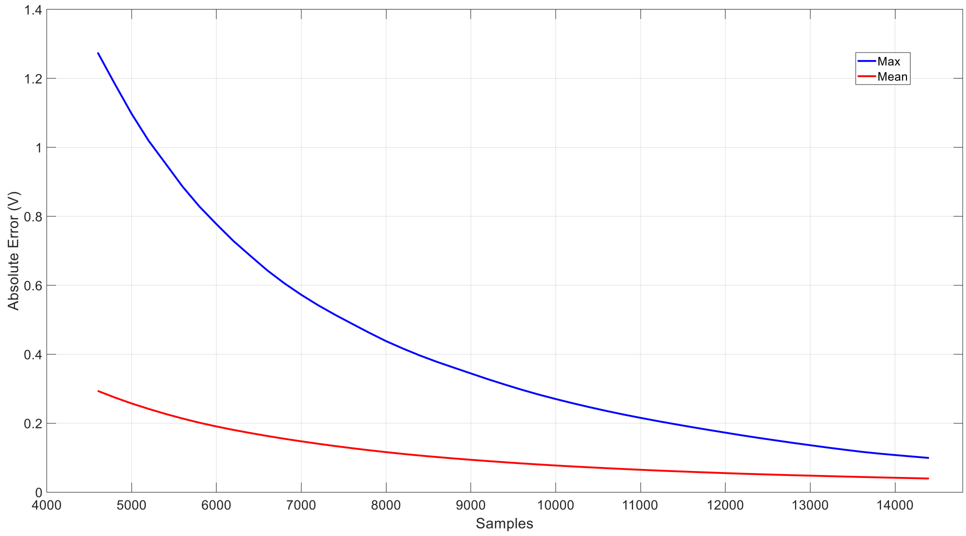

Figure 4.6. Variation of absolute errors obtained by curve fitting method.

Figure 4.6 shows the absolute error obtained at the end of the 2nd hour with curve fitting as the duration of the experiment (measured signal) gets shorter. Errors represent the maximum and mean values of the estimation errors of experiments performed with 52 soil samples. For example, if the experiment is completed with 8.400 samples (1 hour and 10 minutes), the end of 2 hours can be calculated with an average error margin of 0.11V (maximum 0.39V) with the curve to be fitted. As a result of the tests performed with the curve fitting method, it was seen that the 2-hour experiment period could be reduced to 1 hour and 10 minutes.

4.2. Prediction with SVR

The model created in the time shortening tests with SVR is trained differently for each soil signal. Soil signals are divided into training and testing in itself. According to the characteristics of the signs, the train and test rates may vary.

While the sample at time t_i in the signal piece reserved for model training is determined as the input value, the sample at time t_{i+1} is given as the output value. This training continues for the length of the training piece and the training of the SVR model is completed. The input and output values used in SVR training are given in Table 4.1.

Table 4.1. Input and output values used for the SVR train.

Input Values	Output Values
x_0	x_1
x_1	x_2
x_2	x_3
x_3	x_4
. . .	. . .
x_{n-m-1}	x_{n-m}

Relationship between samples training is completed by performing SVR training as shown in Table 4.1. The SVR model, which is ready for testing, is subjected to estimation by shifting 1 time step in the test process, as in the train process. The estimated value for the time t_{n-m+1} is given as the input value for the time t_{n-m+2}, and the entire soil signal is estimated. The absolute error values between the obtained prediction values and the test part of the soil signal are calculated. This process was repeated for all soil signals, and these signals were estimated in a longer time.

The average absolute error and maximum absolute error values obtained as a result of the tests performed are given in Figure 4.7.

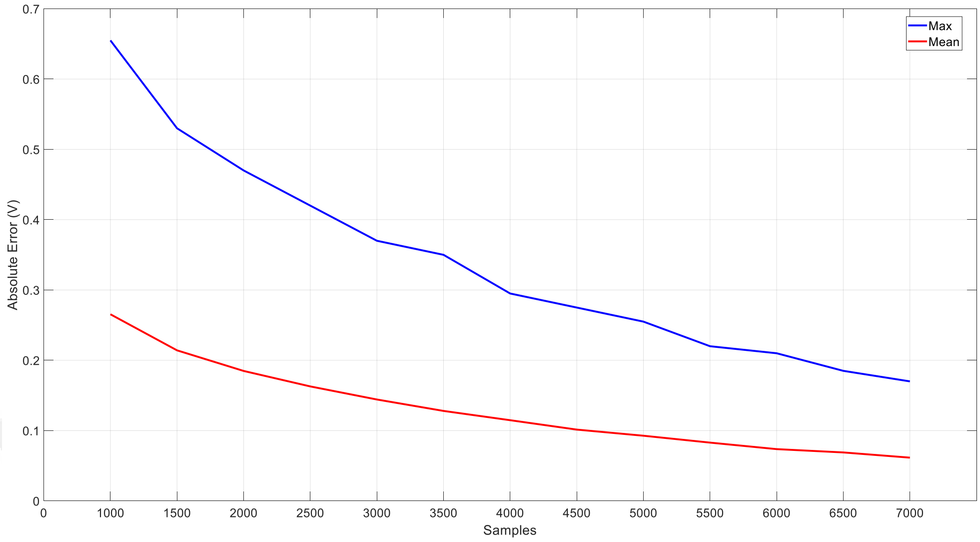

Figure 4.7. Variation of absolute errors obtained by SVR method.

As seen in Figure 4.7, all soil signals were tried to be estimated between 1000 and 7000 samples. It is seen that the absolute error values between the predicted values and the soil signals decrease as the sample size separated as trains increases. It is seen that the absolute error values of the obtained results are acceptable (MAE=0.2V, MaxAE=0.4V) starting from the 3000th sample. In this case, it has been observed that the 2-hour (14400 samples) test time required for the soil texture analysis experiment in the SVR method can be shortened to 25 minutes (3000 samples) with a certain margin of error.

4.3. Prediction with MLP

While making time series forecasting with MLP method, unlike traditional MLP, a signal is separated as train and test in itself. According to the characteristic of the signal, the train and test rates may vary. The train dataset is used for training the MLP, while the test dataset is used to calculate the errors of the predicted values. In Table 4.2, the structure of the dataset allocated during the MLP train process on the $A = [x_0, x_1, x_2, \dots, x_n]$ dataset is given.

Table 4.2. Example input and output values used for the train.

Input Values	Output Values
x_0, x_1, x_2	x_3
x_1, x_2, x_3	x_4
x_2, x_3, x_4	x_5
x_3, x_4, x_5	x_6
x_4, x_5, x_6	x_7
. . .	. . .
$x_{n-m-3}, x_{n-m-2}, x_{n-m-1}$	x_{n-m}

As shown in Table 4.2, the dataset separated as a train is divided into segments of certain lengths within itself. While this segment is given as input values to MLP architecture, the first value after the segment is given as output value. In this case, the input layer of the MLP is the size of the separated segment and the output layer is 1. Neural networks working in this way are called sequence-to-one (seq2one) regression in the literature. This process was repeated until the last value of the part separated as a train by shifting the segment 1 time step. In this way, the training of the MLP structure was completed with the supervised learning method. In the estimation phase, input, output and actual values are given below.

Table 4.3. Input, output and absolute error formulas used in testing.

Input Values	Output Values	Absolute Error (AE)
$x_{n-m-2}, x_{n-m-1}, x_{n-m}$	$\hat{y}_{n-m+1}$	$\lvert y_{n-m+1} - \hat{y}_{n-m+1}\rvert$
$x_{n-m-1}, x_{n-m}, \hat{y}_{n-m+1}$	$\hat{y}_{n-m+2}$	$\lvert y_{n-m+2} - \hat{y}_{n-m+2}\rvert$
. . .	. . .	. . .
$\hat{y}_{n-3}, \hat{y}_{n-2}, \hat{y}_{n-1}$	$\hat{y}_n$	$\lvert y_n - \hat{y}_n\rvert$

As shown in Table 4.3, while estimating the soil signals, 1 time post step was estimated in each iteration. The estimated y value at the end of each iteration was given as the last value of the input segment in the next iteration, and the segment length was kept constant by deleting the first value of the segment. In this way, the whole signal was estimated with a step size of 1 time step. The MLP architecture used in the study is designed as 1 input, 1 hidden and 1 output layer. The hidden layer has 100 neurons, while the output layer has a single neuron, the input layer varies according to the segment length. The shorter the length of the segments to be given as input values, the longer the MLP training process. In order to find the optimal segment length, the segment length was tested on 5 soil signals by increasing the segment length by 50 samples during each test in the [50;1000] sample range. In the tests performed, the train and test datasets were set to 50%. In other words, training made with the first 7200 samples and the remaining 7200 samples were estimated. The absolute error variation with segment length obtained with the results of the test is given in Figure 4.8.

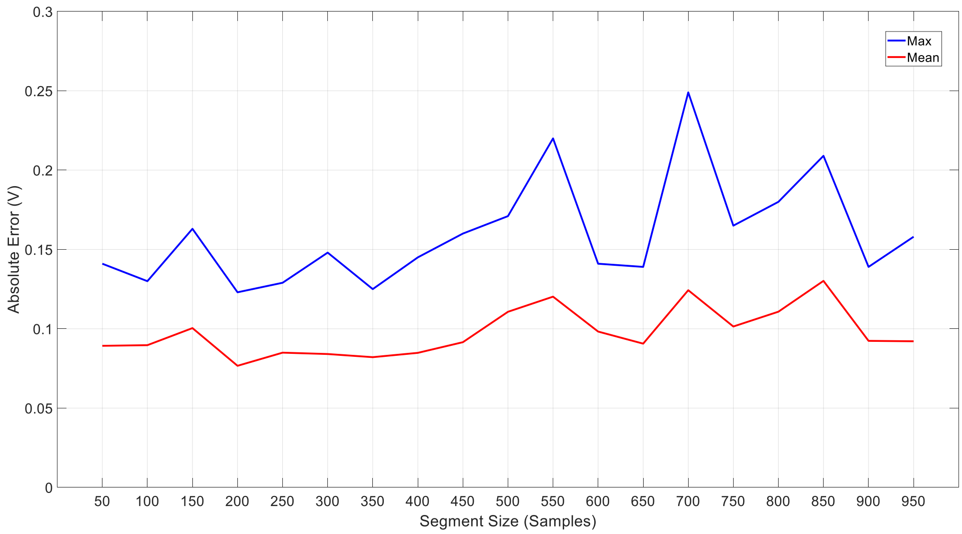

Figure 4.8. Average and maximum absolute error variation with segment lenghts.

There is a difference between segment lengths and successful estimation shown in Figure 4.8. As the segment length or the number of training repetitions increases in MLP, the prediction success increases. However, when estimating in time series, as the segment length increases, the estimation success does not change according to a certain pattern, since the MLP trains with fewer iterations. In other words, the given graphic also shows the change in success between the number of inputs in MLP and the training repetition. Since the segment length that can be estimated most successfully is 200 samples, the input size was chosen as 200 samples in the estimation study with the MLP method. Unlike the curve fitting method, instead of shortening the last part of the real curve (14400th sample) for time shortening, estimation was made by increasing 500 samples from the 1000th sample. The absolute error variation graphs obtained as a result of the tests are given in below.

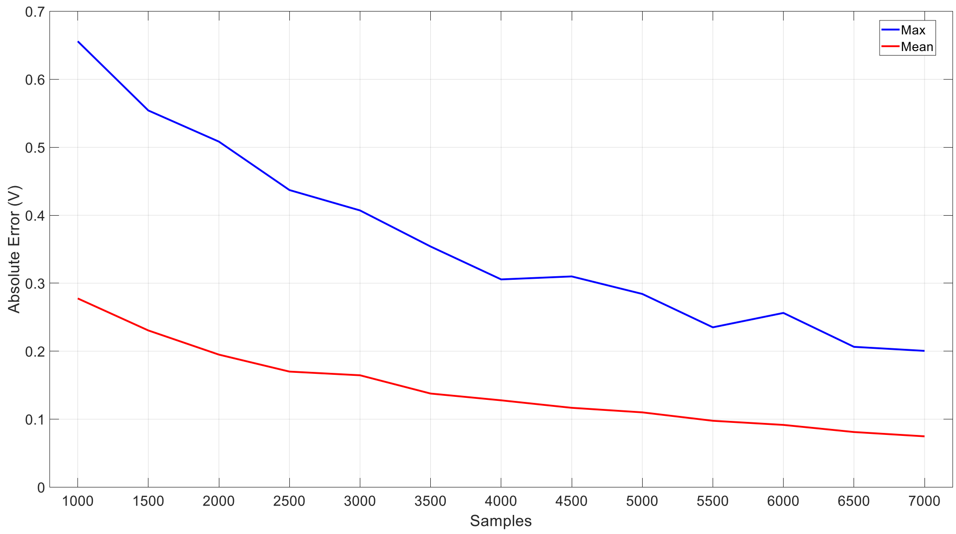

Figure 4.9. Average and maximum absolute error variation with segment lenghts.

When Figure 4.9 is examined, it is seen that the acceptable absolute errors are in the 3000th sample in order to shorten the soil texture analysis experiment time as a result of the tests performed with 200 samples input. As a result of the tests performed with MLP, it was seen that the 2-hour soil texture analysis time could be shortened to 25 minutes.

4.4. Prediction with LSTM

In the LSTM method, as in the MLP method, the signals are divided into two as train and test. The train dataset is used for training the LSTM architecture, while the test dataset is used to calculate the absolute errors of the predicted values. Like the MLP method, it establishes the connection between the first input and the second input, thanks to the forget gate in LSTM. With the forget gate, the test dataset inputs can be connected each other and given to the LSTM structure. Since the LSTM method is a method that is dependent on standardization due to its structure, the dataset has been standardized. The standardization process was carried out according to equations (14) – (16).

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{14}$$

$$std\_dev = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \tag{15}$$

$$dataset\_standardized = \frac{\sum_{i=1}^{n} x_i - \mu}{std\_dev}, \tag{16}$$

Equation (14) calculates the mean of the dataset (μ), while equation (15) calculates the standard deviation of the dataset. The standardization process of the data is completed by subtracting the μ value from all the data and dividing by the standard deviation value with equation (16). The general view between input and output for training LSTM architecture is given below.

Table 4.4. Inputs and outputs used for LSTM model training.

Input Values	Output Values
$x_0, \dots, x_{199}$	x_{200}
$x_1, \dots, x_{200}$	x_{201}
$x_2, \dots, x_{201}$	x_{202}
$x_3, \dots, x_{202}$	x_{203}
$x_4, \dots, x_{203}$	x_{204}
. . .	. . .
$x_{n-m-200}, \dots, x_{n-m-1}$	x_{n-m}

As shown in Table 4.4, the LSTM structure is trained as 200 input and 1 output, and the connection between the inputs is determined by the forget gate. After the training process was completed, the estimation process was done with the same logic. The estimated value $\widehat{y_i}$ at time t_i is used as the input value for time

t_{i+1}. After all values have been estimated, these values are unstandardized by equation (17).

$$y_i = (\hat{y}_i * std\_dev) + \mu, \tag{17}$$

The actual values of the $\hat{y}_i$ values estimated by equation (17) were calculated. The general state of the estimation process is given in Table 4.5.

Table 4.5. Input, output ve absolute error formulas used for testing in LSTM method.

Input Values	Output Values	Absolute Error (AE)
$x_{n-m-200}, \cdots, x_{n-m}$	$\hat{y}_{n-m+1}$	$\lvert y_{n-m+1} - \hat{y}_{n-m+1}\rvert$
$x_{n-m-199}, \cdots, \hat{y}_{n-m+1}$	$\hat{y}_{n-m+2}$	$\lvert y_{n-m+2} - \hat{y}_{n-m+2}\rvert$
$x_{n-m-198}, \cdots, \hat{y}_{n-m+2}$	$\hat{y}_{n-m+3}$	$\lvert y_{n-m+3} - \hat{y}_{n-m+3}\rvert$
. . .	. . .	. . .
$\hat{y}_{n-201}, \cdots, \hat{y}_{n-1}$	$\hat{y}_n$	$\lvert y_n - \hat{y}_n\rvert$

When the formulas given in Table 4.5 are examined, it is seen that there are many similarities between the LSTM structure and the MLP structure. The difference between MLP and LSTM is that LSTM creates the connection itself between the inputs it receives sequentially. For this reason, there is no need to segment the train dataset in tests performed with LSTM. The LSTM architecture used for the estimation process is given below.

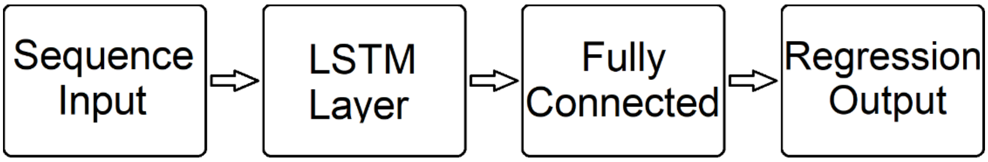

Figure 4.10. LSTM architecture used in thesis.

According to Figure 4.10, there is a sequence input layer, an LSTM hidden layer with 200 neurons, and a regression layer with 1 output. The LSTM hidden layer and the regression layer are fully connected. The LSTM estimation results made according to the given architecture are given in Figure 4.11.

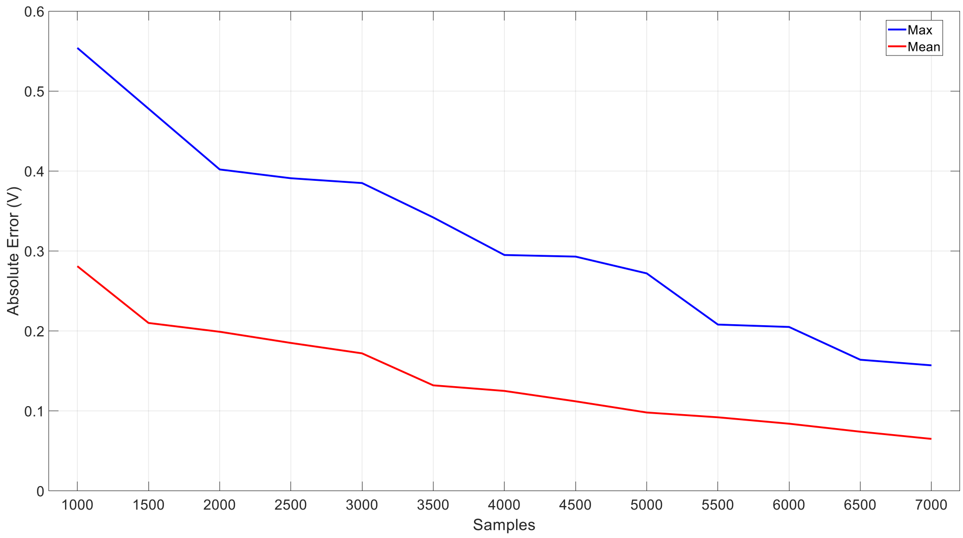

Figure 4.11. Variation of absolute errors obtained by LSTM method.

When Figure 4.11 is examined, it is seen that the maximum and average absolute errors calculated from the 2000th sample as a result of the tests performed with the LSTM method are smaller than the acceptable values. In this case, it has been observed that the 2-hour LGB test time can be shortened to 16.5 minutes (2000 samples) with the LSTM method.

4.5. Estimation with SVR

In this section, which aims to estimate the sand, silt and clay ratios of the soils tested with LGB. One of the successful and popularly known SVR in the literature has been used. Regression analyzes were performed for each of the three soil components (sand, silt and clay) separately. Therefore, in the datasets prepared separately for each soil component, the output information is in the form of a single attribute showing the ratio of the relevant soil component. In other words, the input part of the main dataset is prepared with 10 features extracted from the signals obtained from all the replica experiments of 52 soils, the sand ratios for the sand regression experiment are added to these inputs and the machine learning method is applied. The same data preparation is then repeated for silt and clay.

In the regression analysis study, the SVR method, which does not need parameter adjustment, works faster than modern regression methods such as ANN, and does not need to repeat the training due to the optimization method it uses, was used. In Figure 4.12, the mean absolute error (MAE) values for sand, silt and clay estimations for 52 soils calculated with SVR are given.

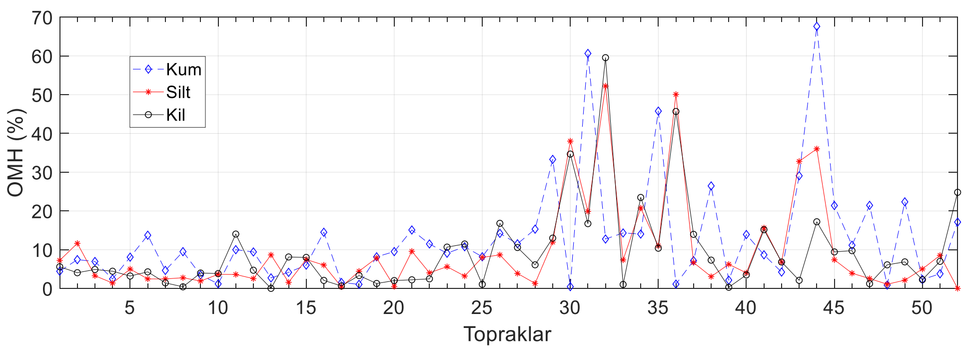

Figure 4.12. MAE values estimations obtained with SVR

In Figure 4.12, the absolute error rates obtained as a result of the experiments performed for all soils with the SVR are shown separately for sand, silt and clay. Considering the 10% error rate, which was also taken into account in the hydrometer tests, the error rate was 28 for sand, 41 for silt and 36 for clay, with a maximum error rate of 10%. Another remarkable point in the figure is that the component estimates of some soils are realized with very high error. When these soils with a high error rate were examined, it was understood that there were unique samples in the data set, so it was determined that high error estimation for these soils was a natural result.

The results obtained by the time shortening methods have been predicted to be able to predict the components in the soil in a shorter time thanks to the successful results. Based on this prediction, SVR has been studied to predict soil components in a shorter time. In the case of a measurement shorter than 70 minutes for the curve fitting method, 25 minutes for the SVR method, 25 minutes for the MLP method, and 16.5 minutes for the LSTM method, the times determined in the time shortening methods are less than the acceptable error rate, as well as the error in predicted sand, silt and clay estimations to increase. The mean absolute error change in the estimation of the soil components of the time shortening with SVR on 52 soil dataset used in the duration shortening tests is presented in Table 4.6.

Table 4.6. The effect of time shortening on the estimation of soil components.

Time (Samples)	Clay MAE	Silt MAE	Sand MAE
14400	5,97	4,48	5,66
7200	6,07	4,47	5,64
3600	6,04	4,13	5,54
1800	6,26	4,02	5,48
900	6,59	3,98	5,45
400	7,16	3,60	6,25
200	7,90	3,66	6,84
80	6,58	3,75	8,92

According to Table 4.6, 900 samples (450 seconds) for sand, 400 samples (200 seconds) for silt and 14,400 samples (2 hours) for clay have the lowest MAE. When Table 4.6 is examined, it is seen that 14400 samples are required for the most accurate estimation. However, the clay ratio causes the required number of samples. Here, it was observed that it would be beneficial to reduce the test time to 3,600 samples (30 minutes) with a small loss (with a margin of error of 0.07) for the clay component. In this way, it has been observed that the soil components can be calculated with an acceptable margin of error, while on the other hand, the test period can be shortened by 75%. Therefore, after the first 450 seconds (900 samples) of the soil signal, it seems to be used only to better determine the clay ratio.

5. RESULTS AND DISCUSSIONS

With today's technology, all fields have developed and continue to develop. Computers and machine learning are among the main reasons for this development. As in other fields, agriculture is one of the fields that develop with computers and machine learning. Soil and soil type, which are the most important resources of agriculture, are the most important factors for the quality of the agricultural product to be produced. The LGB device blended soil analysis and machine learning and gave successful results in determining soil type with inexpensive and easy-to-find equipment.

In this thesis, however, it is aimed to shorten the soil analysis experiment time with LGB device. Tests were performed to estimate the final sample (14400^{th} sample) using curve fitting, SVR, MLP and LSTM methods for experimental time shortening with an acceptable margin of error. In the tests, signals obtained from 52 different soil samples analyzed with LGB device were used. The sensitivity of the LGB device was chosen as 0.2V acceptable MAE and 0.4V as MaxAE. In addition, as a result of time shortening studies, the SVR method was used for the estimation of soil components.

For the curve fitting method, the signals obtained from the soil analysis experiments were cut from the last 200 samples (approximately 1.5 minutes) to find the shortest piece of marker. In the curve fitting method, the second-order exponential function with the most suitable characteristic for the soil signals with an average R-square value of 0.9788 was chosen. The shortest sign obtained with this method is 8400 samples, and it has been observed that the 2-hour LGB experiment period can be shortened to 1 hour and 10 minutes.

In the SVR method, all soil signals are divided into train and test. SVR model is selected as single input and single output. In the SVR method, the signals are estimated by extending 500 samples from the first 1000 samples to 7000 samples. As a result of the tests, it was seen that a 3000-sample soil signal would be sufficient for the estimation of the 14400th sample. In this case, it is understood that the soil analysis test time can be shortened to approximately 25 minutes.

In the MLP method, the signals are divided into train and test. MLP architecture is selected as 1 input layer with 200 inputs, 1 hidden layer with 100 neurons, and 1 output layer. In order to determine the number of neurons in the input layer, experiments were carried out on 5 randomly selected soil signals, up to 50 inputs and 950 inputs, and the lowest average error was obtained with the 200-input MLP architecture. In the MLP method, the estimation errors are calculated by extending the signals from the beginning instead of shortening them from the end. As a result of the tests, it has been seen that the 14400th sample can be estimated with a 3000-sample soil signal. In this case, it is understood that the soil analysis test time can be shortened to approximately 25 minutes.

In the LSTM method, which is another machine learning method, each soil signal is divided into test and train in itself, as in the MLP method. Tests were carried out with the LSTM architecture, which was determined as a single-input sequence layer, a 200-neuron LSTM layer and a fully connected layer, and a single-output regression layer. Thanks to the forget gate in the LSTM architecture, the effects of the inputs on each other did not need to be changed in the size of the input layer. As a result of the tests performed with the LSTM method, soil texture analysis could be performed with a 2000-sample signal and soil texture analysis experiment time can be shortened to 16.5 minutes.

Curve fitting, SVR, MLP and LSTM methods were used for the time shortening study. MAE and MaxAE values of these methods were calculated. The comparison of the results obtained in the time shortening study is given in Table 5.1.

Table 5.1. The comparison of the time shortening results.

Signal Size (Samples)	Curve Fitting		SVR		MLP		LSTM	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max
1000	-	-	0.27	0.66	0.28	0.66	0.28	0.55
1500	-	-	0.21	0.53	0.23	0.55	0.21	0.48
2000	-	-	0.19	0.47	0.20	0.51	0.20	0.40
2500	-	-	0.16	0.42	0.17	0.44	0.18	0.39
3000	-	-	0.14	0.37	0.16	0.41	0.17	0.38
3500	-	-	0.13	0.35	0.14	0.35	0.13	0.34
4000	-	-	0.12	0.30	0.13	0.31	0.12	0.30
4500	0.29	1.28	0.10	0.28	0.12	0.31	0.11	0.29
5000	0.26	1.10	0.09	0.26	0.11	0.28	0.10	0.27
5500	0.22	0.92	0.08	0.22	0.10	0.24	0.09	0.21
6000	0.20	0.78	0.07	0.21	0.09	0.26	0.08	0.20
6500	0.18	0.66	0.07	0.18	0.08	0.21	0.07	0.17
7000	0.15	0.57	0.06	0.17	0.07	0.20	0.06	0.16
7500	0.13	0.50	-	-	-	-	-	-
8000	0.12	0.44	-	-	-	-	-	-
8500	0.10	0.39	-	-	-	-	-	-
9000	0.09	0.34	-	-	-	-	-	-
9500	0.08	0.31	-	-	-	-	-	-
10000	0.08	0.27	-	-	-	-	-	-
10500	0.07	0.24	-	-	-	-	-	-
11000	0.07	0.22	-	-	-	-	-	-
11500	0.06	0.19	-	-	-	-	-	-
12000	0.06	0.17	-	-	-	-	-	-
12500	0.05	0.15	-	-	-	-	-	-
13000	0.05	0.14	-	-	-	-	-	-
13500	0.05	0.12	-	-	-	-	-	-

14000	0.04	0.11	-	-	-	-	-	-
14400	0.04	0.10	-	-	-	-	-	-

As can be seen in Table 5.1, no test was performed for soil signals shorter than 4500 samples for curve fitting method and more than 7000 samples for machine learning methods. The reason why the test is not performed on shorter soil signals in the curve fitting method is that the results obtained with 4500 samples are much higher than the acceptable absolute error values. In machine learning methods, tests were not continued on soil signals with more than 7000 samples, due to the presence of the shortest signal segment that can be estimated.

It has been predicted that the soil component ratios can be determined in a shorter time with the results obtained by the time shortening studies. Based on this, an estimation study was made on soil components with SVR and it was understood that 900 samples were sufficient for sand and silt estimation, and a 2-hour soil signal was needed for clay. Mean absolute error values obtained from soil particle estimation tests with SVR are given in Figure 5.1.

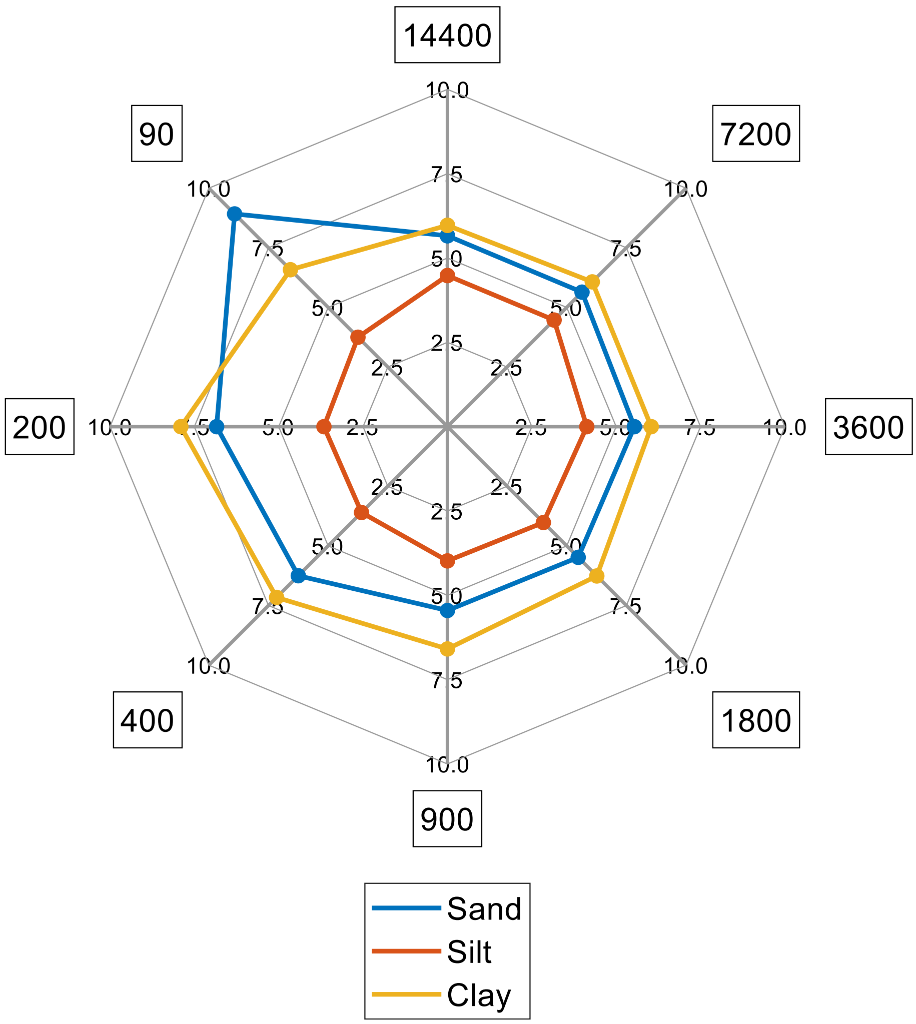

Figure 5.1. MAE values for sand, silt and clay estimates obtained with SVR.

When Figure 5.1 is examined, it was concluded that the clay estimation could be shortened to 3600 samples with a small margin of error, and in this case, a 30-minute experiment time would be sufficient for the determination of soil components ratios.

With this thesis, it was seen that the soil texture analysis time could be shortened. In machine learning methods used for time series estimation, more successful results can be obtained by increasing the number of hidden layers and/or the number of neurons in the hidden layer. It is predicted that the Bidirectional Long Short-Term Memory (BiLSTM) method, which is another machine learning method used for time series forecasting, can provide more successful results and shorten the duration of the soil texture analysis experiment, thanks to both its feed-forward and feedback structure. Soil component estimation studies can be enriched, and higher success can be achieved by using architectures such as MLP, LSTM, which are among other machine learning methods with more complex architecture.

REFERENCES

AFNOR (Association Française de Normalisation), 1979. Granulome´trie. Analyse granulome´trique des poudres fines sur tamiseuse a` de´pression d'air. Doc. NF-X-11- 640; and subsequent documents: 1981a: NF-X-11-683; 1981b: NFX-11-630; 1982: NF-X-11-642; 1983c: NF-X-31-107; 1990: NF-X-11-661. Paris: AFNOR.

ALLEN, T., DAVIES, R., and SCARLETT, B., 1996. Particle Size Measurement. (2 vols.) London: Chapman and Hall.

ATTENBERG, A., 1916. Die Klassifikation der humusfreien und der humusarmen Mineralboden Schwedens nach den Konsistenzverhaltnissen derselben. Int. Mitt. Bodenkd. 6: 27–37.

AVERY, B. W., and BASCOMB, C. L., 1982. Soil Survey Laboratory Methods. Tech. Monogr. No. 6, Soil Survey. Harpenden, U.K., p. 18 –19.

BARTH, H. G., and SUN, S. T., 1985. Particle size analysis. Analytical Chemistry. 57: 151R-175R.

BOSER, B. E., GUYEN, I. M. and VAPNIK, V. N. 1992. "A training algorithm for optimal margin classifiers". Pittsburgh: Proceedings of the fifth annual workshop on Computational learning theory, 144–152.

BOX, G.E.P., JENKINS G.M., REINSEL, G.C. and LJUNG G.M., "Time Series Analysis: Forecasting and Control", John Wiley & Sons, Inc., Hoboken, New Jersey, United States of America, 712 pages.

BSI (British Standards Institution), 1981. Code of Practice for Site Investigation, Standard BS5930; and subsequent Standards: 1993: BS3406 (Part 4); 1998: BS7755 –5.4. London.

BURT, R., REINSCH, T. G., and MILLER, W. P., 1993. A micro-pipet method for water-dispersible clay. Commun. Soil Sci. Plant Anal. 24: 2531–2544.

CARONI, E., and MARAGA, F., 1983. Misure granulometriche in alvei naturali con un compasso registratore. Geol. Appl. Idrogeol. 18: 19 –31.

CARTER, M. R., 1993. Soil Sampling and Methods of Analysis. Boca Raton, FL: Lewis Publishers.

CHO, K., MERRIENBOER, B.V., GÜLÇEHRE, Ç., BAHDANAU, D., BOUGARES, F., SCHWENK, H., and BENGIO, Y., 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.

CORTES, C. and VAPNIK, V., 1995. "Support-vector networks", Machine Learning, 20, 273 – 297.

DIN (Deutsches Institut für Normung), 1983. Partikelgrossenanalyse: Sedimentationanalyse in Schwerefeld; Pipette-Verfahren. DIN 66115, Berlin.

GEE, G. W., and BAUDER, J. W., 1986. Particle size analysis. In: Methods of Soil Analysis, Part I, 2nd ed. Madison, WI: Am. Soc. Agron. p. 383 – 411.

GERS, F.A., and SCMIDHUBER, J:, 2000. Recurrent nets that time and count. International Joint Conference on Neural Networks (IJCNN).

GREFF, K., SRIVASTAVA, R. K., KOUTNIK, J., STEUNEBRINK, B. R. and SCHMIDHUBER, J., 2016, LSTM: A search space odyssey, IEEE Transactıons On Neural Networks And Learnıng Systems, Vol. 28, No. 10, p. 2222-2232.

GRIFFITHS, J. C., 1967. Scientific Method in Analysis of Sediments.New York: McGraw-Hill.

GROUT, H., TARQUIS, A. M., and WIESNER, M. R., 1998. Multifractal analysis of particle size distributions in soil. Environ. Sci. Technol. 32: 1176 –1182.

GUYON, I., BOSER, B., VAPNIK, V., 1993. "Automatic capacity tuning of very large VC-dimension classifiers". Advances in Neural Information Processing Systems 5, 147–155.

HEAD, K. H., 1992. Manual of Soil Laboratory Testing, Vol. I. Soil Classification and Compaction Tests. London: Pentech Press.

HUYEN, N.T., HA, D.D. and HIE, P.T., “Buried objects detection in heterogeneous environment using UWB systems combined with curve fitting method”, ICT Express, Accepted on 23 June 2020.

ISO (International Standards Organisation), 1973. Guide to the Choice of Series of Preferred Numbers and Series Containing More Rounded Values of Preferred Numbers. Doc. ISO-497-1973-(E) (1st ed.); and subsequent documents: 1977: ISO- 387-(E) (1st ed.); 1998: ISO-11277-1-(E) (1st ed.). Geneva.

JOZEFOWICZ, R., ZAREMBA, W. and SUTSKEVER, I., 2015. An Empirical Exploration of Recurrent Network Architectures. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2342-2350.

KLUTE, A., 1986. Methods of Soil Analysis, Part I. Physical and Mineralogical Methods. 2nd ed. Madison, WI: Am. Soc. Agron.

KOUTNIK, J., GREFF, K., GOMEZ, F. and SCHMIDHUBER, J., 2014, A Clockwork RNN, Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1863-1871

LAXTON, J. L., 1980. A method for estimating the grading of boulder and cobble grade material. In: IGS Short Communications. Report Inst. Geol. Sci. No 80/1. London: HMSO. p. 31–35.

MAEDA, T., TAKENAKA, H., and WARKENTIN, B. H., 1977. Physical properties of allophane soils. Adv. Agron. 29: 229 –264.

MASUM, S., LIU, Y. and CHIVERTON, J., 2018. Multi-step Time Series Forecasting of Electric Load Using Machine Learning Models. Springer International Publishing AG, part of Springer Nature 2018, ICAISC 2018, LNAI 10841, pp. 148–159.

METZ, R., 1985. The importance of maintaining horizontal sieve screens when using a RoTap. Sedimentology 32: 613 – 614.

MOGHADAM, I.P., AFRAND, M., HAMAD, S.M., BARZİNJY, A.A. and TALEBİZADEHSARDARİ, P., “Curve-fitting on experimental data for predicting the thermal-conductivity of a new generated hybrid nanofluid of graphene oxide-titanium oxide/water”, Physica A: Statistical Mechanics and its Applications, 548, 2020, 122140.

NADEAU, P. H., 1985. The physical dimensions of fundamental clay particles. Clay Minerals 20: 449 –514.

NADEAU, P. H., WILSON, M. J., McHARDY, W. J., and TAIT, J. M., 1984. Interparticle diffraction: A new concept of interstratification of clay minerals. Clay Mineral. 19: 757–770.

ODÉN, S., 1915. Eine neue Methode zur mechanischen Bodenanalyse. Int. Mitt. Bodenanal. 5: 257–311.

ORHAN, U. ve KILINÇ, E., "Estimating soil texture with laser-guided Bouyoucos", Automatika, 61:1, 1-10, 2020.

RINGROSE-VOASE, A., and BULLOCK, P., 1984. The automatic recognition and measurement of soil pore types by image analysis and computer programs. J. Soil Sci. 35:673 – 684.

RITTER VON RITTINGER, P., 1867. Lehrbuch der Aufbereitungskunde. Berlin: Ernst & Korn Verlag.

SCHÖLKOPF, B., BURGES, C. and VAPNIK, V., 1995. “Extracting support data for a given task”. Menlo Park, CA: First International Conference on Knowledge Discovery & Data Mining, AAAI Press.

SCHÖLKOPF, B., BURGES, C. and VAPNIK, V., 1996. “Incorporating Invariances in Support Vector Learning Machines”. Bochum: Artificial Neural Networks - ICANN 96.

SSCHÖLKOPF, B. and SMOLA, A. 2002. “Support Vector Machines and Kernel Algorithms”, Encyclopedia of Biostatistics, 5328-5335.

SHABBIR, S., SATYANARAYANA, B. and SREERAMULU, K., “Characterization of Hyperelastic Material by Experimental Tests and Curve Fitting”, IConAMMA, 2018.

SHEN, F., CHAO, J. and ZHAO, J., 2015. Forecasting exchange rate using deep belief networks and conjugate gradient method. National Key Laboratory for Novel Software Technology, and Department of Computer Science and Technology, Nanjing University, China. Neurocomputing 167 (2015) 243–253.

SHIOZAWA, S., and CAMPBELL, G. S., 1991. On the calculation of a mean particle diameter and standard deviation from sand, silt and clay fractions. Soil Science 152:427– 431.

SMITH, K., 2001. Soil and environmental analysis physical methods. (2.nd ed.). New York: Marcel Dekker.

STABINGER, H., LEOPOLD, H., and KRATKY, O., 1967. Eine neue Präzisionsmethode zur Bestimmung der Dichte von Flüssigkeiten. Monatsh. Chem. 98:436 – 438.

SUR, H. S., and KUKAL, S. S., 1992. A modified hydrometer procedure for particle size analysis. Soil Science 153:1– 4.

SYVITSKI, J. P. M., 1991. Principles, Methods, and Applications of Particle Size Analysis. Cambridge, U.K.: Cambridge Univ. Press.

TANNER, C. B., and JACKSON, J. J., 1947. Nomographs of sedimentation times for soil particles and gravity or centrifugal acceleration. Soil Sci. Soc. Am. Proc. 12:60 – 65.

TOVEY, N. K., and SMART, P., 1982. Electron Microscopy of Soils and Sediments: Techniques. Oxford: Clarendon Press.

TYLER, S. W., and WHEATCRAFT, S. W., 1992. Fractal scaling of soil particlesize distributions: Analysis and limitations, Soil Sci. Soc. Am. J. 56:362–369.

USDA (United States Department of Agriculture), 1996. Soil Survey Laboratory Methods Manual. Soil Survey Investigations Report No. 46, Version 3.0. Washington, D.C.: Soil Conservation Service.

VAPNIK, V., GOLOWICH, S. E. and SMOLA, A. 1997. "Support vector method for function approximation, regression estimation and signal processing". Cambridge: Advances in Neural Information Processing Systems 9, 281-287.

VAPNIK, V. and LERNER, A. 1963. "A Pattern Recognition Using Generalized Portrait Method", Automation and Remote Control, 24(6), 774 – 780.

WANG, J. and FENG, L., "Curve-fitting models for fossil fuel production forecasting: Key influence factors", Journal of Natural Gas Science and Engineering, 32, 2016, pp. 138-149.

CURRICULUM VITAE

Ferhat ALBAYRAK was born in Antalya. He completed his elementary education at Namık Kemal İlköğretim Okulu and highschool education at Antalya Lisesi. He received BSc degree from the department of Computer Engineering of Sakarya University in 2015. Since 2020, he has been working as as a research assistant at Computer Engineering Department of Çukurova University in Adana.