APPLICATION OF MACHINE LEARNING TECHNIQUES TO ESTIMATE UNSOAKED CALIFORNIA BEARING RATIO IN EKITI CENTRAL SENATORIAL DISTRICT

Abstract

This paper investigates the relationship between soil physical properties and the Un-soaked California Bearing Ratio (USCBR) of soil found in Ekiti State Central Senatorial District (ESCSD), which includes Natural Moisture Content (NMC%) Percentage Fines, Specific Gravity (SG) and Consistency Limits (LL%, PL%, & PI %). The database was prepared in the laboratory by conducting tests on ninety-nine (99) soil samples which were obtained in a burrowed pit found in the Central Senatorial District of Ekiti State. An R version 4.0.5 and R studio version 1.2.5033 was used to analyze the Artificial Neural Networks (ANNs) and Least Square Regression (LSR) in order to develop a simplified CBR model. In both models, independent layer containing six nodes (soil physical properties) and the dependent layer containing a single node (i.e. CBR) were taken. The descriptive analysis for training and testing was performed; boxplots of the variables were plotted and; sensitivity analysis was carried out. The capacity of the developed equation was evaluated in terms of error metrics MSE and RMSE. The analysis showed that both ANN and MLR models predicted CBR close to the laboratory value. However, the model without the percentage passing sieve 200 (MIC) is the best, having Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values of 614.1707 and 627.5754, respectively. From the error metrics analysis, the results showed that PL and LL is the most influential variable that affects the developed CBR model's output.

Keywords: *Ado- Ekiti, Artificial Neural Networks (ANNs), Error metrics, Least Square Regression, Soil Physical Properties, Machine Learning.*

1.0. Introduction

Highway Construction is the most critical segment in road execution processes. The highway is a relatively reliable crust built over the natural soil to support the wheel and traffic loads and provide a hard, durable and abrasion-resistant stratum (Ramasubbarao, Sankar 2013, Kamal, Bas 2021). A flexible highway consists of a series of layers involving a base course, sub-base and wearing course, which lies on the natural ground level. The subgrade is seen as the essential section of the highway (Yu, Wang et al. 2017). It is needful that the natural ground should be levelled appropriately to achieve a firm surface and to attain its maximum carrying capacity for the load of the above section of the highway and the moving load of vehicles (Venkatasubramanian, Dhinakaran 2011). It has become sacrosanct to predict the strength of subgrade soil on which the whole load of the pavement rests from the preceding. For this, California Bearing Ratio (CBR) is one of the most commonly applied approaches that are being used to evaluate the stiffness modulus, strength of the subgrade soil and designing the thickness of each segment of the pavement (Janjua, Chand 2016). When the natural ground level has a higher CBR value, it implies that the platform has more capacity to withstand more vehicular load coming over it. Ultimately, the depth of the highway segment will be small and vice versa (Shirur, Hiremath 2014). The springing-up of various infrastructural developments in most of the urban centres in Nigeria and within the Ekiti State Central Senatorial District (ESCSD) has prompted the subsoil evaluation approach for the primary purpose of safety, economy and durability of the Civil engineering infrastructure. An indirect soil testing is borne out of the lack of functional detailed geotechnical testing equipment in most of the technical and engineering institutions in Nigeria. The methods that allow the estimation of soil indexes of tropical soil or the cost needed to evaluate them are beneficial because of the few available state-of-the-art equipment for testing, Soil sampling, along with laboratory or field measurement, is very costly and time-consuming. Also, due to temporal and spatial variations, an appropriate assessment of the properties involves massive spending (Namdarvand, Jafarnejadi et al., 2013). Hence, many researchers attempt to obtain the possible relationship between costly measured properties with the other soil properties, which in pedology, are known as transfer functions. The transfer functions convert the primary data obtained from soil studies into properties obtained from soil (it involves much cost and time). These transfer functions have been developed to estimate the chemical, biological and mechanical properties of the soil. A lateritic soil can be vividly explained as tropical weathering products, with red, reddish-brown, or dark brown colour, with

or without nodules or concretions, generally (but not exclusively) found below hardened ferruginous crusts or hardpan (Huat, Gue et al. 2004). Various researches have been performed on the geotechnical parameters and laterite soil behaviour. Soil survey ranges from visual inspection of trail pits to an extensive borehole investigation with deep and numerous boreholes and extensive sampling and testing of the soil, usually by evaluation experts.

Soil investigation is an essential aspect that must be carried out before starting the actual construction work. Tests to examine the engineering properties of soil in situ are a valuable means of investigation since these parameters can be determined directly without the disturbing effect of boring and sampling (Fall, Tisot et al. 1995). The CBR test consumes more time and saps more energy in the laboratory, particularly when soil unsuitability cannot be ascertained. To solve this problem, it is essential to establish a relationship between the CBR values of the soil with its physical parameters such as (% fines), Consistency limits: PL (Plastic Limit) LL (Liquid Limit), SG (Specific Gravity), and NMC (Natural Moisture Content). The study has shown a relationship between the CBR value of Ekiti Central Senatorial District soil and its physical properties. Least Square Regression (LSR) and Artificial Neural Networks (ANN) models were utilised to detect this. A correlation has been established with a simplified model in this research to estimate an un-soaked California Bearing Ratio of tropical lateritic soil.

2. Methodology

2.1. Multiple Linear Regression Analysis

An MLR expresses the correlation between one continuous output variable and two or more input variables. CBR is the output variable, and others are input variables in this study. In the model, the CBR value is the subject of all other physical parameters. The expression is as shown below:

CBR = f (NMC, LL, PL, GS % Fines) ……..... (2. 1)

The equation will be created as follows:

$$Y = bo + b1x1 + b2x2 + b3x3 + b4x4 + \ldots\ldots. Bnxn \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2)$$

Where bo, b1, b2, b3, b4, bn are constants, Y is CBR and, x1, x2, x3, x4, xn are soil properties considered for analysis. These constant values can be obtained using available data Analysis R software to get a suitable equation (Rakaraddi, Gomarsi 2015).

2.2. Artificial Neural Network Model

In the past decades, due to difficulties in solving complex engineering system, Artificial Neural Network (ANN) has been applied by many researchers to analyze the human brain and nervous system behavior. ANN structure has three main layers: a collection of input nodes, a layer or layers of hidden nodes, and a collection of output nodes. The possibility of using ANN for the estimation of CBR was investigated by building various appropriate ANN models. Variables (parameters) that belong to two categories of soil index properties, which reflect compaction properties and plasticity, are implemented. Therefore, entirely five basic soil parameters such as percentage passing 75-micron sieves (MIC); Consistency Limits (PL %, LL % & PI %), Specific Gravity (SG), and Natural Moisture Content (NMC %), were considered as independent variable indices for the ANN models. The online available R Software was used to perform the computations being essential. To develop the best proper ANN architecture in each model, the Neurons number in the hidden layer and different transfer functions were attempted to get the best prediction of CBR values. They have been, consequently, varied until the convergence was achieved in the mean squared error.

3.0. Experimental Programme

Among other tests, to get data for the model's establishment, 99 CBR test data belonging to different soil groups were obtained. Tests were performed for the viability evaluation of the soils for using it as a base material. Therefore, tested soil samples were obtained in Ekiti Central Senatorial District in Ekiti State, Nigeria. The tests were carried out on samples at the Federal Polytechnic Ado-Ekiti Soil Mechanic Laboratory, Civil Engineering Department, according to AASHTO and BS 1377 (1990) Specifications and R version 4.0.5 and R studio version 1.2.5033 was used for this research work [R Core Team 202]

4.0. RESULTS AND ANALYSIS

4.1. SOIL CLASSIFICATION

The analysis was carried out based on the data obtained from 99 sample locations in the Ekiti State Central Senatorial District (Nigeria). The soil data collected were processed on soil index

and strength properties in the community mentioned above. Fig.1 shows the research areas location, which reveals the urban area of Ado-Ekiti and its environment. Table SM1 (a) to Table SM 1(e) shows the test results survey and their classification based on AASTO and USC system. The soil samples of the study area are characterised as low plastic, sandy, gravelly clay, and others, as medium compressible soil according to the Unified Soil Classification System (USCS). In the AASHTO classification system, the soil samples are essentially graded as 'Excellent' to 'Good' (A-2-4), 'Fair' to 'Poor' (A-2-6) and 'Clayey soil' (A-6 - A-7), respectively.

4.2. ANALYSIS ON EFFECTS OF NMC, % PASS 75 MICRONS (MIC), GS, LL AND PL ON CBR USING NEURAL NETWORK

Tables SM 2 and Table SM 3 represent the summary of the descriptive statistics on the training and test data sets. The mean and median in all the predictor variables are close to one another; this means that the consistency is moderate. The same pattern is witnessed in the response variables. The small values of standard deviation show that the values cluster around the mean very well. The higher value of sample variance 24.3483, 285.1336, 109.3454, 78.6479, 417.7472, and NMC, MIC, LL, PL, and CBR, indicate a large spread on the data; this shows that many of the values do not cluster around the mean but spread out farther away. On the other hand, Kurtosis measures the amount of probability in the tails reported in the tables above. It can be observed that all the variables have their Kurtosis less than 3; this means that the data sets do not have heavy tails but are typically distributed.

Skewness is a measure of the distribution of values around the mean. If skewness is less than -1 or greater than 1, then the distribution is highly skewed. If it is between -0.5 and -1 or between 0.5 and 1, it is moderately uneven. From the values of skewness reported in the above tables, it can be observed that skewness is not evident in NMC, % passing 75 microns, and GS, but pronounced in LL, PL, and CBR. Some are skewed to the left, while others to the right, therefore the skewness fell within moderately skewed description. As shown in Figure SM 1, the chart reveals that the values are coherent in each of the variables and among the variables. Outliers are found in CBR, NMC, LL, and PL

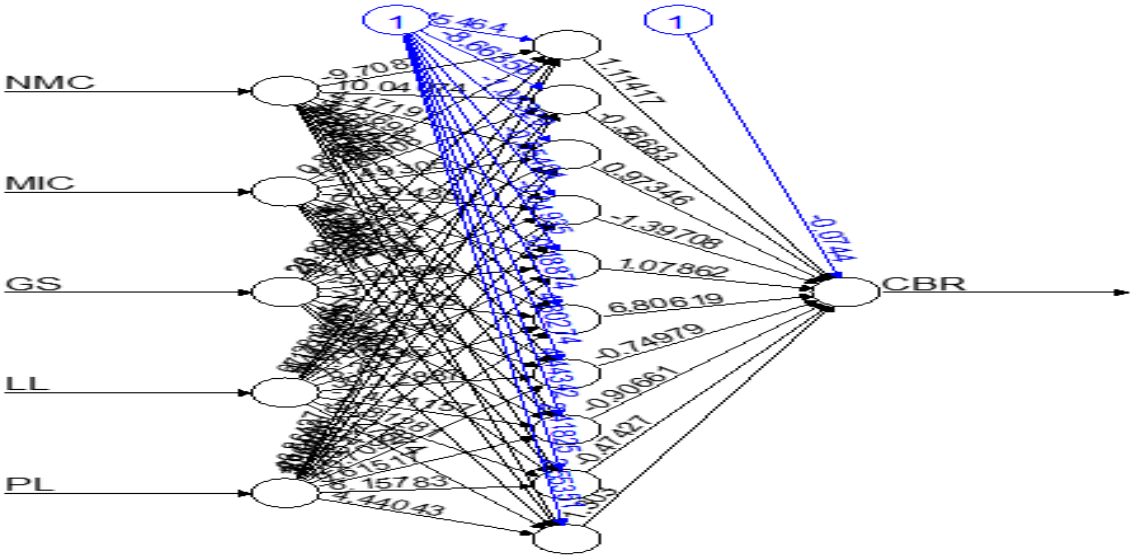

Figure 1 Neural network plot of CBR

The leftmost nodes, that is, the input node, are the raw data variables as shown in Figure 1: NMC, % Passing 75-micron sieve, GS, LL, and PL. the arrows in black and the associated values are the weights which are the contributions of the variables to the next node. The middle nodes (no matter how many) are the hidden nodes. The blue lines stand for the bias weights. Each of these nodes constitutes a component that the network is learning to recognise. The far-right node is the output node; it is the final output of the neural network. The sole role is to supply all the nodes with a trainable constant value (in addition to the primary inputs that the node receives). This can be achieved with a single bias node with connections to N nodes or N bias nodes, each with a single link; the result should be the same. Weights in an ANN are the most critical factor in converting input to impact the output. This is similar to slope in linear regression, where weight is multiplied to the input to form the result. Weights are numerical parameters that determine how strongly each of the neurons affects the other. For a typical neuron, if the inputs are *x1*, *x2*, and *x3*, then the synaptic weights to be applied to them are denoted as *w1*, *w2*, and *w3*.

Dependent is $y = f(x) = \sum x_i w_i$ Where *i* starts from 1 to n inputs.

Bias is like the intercept added in a linear equation. An additional property is used to adjust the dependent variable and the weighted sum of the independent variable to the neuron.

The processing done by the neuron is as shown below:

*Dependent = sum (weight*Independent) + bias …… eqn*

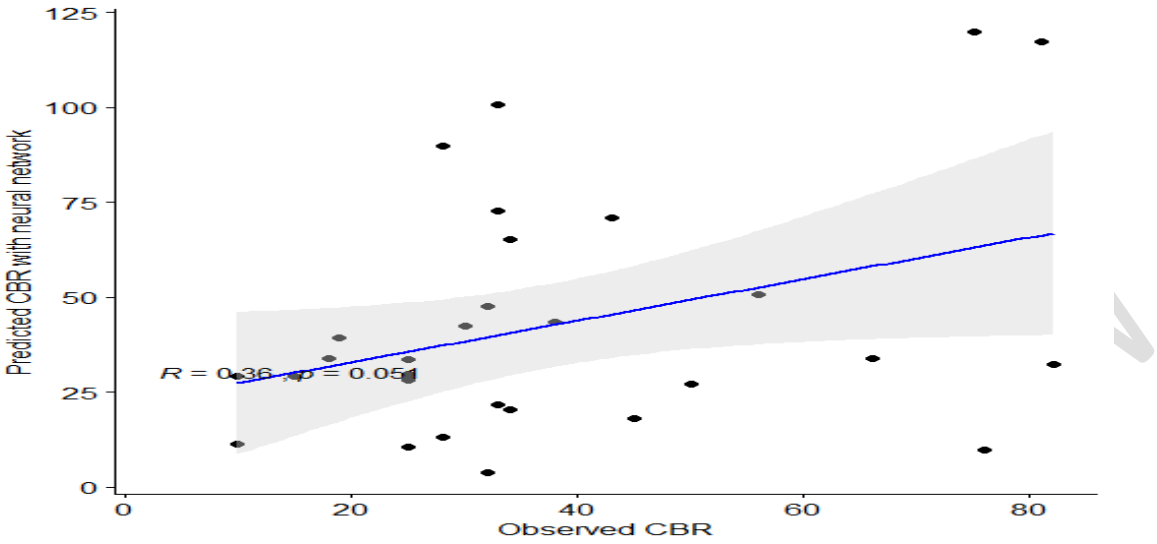

Fig. 2: Correlation between the predicted and the observed of the full model using Neural Network method. It can be summarised from *fig. 2* that some observations cluster around the line though many are found further away from it. It can be summarised from *fig SM. 3* that some observations cluster around the line, though many are found further away from it.

4.3. Effect of NMC, MIC, GS, LL, and PL on CBR Using Multiple Linear Regression Models

Variables LL and PL are the only inputs contributing significantly to the output CBR at 5% level. However, the estimates are low, as shown in *Fig SM 5* and *Fig SM 4*, representing the residual versus fitted and the Q-Q plots showing linearity and normality by horizontal and almost straight lines.

5.0. Sensitivity Analysis

5.1. Effect of Reduction of Covariates on the Model

Table 1: Removal of covariate GS from the equation

	Estimate	Std. Error	t value	P-value
(Intercept)	51.025	10.466	4.875	0
NMC	-0.266	0.489	-0.544	0.588
MIC	0.006	0.161	0.037	0.971
LL	-0.924	0.349	-2.645	0.010
PL	1.015	0.425	2.387	0.020

Source: Authors Laboratory works

Table 2: Removal of covariate NMC from the model

	Estimate	Std. Error	T-value	P-value
(Intercept)	81.429	41.623	1.956	0.055
MIC	-0.024	0.163	-0.146	0.885
GS	-13.175	16.283	-0.809	0.421
LL	-0.89	0.352	-2.527	0.014
PL	0.978	0.428	2.285	0.026

Source: Authors Laboratory works

None of the covariates is significant, but GS contributes most, though not significantly, according to *Table SM 6* compared with *Table 1* where PL and LL contribute significantly to CBR, having the lowest p-values and highest estimates. *Table SM 8*: Removal of covariate %passing 75-micron sieve from the equation showed that LL and PL contribute significantly to CBR, having the lowest p-values and highest estimates. Also, *Table 2:* Removal of covariate NMC from the model reveals that LL and PL contribute significantly to CBR, having the lowest p-values and highest estimates.

Table 3: Information criteria to select the best model

Model	**AIC**	**BIC**
Full model (Least square)	616.1500	631.7900
without PL	619.3470	632.7516
without LL	620.4648	633.8695
without GS	614.9783	628.3829
without MIC (% passing 75 MIC)	614.1707	627.5754
without NMC	614.5941	627.9988

Source: Authors Laboratory works

According to (Robert et al. 1995) and (Ying Tang 2005), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) model selection criteria should be adopted in the field of practice. Therefore, from the table above, the model with the least values of information criteria was selected as the best, which was the model without the percentage passing 75-micron sieve (Silt & Clay) having Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values 614.1707 and 627.5754, respectively.

6.0. Error Metrics

6.1. Mean Square Error

Mean square error (MSE): It measures the average of the error square, the average square difference between the estimated and actual values.

$$MSE = \frac{\sum_{i=1}^{n}\left(A_t - F_t\right)^2}{n}$$, where A = actual values, F = forecasts or prediction, n = number of Observations, t = period

6.2. Root Mean Square Error

The root means square error (RMSE) is a quadratic scoring rule that measures the average magnitude of the error. It is the square root of the average of squared differences between forecast and actual, i.e., the MSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(A_t - F_t\right)^2}{n}}$$

6.3. Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) is the mean of absolute errors over the actual observed values.

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|}{n} \times 100$$

6.4. Mean Absolute Error

Mean Absolute Error (MAE) estimates the mean magnitude of the errors in a set of predictions without considering their direction. It is the average over the test sample of the absolute differences between forecast and actual observation where all individual differences have equal weight.

$$MAE = \sum_{i=1}^{n}\left|A_t - F_t\right|$$

Table 4: Testing the predicting power of the models using Error metrics.

Model	R square	MAE	MSE	RMSE	MAPE
Full model (Neural network)	0.0362	24.0078	919.1906	30.3182	0.6919
Full model (Least square)	0.1231	15.2178	370.9686	19.2605	0.4921
Without PL	0.0545	16.2215	**405.7333**	**20.1428**	0.5412
Without LL	0.0390	**17.0423**	388.0133	19.6981	**0.5923**
Without GS	0.1125	15.2816	378.3501	19.4512	0.4967
Without MIC	0.1231	15.1778	371.9025	19.2848	0.4896
Without NMC	0.1174	14.8828	370.8554	19.2576	0.4853

Source: Authors Laboratory works

From Table 4, using the error metrics MSE and RMSE, it can be seen that PL is the most influential variable that affects the output CBR.

6.5. Prediction Using the Best Model

With reference to Tables 1 and Table 2, the prediction model is as follows:

$$CBR = 85.737 - 0.315 * NMC - 13.915 * GS - 0.878 * LL + 0.944 * PL$$

7.0. Conclusions

The following conclusions were drawn after the statistical analysis:

i. The descriptive analysis showed that all the variables have their Kurtosis less than 3; the data sets do not have heavy tails but are usually distributed.
ii. The skewness is not evident in NMC, % passing 75 microns, and GS, but pronounced in LL, PL, and CBR. Some are skewed to the left while others to the right; therefore, the skewness fell within moderately skewed.
iii. The box plot of the variables reveals that the values are coherent in each of the variables and among the variables.
iv. The diagnostic properties of residuals on the entire model showed residual versus fitted. The Q-Q plots showed linearity and normality by their horizontal and almost straight lines, respectively.
v. The sensitivity analysis showed that LL and PL contribute significantly to CBR, having the lowest p-values and highest estimates.
vi. The model with the least values of information criteria was selected as the best, which was the model without the percentage passing 75-micron sieve having Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values 614.1707 and 627.5754, respectively.
vii. The error metrics MSE and RMSE PL was the most influential variable that affects the output CBR.
viii. With reference to tables 7 and 9, the simplified prediction model was correlated as follows:
$$CBR = 85.737 - 0.315 * NMC - 13.915 * GS - 0.878 * LL + 0.944 * PL$$

Refringences

FALL, M., TISOT, J. and CISSE, I.K., 1995Specifications for road design using statistical data, an example of laterite or gravel lateritic soils from Senegal, *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts* 1995, pp. 341A.

FRANK J. FABOZZI, SERGIO M. FOCARDI, SVETLOZAR T. RACHEV and BALA G. ARSHANAPALLI. (2014). *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications.* John Wiley & Sons, Inc.

HUAT, B.B., GUE, S.S. and ALI, F.H., 2004. *Tropical residual soils engineering.* CRC Press.

JANJUA, Z.S. and CHAND, J., 2016. Correlation of CBR with Index properties of soil. *International Journal of Civil Engineering and Technology (IJCIET)*, **7**(5), pp. 57-62.

KAMAL, I. and BAS, Y., 2021. Materials and technologies in road pavements-an overview. *Materials Today: Proceedings*, .

NAMDARVAND, F., JAFARNEJADI, A. and SAYYAD, G., 2013. Estimation of soil compression coefficient using artificial neural network and multiple regressions. *Int Res J Appl Basic Sci*, **4**(10), pp. 3232-3236.

RAKARADDI, P.G. and GOMARSI, V., 2015. Establishing relationship between CBR with different soil properties. *International journal of research in engineering and technology*, **4**(2), pp. 182-188.

RAMASUBBARAO, G.V. and SANKAR, G.S., 2013. Predicting soaked CBR value of fine grained soils using index and compaction characteristics. *Jordan Journal of Civil Engineering*, **159**(3164), pp. 1-7.

ROBERT E. KASS and ADRIAN E. RAFTERY, "Bayes Factors," *Journal of the AmericanStatistical Association* 90, no. 430 (June 1995): pp 773–795

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R project.org

SHIRUR, N.B. and HIREMATH, S.G., 2014. Establishing relationship between CBR value and physical properties of soil. *IOSR journal of mechanical and civil engineering*, **11**(5), pp. 26-30.

VENKATASUBRAMANIAN, C. and DHINAKARAN, G., 2011. ANN model for predicting CBR from index properties of soils. *International Journal of Civil & Structural Engineering*, **2**(2), pp. 614-620.

YING YANG, "Can the Strengths of AIC and BIC Be Shared?" *Biometrika* 92, no. 4(December 2005): pp 937–950.

YU, H., WANG, Y., ZOU, C., WANG, P. and YAN, C., 2017. Study on subgrade settlement characteristics after widening project of highway built on weak foundation. *Arabian Journal for Science and Engineering*, **42**(9), pp. 3723-3732.