Procedia Engineering

Volume 143, 2016, Pages 566–573

Advances in Transportation Geotechnics 3 . The 3rd International Conference on Transportation Geotechnics (ICTG 2016)

A Data-Driven Approach for q_u Prediction of Laboratory Soil-Cement Mixtures

Joaquim Tinoco[1*], António Alberto[2], Paulo da Venda[2], António Gomes Correia[1], and Luís Lemos[2]

[1]*University of Minho, Campus de Azurém, Guimarães, Portugal*
[2]*University of Coimbra, Coimbra, Portugal*
jtinoco@civil.uminho.pt, aalberto@dec.uc.pt, pjvo@dec.uc.pt, agc@civil.uminho.pt, llemos@dec.uc.pt

Abstract

In this paper a new data-driven approach is proposed for uniaxial compressive strength (q_u) prediction of laboratory soil-cement mixtures. The proposed model is able to predict q_u over time under different conditions, e.g. different cement contents or soil types, and can be applied at the pre-design stage. This means that the model can be applied previously to the preparation of any laboratory formulation. The designer only needs to collect information about the main geotechnical soil properties (grain size, organic matter content, among other) and select the binder composition to prepare the mixture.

Based on a sensitivity analysis, the key model variables were identified and its effect quantified. Thus, it was caught by the model the most relevant variables in q_u prediction over time and very high prediction capacity with an overall regression coefficient higher than 0.95.

Keywords: Soil-cement mixtures; Laboratory formulations; Uniaxial compressive strength; Data mining; Neuronal networks; Sensitivity analysis

1 Introduction

The uniaxial compressive strength (q_u) of soil-cement mixtures is a fundamental design parameter necessary for many transportation geotechnics applications. This mechanical property is obtained through laboratory tests requiring time, which is generally very limited. Consequently, is very useful to have at this stage, at least pre-design, prediction tools to obtain this design parameter. However, this not taken into account the number of variables that affect q_u and obviously the traditional statistical analysis is unable to deal with.

* Corresponding author e-mail: jtinoco@civil.uminho.pt

Selection and peer-review under responsibility of the Scientific Programme Committee of ICTG 2016

doi:10.1016/j.proeng.2016.06.073

Aiming to overcome this limitation, a first and successful attempt was recently made, taking advantage of the high learning capabilities of Data Mining (DM) techniques (Tinoco et al.; 2011; Tinoco et al., 2014; Gomes Correia et al, 2014). Although a good performance have been achieved for both strength and stiffness prediction of laboratory soil cement mixtures (see Figure 1), there are some limitations that still need to be overcome. In particular, the model dependence on the mixtures proprieties, such its porosity, is one of its main drawbacks. As can be observed in Figure 2, which shows and compare the relative importance of the input variables in q_u and E_0 (young modulus) prediction, the mixture porosity has a relative importance around 15% in q_u prediction and higher than 20% in E_0 prediction.

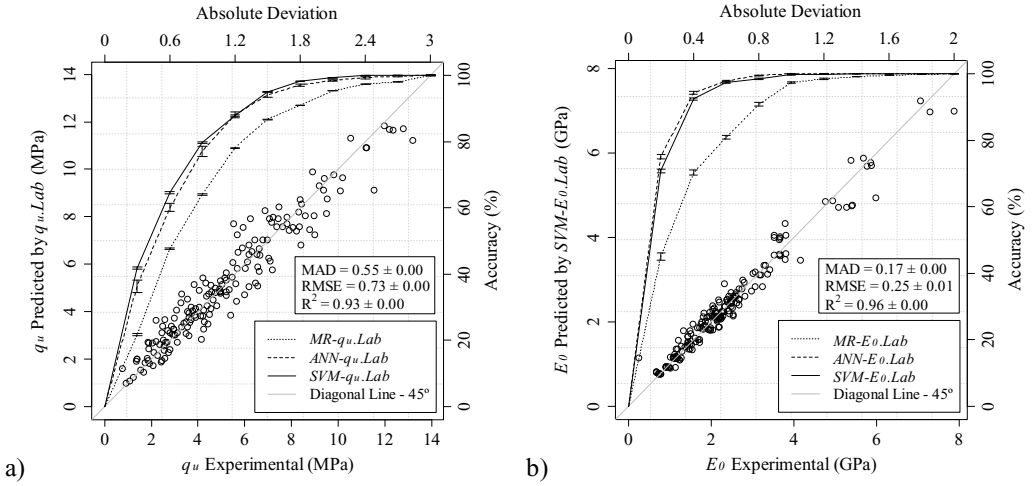

Figure 1: Data mining models performance in laboratory soil cement mixtures - mechanical properties prediction (Gomes Correia et al, 2014): a) q_u and b) E_0

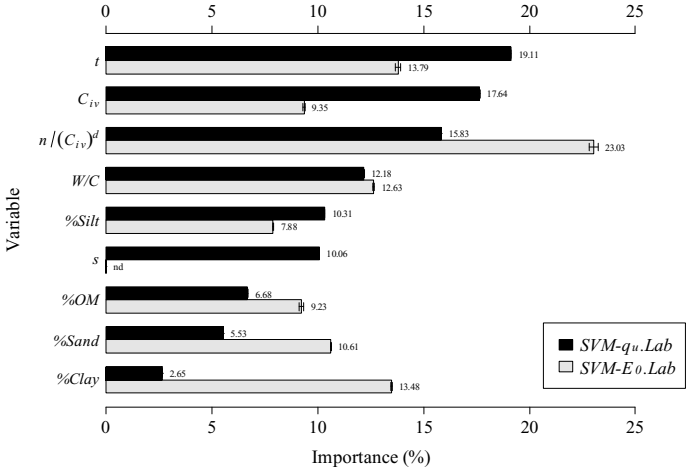

Figure 2: Comparison of the relative importance of each input variable in laboratory soil cement mechanical properties prediction according to SVM-q_u.Lab and SVM-E_0.Lab models (Tinoco et al., 2014).

Aiming to overcome such models dependence of the final mixtures properties, in this paper a new model (data-driven approach) is proposed for q_u prediction over time which is independent of final mixtures properties.

The proposed model, based on advanced statistics analysis, allows to estimate q_u over time based on 8 input variables such as the cement content, soil grain size distribution or type of cement. A cross-validation approach under 5 runs was applied for model generalization assessment.

2 Models and Data

2.1 Data Mining Algorithms

As previously underlined, the proposed models were developed using data mining techniques (Witten and Frank, 2005), namely support vector machines (SVM) (Smola and Scholkopf, 2004), artificial neural networks (ANN) (Kenig et al, 2001) and multiple regression (MR), which has been successful applied in many different scientific domains (Domingos, 2012; Goh and Goh, 2007; Kewley et al., 2000; Cortez et al., 2009) to solve real problems that traditional analysis are unable to deal with. These techniques, characterized by high learning capabilities, make use of computational tools to extract useful knowledge from raw data (Fayyad et al., 1996). For a baseline comparison, the classic Multiple Regression (MR) method was also tested.

ANN and SVM models were implemented using the rminer library of the R tool (Cortez, 2010). For ANN, it was adopted the multilayer perceptron that contains only feedforward connections, with one hidden layer with H hidden units with logistic functions: $1/(1+e^{-x})$. To find the best value for H, a grid search of $\{1, 2, \ldots, 10\}$ was used. For SVM, the methodology proposed by Huang et al. (2007) for model selection (i.e. to select the best values of the hyperparameters C, ε and γ) was applied.

2.2 Models Assessment

Models performance was evaluated based on the difference between experimental and predicted values for all N examples. In a model for which such difference is close to zero a high accuracy is expected. In particular, three different metrics were calculated (Tinoco et al., 2011): Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Correlation (R^2). Low values of MAE and RMSE, as well as an R^2 close to the unit value should be interpreted as high predictive capacity of the model. The Regression Error Characteristic (REC) curve (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, was also adopted during model performance analysis. For validation purposes, a 5 runs were applied under a 10-fold cross-validation schema.

Frequently, the absence of explanatory power of complex DM algorithms, such as SVM or ANN, is pointed out as their main drawback. In order to overcome this problem, Cortez and Embrechts (2013) proposed a novel visualization approach based on a sensitivity analysis (SA) method. SA is a simple method that is applied after the training phase and that measures the model responses when a given input is changed, allowing the measurement of the relative importance of each input variable, as well as its average effect on the target variable.

In the present work, 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, two important visualization techniques can be computed. The input importance barplot shows the relative influence (R_a) of each input variable in the model. To measure this effect, first the gradient metric (g_a) for all inputs was calculated. After that, the relative influence was computed.

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

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

To analyze the average impact of a given input in the fitted model, the Variable Effect Characteristic (VEC) curve can be used. For a given input variable, the VEC curve plots the attribute *L* level values (*x*-axis) versus the SA responses (*y*-axis).

2.3 Database

For models training, a database with 269 records was taken. These samples make part of a laboratory study aiming to define the binder mixture to obtain the best technical, economical and environmental soil stabilization (Correia 2011; Venda Oliveira et al., 2012; Venda Oliveira et al., 2013; Venda Oliveira et al., 2014; Correia et al., 2015). The soils selected were from Coimbra area (located near Coimbra city, Portugal), ranging from cohesive to cohesionless soils, organic to non-organic soils, presenting different geotechnical properties. Fourteen different binders were tested, including Portland cement, slag, fly ash, lime and silica fume, applied individually or combined.

As models input a set of 8 variables were selected. The definition of such variables took into account the empirical knowledge related to soil cement mixtures behavior, particularly concerning to the q_u evolution with time (Sariosseiri and Muhunthan, 2009; Lee et al., 2005; Lorenzo and Bergado, 2004; Chen and Wang, 2006). Also the feedback obtained from the learning process was used in the input variables selection. Bellow are listed all the 8 input variables considered in the models for q_u prediction.

- Soil clay content (%) – *%Clay*
- Soil organic matter content (%) – *%OM*
- Relation between water and cement content – ω_0/a_w
- Cement dosage (kg/m^3) – kg/m^3
- Age of the mixture (days) – t
- Coefficient related with the cement type – s
- Percentage of cement (%) – C
- Coefficient related with the secondary binder – L_2

Table 1 summarizes the main statistics of all 8 model inputs as well as of the output variable.

Variable	Minimum	Maximum	Mean	Standard deviation
%Clay	0.00	25.00	9.42	4.67
%OM	0.00	19.40	7.98	4.73
ω_0/a_w	0.63	10.91	4.52	1.80
kg/m^3	57.30	500.00	158.74	69.07
t (days)	3.00	360.00	31.88	33.36
s	0.20	0.38	0.23	0.07
C	0.50	1.00	0.78	0.11
L_2	0.00	28.50	20.73	12.07
q_u (MPa)	0.10	3.77	1.14	0.88

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

3 Results

This section summarizes some of the most relevant results achieved during the q_u study through the application of advanced statistics analysis.

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

Model	Hyperparameter	time (s)
MR	-	0.31 ± 0.03
ANN	$H = 7 \pm 2$	27.87 ± 2.06
SVM	$\gamma = 0.42 \pm 0.06$; $C = 5.04 \pm 0.27$; $\varepsilon = -4.74 \pm 0.26$	18.71 ± 0.15

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

Table 3 shows and compare the performance of MR, ANN and SVM algorithms in q_u prediction, based on MAE, RMSE and R^2 metrics (mean values and respective 95% level confidence intervals according to a t-student distribution). From its analysis, we observe that SVM performs much better than MR, having achieved a R^2 around 0.93. ANN attained a very similar performance with an R^2 around 0.91. Also based on MAE and RMSE metrics, a good performance is confirmed from ANN and SVM (in both cases very low values where achieved).

Model	MAE	RMSE	R^2
MR	0.35 ± 0.00	0.56 ± 0.01	0.59 ± 0.01
ANN	0.16 ± 0.02	0.26 ± 0.03	0.91 ± 0.02
SVM	0.15 ± 0.01	0.23 ± 0.03	0.93 ± 0.01

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

REC curves depicted on Figure 3a confirms the high performance of ANN and SVM algorithms in q_u predictions and show that, for example, more than 95% of the records can be accurately predicted within an absolute deviation lower than 0.50 MPa, and that all ANN and SVM prediction have an absolute deviation lower than 1.00 MPa.

Although *ANN-q_u.Lab_new* and *SVM-q_u.Lab_new* models present a very high performance, we observed that q_u prediction accuracy can be improved through the calculation of the average of *ANN-q_u.Lab_new* and *SVM-q_u.Lab_new* prediction. With this trick, an R^2 higher than 0.95 is achieved as well as an RMSE very close to 0.19MPa.

Figure 3b shows the relation between observed values and the average of *ANN-q_u.Lab_new* and *SVM-q_u.Lab_new* models prediction, from which is observed a high prediction capacity of the models.

Based on a GSA, the average relative importance of each input variable was quantified. From Figure 4a, which plots the relative importance of each input variable according to *ANN-q_u.Lab_new* (mean value and the correspondent t-student 95% confidence interval), it is observed that cement dosage, relation between water and cement content and age of the mixture are the three most influent variables in q_u development, weighing around 45% in the model. With a weigh around 28% appears the soil influence (%Clay and %OM). These set of 5 variables incorporate the well known influence (from empirical studies) of the cement and water content, age of the mixtures as well as the soil properties.

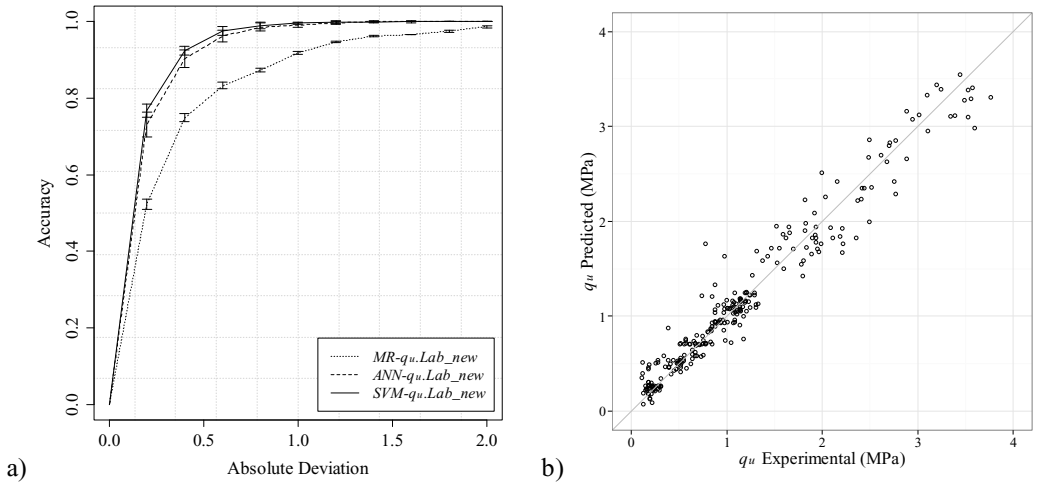

Figure 3: Models performance: a) REC curves of *MR-q_u.Lab_new*, *ANN-q_u.Lab_new* and *SVM-q_u.Lab_new* models; b) Scatterplot of the average of *ANN-q_u.Lab_new* and *SVM-q_u.Lab_new* predictions.

Figure 4b depicts the effect of the most relevant variable in q_u prediction according to *ANN-q_u.Lab_new* model, showing, as expected, an increase of q_u with the cement content (kg/m^3). The VEC curve of the cement dosage has two stretches with high growth rate separated by a stretch with low growth rate between 150 and 300 kg/m^3. This behavior already have been observed in experimental studies (Horpibulsuk 2001; Correia, 2011; Zhang et al 2013) and can be explained by the existence of a transitional zone where the stabilized soil starts to lose its identity (soil particles linked by cementitious products), gradually transforming into a cement mortar (a hardened paste with soil particles embedded in it).

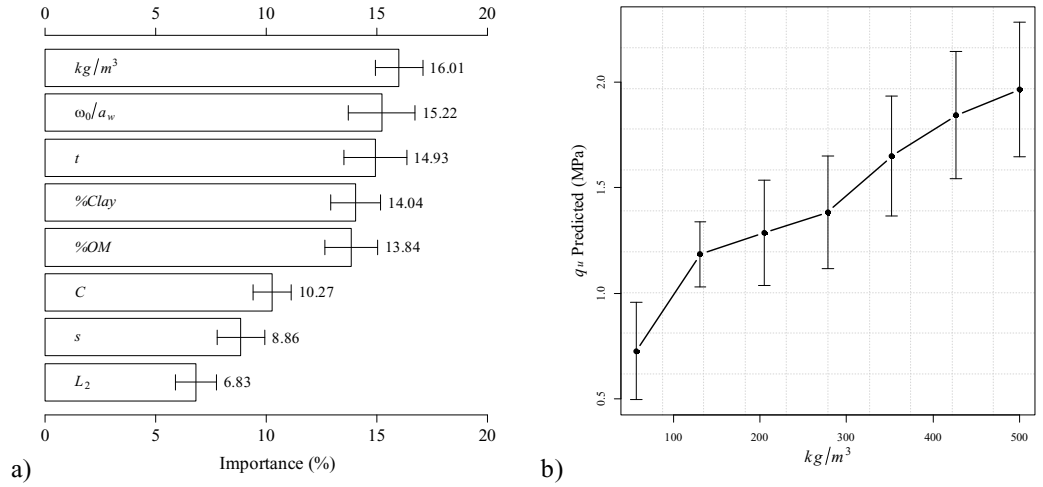

Figure 4: Global sensitivity analysis results over *ANN-q_u.Lab_new* model: a) Relative importance of each input variable; b) VEC curve of cement dosage (kg/m^3) variable.

4 Conclusions

In this work it was evidenced the high learning capabilities of data mining algorithms, particularly artificial neural networks and support vector machines and its great contribute to help in solving complex problems in geotechnical field when the traditional statistical tools are unable to deal with. Particularly, a data driven approach able to predict uniaxial compressive strength (q_u) of laboratory soil cement mixtures with high accuracy (R^2 higher than 0.95) was developed. Additional to its accuracy, the proposed approach has the advantage to be applied during the pre-design stage since only depends on parameters not requiring experimental measurements.

Supported on a detailed sensitivity analysis it were identified the key variables in q_u development, where the cement dosage, the relation between water and cement content and age of the mixture were ranked as the three most influent. The effect of cement dosage in q_u increase and its evolution with time was also observed and quantified, which is in agreement with the empirical knowledge related with soil cement mixtures.

5 Acknowledgements

The authors would like to express their thanks to CIMPOR, SIKA Portugal and CALCIDRATA for supplying the binders used in the work and to the institutions that supported the research financially: Universities of Minho and Coimbra, ISISE, CIEPQPF and ACIV.

References

A. Goh, S. Goh, Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data, Computers and Geotechnics 34 (2007) 410–421.

A Gomes Correia, J. Tinoco, and Paulo Cortez, Use of data mining in design of soil improvement by jet grouting, in Proceedings of Second International Conference on Information Technology in Geo-Engineering (ICITG 2014), 2014, pp. 43–63.

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

C.M. Huang, Y.J. Lee, D.K.J. Lin, and S.Y. Huang, Model selection for support vector machines via uniform design, Computational Statistics & Data Analysis, 52:1 (2007) 335–346.

A.A.S. Correia, Applicability of Deep Mixing Technique to the Soft Soil of Baixo Mondego, Ph.D. Thesis, University of Coimbra, (in Portuguese) (2011).

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

S. Horpibulsuk, Analysis and assessment of engineering behavior of cement stabilized clays. PhD Thesis, Saga University (2001).

F. Lee, Y. Lee, S. Chew, K. Yong, Strength and modulus of marine clay-cement mixes, Journal of Geotechnical and Geoenvironmental Engineering 131 (2005) 178–186.

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

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

H. Chen, Q. Wang, The behaviour of organic matter in the process of soft soil stabilization using cement, Bulletin of Engineering Geology and the Environment 65 (2006) 445–448.

I. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, second edition, Morgan Kaufmann, 2005.

J. Bi, K. Bennett, Regression error characteristic curves, in Proceedings of the twentieth international conference on machine learning, 2003, pp. 43–50.

J. Tinoco, A. Gomes Correia, and P. Cortez, Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time, Construction and Building Materials 25:3 (2011) 1257–1262.

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

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems 47 (2009) 547–553.

P. Cortez, Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool, in proceedings of Advances in Data Mining - Applications and Theoretical Aspects 10th Industrial Conference on Data Mining (ICDM 2010), Lecture Notes in Artificial Intelligence 6171, 2010, pp. 572-583.

P. Cortez, M. Embrechts ,Using sensitivity analysis and visualization techniques to open black box data mining models, Inform Sci, 225 (2013) 1–17.

P. Domingos, A few useful things to know about machine learning, Communications of the ACM 55 (2012) 78–87.

R. Kewley, M. Embrechts, C. Breneman, Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks, IEEE Transactions on Neural Networks 11 (2000) 668–679.

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

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM 39 (1996), 27–34.

P.J. Venda Oliveira, A.A.S. Correia, M.R. Garcia, Effect of stress level and binder composition on secondary compression of an artificially stabilized soil, Journal of Geotechnical and Geoenvironmental Engineering, 139:5 (2013) 810-820.

P.J. Venda Oliveira,; A.A.S. Correia, M.R. Garcia, Effect of organic matter content and curing conditions on creep behaviour of an artificially stabilized soil, Journal of Materials in Civil Engineering, 24:7 (2012) 868-875.

P.J. Venda Oliveira, A.A.S. Correia, T.J.S. Lopes, 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 (2014) 04014053.

R. Zhang, A. Santoso, T. Tan, K. Phoon, Strength of High Water-Content Marine Clay Stabilized by Low Amount of Cement, Journal of Geotechnical and Geoenvironmental Engineering, 139:12 (2013) 2170–2181.