Contents lists available at **SCCE**

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Prediction of Compression Index of Marine Clay Using Artificial Neural Network and Multilinear Regression Models

Ramachandiran Saisubramanian[1*], V. Murugaiyan[2]

1. Research Scholar, Department of Civil Engineering, Pondicherry University, India
2. Professor, Department of Civil Engineering Pondicherry University, India

Corresponding author: *saisubrmanian@pec.edu*

https://doi.org/10.22115/SCCE.2021.287537.1324

ARTICLE INFO

Article history:
Received: 23 May 2021
Revised: 23 November 2021
Accepted: 30 November 2021

Keywords:
Compression index;
Marine clay;
Artificial neural network;
Multilinear regression.

ABSTRACT

Compression Index (CI) is one of the frequently used soil parameters for the determination of possible settlement. In this study, the Compression Index of Marine clay is predicted using Artificial Neural network (ANN). Marine clay samples were collected from eight boreholes located at distance varying from 0.5 Km to 2.5 Km landward from the coastline of Pondicherry. The depth of boring was up to 12m. These samples were used for determining the Plastic Limit (PL), Liquid Limit (LL) and the Natural Moisture Content (NMC) and these were taken as input parameters for computing CI. These input parameters are taken as 'data set 1'. Similar properties of soil from over 51 boreholes were considered for analysis designated as 'Data set 2'where the depth of sampling was up to 52. These were located at a distance up to 5.0 Km from the shoreline of Puducherry distributed across the town covering a length of over 5.0 km. In Data set 2, the LL, PL, Plasticity index (PI) Specific Gravity (G), Swell Percentage, 'N'value and the ratio of PL/LL of the soil samples were taken as input parameters for prediction of CI. The input variables were reduced in successive iterations to determine their influence in the prediction of CI. Multilinear Regression Models using the same set of inputs was compared with that of ANN. Both the analysis methods indicated that the LL and PL of soil are not only easy to determine but are competent to predict CI with a high degree of accuracy.

How to cite this article: Saisubramanian R, Murugaiyan V. Prediction of compression index of marine clay using artificial neural network and multilinear regression models. J Soft Comput Civ Eng 2021;5(4):114–124. https://doi.org/10.22115/scce.2021.287537.1324

1. Prediction models

It has always been the endeavour of geotechnical engineers to simplify rigorous testing by establishing a predictive model for soils using the basic parameters. Many attempts have been made to establish a relationship among basic parameters of expansive soils, marine clay and stabilized soils to measure strength and compressibility using statistical measures such as regression analysis, correlation index and random field theory [1–4]. The relationship between the properties like Liquid limit, moisture content and friction angle (Ø) exhibit a distinct range when the prediction of N value is related to the measured value using Swedish sounding test [5]. Principal component analysis and its findings indicate the factor loadings among the variables. When these parameters are used for predicting strength indicators using Artificial Neural Networks (ANNs) yield closer and reliable predictions [6].

A probabilistic approach for the determination of Compression Index (CI) adopting the Bayesian approach [7] for marine clay is reported to give a better fit for the data from various sites in South Korea. Isotache interpretation was used [8] for determining consolidation behaviour of the long-term behaviour of clays in various parts of the world reaching depths up to 300m. Long-term consolidation properties are satisfactorily done using the Isotach model [9].

Artificial Neural Network (ANN) was adopted for assessing the CI using the data from various sites in South Korea demonstrated that ANN is a better and an accurate tool than many empirical formulae [10]. A similar comparison of prediction of CI using ANN reported for various soil data in the Middle East also correlates well with the accuracy in prediction of CI using ANN [11]. The use of ANN has also been accurate and reliable for the prediction of other parameters like free swell index [12] and maximum dry density (MDD) and Optimum Moisture Content (OMC) [6]. The least-square support vector machine depends on the regression method for the prediction of CI [13]. Statistical analysis and modelling of 130 soil sampling data in Iran have been developed for prediction of compression index and it is reported that the Root Mean Squared Error (RMSE) is 0.08 at its maximum [14].

The compression index is often related to the void ratio and many empirical relationships were proposed. However, the empirical relationship proposed is validated based on the correlation coefficient. It has been reported that these equations exhibit significance with a low value of void ratio only. A detailed study considering over 1700 data was done for proposing models relating compression index and void ratio for normally consolidated soils. Apart from the R^2 values and RMSE values, Summed Square of Residuals (SSE) value was also used to indicate the reliability of the empirical relationship [15]. However, the empirical relationship proposed is validated based on the correlation coefficient. It has been reported that these equations exhibit significance with a low value of void ratio only.

2. Basic structure and advantages of ANN

The Basic structure of ANN consists of artificial neurons similar to that of 'biological neurons' of the human brain that are grouped into layers. A more common structure would be an input layer, one or more hidden layers, and an output layer [16,17]. A human brain works by making the right connections and that this forms the basis of the working of the ANN model. ANN has multiple nodes that interact with the other. Every link is associated with a weight and the arrow describing the flow of information indicates the link. If the output is good, the weights assigned are considered appropriate. The results of ANNs when compared with multilinear regression models and improvement of MLR models have also been worked out. [18]. Hence, ANN is considered as most 'robust' system.

3. Sampling and data acquisition

Eight investigation boreholes were located along the coastline of Puducherry, India, covering a distance of about 55 Km. In addition, an active saltpan was also considered for determining the properties, located about 50 m from the borehole location BH1. The location of the investigation boreholes was fixed taking into account the coastal formations, geological information and analysis of the soil data available from the various project investigations. For identification, each borehole is assigned a symbol, BH1 to BH 8 to represent the eight boreholes. The location details are given in Table1. Each of the boreholes had marine clay layers occurring at varying depths. The samples of marine clay were tested for their properties and these are taken as 'Data set1'for analysis and for comparison. The soil properties at different locations in various boreholes were assigned a unique ID to relate the properties to the exact location of the occurrence of marine clay. The identity followed SP1 to SP28 from borehole 1 to borehole 8 in the order of Boreholes. These are presented in Table 2.

Table 1
Location details of Field Sampling sites (Dataset1).

Field Sampling site	Latitude	Longitude
1	$12^o12'49''$N	$79^o58'17''$E
2	$11^o57'22''$N	$79^o49'32''$E
3	$11^o55'54''$N	$79^o49'22''$E
4	$11^o57'22''$N	$79^o49'32''$E
5	$11^o54'2''$N	$79^o48'43''$E
6	$11^o52'48''$N	$79^o48'1''$E
7	$11^o52'42''$N	$79^o47'49''$E
8	$11^o57'22''$N	$79^o49'32''$E

Table 2
Unique ID for Field Sampling Stations.

Borehole Location	Depth in m	Unique ID	Borehole Location	Depth in m	Unique ID
BH1	7	SP1	BH4	10	SP15
BH1	8	SP2	BH5	10	SP16
BH1	9	SP3	BH5	11	SP17
BH1	10	SP4	BH5	12	SP18
BH1	11	SP5	BH6	7	SP19
Salt pan	0.1	SP6	BH6	8	SP20
BH2	2	SP7	BH6	9	SP21
BH2	3	SP8	BH6	10	SP22
BH2	4	SP9	BH6	11	SP23
BH2	5	SP10	BH6	12	SP24
BH3	7	SP11	BH8	1	SP25
BH4	3	SP12	BH8	7	SP26
BH4	5	SP13	BH8	8	SP27
BH4	9	SP14	BH8	9	SP28

The soil samples were subjected to statistical analysis and the descriptive analysis was done using the software XLstat Version 2016. The findings of the soil parameters are given in Table 3.

Table 3
Descriptive Analysis of soil properties from field investigation sites-Data set 1.

Statistic	LL (%)	PL (%)	NMC (%)	CI
Number of observations	28	28	28	28
Minimum	42.00	22.00	30.00	0.29
Maximum	75.00	43.00	55.00	0.59
1st Quartile	60.00	31.00	38.00	0.45
Median	64.00	32.00	40.00	0.49
3rd Quartile	70.50	38.00	43.00	0.55
Mean	64.11	33.71	40.61	0.49
Standard deviation (n)	7.50	4.92	5.63	0.07
Variation coefficient	0.12	0.15	0.14	0.14
Skewness (Pearson)	-0.85	0.04	0.53	- 0.83
Kurtosis (Pearson)	0.84	-0.27	0.30	0.77
Standard error of the mean	1.44	0.95	1.08	0.01

The results indicate that the variation of LL and CI follows a left tailed distribution since the skewness is <1, while PL, NMC follow a right-tailed distribution.since skewness is >1. The mean value of 0.489 is located close to the coast with the first quartile of 0.49 – which indicate that more than 75% of the soil tested indicates similar status. There are two values - 0.29 and 0.37 and can be considered outliers.

To have a better comprehension of the soil variations, a set of soil parameters already determined for meeting the project-specific requirements such as that for the construction of bridges, the road over bridges, and multistoried buildings by government agencies a were considered and another data set, assigned as ‘Data set 2’ was considered to depict the variation of soil profile more closely. The total number of clay sample properties that are relevant for the determination of the compression index yielded 200 data points from about 51 boreholes located across Puducherry representing a distance varying from 10m to 5 Km from the coastline. The descriptive analysis of the data set2 is given in Table 4.

Table 4
Soil properties from Project sites for prediction of CI. –Data set 2.

Statistic	LL	PL	PI	G	SWELL %	CI	N	PL/LL
Number of observations	200	200	200	200	200	200	200	200
Minimum	20.230	15.400	1.890	2.540	0.010	0.072	1.000	0.357
Maximum	74.000	62.000	37.000	2.730	14.483	0.448	100.000	0.943
1st Quartile	38.000	23.038	12.133	2.640	0.953	0.196	5.000	0.491
Median	49.000	26.000	20.000	2.670	3.228	0.273	9.000	0.566
3rd Quartile	54.740	30.250	25.000	2.690	5.565	0.313	16.000	0.698
Mean	46.782	27.708	19.073	2.659	3.955	0.257	14.030	0.609
Standard deviation (n)	11.018	7.537	8.924	0.040	3.470	0.077	16.831	0.146
Skewness (Pearson)	-0.377	1.563	-0.177	-0.91	0.925	-0.377	3.048	0.537
Kurtosis (Pearson)	-0.437	3.581	-0.779	0.094	0.092	-0.437	10.782	-0.787

In the ‘data Set 2’ LL has a third quartile value of 54.782%, with a mean of 46.82%. The PI has a mean of 25%, with the third quartile value of 25%. When these two properties are considered, it can be concluded that the clay is exhibiting high plasticity in 75% of the data points.

In the soil properties tabulated above, the Swell percentage is computed from the Plasticity index of soil. The Swell percentage is computed using the relationship proposed by Carter and Bentley [16].

$Swell\ (\%) = (60k(PI^{2.44})$ where k is a dimensionless constant equal to 3.6×10^{-5} (1)

Computation of CI for remoulded clays is computed using the formula

$CI = 0.009(LL - 10)$ Proposed by Terzagi and Peck (1967) (2) for normally consolidated clays and are used in data set 1. The idea behind this approach is that in ‘Data set 1’, the maximum depth of sampling is limited to 12m and the clay layers are expected to be normally consolidated.

The empirical formula

$CI = 0.007(LL - 10)$ proposed by Skempton (1994) (3)

for remoulded clays is used for computing CI for ‘Data set 2’, in which the depth of sampling varied up to 52m.

The choice of input variables is based on the potential for affecting the expansive nature of marine clay. This is identified by correlation analysis. Those parameters that have more than 0.5 .coefficient of correlation has been chosen as the input parameters. The correlation coefficient of input variables is given in Table5.

Table 5
Correlation analysis highlighting the significant input parameters.

Parameter	LL (%)	PL (%)	NMC (%)	PI	Swell %
CI (for Data Set 1)	1.00	0.92	0.66	-	-
CI (For data set 2)	1.00	0.593	-	0.733	0.643

4. Prediction of CI using ANN analysis and MLR models

The ANN analysis is done using the software SPSS version 21considering all the data in dataset 1. The default settings in the software consider 70% of the data for training. The remaining 30% of the data is used for testing. The default provision as in the software was used for the analysis and adopted throughout. Nine trials were done for the prediction of CI in Data set 1. The architecture, number of data considered for testing and training are in Table 6.

Table 6
ANN architecture-dataset1.

SL. NO	TRAINING	TESTING	ANN ARCHITECTURE
1	22(78.69%)	6(21.4%)	3-2-1
2	22(78.69%)	6(21.4%)	3-1-1
3	18(64.30%)	10(35.7%)	3-2-1
4	19(69.97%)	9(32.7%)	3-2-1
5	24(85.70%)	4(14.3%)	3-3-1
6	18(64.30%)	10(35.7%)	3-5-1
7	22(78.69%)	6(21.4%)	3-1-1
8	17(60.77%)	11(39.3%)	3-1-1
9	21(75.0%)	7(25. %)	3-2-1

The prediction gave a value with a very high R^2 value. The observed values of R^2, MAE and RMSE are tabulated in Table 7.

The performance of the MLR and ANN models were evaluated using three indices, namely; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R^2). The mean error reflects the proximity of the predicted values with that of observed values. RMSE signifies the comprehensive difference between the observed and predicted values. The measure of the total variance in respect of observed values is represented by R^2 values.

The MAE and RMSE can be determined by the mathematical equations;

$$MAE = \frac{1}{n}\sum_{i=1}^{n} mod\left(x_{oi} - x_{pi}\right) \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} mod\,(x_{oi} - x_{pi})^2} \tag{5}$$

Where x_{oi} is – observed value and x_{pi} is the predicted value.

The model that carries the maximum coefficient of determination, as well as the minimum value of MAE and RMSE, will be the best-fit model. Accordingly, ANN5 is the best prediction model using ANN

Table 7
Data set 1-ANN prediction value measure.

Output measure	ANN1	ANN2	ANN3	ANN4	ANN5	ANN6	ANN7	ANN8	ANN9
R2	0.9900	0.9880	0.9880	0.9940	0.988	0.972	0.994	0.972	0.9920
RMSE	0.0080	0.0093	0.0082	0.0046	0.007	0.012	0.005	0.014	0.0073
MAE	0.0021	0.0050	0.0025	0.0021	0.003	0.007	0.002	0.003	0.0039

The ANN4 model has the highest R^2 value while having the least RMSE and MAE and hence is taken as having the set of best-predicted values. A scatter diagram of the values obtained is given in figure.1 for dataset 1.

In MLR models, a linear relationship of one dependent variable with a set of independent variables is determined. This relies on the method of least squares with the sum of the square of error should be the minimum. The MLR model for data set 1, with CI as the dependent and PL, LL and NMC as the independent variables obtained are given below.

$$CI = 9.253E^{-02} + 9.31E^{-03} \times LL - 2.988E^{-04} \times PL - 1.3159E^{-04} \times NMC \tag{6}$$

Figure 2 shows the scatter diagram between predicted values and observed values using the MLR model with an R^2 value of 0.99 and an RMSE value of 0.002.

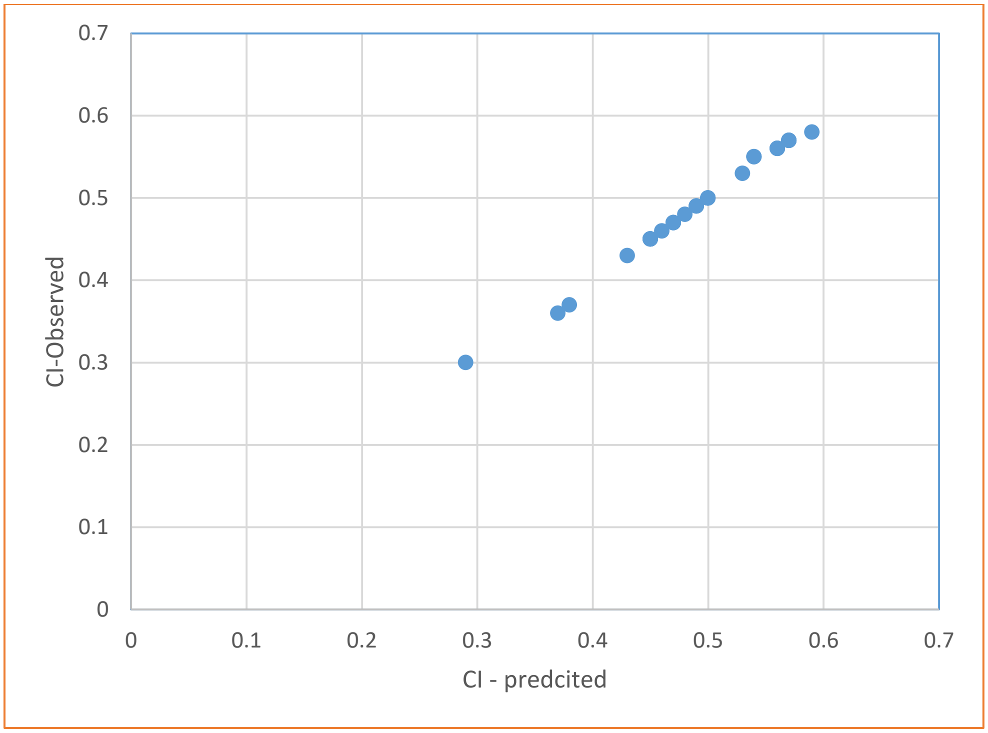

Fig. 1. Scatter diagram of best-predicted values using ANN for data set 1.

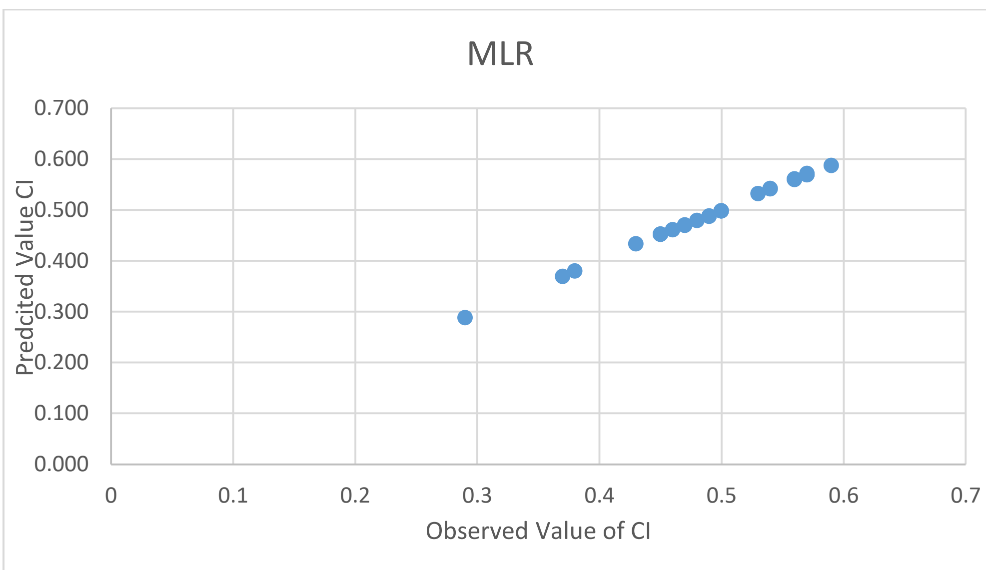

Fig. 2. Scatter diagram of CI using e MLR model for data set1.

Similarly, as was done for the field data, the ‘data set 2’ was used for predicting CI using ANN and MLR models considering the properties in all the 200 data points. In this exercise, CI is considered as a dependent variable. All other parameters LL, PL, PI, G, Swell %, N, PL/LL were considered as independent variables in the first trial for prediction of CI. Subsequently, in the second trial, the independent variables have been reduced to five independent variables LL, PL, PI, G and swell percentage. The third trial was done taking into account LL, PL, PI and swell percentage. In the fourth trial, LL, PL and swell %.In.The fifth trial was done considering LL, PI and G. Sixth trial was one with two independent variables LL and PL alone. The idea of reducing the number of independent variables is to see how the independent variables are competent enough to predict CI. It is seen that LL and PL/ LL are appreciably competent to predict the CI. The best prediction model is selected from the ANN trials by considering the least of the values of MAE and RMSE in each trial.R^2 value should be the highest. The trial in which the MAE, RMSE are least and R^2 value is highest is considered as the best trial. Table 6 shows the ANN architecture for the best-fit model for all trials done.

Table 8
The architecture of ANN Model for the Best-predicted Values of CI for 'Data Set 2'.

Sl.no	Model	ANN Architecture	No of Parameters	Combination of Input Parameters	Output Parameter	MAE	RMSE	R^2
1	ANN7	7-4-1	7	LL, PL, PI, G, SWELL%, N and PL/LL	CI	0.0018	0.0043	0.997
2	ANN5	5-3-1	5	LL, PL, PI, G and SWELL%	CI	0.0010	0.0032	0.998
3	ANN4	4-3-1	4	LL, PL, PI and SWELL%	CI	0.0014	0.0037	0.997
4	ANN2	2-2-1	2	LL and PL	CI	0.0013	0.0037	0.998
5	ANN3	3-2-1	3	LL, PL and SWELL%	CI	0.0014	0.0037	0.997
6	ANN3K	3-2-1	3	LL, PI and G	CI	0.0019	0.0044	0.996

The architecture for the best fit model based on the measure of least values of MAE, RMSE and the highest value of R^2 indicate that CI can be predicted using LL and PL the two fundamental properties as independent input variables. The scatter diagram of the best-fit ANN model is given in figure3 for data set 2.

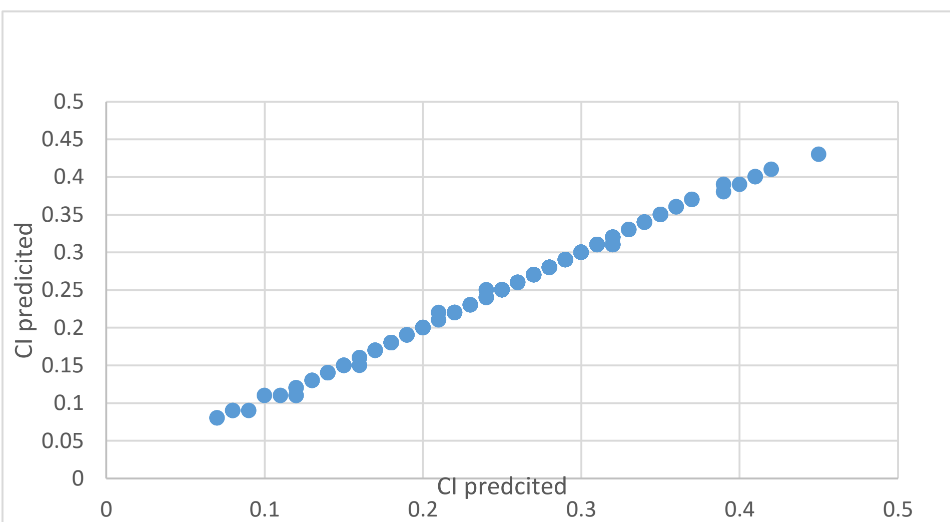

Fig. 3. Best fit model using ANN with LL and PL as independent variables for data set 2.

Table 9
MLR models for prediction of CI –'Data set 2'.

Sl. No	No. of variables	Input Variables	MLR Models	R^2
1	7	LL,PL,PP,G,SWELL%,N,PL/LL	$CI = -0.067 + 0.004(LL) + 0.003(PL) - 0.001(G) + 7.32 \times 10^{-6}\,(N) - 0.01(PL/LL)$	0.999
2	5	LL,PL,PI,G,SWELL%	$CI = -0.067 + 0.004(LL) + 0.003(PL) + 0.002(PI) - 0.001(G)$	0.999
3	4	LL,PI,PL,SWELL%	$CI = -0.070 + 0.004(LL) + 0.003(PL) + 0.002(PI)$	0.999
4	3	LL, PL, SWELL%	$CI = -0.069 + 0.007(LL)$	0.998
5	3	LL, PI, G	$CI = -0.067 + 0.007(LL) - 9.379 \times 10^{-6}(PI) - 0.001(G)$	0.998
6	2	LL, PL	$CI = -0.070 + 0.007(LL) + 2.407 \times 10^{-5}\,(PL)$	0.998

5. Conclusion

- The Compression Index (CI) is derived considering the state of clay as 'normally consolidated' for all the field investigation stations. The mean value of 0.489 is located close to the coast with the first quartile of 0.49 – which indicate that more than 75% of the soil tested indicates similar status. There are two values - 0.29 and 0.37 and can be considered outliers.
- 'Data Set 2' indicate that LL has a third quartile value of 54.782%, with a mean of 46.82%. The PI has a mean of 25%, with the third quartile value of 25%. When these two properties are considered, it can be concluded that the clay is exhibiting high plasticity in 75% of the data points.
- MLR and ANN model for the field data and the project data for prediction of CI indicate a high accuracy with R^2 values above 0.98 in all cases. The ANN is adopted taking the PL, LL and NMC into account, as these properties are determined easily. In 'Dataset 2', the soil properties such as PL, LL, PI, G, Swell%, N and the ratio of PL/LL are considered for the analysis. Five iterations are done to get the best value of prediction. The inputs were reduced to five by removing N and PL/LL as they are considered the least significant (from the MLR equation). This process of reduction of input is adopted for considering four values by removing G from the earlier set of five values. Two sets of properties: i) LL, PI, G and ii) LL, PL, Swell% - was adopted in the next trial. Further, the input was reduced to LL and PL. Both the MLR and ANN indicated that it is possible to predict CI from the two fundamental properties that are LL and PL, which have a R^2value of 0.998.
- ANN and Multilinear models for the prediction of CI from the fundamental values of LL and PL showed that these two soil parameters are very competent to predict CI as much as the combinations of additional independent variables such as PI, Swell %, G and N value.

References

[1] Wroth CP, Wood DM. The correlation of index properties with some basic engineering properties of soils. Can Geotech J 1978;15:137–45. https://doi.org/10.1139/t78-014.

[2] Fenton GA, Griffiths D V. Bearing-capacity prediction of spatially random c – ϕ soils. Can Geotech J 2003;40:54–65. https://doi.org/10.1139/t02-086.

[3] Yoon GL, Kim BT, Jeon SS. Empirical correlations of compression index for marine clay from regression analysis. Can Geotech J 2004;41:1213–21. https://doi.org/10.1139/t04-057.

[4] Soleimani S, Rajaei S, Jiao P, Sabz A, Soheilinia S. New prediction models for unconfined compressive strength of geopolymer stabilized soil using multi-gen genetic programming. Measurement 2018;113:99–107. https://doi.org/10.1016/j.measurement.2017.08.043.

[5] Makato K, Khang TT. Relationships between N value and parameters of ground strength in the South of Vietnam. Geotech Sustain Dev Hanoi 2013:1–7.

[6] Hossein Alavi A, Hossein Gandomi A, Mollahassani A, Akbar Heshmati A, Rashed A. Modeling of maximum dry density and optimum moisture content of stabilized soil using artificial neural networks. J Plant Nutr Soil Sci 2010;173:368–79. https://doi.org/10.1002/jpln.200800233.

[7] Yan WM, Yuen K-V, Yoon GL. Bayesian Probabilistic Approach for the Correlations of Compression Index for Marine Clays. J Geotech Geoenvironmental Eng 2009;135:1932–40. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000157.

[8] Watabe Y, Udaka K, Morikawa Y. Strain Rate Effect on Long-Term Consolidation of Osaka Bay Clay. Soils Found 2008;48:495–509. https://doi.org/10.3208/sandf.48.495.

[9] Watabe Y, Udaka K, Nakatani Y, Leroueil S. Long-term consolidation behavior interpreted with isotache concept for worldwide clays. Soils Found 2012;52:449–64. https://doi.org/10.1016/j.sandf.2012.05.005.

[10] Park H Il, Lee SR. Evaluation of the compression index of soils using an artificial neural network. Comput Geotech 2011;38:472–81. https://doi.org/10.1016/j.compgeo.2011.02.011.

[11] Farzin Kalantary. Prediction of compression index using artificial neural network. Sci Res Essays 2012;7. https://doi.org/10.5897/SRE12.297.

[12] Dutta RK, Singh A, Rao TG. Prediction of Free Swell Index for the Expansive Soil Using Artificial Neural Networks. J Soft Comput Civ Eng 2019;3:47–62. https://doi.org/10.22115/scce.2018.135575.1071.

[13] Samui P, Das S, Kim D, Yoon GL. Determination of compression index for marine clay: a least square support vector machine approach. Int J Adv Soft Comput Appl 2011;3.

[14] Farzi M. Suggesting a Formula to Calculate the Compression Index in Ahvaz. Indian J Sci Technol 2017;10:1–8. https://doi.org/10.17485/ijst/2017/v10i32/106232.

[15] Al-Khafaji A, Buehler A, Druszkowski E. Validation of Compression Index Approximations Using Soil Void Ratio, 2019, p. 42–52. https://doi.org/10.1007/978-3-030-01941-9_4.

[16] Pourzangbar A, Losada MA, Saber A, Ahari LR, Larroudé P, Vaezi M, et al. Prediction of non-breaking wave induced scour depth at the trunk section of breakwaters using Genetic Programming and Artificial Neural Networks. Coast Eng 2017. https://doi.org/10.1016/j.coastaleng.2016.12.008.

[17] Pourzangbar A, Saber A, Yeganeh-Bakhtiary A, Ahari LR. Predicting scour depth at seawalls using GP and ANNs. J Hydroinformatics 2017. https://doi.org/10.2166/hydro.2017.125.

[18] Noori R, Khakpour A, Omidvar B, Farokhnia A. Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. Expert Syst Appl 2010;37:5856–62. https://doi.org/10.1016/j.eswa.2010.02.020.