Proceedings of the XVII ECSMGE-2019
Geotechnical Engineering foundation of the future
ISBN 978-9935-9436-1-3

doi: 10.32075/17ECSMGE-2019-0598

XVII European Conference
ECSMGE 2019
Reykjavik
on Soil Mechanics
and Geotechnical Engineering

Data-driven approach to predict unconfined compression strength of laboratory soil stabilized with cementitious binders

Approche basée sur des données pour prédire la résistance à la compression non confinée des sols stabilisés en laboratoire avec des liants à base de ciment

J. Tinoco
ISISE – Institute for Sustainability and Innovation in Structural Engineering/ALGORITMI Research Center, University of Minho, Guimarães, Portugal

A. Alberto
CIEPQPF - Research Centre on Chemical Processes and Forest Products Engineering,University of Coimbra, Coimbra, Portugal

P.J. Venda Oliveira, L. Lemos
ISISE – Institute for Sustainability and Innovation in Structural Engineering,University of Coimbra, Coimbra, Portugal

A. Gomes Correia
ISISE – Institute for Sustainability and Innovation in Structural Engineering,University of Minho, Guimarães, Portugal

ABSTRACT: Uniaxial compressive strength (q_u) of soil stabilized with cementitious binders is a key feature for design purposes. However, its measurement requires extensive laboratory tests, which is time and resources consuming. Accordingly, aiming to make this process faster and cheaper, this paper presents a novel approach for q_u estimation of soil stabilized with cementitious binders based on soft computing techniques, particularly Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs). For models training, a database comprising 444 records, encompassing cohesionless to cohesive and organic soils, different binder types, mixture conditions and curing time was compiled. The results show a promising performance in q_u prediction of laboratory soil-cement mixtures, being the best results achieved with the SVM model ($R^2 = 0.94$). In addition, by averaging SVM and ANN predictions a slightly better accuracy can be achieved ($R^2 = 0.95$). Through the application of a sensitivity analysis over the fitted models, it is measured the relative importance of each model attributes, which highlighted the major effects of water/cement ratio, cement content, organic matter content and curing time, which are known as preponderant in soil-cement mixtures behaviour.

RÉSUMÉ: La résistance en compression uniaxiale (q_u) des sols stabilisés avec liants à base de ciment est un élément très important pour le projet. Toutefois, sa mesure nécessite des essais intensifs en laboratoire, qui demande du temps et des ressources. Pour permettre un processus plus rapide et moins cher, ce travail présente une nouvelle approche pour l'estimation de q_u des sols stabilisés avec des liants à base de ciment, basée sur des

techniques informatiques, (en particulier) "Support Vector Machines" (SVMs) et "Artificial Neural Networks" (ANNs). Les modèles sont utilisés avec une base de données comprenant des 444 données, englobant sols non cohésifs, cohésifs et organiques, différents types des liants, différents conditions de mélange et des temps de durcissement. Les résultats montrent une performance prometteuse dans la prédiction de q_u avec des mélanges de sol-ciment préparés en laboratoire, et les meilleurs résultats sont obtenus avec le modèle SVM ($R^2 = 0.94$). En complément, avec la moyenne de SVM et ANN sont obtenus prédictions avec une précision légèrement meilleure ($R^2 = 0.95$). Avec l'implémentation d'une analyse de sensibilité sur les modèles utilisés, on mesure l'importance relative des attributs de chaque modèle, qui a souligné l'importance du rapport eau/ciment, le teneur du ciment, le teneur de la matière organique et le temps de durcissement, qui sont connu comme les plus prépondérant dans le comportement de mélanges de sol-ciment.

Keywords: Soil-cement mixtures; jet grouting; deep soil mixing; soft computing; sensitivity analysis

1 *INTRODUCTION*

Mechanical properties study of soil-cement mixtures is a complex task due to high number of parameters involved. Over the last decades, several researches have been conducted, following different approaches but with the same purpose of a better understand of soil-cement mixtures behaviour over time.

Concerning to uniaxial compression strength (q_u, *MPa*), this mechanical property is obtained through laboratory tests that involves time and resources consuming, which are generally very limited. Therefore, it is important to reduce the number of laboratory tests without compromising safety or confidence issues. A common practice is to prepare (before construction works) and test some laboratory samples aiming to simulate the field conditions. These samples, prepared with the same soil, cement and water used in the field, will give an important idea about the behaviour of the in field mixture. However, this laboratory samples also represent an important cost for the project and therefore should be minimized.

This scenario underlines the necessity, at least upon at a pre-design stage, to have available prediction tools to obtain the best design parameters. However, due to the high number of parameters affecting the behaviour of soil-cement mechanical properties, in particularly the q_u, the traditional statistical analysis are unable to deal with.

Aiming to overcome this limitation, a first and successful attempt have been recently made, taking advantage of the high learning capabilities of Data Mining (DM) techniques (Tinoco et al., 2014; Gomes Correia et al., 2014. Although a good performance have been achieved in q_u prediction of laboratory soil-cement mixtures with an $R^2 = 0.93$ (see Gomes Correia et al. (2014) for more details), there are some limitations that still need to be overcome. In particular, the model dependence on the mixture properties, such as its porosity, is one of its main drawbacks. As can be observed in Figure 1, which shows the relative importance of each input variables in q_u prediction, the mixture porosity (only measured after mixture preparation) has a relative importance higher than 15%. Moreover, these models were developed based on a database regarding soil-cement samples covering mostly high cement dosages (Gomes Correia et al., 2014).

Hence, aiming to eliminate models dependence on the final mixtures properties, namely its porosity, as well as increase their applicability domain, a new data-driven model is here proposed for q_u prediction over time without considering any property of the final soil-cement mixture and covering a larger range of cement contents. For that, a set o ten input

variables such as the cement content, soil grain size distribution or type of binder was select to feed the models.

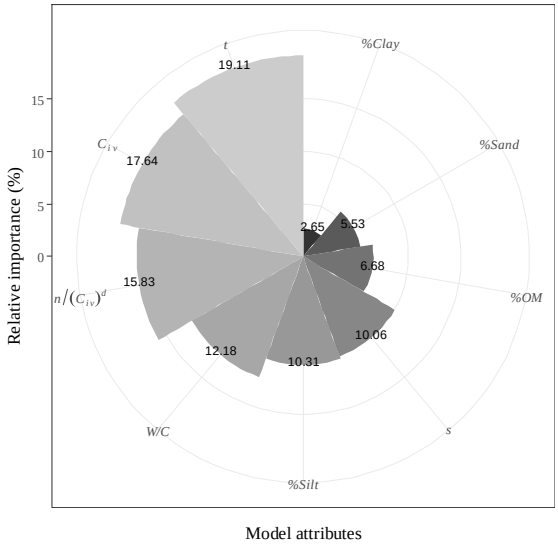

Figure 1. Relative importance of each input variable in q_u prediction of laboratory soil-cement mixtures according to SVM algorithm.

2 *METHODOLOGY*

2.1 Modelling

For q_u modelling it was followed a data driven approach where three different DM algorithms were fitted to a database previously compiled and prepared containing unconfined compression tests results related to laboratory soil-cement mixtures, as well as a set of ten input variables related to the soil and cement characteristics used to prepare the mixture. In particular, two of the high flexible learning DM algorithms were trained, namely Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs). Bellow is presented a brief overview of the two DM algorithms applied in this study, highlighting the adopted parameters for each one.

Initially developed for classification tasks (Cortes and Vapnik, 1995), SVMs were latter adapted to regression tasks thanks to the introduction of ε-insensitive loss function (Smola and Schölkopf, 2004). The main purpose of the SVMs is to transform input data into a high dimensional feature space using non-linear mapping. This transformation depends on a kernel function. In this work the popular Gaussian kernel was adopted. In this context, its performance is affected by three parameters: γ, the parameter of the kernel; C, a penalty parameter; and ε (only for regression), the width of an ε-insensitive zone. The heuristics proposed by Cherkassky and Ma (2004) were used to define the first two parameter values, $C = 3$ (for a standardised output) and $\varepsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \cdot \sum_{1=1}^{N}(y_i - \hat{y}_i)^2$, y_i is the measured value, $\hat{y}_i$ is the value predicted by a 3-nearest neighbour algorithm and N is the number of examples. A grid search of $2^{\{-15;-11;-7;-3;1\}}$ was adopted to optimise the kernel parameter γ, under an internal threefold cross-validation scheme.

Concerning to ANNs, they are a method of artificial intelligence, which seeks to simulate the biological structure of the human brain and nervous system through their architecture (Kenig et al., 2001). ANNs are a technique capable of modelling complex non-linear mappings and is robust in exploration of data with noise. In this study the multilayer perceptron that contains only feedforward connections, with one hidden layer containing H processing units, was adopted. Because the network's performance is sensitive to H (a trade-off between fitting accuracy and generalisation capability), it was adopted a grid search (similar to the one used for SVM) of $\{0; 2; 4; 6; 8\}$ during the learning phase to find the best H value. Such grid search only considered training data, dividing it into fitting (70%) and validation data (30%), where the validation error was used to select the best H. After selecting the best H value, the ANN is retrained with the whole training data. The neural function of the hidden

nodes was set to the popular logistic function $1/(1+e^{-x})$.

The R statistical environment (R Team, 2009) and the rminer package (Cortez, 2010), were used to conduct all experiments.

2.2 Model Evaluation

For models comparison and accuracy measurement, three metrics currently used in regression problems were calculated (Hastie et al., 2009): Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coeficient of correlation (R^2). A low value of MAE and RMSE and an R^2 close to the unit value means a higher predictive capacity. The main difference between MAE and RMSE is that the latter one is more sensitive to extreme values since it uses the square of the distance between the real and predicted values (Tinoco et al., 2014). In addition to this three metrics it was taken also advantage of Regression Error Characteristic (REC) curve proposed by Bi and Bennett (2003), which plots the error tolerance on the *x*-axis versus the percentage of points predicted within the tolerance on the *y*-axis, allowing a quick and easy comparison of different DM models. For models generalization purposes, a cross-validation (k-fold = 10) approach (Hastie et al., 2009) was applied and the entire process was repeated 5 times.

Understanding what was learned by the models is also a key point in any data driven project. Since data driven models, particularly SVM or ANN that rely on complex statistical analysis and are frequently referred to as "black boxes", are mathematically very complex it urges the necessity to "open" such models in order to facilitate its understanding. Aiming to overcome this drawback, Cortez and Embrechts (2013) proposed a novel visualization approach based on sensitivity analysis (SA), which is used in this work. SA is a simple method that is applied after the training phase and measures the model responses when a given input is changed, allowing the quantification of the relative importance of each attribute as well as its average effect on the target variable. In particular, it was applied the Global Sensitivity Analysis (GSA) method (Cortez and Embrechts, 2013), which is able to detect interactions among input variables. This is achieved by performing a simultaneous variation of *F* inputs. Each input is varied through its range with *L* levels and the remaining inputs fixed to a given baseline value. In this work, it was adopted the average input variable value as a baseline and set $L = 12$, which allows an interesting detail level under a reasonable amount of computational effort.

With the sensitivity response of the GSA, the input importance barplot can be ploted, which shows the relative influence (R_a) of each input variable in the model (from 0% to 100%). The rational of GSA is that the higher the changes produced in the output, the more important is the input. To measure this effect, first the gradient metric (g_a) for all inputs was calculated. After that, the relative influence was computed according to the following equation:

$$R_a = g_a/g_i \cdot 100(\%), \quad where \ g_a = \sum_{j=2}^{L} \left|\hat{y}_{a,j} - \hat{y}_{a,j-1}\right|/(L-1) \qquad (1)$$

where *a* denotes the input variable under analysis and $\hat{y}_{a,j}$ is the sensitivity response for $x_{a,j}$.

2.3 Database

For models training and testing purposes, a database with 444 records was collected and compiled. These samples make part of different laboratory studies carried out on Universities of Minho and Coimbra (Tinoco et al., 2014; Venda Oliveira et al., 2014; Correia et al., 2015). The soils used in the preparation of the laboratory samples were collected from eight test sites. One of them is Coimbra area (located in Portugal), ranging from cohesive to cohesionless soils, organic to nonorganic soils, presenting different

geotechnical properties. Fourteen different binders were tested, including Portland cement, slag, fly ash, lime and silica fume, applied individually or combined. Concerning to the seven remaining sites, all of them are of clayey nature, containing different percentages of sand, silt, clay and organic matter (Gomes Correia et al., 2014). These samples were prepared with cement type CEM I 42.5R (Portland cement with 100% clinquer) and CEM II 42.5R (composed Portland cement with ≥65% clinquer). In addition, a couple of samples were also prepared with pozzolanic cement (CEM IV/A 35.5R with ≥20% clinquer).

A set of 10 variables were selected to models input. The definition of such variables took into account the empirical knowledge related to soil-cement mixtures behavior, particularly concerning to the q_u evolution over time (Sariosseiri and Muhunthan, 2009; Lorenzo and Bergado, 2004). Bellow are listed all 10 input variables considered in this study for q_u prediction.

- *%Clay* – Clay content (%)
- *%Sand* – Sand content (%)
- *%Silt* – Silt content (%)
- *%OM* – Organic matter content (%)
- ω_0 – Water content (%)
- a_w – Cement content (%)
- *W/C* – Water/Cement ratio
- t – Age of the mixture (days)
- C_s – Coefficient related with the binder type
- L_2 – Coefficient related with a secondary binder

Table 1 summarizes the main statistics of all 10 inputs variables as well as of the output variable, showing the wide range of cement content as well as the q_u values.

Table 1. Summary of the main statistics of the input and output variables used in q_u prediction

Variable	Minimum	Maximum	Mean	Standard deviation
%Clay	0.00	45.00	19.84	14.31
%Sand	0.00	99.00	22.97	22.10
%Silt	1.00	79.00	57.17	18.15
%OM	0.00	19.40	5.87	4.64
ω_0	7.17	113.05	64.96	24.48
a_w	3.00	284.32	55.42	69.21
W/C	0.63	10.91	3.30	2.05
t	3.00	90.00	25.71	15.42
C_s	0.20	0.38	0.22	0.06
L_2	0.00	1.00	0.61	0.49
q_u	0.10	13.19	2.77	2.72

3 *RESULTS AND DISCUSSION*

The average hyperparameters and fitting time values (and respective 95% level confidence intervals according to a *t-student* distribution) of the two DM algorithms trained for q_u prediction of laboratory soil-cement mixtures (i.e. ANN and SVM) are shown in Table 2.

The achieved results shows a promising performance in q_u prediction of laboratory soil-cement mixtures based on the set of inputs selected that not include any information about the mixture properties. In fact, as shown in Table 3, both ANN and SVM algorithms (further refered only as ANN.Lab and SVM.Lab for shorten) were able to predict q_u very accurately, haven achieved an $R^2 = 0.94$.

Based on MAE or RMSE it is possible to observe that the SVM.Lab is able to predict q_u with a slightly higher accuracy when compared with ANN.Lab.

Table 2. Hyperparameters and computation time for each fitted model.

Model	Hyperparameter	Time (s)
ANN.Lab	$H = 7 \pm 1$	17.18 ± 0.45
SVM.Lab	$\gamma = 0.21 \pm 0.04;\ C = 4.84 \pm 0.22;\ \varepsilon = -5.57 \pm 0.50$	9.63 ± 0.28

Although ANN.Lab and SVM.Lab models present a very high performance, it was observed that q_u prediction accuracy can be improved by averaging ANN.Lab and SVM.Lab predictions (ANN&SVM.Lab for shorten). With this trick, an $R^2 = 0.95$ is achieved as well as an RMSE very close to 0.61 MPa (see Table 3). Figure 2, that plots the REC curves of each model, illustrates this slightly better performance in q_u prediction by averaging ANN.Lab and SVM.Lab predictions.

Table 3. Models performance comparison based on metrics MAE, RMSE and R^2.

Model	MAE	RMSE	R^2
ANN.Lab	0.46 ± 0.02	0.69 ± 0.05	0.94 ± 0.01
SVM.Lab	0.43 ± 0.01	0.67 ± 0.03	0.94 ± 0.01
ANN&SVM.Lab (average)	0.41 ± 0.01	0.61 ± 0.02	0.95 ± 0.00

Figure 3 depicts the histogram of the prediction error according to ANN&SVM.Lab model. As shown, only few prediction have a deviation higher than 1MPa, which represent a very high performance. The two sashed line in the graph represent the 5% and 95% quantiles, that corresponds to a deviation of -0.79MPa and 1.03MPa respectively.

From an engineering point of view, in addition to the model accuracy it is also important to understand what have been learned by it, particularly when dealing with ANN and SVM algorithms that are mathematically very complex. With this in mind, a GSA (Cortez and Embrechts, 2013) methodology was applied over the models in order to measure the influence of each model attribute in q_u prediction.

Figure 4 plots the relative importance of each input variable, showing that *W/C* is the most relevant variable in q_u prediction according to both ANN.Lab and SVM.Lab models, with a relative importance higher than 20%. The three next key variables are, according to SVM.Lab model, a_w, *%OM* and *t*. Based on ANN.Lab, the ranking is slightly different, being ω_0, *%Silt* and *%Sand* the next three most influent variables after *W/C*. Comparing both ANN.Lab and SVM.Lab models, the last one seems to be more realistic. In fact, among the four most relevant variables, SVM.Lab model includes the influence of the water and cement contents (*W/C* and a_w), soil organic matter content (*%OM*) and age of the mixture (*t*), which are known as preponderant in soil-cement mixtures behaviour (Lorenzo and Bergado, 2004; Consoli et al., 2011). According to ANN.Lab model, the effect of the cement content is less representative (only present on *W/C*) and the effect of the cure time only takes the sixth position in the ranking (less than 10%). As well known, the age of the mixture is one the most influent variables in soil-cement mixtures behaviour. Thus, considering models accuracy as well as the relative importance of each variable, SVM.Lab seems to be a better choice to estimate q_u development over time of laboratory soil-cement mixtures.

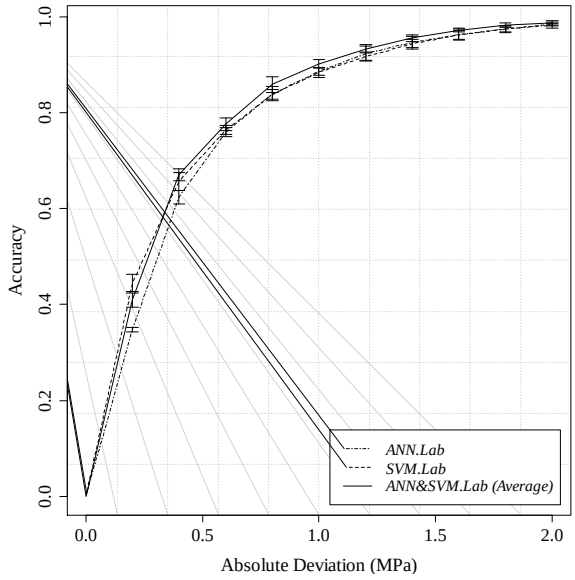

Figure 2. Comparison of ANN.Lab, SVM.Lab, and ANN&SVM.Lab performance in q_u prediction of laboratory soil-cement mixtures based on REC curves.

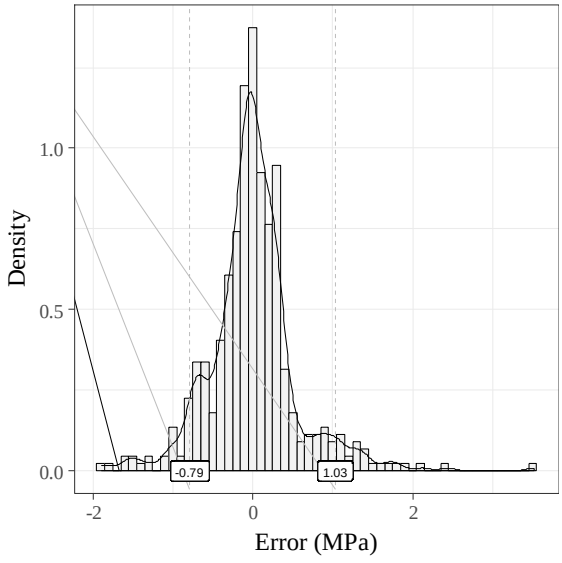

Figure 3. Histogram of the ANN&SVM.Lab prediction errors.

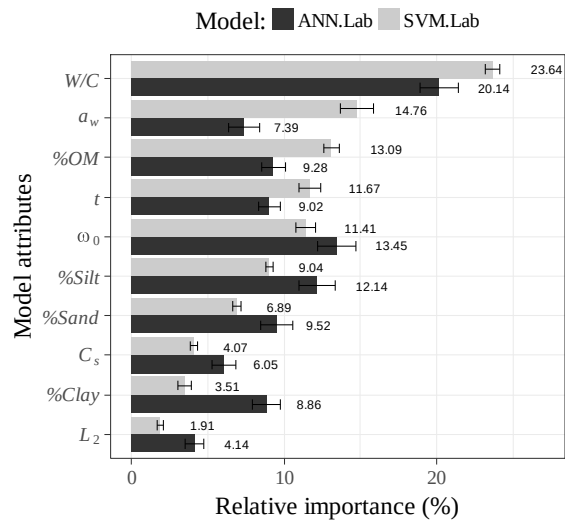

Figure 4. Comparison of the relative importance of each input variable based on a GSA.

4 *CONCLUSIONS*

A dada driven approach is proposed for uniaxial compressive strength (q_u) prediction of laboratory soil-cement mixtures. The proposed models, supported on a representative database comprising 444 records, are able to predict q_u over time with a very promising accuracy ($R^2 = 0.95$). In addition, only information available during the project stage, such as soil properties, binder and water content, is taken as model inputs. This way, the project design can calculate the expected q_u for different scenarios (formulations) without the need to prepare/test any sample. As a result a better optimization of the available resources can be done and consequently important economic benefits can be achieved.

The key variable in q_u prediction over time were also identified based on a global sensitive analysis (GSA). It was observed that the water/cement ratio (*W/C*) is the most relevant variable followed by cement content, soil organic matter content and age of the mixture.

5 *ACKNOWLEDGEMENTS*

This work was supported by FCT – "Fundação para a Ciência e a Tecnologia", within ISISE, project UID/ECI/04029/2013, and within CIEPQPF, project EQB/UI0102/2014, as well Project Scope: UID/CEC/00319/2013 and through the post-doctoral Grant fellowship with reference SFRH/BPD/94792/2013. This work was also partly financed by FEDER funds through the Competitivity Factors Operational Programme - COMPETE and by national funds through FCT within the scope of the projects POCI-01-0145-FEDER-007633, POCI-01-0145-FEDER-007043 and POCI-01-0145-FEDER-028382.

6 *REFERENCES*

Bi, J., Bennett, K.: Regression error characteristic curves. In: Proceedings of the Twentieth International Conference on Machine Learning, pp. 43-50. AAAI Press, Washington, DC, USA (2003)

Cherkassky, V., Ma, Y.: Practical selection of svm parameters and noise estimation for svm regression. Neural Networks 17(1), 113-126 (2004)

Consoli, N.C., Rosa, D.A., Cruz, R.C., Dalla Rosa, A.:Water content, porosity and cement content as parameters controlling strength of artificially cemented silty soil. Engineering Geology 122(3-4), 328-333 (2011)

Correia, A.A., Oliveira, P.J.V., Custódio, D.G.: Effect of polypropylene fibres on the compressive and tensile strength of a soft soil, artificially stabilised with binders. Geotextiles and Geomembranes 43(2), 97-106 (2015)

Cortes, C., Vapnik, V.: Support vector networks. Machine Learning 20(3), 273-297 (1995)

Cortez, P.: Data mining with neural networks and support vector machines using the r/rminer tool. In: 10th Industrial Conference on Data Mining, pp. 572-583. LNAI 6171, Springer, Berlin, Germany (2010)

Cortez, P., Embrechts, M.: Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences 225(Mar), 1-17 (2013)

Gomes Correia, A., Tinoco, J., Cortez, P.: Use of data mining in design of soil improvement by jet grouting. In: Second International Conference on Information Technology in Geo-Engineering (ICITG 2014), pp. 43-63. IOS Press,, Durham, UK (2014)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer-Verlag New York (2009)

Kenig, S., Ben-David, A., Omer, M., Sadeh, A.: Control of properties in injection molding by neural networks. Engineering Applications of Artificial Intelligence 14(6), 819-823 (2001)

Lorenzo, G., Bergado, D.: Fundamental parameters of cement-admixed clay-new approach. Journal of Geotechnical and Geoenvironmental Engineering 130(10), 1042-1050 (2004)

R Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Autria (2009). Web site: http://www.r-project.org/

Sariosseiri, F., Muhunthan, B.: Effect of Cement Treatment on Geotechnical Properties of some Washington State Soils. Engineering Geology 104(1-2), 119-125 (2009)

Smola, A., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199-222 (2004)

Tinoco, J., Gomes Correia, A., Cortez, P.: A novel approach to predicting young's modulus of jet grouting laboratory formulations over time using data mining techniques. Engineering Geology 169(Feb), 50-60 (2014).

Venda Oliveira, P.J., Correia, A.A., Lopes, T.J.: Effect of organic matter content and binder quantity on the uniaxial creep behavior of an artificially stabilized soil. Journal of Geotechnical and Geoenvironmental Engineering 140(9), 04014053 (2014)