Estimation of Compaction Characteristics Using Machine Learning Techniques

Akinwamide Joshua Tunbosun
Department of Civil Engineering, Faculty of Engineering
The Federal Polytechnic
Ado Ekiti, Ekiti State, Nigeria
akinwamidejoshua2@gmail.com

Jacob Odeh Ehiorobo, Ph.D. Osuji Sylvester Obinna. Ph. D. and Ebuka Nwankwo. Ph.D.
Department of Civil Engineering, Faculty of Engineering
University of Benin
Benin City, Edo State, Nigeria
jacehi@uniben.edu, osujisyl@yahoo.com, ebuka.nwankwo@uniben.edu

Abstract

Application of Machine learning which is the main driver of Artificial Intelligence is gradually gaining ground in the field of geotechnical engineering for prediction and forecasting. In the present research, efforts are made to explore some machine learning techniques such as: Support Vector Machine, Random Forest, Artificial Neural Networks, M5 Tree and Multiple Linear Regression models to correlate maximum dry density and optimum moisture content with somesoil index properties where the best predicting model was determined. Four hundred and eighty (480) soil samples were obtained and divided into data set using training and validation of the developed models from the basic soil parameters. The Goodness of fit between the Actual and the Predicted Valuesshowed that the values of Root Mean Square Error 253.84, 295.44, 218.08, 101.63, 211.12 and 3.91, 4.55, 3.54, 1.67, 20.13. are from Support Vector Machine; Random Forest, Artificial Neural Networks, M5 Tree and Multiple Linear Regression models for Maximum Dry Density and Optimum Moisture Content values respectively. The least values 101.63 and 1.67 was observed from Random Forest for Maximum Dry Density and Optimum Moisture Content. Similarly, the correlation coefficient valuesrange between 0.15 – 0.96 and 0.2 ─0.96. From the foregoing, the correlation coefficient value proved the Random Forest model as the bestfor estimating Maximum Dry Density and Optimum Moisture Content while the model having the least correlation coefficient is the Multiple Linear Regression model for Maximum Dry Density and Support Vector Machine for Optimum Moisture Content respectively. It's concluded that this research work will be an excellent guidetoConstructors, Future planners and Civil engineers for estimation of unavailable data, and for cross checking the observed values particularly at the project preliminary stageswithin the study area.

Key words: Compaction characteristics, Correlation coefficient, Machine Learning, Optimum Moisture Content and Maximum Dry Density.

1. Introduction

Holtz et al. (2010) defines soil compaction as a mechanical process which accounts for the increase in the density of soil by reducing the air volume from the pore spaces. According to him, this results in the changes that occur in pore space size, particle distribution, and the strength of the soil. Rollings and Rollings (1996) opined that, the major purpose of the compaction process is the increase in strength and stiffness of the soil by reduction in the compressibility and decrease in the permeability of the soil mass by its porosity. The type of soil and the grain sizes of the soil play a significant role in compaction process as a reduction in pore spaces within the soil increases the bulk density. Soil types with higher percentages of clay and silt have a lower density than coarse-grained soil since they naturally have more pore spaces. Going by Hausman (1990) the compaction curve obtained in the laboratory tests or field compaction is a representation of the typical moisture-density curve which clarifies the compaction characteristics theory. The importance of this property is well appreciated in the construction of earth dams and other earth filling projects. It is a vital process and is employed during the construction projects such as; highway, railway

subgrades, airfield pavements, landfill liners and in earth retaining structures like Tailings Storage Facility (TSF). Considerable time, effort and cost is used during a compaction test in order to determine the optimal properties i.e. maximum dry unit weight and optimum water content hence, there is the need to develop predictive models using simple soil tests like Atterberg limit tests and Gradation tests especially, when these are known already from project reports, bibliographies, and from database of the engineering properties of quarried soil within the geographical area or soil samples of similar properties. Past authors have made comparison between artificial neuron network and multiple linear regression in the field of Geotechnical Engineering; for example, (Harini*etal* 2014) compared them for estimation of California Bearing Ratio(CBR) of fine-grained soils; (Boadu *e.tal.* 2013), (Siddiqui*etal* 2014) tried to estimate geotechnical indices from electrical measurements using both models and compared theirresults. Measurement of the compaction characteristics and California Bearing Ratio (CBR) of soil inthe laboratory is neither time-efficient nor cost-efficient.The growing need for a predictive model as an alternativeto laboratory testing was the impetus for motivation in thisproject. In a geotechnical engineering project, an accurate prediction of the maximum dry density (MDD), optimum moisture content (OMC), CBR for soaked and Unsoaked will not only save time, but will also help reduce the costs, cut down on the use of resources, and lessen the required human labor.

The predicted maximum dry unit weight and optimum water content can be used for the preliminary design of the project. Correlations are generally derived with the help of statistical methods using data from extensive laboratory or field testing. Linear Regression (LR) Analysis, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and M 5 model trees (M5P); are some of the types of machine learning techniques. These techniques learn from data cases presented to them to capture the functional relationship among the data even if the fundamental relationships are unknown or the physical meaning is tough to explain. These techniques are being used globally to solve different civil engineering problems (Anbazhagan et al. 2016), (Singh et al. 2017), (Prasad et al. 2017). Compaction characteristic are extensively used for the design of earthen dams, embankments, pavements, landfill liners and foundation of various Civil Engineering structures. Most of these parameters are determined in the laboratory and some are estimated on the field. Their calculation requires a specific laboratory equipment; an experienced geotechnical engineer, with a team of skilled technicians. Thus, determination of these parameter is costly and time consuming. Also, soil is a highly erratic material as its performance is based on the processes due to which it is formed. Hence, correlations developed for one region may not be applicable for the other. This ascertains the need to develop region-based correlations to predict geotechnical properties. In the present study compaction parameters have been estimated using MLR, SVM, ANN, M5 TREE and RFfrom Some geotechnical datawhich were obtained from the three Senatorial Districts of Ekiti Southwestern Nigeria for development of more accurate models.

2. Material and Methods

Four Hundred and Eighty (480) soils dataset used for this research were obtained from the three Senatorial Districts of Ekiti State, at an average depth of 1.2 m were tested in compliance with BS 1377 (1990) at the material testing laboratory of Federal Polytechnic Ado-Ekiti, to determine both the index properties (NMC, Gs, Gravel, Sand, Fine, PL and LL) and the compaction characteristics (OMC and MDD) respectively. The developed geotechnical data base has information for 480 trial pit'slocations for developments of different models and statistical relationships using 70 % of the data for models building /training and 30 % for model evaluation. Partitioning was performed on the data for variable standardization. R version 4.0.5 (R core team, 2021) and R studio version 1.2.5033 were used for statistical analysis. The Correlation Coefficient R and the root Mean Square Error (RMSE) were used to measure the performance of the predictions since the Correlation coefficient is a key function to establish a relative relationship between the expected and the observed data Shahin et al, 2008; Smith, 1986 this was established by plotting the experimented and predicted values on vertical and horizontal axis respectively Pinheiro et al, 2008 forthe developed equations to measure their individual efficiency. The research methodology flow chart is as shown in figure 1

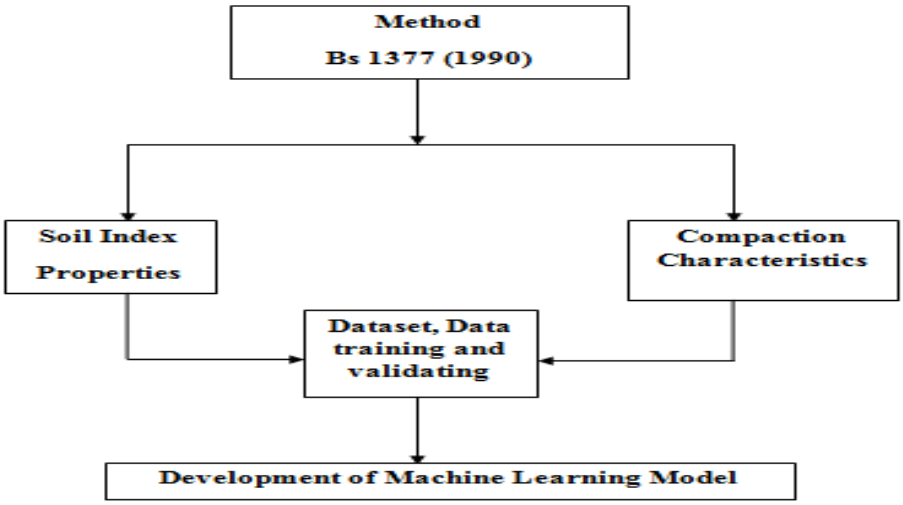

Figure 1: Development of compaction characteristics procedural flow chart

3. Results and Discussion

3.1 Measurement of Interrelationship among the Predictor

It is established that there should not be any significant relationship among the independent variables for prediction when using Multiple Linear Regression. Figure 2a showed a Scatter matrix showing interrelationship among the predictors where the Lower triangles provide scatter plots and upper triangles provide correlation values.**Gravel and Fines**, **Sand and Fines** are highly correlated. Similarly, **LL and PL** are also highly correlated which leads to multi co-linearity issues. Predicting the model based on this dataset may be erroneous. However, one way of handling these kinds of issues is based on principal component analysis (PCA) as shown in figure 2a and 2b . In the scatter matrix all the obvious relationship among the input variables is gone, hence there exists no multi co-linearity among them. Zero correlation means there is no significant relationship among the predictors. This serves as a good foundation for multiple linear regression analysis.

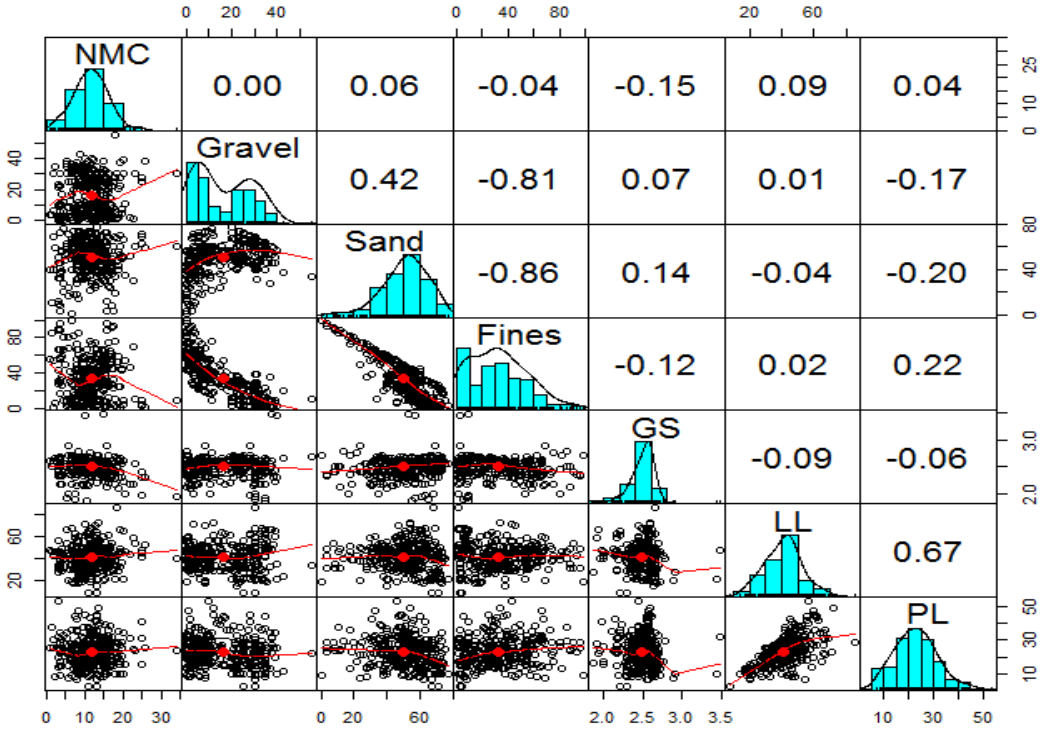

Fig. 2a: Scattered matrix showing interrelationship among the predictors

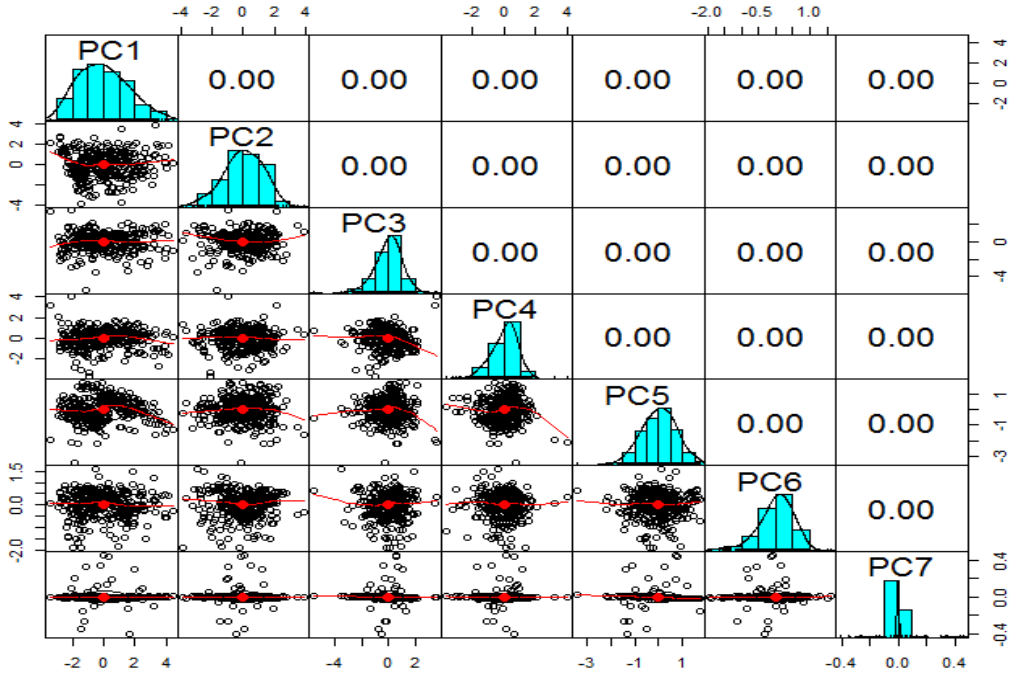

Fig. 2b: Scattered matrix showing no relationship among the predictors

3.1.1 Principal Component Analysis (PCA)

The principal components are the linear combinations of the original variables that account for the variance in the data. It is based only on the independent variables, so we removed the response variable from the dataset. The maximum number of components extracted always equals the number of variables, the eigenvectors, which comprised of coefficients used to calculate the principal component scores. The coefficients indicate the relative weight of each variable in the PCA. The eighth variable was then removed (dependent) from the dataset as shown in figure 1a and 1b for OMC and MDD respectively.

Table 1: Eigen vectors from the PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
NMC	-0.0106	-0.1877	-0.7041	0.6618	-0.1735	0.0291	0.0022
Gravel	-0.4944	-0.1851	0.0097	-0.2370	-0.6940	0.0826	0.4202
Sand	-0.5243	-0.1413	-0.0184	0.1055	0.6744	0.0381	0.4872
Fines	0.6049	0.1916	0.0104	0.0630	-0.0483	-0.0704	0.7655
GS	-0.1361	0.1102	0.6698	0.6996	-0.1694	-0.0515	-0.0021
LL	0.1431	-0.6968	0.1244	-0.0292	0.0128	-0.6910	-0.0007
PL	0.2751	-0.6133	0.1991	0.0233	0.0471	0.7112	0.0000

Table 2 Eigen Analysis of the Correlation Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.596	1.2718	1.0581	0.919	0.75974	0.54974	0.0634
Proportion of variance	0.3639	0.2311	0.1559	0.1208	0.08055	0.04317	0.00057
cumulative proportion	0.3639	0.5949	0.7549	0.8757	0.95625	0.99943	1.00000

The first principal components explain the variability around 36%, second 23%, third 16%, and fourth 12%. Summarily, the first four principal components capture the approximately 88% of the variability which is a majority of the variability as shown in table 2. In this case, the first four components capture the majority of the variability, while the remaining components contribute negligible variability. In these results, the scores for the first four principal components can be calculated from the standardized data using the coefficients listed under PC1 to PC4 as shown in Table 1 and 2 with figure 3 as showed in the screen plot the proportion of variance for picking or selecting the PCA.

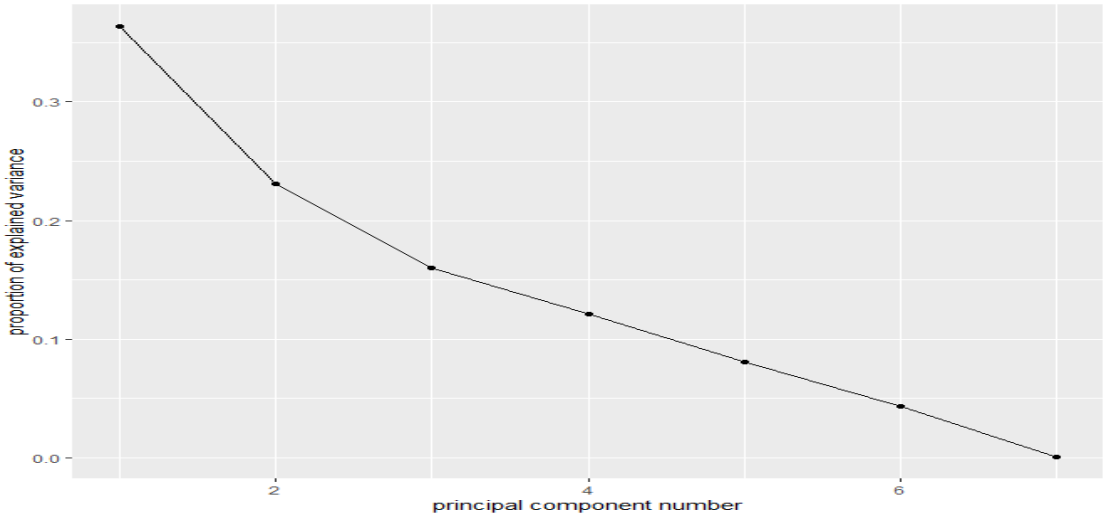

Fig. 3: Principal Component Analysis (PCA) Scree plotshowing relevant components or factors to be considered

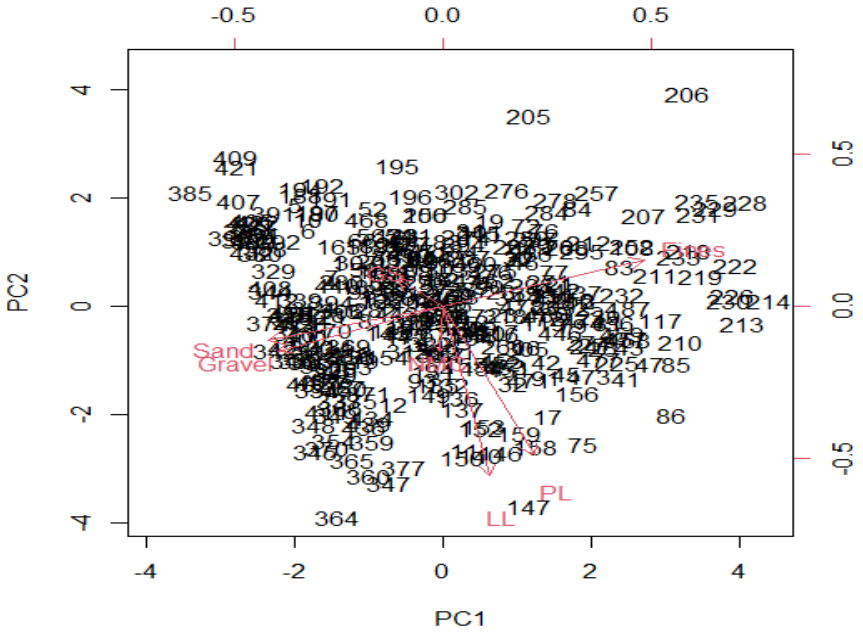

Fig. 4: Bi-plot of the components

3.1.2. Principal Components Analysis (PCA) Bi-plot

The loading plot was used to identify which variables have the largest effect on each component. Loadings can range from -1 to 1. Loadings close to -1 or 1 indicate that the variable has strongly influenced the component while Loadings close to 0 indicate that the variable has a weak influence on the component. Evaluating the loadings can also help to characterize each component in terms of the variables. In this case, Fine has a high positive relationship with the PC1 while PL and LL also have high negative relationship with PC2 as shown in figure 4.

3.1.3. The derived linear model from PC1, PC2, PC3 and PC4

The derived Model is given as theoretical and estimated model in equations (1) to (2) and equations (3) to (4) for MDD and OMC respectively, where the first four principal components were adopted since the majority of the information are present in the four components. The resulting four component score variables are representative and can be used in place of the seven original variables with a 12% loss of information.

The theoretical model for MDD

$$\text{MDD} = \boldsymbol{\alpha} + \boldsymbol{\beta_1}(\text{PC1}) + \boldsymbol{\beta_2}(\text{PC2}) + \boldsymbol{\beta_3}(\text{PC3}) + \boldsymbol{\beta_4}(\text{PC4}) + \boldsymbol{\epsilon} \quad (1)$$

The estimated model with actual coefficients for MDD

MDD = 1907.25 - 1.13(PC1) + 10.22{PC2}) + 12.48{PC3}) - 35.24{PC4}) (2)

The theoretical model for OMC

$$\text{OMC} = \boldsymbol{\alpha} + \boldsymbol{\beta_1}(\text{PC1}) + \boldsymbol{\beta_2}(\text{PC2}) + \boldsymbol{\beta_3}(\text{PC3}) + \boldsymbol{\beta_4}(\text{PC4}) + \boldsymbol{\epsilon} \quad (3)$$

The estimated model with actual coefficients for MDD

OMC = 14.23 + 1.01{PC1}) + 0.5{PC2}) - 0.29{PC3}) - 0.24{PC4}) (4)

3.2 Measures of Accuracy between the Actual and the Predicted Values (Goodness of fit)

The Correlation Coefficient R and the Root Mean Square Error (RMSE) are the major yardsticks that are usually adopted to measure the performance of any prediction where the Correlation coefficient is a key function to establish a relative relationship between the expected and the observed data (Shahin et al, 2008). (Smith, 1986) prepared the following guide to measure R ─/R/≥0.8 Strong correlation, ─0.2< /R/ < 0.8 Correlation exists, /R/≤0.2 Weak correlation and /R/ = 0 No correlation. It is obvious from the values of RMSE 253.84, 295.44, 218.08, 101.63, 211.12 and 3.91, 4.55, 3.54, 1.67, 20.13 are from; MLR, ANN, SVM, Random Forest, and M5 model for MDD and OMC values respectively. The least values 101.63 and 1.67 were observed from random forest (RF) for MDD and OMC.

Similarly, the R values range between 0.15 – 0. 96 and 0.2 ─0.96 as reflected in table 3 to table 6 and figures 7 and 8, which established the relationship among the predicted and the actual OMC and MDD; using the five machine learning models. The correlation coefficient values deduced the Random Forest Model for OMC and MDD as the best, while the model having the least correlation coefficient is the Multiple Linear Regression (MLR) model for MDD and Support Vector Machine (SVM) for OMC respectively.

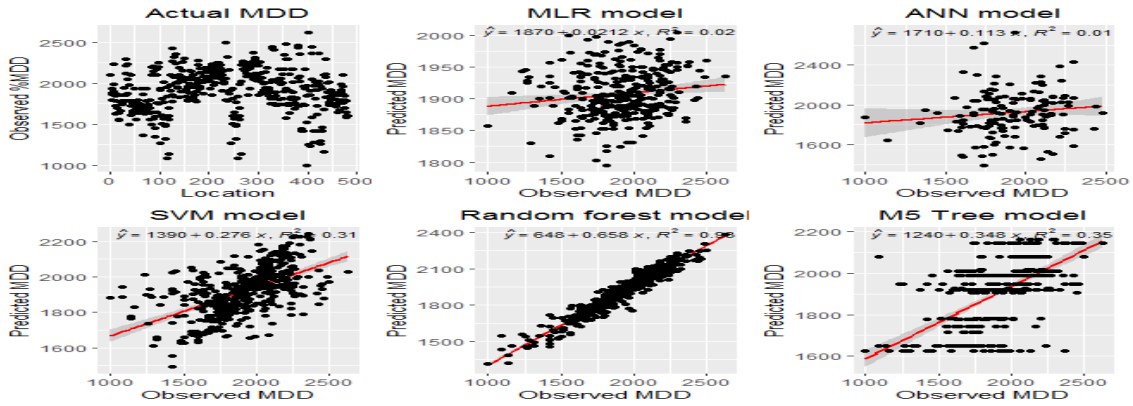

Fig. 5: Scattered plots for the performance analysis of the models for predicting Maximum Dry Density (MDD kg/m^3)

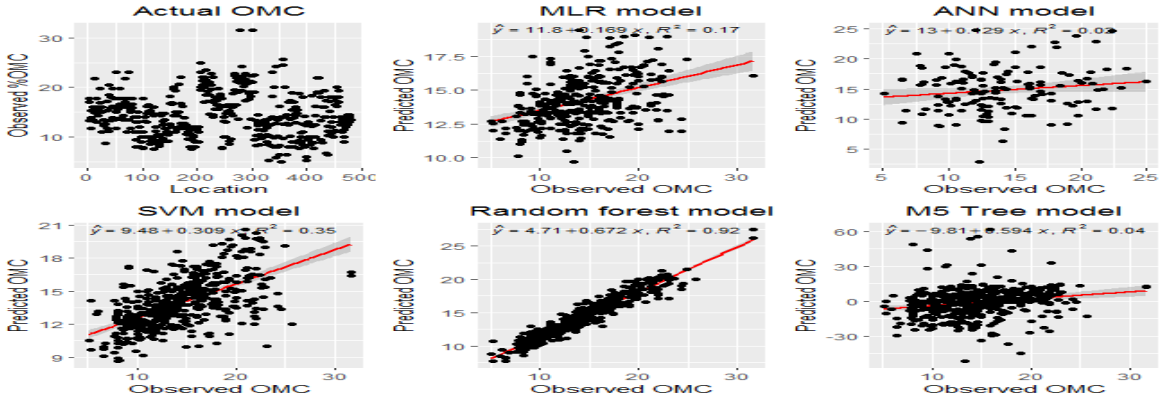

Fig 6: Scatter plots showing performance of the model in terms of correlation coefficients for Optimum

Table 3: Measure of accuracy (Goodness of fit) for OMC

Techniques	Soil	Goodness of fit					
	Indices	ME	MAE	MSE	RMSE	R	R^2
MLR	OMC	0	3.15	15.25	3.91	0.41	0.17
ANN	OMC	0.48	3.7	20.69	4.55	0.37	0.14
MSTREE	OMC	0.42	2.58	12.56	3.54	0.59	0.35
R F	OMC	0.01	1.28	2.28	1.67	0.96	0.92
SVM	OMC	-15.6	17.14	405.14	20.13	0.2	0.04

Table 4: Measure of accuracy (Goodness of fit) for MDD

Techniques	Soil	Goodness of fit					
	Indices	ME	MAE	MSE	RMSE	R	R^2
MLR	MDD	0.00	201.29	64435.71	253.84	0.15	0.02
ANN	MDD	0.14	222.73	87286.55	295.44	0.28	0.08
MSTREE	MDD	1046.96	1503.43	360.32	1898.21	0.10	0.01
R F	MDD	-3.45	76.73	10329.04	101.63	0.96	0.93
SVM	MDD	14.62	149.39	47557.59	218.08	0.56	0.31

Table 5: Predictions from the five Machine Learning (ML) models For OMC

ACTUAL OMC	Pred OMC MLR	Pred OMC SVM	Pred OMC RF	Pred OMC MS TREE
13.3	13.525	14.602	14.087	-1.433
17.7	12.795	13.46	16.615	0.8769
17	12.916	12.905	15.171	-10.77
13.9	13.173	12.894	13.487	-10.3
11.8	12.89	11.732	13.116	1.6723
15.3	13.46	11.861	14.269	-1.818

Table 6: predictions from the five machine learning (ML models for MDD)

ACTUAL MDD	Pred MDD MLR	Pred MDD SVM	Pred MDD RF	Pred MDD MS TREE
2100	1919.7	1874.7	1992.6	4346.2
1935	1919	1890	1921.7	3829
1867	1926.6	1906.8	1878.5	4511
1887	1923.7	1925.4	1916.5	3888.3
1798	1933.6	2050.2	1860.4	2852.4
1994	1953.1	1998.2	1952	3397.5

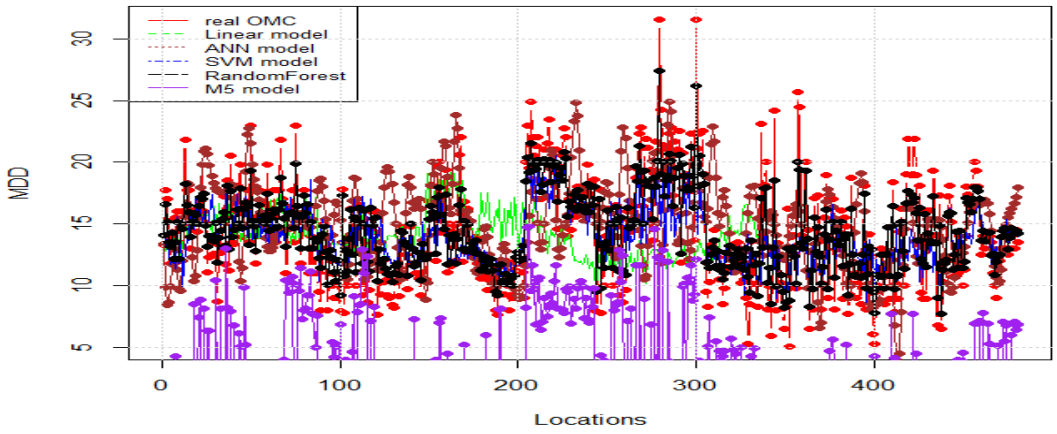

Fig.7: Line plot showing the movement of the observed and the predicted

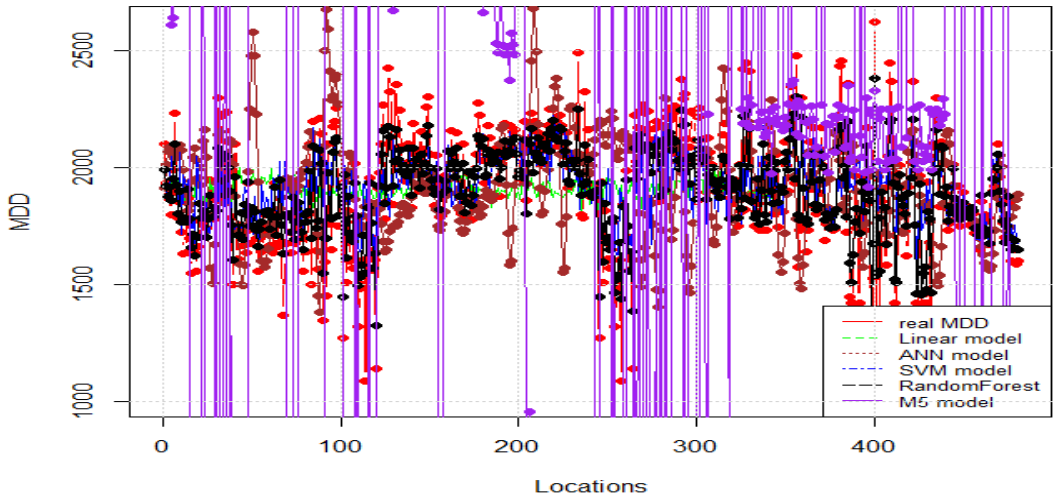

Fig.8: Line plot showing the movement of the observed and the predicted

3.2 Comparison of Models

The predicted values generated by Random Forest (RF) model seems to move side by side with the actual MDD and OMC while the MS Tree gave a worst performance as shown in figure 7 and 8 and table 3 and table 4 respectively, where the coefficient of determination R^2 gave 0.92 and 0.93. From the foregoing the results suggest a good model and of course the best among the five applied. for Random Forest (RF) while MS Tree gave the worst model.

4. Conclusions

It is observed from the results that machine learning techniques has an excellent contribution in geotechnical engineering. The multiple linear regression model which has been in existence over ten decades now, can no longer handle huge data that we use in today's geotechnical data analysis challenges. The support vector machine performed better than the MLR. ANNs while M5 tree model exhibits steps of jumped phenomenon in the predicted values of the response variable. However, it is noteworthy that Random Forest came out as the best machine learning techniques for the estimation of compaction characteristics in this research work using the correlation and the performance tests. The developed model in the present work relates MDD and OMC with some soil index

properties. The results reveal a high correlation coefficient R and could judiciously be used for estimating OMC and MDD of a regional soil. It gives a very good estimate of MDD and OMC without actually performing the test.

References

Anbazhagan, P., Uday,A., Moustafa,S.S.R., and Nassir, S.N.A., Correlation of densities with shared wave velocities and SPT values*." J. Geophys. Engg*.**13 (3)**: 320–341, 2016

BoaduF.K., Owusu-Nimo, F., Achampong, F., Ampadu S.I., Artificial neural network and statistical models for predicting the basic geotechnical properties of soils from electrical measurements, Near Surface Geophysics 11(6): 599-612. 2013

Hausmann, M., Engineering Principles of Ground Modification. USA: McGraw-Hill Publishing Company, 1990.

Holtz, R. D., and Kovacs, W. D., An Introduction to Geotechnical Engineering. USA: Prentice- Hall.Piñeiro G, 2010, Perelman, S., Guerschman, J.P., and Paruelo, J.M., "How to evaluate models: observed vs. predicted or predicted vs. observed? Ecological Modelling 216: 316-322, 2008.

Prasad, H.D., Puri, N., and Jain, A., "Prediction of in-place density of soil using SPT N values." Proc. National Conference on Recent Advances in Mechanical Engineering, 2017.

Proc. National Conference on Numerical Modelling in Geomechanics, Kurukshetra R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R project.org/, 2021.

Rollings, M., and Rollings, R. R., Geotechnical materials in construction. USA: McGraw- Hill, 1996.

Roorkee Prasad H.D., Puri, N., and Jain, A., "Prediction of Compression Index of Clays Using Machine Learning Techniques", 2017.

Shahin M. A., Jaksa M. B., and Maier H.R, State of the art of artificial neural networks in geotechnical engineering, Electronic Journal of Geotechnical Engineering **8**: 1-26, 2008.

Siddique, R.., Aggarwal, P., and Aggarwal, Y."Prediction of compressive strength of self- compacting concrete containing bottom ash using artificial neural networks". Advances in Engineering Software42(10):780-786. 2011

Singh, B., Sihag, P., and Singh, K.., "Modelling of impact of water quality on infiltration rate of soil by random forest regression". Modeling Earth Systems and Environment **3 (3)**: pp 999-1004, 2017.

Smith, G. N., "Probability and statistics in civil engineering: an introduction", London: Collins, 1986.

Taghi, M., Sattari, M. P., Halit, A., and Fazli, o., M5 Model Tree Application in Daily River Flow Forecasting in Sohu Stream, Turkey. Water Resources, 2013, Vol. 40, No. 3, pp. 233–242. © Pleiades Publishing, Ltd, 2010.

Biographies

Engr. Akinwamide Joshua is a Ph.D. Researcher in the Department of Civil Engineering University of Benin, Edo-State Nigeria, with a broad knowledge in geotechnical engineering/Civil Engineering Materials. He is a lecturer at the Federal Polytechnic, Ado-Ekiti, Ekiti -State Nigeria. He is a Registered Civil Engineer by Council of Regulation of Engineering in Nigeria (COREN). A corporate Member Nigerian Association of Technologist in Engineering (NATE). He holds M. ENG in Geotechnical Engineering. He has over 15years working experience as a Geotechnical engineer and Quality Control and Quality Assurance Specialist.as a researcher he has published h over 15 papers in peer-reviewed journals, conferences and workshops at both local and international level.

Engr. Prof. Jacob Odeh Ehiorobo is a professor of applied geomatic, Water resources and environmental system engineering. He is a professor in the Department of Civil Engineering, University of Benin, Edo- State Nigeria. He has held several administrative positions in the university such as Dean of Faculty ofEnvironmental Studies l, Director of External Linkage for the University and Deputy Vice Chancellor Administration. Upon graduating with Bsc, Msc and PhD as a prolific researcher he has published over 100 papers in peer-reviewed journals, conferences and workshops at both local and international level. Engr. Prof. Jacob Odeh Ehiorobo research interest includes: Deformation Surveys and Analysis, Precise engineering surveys, GNSS Positioning and Geodetic Surveys. Remote sensing and GIS Applications in Disaster Monitoring and Control, Water Resources Modelling and Environmental Hazards Analysis, Highway and Transportation including Automatic Vehicle location.

Engr. Prof. Sylvester Obinna Osujiis a Professor of Structural engineering in the Department of Civil engineering, University of Benin, Edo-State Nigeria and currently a Visiting Professor on Accumulated Leave at the Department of Civil Engineering, College of Technology, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria. Engr. Prof. Sylvester Obinna Osuji hasheld several administrative position in the University of Benin, Edo- State such as : Head of department of Civil Engineering, Sub -Dean School of Engineering and Deputy Director

procurement for the University of Benin. Upon graduating with BSC Civil engineering, M.ENG Structural engineering and PhD in structural Engineering. He has published more than 70 papers in peer-reviewed journals, conferences and workshops at both local and international level. Engr. Prof. Sylvester Obinna Osujiis also a practicing lawyer who holds LLB (Law), (University of Benin, Benin City, Nigeria) – 2006., B.L., (Nigerian Law School, Abuja) – 2007and L.L.M (Masters of Law) (University of Benin, Benin City, Nigeria) – 2009

Engr. Dr. Ebuka Nwankwo is an associate professor in the Department of Civil Engineering, faculty of Engineering University of Benin. Dr Nwankwo graduated with a first-class honour in Civil Engineering in 2005 from the Federal University of Technology Owerri. After spending some time in the industry, Dr Nwankwo proceeded for his MSc and PhD in Structural Engineering from the Imperial College London. In 2014, he was awarded a PhD from Imperial College. Dr. Nwankwo has published over 35 articles in peer review journals. Dr Nwankwo has been a visiting researcher at the University of Liverpool. He is COREN registered and member of the Nigerian Society of Engineers (NSE). He has been involved in the training and mentoring of young engineers for NSE. He has a wide range of experience in civil engineering design and project management. He has been called up to give his expert opinions on my projects within and outside Nigeria. Dr Nwankwo has also worked as professional structural engineer in Paris. He has been involved in the design of high-rise structures and geotechnical investigations for civil infrastructure.