ORIGINAL PAPER

Artificial Neural Networks: A Solution to the Ambiguity in Prediction of Engineering Properties of Fine-Grained Soils

Viji K. Varghese · Shemy S. Babu ·
R. Bijukumar · Sobha Cyrus ·
Benny Mathews Abraham

Abstract Determination of soaked california bearing ratio (CBR) and compaction characteristics of soils in the laboratory require considerable time and effort. To make a preliminary assessment of the suitability of soils required for a project, prediction models for these engineering properties on the basis of laboratory tests—which are quick to perform, less time consuming and cheap—such as the tests for index properties of soils, are preferable. Nevertheless researchers hold divergent views regarding the most influential parameters to be taken into account for prediction of soaked CBR and compaction characteristics of fine-grained soils. This could be due to the complex behaviour of soils—which, by their very nature, exhibit extreme variability. However this disagreement is a matter of concern as it affects the dependability of prediction models. This study therefore analyses the ability of artificial neural networks and multiple regression to handle different influential parameters simultaneously so as to make accurate predictions on soaked CBR and compaction characteristics of fine-grained soils. The results of simple regression analyses included in this study indicate that optimum moisture content (OMC) and maximum dry density (MDD) of fine-grained soils bear better correlation with soaked CBR of fine-grained soils than plastic limit and liquid limit. Simple regression analyses also indicate that plastic limit has stronger correlation with compaction characteristics of fine-grained soils than liquid limit. On the basis of these correlations obtained using simple regression analyses, neural network prediction models and multiple regression prediction models—with varying number of input parameters are developed. The results reveal that neural network models have more ability to utilize relatively less influential parameters than multiple regression models. The study establishes that in the case of neural network models, the relatively less powerful parameters—liquid limit and plastic limit can also be used effectively along with MDD and OMC for better prediction of soaked CBR of fine-grained soils. Also with the inclusion of less significant parameter—liquid limit along with plastic limit the predictions on compaction characteristics of fine-grained soils using neural network analysis improves considerably. Thus in the case of neural network analysis, the use of relatively less influential input

V. K. Varghese (✉)
PWD Investigation & Planning (Roads) Sub Division,
Civil Station, Kakkanad, Kochi 682030, Kerala, India
e-mail: vijivalex@rediffmail.com; v_viji@hotmail.com

S. S. Babu
Kerala Highway Research Institute, PWD, Kariyavattom,
Thiruvananthapuram 695581, Kerala, India

R. Bijukumar
PWD Roads Section, Parassala, Thiruvananthapuram,
Kerala, India

S. Cyrus · B. M. Abraham
Department of Civil Engineering, School of Engineering,
Cochin University of Science and Technology,
Kochi 682022, Kerala, India

parameters along with stronger parameters is definitely beneficial, unlike conventional statistical methods—for which, the consequence of this approach is unpredictable—giving sometimes not so favourable results. Very weak input parameters alone need to be avoided for neural network analysis. Consequently, when there is ambiguity regarding the most influential input parameters, neural network analysis is quite useful as all such influential parameters can be taken to consideration simultaneously, which will only improve the performance of neural network models. As soils by their very nature, exhibit extreme complexity, it is necessary to include maximum number of influential parameters—as can be determined easily using simple laboratory tests—in the prediction models for soil properties, so as to improve the reliability of these models—for which, use of neural networks is more desirable.

1 Introduction

Large-scale constructions are taking place all over the world for which huge quantities of filling/stabilization materials are required. For this purpose, soils and industrial by-products such as fly ash, quarry fines etc. are collected from extensive areas. Such soils/materials may have large variations in their engineering properties. Proper estimation of various engineering properties (such as soaked CBR, MDD, OMC etc.) of soils/filling materials used for construction work is essential to ensure satisfactory performance of structures built over them. To obtain the compaction characteristics from laboratory tests, considerable time and effort is required. The soaked CBR test which is an empirical measure for the evaluation of sub grade strength of roads and pavements is not only laborious and time consuming, but also requires costly equipment. Hence while planning various construction projects, due to limited resources and time available, few laboratory tests on soaked CBR and compaction characteristics of soils are conducted.

Therefore the soil investigation data obtained are quite insufficient in many cases.

Under these circumstances, if the estimation of the soaked CBR/compaction characteristics of fine-grained soils could be developed on the basis of some laboratory tests—which are simple, speedy and cheap, it shall be useful to engineers. Thus, for a preliminary assessment of the suitability of the soils/filling material required for an earthwork project, determination of compaction characteristics/soaked CBR is important.

Even though development of prediction models for engineering properties of fine-grained soils is beneficial, it is found that researchers differ widely in their opinion regarding the input parameters to be chosen for these prediction models. This could be due to the complex behaviour of soils which, by their very nature, exhibit extreme variability. However this disagreement is a matter of concern as it hampers the use of prediction models. Hence this study focuses on use of various influential input parameters simultaneously, for development of prediction models, using two analytical tools—neural networks and multiple regressions.

2 Literature Review

A detailed literature review was conducted prior to the development of prediction models for engineering properties of fine-grained soils using regression analysis and artificial neural network analysis. The review investigated the current state of knowledge regarding soaked CBR and compaction characteristics of fine-grained soils and examined the various correlations already proposed for these engineering properties of soils. The suitability of regression and neural networks for development of prediction models in geotechnical engineering was analysed. The review also explored the steps to be taken to improve the performance of the prediction models and the various criteria to be adopted for the determination of goodness-of-fit between the predicted and observed data.

Pandian et al. (1997) proposed a method to predict the compaction characteristics of fine-grained soils in terms of the liquid limit. Studies conducted by Sridharan and Nagaraj (2005) to discover which of the index properties of fine-grained soils correlate well with the compaction characteristics proved that that the compaction characteristics do not correlate well

with either the liquid limit or the plasticity index of the soils. However, the plastic limit was found to correlate well with the compaction characteristics, namely optimum moisture content and maximum dry unit weight. Roy and Chattopadhyay (2006) derived a correlation for predicting OMC and MDD on the basis of liquid limit of a soil.

Agarwal and Ghanekar (1970) developed a CBR model considering OMC and liquid limit of cohesive soils. Roy et al. (2009) derived correlation for CBR of cohesive soils on the basis of compaction characteristics. Patel and Desai (2010) proposed a correlation between plasticity index, MDD and OMC for soaked CBR of alluvial soils. A study to check the validity of available correlations between CBR and other properties of soils was made by Roy et al. (2007) in which only partial agreement was found between the predicted and tested values. Datta and Chattopadhyay (2011) found that predicted values from correlation given by Patel and Desai (2010) agree with the tested values particularly for CI soils. But for other soils the predicted values are much lower than the tested values.

Studies on applications of artificial neural networks in geotechnical engineering by Shahin et al. (2008) establish that ANNs are well suited to modelling the complex behaviour of most geotechnical engineering materials which, by their very nature, exhibit extreme variability. ANNs learn from data examples presented to them in order to capture the subtle functional relationships among the data even if the underlying relationships are unknown or the physical meaning is difficult to explain. This is in contrast to most traditional empirical and statistical methods which need prior knowledge about the nature of the relationships among the data (Shahin et al. 2008).

Shahin et al. (2008) state that the purpose of ANNs is to non-linearly interpolate (generalize) in high-dimensional space between the data used for calibration. Unlike conventional statistical models, ANN models generally have a large number of model parameters (connection weights) and can therefore *over fit* the training data, especially if the training data are noisy. In other words, if the number of degrees of freedom of the model is large compared with the number of data points used for calibration, the model might no longer fit the general trend, as desired, but might learn the idiosyncrasies of the particular data

points used for calibration leading to 'memorization' rather than 'generalization'. As ANNs have difficulty extrapolating beyond the range of the data used for calibration, in order to develop the best ANN model, given the available data, all of the patterns that are contained in the data need to be included in the calibration set. For example, if the available data contain extreme data points that were excluded from the calibration data set, the model cannot be expected to perform well, as the validation data will test the model's extrapolation ability, and not its interpolation ability.

In order to ensure that over-fitting does not occur, Smith (1993) and Amari et al. (1997) suggested that the data be divided into three sets; training, testing and validation. The training set is used to adjust the connection weights. The testing set measures the ability of the model to generalize and the performance of the model using this set is checked at many stages of the training process. Training is stopped when the error of the testing set starts to increase. The testing set is also used to determine the optimum number of hidden layer nodes and the optimum values of the internal parameters (learning rate, momentum term and initial weights). The validation set is used to assess model performance once training has been accomplished. According to Taylor et al. (2003) this approach can also be used to ensure that the network converged to the global minimum instead of local minima.

Hecht-Nielsen (1989) provides proof that a single hidden layer of neurons, operating a sigmoidal activation function, is sufficient to model any solution surface of practical interest. To the contrary, Flood (1991) states that there are many solution surfaces that are extremely difficult to model using a sigmoidal network using one hidden layer. In addition, some researchers (Flood and Kartam 1994; Ripley 1996; Sarle 1994) state that the use of more than one hidden layer provides the flexibility needed to model complex functions in many situations. Lapedes and Farber (1988) provide more practical proof that two hidden layers are sufficient, and according to Chester (1990), the first hidden layer is useful to extract the global features of the training patterns. However, Masters (1993) states that using more than one hidden layer often slows the training process dramatically and increases the chance of getting trapped in local minima.

Table 1 Database for soaked CBR prediction of fine grained soils

Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%)	Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%)
Training data											
1	57.0	27.0	23.20	1.652	1.40	49	36.0	19.0	16.00	1.700	5.50
2	66.0	29.0	20.50	1.660	3.80	50	37.0	20.5	16.50	1.710	4.12
3	56.0	24.0	18.80	1.680	4.10	51	37.0	20.0	17.88	1.530	3.00
4	57.0	25.0	18.50	1.690	4.30	52	35.5	21.5	16.50	1.710	2.65
5	61.0	28.0	19.00	1.640	3.00	53	36.0	21.5	17.00	1.750	4.00
6	26.5	12.5	14.00	1.870	7.80	54	43.0	23.0	15.90	1.634	4.98
7	26.5	17.0	13.00	1.826	10.86	55	36.5	20.0	14.90	1.712	4.64
8	38.0	21.0	20.00	1.640	2.50	56	40.5	19.5	17.00	1.650	2.38
9	38.5	20.6	21.00	1.640	2.40	57	46.5	20.0	18.00	1.700	5.64
10	23.5	15.0	14.00	1.768	10.88	58	44.5	23.0	21.90	1.690	4.12
11	25.5	13.5	13.00	1.810	13.00	59	42.0	21.0	20.00	1.640	3.50
12	58.0	26.0	17.50	1.710	4.50	60	38.0	19.0	18.68	1.700	3.00
13	60.0	30.0	17.30	1.500	2.50	61	29.5	17.5	15.00	1.785	6.31
14	65.0	24.0	16.80	1.720	3.90	62	43.0	22.0	19.00	1.680	4.00
15	64.0	27.0	22.00	1.580	1.80	63	47.0	24.0	20.50	1.730	3.20
16	60.0	25.0	19.00	1.670	4.00	64	40.5	18.0	18.80	1.780	4.45
17	26.5	17.0	12.80	1.827	11.50	65	40.0	20.0	17.00	1.710	2.50
18	56.0	28.0	17.00	1.700	4.60	66	45.0	20.0	18.60	1.780	4.45
19	78.0	26.0	24.00	1.607	2.40	67	47.0	21.0	19.00	1.702	2.65
20	28.5	16.5	14.50	1.827	7.30	68	36.5	22.0	16.00	1.700	4.85
21	37.0	17.9	16.50	1.706	3.20	69	41.0	23.0	20.90	1.648	3.91
22	29.0	18.0	13.40	1.761	11.00	70	48.8	20.8	18.70	1.694	3.50
23	41.6	21.0	19.40	1.716	3.90	71	40.4	21.9	18.50	1.651	4.41
24	24.0	14.0	16.50	1.779	7.75	72	48.3	22.5	15.50	1.722	3.90
25	59.0	27.0	21.10	1.500	1.80	73	36.0	21.5	17.00	1.750	4.00
26	27.5	16.5	14.00	1.795	7.27	74	46.0	20.2	16.80	1.660	3.50
27	30.0	16.0	17.60	1.790	6.50	75	25.0	14.0	15.00	1.860	8.50
28	47.5	22.0	19.00	1.700	2.30	76	43.0	18.0	19.00	1.692	3.36
29	37.5	21.0	16.50	1.690	2.50	77	42.0	17.0	17.40	1.690	4.58
30	31.0	17.0	15.60	1.810	8.50	78	39.0	20.0	15.00	1.767	4.50
31	43.0	19.2	17.70	1.620	3.00	79	43.0	22.0	18.00	1.570	2.24
32	25.0	18.0	12.00	1.844	18.30	80	42.5	20.0	19.50	1.680	4.80
33	39.0	20.0	21.10	1.640	2.20	81	48.9	20.5	18.20	1.752	3.25
34	26.0	15.0	14.50	1.780	10.00	82	44.4	22.0	18.60	1.760	3.75
35	33.0	20.0	14.30	1.787	7.30	83	53.0	24.0	13.00	1.740	4.50
36	58.0	28.0	20.00	1.650	2.50	84	53.0	26.0	25.50	1.560	1.69
37	24.5	12.5	14.00	1.853	7.90	85	58.5	25.0	22.00	1.632	2.51
38	38.7	21.0	19.20	1.670	2.80	86	28.0	19.0	13.90	1.832	10.25
39	60.0	28.0	22.00	1.634	2.00	87	59.5	27.0	19.80	1.510	2.00
40	23.0	15.0	11.50	1.890	15.10	88	28.0	18.0	16.80	1.840	7.50
41	59.0	24.0	16.00	1.730	4.70	89	58.0	25.0	15.90	1.750	4.90
42	31.0	12.0	16.50	1.890	7.00	90	63.0	27.0	18.70	1.640	3.20
43	39.0	18.0	19.50	1.830	5.80	91	31.5	17.5	14.50	1.794	5.60

Table 1 continued

Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%)	Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%)
44	38.0	19.0	20.00	1.780	4.00	92	29.0	19.0	14.00	1.810	9.18
45	35.0	18.0	17.60	1.820	6.50	93	65.0	24.0	19.50	1.630	3.50
46	62.0	24.0	20.50	1.610	2.30	94	25.5	14.5	17.90	1.750	6.80
47	27.0	16.0	16.40	1.790	5.90	95	55.5	21.0	20.00	1.633	1.70
48	60.0	29.0	21.00	1.620	2.50						
Testing data											
1	37.0	20.0	21.00	1.630	3.21	6	59.0	25.0	14.90	1.780	5.10
2	42.0	21.0	19.89	1.520	3.28	7	60.0	25.0	19.00	1.670	4.00
3	42.5	17.0	19.70	1.660	3.00	8	26.5	15.0	18.00	1.710	5.00
4	52.0	26.0	14.60	1.810	5.60	9	37.0	21.5	21.40	1.550	3.14
5	56.0	24.0	14.80	1.790	5.20						
Validation data											
1	55.0	25.0	15.00	1.795	4.88	5	54.0	26.0	14.10	1.840	5.80
2	63.5	25.5	15.90	1.800	3.52	6	57.0	27.0	15.30	1.760	5.10
3	62.0	30.0	15.70	1.770	3.90	7	39.0	18.5	20.85	1.630	2.40
4	53.0	23.0	14.40	1.820	5.40	8	47.0	25.3	22.60	1.500	2.26

According to Karsoliya (2012), practically, it is very difficult to determine a good network topology just from the number of inputs and outputs. It depends critically on the number of training examples and the complexity of the classification trying to learn. There are problems with one input and one output that require millions of hidden units, and problems with a million inputs and a million outputs that require only one hidden unit, or none at all. However, this is true fact that by taking suitable number of hidden layers and the number of neurons in each hidden layer, better results in less training time can be obtained. Many researchers develop approach to estimate the number of neurons and hidden layers requirement for a neural network but the approximation also get dependable on the type of the database samples for which the network is designed. By increasing the number of hidden layers up to three layer, accuracy can be achieved up to great extent but complexity of the neural network and training time is increased many folds. If accuracy is the main criteria for designing a neural network then hidden layers can be increased. According to Sarle (1995) if early stopping method is adopted, it is essential to use lots of hidden units to avoid bad local optima. As mentioned in this paper, early stopping method is used in this case to develop the neural network models.

The method most commonly used for finding the optimum weight combination of feed-forward MLP neural networks is the back-propagation algorithm (Rumelhart et al. 1986) which is based on first-order gradient descent. The use of global optimization methods, such as simulated annealing and genetic algorithms, have also been proposed (Hassoun 1995). The advantage of these methods is that they have the ability to escape local minima in the error surface and, thus, produce optimal or near optimal solutions. However, they also have a slow convergence rate. Ultimately, the model performance criteria, which are problem specific, will dictate which training algorithm is most appropriate. Breiman (1994) recommended that if training speed is not a major concern, there is no reason why the back-propagation algorithm cannot be used successfully.

Tang et al. (1991) found that when the number of input variables increases, the forecasting ability of the neural network improves. It is likely that more accurate forecasts are produced when more information has been provided by the increased number of input variables. However, there is a trade-off between the accuracy of the model and the model complexity in terms of the number of input variables. To enhance accuracy of the forecast, the size of the training set should be relatively large when the number of input

Table 2 Database for OMC and MDD prediction of fine-grained soils

Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)	Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)
Training data									
1	47.5	22.0	19.00	16.6666300	61	39.0	20.0	21.10	16.0783960
2	28.0	18.0	16.80	18.0391760	62	85.0	38.0	28.00	13.1960494
3	73.4	51.9	44.40	11.0980148	63	46.0	20.2	16.80	16.2744740
4	61.0	28.0	19.00	16.0783960	64	46.5	20.0	18.00	16.6666300
5	48.8	20.8	18.70	16.6078066	65	43.0	23.0	15.90	16.0195726
6	42.0	21.0	20.00	16.0783960	66	59.0	24.0	16.00	16.9607470
7	65.0	24.0	16.80	16.8627080	67	42.5	20.0	19.50	16.4705520
8	45.0	26.0	18.00	15.9019258	68	39.0	24.0	19.00	16.4999637
9	23.5	15.0	14.00	17.3332952	69	43.0	22.0	19.00	16.4705520
10	37.5	21.0	16.50	16.5685910	70	35.0	18.0	17.60	17.8430980
11	40.0	20.0	17.00	16.7646690	71	27.0	16.0	16.40	17.5489810
12	23.4	15.1	12.80	19.0979972	72	44.0	24.0	20.00	16.1999644
13	26.5	12.5	14.00	18.3332930	73	36.5	20.0	14.90	16.7842768
14	48.3	22.5	15.50	16.8823158	74	62.0	24.0	20.50	15.7842790
15	58.0	26.0	17.50	16.7646690	75	48.9	20.5	18.20	17.1764328
16	31.5	17.5	14.50	17.5881966	76	57.0	27.0	23.20	16.1960428
17	36.5	22.0	16.00	16.6666300	77	61.0	42.0	29.00	13.4999703
18	58.7	45.2	39.70	12.1960516	78	59.0	27.0	21.10	14.7058500
19	64.0	27.0	22.00	15.4901620	79	59.5	27.0	19.80	14.8038890
20	29.0	19.0	14.00	17.7450590	80	41.6	21.0	19.40	16.8234924
21	38.5	20.6	21.00	16.0783960	81	43.0	18.0	19.00	16.5881988
22	38.0	19.0	18.68	16.6666300	82	42.0	24.0	19.20	16.4999637
23	44.0	24.0	20.00	16.1960428	83	57.0	25.0	18.50	16.5685910
24	41.0	23.0	20.90	16.1568272	84	60.0	29.0	21.00	15.8823180
25	63.0	27.0	18.70	16.0783960	85	66.0	29.0	20.50	16.2744740
26	27.5	16.5	14.00	17.5980005	86	43.0	19.2	17.70	15.8823180
27	25.0	14.0	15.00	18.2352540	87	29.0	15.0	14.57	16.7646690
28	28.2	17.6	13.70	18.3000000	88	47.0	24.0	20.50	16.9607470
29	56.0	28.0	17.00	16.6666300	89	36.0	21.5	17.00	17.1568250
30	44.5	23.0	21.90	16.5685910	90	39.0	20.0	15.00	17.3234913
31	29.5	17.5	15.00	17.4999615	91	71.0	29.7	22.00	14.5195759
32	38.0	19.0	20.00	17.4509420	92	44.4	22.0	18.60	17.2548640
33	39.0	29.5	24.00	15.1960450	93	25.0	18.0	12.00	18.0783916
34	30.0	16.0	17.60	17.5489810	94	56.0	24.0	18.80	16.4705520
35	47.0	21.0	19.00	16.6862378	95	69.0	31.0	24.00	15.6960439
36	36.0	19.0	16.00	16.6666300	96	53.0	26.0	25.50	15.2940840
37	60.0	25.0	19.00	16.4019247	97	28.5	16.5	14.50	17.9117253
38	25.5	14.5	17.90	17.1568250	98	37.0	20.5	16.50	16.7646690
39	58.0	25.0	15.90	17.1568250	99	48.0	21.3	21.20	15.9999648
40	42.0	17.0	17.40	16.5685910	100	23.0	15.0	11.50	18.5293710
41	33.0	20.0	14.30	17.5195693	101	25.5	13.5	13.00	17.7450590
42	31.0	12.0	16.50	18.5293710	102	36.0	24.9	21.00	16.3000000
43	59.8	27.5	20.80	16.7646690	103	55.5	21.0	20.00	16.0097687
44	40.5	18.0	18.80	17.4509420	104	60.0	25.0	19.00	16.3725130

Table 2 continued

Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)	Sl. No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)
45	60.0	30.0	17.30	14.7058500	105	60.0	28.0	22.00	16.0195726
46	24.5	12.5	14.00	18.1666267	106	26.5	17.0	13.00	17.9019214
47	37.0	18.0	16.20	17.9019214	107	65.0	24.0	19.50	15.9803570
48	39.0	18.0	19.50	17.9411370	108	68.0	49.7	29.40	12.9411480
49	45.0	20.0	18.60	17.4509420	109	26.5	17.0	12.80	17.9117253
50	28.0	19.0	13.90	17.9607448	110	36.0	21.5	17.00	17.1568250
51	43.0	22.0	18.00	15.3921230	111	37.0	20.0	17.88	14.9999670
52	55.0	30.0	21.00	15.9509453	112	58.5	32.1	28.00	14.1960472
53	29.0	18.0	13.40	17.2646679	113	24.0	14.0	16.50	17.4411381
54	40.4	21.9	18.50	16.1862389	114	58.0	28.0	20.00	16.1764350
55	38.7	21.0	19.20	16.3725130	115	78.0	26.0	24.00	15.7548673
56	53.0	24.0	13.00	17.0587860	116	38.0	21.0	20.00	16.0783960
57	28.0	21.0	14.30	17.2000000	117	26.0	15.0	14.50	17.4509420
58	35.5	21.5	16.50	16.7646690	118	112.0	43.0	25.00	14.5195759
59	37.0	17.9	16.50	16.7254534	119	58.5	25.0	22.00	15.9999648
60	40.5	19.5	17.00	16.1764350	120	31.0	17.0	15.60	17.7450590
Testing data									
1	56.0	24.0	14.80	17.5490000	9	39.0	18.5	20.85	15.9804000
2	59.0	25.0	14.90	17.4509000	10	37.0	20.0	21.00	15.9804000
3	60.0	25.0	19.00	16.3725000	11	42.0	21.0	19.89	14.9019000
4	89.0	47.0	32.00	14.3000000	12	42.5	17.0	19.70	16.2745000
5	52.0	27.0	17.00	17.6600000	13	35.8	22.5	18.60	16.6623000
6	50.0	18.0	16.04	15.9600000	14	53.0	23.0	14.40	17.8431000
7	32.5	32.5	15.50	18.0392000	15	22.3	16.2	13.80	18.0914000
8	57.0	27.0	15.30	17.2549000					
Validation data									
1	55.0	25.0	15.00	17.5980000	6	44.0	22.0	17.49	16.2600000
2	54.0	26.0	14.10	18.0392000	7	52.0	26.0	14.60	17.7451000
3	26.5	15.0	18.00	16.7647000	8	63.3	51.5	30.53	13.7255000
4	37.2	24.1	17.80	16.5838000	9	32.8	19.2	20.19	16.5686000
5	100.0	27.0	34.00	13.5980000	10	28.1	15.7	13.80	18.1385000

variables increases. Tang et al. also suggested that the amount of data can affect the performance of the forecast technique, and more training data typically means a more accurate forecast. However, even with a small amount of time series data as an input, the neural network can perform reasonably well.

The coefficient of correlation (r), the root mean square error ($RMSE$) and the mean absolute error (MAE) are the main criteria that are often used to evaluate the prediction performance of ANN models. The coefficient of correlation is a measure that is used to determine the relative correlation and the goodness-of-fit between the predicted and observed data. Smith (1986) suggested the following guidelines for values of $|r|$ between 0.0 and 1.0: $|r| \geq 0.8$—strong correlation exists between two sets of variables; $0.2 < |r| < 0.8$—correlation exists between the two sets of variables and $|r| \leq 0.2$—weak correlation exists between the two sets of variables.

Willmott and Matsuura (2005) examined the relative abilities of root mean square error ($RMSE$) and mean absolute error (MAE)—to describe average model-performance error. The findings indicated that MAE is a more natural measure of average error and
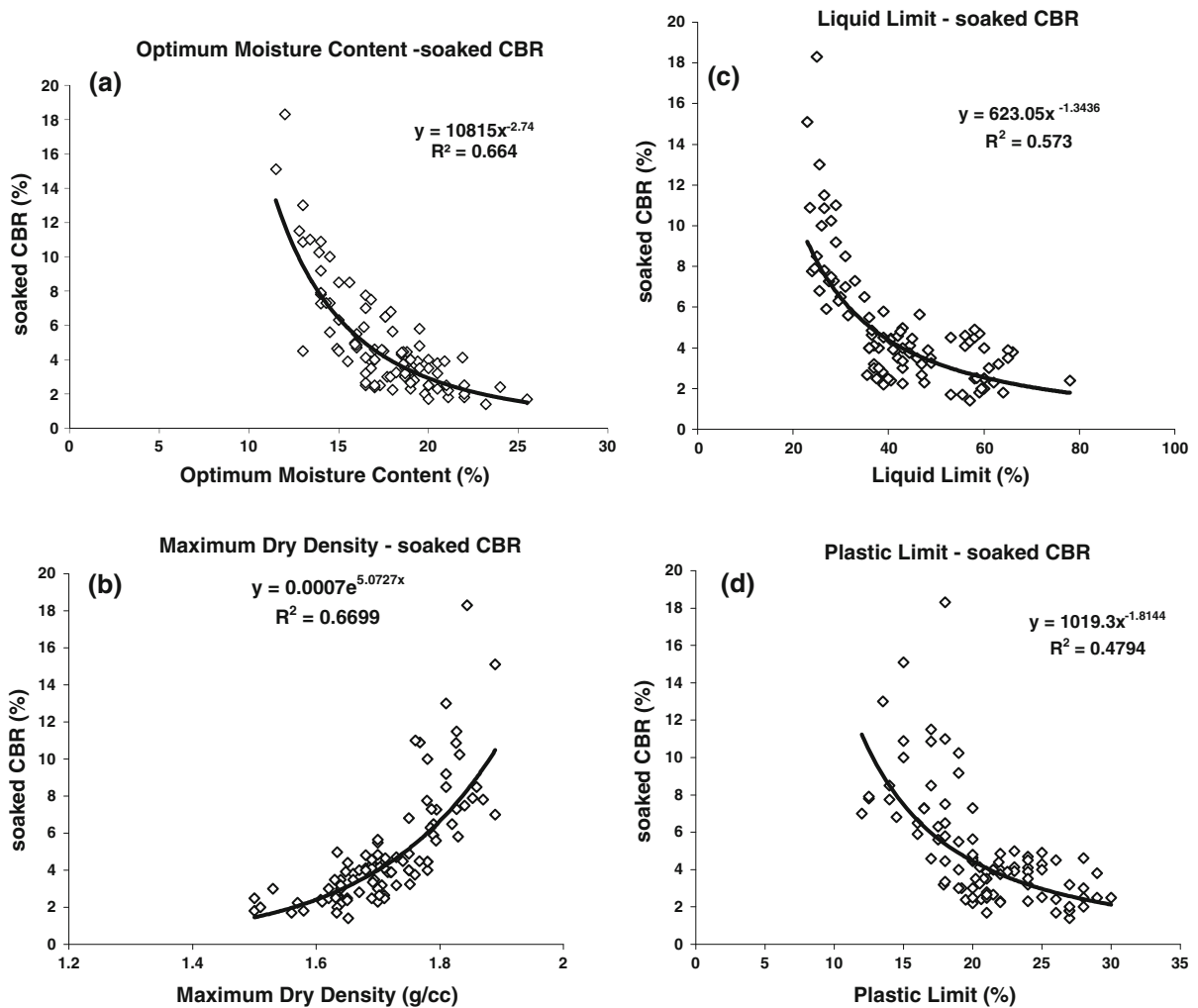
Fig. 1 a Optimum Moisture Content versus soaked CBR, **b** Maximum Dry Density versus soaked CBR, **c** Liquid Limit versus soaked CBR, **d** Plastic Limit versus soaked CBR

(unlike *RMSE*) is unambiguous. Dimensioned evaluations and inter-comparisons of average model-performance error, therefore, should be based on *MAE*.

3 Methodology

The analytical tools used for the study include statistical method such as Regression Analysis and Artificial Intelligence method such as Artificial Neural Network Analysis. EXCEL, the spreadsheet program in Microsoft's popular Office software package, has been used to perform Regression Analysis. Artificial Neural Network Analysis has been performed using

the software package MATLAB. The recommendations of researchers referred in the literature review are also taken to consideration to develop prediction models for CBR and compaction characteristics of fine-grained soils. The fine-grained soil databases used for prediction of soaked CBR, MDD and OMC, collected from various published literatures are given in Table 1 and 2.

The parameters potentially influential to CBR/compaction characteristics of fine grained soils found using simple regression, that can also be determined easily by simple laboratory tests are chosen for prediction. Initially prediction models are developed using only the most effective input parameters—
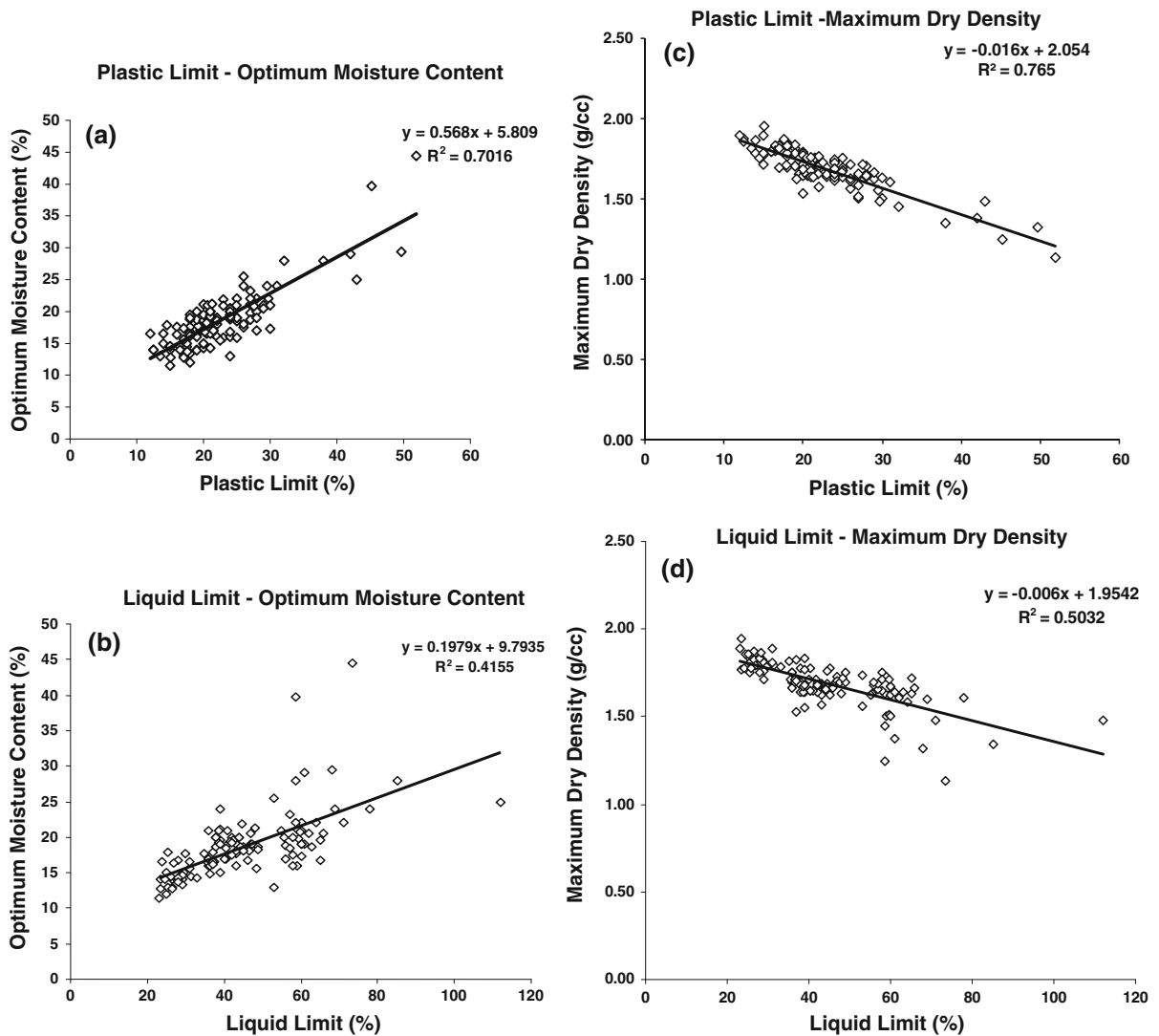
Fig. 2 a Plastic Limit versus Optimum Moisture Content, **b** Liquid Limit versus Optimum Moisture Content, **c** Plastic Limit versus Maximum Dry Density, **d** Liquid Limit versus Maximum Dry Density

determined by means of simple regression. Later other predication models are developed employing less effective input parameters also-along with more effective parameters. The performance of prediction models are closely observed to derive a conclusion regarding the strength and reliability of the two tools.

3.1 Prediction of Soaked CBR of Fine-Grained Soils Using Regression and Neural Networks

The input parameters OMC, MDD, PL and LL are chosen to develop prediction models for CBR of fine-grained soils using regression analysis and artificial neural network analysis.

3.1.1 Simple Regression Analysis

Simple regression analysis was performed prior to multiple regression analysis and neural network analysis in order to identify useful input parameters and to avoid irrelevant/weakly correlated input parameters so as to reduce the chance of neural networks being caught in local minima. To determine the relationship between soaked CBR of fine grained soils

Table 3 Equations derived using simple regression analysis

Predicted engineering properties of fine-grained soils (y)	Input Data (x)	Simple Regression Equations
Soaked CBR (%)	OMC (%)	$Y = 10815x^{-2.744}$
	MDD (g/cc)	$Y = 0.0007e^{5.0727x}$
	PL (%)	$Y = 1019.3x^{-1.8144}$
	LL (%)	$Y = 623.05x^{-1.3436}$
OMC (%)	PL (%)	$Y = 0.568x + 5.809$
	LL (%)	$Y = 0.1979x + 9.7935$
MDD(kN/m^3)	PL (%)	$Y = -0.0163x + 2.0548$
	LL (%)	$Y = -0.006x + 1.9542$

and their respective OMC, MDD, plastic limit (PL) and liquid limit (LL) simple regression analysis was used. For soaked CBR, the results of simple regression analyses shown in Fig 1a–d) indicate that in the case of fine-grained soils, MDD and OMC are stronger parameters on soaked CBR than other parameters such as LL and PL. The results of simple regression analyses for compaction characteristics, shown in Fig 2a–d) indicate that in the case of fine-grained soils, the PL bears better correlation with the compaction characteristics than LL. The equations for soaked

CBR, MDD and OMC derived using simple regression analyses are given in Table 3.

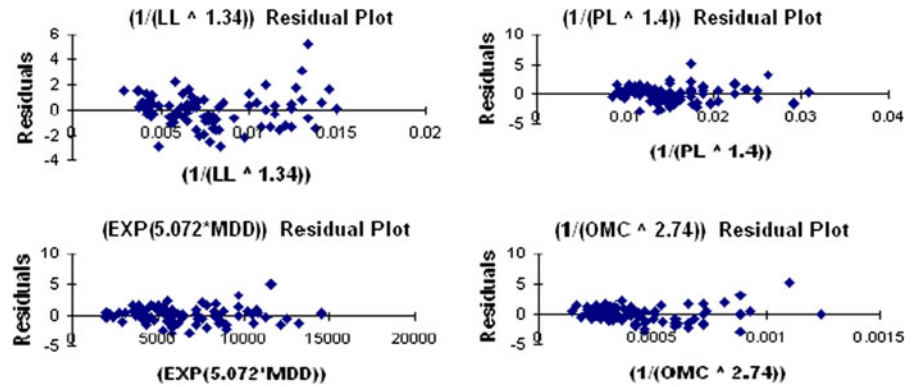### 3.1.2 Multiple Regression Analysis

The simple regression analyses of LL, PL, OMC and MDD with soaked CBR of fine grained soils indicate that, among these four parameters, MDD and OMC are the most effective parameters on soaked CBR of fine grained soils. Based on the results of simple regression analyses, four variables, namely $(1/(LL^{1.34})$, $(1/(PL^{1.4})$, $(1/(OMC^{2.74})$ and $(EXP(5.072*MDD))$ were used to predict the soaked CBR of fine grained soils (due to the combined effect of various parameters in the multiple regression model, minor modification in the parameter $(1/(PL^{1.81})$—obtained using simple regression analysis, was found necessary. When multiple regression analysis was carried out, the performance of the multiple regression model with parameter $(1/(PL^{1.4})$ was found to be much better than that with parameter $(1/(PL^{1.81})$ and hence the same was adopted). The summary output is shown in Table 4.

In this case the correlation coefficient (Multiple R) is 0.914, coefficient of determination R^2 is 0.835 and adjusted R^2 is 0.827. Hence the four independent variables together explain 83.5 % of total variation in dependent variable CBR (df-degree of freedom,

Table 4 Summary output of multiple regression analysis using 4 variables

Multiple R					0.91370265
R Square					0.83485253
Adjusted R Square					0.82751264
Standard Error					1.27651724
Observations					95

	df	SS	MS	F	Significance F
ANOVA					
Regression	4	741.3678158	185.342	113.7419	$2.45735E \times 10^{-34}$
Residual	90	146.6546642	1.6295		
Total	94	888.02248			

	Coefficients	Standard error	t stat	P value	Lower 95 %	Upper 95 %
Intercept	-1.1885529	0.479827848	-2.477	0.015116	-2.141814608	-0.23529
1/LL$^{1.3}$	423.847882	103.0954142	4.11122	$8.68E \times 10^{-05}$	219.0308522	628.6649
1/PL$^{1.4}$	-130.72343	62.48999634	-2.0919	0.039263	-254.8707156	-6.57615
1/OMC$^{2.74}$	6708.64064	1024.497395	6.54823	$3.51E \times 10^{-09}$	4673.297789	8743.983
EXP(5.072*MDD)	0.00030813	$8.61066E \times 10^{-05}$	3.57842	0.000559	0.00013706	0.000479

**Fig. 3** Residual Plots of four input variables for prediction of soaked CBR

Table 5 Equations for prediction of engineering properties of fine-grained soils using multiple regression analysis

Predicted engineering properties of fine-grained soils	Input data	Equations derived using multiple regression analysis
Soaked CBR (%)	OMC and MDD (g/cc)	$CBR = -1.4501 + 8907.3739 \times (1/OMC^{2.74}) + 0.000384 \times (EXP(5.072*MDD))$
	LL, PL, OMC and MDD(g/cc)	$CBR = -1.1886 + 423.85 \times (1/LL^{1.34}) - 130.72 \times (1/PL^{1.4}) + 6708.641 \times (1/OMC^{2.74}) + 0.00031 \times (EXP(5.072 \times MDD))$
OMC (%)	PL	$OMC = 5.809 + 0.5680 \times PL$
	PL and LL	$OMC = 5.9257 + 0.6134 \times PL - 0.0255 \times LL$
MDD(kN/m^3)	PL	$MDD = 20.145 - 0.1597 \times PL$
	PL and LL	$MDD = 20.159 - 0.157 \times PL - 0.0013 \times LL$

SS-sum of squares, ms-mean square error). However when degrees of freedom lost are also taken into consideration, the four independent variables together explain only 82.75 % of total variation in dependent variable. Thus, truly speaking, adjusted R^2 indicate the adequacy of the model as it takes into account the deviations as well as degrees of freedom. If adjusted R^2 is fairly close to 1, the overall model can be considered adequate to fit the data. But it does not mean that there are no insignificant parameters in it. The residual plots of the four variables shown in Fig. 3 indicate that there is no obvious correlation between the residuals and each independent variable. However, the fact that the residuals look random and that there is no obvious correlation with the variable does not necessarily mean by itself that the model is adequate. More tests are needed.

The overall significance test of the regression model is tested by F-statistic. The column F gives the overall F-test of H0: coefficients = 0 versus Ha: at least one of the coefficients is not equal to zero. Excel computes F as F = [Regression SS/(k−1)]/[Residual SS/(n−k)], where k is the number of regressors including the intercept. The decision rule is when the computed F is greater than critical F, reject the null hypothesis. The computed F (113.7419) in this case is greater than the critical F (3.01—at 0.01 significance level). The closer the significance level is to 0 %, the stronger is the evidence against the null hypothesis and hence the null hypothesis may be rejected. The value of Significance F in ANOVA output is $2.457E \times 10^{-34}$. Thus the significance F value is much less than 0.01 (hence we reject H0 at significance level 0.01). The corresponding level of confidence is 0.9999. Therefore at least one of coefficients −1.1886, 423.85, −130.72, 6708.641 and 0.00031 is significant for the model. The statistical significance of independent variables and intercept in the multiple regression model is then ascertained by computed t values. This t value of each coefficient is obtained by dividing estimated coefficient by its standard error under the null hypothesis that the coefficient is zero. The computed t static value of all the variables and intercept are greater than critical t value (2), at 0.05
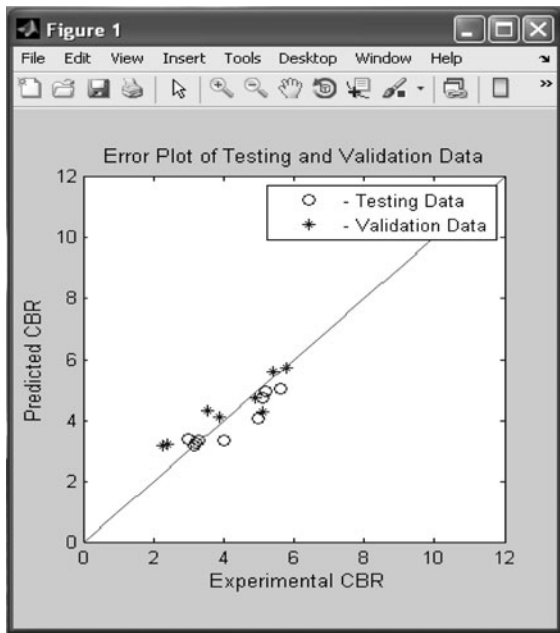
Fig. 4 *Error plot* of Testing and Validation data of soaked CBR prediction in ANN Analysis

level of significance. This implies that all the variables and intercept in the multiple regression model are statistically significant at 0.05 level of confidence.

The P value approach is also used for evaluating the contribution of individual independent variables. The P value in the regression output indicates the significance of individual coefficient in the regression model. The regression output indicate that $P_{intercept} = 0.0151$, which corresponds to fairly high confidence level, $1 - 0.0151 = 0.9849$. This suggests that the parameter -1.1886 is significant. The regression output also indicates that confidence levels for other parameters are high, which means that they are also significant. For the regression model, the lower and upper 95 % limits for intercept and variables do not include zero. This agrees with the previous conclusions made about their significance. Hence the multiple regression model obtained, soaked $CBR = -1.1886 + 423.85 \times (1/LL^{1.34}) - 130.72 \times (1/PL^{1.4}) + 6708.641 \times (1/OMC^{2.74}) + 0.00031 \times (EXP (5.072 \times MDD))$ is adequate to fit the experimental data. Various prediction models for engineering properties of fine-grained soils thus derived using Multiple Regression Analysis are given in Table 5.

3.1.3 Artificial Neural Network Analysis

The chosen data is divided into three groups, for training, testing and validation, so that the generalization capacity of network could be checked after training phase. The training set is the largest set and is used by neural network to learn patterns present in the data. The testing set is used to evaluate the generalization ability of a supposedly trained network. A final check on the performance of the trained network is made using validation set. Mean Absolute Error (MAE) is used as a measure of error made by the network.

For prediction of soaked CBR using Artificial Neural Network Analysis the four input variables used are the OMC expressed in percentage, MDD expressed in g/cc, PL expressed in percentage and LL expressed in percentage. Hence the input layer has four neurons. The only output is the soaked CBR of

Table 6 Artificial neural network analysis

Predicted engineering properties of fine-grained soils	Input data	Artificial neural network analysis using feed forward back propagation neural network					
		Network structure	Activation function	Learning rate	Epochs	MAE-Testing Data	MAE-Validation Data
Soaked CBR (%)	OMC and MDD (g/cc)	2-81-81-1	logsig-logsig-logsig-logsig	0.005	10000	1.0446	1.0697
	LL, PL, OMC and MDD (g/cc) ((G/CC(g/cc)	4-100-100-1	logsig-logsig-logsig-logsig	0.005	3500	0.3634	0.5006
OMC (%)	PL	1-120-1	logsig-logsig-logsig	0.005	10000	3.5135	4.1038
	PL and LL	2-100-50-1	logsig-logsig-logsig-logsig	0.005	35000	1.9455	1.7297
MDD (kN/m³)	PL	1-50-1	logsig-logsig -logsig	0.005	2000	1.1691	1.2661
	PL and LL	2-100-50-1	logsig-logsig –logsig-logsig	0.005	8500	0.8577	0.8471

Table 7 Prediction of properties of fine-grained soils using multiple regression analysis and artificial neural network analysis—Comparison between the experimental and predicted values

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%) (experimental)	Multiple regression analysis		Artificial neural network analysis	
						Soaked CBR (%) (predicted using 2 variables)	Absolute error	Soaked CBR (%) (predicted using 2 variables)	Absolute error
(i) Prediction of soaked CBR of fine-grained soils using 2 variables									
1	55.0	25.0	15.00	1.795	4.88	7.3399	2.4599	4.9742	0.0942
2	63.5	25.5	15.90	1.800	3.52	6.6411	3.1211	4.3866	0.8666
3	62.0	30.0	15.70	1.770	3.90	6.3017	2.4017	4.3473	0.4473
4	53.0	23.0	14.40	1.820	5.40	8.4383	3.0383	6.5519	1.1519
5	54.0	26.0	14.10	1.840	5.80	9.2113	3.4113	8.0835	2.2835
6	57.0	27.0	15.30	1.760	5.10	6.4963	1.3963	4.4954	0.6046
7	39.0	18.5	20.85	1.630	2.40	2.2102	0.1898	3.8910	1.4910
8	47.0	25.3	22.60	1.500	2.26	1.0591	1.2009	3.8781	1.6181
Mean absolute error (MAE)							**2.1524**		**1.0697**
Correlation coefficient (r)							**0.93**		**0.79**

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (g/cc)	Soaked CBR (%) (experimental)	Multiple regression analysis		Artificial neural network analysis	
						Soaked CBR (%) (predicted using 4 variables)	Absolute error	Soaked CBR (%) (predicted using 4 variables)	Absolute error
(ii) Prediction of soaked CBR of fine-grained soils using 4 variables—liquid limit, plastic limit, maximum dry density and optimum moisture content									
1	55.0	25.0	15.00	1.795	4.88	6.1319	1.2519	4.7505	0.1295
2	63.5	25.5	15.90	1.800	3.52	5.3038	1.7838	4.3337	0.8137
3	62.0	30.0	15.70	1.770	3.90	5.3621	1.4621	4.0992	0.1992
4	53.0	23.0	14.40	1.820	5.40	6.9039	1.5039	5.6181	0.2181
5	54.0	26.0	14.10	1.840	5.80	7.7111	1.9111	5.7186	0.0814
6	57.0	27.0	15.30	1.760	5.10	5.5240	0.4240	4.2969	0.8031
7	39.0	18.5	20.85	1.630	2.40	2.5698	0.1698	3.2343	0.8343
8	47.0	25.3	22.60	1.500	2.26	1.7559	0.5041	3.1855	0.9255
Mean absolute error (MAE)							**1.1263**		**0.5006**
Correlation coefficient (r)							**0.95**		**0.93**

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m^3)	Multiple regression analysis		Artificial neural network analysis	
					Predicted OMC (%)	Absolute error	Predicted OMC (%)	Absolute error
(iii) Prediction of OMC using 1 variable—plastic limit								
1	55.00	25.00	15.00	17.5980	20.0096	5.0096	20.0207	5.0207
2	54.00	26.00	14.10	18.0392	20.5776	6.4776	20.5848	6.4848
3	26.50	15.00	18.00	16.7647	14.3294	3.6706	15.6398	2.3602
4	37.20	24.10	17.80	16.5838	19.4984	1.6984	19.5093	1.7093
5	100.00	27.00	34.00	13.5980	21.1457	12.8543	21.1325	12.8675
6	44.00	22.00	17.49	16.2600	18.3055	0.8155	18.3549	0.8649
7	52.00	26.00	14.60	17.7451	20.5776	5.9776	20.5848	5.9848
8	63.30	51.50	30.53	13.7255	35.0622	4.5322	29.9707	0.5593
9	32.80	19.20	20.19	16.5686	16.7151	3.4749	17.0273	3.1627

Table 7 continued

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)	Multiple regression analysis		Artificial neural network analysis	
					Predicted OMC (%)	Absolute error	Predicted OMC (%)	Absolute error
10	28.10	15.70	13.80	18.1385	14.7270	0.9270	15.8235	2.0235
Mean absolute error (MAE)						**4.5438**		**4.1038**
Correlation coefficient (r)						**0.59**		**0.59**
(iv) Prediction of OMC using 2 variables—plastic limit and liquid limit								
1	55.00	25.00	15.00	17.5980	19.8563	4.8563	13.7481	1.2519
2	54.00	26.00	14.10	18.0392	20.4953	6.3953	13.7647	0.3353
3	26.50	15.00	18.00	16.7647	14.4502	3.5498	17.7837	0.2163
4	37.20	24.10	17.80	16.5838	19.7591	1.9591	17.1725	0.6275
5	100.00	27.00	34.00	13.5980	19.9332	14.0668	30.5504	3.4496
6	44.00	22.00	17.49	16.2600	18.2971	0.8071	16.7561	0.7339
7	52.00	26.00	14.60	17.7451	20.5464	5.9464	13.8311	0.7689
8	63.30	51.50	30.53	13.7255	35.9006	5.3706	27.1030	3.4270
9	32.80	19.20	20.19	16.5686	16.8657	3.3243	17.6886	2.5014
10	28.10	15.70	13.80	18.1385	14.8387	1.0387	17.7849	3.9849
Mean absolute error (MAE)						**4.7315**		**1.7297**
Correlation coefficient (r)ß						**0.55**		**0.97**

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m³)	Multiple regression analysis		Artificial neural network analysis	
					Predicted MMD (kN/m³)	Absolute error	Predicted MMD (kN/m³)	Absolute error
(v) Prediction of maximum dry density using 1 variables—plastic limit								
1	55.00	25.00	15.00	17.5980	16.1515	1.4465	16.3144	1.2836
2	54.00	26.00	14.10	18.0392	15.9918	2.0474	16.1898	1.8494
3	26.50	15.00	18.00	16.7647	17.7489	0.9842	17.3753	0.6106
4	37.20	24.10	17.80	16.5838	16.2953	0.2885	16.4268	0.1570
5	100.00	27.00	34.00	13.5980	15.8320	2.2340	16.0671	2.4691
6	44.00	22.00	17.49	16.2600	16.6307	0.3707	16.6840	0.4240
7	52.00	26.00	14.60	17.7451	15.9918	1.7533	16.1898	1.5553
8	63.30	51.50	30.53	13.7255	11.9185	1.8070	10.6612	3.0643
9	32.80	19.20	20.19	16.5686	17.0780	0.5094	16.9987	0.4301
10	28.10	15.70	13.80	18.1385	17.6371	0.5014	17.3213	0.8172
Mean absolute error (MAE)						**1.1942**		**1.2661**
Correlation coefficient (r)						**0.63**		**0.63**
(vi) Prediction of maximum dry density using 2 variables—plastic limit and liquid limit								
1	55.00	25.00	15.00	17.5980	16.1437	1.4543	16.2394	1.3586
2	54.00	26.00	14.10	18.0392	15.9876	2.0516	16.2258	1.8134
3	26.50	15.00	18.00	16.7647	17.7551	0.9904	18.9431	2.1784
4	37.20	24.10	17.80	16.5838	16.3086	0.2752	16.3955	0.1883
5	100.00	27.00	34.00	13.5980	15.7700	2.1720	13.0438	0.5542
6	44.00	22.00	17.49	16.2600	16.6303	0.3703	16.3777	0.1177
7	52.00	26.00	14.60	17.7451	15.9902	1.7549	16.2473	1.4978
8	63.30	51.50	30.53	13.7255	11.9614	1.7641	13.9807	0.2552

Table 7 continued

Sl No.	LL (%)	PL (%)	OMC (%)	MDD (kN/m^3)	Multiple regression analysis		Artificial neural network analysis	
					Predicted MMD (kN/m^3)	Absolute error	Predicted MMD (kN/m^3)	Absolute error
9	32.80	19.20	20.19	16.5686	17.0857	0.5171	16.7045	0.1359
10	28.10	15.70	13.80	18.1385	17.6428	0.4957	18.5102	0.3717
Mean absolute error (MAE)						**1.1846**		**0.8471**
Correlation coefficient (r)						**0.63**		**0.77**

soil and therefore the output layer has only one neuron. In this study, a feed forward back propagation neural network with two hidden layers (with hundred neurons each) gave good results. Logistic sigmoid transfer function is used for all the layers and only the output is normalized to get desired results. Since Logistic sigmoid transfer function (often called squashing functions) compress the infinite input range to finite output range normalization of inputs was not found necessary. Learning rate of 0.005 is found to be suitable for good performance. Training goal for the networks is set to 10^{-4}. The networks are trained for fixed number of epochs.

During the training phase, the system adjusts its connection/weight strengths in favour of the inputs that are most effective in determining a specific output. It is possible that repeated training iterations successively improve performance of the network on training data by memorizing training samples. Nevertheless, the resulting network may perform poorly on test data. The use of validation set in the study is an important guard against this overtraining or over fitting network. To avoid overtraining, the error (maximum absolute error and/or mean absolute error) on the test set is monitored. Training is continued as long as the error on the test set decreased. It is terminated when the error on the test set started increasing again. Thus the training is halted when the error on testing dataset is lowest (early stopping method) even though the error on training dataset is found to decrease further as training is continued. As early stopping method is adopted, in order to escape local minima—large numbers of hidden nodes are used. The Mean Absolute Errors of testing and validation data of the model are 0.3634 and 0.5006, respectively. The error plot of testing and validation data is given in Fig. 4. Table 6 indicates different

prediction models thus derived for soaked CBR, MDD and OMC using Artificial Neural Network Analysis.

3.2 Comparison of Results of Multiple Regression Analysis and Neural Network Analysis

To study how well the analytical values "fit" the experimental values, the mean absolute errors (MAE) and correlation coefficients (r) of the predicted values are determined. The results of prediction of CBR and compaction characteristics using multiple regression analysis and neural network analysis are shown in Table 7. To evaluate the performances of multiple regression models and ANN models, both MAE and correlation coefficient are taken to consideration. Figure 5a, b illustrate the neural network-multiple regression error plots for prediction of soaked CBR of fine-grained soils using 2 variables (MDD, OMC) and 4 variables (MDD, OMC, PL and LL). Similar error plots for prediction of OMC and MDD of fine-grained soils using 1 variable (PL) and 2 variables (PL and LL) are shown in Fig. 6a–d). The performances of both neural network models and multiple regression models used for prediction of soaked California Bearing Ratio (CBR)/Compaction characteristics of fine-grained soils are summarized in Table 8.

The study show that for prediction of soaked CBR of fine-grained soils, using 4 variables—MDD, OMC, PL and LL, though the predicted and the experimental values strongly agree for neural network models as well as multiple regression models the predictions using neural network models are more accurate. Also, fairly good predictions on soaked CBR could be made using the two strong parameters—MDD and OMC alone, in which case also neural network models are more accurate (though the mean absolute errors of ANN models for the prediction of soaked CBR of fine-

(a) Soaked CBR prediction (2 variables)
Multiple regression vs Neural networks - Error plot

(b) Soaked CBR prediction (4 variables)
Multiple regression vs Neural networks - Error plot
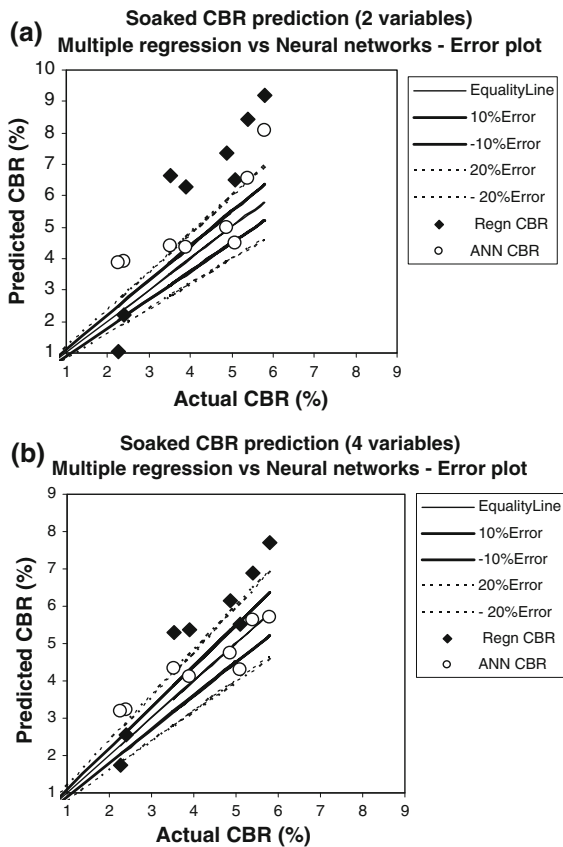
Fig. 5 a Soaked CBR prediction (2 variables) Multiple regression versus Neural networks—*Error plot*, **b** Soaked CBR prediction (4 variables) Multiple regression versus Neural networks—*Error plot*

grained soils using 4 input variables and 2 input variables are lesser than the corresponding multiple regression models, the correlation coefficients of multiple regression models are higher than ANN models -this could at times happen due to smaller sample size and presence of outliers). In the case of prediction of compaction characteristics using two variables PL and LL, neural network models are far more accurate than multiple regression models. The results thus prove that neural network analysis has more potential as a forecasting tool than multiple regression analysis.

The results also indicate that, when the input parameters PL and LL are used along with the two strong parameters OMC and MDD—the accuracy of prediction of soaked CBR increases significantly in the case of neural network analysis while moderate improvement in accuracy is observed for prediction of

soaked CBR using multiple regression model. In the case of prediction of OMC, with the use of additional input parameter LL along with PL the accuracy of prediction of neural network model increase considerably while the use of additional input parameter affects the multiple regression model adversely. The additional input parameter—LL also improves the accuracy of prediction of MDD using neural network model notably, while the increase in accuracy of prediction of multiple regression model with the use of additional parameter is negligible. The superior performance in the case of neural network models—with more number of influential parameters—is obvious in individual results of validation data depicted in the error plots.

It can be inferred from the results that in the case of neural network analysis, the use of relatively less influential input parameters along with stronger parameters is definitely beneficial, unlike conventional statistical methods—for which, this approach is risky—with either damaging effect or less benefit. The various input parameters used for prediction of dependent variable need not have strong and equal influence on the dependent variable, for ANN to develop a good model. But in the case of multiple regression models, this approach need not always produce positive results. With the addition of more input parameters, the information available to neural networks increase, which in turn improves the performance of neural networks. ANN models are capable of extracting even the subtle relationship of the dependent variable to the weak independent variables in a better manner than multiple regression models. Only very weak parameters need to be avoided in the case of neural network analysis. However use of less influential parameters alone in ANN models is not likely to give favorable results. The behaviors of soils as such are very complicated and hence to safeguard the dependability of prediction models for soil properties it is always preferable to use more number of influential parameters—which can be done safely only using neural networks. Those parameters which can be easily determined using simple laboratory tests—having moderate correlation with predicted variable can also be included in neural network analysis to improve the prediction results.

On comparing the predictions of engineering properties of fine-grained soils made using regression analysis and neural network analysis it is understood
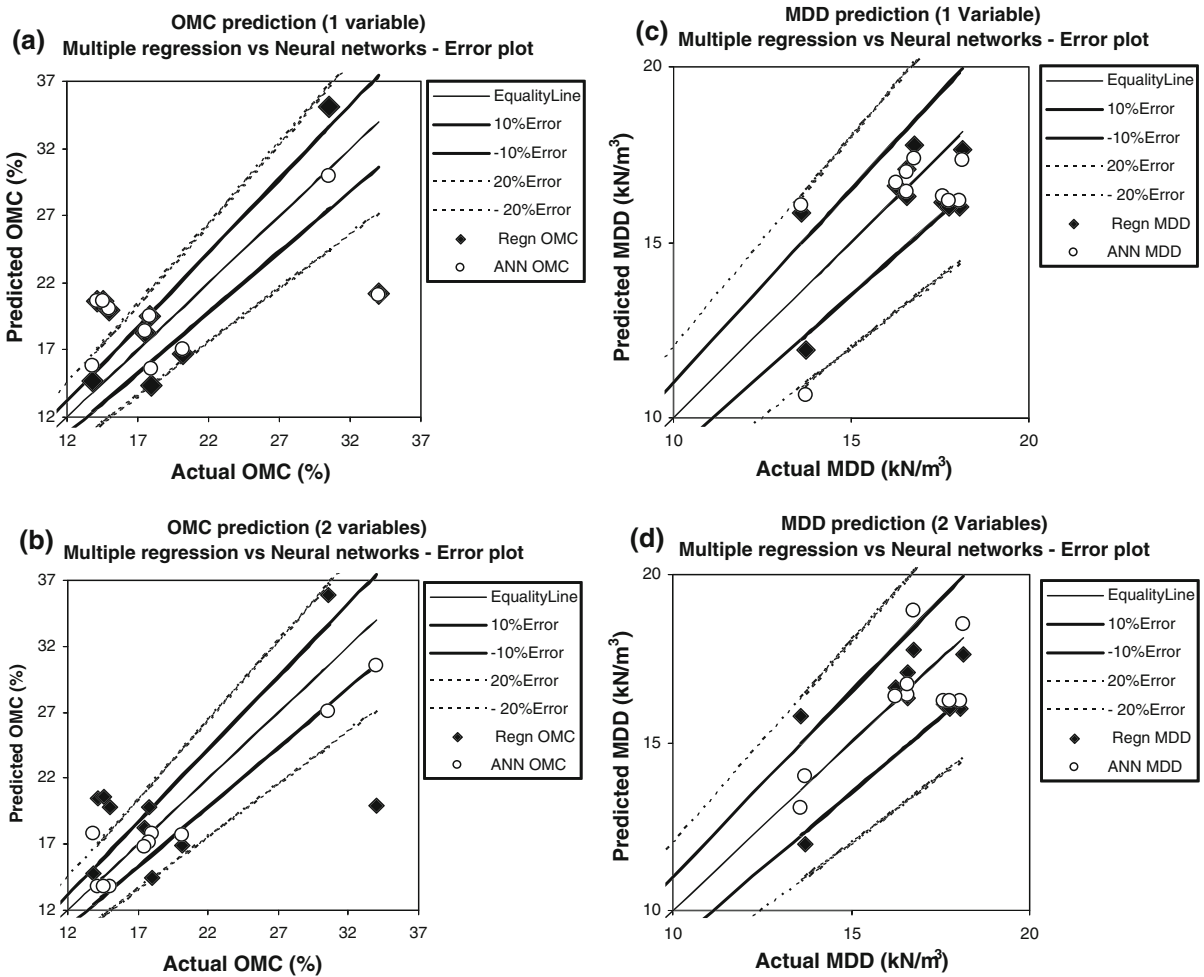
Fig. 6 **a** OMC prediction (1 variable) Multiple regression versus Neural networks—*Error plot*, **b** OMC prediction (2 variables) Multiple regression versus Neural networks—*Error plot*, **c** MDD prediction (1 variable) Multiple regression

versus Neural networks—*Error plot,* **d** MDD prediction (2 variables) Multiple regression versus Neural networks—*Error plot*

that with a uniformly representative database and sufficient number of input variables, ANN is capable of making more accurate predictions. To get most accurate predictions, the database used for fine-grained soils should be well representative which includes soils belonging to different groups such as CH, CI and CL as well as MH, MI and ML. In this study, the database (inclusive of testing and validation data) used for the predictions of CBR values of fine-grained soils included 112 soil samples while the database used for the predictions of compaction characteristics of fine-grained soils consisted of 145 soil samples. Though these databases cannot be claimed as perfect uniformly representative large

databases, representing all possible cases, the results indicate that reasonably accurate predictions can be made even with lesser data. The performances of artificial neural networks will definitely improve with a larger and more representative database and sufficient number of input variables. However, neural network analysis lacks transparency—the output is obtained as numerical values. No information can be gathered about the effect of each input variable on the predicted variable. Neural networks may even be termed as "black boxes" for this inability. For regression analysis, the output obtained in the form of equations/trend lines gives an overall idea about the influence of each input variable on the predicted

Table 8 Performance evaluation of multiple regression analysis and artificial neural network analysis

Predicted engineering properties of fine-grained soils	Input data	Multiple regression analysis		Artificial neural network analysis		
		Correlation coefficient (r) of validation data	Mean absolute error (mae) of validation data	Correlation coefficient (r) of validation data	Mean absolute error (MAE)	
					Testing data	Validation data
Soaked CBR (%)	OMC and MDD (g/cc)	0.93	2.1524	0.79	1.0446	1.0697
	LL, PL, OMC and MDD (g/cc)	0.95	1.1263	0.93	0.3634	0.5006
OMC (%)	PL	0.59	4.5438	0.59	3.5135	4.1038
	PL and LL	0.55	4.7315	0.97	1.9455	1.7297
MDD (kN/m^3)	PL	0.63	1.1942	0.63	1.1691	1.2661
	PL and LL	0.63	1.1846	0.77	0.8577	0.8471

variable. While the training time/processing time required for obtaining the best neural network model can vary, the regression analysis take definite time.

4 Summary

Predictions of engineering properties (namely soaked CBR, MDD and OMC) of fine-grained soils using neural network models are better than those using regression models. Accurate predictions on CBR of fine-grained soils can be obtained using neural networks with four numbers of influential input parameters namely OMC, MDD, PL and LL. Also precise predictions on compaction characteristics of fine-grained soils can be achieved with two numbers of influential input parameters namely PL and LL. In the case of neural network analysis, the use of relatively less influential input parameters along with stronger parameters is definitely beneficial, unlike conventional statistical methods—for which, this approach need not always return positive/favourable results. The various input parameters used for prediction of dependent variable should not necessarily have strong and equal influence on the dependent variable, for ANN to develop a good model. The behaviours of soils as such are very complicated and hence to safeguard the dependability of prediction models for soil properties it is always preferable to use more number of influential parameters—which can be done securely only using neural networks. However, ANN models give less insight into the underlying physical relationship between each input variable and predicted variable.

5 Conclusions

- Predictions on soaked CBR and compaction characteristics of fine-grained soils using neural network models are better than those using regression models.
- Accurate predictions on soaked CBR of fine-grained soils can be obtained using neural networks with four numbers of influential input parameters namely OMC, MDD, PL and LL. Also precise predictions on compaction characteristics of fine-grained soils can be achieved with two numbers of influential input parameters namely PL and LL.
- The performance of neural networks improves as the number of input parameters increase. Thus for neural network analysis, the use of relatively less influential input parameters along with stronger parameters is definitely beneficial, unlike multiple regression—for which, this approach is sometimes not so advantageous. In other words, in the case of neural network analysis, it is more effective to use all influential parameters—which can be determined easily by simpler laboratory tests—simultaneously, while this approach need not always give better results for conventional statistical methods. With the addition of more input parameters, the information available to neural networks increase, which in turn improves the performance of neural networks. The various input parameters used for prediction of dependent variable need not have strong and equal influence on the dependent variable, for ANN to develop a good model. Hence

very weak input parameters alone need to be avoided in neural network analysis.

- The behaviours of soils as such are very complicated and hence to safeguard the dependability of prediction models for soil properties it is always preferable to use more number of influential parameters in the prediction models—which can be done safely only using neural networks.

References

Agarwal KB, Ghanekar KD (1970) Prediction of CBR from plasticity characteristics of soil. In: Proceedings of the 2nd south-east Asian conference on soil engineering, Singapore, pp 571–576

Amari SI, Murata N, Muller KR, Finke M, Yang HH (1997) Asympotic statistical theory of overtraining and cross-validation. IEEE Trans Neural Netw 8(5):985–996

Breiman L (1994) Comment on neural networks: a review from a statistical by B. Cheng and D. M. Titterington. Stat Sci 9(1):38–42

Chester DL (1990) Why two hidden layers are better than one. Internat Joint Con Neural Netw 1:265–268

Datta T, Chattopadhyay BC (2011) Correlation between CBR and index properties of soil. In: Proceedings of Indian Geotechnical Conference, Kochi

Flood I (1991) A Gaussian-based neural network architecture and complementary training algorithm. In: Proceedings of the International Joint Conference on Neural Networks, New York, pp 171–176

Flood I, Kartam N (1994) Neural networks in civil engineering I: principles and understanding. J Comput Civil Eng 8(2):131–148

Hassoun MH (1995) Fundamentals of artificial neural networks. MIT Press, Cambridge

Hecht-Nielsen R (1989) Theory of the back-propagation neural network. In: Proceedings of the International Joint Conference on Neural Networks, Washington, DC, pp 593–605

Karsoliya S (2012) Approximating number of hidden layer neurons in multiple hidden layer BPNN Architecture. Internat J Eng Trends Technol 3(6):714–717

Lapedes A and Farber R (1988) How neural networks work. Neural information processing systems. American Institute of Physics, College Park, pp 442–456

Masters T (1993) Practical neural network recipes in C++. Academic Press, San Diego

Pandian NS, Nagaraj TS, Manoj M (1997) Re-examination of compaction characteristics of fine-grained soils. Geotechnique 47(2):363–366

Patel RS, Desai MD (2010) CBR Predicted by index properties of soil for alluvial soils of South Gujarat, Indian Geotechnical Conference, Proc. IGC I: 79–82

Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge

Roy TK, Chattopadhyay BC (2006) Prediction of compaction characteristics of subgrade materials for the road. Proc Indian geotech conf Chennai 2:737–740

Roy TK, Chattopadhyay BC, Roy SK (2007) Prediction of CBR for subgrade of different materials from simple test, Proceedings of the International conference on civil engineering in the new millennium—opportunities and challenges—Kolkata

Roy TK, Chattopadhyay BC, Roy SK (2009) Prediction of CBR from compaction characteristics of cohesive soil. Highw Res J 77–88

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representation by error propagation. Parallel distributed processing. MIT Press, Cambridge

Sarle WS (1994) Neural networks and statistical models. In: Proceedings of the 19th Annual SAS Users Group International Conference, Cary, NC: SAS Institute, pp 1538–1550

Sarle WS (1995) Stopped training and other remedies for overfitting. In: Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, pp 352–360

Shahin MA, Jaksa MB, Maier HR (2008) State of the art of artificial neural networks in geotechnical Engineering. EJGE paper

Smith GN (1986) Probability and statistics in civil engineering: an introduction. Collins London

Smith M (1993) Neural networks for statistical modeling. Van Nostrand Reinhold, New York

Sridharan A, Nagaraj HB (2005) Plastic limit and compaction characteristics of fine-grained soils. Ground Improvement 9:17–22

Tang Z, Almeida C, Fishwick PA (1991) Time series forecasting using neural networks versus Box-Jenkins methodology. Simulation 57(5):303–310

Taylor B, Darrah M, Moats C (2003) Verification and validation of neural networks: a sampling of research in progress. Proceedings of SPIE 5103:8–16

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82