Prediction of Compression Index of Soft Soils from the Brazilian Coast Using Artificial Neural Networks and Empirical Correlations

A.G. Oliveira Filho, L.B. Totola, K.V. Bicalho, W.H. Hisatugu

Abstract. This paper aims to explore the potential use of artificial neural networks (ANNs) to predict the compression index (C_C) of soft soils from the Brazilian coast. Results from 225 standard consolidation (oedometer) tests and the corresponding soil index properties (*i.e.*, initial void ratio, natural water content and Atterberg limits) of a wide variety of fine-grained soils reported in the literature were compiled and investigated herein. The ANN prediction performance is compared with linear empirical correlations created from the database investigated. In addition, correlations presented in the literature are also used and evaluated through different statistical techniques. Overall, for the organized dataset, the ANN outperformed the empirical correlations, highlighting the fragility and limitations of single and multiple variable linear empirical correlations.

Keywords: artificial neural network, compression index, empirical correlations, soft soils, statistical analysis.

1. Introduction

The idealized linear relationship between void ratio and log of effective consolidation pressure of a normally consolidated clay that applies over a range of stresses and void ratios defines the compression index (C_c). The C_c obtained from the consolidation test on clay may be a useful engineering approximation for predictions of consolidation settlement of normally consolidated natural clays. At a given value of effective consolidation pressure, the void ratio of a normally consolidated natural clay depends on the nature and amount of clay minerals present, as indicated by the liquid limit (Skempton, 1970). The greater the liquid limit, the higher the void ratio in a soil. Moreover, previous published empirical relationships between C_c and soil index properties are often used during preliminary investigation of suitability of a foundation site during planning stages.

While conducting the laboratory test is indispensable, it is also relatively time-consuming. In addition, sufficient undisturbed field samples are often difficult and costly to obtain. For these reasons, numerous studies have been made to predict the C_C from soil index properties, obtained from tests more easily carried out (Djoenaidi, 1985). Many researchers have published empirical correlations estimating C_C from soil index properties around the world (*e.g.* Terzaghi & Peck, 1967; Azzouz *et al.*, 1976; Ozer *et al.*, 2008; Kalantary & Kordnaeij, 2012; McCabe *et al.*, 2014; and Kootahi & Moradi, 2016) and for Brazilian soft soils (Futai *et al.*, 2008; Coutinho & Bello, 2014; Baroni & Almeida, 2017). However, empirical correlations may not be applied to soils elsewhere without consideration of soil origin, and the multiplicity of existing empirical correlations indicates the need of evaluation criteria for their use.

The artificial neural networks (ANNs) technique has been used in geotechnical engineering for prediction of engineering properties of soils based on previously known index properties of these soils. The work of Rumelhart *et al.* (1986) on the backpropagation algorithm is a milestone in the use of ANN in civil engineering studies. Further studies on the application of ANN in geotechnical engineering include the prediction of properties like the hydraulic conductivity in clays (Goh, 1995), the optimum water content and the corresponding maximum dry density of the soil (Najjar *et al.*, 1996) and the residual friction angle prediction of clays (Das & Basudhar, 2008). ANNs were also used for soils settlement estimation (Nejad *et al.*, 2009; Benali *et al.*, 2013) and shear strength parameter prediction (Khanlari *et al.*, 2012).

Due to their learning capacity, ANNs are less influenced by the natural variability of C_C and therefore are a potential tool in estimating the parameter. The use of ANN for the C_C prediction is presented in some studies (Ozer *et al.*, 2008; Park & Lee, 2011; Kalantary & Kordnaeij, 2012; Kurnaz *et al.*, 2016) and all of them presented satisfactory results.

This paper aims to explore the potential use of a computer-based modelling technique namely ANN to predict

Amandio Gonçalves de Oliveira Filho, M.Sc., Departamento de Engenharia Civil, Universidade Federal do Espírito Santo, Vitória, ES, Brazil. e-mail: amandio.oliveira@aluno.ufes.br.
Lucas Broseghini Totola, M.Sc., Departamento de Engenharia Civil, Universidade Federal do Espírito Santo, Vitória, ES, Brazil. e-mail: lucas.totola@aluno.ufes.br.
Kátia Vanessa Bicalho, Ph. D., Full Professor, Departamento de Engenharia Civil. e-mail: katia.bicalho@ufes.br.
Willian Hiroshi Hisatugu, D.Sc., Full Professor, Departamento de Computação e Eletrônica. e-mail: wilian.hisatugu@ufes.br.
Submitted on June 1, 2019; Final Acceptance on Januray 29, 2020; Discussion open until August 31, 2020.
DOI: 10.28927/SR.431109

C_C using measured index soil properties. A collected database containing results of 295 standard oedometer laboratory tests and corresponding soil index properties, carried out on a wide variety of fine-grained soils from Brazilian coast and reported by different researchers was compiled by the authors. Thus, a wide range of soils and soil properties, including low and high plasticity soils from the Brazilian coast, are investigated. The ANN ability to overcome limitations of single and multiple variables linear correlations is evaluated by comparing ANNs predictions performances with single empirical correlations created for the database investigated. In addition, results of correlations presented in the literature are also evaluated through different statistical techniques. These empirical correlations use simple or multiple variables to predict C_C from index properties such as natural water content (w_n), initial void ratio (e_0) and liquid limit determined by the Casagrande method (LL_{CUP}).

2. Material and Methods

2.1. Database soil description

The results of standard consolidation (or oedometer) tests and the corresponding index properties of 295 soft soils from different deposits of the Brazilian coast are investigated. The dataset reported by different researchers (Table 1) and compiled in this paper are derived from academic studies providing a high-quality database of laboratory consolidation tests. The data include test results on low and high plasticity soils from six Brazilian coastal states, up to 3000 km away from each other: Espírito Santo (ES), Santa Catarina (SC), Pernambuco (PE), Rio de Janeiro (RJ), Rio Grande do Sul (RS) and São Paulo (SP). The standard oedometer test in Brazil is carried out according to ABNT NBR-12007 (ABNT, 1990).

The predictive model capacity is highly dependent on experimental database quality. Laboratory data may contain inaccuracies associated with experimental laboratory errors. For this reason, and assuming the oedometer tests are performed on fully saturated samples (S = 100 %), some of the soils presented data inconsistency and were removed from the investigation. In addition, Tukey's Rule was applied for identification and exclusion of systematic bias or outliers. The outliers are values below Q1 - 1.5*(Q3 - Q1) or above Q3 + 1.5*(Q3 - Q1), where Q3 and Q1 are the first and third quartile of the dataset, respectively.

From the mentioned preprocessing, 70 out of 295 samples were removed. Table 2 presents the statistical properties of the 225 remaining samples from the dataset. Besides C_C, the soil index properties examined are the natural water content (w_n), the initial void ratio (e_0), the plasticity index (PI) and the liquid limit determined by the Casagrande method (LL_{CUP}).

The fine-grained particles of a soil govern the compressibility, constituting, mainly, the silt and clay fractions, with particle sizes range smaller than about 200 mesh sieve size (0.074 mm). From liquid limit (LL_{CUP}) and plasticity index (PI) values, it is possible to classify these fines through the Casagrande plasticity chart, which is used by the Unified Soil Classification System (USCS) as shown in Fig. 1. The chart analysis shows the heterogeneity of Brazilian coast fine-grained soils with a wide range of LL_{CUP} values. These values suggest large variability of clay mineral groups within the investigated database. Most of the investigated soils (*i.e.*, 88 %) are classified either as high plasticity clays (CH) or as high plasticity silts and organics soils (MH-OH).

Table 2 - Statistical description of the selected 225 experimental results.

Variable	Minimum	Maximum	Mean	Standard deviation
LL_{CUP} (%)	25	211	94.46	43.18
PI (%)	4	136.1	54.92	29.55
w_n (%)	29	221.34	97.5	41.53
e_0	0.73	5.66	2.54	1.03
C_c	0.09	3.27	1.3	0.71

Table 1 - References sources and location of the soil samples investigated from Brazilian coast.

Reference	Location	n**
Baran (2014)	Araranguá, Florianópolis, Itajaí, Palhoça, Penha and Tubarão (SC), Recife (PE) and Rio Grande (RS)	109
Kootahi & Moradi (2016)	Juturnaíba, Macaé and Rio de Janeiro (RJ), Recife (PE)	50
UFES* Geotechnical Laboratory	Grande Vitória (ES)	56
Póvoa (2016)	Macaé (RJ)	6
Queiroz (2013)	Itaguaí (RJ)	8
Silva (2013)	Duque de Caxias (Sarapuí and REDUC) and Queimados (RJ), Recife and Suape (PE), Florianópolis (SC) and Santos (SP)	66

*Federal University of Espírito Santo.
**Number of soil samples.

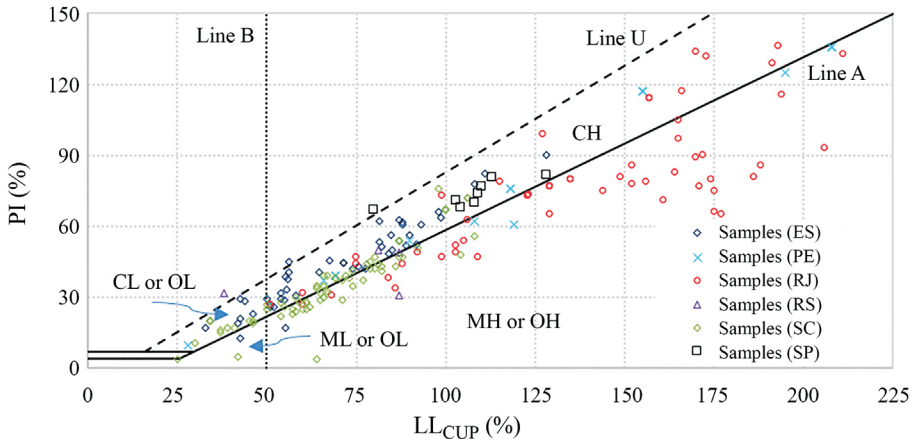

Figure 1 - Casagrande plasticity chart of investigated soil samples.

Table 3 presents a statistical summary of C_C for each soil class. As expected, the soils within the high plasticity classes have the higher values of C_C. These classes also present a wide range of C_C values (*i.e.* from 0.17 to 3.08 and 0.44 to 3.27, respectively).

The Pearson correlation coefficient was used to evaluate the correlation among the used properties. From Table 4, w_n and e_0 showed relatively strong correlation with C_C (*i.e.*, 0.87 and 0.86, respectively). In addition, the pairs w_n - e_0 and LL$_{CUP}$-PI showed strong correlations which may affect ANN performance.

Table 4 - Correlation among the properties investigated.

	C_C	LL$_{CUP}$	PI	w_n	e_0
C_C	1				
LL$_{CUP}$	0.82	1		Symmetrical	
PI	0.79	0.93	1		
w_n	0.87	0.71	0.71	1	
e_0	0.86	0.70	0.70	0.99	1

2.2. Existing correlations investigated

As presented in Table 5, several empirical correlations have been previously published by different investigators for estimating C_C values from different local sites. Most of these equations estimate C_C from single-variable regressions with w_n (C1-C4), LL$_{CUP}$ (C6-C8) or e_0 (C10 - C12). Multiple regression correlations (C14-C15) are also presented and account for both mineralogical composition and soil structure influence on C_C. It should be noted that all these correlations are in linear form and show a direct and positive relationship between C_C and the input properties investigated.

2.3. Artificial neural networks

ANNs are characterized as artificial intelligence (AI) techniques inspired by the structure of the human brain to simulate its operation in computational systems in a simplified way. The neural networks are distinguished by performing three essential operations: learning and storing knowledge; applying the knowledge acquired in solving proposed problems; as well as acquiring new knowledge from constant learning (Khanna, 1990).

The artificial neuron is the basic processing element of an ANN. A neuron model is formed by a set of input connections (x_j), synaptic weights (w_{kj}), where k is the number of input neurons and j corresponds to the input stimulus; and the bias (b_k), a weighting parameter which can increase or decrease the value of the linear combination of inputs of the neuron activation function ($f.$). Figure 2 presents a simplified model of an artificial neuron, where (u_k) represents the linear combination of input signals, and (y_k) corresponds to the output value of the neuron, adapted from Haykin (2001). Thus, the input weighting process represents the learning rate acquired by an ANN. The weights are adjusted as the input dataset is presented to the network. The supervised learning process in an ANN is based on the

Table 3 - C_C Statistical summary by USCS Division of investigated soil samples.

USCS	Samples	Minimum	Maximum	Average	Standard deviation
CH	108	0.17	3.08	1.25	0.62
CL	25	0.09	1.31	0.47	0.29
MH-OH	90	0.44	3.27	1.62	0.69
ML-OL	2	0.25	0.38	0.32	0.09

Table 5 - Selected empirical correlations and their references used to estimate compression index from index properties.

Independent variable(s)	ID	References	Equations	Region
w_n	C1	Azzouz *et al.* (1976)	$C_c = 0.0100w_n - 0.05$	Greek and North-American Clays
	C2	Castello & Polido (1986)	$C_c = 0.014w_n - 0.17$	Brazilian Coast soft clays (ES)
	C3	Coutinho & Bello (2014)	$C_c = 0.014w_n - 0.094$ ($w_n < 200$) $C_c = 0.004\ w_n + 1.738$ ($w_n > 200$)	Brazilian Coast Marine clays (PE)
	C4	Kootahi & Moradi (2016)	$C_c = 0.012(w_n - 7.75)$	Marine fine-grained soils worldwide
LL_{CUP}	C6	Terzaghi & Peck (1967)	$C_c = 0.009(LL_{CUP} - 10)$	All clays
	C7	Castello & Polido (1986)	$C_c = 0.01(LL_{CUP} - 8)$	Brazilian Coast Marine clays (ES)
	C8	Kootahi & Moradi (2016)	$C_c = 0.012(LL_{CUP} - 8)$	Marine fine-grained soils worldwide
e_0	C10	Azzouz *et al.* (1976)	$C_c = 0.400\ e_0 - 0.100$	Greek and North-American Clays
	C11	Castello & Polido (1986)	$C_c = 0.228\ e_0 + 0.22$	Brazilian Coast Marine clays (ES)
	C12	Kootahi & Moradi (2016)	$C_c = 0.510(e_0 - 0.33)$	Marine fine-grained soils worldwide
e_0, LL_{CUP}, w_n	C14	Azzouz *et al.* (1976)	$C_c = 0.37(e_0 + 0.003LL_{CUP} + 0.0004w_n\text{-}0.34)$	Greek and American Clays
e_0, LL_{CUP}	C15	Kootahi & Moradi (2016)	$C_c = 0.374(e_0 + 0.01LL_{CUP} - 0.47)$	Marine fine-grained soils worldwide

adjustment of the synaptic weights so that the output value is the closest possible to the expected value.

The activation function (*f.*) has the objective of limiting the input signals of the network in a specific range, usually between [0;1] or [-1;1], to generate the output neuron from the input values x_i of the network and the adjusted weights. The most used functions in geotechnical research are log-sigmoid, tan-sigmoid and linear. The ANN architecture is the way the network presents the arrangement of its neurons (Fig. 2). The structure can be in a single hidden layer or in several layers. The layers located between the entry and exit layers are called intermediate or hidden layers.

2.3.1. Backpropagation algorithm

The multilayer perceptron (MLP) is a multilayered artificial neural network composed of sigmoidal activation functions in the hidden layers. In this type of architecture, the hidden layer uses a non-linear activation function, such as the sigmoidal function, giving the network a genuinely non-linear model. For MLP, unlike simple Perceptron, the error *e* is not obtained simply from the difference between the desired output and the output calculated by the network because there are now intermediate layers. Hence, for the training stage, Rumelhart *et al.* (1986) proposed the backpropagation algorithm, one of the most used in practical applications of ANN.

The algorithm principle is to estimate the error of the intermediate layers by estimating the effect caused in the output layer error, using the descending gradient. The error is thus backpropagated in the network to correct the synaptic weights of the hidden layers. For this reason, the activation functions need to be continuous, differentiable, such as logistic functions, and hyperbolic tangent (Braga *et al.*, 2001). The Levenberg-Marquardt (LM) algorithm, an optimization of the backpropagation algorithm, is an iterative numerical optimization technique capable of locating the minimum of a function expressed as the sum of squares of other nonlinear functions and is widely used in ANN studies.

The LM backpropagation is an adaptive network, where each node in the network has the same node function. LM function uses the Jacobian matrix for calculations that assume the performance as a mean or sum of squared error (Kannaiyan *et al.*, 2019). The LM algorithm can be better explained in Hagan & Menhaj (1994) and Raina *et al.* (2009).

2.3.2. Proposed ANN models

The potential of ANN to estimate the compression index (C_C) is investigated by developing different ANNs models. The simulations in this study have been carried out in the *MATLAB* environment. Using the toolbox *nftool*, the ANNs were trained with the Levenberg-Marquardt (LM) training algorithm.

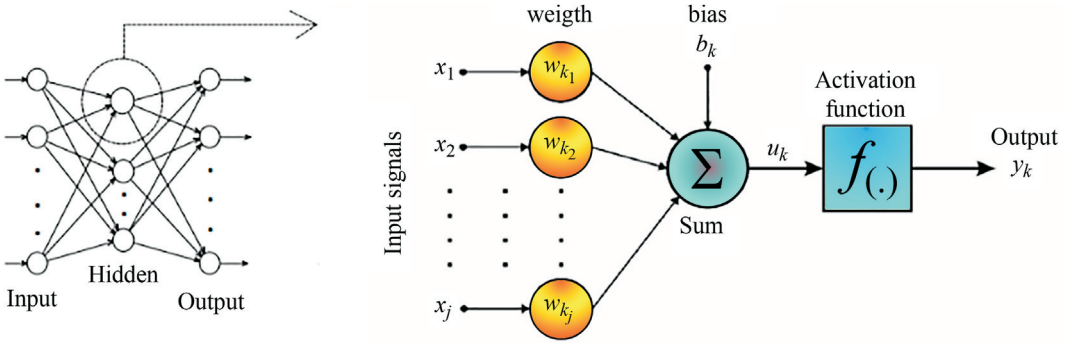

Figure 2 - ANN Architecture and Non-linear neuron model adapted from Haykin (2001) and Shahin *et al.* (2001).

Shahin (2013) points out that in most geotechnical problems, the architecture of an ANN is usually obtained through trial and error. Hornik *et al.* (1989) demonstrated a single hidden layer as sufficient to approximate any continuous function. Caudil (1988) points out that the number of hidden neurons (*H*) in a network with a single layer is a function of the number of input variables (*I*), as shown in Eq. 1.

$$H < (2I + 1) \tag{1}$$

On this hand, the networks in this study were trained with a single hidden layer, with the number of neurons in the hidden layer varying (2, 4, 6 and 10), and using four input parameters: LL_{CUP}, PI, e_0 and w_n. The log-sigmoid activation function in the hidden layer and the linear function in the output layer were used. In the training stage, 80 % of the total soil samples were used (177 selected experimental data). From these, 70 % were used for training, 15 % for validation and 15 % for testing. It is important to highlight that in *nftool toolbox,* these three steps make up the network training. Overall, in the first step, the network adjusts the synaptic weights minimizing the error function (*i.e.*, the mean squared error, MSE, Eq. 2). The validation samples are used to measure the generalization of the networks, and their errors are used to correct the synaptic weights and stop the training process. The testing samples are independent measures of the ANN. The remaining soil samples (20 %) were reserved for the cross-validation test (48 selected data). In this stage, the generalization capacity of the model is evaluated. The soil samples used for the cross-validation test were not included in the training stage.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(C_{c,\,measured} - C_{c,\,predicted})^2 \tag{2}$$

The selection of samples for training and cross-validation subsets followed the representativeness criteria including different domains of variation of C_c, a strategy distribution of the samples as discussed by Fortin *et al.* (1997). Thus, all the ranges of the histogram shown in Fig. 3 are represented in both training and cross-validation sets.

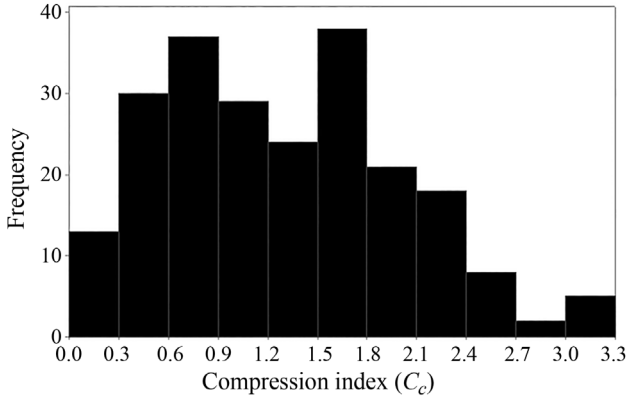

Figure 3 - Compression index histogram of the selected experimental results.

3. Results

3.1. Statistical evaluation methodology

It is not easy to choose the best estimation method to use, requiring criteria in its selection, which suggests the use of statistical techniques to evaluate the different methods used in geotechnical investigations. As a methodology to evaluate the estimation capacity of empirical correlations and ANNs, some statistical criteria are assessed: (i) the root mean square error (*RMSE*), (ii) the estimated and measured compression index ratio (*K*), (iii) the ranking index (*RI*) and (iv) the ranking distance (*RD*). The methodology is based on previous publications (Briaud & Tucker, 1988; Giasi *et al.*, 2003; Ozer *et al.*, 2008; Onyejekwe *et al.*, 2015; Güllü *et al.*, 2016).

The *RMSE* is the root mean square error of the difference between the estimated and measured values, which consequently attributes greater weight to the largest errors. Values close to zero indicate better model performance. The *RMSE* is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(C_{c,\,estimated} - C_{c,\,measured})^2} \tag{3}$$

where n is the number of observations, $C_{C,\,estimated}$ is the value predicted by the empirical correlations or ANNs, and $C_{C,\,measured}$ is the experimental C_C value obtained from the oedometer test.

The average and standard deviation of $C_{C,\,estimated}$ and $C_{C,\,measured}$ ratio (K) assesses how an equation underestimates or overestimates a value and compose accuracy and precision measurement parameters of the *RI* and *RD* methods. K is calculated as:

$$K = \frac{C_{c\,,\,estimated}}{C_{c\,,\,measured}} \tag{4}$$

Briaud & Tucker (1988) point out that the accuracy of K in evaluating the predictive capacity of a method is represented by the average of K. The precision of the method, is given by the scatter of the estimated values around the average of K, which is measured by the standard deviation of K. Theoretically, the factor K varies between zero and infinity, with an optimal value equal to one (Briaud & Tucker, 1988). Values of ($K < 1$) and ($K > 1$) indicate if results are underestimated or overestimated, respectively (Abu-Farsakh & Titi, 2004). The best estimated results are associated with the average (μ) of K close to one and standard deviation (*SD*) of K close to zero (Abu-Farsakh & Titi, 2004; Güllü *et al.*, 2016).

The ranking index (*RI*) was proposed by Briaud & Tucker (1988) to alleviate the problem of non-symmetric distribution of K values. The *RI* is a general index that relates μ and *SD* of all the K values of a group of estimates of a variable and provides a judgment of the accuracy and precision of the estimation. For the evaluation, low values of *RI* indicate good performance of the prediction model. The *RI* is determined by the formula:

$$RI = \left|\mu(\ln[K])\right| + SD(\ln[K]) \tag{5}$$

The ranking distance (*RD*) is an alternative general index proposed for assessing the quality of a calculation method, assessing the accuracy and precision of an estimation. As *RI*, *RD* also considers μ and *SD* of all the K values of the analyzed data (Giasi *et al.*, 2003). For assessment, low *RD* values indicate high accuracy and precision in the estimated values. The *RD* is determined as follows:

$$RD = \sqrt{(1-\mu_{[k]})^2 + (SD_{[k]})^2} \tag{6}$$

The performance is also evaluated through the coefficient of correlation for equations (R^2). In general, R^2 value has the objective of evaluating the relationship between two variables, from "n" observations of those variables, indicating how much the independent variable can be explained by the fixed variable. The R^2 values close to 1 indicate a better correlation between two variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(C_{c\,,\,estimated\,,\,i} - C_{c\,,\,measured\,,\,i}\right)^2}{\sum_{i=1}^{n}\left(C_{c\,,\,measured\,,\,i} - \overline{C}_{c\,,\,measured}\right)^2} \tag{7}$$

where $C_{C,\,measured}$ = input value, $C_{C,\,estimated}$ = estimated output value, $\overline{C}_{c,\,measured}$ = average input values and n = number of variables.

3.2. New empirical correlations proposed

Based on Pearson correlation values, Table 4, there is a strong relationship between C_C and index properties w_n, e_0 and LL_{CUP}. For this reason, three new single linear empirical correlations have been developed to estimate C_C from these properties for the Brazilian coast soft soils. The correlations have been created in the statistical software *Minitab*. The same 177 samples used for ANN training were used to create the empirical correlations. The remaining 20 % (48 samples) will be used as cross-validation test for both ANNs models and correlations created.

The Kolmogorov-Smirnov hypothesis test has been used to evaluate the residuals normality and homoscedasticity diagnosis has been performed to ensure empirical correlations validity. At the time the normality hypothesis was denied, points outside the 95 % interval have been removed (outliers) and the analysis has been repeated (Berger & Zhou, 2014). After statistical analysis, three new simple adjustment empirical correlations were determined for investigated Brazilian coast soft soils as follows (Table 6). All of them present a p-value (> 0.05) which indicates a good adherence to normal distribution.

Figures 4a, 4b and 4c present graphically the empirical correlations created and those from Table 5, with C_C as a function of w_n, e_0 and LL_{CUP}, respectively. It is noticed that empirical correlations could yield to the different results in a wide range of variability, which shows the particularity of the correlations with the local geological sites of the selected soil samples for the modelling.

3.3. Trained ANN performance

The C_C prediction capacity for Brazilian coast soft soils of the trained ANNs were evaluated by different statistical parameters (*i.e.*, *RMSE*, *K*, *RI*, *RD* and R^2). The predicted C_C values are compared to the measured C_C values determined from laboratory oedometer test. The results obtained by the four trained ANNs are summarized in the Tables 7 and 8 for the training and cross-validation sets,

Table 6 - Correlations proposed for the investigated database.

ID	Variable(s)	Equation	R^2
C5	w_n	$C_C = 0.01601w_n - 0.3209$	0.83
C9	LL_{CUP}	$C_C = 0.6155e_0 - 0.3521$	0.55
C13	e_0	$C_C = 0.01581LL_{CUP} - 0.138$	0.83

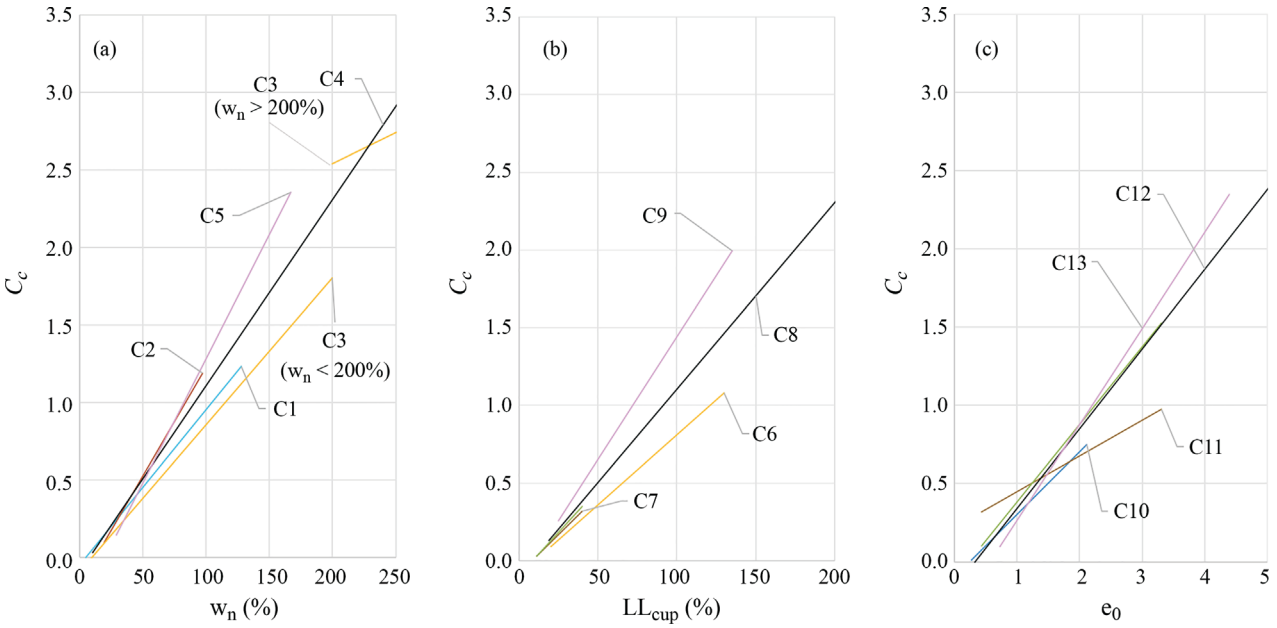

Figure 4 - Graphical representation of empirical correlations for C_C prediction using (a) w_n, (b) LL_{CUP}, and (c) e_0.

Table 7 - Statistical measurements for ANNs performances (training set).

ID	Input parameters	Hidden neurons	RMSE	*K*				*RI*	*RD*	R^2
				% < 1	% > 1	μ	*SD*			
ANN1	PI, e_0, LL_{CUP}, w_n	2	0.31	46.3	53.7	1.13	0.45	0.38	0.47	0.80
ANN2		4	0.262	46.9	53.1	1.10	0.41	0.34	0.43	0.86
ANN3		6	0.30	46.9	53.1	1.15	0.44	0.40	0.46	0.82
ANN4		10	0.28	52.0	48.0	1.07	0.38	0.32	0.38	0.84

Table 8 - Statistical measurements for ANNs performances (cross-validation set).

ID	Input parameters	Hidden neurons	*RMSE*	*K*				*RI*	*RD*	R^2
				% < 1	% > 1	μ	*SD*			
ANN1	PI, e_0, LL_{CUP}, w_n	2	0.28	52.1	47.9	1.10	0.35	0.33	0.36	0.88
ANN2		4	0.26	45.8	54.2	1.14	0.33	0.34	0.35	0.90
ANN3		6	0.30	45.8	54.2	1.15	0.43	0.40	0.46	0.87
ANN4		10	0.29	54.2	45.8	1.07	0.29	0.28	0.30	0.87

respectively. Each ANN was trained varying the number of neurons in the hidden layer.

The high values of R^2 associated with low *RMSE* values are evidence of good statistical performance of the ANNs trained for C_C prediction. The R^2 values range between 0.80-0.90 and *RMSE* values range between 0.26-0.30 considering the cross-validation sets. The cross-validation set results prove the ANN ability to generalize the acquired knowledge.

In terms of precision, assessed by the *K* values, a balance between the percentage values of ($K > 1$) and ($K < 1$) is observed for the trained ANNs, with a slight tendency to overestimate the C_C for most of the models ($\mu > 1$).

Overall, the increase in number of neurons in the hidden layer has not provided significant improvement in the C_C prediction ability for the investigated experimental dataset. It is suggested that the use of 4 neurons is enough. Thus, in this study, good prediction performance was reached using a number of hidden neurons equal to the number

of inputs of the model, which agrees with the work of Ozer *et al.* (2008).

3.4. Empirical correlations performance

The C_C prediction capacity for Brazilian coast soft soils of both empirical correlations developed from this study and the correlations presented in the literature also were evaluated by statistical parameters (*i.e.*, *RMSE*, *K*, *RI*, *RD* and R^2). The cross-validation set is used to illustrate and compare the behavior of empirical correlations and ANN models. Table 9 shows the results.

The common behavior for the empirical correlations was a tendency of underestimate C_C ($\mu_{[k]} < 1$), especially those correlating C_C and LL_{CUP}. Also, the empirical correlations tended to underestimate C_c for soil samples of more compressible soils (higher C_c values). The LL_{CUP} values are strongly affected by clay mineralogy. Thus, there are limitations in the application of the investigated empirical correlations to soils from different geological origin or to soils with a C_C range outside the limits of the data from which the correlation was created.

For the reason cited, the empirical correlations proposed for this study database C5 and C13 (with w_n and e_0, respectively) presented the lower *RMSE*, *RI* and *RD* values. These two soil index properties have the strong correlation with C_C, which explains how they better explain C_C variation. In the same way, the multi-variable correlation C15, proposed by Kootahi & Moradi (2016) (for marine soils around the world and similar to soils in this study database) had a reasonable statistical performance. Also, the correlation was not extrapolated.

From Tables 7, 8 and 9, none of the correlations have satisfied all the evaluation criteria concomitantly and the trained ANNs presented statistical performance more consistent than the empirical correlations. For a general assessment, low values of *RI and RD* close to each other correspond to higher accuracy estimation, which is observed for ANNs and empirical correlations C5, C13 and C15, as shown in Fig. 5. Even though the ANN results regarding *RI* and *RD* are similar with these 3 correlations (two of them proposed for this study database), the *RMSE* for the ANNs can be up to 35 % lower.

The results reinforce the need for using more than one statistical parameter in the evaluation of different estimation methods. Analyzing only one parameter alone can lead to erroneous and meaningless conclusions. However, in this specific study, the values of *RD* and *RI* did not add conclusions and the analysis could be limited to *RMSE* and the values of *K* and $\mu_{[K]}$. Even though Güllü *et al.* (2016) point out that the *RD* index assigns equal weight to both accuracy and precision of estimation and provides more information than the *RI* index and *RMSE*, a crescent linear variation of both $SD_{[K]}$ with *RD* index and *RMSE* and $\mu_{[K]}$ with *RI* index is observed. That said, they seem to indicate the same behavior.

Figures 6 and 7 present the results when the performance by soil class is evaluated according to USCS. As shown in Table 3, there is limited data for ML-OL and CL samples, so only the CH, MH-OH samples are analyzed.

Table 9 - Statistical summary of the C_C estimated from empirical correlations cross-validation set.

Variable(s)	ID	*RMSE*	*K*				*RI*	*RD*	R^2
			% < 1	% > 1	μ	*SD*			
w_n	C1	0.56	79.2	20.8	0.83	0.37	0.60	0.41	0.85
	C2	0.33	58.3	41.7	1.04	0.42	0.35	0.42	
	C3	0.55	77.1	22.9	0.86	0.41	0.59	0.44	
	C4	0.41	75.0	25.0	0.95	0.41	0.45	0.41	
	C5	**0.32**	**58.3**	**41.7**	**1.03**	**0.37**	**0.33**	**0.37**	
LL_{CUP}	C6	0.73	89.6	10.4	0.67	0.25	0.82	0.42	0.77
	C7	0.63	87.5	12.5	0.77	0.30	0.69	0.38	
	C8	0.48	64.6	35.4	0.92	0.36	0.51	0.37	
	C9	0.40	31.3	68.8	1.20	0.47	0.48	0.51	
e_0	C10	0.57	79.2	20.8	0.82	0.36	0.38	0.40	0.82
	C11	0.78	79.2	20.8	0.81	0.53	0.80	0.56	
	C12	0.39	68.8	31.3	1.00	0.41	0.41	0.41	
	C13	**0.35**	**60.4**	**39.6**	**1.02**	**0.37**	**0.36**	**0.38**	
e_0, LL_{CUP}, w_n	C14	0.54	79.2	20.8	0.83	0.33	0.58	0.38	0.84
e_0, LL_{CUP}	**C15**	**0.36**	**62.5**	**37.5**	**0.99**	**0.36**	**0.36**	**0.36**	**0.86**

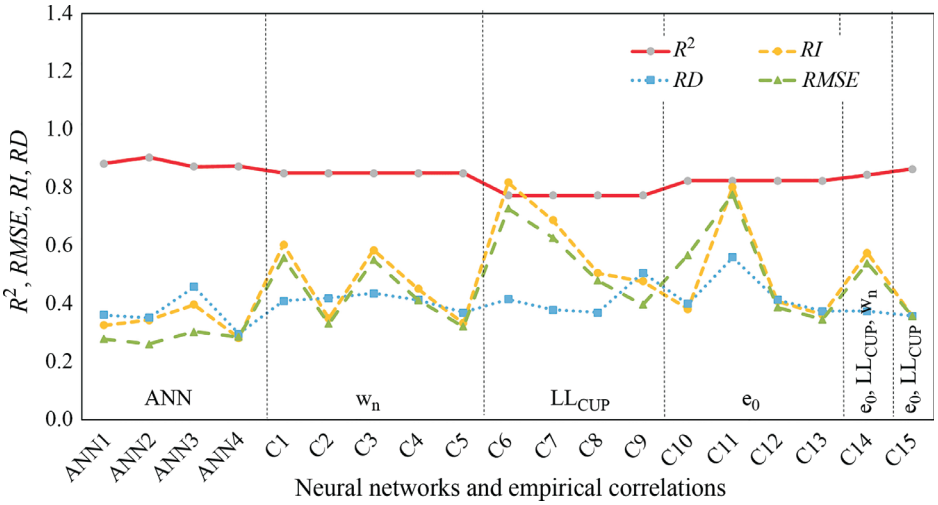

Figure 5 - Results of R^2, RMSE, RI and RD evaluation for each empirical correlation and ANNs for cross-validation set.

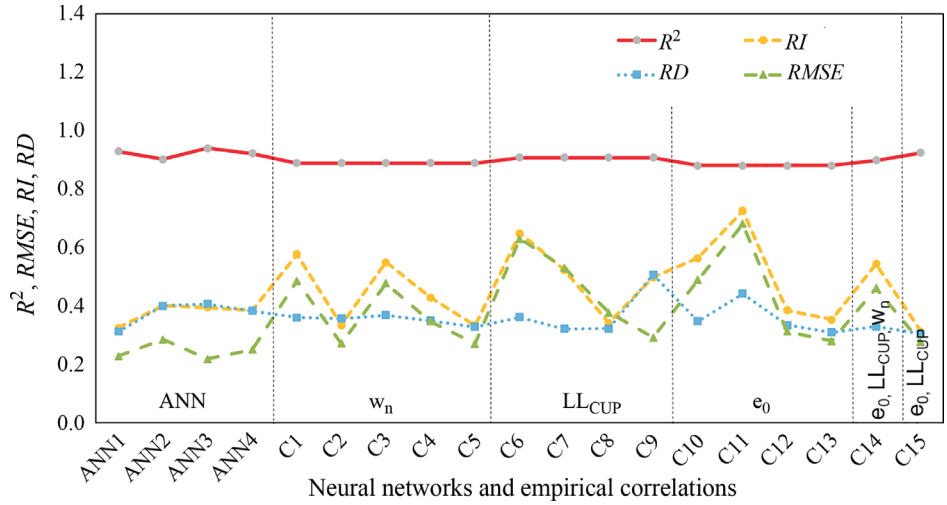

Figure 6 - Results of R^2, RMSE, RI and RD by each empirical correlation and ANNs for CH soils for cross-validation set (n = 26).

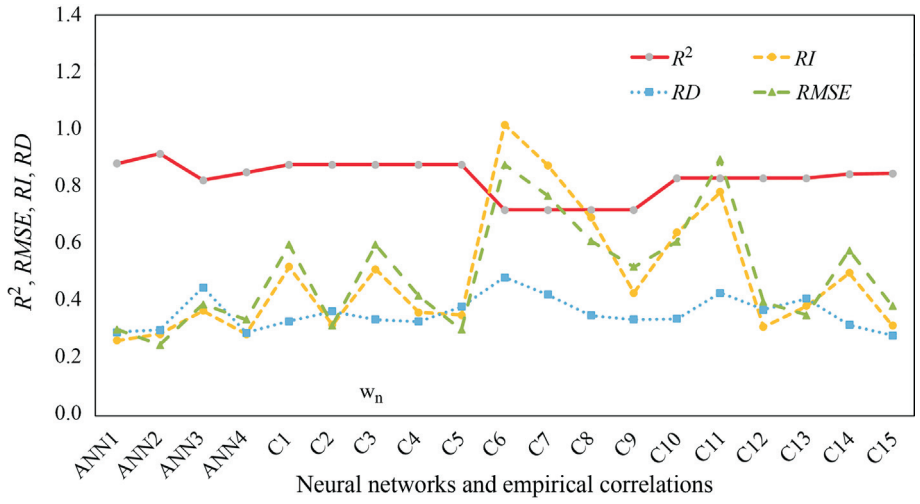

Figure 7 - Results of R^2, RMSE, RI and RD by each empirical correlation and ANNs for MH-OH soils for cross-validation set (n = 16).

Overall, both CH and MH-OH show similar results when analyzed by ANNs, however, the MH-OH class presented best precision and accuracy, as reveled by lower *RI* and *RD* values. The empirical correlations seem to follow the same tendency of the ANNs. Again, C5 and C15 are the correlations which better approximate the minimum values of *RMSE, RI* and *RD,* for both soil sample classes.

The ANN results grouped by Brazilian states are shown in Fig. 8. Due to the reduced number of cross-validation samples for SP, RS and PE Brazilian states, a statistical evaluation for these geological sites was not possible. In terms of *RMSE*, *RD* and *RI*, the best minimum adjustments between the curves occurred for soil samples from RJ and SC, indicating highest accuracy and precision

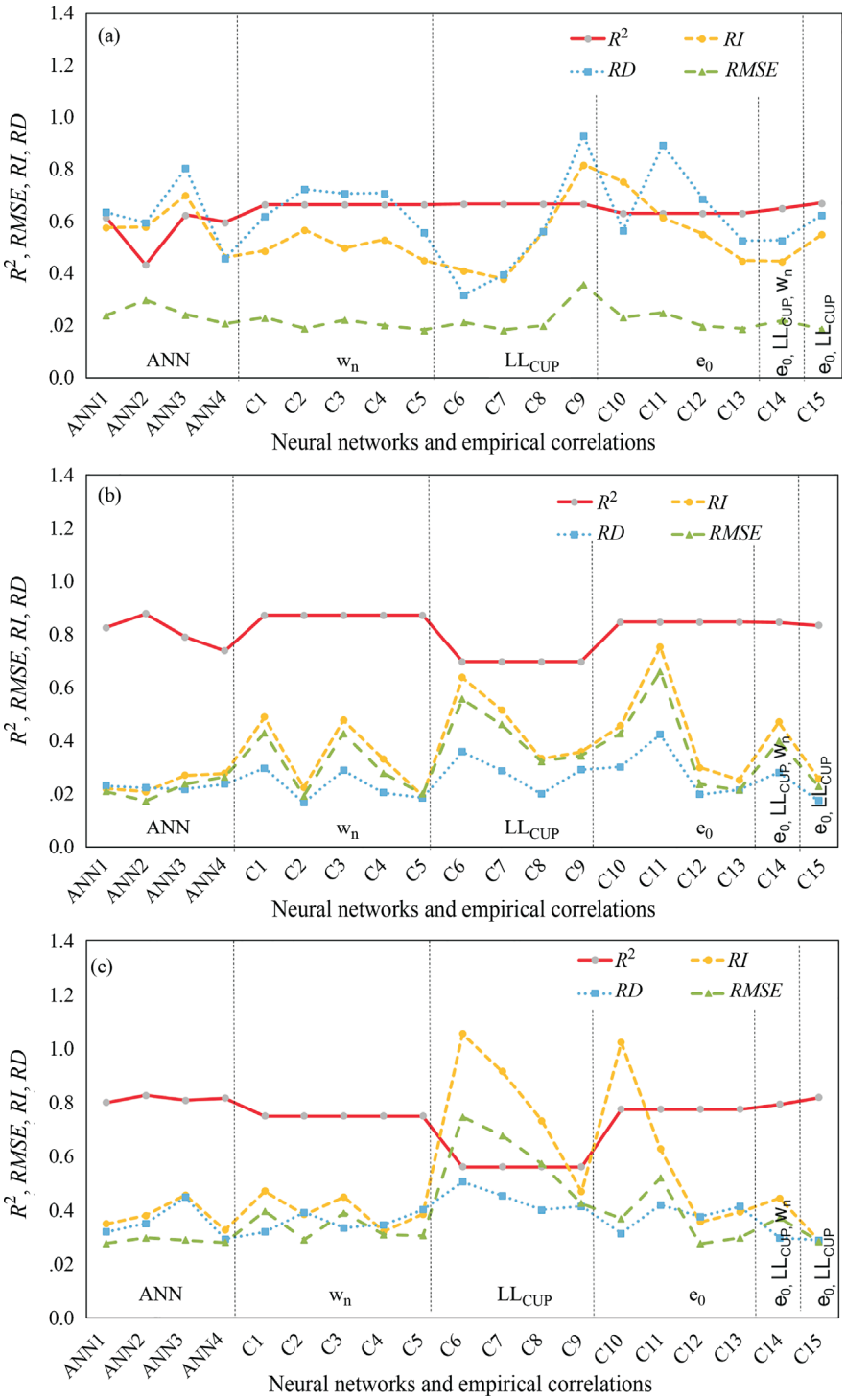

Figure 8 - Results of R^2, RMSE, RI and RD by empirical correlations and ANNs for cross-validation samples by (a) ES (n = 10), (b) RJ (n = 13) and (c) SC (n = 15).

of the predicted C_c values, especially for the network ANN2. The C_c interval variation for ES soil samples is 0.13-1.53, against 0.29 - 3.08 for RJ and 0.09 - 2.20 for SC samples. These observations indicate best performances of the ANN estimations for soils with higher compressibility.

In addition, Fig. 9 compares measured and estimated C_c for the empirical correlations and trained ANNs with the best and worst statistical performance for cross-validation sets. It is possible to notice that ANN results are close to each other, while greater divergence is noted for empirical correlations. Overall, both, empirical correlations and ANNs tend to underestimate higher C_c values ($C_c > 1$). According to the calculated statistical parameters, the best distribution of the C_c estimated by the ANNs around the equal line for any value of C_c can be graphically observed in all cases. The C_c empirical correlations with w_n present results closer to the equal C_c measured line than correlations with e_0 and LL_{CUP}, corroborating the statistical parameters previously evaluated.

4. Conclusions

In this paper, the performance of ANNs and widely used single and multi-variable empirical equations for compression index estimation was evaluated using a dataset of 225 fine-grained soils with a wide range of soil properties (*i.e.*, LL_{CUP} values ranging from 25 up to 200 %) from six Brazilian coastal states. Different networks have been trained with 2, 4, 6 and 10 neurons in a single hidden layer. The ANN training used the Levenberg-Marquardt algorithm (LM), the log-sigmoid activation function in the hidden layer and the linear function in the output layer. In addition, new empirical correlations were proposed for Brazilian coast soft soils using the least squares regression and residual analysis test techniques. The performance of both empirical correlations and ANNs have been evaluated through statistical techniques that include: (i) the root mean square error (*RMSE*), (ii) the ratio of the estimated to measured compression index (*K*), (iii) the ranking index (*RI*) and (iv) the ranking distance (*RD*).

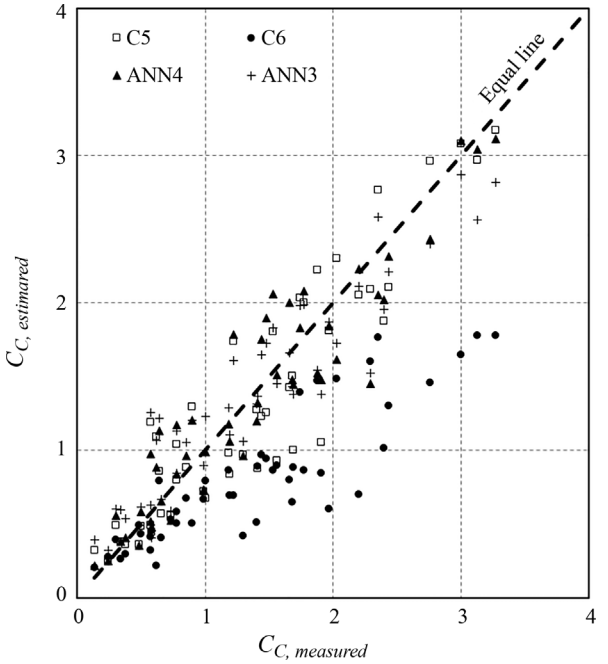

Figure 9 - Comparison between measured and estimated C_c for empirical correlation and ANN for cross-validation set (n = 48) by best (C5-ANN4) and worst (C6-ANN3) performances.

Overall, the proposed ANN models outperformed the empirical correlations investigated, which is proved by the statistical parameters used. The minimum *RMSE* was 0.26 for the trained ANNs, 0.32 for the single empirical correlations created and 0.36 for the empirical correlations from the literature. The main reason for that is the underestimation of C_c for samples of more compressible soils by the empirical correlations. Among the input properties, the empirical correlations proposed in this study correlating C_c with w_n (C5) and e_0 (C13) showed the best estimation results. Also, the better performance of proposed correlations over those from the literature proves the influence of soil geological origin on the prediction capacity performance.

It is noteworthy that the empirical correlations are usually better applied to the modelling soil sample sites, and several standard oedometer tests using different soil samples in an investigated site are required for determination of C_c due to the observed heterogeneity of Brazilian coast fine-grained soils. By their ability to learn, adapted ANNs are less influenced by the site natural spatial variability. Moreover, the ANN method can always be updated by presenting new training soil examples as new data with measured C_c and corresponding index properties become available. Thus, these presented results reveal that the adapted ANNs created for estimation of soft soils C_c from the Brazilian coast have potential application as an alternative to the empirical correlations during preliminary investigation of suitability of a foundation site during planning stages.

In addition, the authors propose to send to readers the *MATLAB .mat* file, which contains the synaptic weights from the trained ANN4, which can be used in *MATLAB* environment, for prediction of compression index from Brazilian soil samples in preliminary investigation studies.

Acknowledgments

The first author is grateful to Petrobras for their support. The second author is grateful to CAPES for financial support. The third author is grateful to CNPq for financial support. The authors are grateful to the editor and reviewers for their interest in our research and for helpful comments that considerably improved our manuscript.

References

ABNT (Associação Brasileira de Normas Técnicas) NBR 12007, (1990). Soil - Unidimensional Consolidation

Test - Procedure. ABNT, Rio de Janeiro, Brazil, 15 p. (in Portuguese).

Abu-Farsakh, M.Y. & Titi, H.H. (2004). Assessment of direct cone penetration test methods for predicting the ultimate capacity of friction driven piles. Journal of Geotechnical and Geoenvironmental Engineering, 130(9):935-944.

Azzouz, A. S.; Krizek, R. J. & Corotis, R. B. (1976). Regression analysis of soil compressibility. Tokyo. Soils Foundations, 16(2):19-29.

Baran, K.R., (2014). Compressibility Geotechnical Properties of a Soft Clay From Itajaí-SC. Master Dissertation, Department of Civil Engineering, Technological Center, Federal University of Santa Catarina, Santa Catarina, 335 p.

Baroni, M. & Almeida, M.S.S. (2017). Compressibility and stress history of very soft organic clays. Proceeding of Institution of Civil Engineers - Geotechnical Engineering, 170(2):148-160. Thomas Telford Ltd.

Berger, V.W. & Zhou, Y.Y. (2014). Kolmogorov-Smirnov Test: Overview. Wiley Statsref: Statistics Reference Online. John Wiley & Sons, Ltda, New York, pp. 1-5.

Benali, A.; Nechnech, A. & Bouzid, A. (2013). Principal Component Analysis and Neural Networks for Predicting the Pile Capacity Using SPT. International Journal of Engineering and Technology, 5(1):162-169.

Braga, A.P.; Carvalho, A.P. & Ludermir, T.B. (2001). Neural Networks: Theory and Applications, 2nd ed. LTC, Rio de Janeiro, 226 p.

Briaud, J.L. & Tucker L.M. (1988). Measured and predicted axial load response of 98 piles. J Geotech Eng (ASCE), 114(9):984-1001.

Castello, R.R. & Polido, U.F. (1986). Some characteristics of consolidation of marine clays from Vitória, ES. In. VIII Brazilian Congress of Soil Mechanics and Foundation Engineering, Rio Grande do Sul, pp. 149-159.

Caudil, M. (1988). Neural networks primer, Part III. AI Expert, 3(6):53-59.

Coutinho, R.Q. & Bello, M.I.M.C.V. (2014). Geotechnical characterization of Suape soft clays, Brazil. Soils and Rocks, 37(3):257-276.

Das, S.K. & Basudhar, P.K. (2008). Prediction of residual friction angle of clays using artificial neural network. Engineering Geology, 100(3-4):142-145.

Djoenaidi, W.J. (1985). Compendium of Soil Properties and Correlations. Doctorate Thesis, The School of Civil and Mining Engineering, University of Sydney, Sydney, 788 p.

Fortin, V.; Ouarda, T.B.M.J. & Bobée, B. (1997). Comment on "The use of artificial neural networks for the prediction of water quality parameters" by HR Maier and GC Dandy. Water Resources Research, 33(10):2423-2424.

Futai, M.M., Almeida, M.S.S.; Lacerda, W.A. & Marques, M.E.S. (2008). Laboratory behavior of Rio de Janeiro soft clays. Index and compression properties- part 1. Soils and Rocks, 3(2):69-75.

Giasi, C.I.; Cherubini, C. & Paccapelo, F. (2003). Evaluation of compression index of remoulded clays be means of Atterberg limits. Bulletin of Engineering Geology and the Environment, 62(4):333-340.

Goh, A.T.C. (1995). Modeling soil correlations using neural networks. Journal of Computing in Civil Engineering, ASCE, 9(4):275-278.

Güllü, H.; Canakci, H. & Alhashemy, A. (2016). Use of ranking measure for performance assessment of correlations for the compression index. European Journal of Environmental and Civil Engineering, 22(5):578-595.

Hagan, M.T. & Menhaj, M.B. (1994). Training feedforward networks with the Marquardt algorithm. IEEE transactions on Neural Networks, 5(6):989-993.

Haykin, S. (2001). Neural Networks. A Comprehensive Foundation, 2nd ed. Pearson Education, McMaster University, Hamilton, Ontario, Canada, 823 p.

Hornik, K.; Stinchcombe, M. & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359-366.

Kalantary, F. & Kordnaeij, A. (2012). Prediction of compression index using artificial neural network. Scientific Research and Essays, 7(31):2835-2848.

Kannaiyan, M.; Govindan, K. & Raghuvaran, J.G.T. (2019). Prediction of specific wear rate for LM25/ZrO2 composites using Levenberg-Marquardt backpropagation algorithm. Journal of Materials Research and Technology, 9(1):530-538.

Khanlari G.R.; Heidari, M.; Momeni, A.A. & Abdilor, Y. (2012). Prediction of shear strength parameters of soils using artificial neural networks and multivariate regression methods. Engineering Geology, 131-132(1):11-18.

Khanna, T. (1990). Foundations of Neural Networks. Addison-wesley Publishing Company, New York, 196 p.

Kootahi, K. & Moradi, G. (2016). Evaluation of compression index of marine fine-grained soils by the use of index tests. Marine Georesources & Geotechnology, 35(4):548-570.

Kurnaz, T.F.; Dagdeviren, U.; Yildiz, M. & Ozkan, O. (2016). Prediction of compressibility parameters of the soils using artificial neural network, 5(1801):1-11.

McCabe, B.A.; Sheil, B.B.; Long, M.M.; Buggy, F.J. & Farell, E.R. (2014). Empirical correlations for the compression index of Irish soft soils. Geotechnical Engineering, 167(6):510-517.

Najjar, Y.M.; Basheer, I.A. & Naouss, W.A. (1996). On the identification of compaction characteristics by neuronets. Computers and Geotechnics, 18(3):167-187.

Nejad, F.P.; Jaksa, M.B.; Kakhi, M. & McCabe, B.A. (2009). Prediction of pile settlement using artificial neural networks based on standard penetration test data. Computers and Geotechnics, 36(7):1125-1133.

Onyejekwe, S.; Kang, X. & Ge, L. (2015). Assessment of empirical equations for the compression index of fine-grained soils in Missouri. Bulletin of Engineering Geology and the Environment, 74(3):705-716.

Ozer, M.; Isik, N.S. & Orhan, M. (2008). Statistical and neural network assessment of the compression index of clay-bearing soils. Bull Eng Geol Environ., 67(4):537-545.

Park, H.I. & Lee, S.R. (2011). Evaluation of the compression index of soils using an artificial neural network. Computers and Geotechnics, 38(4):472-481.

Póvoa, L.M.M. (2016) Geotechnical Characterization of a Soft Soil Deposit in Lowland Area Located in Macaé-RJ. Master Degree Dissertation, Civil Engineering, Federal University of Norte Fluminense Darcy Ribeiro (UENF). Campos dos Goytacazes, Rio de Janeiro, 156 p.

Queiroz, C.M. (2013). Geotechnical Properties of a Soft Clay Deposit in the Region of Itaguaí-RJ. Master Degree Dissertation, Geotechnics and Transprotation Course, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, 118 p.

Raina, R.; Madhavan, A. & Ng, A.Y. (2009). Large-scale deep unsupervised learning using graphics processors. Proc. 26th annual international conference on machine learning. ACM, pp. 873-880.

Rumelhart, D.E.; Hilton, G.E. & Williams, R.J. (1986). Learning representation by back-propagation errors. Nature, 323(6088):533-536.

Shahin, M.A.; Jaksa, M.B. & Maier, H.R. (2001). Artificial neural network applications in geotechnical engineering. Australian geomechanics, 36(1):49-62.

Shahin, M.A. (2013). Artificial intelligence in geotechnical engineering: Applications, modeling aspects, and future directions. In: X.S. Yang *et al*., Metaheuristics in Water, Geotechnical and Transport Engineering. 1st ed. Elsevier, Amsterdan, pp. 169-204.

Skempton, A.W. (1970). The consolidation of clays by gravitational compaction. Q. J. Geol. Sot. 125(1-4):373- 411.

Silva, D.M. (2013). Estimation of the Compressive Index of Soft Clays of the Brazilian Coast From Characterization Tests, Master Dissertation, Department of Civil Engineering, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, 182 p.

Terzaghi, K. & Peck, R.B. (1967). Soil Mechanics in Engineering Practice. 2nd ed. Wiley, New York, 512 p.