Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 3079 – 3093

doi:10.3906/elk-1905-179

Modeling compaction parameters using support vector and decision tree regression algorithms

Abdurrahman ÖZBEYAZ[1,*], **Mehmet SÖYLEMEZ**[2]

[1]Department of Electrical and Electronic Engineering, Faculty of Engineering, Adıyaman University, Adıyaman, Turkey

[2]Department of Civil Engineering, Faculty of Engineering, Adıyaman University, Adıyaman, Turkey

Received: 31.05.2019 • **Accepted/Published Online:** 16.10.2019 • **Final Version:** 25.09.2020

Abstract: Shortening the periods of compaction tests can be possible by analyzing the data obtained from previous laboratory tests with regression methods. The regression analysis applied to current data reduces the cost of experiments, saves time, and gives estimated outputs. In this study, the MLS-SVR, KB-SVR, and DTR algorithms were employed for the first time for the estimation of soil compaction parameters. The performances of these regression algorithms in estimating maximum dry unit weight (MDD) and optimum water content (OMC) were compared. Furthermore, the soil properties (fine-grained soil, sand, gravel, specific gravity, liquid limit, and plastic limit) were employed as inputs in the study. The data used for the study were supplied from the experimental soil tests from small dams in Niğde, a province in the southern part of Central Anatolia, Turkey. Polynomial-based KB-SVR yielded the best R-values with 0.93 in the prediction of both OMC and MDD. Moreover, in the multioutput estimation model, polynomial and RBF-based KB-SVR methods were successful with 0.98 and 0.99, respectively. Additionally, while the MSE value was 1.33 in the estimation of OMC, this value was 0.04 in the estimation of MDD. Accordingly, MDD was the most successfully estimated parameter in all processes. It was concluded that through the algorithms used in this study, the prediction of soil compaction parameters could be possible without the need for further laboratory tests.

Key words: Regression, compaction, soil index parameters, maximum dry unit weight, optimum water content, support vector machine, decision tree

1. Introduction

Compaction is the process of increasing soil density by applying mechanical energy, thus leading to the expulsion of air voids in the soil. This, in turn, produces an increase in the shear strength and a decrease in the consolidation and permeability of soils [1]. Compaction is widely used in many important engineering projects such as roads, airfields, earth dams, and landfill construction. With the help of the compaction process, voids under static and dynamic loads on floors are reduced; wear can be reduced or delayed; liquefaction features can be eliminated; volume changes due to frost, swelling, shrinkage, etc. can be controlled by reduced permeability; and the soil can be provided with a more stable structure. While laboratory density is determined by performing the Proctor compaction test on several soil samples with different water contents, the determination of the maximum dry unit weight (MDD) and optimum water content (OMC), which is a time-consuming process, requires a considerable amount of material and expert operators in the laboratory.

*Correspondence: aozbeyaz@adiyaman.edu.tr

Regression analysis is used for determining a cause-and-effect relationship between two or more variables, and it is employed to make estimations about the subject of a study using the relationships among those variables. It is possible to find cause-and-effect relationships between, for example, income and expense, age and height, or grade and average by this method. In the regression analysis method, a mathematical model, called the regression model, is employed to determine the relationship between two or more variables. When we examined the published literature, we observed several studies predicting MDD and OMC [2–16]. However, there are no studies comparing regression analysis algorithms with each other. In one study, compaction parameters were estimated from soil types using regression analysis in a study [9]. In this study, only an artificial neural network (ANN) was employed. In another study, an ANN was also used for predicting OMC and MDD [7]. In a further study, multilinear regression analysis was used to predict OMC and MDD parameters [17]. In another study, the effective stress parameters of unsaturated soils were predicted using an artificial neural network regression algorithm [18]. When we compare the studies mentioned above with our study, we see that our study is the first to use the decision tree and SVM algorithms to estimate compaction parameters.

Using multivariate support vector regression (MLS-SVR), kernel-based support vector regression (KB-SVR), and the decision tree algorithm (DTR), the present study aimed to estimate compaction parameters without performing laboratory experiments. We preferred these algorithms because of their successful performance in previous regression analysis and classification studies. In this study, data were used from 126 compaction and soil classification experiments gathered by Gunaydin [9]. These data included different soil types: clay of high plasticity (CH), clay of intermediate plasticity (CI), clay of low plasticity (CL), clayey gravels (GC), silty gravels (GM), silt of high plasticity (MH), silt of intermediate plasticity (MI), silt of low plasticity (ML), and clayey sands (SC).

The rest of the study is organized as follows: Section 2 covers the data and some information about soil types and their characteristics, and information about the three methods employed. Section 3 presents the findings and discussions. Finally, Section 4 presents the concluding remarks.

2. Materials and methods

2.1. Soil types

This study used the results of 126 compaction and soil classification tests of nine different soil types [clay of high plasticity (CH), clay of intermediate plasticity (CI), clay of low plasticity (CL), clayey gravels (GC), silty gravels (GM), silt of superior plasticity (MH), silt of intermediate plasticity (MI), silt of low plasticity (ML), and clayey sands (SC)] used in the construction of small dams in the vicinity of Niğde. The data were gathered by Gunaydin [9]. More than one soil type was employed in this study since the diversity of soil types characterizes the variety of applications in the geotechnical field. The purpose of this study is to discuss and estimate the compaction parameters of these different soil types.

2.2. Characteristics of soils and compaction parameters

The concepts of grading, consistency limits (liquid limit and plastic limit), density, and compaction parameters (MDD, OMC) are the oldest and most fundamental concepts in soil mechanics. These characteristics are commonly used to identify, classify, and assess soil properties. Grades of gravel can be readily appreciated even by the most untrained eye, and gravel is a somewhat different material from sand. Likewise, fine-grained (silt) soils and clay are also distinguishable. It is not just the particle size but also the particle-size distribution

that is important in a particular type of soil. Thus, the grading of soil determines many of its characteristics. Density is usually of primary consideration where density values are used directly, for example, to calculate the earth pressure behind retaining walls or basements, since it is the combined mass of soil and water that determines the pressure. The notion of soil consistency limits stems from the fact that soil can exist in any of four states, depending on its moisture content. However, this study investigated only two states (the liquid and plastic limits). The liquid and plastic limits represent the moisture contents at the borderline between the plastic and liquid phases, and the semisolid and solid phases (Figure 1a). The compaction parameters, which are the maximum dry unit volume and the optimum water content, are determined by the Proctor test under laboratory conditions. The ideal amount of water needed to obtain the MDD is called the water content (Figure 1b). Assessing the OMC using the MDD is essential in most engineering projects, and using these data, the required engineering parameters are determined for compressing the soils.

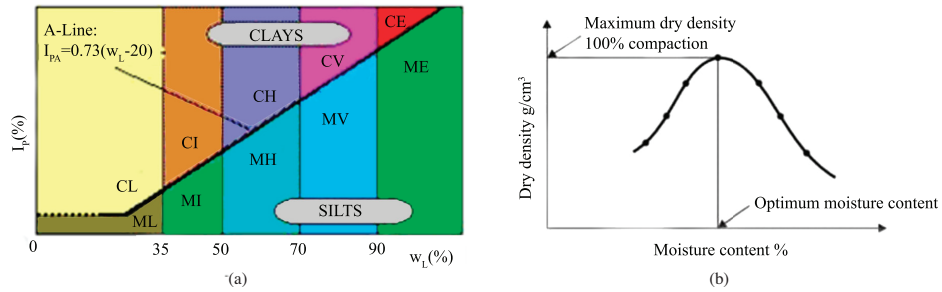

Figure 1. Atterberg (consistency) limits (a), compaction curve (b).

2.3. Data

Data were gathered from the experimental tests of soil obtained from the small dams of Niğde in Turkey. The summarized values and some statistics belonging to the data used in the analysis section are shown in Table 1. According to the unified soil classification system (USCS), soils are classified as clay of high plasticity (CH), clay of intermediate plasticity (CI), clay of low plasticity (CL), clayey gravels (GC), silty gravels (GM), silt of high plasticity (MH), silt of intermediate plasticity (MI), silt of low plasticity, (ML), and clayey sands (SC). One hundred twenty-six datasets were analyzed using SVM and decision tree algorithms.

2.4. Multivariate support vector regression (MLS-SVR)

A single-output regression model can be extended to a multioutput regression model [19]. Multivariable regression maps the multiple input data to a multivariate input space in the learning stage [20]. This method is supported by some regression algorithms. The support vector machine (SVM) algorithm is one of them. Least square SVM (LS–SVM) is a version of SVM and was initially introduced by Suykens and Vandewalle [21]. LS–SVM has been proved to be a beneficial and promising method. There are some advantages of LS–SVM, one of which is that LS–SVM uses a linear equation that is simple to solve and good for computational time-saving [22]. The multioutput regression aims to predict an output vector y that is a high-dimensional matrix from a given input vector x. In other words, the multioutput regression problem can be formulated as a learning method mapping the data from a one-dimensional space to a multidimensional space. The multioutput LS–SVR (MLS–SVR) solves this problem by finding the following: $W = (w_1, w_2, ..., w_m) \in R$ and $b = (b_1, b_2, ..., b_m)^T \in$

Table 1. The data abstracted in the study.

-	Characteristics of soils (inputs)						Compaction parameters (outputs)	
-	Fine-grained%	Sand%	Gravel%	G_S	W_L%	W_P%	OMC%	MDD g/cm^3
1	64	31	6	2.69	55	24	20.8	1.53
2	74	24	2	2.70	57	23	25.0	1.43
3	73	24	3	2.71	52	24	21.5	1.56
...	...	...	...	...	...	...	...	...
124	24	49	27	2.68	32	16	10.0	2.03
125	42	52	6	2.70	31	16	10.5	1.97
126	47	42	11	2.76	41	25	15.6	1.79
Max	83	71	67	2.85	57	30	26.0	2.09
Min	13	15	0	2.58	23	14	7.6	1.43
Avg	51	37	12	2.72	40	21	16.3	1.78

R. W is a multidimensional vector and b is a single-dimensional vector that minimizes the objective function given in the following equation:

$$\min_{(W\in R^m xn, b\in R^m)} \varphi(W,I) = 1/2trace(W^TW) + \gamma 1/2trace(I^TI)]. \quad (1)$$

2.5. Kernel-based support vector regression (KB-SVR)

SVM was first developed by Vapnik in 1995. It has been successfully applied to a number of real-world problems related to regression [22]. The concept of SVM was introduced for desirable margin classifiers in the context of statistical learning theory. The fundamental idea in SVM is to search for the optimal hyperplane separating clusters in such a way that the samples within one category of the target variable fall on one side of the plane, while samples within the other category fall on the other side [23]. The basic theory of the SVM regression method can be briefly summarized as follows: consider a regression problem with its training data of the form x_i,y_i, where i = 1,2,...L, $x_i \in R$, $y_i \in R$; here x_i is a sample value of the input vector x consisting of N training patterns, and y_i is the corresponding value of the desired model output. The actual model output y_i is expressed as a linear function, f(x) given by the following equation [24]:

$$y_i = f(x) = w_T\varphi(x) + b. \quad (2)$$

In the SVM algorithm, kernels are used to compute the elements of the gram matrix (gram matrix determines the vectors v_i up to isometry) used for some specified functions. These functions are used to map the training data into the kernel space. If the data are not linearly separated, kernel functions (φ) are employed to analyze data by moving them to higher-dimensional spaces. However, the corresponding conversions can be done with only one function instead of $\varphi_1(x),\varphi_2(x),...,\varphi_n(x)$ functions. Mercer's theorem can be useful to assess the kernel functions. According to this theorem, if there is a φ mapping written in the form of $K(x_i,x_j) = \varphi(x_i)^T,...,\varphi(x_j)$, this equation is expressed as a positive definite symmetric kernel function. The following kernel functions given in Table 2 are used in the study.

Table 2. Kernel functions used in SVM regression.

Linear kernel	$K(x_i,x_j)= x_i^T x_j$
Polynomial kernel	$K(x_i,x_j,c,d) = (c+x_i^T x_j)^d$
Radial basis kernel	$K(x_i,x_j,c,d) = \exp \frac{\lvert x_i - x_j \rvert^2}{2\sigma^2}$

2.6. Decision tree regression (DTR)

Decision trees resemble flow charts. In a decision tree, in which each attribute is represented by a node, branches and leaves are the elements of the structure. In a tree, the final structure is expressed as a leaf; the topmost structure is expressed as the root, and the structure between the leaf and root is expressed as a branch [25]. Many methods like entropy-based algorithms, regression trees, and memory-based classification models have been developed for generating decision trees in the published literature. One of the most important problems in these methods is to assess the criterion for branching. Furthermore, decision trees can often be complex. In a decision tree, it may be possible to replace a leaf with a subtree. This process is called the pruning of a decision tree. Thus, predictable error rates can be decreased and the quality of the regression model can be increased by replacing a leaf with a subtree in an algorithm [26].

In the present study, decision tree regression (DTR) was used to estimate the soil compaction parameters. Since the data included two different dependent variables, two decision tree models were employed in this method. Decisions developed the tree using the MSE (mean squared error) as a splitting criterion. DTR models were regression trees with binary splits. In the models, the maximum number of splits was determined as data size minus 1. Furthermore, since models originated from the same parent node, the sums of their risks were greater than or equal to the risk associated with the parent node. Moreover, there were nine node splits in the decision tree for OMC estimation and fifteen for MDD estimation, and each leaf had two observations per tree leaf. The output of the model tree included the optimal sequence of pruned subtrees, but in our model, no prune levels were specified. We used a tenfold cross-validation model in the DT regression.

3. Findings and discussion

In this study, three different regression models were used in estimating soil index parameters. These were multivariate support vector regression (ML–SVR), kernel-based support vector regression (KB–SVR), and decision tree regression (DTR) methods. In these methods, fine-grained, sand, gravel, specified weight (G_S), liquid limit (W_L), and plastic limit (W_P) were employed as the input variables. The maximum dry unit weight (MDD) and optimum water contents (OMC) were the variables to be predicted. Since there were two outputs, the different estimation models had to be developed for each method. Furthermore, these two outputs were sometimes employed as a single output because a multiinput–multioutput regression model was supported in some methods. The correlation matrices that belong to raw data are given in Figure 2. In the figure, a robust correlation is seen between W_L and OMC, and its R-value is 0.82. Besides, there is a strong relation between W_L and MDD, and its R-value is 0.76. However, the correlations between G_S and OMC and between Gs and MDD are very poor. At first glance it can be said that the liquid limit is meaningful in the estimation of OMC and MDD but not for G_S.

In the MLS-SVR model, the regression analysis primarily used three methods: GridMLSSVR, MLSSVR-Train, and MLSSVRPredict [20]. The GridMLSSVR function was applied to find the best gamma, lambda,

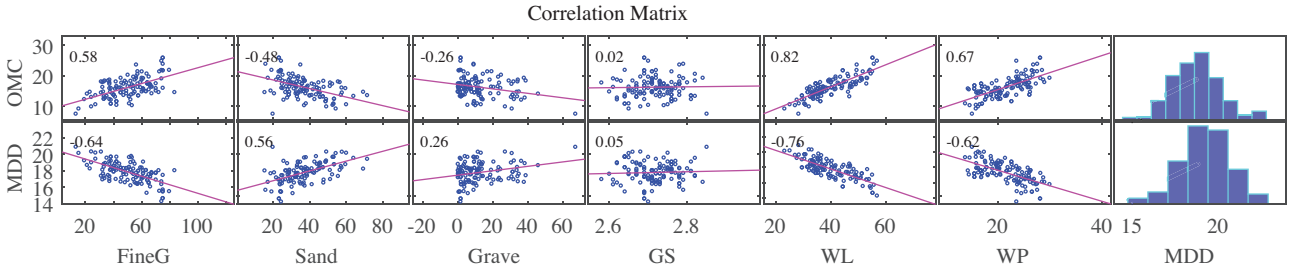

Figure 2. Correlation matrices related to raw data between inputs and (a) OMC and (b) MDD outputs.

and p parameters, which are used as inputs to the MLSSVRTrain function. Gamma and lambda were the positive, real, regularized variables and p was the favorable hyperparameter of the radial basis kernel (RBF) function. The MLSSVRTrain function was a learning algorithm. This function mapped the training output space according to the training input data. The MLSSVRPredict procedure estimated the test input values according to the model set in the training process. In the ML–SVR implementation, the regression analysis was performed for ten iterations and also the mean squared error (MSE) was calculated for each iteration. Figures 3a, 3b, and 3c show the regression plot, the MSE graph, and the R-values for ten iterations, respectively. In the figures, it is observed that the MLS–SVR method produced poor regression ratios (between 0.25 and 0.64). Moreover, the average R-value is 0.43 and the Pearson correlation values change across different iterations. The minimum and maximum values are seen to vary between 0.2 and 0.6. Furthermore, it is also inferred from the MSE graph that the error rates in OMC estimation are higher than in MDD estimation. Therefore, this method was more successful in MDD estimation.

Another regression method is kernel-based support vector regression (KB–SVR). In this method, linear, polynomial, and radial basis functions (RBF) were employed as the kernel functions. Furthermore, three distinctive SVR models and three different types as output were employed for each kernel function. K-fold cross-validation was used to predict the outputs in SVR. K-fold means a positive integer (>1) specifying the number of folds or groups (K) to be used for cross-validation. In the analysis, the preferred value for K was 30% of the data, and thus the losses obtained by the cross-validation were calculated. For every fold, performance loss was computed. According to the results, regression losses were 2.90 in OMC estimation and 0.41 in MDD when the linear kernel function was used. Besides, when the polynomial kernel was used, this value was 3.37 in OMC and 0.12 in MDD. When the RBF was used, the loss was 4.08 in OMC and 0.18 in MDD. Furthermore, the R-values between the output and predicted values were calculated for ten iterations. The regression results between outputs and predictions are given in Figure 4. In the figure, regression values between the outputs and predictions are shown for each applied kernel function. Furthermore, the Pearson correlation coefficient and mean squared error (MSE) were calculated over 30% of the outputs that were used as the prediction. In Figure 4, while histograms appear in the matrix diagonal, scatter plot pairs appear on the diagonal. The slopes of the least-squares reference lines in the scatter plots are equal to the displayed correlation coefficients. Moreover, it is observed that there are strong correlations between real outputs and predictions. The best R-value is obtained as 0.91 in OMC estimation when the linear kernel is used, as 0.93 in both OMC and MDD when the polynomial kernel is used, and as 0.90 in MDD when the RBF kernel is used. However, lower values are observed in R-values in the opposite cases. Considering these findings, it can be said that the polynomial kernel outperformed the other kernel functions. Moreover, in Figure 4, it is observed that correlations are more robust in OMC predictions than in MDD predictions. The graphics of the Pearson correlations and MSE values

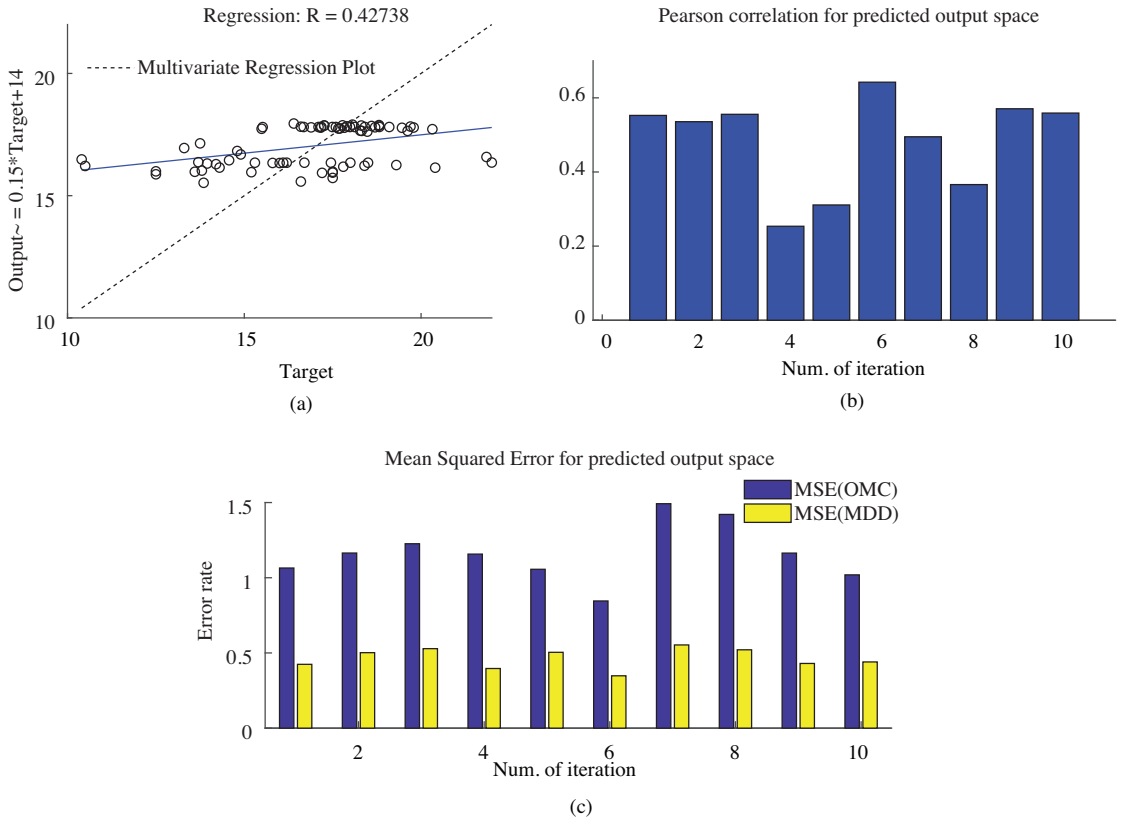

Figure 3. (a) Regression plot; (b) R and (c) MSE values for ten iterations in ML-SVR method.

are shown in Figure 5. The successes are also seen in MSE values. That is, these values displayed in the figures are seen to decrease in MDD prediction compared to OMC prediction. As a conclusion, it can be said that MDD prediction is better in most cases. It can be deduced from the figure that the Pearson correlations and MSE values are superior in all iterations in MDD estimation and the best results are obtained for the polynomial kernel. Moreover, the performances of linear and RBF kernel functions fall behind in this regard. The average result values of Pearson's correlations and MSE values are given in Table 3.

Table 3. The average results in KB-SVR analysis.

Kernel type (SVR)	OMC			MDD		
-	R-values	Pearson corr.	MSE (avg.)	R-values	Pearson corr.	MSE (avg.)
Linear kernel	0.91	0.87	1.33	0.84	0.87	0.45
Polynomial kernel	0.93	0.87	1.54	0.93	0.87	0.04
RBF kernel	0.89	0.79	1.72	0.90	0.81	0.05

In SVR analysis, we also calculated the R-values by combining two outputs (OMC and MDD) as one output using linear, polynomial, and RBF kernel functions. The plots of the obtained results are given in Figure 6. In the multioutput regression studies, the R-values are obtained as 0.88, 0.98, and 0.99 for linear, polynomial, and RBF kernel functions, respectively. In the single-output regression studies, R-values

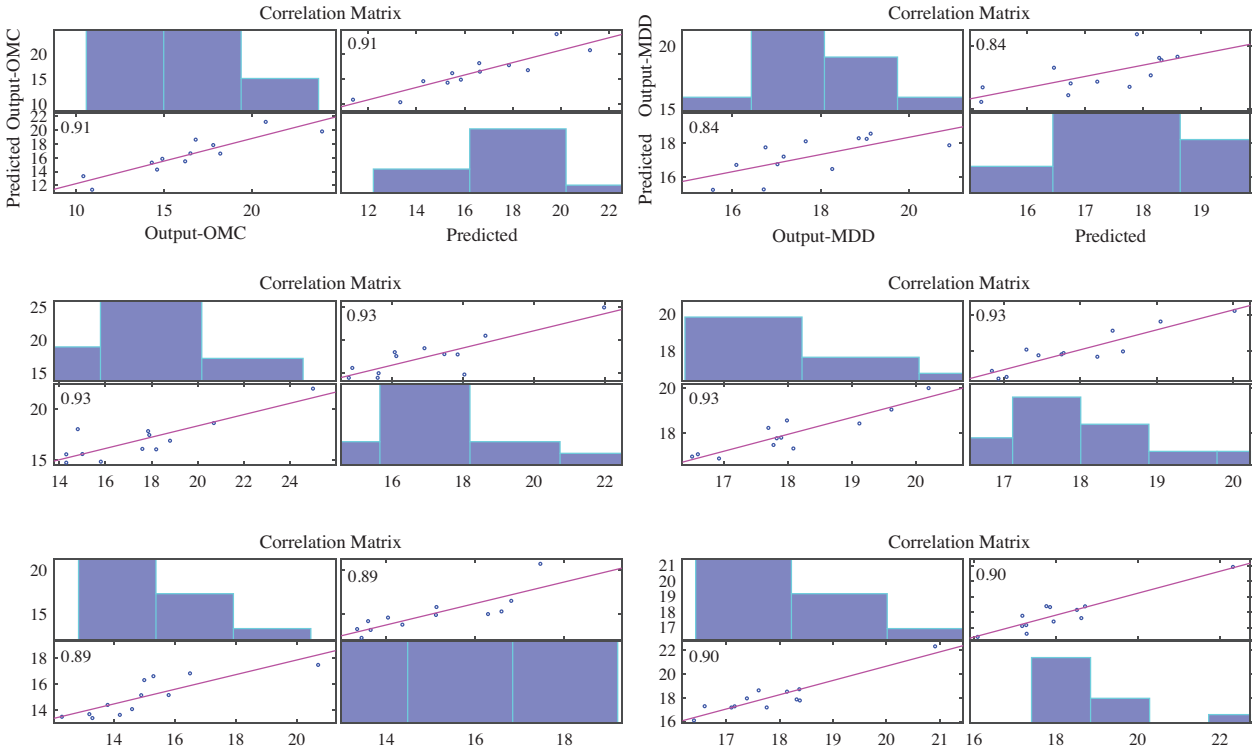

Figure 4. Correlations between the real outputs and the predictions for (a, b) linear, (c, d) polynomial, and (d, e) RBF kernel functions. (a, c, e) correspond to OMC and (b, d, f) to MDD estimations.

are obtained as 0.91 and 0.84 in OMC and MDD predictions, respectively, when the linear kernel is used. Furthermore, the R-value is 0.93 both in OMC and MDD when the polynomial kernel is employed. Lastly, the R-values are 0.89 and 0.90 in OMC and MDD, respectively, when the RBF kernel is used. The overall obtained R-values are summarized in Table 4.

Table 4. Comparing the R-values obtained in multi- and single-output kernel based-SVR analysis.

		Linear		Polynomial		RBF	
R-values	Multioutput	0.88		0.98		0.99	
	Single-output	OMC	MDD	OMC	MDD	OMC	MDD
		0.91	0.84	0.93	0.93	0.89	0.90

In the DTR method, we had two prediction models for two prediction outputs (OMC and MDD). The decision tree models were obtained from ten different training iterations. The split size number changed according to the decision tree models. We observed that the default numbers of splits were 9, 11, 8, 7, 9, 9, 9, 10, 11, and 8 in OMC estimation. These values were 15, 10, 8, 7, 8, 8, 7, 8, 7, and 8 in MDD prediction. Besides, the averages of imposed splits were 9 and 15 in OMC and MDD, respectively. In the decision tree models, each node had two different leaves. In the analysis process, we employed tenfold cross-validation method. Cross-validation losses for the partitioned regression model were 4.95 and 0.77 in the OMC and MDD estimations, respectively, and these losses were suitable for the selected pruning levels. By default, the number of imposed splits was one less than the number of leaves. The histograms of the number of imposed splits on

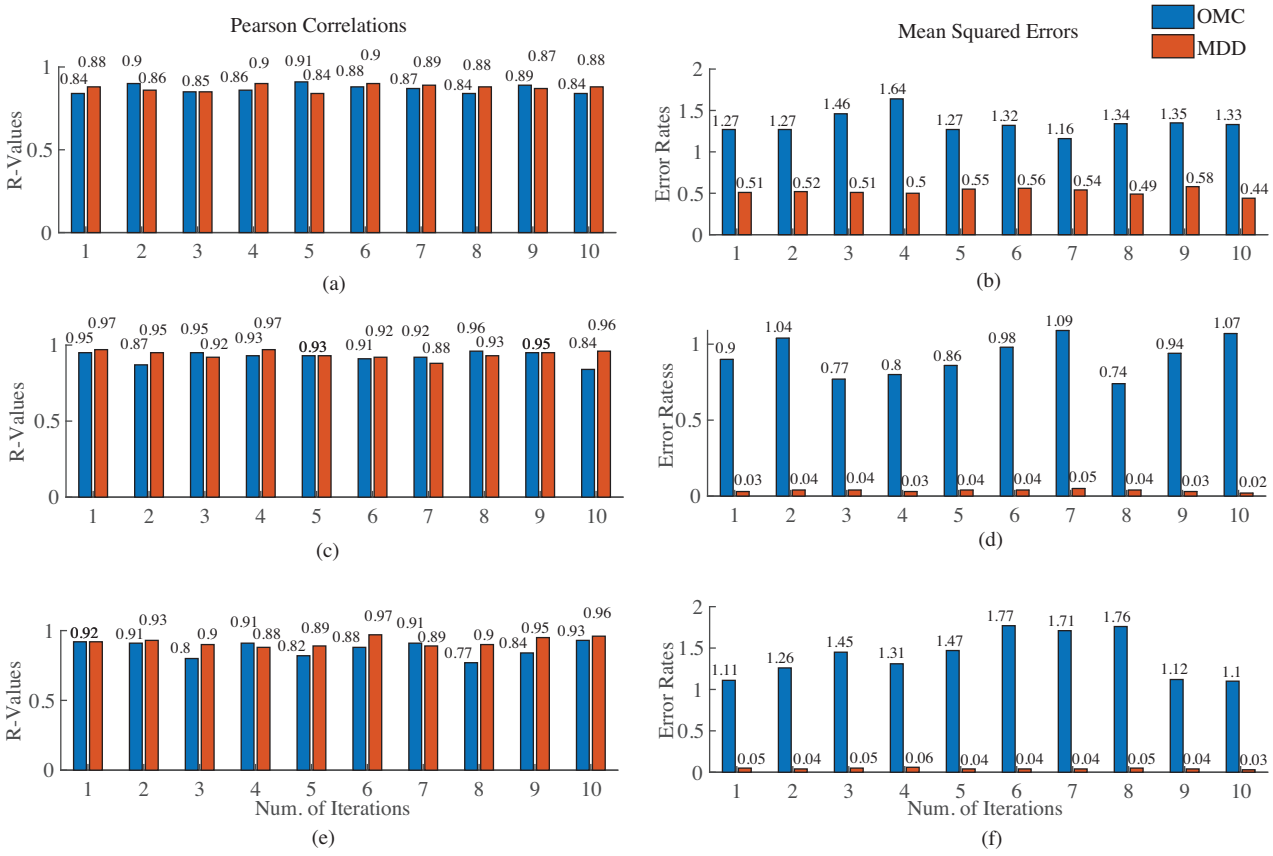

Figure 5. Pearson's correlations (a, c, e) and MSE values (b, d, f) obtained at ten iterations in kernel-based SVR analysis. (a, b) belong to linear, (c, d) belong to polynomial, and (e, f) belong to RBF kernel functions.

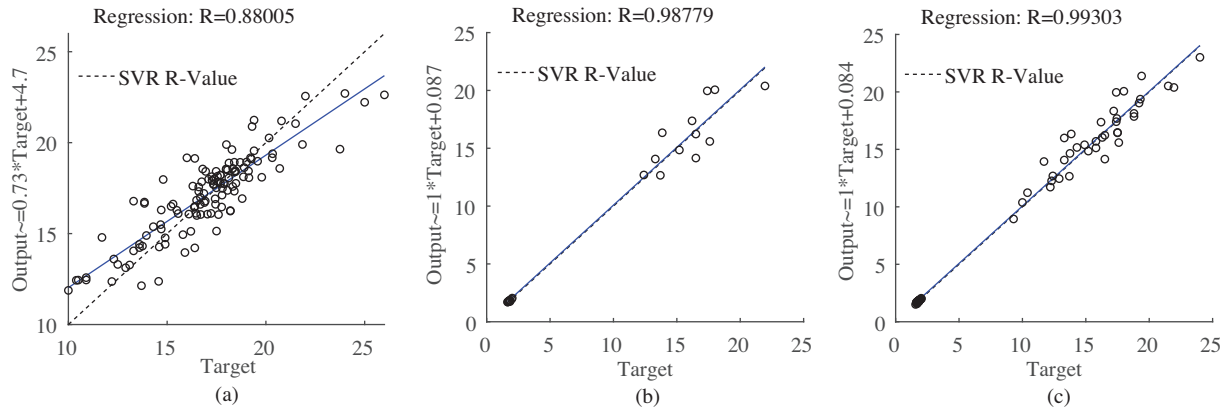

Figure 6. Regression plots for (a) linear, (b) polynomial, and (c) RBF kernel functions in multivariable SVR analysis.

the trees are given in Figure 7. In this method, DTR was developed both for multioutput (MDD and OMC used as single output together) and single-output models (MDD and OMC used as output one by one). The tree models are given in Figure 8. In the developed models, 50% of data were employed for training, and the rest of the data were used for validation.

According to the developed DTR models, some inputs are found at more than one node. In the OMC prediction model, it is seen that the most decisive input values are in plastic limit (W_P) and gravel. Moreover, liquid limit (W_L) and gravel are also important nodes in the decision mechanism of the OMC estimation. In addition, if W_P is greater than 21.83% and gravel is less than 2.64%, the OMC is immediately estimated. The longest part of the decision mechanism is for the large values of W_L and gravel. In the MDD estimation model, the most decisive input values are plastic limit (W_P) and fine-grained soil. Liquid limit (W_L), sand, and G_S are also important nodes in the decision mechanism of the MDD model. In addition, in this model, if (W_P) is greater than 21.8% and fine-grained soil is higher than 71.3%, or if W_P is lower than 21.8% and W_L is lower than 29.5%, the MDD is immediately calculated. In addition to those results, five and eight different prune levels have been tried in the OMC and MDD estimations, respectively. The prediction models for the calculated prune levels and the validation data are shown in Figure 9. When the out-of-sample predictions are examined, it is seen that the prediction strength is weakened as the pruning levels decreased.

In the DTR analysis process, we obtained the estimated values and compared the five and eight predicted outputs at each pruning level for the OMC and MDD estimations. Prediction values were also calculated at each pruning level for test data, and these values are given in Table 5. In this table, L-0 indicates the pruning level at 0. The last two rows are the average and R-values. The R-values indicate the correlation between the real outputs and the predicted values at each pruning level. In the table, it is observed that the OMC estimation is better performed than the MDD estimation. Regression plots are given in Figure 10. According to the figure, R-values are observed to be 0.73, 0.44, and 0.97 in OMC, MDD, and multioutput estimations, respectively.

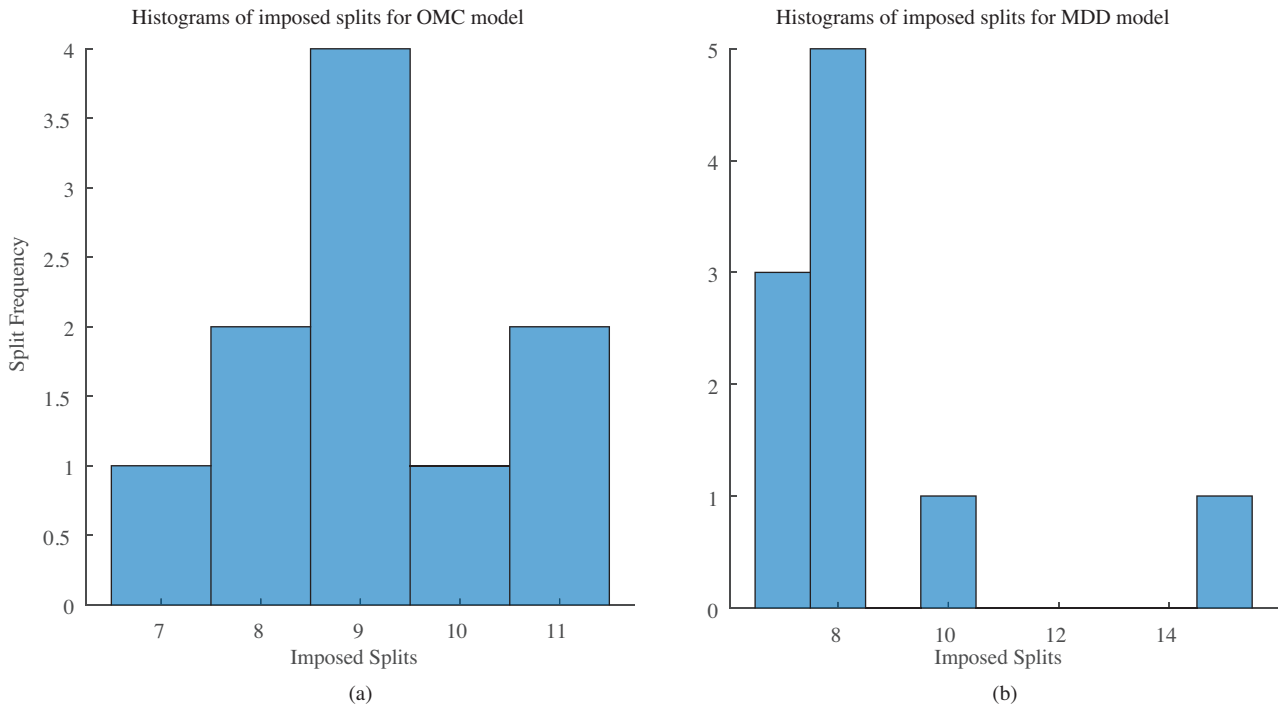

Figure 7. Histograms of imposed splits in OMC (a) and in MDD (b) estimations.

When all analysis processes are evaluated together, it is observed that the most successful method among the regression applications is the polynomial-based KB-SVR method. Furthermore, other KB-SVR methods are also more successful than other methods in the compaction analysis. These situations show that the KB-

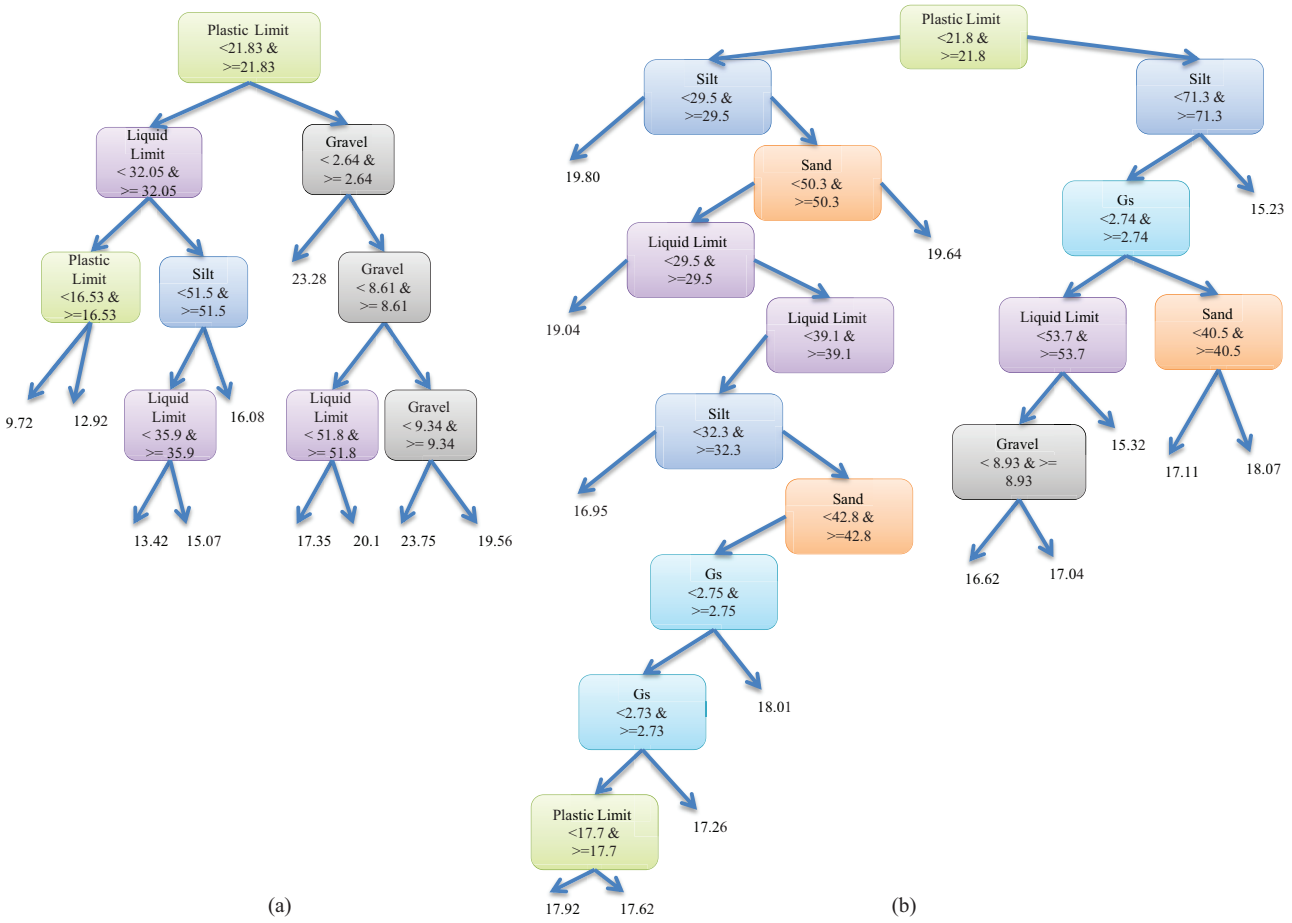

Figure 8. Histograms of imposed splits in OMC (a) and in MDD (b) estimations.

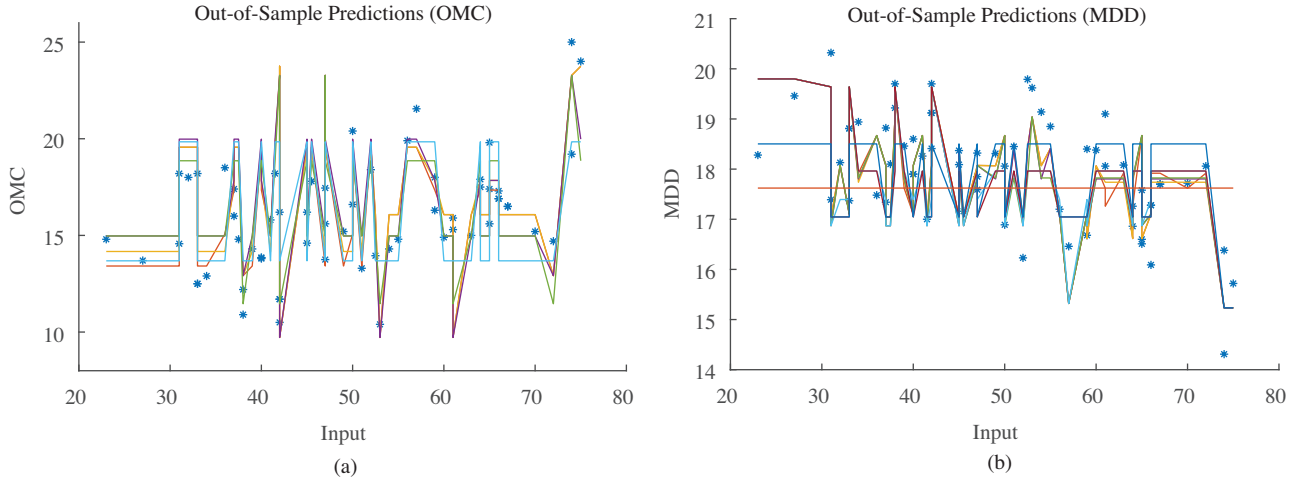

Figure 9. Prune levels in OMC (a) and in MDD (b) estimations.

SVR method is an efficient method for compaction tests that do not require any laboratory environment. Moreover, even in the analysis of the multioutput process, it is observed that KB-SVR methods have good

Table 5. The real values of OMC and MDD, their prediction values in each pruning level, and R-values.

Data num.	Output OMC	Predicted OMC					Output MDD	Predicted MDD							
		L-0	L-1	L-2	L-3	L-4		L-0	L-1	L-2	L-3	L-4	L-5	L-6	L-7
1	21,55	19,96	20,36	20,77	20,77	19,94	1.65	1.63	1.63	1.63	1.63	1.63	1.63	1.60	1.56
2	15,20	14,44	20,36	20,77	20,77	19,94	1.77	1.84	1.84	1.84	1.84	1.84	1.84	1.80	1.77
3	13,60	14,44	20,36	20,77	20,77	19,94	1.80	1.84	1.84	1.84	1.84	1.84	1.84	1.80	1.77
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
61	14.30	14.44	14.01	14.01	14.01	13.31	1.78	1.84	1.84	1.84	1.84	1.84	1.84	1.80	1.77
62	17.45	14.44	14.01	14.01	14.01	13.31	1.76	1.84	1.84	1.84	1.84	1.84	1.84	1.80	1.77
63	10.00	14.44	14.01	14.01	14.01	13.31	2.03	1.84	1.84	1.84	1.84	1.84	1.84	1.80	1.77
Avg.	16.05	16.54	16.04	15.56	16.03	15.55	1.78	1.80	1.75	1.69	1.74	1.69	1.64	1.59	1.54
R-Val.		0.74	0.73	0.76	0.77	0.68		0.44	0.45	0.44	0.42	0.39	0.34	0.29	0.28

regression performances. Therefore, we can say that this method is better than the other methods employed throughout the study. It is also concluded that the MDD estimation shows high performance compared to the OMC estimation throughout all regression applications. We can argue that these two compaction parameters are linearly independent of each other, and also the MDD data have a more determinative feature in the compaction analysis. In addition, it is observed that the MLS-SVR analysis processes are not successful in compaction tests. Lastly, in the study, it is decided that the DTR analysis is good, but not as adequate as KB-SVR. However, it is observed that this method has a high performance when the multioutput model is used in the compaction parameters' estimation. Therefore, it is concluded that while DTR is suitable in the nonlinear regression studies involving nominal data, KB-SVR is successful in the regression applications with numerical values. A comparison of all the findings obtained from the study is given in Table 6.

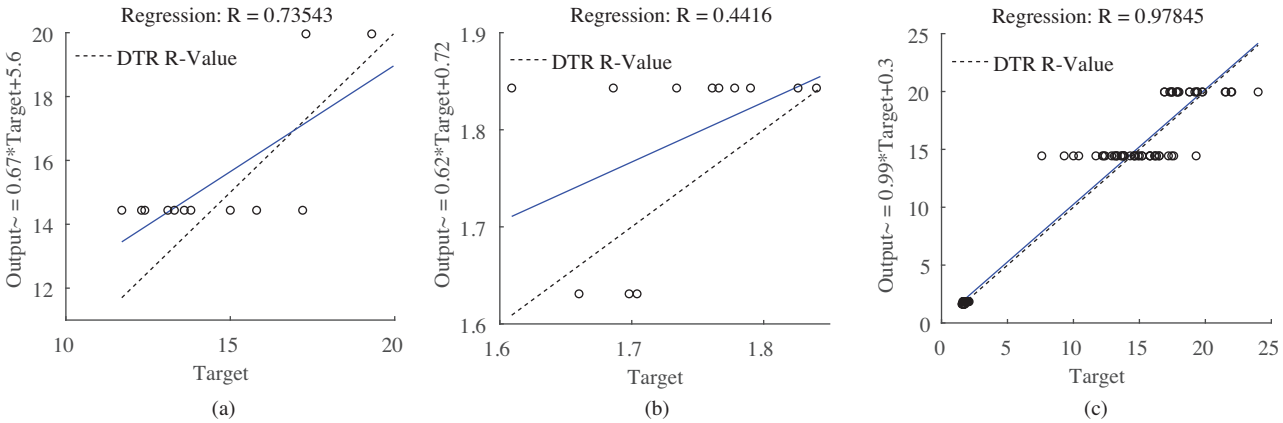

Figure 10. R-values of (a) OMC regression model, (b) MDD regression model, and (c) combination of two models in decision tree algorithm.

4. Conclusions

Compaction parameter estimation is an essential process in soil index studies because, with regression modeling, we can obtain some crucial outputs for adjusted input values without the need for laboratory tests. In this study, the MDD and OMC values were estimated using three different regression methods: multivariate support vector

Table 6. A comparison of all the findings.

-	Single-output (R-values)		Multioutput (R-values)
-	OMC	MDD	OMC-MDD
ML-SVR	-	-	0.42
KB-SVR (linear)	0.91	0.84	0.88
KB-SVR (polynomial)	0.93	0.93	0.98
KB-SVR (RBF)	0.89	0.90	0.99
DTR	0.73	0.44	0.97

machine (MLS-SVR), kernel-based support vector machine (KB-SVR), and decision tree regression (DTR). A thorough search of the relevant literature showed that the present study was the first to use these algorithms for the estimation of compaction parameters.

The data used in the study were obtained from experiments conducted on soil tests in the small dams of Niğde, Turkey [9]. The MLS-SVR, KB-SVR, and DTR algorithms were separately applied to the data and the estimations were carried out independently of each other. Then all performances were compared. According to the obtained results, the average R-value was 0.43 when the MLS-SVR method was applied. In this method, the MSE value was higher in the OMC estimation than in the MDD estimation. Moreover, in the application of KB-SVR, three different kernel functions were used. Among these functions, the polynomial kernel was the best with 0.93 in the estimations, and the linear kernel was good with 0.91 in the OMC prediction. Also, in the KB-SVR method, MSE values were better in the MDD prediction than in the OMC prediction. In addition, when two outputs were used as one output in the KB-SVR method, R-values were the best at 0.98 and 0.99 in the polynomial and RBF kernel functions, respectively. Lastly, in the DTR method, R-values were obtained as 0.73, 0.44, and 0.97 in the estimation of OMC, MDD, and multioutput, respectively. We concluded that the W_P and silt are a decisive factor in the estimation process in the DTR method.

Consequently, when the proposed regression analysis methods were compared to each other, it was seen that the KB-SVR outperformed with an R-value of 0.93 in the estimation of the single parameter. Moreover, MDD estimation was more successful than OMC prediction when all MSE results are considered. The KB-SVR method was the best in all processes because this method could efficiently perform nonlinear analysis and transform the inputs implicitly to a multidimensional featured space. Moreover, when we compared the best results with the published literature, we saw that this success was consistent with the findings in the literature. In this study, we observed that the algorithms initially proposed were successful in the estimation of compaction parameters.

Computer code availability

The application was developed in MATLAB. It is intended to be open source and is available by contacting the authors, or alternatively downloadable from GitHub (2019), regression algorithms for compaction [online]. Website https://github.com/aozbeyaz/SC [accessed 31.10.2019]. Data are also accessible under the code files.

Acknowledgment

The authors would like to thank Prof. Dr. Osman Günaydın for his permission to use the compaction data.

References

[1] Holtz RD, Kovacs WD. Compaction. An Introduction to Geotechnical Engineering. Upper Saddle River, NJ, USA: Prentice Hall, 1981, pp. 109–161.

[2] Korfiatis GP, Manikopoulos CN. Correlation of maximum dry density and grain size. Journal of the Geotechnical Engineering Division 1982; 108 (9): 1171–1176.

[3] Wang MC, Huang CC. Soil compaction and permeability prediction models. Journal of Environmental Engineering 1984; 110 (6): 1063–1083. doi: 10.1061/(ASCE)0733-9372(1984)110:6(1063)

[4] Basheer IA. Empirical modeling of the compaction curve of cohesive soils. Canadian Geotechnical Journal 2001; 38 (1): 29–45. doi: 10.1139/t00-068

[5] Omar M, Shanableh A, Basma A, Barakat S. Compaction characteristics of granular soils in United Arab Emirates. Geotechnical and Geological Engineering 2003; 21 (3): 283–295. doi: 10.1023/A:1024927719730

[6] Suits LD, Sheahan T, Nagaraj T. Rapid estimation of compaction parameters for field control. Geotechnical Testing Journal 2006; 29 (6): 100009. doi: 10.1520/GTJ100009

[7] Sinha SK, Wang MC. Artificial neural network prediction models for soil compaction and permeability. Geotechnical and Geological Engineering 2008; 26 (1): 47–64. doi: 10.1007/s10706-007-9146-3

[8] Tekinsoy MA, Kayadelen C, Keskin MS, Söylemez M. An equation for predicting shear strength envelope with respect to matric suction. Computers and Geotechnics 2004; 31 (7): 589–593. doi: 10.1016/j.compgeo.2004.08.001

[9] Günaydın O. Estimation of soil compaction parameters by using statistical analyses and artificial neural networks. Environmental Geology 2009; 57 (1): 203–215. doi: 10.1007/s00254-008-1300-6

[10] Isik F, Ozden G. Estimating compaction parameters of fine- and coarse-grained soils by means of artificial neural networks. Environmental Earth Sciences 2013; 69 (7): 2287–2297. doi: 10.1007/s12665-012-2057-5

[11] Ören AH. Estimating compaction parameters of clayey soils from sediment volume test. Applied Clay Science 2014; 101: 68–72. doi: 10.1016/j.clay.2014.07.019

[12] Lubis AS, Muis ZA, Hastuty IP, Siregar IM. Estimation of compaction parameters based on soil classification. IOP Conference Series: Materials Science and Engineering 2018; 2018: 306. doi: 10.1088/1757-899X/306/1/012005

[13] Al-Khafaji AN. Estimation of soil compaction parameters by means of Atterberg limits. Quarterly Journal of Engineering Geology 1987; 26 (93): 359–368.

[14] Najjar YM, Basheer IA. Utilizing computational neural networks for evaluating the permeability of compacted clay liners. Geotechnical and Geological Engineering 1996; 14 (5): 193–212. doi: 10.1007/BF00452947

[15] Hausmann MR. Engineering Principles of Ground Modification. New York, NY, USA: McGraw-Hill, 1990.

[16] Kiefa MAA. General regression neural networks for driven piles in cohesionless soils. Journal of Geotechnical and Geoenvironmental Engineering 1998; 124 (12): 1177–1185. doi: 10.1061/(ASCE)1090-0241(1998)124:12(1177)

[17] Sivrikaya O. Models of compacted fine-grained soils used as mineral liner for solid waste. Environmental Geology 2008; 53 (7): 1585–1595. doi: 10.1007/s00254-007-1142-7

[18] Kayadelen C. Estimation of effective stress parameter of unsaturated soils by using artificial neural networks. International Journal for Numerical and Analytical Methods in Geomechanics 2008; 32 (9): 1087–1106. doi: 10.1002/nag.660

[19] An X, Xu S, Zhang LD, Su SG. Multiple dependent variables LS-SVM regression algorithm and its application in NIR spectral quantitative analysis. Guang pu xue yu guang pu fen xi 2009; 29 (1): 127–130 (in Chinese).

[20] Xu S, An X, Qiao X. Multi-output least-squares support vector regression machines. Pattern Recognition Letters 2013; 34 (9): 1078–1084. doi: 10.1016/j.patrec.2013.01.015

[21] Suykens JAK, Vandewalle J. Least squares support vector machine classifier. Neural Processing Letters 1999; 9 (3): 293–300. doi: 10.1023/A:1018628609742

[22] Hwang C. Multioutput LS-SVR based residual MCUSUM control chart for autocorrelated process. Journal of the Korean Data and Information Science Society 2016; 27 (2): 523–530.

[23] Hariri-Ardebili MA, Pourkamali-Anaraki F. Support vector machine based reliability analysis of concrete dams. Soil Dynamics and Earthquake Engineering 2018; 104: 276–295. doi: 10.1016/j.soildyn.2017.09.016

[24] Zhang H, Gao M. The application of support vector machine (SVM) regression method in tunnel fires. Procedia Engineering 2018; 211: 1004–1011. doi: 10.1016/j.proeng.2017.12.103

[25] Quinlan JR. Induction of decision trees. Machine Learning 1986; 1 (1): 81–106. doi: 10.1007/BF00116251

[26] Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, NJ, USA: Wiley, 2003.