International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Estimating California Bearing Ratio Using Decision Tree Regression Analysis Using Soil Index and Compaction Parameters

**Osman GÜNAYDIN[1], Abdurrahman ÖZBEYAZ *[2], Mehmet SÖYLEMEZ[3]**

**Submitted: 14/12.2018 Accepted : 21/02/2019 Published: 20/03/2019**

***Abstract:*** California Bearing Ratio is used as an index of soil strength and bearing capacity. In the machine learning theory, a decision tree algorithm can help us to define preferences, risks, benefits and targets. In this study, the decision tree algorithm was employed for estimating California Bearing Ratio from the soil index and compaction parameters. There were seven inputs and one output in the study. In the analysis, we employed gravel, sand, fine grain, liquid limit, plastic limit, maximum dry unit weight and optimum water as inputs and California Bearing Ratio as output. The number of data was 124. In the decision tree algorithm, data were divided two for train and test groups. In addition, 10-fold cross validation process was applied to data in the analysis. Consequently, fine grain values used as input in the study were carried out to be very determinative for regression analysis. Decision tree regression analysis estimation indicated strong correlation (R = 0.89) between the output and target. It has been shown that the correlation equations obtained as a result of regression analysis are in satisfactory agreement with the test results.

***Keywords:*** *California Bearing Ratio, Regression, Decision Trees, Machine Learning*

## 1. Introduction

California bearing ratio (CBR) is an empirical test and widely applied in design of flexible pavement over the world. This method was developed during 1928-29 by the California Highway Department. Use of CBR test results for design of roads, introduced in USA during 2nd World War and subsequently adopted as a standard method of design in other parts of the world, has recently being discouraged in some advanced countries because of the imperialness of the method [1]. When some construction projects such as road set, road of airstrip, footway are deciding to implement, suitable sub-base constructions have to be built pre-construction. Moreover, some certain sub-base construction properties such as bearing capacity, lodgement and eruption have to be met in these projects. That's why, an approach of assessment of such excavation is crucial and takes important part among geotechnical and road engineering works. When the stiffness and shear strength of subgrade are assessed, CBR method is commonly employed in tests. Moreover, it is an indirect measure demonstrating comparison of the strength of subgrade material to the strength of standard crushed rock [2].

Thanks to advances in information technology, we can record very large amounts of data. Many recorded data is waiting to be analysed by experts. Different things are expected from the analysis of each of these data. Machine learning helps us to find solutions for many difficult real-world problems related to engineering, medicine and social science. Since the analysis of very high quantities of recorded data cannot be analysed with human power, this data can be analysed with the help of machine learning algorithms that are the result of technological development [3]. Regression investigates to predict the relations between output and input variable by automatically in machine learning. Much data is used in the regression when doing this. The aim of regression analysis is to estimate the output variables from new samples [3]–[6]. In literature, linear regression, support vector regression, multilayer perception (MLP), K-nearest neighbour (KNN) and the decision tree methodologies are usually employed for regression analysis. In this study, we prefer decision tree (regression tree) method for estimating some values belonging to the laboratory works in the civil engineering field.

A decision tree is a hierarchical data structure implementing the divide-and-conquer strategy. It is an efficient nonparametric method, which can be used both for classification and regression [3]. Although regression trees are not as popular as classification trees, they are highly competitive with different machine learning algorithms and are often applied to many real-life problems [6]. Regression tree is a type of the machine learning tools that can satisfy both good prediction accuracy and easy interpretation, and therefore, have received extensive attention in the literature. Regression tree uses a tree-like graph or model and is built through an iterative process that splits each node into child nodes by certain rules, unless it has a terminal node that the samples fall into. A regression model is fitted to each terminal node to get the predicted values of the output variables of new samples [4].

CBR test applied in the laboratory requires a large soil sample and is a challenge as well as time consuming. In obtaining representative CBR values, engineers may face with some challenges. Furthermore, there are some difficulties about soil investigation in laboratory works related to limited budget and poor planning conditions. Furthermore, the results sometimes are not accurate due to poor quality of skill of the technicians testing the soil samples in the laboratory. All these problems may result in serious delay in the progress of the project, and ultimately it may lead to escalation of the project cost. Therefore, some prediction models have been proposed to fit unknown values without use of

[1] *Civil Engineering, Adiyaman University, Adiyaman, TURKEY*
*ORCID: 0000-0001-7559-5684*
[2] *Electrical Electronics Engineering, Adiyaman University, Adiyaman, TURKEY*
*ORCID: 0000-0002-2724-190X*
[1] *Civil Engineering, Adiyaman University, Adiyaman, TURKEY*
*ORCID: 0000-0001-8684-9117*
* *Corresponding Author: Email: aozbeyaz@adiyaman.edu.tr*

experimental procedures. In the literature studies, Yildirim and Gunaydin [2] and Yildirim [7], taking the power of Artificial Intelligence methods in solving complex system, Artificial Neural Network (ANN) and regression analyse methods were applied for the prediction of CBR. Yang et al. [4] have developed a regression tree approach using mathematical programming. In this study, we employed decision Tree regression model to fit some values CBR test.

Throughout the study, the decision tree regression was utilized to predict CBR values without making any experiment in laboratory. Effortless, easy to interpret, having understandable rules and having discrete attribute values was the motivation factor in the study. For this purpose, one hundred twenty four (124) compaction and soil classification test results of ten different soil types (CH, CI, CL, GC, GM, GP-GC, MH, MI, ML, SC) were collected from the public high ways of Turkey's various regions. In the study, we aimed to make a simple regression analyses showing the relationship between the CBR and sieve analysis, Atterberg limits, maximum dry unit weight ($\gamma k_{max}$) and optimum moisture content (OMC) by using the collected data in this study. For this purpose, decision tree regression was applied for the prediction of CBR test results using data collected in the present study. In the estimation of the CBR, decision tree regression was employed for the first time in this study.

## 2. Materials and Methods

In this section, the materials and methods that we employed in our study are described.

### 2.1. Data

The experimental data were collected from the results of soil mechanics laboratory tests of the public highways in Turkey's seven different regions. The first group consisting of 62 data was used to train the network and to develop different decision tree models. The second group consisting of 62 data was used to validate and test the accuracy of the developed models. Brief information about data is given in Table 1. According to the unified soil classification system (USCS), grounds are classified as CH, CI, CL, GC, GM, GP-GC, MH, MI, ML and SC. These are soil types. One hundred twenty-four (124) data were employed in regression analyses using decision tree.

**Table 1.** Statistical values of the data used in the study

| | *Grav. (%)* | *Sand (%)* | *FG. (%)* | *$W_L$ (%)* | *$W_P$ (%)* | *$\gamma k_{max}$ (%)* | *OMC (%)* | *CBR (%)* |
|---|---|---|---|---|---|---|---|---|
| Max | 78,00 | 49,00 | 99,10 | 89,00 | 43,00 | 2,19 | 40,20 | 23,00 |
| Min | 0,00 | 0,90 | 10,00 | 20,00 | 11,00 | 1,21 | 7,20 | 0,00 |
| Avrg. | 13,05 | 18,50 | 68,44 | 43,03 | 22,33 | 1,66 | 19,51 | 6,15 |
| Hydr. | 0,00 | 17,50 | 77,50 | 40,00 | 20,50 | 1,64 | 18,80 | 4,00 |
| Skew | 1,43 | 0,30 | -0,86 | 1,04 | 1,03 | 0,29 | 0,75 | 1,33 |
| Kurt. | 0,60 | -0,80 | -0,58 | 1,05 | 0,65 | -0,04 | 0,69 | 0,74 |
| Var. | 459,04 | 123,02 | 722,26 | 191,04 | 52,80 | 0,04 | 45,70 | 29,41 |
| Std. | 21,43 | 11,09 | 26,87 | 13,82 | 7,27 | 0,21 | 6,76 | 5,42 |

In the Table1, Max, Min, Avrg., Hydr., Skew, Kurt., Var., and Std. are referred as maximum, minimum, average, hydrangea, skewness, kurtosis, variance and standard deviation values of one hundred twenty-four (124) data employed in the study. Moreover, FG, $W_L$, $W_P$, $\gamma k_{max}$, OMC and CBR refer to fine grain, liquid limit, plastic limit, maximum dry unit weight, optimum water and California Bearing Ratio respectively.

### 2.2. California Bearing Ratio (CBR)

CBR is defined as the ratio which is between the resistance against the sinking of a penetration piston into the soil with 1.27 mm/min (0.05 in./min) velocity and the resistance is shown by a standard crushed rock sample for the same penetration depth [2], [7], [8]. CBR resistance, defined as the ratio between the applied stress (unit-strength), according to a specific energy compressed soil in a predetermined moisture content on the speed control to a sunk penetration piston to reach the required depth, the applied standard tension, in the experiment by using listed crushed rock for the piston to reach into the same depth [9].

The CBR test can be performed in either the laboratory or the field. The laboratory CBR test is described in ASTM D 1883-99 (2003) and the field CBR test is described [10]. In the laboratory, the CBR test is typically performed on compacted soil samples, while in the field, the CBR test would be performed at the ground surface, or on a level surface excavated in a test pit, trench, or bulldozer cut [11].

$$CBR = \frac{Applied\ Stress\ in\ Experiment(or\ Load)}{Standart\ Stress(or\ Load)} x100$$

### 2.3. Decision Tree

A decision tree is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree is composed of internal decision nodes and terminal leaves. This method can be used for both classification and regression. A regression tree is constructed in almost the same manner as a classification tree, except that the impurity measure that is appropriate for classification is replaced by a measure appropriate for regression. Let us say for node $m$, $X_m$ is the subset of $X$ reaching node $m$, namely; it is the set of all $x \epsilon X$ satisfying all the conditions in the decision nodes on the path from the root until node $m$. We define;

$$b_m(x) = \begin{cases} 1, & if\ x \in X_m: x\ reaches\ node\ m \\ 0, & otherwise \end{cases} \tag{1}$$

Good split of a tree is decided by the mean square error from estimated value. Let $g_m$ is predicted value in node $m$ in regression.

$$\begin{aligned} E_m &= \frac{1}{N_m}\sum_t (r^t - g_m)^2 b_m(x^t) \\ N_m &= |X_m| = \sum_t b_m(x^t) \end{aligned} \tag{2}$$

$E_m$ is related to variance at $m$. The mean of the required outputs of samples reaching the node is used in a node.

$$g_m = \frac{\sum_t b_m(x^t) r^t}{\sum_t b_m(x^t)} \tag{3}$$

If error is acceptable for a node ($E_m < \theta_r$), a leaf node is created, and it stores $g_m$ value. Namely, a piecewise constant approximation with discontinuities is constructed at leaf boundaries. If the error is not acceptable, data reaching node $m$ is split further such that the sum of the errors in the branches is minimum[3], [12].

## 3. Experimental Results

In this study, the decision tree algorithm was applied to experimental data collected from different regions of Turkey. We employed Matlab platform in the analyses studies. Decision tree grows the tree using MSE (mean squared error) as the splitting criterion. Decision tree model is a regression tree with binary splits. In our model, maximum number of splits were assessed as data size minus one. In the model, merge leaves was *'on'* mode and so the model was merged, because the model originated from the same parent node, and the sum of their risk values was greater or equal to the risk associated with the parent node. Moreover, each splitting node in the decision tree had ten. Moreover, each leaf had two observations per tree leaf. The output of the model tree included the optimal sequence of pruned sub trees but in our model, any prune levels weren't specified.

We applied tenfold cross-validation model in the regression analysis because ten folds mean 90% of full data used for training in each fold test. This was a compromise practically motivated by: 90% is not too far from full 100%, which means that cross-validation produces a fair estimation of test performance[13]. Cross-validation loss of the partitioned regression model was 5.92. When this loss was compared the pruning levels, this values is better at the selected pruning level. We had one prediction model, and finally we had one predicted outputs for both CBR. Moreover, the number of input variables was preferred as seven in the analysis. The seven input variables used in the model were gravel (G, %), sand (S, %), fine-grain (FG, %), liquid limit (WL, %), plastic limit (WP, %), maximum dry unit weight ($\gamma k_{max}$, %) and optimum water (OMC %). The number of output variable was one, and it was CBR. In the analysis, the models were obtained among ten different trained models, and we preferred the first tree in the trained model trees list. The number of split size was changed in the trained tree list. We observed the default number of splits as 10, 11, 11, 12, 10, 11, 13, 8, 12 and 12 according to the trained model list. The average number of splits was eleven. This model was automatically generated according to obtained data. The histogram of the number of imposed splits on the trees is given in Figure 1. By default, the number of imposed splits is one less than the number of leaves.

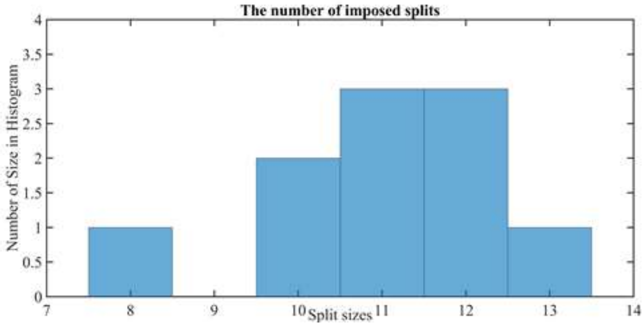

**Figure 1.** The number of imposed splits is shown according to the trained trees list in a histogram.

Hence, regression model was obtained according to the one output. Our decision tree regression model is given in Figure 2.

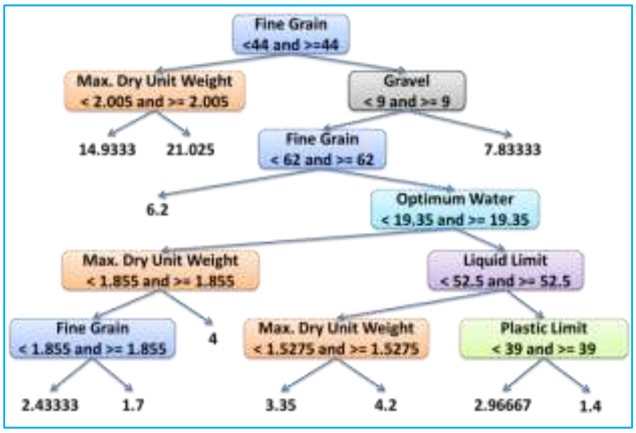

**Figure 2.** Decision tree model is obtained for CBR regression analysis.

The developed decision tree model for the training was improved and tested; validation process was decided according to this decision tree model. When we look at the model given in Figure 2, we can observe some names of inputs included in data at each node. In addition, ten nodes are founded on the model. Each node has two different leaves. The most decisive input values are observed as the fine grain (FG), maximum dry unit ($\gamma k_{max}$) according to the model. Gravel, optimum water (OMC), liquid limit (WL) and plastic limit (WP) are important pieces of the decision mechanism in the model. If the fine grain is fewer than 44% and maximum dry unit weight is fewer than 2.005 gr/cm3 or the fine grain is above 44% and gravel value is above 9%, regression results are immediately calculated in the model. The worst decision input is Sand and secondly Plastic Limit in the data. CBR values are decided according to this model.

In the study, six different prune levels were investigated and predicted out-of-sample responses of regression trees were obtained, and then plotted the results. According to these results, prediction model for six prune levels and the validation data is shown in Figure 3. When we examine the out of sample predictions, we observe that the pruning operations decrease the prediction power in the decision tree analysis.

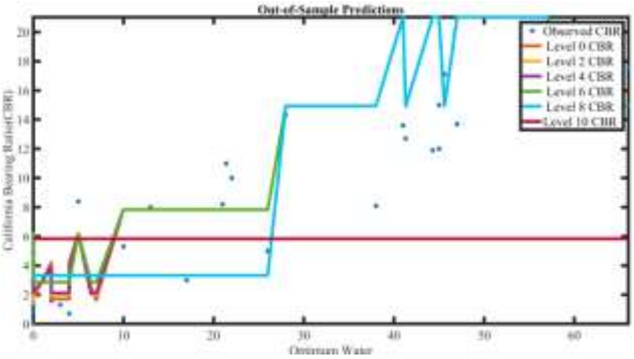

**Figure 3.** Obtained prune levels in decision tree model for CBR regression analysis.

Moreover, at the end of the analyses, we obtained six prune levels as output. We compared the six predicted outputs with the true outputs for each pruning level. In addition, result values are given in Table 2 for each pruning level respectively. Mean squared errors (MSE) were calculated, and these values are given in last column in the table. When Table 2 were scrutinized, the best MSE was 0.89 at level 0 CBR in the model. In addition, this result is shown in Figure 4.

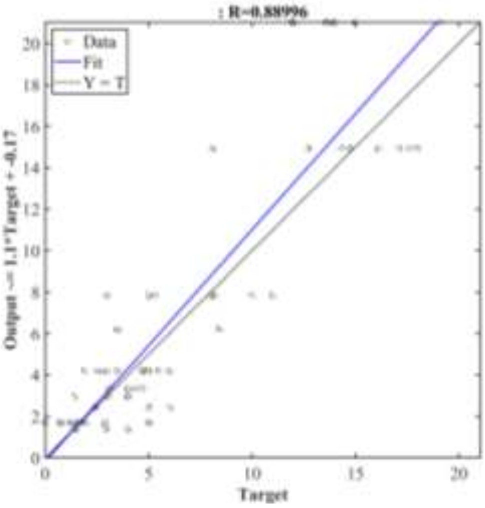

**Figure 4.** R-value at the best pruning level in the model

R-value is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of various determinations for multiple regressions. 0% indicates that the model explains none of the variability of the response data around its mean. When the Figure 4 is analyzed, we see that R-value was found to be 0.89 for CBR output in prediction analyses.

**Table 2.** Measured CBR and mean squared errors in each pruning levels and its R-values.

| **Data Num.** | **Measur. CBR** | **Mean Squared Errors** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Level0 CBR** | **Level1 CBR** | **Level2 CBR** | **Level3 CBR** | **Level4 CBR** | **Level5 CBR** |
| 1 | 5,00 | 10,89 | 9,04 | 8,30 | 4,60 | 2,80 | 0,70 |
| 2 | 1,50 | 0,01 | 0,01 | 1,40 | 1,84 | 3,33 | 18,78 |
| 3 | 1,40 | 0,00 | 0,00 | 1,64 | 2,12 | 3,71 | 19,66 |
| | … | … | … | … | … | … | … |
| 62 | 16,10 | 1,36 | 1,36 | 1,36 | 1,36 | 1,36 | 105,39 |
| 61 | 14,70 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 78,61 |
| 62 | 17,60 | 7,11 | 7,11 | 7,11 | 7,11 | 7,11 | 138,44 |
| **Averages** | | 9,46 | 9,52 | 9,45 | 9,75 | 11,56 | 27,04 |
| **R-Values** | | 0,89 | 0,88 | 0,87 | 0,86 | 0,84 | 4.9e-30 |

## 4. Conclusions

We analyzed estimation of CBR from soil parameters in the regression analysis using decision tree method in the study. The preference reason of this algorithm was that it investigated the dataset at the smaller subsets and other regression algorithms like ANN was already used for CBR test analysis in literature. Fine grain values used as input were observed to be very determinative for regression analyzes. Moreover, decision tree regression analysis estimation indicated strong correlation (R=0.89) between the output and target. Furthermore, it was shown that the correlation equations obtained are satisfactory agreements with the test results and R-values were greatly reduced by pruning. Consequently, it was observed that using some soil parameters together with the compaction parameters increased the reliability of the results in the study, and the success was outperformed than the studies in the mentioned literature.

## Acknowledgements

This study was published as a conference paper at International Advanced Researches and Engineering Congress in 2017.

## References

[1] S. F. Brown, “Soil mechanics in pavement engineering,” *Géotechnique*, vol. 46, no. 3, pp. 383–426, 1996.

[2] B. Yildirim and O. Gunaydin, “Estimation of California bearing ratio by using soft computing systems,” *Expert Syst. Appl.*, vol. 38, no. 5, pp. 6381–6391, May 2011.

[3] E. Alpaydın, *Introduction to Machine Learning*. Cambridge, Massachusetts London England, 2004.

[4] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, “A regression tree approach using mathematical programming,” *Expert Syst. Appl.*, vol. 78, pp. 347–357, Jul. 2017.

[5] H. Sun and X. Hu, “Attribute selection for decision tree learning with class constraint,” *Chemom. Intell. Lab. Syst.*, vol. 163, no. February, pp. 16–23, 2017.

[6] M. Czajkowski and M. Kretowski, “The role of decision tree representation in regression problems - An evolutionary perspective,” *Appl. Soft Comput. J.*, vol. 48, pp. 458–475, 2016.

[7] B. Yildirim, “Kaliforniya Taşıma Oranının Regesyon Analizleri ve Yapay Sinir Ağları ile Belirlenmesi,” Nigde Unviersity, 2009.

[8] T. Taskiran, “Prediction of California bearing ratio (CBR) of fine grained soils by AI methods,” *Adv. Eng. Softw.*, vol. 41, no. 6, pp. 886–892, 2010.

[9] M. Aytekin, *Soil mechanics*. Trabzon, Turkey: Academy Publishing house, 2000.

[10] D. A. 1883-99, “Standard test method for CBR of laboratory-compacted soils,” 2003.

[11] S. K. Das and P. Basudhar, “Prediction of residual friction angle of clay artificial neural network,” *Eng. Geol.*, pp. 142–145, 2008.

[12] Ş. E. Şeker, “Karar Ağacı Öğrenmesi,” pp. 1–7, 2017.

[13] Y. Bengio BENGIOY and Y. Grandvalet YVESGRANDVALET, “No Unbiased Estimator of the Variance of K-Fold Cross-Validation,” *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, 2004.