Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time

Joaquim Tinoco a,⁎, A. Gomes Correia a, Paulo Cortez b

aDepartment of Civil engineering/Environment and Constructions Centre, School of Engineering, University of Minho, Guimarães, Portugal
bDepartment of Information Systems/Algoritmi Centre, School of Engineering, University of Minho, Guimarães, Portugal

ARTICLE INFO

Article history:
Received 30 March 2009
Received in revised form 30 July 2010
Accepted 2 September 2010
Available online xxxx

Keywords:
Ground improvement
Jet grouting
Data mining
Uniaxial compressive strength

ABSTRACT

Jet grouting (JG) is a soil treatment technique which is the best solution for several soil improvement problems. However, JG lacks design rules and quality controls. As a result, the main JG works are planned from empirical rules that are too conservative. The development of rational models to simulate the effects of the different parameters involved in the JG process is of primary importance in order to satisfy the binomial safety-economy that is required in any engineering project. In this paper, we present a new approach to predict the uniaxial compressive strength (UCS) of JG materials based on data mining techniques. This model was developed and verified using data from a JG laboratory formulation that involves the measurement of UCS. The results of the proposed approach are compared with the EC2 analytical model adapted to the JG material, and the advantages of the new approach are highlighted. We show that the novel data-driven model is able to learn (with high accuracy) the complex relationships between the UCS of JG material and its contributing factors.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction and background

Currently, there are several techniques for soil improvement, in which the jet grouting (JG) technology plays a key role [19,20]. This technology is growing rapidly, is characterized by great versatility in a variety of application soil types and treatment geometries (see [1]), and has been applied to several geotechnical engineering tasks [2,3]. The JG technology consists of injecting a high speed grouting of water–cement mixtures, with or without others fluids (air or water), into the subsoil. The fluids are injected through small-diameter nozzles placed on a rod, which is continually rotated and slowly removed toward the surface. Currently adopted JG methods can be classified according to the number of fluids injected into the subsoil: water–cement grout – single-fluid system; air + grout – double-fluid system; and water + air + grout – triple-fluid system.

At present, the design of JG works includes several stages. This paper will focus on the initial stage, where a set of laboratory formulations that include the function of the soil type to be treated and the desired properties are used to select the soil–cement mixture that will be used in the work. In particular, this study allows the definition of the grout water/cement ratio, the amount of cement for cubical meter of soil and the cement type, needed to satisfy the design and economical requirements. The remaining parameters that control the final characteristics of the JG elements, namely, the withdrawal time, the rotation rate, the fluid injection pressures, the effect of the number and the dimensions of the nozzles and the impact of the several fluids (grout, air and water) against the soil, will be evaluated by means of test columns because these parameters are impossible to simulate in laboratory. After the establishment of all parameters, the construction of the columns was initiated. After a period of time, some samples were extracted from these columns and were tested to evaluate the treatment and to determine what adjustments of the parameters were needed. This control procedure, together with the use of the predictions of analytical models of the mechanical behavior of the materials, aims at estimate the physical and mechanical properties of the treated material over time. However, the actual models used to predict the physical and mechanical characteristics of the jet material over time are based on empirical rules with low reliabilities [17]. Therefore, given the high potential of JG technology, there is a need to develop more rigorous and accurate design models. These design models will minimize the number of field tests and optimize the entire constructive process, thereby providing higher technical and economic efficiencies.

This study aims to build a model to predict the uniaxial compressive strength (UCS) of JG material. Determining UCS is a complex geotechnical engineering problem due to the heterogeneous nature of the soils and the large number of parameters involved. Thus, the use of artificial intelligence tools, in particular the application of data mining (DM) techniques [4], has great potential.
within this domain. DM is used to analyze raw (often complex) data and (semi-)automatically extract high-level information for the decision-maker. Despite the success of using DM in several areas (e.g., direct target marketing campaigns, fraud detection or finance predictions) [18], the application of these techniques to geotechnical engineering domain is rather scarce. Some examples of DM applications in the civil engineering domain include: the predicting of the settlement of shallow foundations by a new genetic programming model and the comparison of its results with those of a number of commonly-used traditional methods and artificial neural network-based models [5]; studying the relationship between the swell pressure and soil suction behavior in specimens of Bentonite–Kaolinite clay mixtures with varying soil properties using artificial neural networks [6]; and, predicting the compressive strength of cement conglomerates by means of neural networks [7]. To the best of our knowledge, DM has not yet been applied to JG material data, namely the prediction of UCS.

Fig. 1 gives an overview of the DM process, also known as knowledge discovery in databases (KDD). This process begins with the selection of the database (DB) on which the study will be made, and ends with the presentation of the knowledge discovered.

Regression is one of the main DM goals. This task consists of mapping several input (or independent variables) to a given numeric output (the dependent variable). There are several regression techniques, each one with its own advantages. Multiple regression (MR) is easy to interpret due to its additive nature. Neural networks (NN) and support vector machines (SVMs) are more sophisticated methods, requiring more computation but capable of modeling more complex, nonlinear relationships. Compared to MR, the NN and SVM models are more difficult to interpret. Yet, it is still possible to extract knowledge from NN and SVM models, given in terms of input variable importance [8].

2. Materials and methods

2.1. Jet grouting data

The models were trained and verified using data from huge JG laboratory formulations, prepared at University of Minho to analyze the influence of several parameters in UCS values. These values were obtained in unconfined compression tests with on sample strain instrumentation [9]. The dataset includes 175 records, derived from 35 JG laboratory formulations and 11 input parameters, which are referred as the more influent parameters in mechanics properties studies of soil–cement mixtures [10,11]:

- Water content (%) – w.
- Percentage of sand – %sand.
- Percentage of silt – %silt.
- Percentage of clay – %clay.
- Percentage of organic matter – %OM.

The main statistics on the numerical parameters used are presented in Table 1.

The soils used in the preparation of the JG laboratory formulations came from seven different work sites in Portugal and Spain. The geotechnical properties of the soil were evaluated using identification laboratory tests, and the soil classifications are presented in Table 2. All of the soils were classified as fine-grained soils, they had different percentages of sand, silt, clay and organic matter. The cement types used were CEM I 42.5R, CEM II 42.5R and CEM IV/A 32.5R. All formulations and tests were executed in the Civil Engineering Laboratory of the University of Minho. The box plot graph of the output variables (Fig. 2) shows two outliers and that 75% of the records had strengths lower than 7 MPa with a maximum value around 13 MPa.

2.2. Data mining models

In the present study, three DM techniques were used to model UCS prediction: multiple regression (MR), neural networks (NNs)

<table>
<thead>
<tr>
<th>Soil Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay W/C</td>
<td>0.68</td>
<td>1.12</td>
<td>0.88</td>
<td>0.16</td>
</tr>
<tr>
<td>ro (kg/m³)</td>
<td>1484.11</td>
<td>1916.16</td>
<td>1689.76</td>
<td>118.73</td>
</tr>
<tr>
<td>%sand</td>
<td>0.00</td>
<td>39.00</td>
<td>13.47</td>
<td>11.54</td>
</tr>
<tr>
<td>%silt</td>
<td>33.00</td>
<td>57.00</td>
<td>50.49</td>
<td>5.49</td>
</tr>
<tr>
<td>%clay</td>
<td>22.5</td>
<td>45</td>
<td>35.87</td>
<td>7.74</td>
</tr>
<tr>
<td>%OM</td>
<td>0.40</td>
<td>8.30</td>
<td>2.71</td>
<td>1.81</td>
</tr>
</tbody>
</table>

![Fig. 1. Steps of the DM process – adapted from [4].](image_url)
and support vector machines (SVMs). All of the experiments were conducted using the R Statistical Environment and the RMiner library [12], which facilitated the application of DM techniques in R.

The MR model was defined by Eq. (1) [13], as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

where \(x_i\) are input parameters and \(\beta_0, \beta_1, \ldots, \beta_n\) are coefficients to be adjusted, normally using the least square technique. Due to its additive nature, this model is easy to interpret and it is widely used in regression tasks.

The NN is a computational technique inspired by the structure of the nervous system in the human brain. This technique is capable of modeling complex nonlinear mappings and is robust in its exploration of data with noise. It is particularly useful for problems that do not have an analytical formulation or where explicit, accessible knowledge does not exist. In this work, we adopted the most common NN type, the multilayer perceptron, which uses feedforward connections and one hidden layer with \(H\) processing units. The general model is given by the following equation [13]:

$$\hat{y} = W_{o0} + \sum_{j=1}^{H} \left( \sum_{i=1}^{N} X_i \cdot W_{j1} + W_{j0} \right) \cdot W_{o1}$$

where \(W_{oj}\) represents the weight of the connection from neuron \(j\) to the unit \(i\) (if \(j = 0\), then it is a bias connection), \(o\) corresponds to an output unit, \(f\) is a logistic function, and \(J\) is the number of input neurons. The performance is sensitive to the choice of topology (\(H\)). A NN with \(H = 0\) is equivalent to the MR model. By increasing \(H\), more complex mappings can be performed, yet an excess value of \(H\) will overfit the data, leading to a loss of generality.

The SVM technique was initially proposed for classification tasks (i.e., to model discrete labeled output). After the introduction of the \(\varepsilon\)-insensitive loss function, it was possible to apply SVM to regression problems [14]. SVM has theoretical advantages over NN, such as the absence of local minima in the learning phase, i.e., the model always converges to the optimal solution. The main idea of the SVM is to transform the input data into a high-dimensional feature space by using a nonlinear mapping \(\Phi\), which is normally unknown. Then, the SVM finds the best hyperplane of linear separation within the feature space. This transformation depends on the kernel function adopted. The Gaussian kernel is the most popular one and will be adopted in this work:

$$K(x, x') = e^{-\gamma ||x-x'||^2}, \quad \gamma > 0$$

Under this setup, performance is affected by three parameters: \(\gamma\) - the parameter of the kernel, \(C\) - a penalty parameter, and \(\varepsilon\) - the width of the \(\varepsilon\)-insensitive zone. To reduce the search space, the first two values are set using the heuristics proposed in [15]. The kernel parameter \(\gamma\) has the highest impact on the SVM performance.

Several error metrics can be used to evaluate the performance of regression models, namely the mean absolute error (MAD), the root mean squared error (RMSE) and the coefficient of determination \(R^2\). These coefficients are defined as:

$$\text{MAD} = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where \(y_i\) denotes the desired value, \(\hat{y}_i\) the predicted value, \(\bar{y}\) and \(\bar{y}\) represent the mean of these variables. Lower values of MAD and RMSE correspond to a higher predictive capacity, while the \(R^2\) value should be close to unity. If it is acceptable to commit a few extreme errors, while maintaining the average of prediction with a good approximation, the MAD is the optimal measure. Otherwise, the RMSE measure should be used.

In this study, the NN and SVM hyperparameters (\(H\) and \(\gamma\)) were optimized using a grid search of \(H\) \([2, 4, \ldots, 10]\) and \(\gamma\) \([2^{-15}, 2^{-13}, \ldots, 2^1]\) [13]. To avoid overfitting, the training data was further divided into training (80%) and validation sets (20%). After selecting the best parameter (i.e., the model with the lowest error on the validation set), the model was retrained with all the training data. Then, the importance of the relative was measured, by applying the sensitivity analysis method proposed in [8].

After fitting a model, it is necessary to verify its future performance, i.e., to measure its generalization capability on unseen data. We adopted the Leave-One-Out method for this purpose. This method is especially suitable when the dataset is small (e.g., lower than 100 examples). Under Leave-One-Out, one example at a time is used to test the model, and the remaining data is used to fit the model. Under this scheme, all of the data are used for training and testing. Yet, this method requires approximately \(N\) (the number of data samples) times more computation, because \(N\) models must be fitted. The final generalization estimate is evaluated by computing the MAD, RMSE and \(R^2\) metrics for all \(N\) test samples.

3. Experiments and results

First, we present the adequacy of the EC2 normative model for JG laboratory formulations. Subsequently, we describe the predictions obtained by the model defined with the DM techniques. Finally, the results obtained by a combination of these two models are presented.

3.1. EC2 analytical model

Despite the fact that EC2 analytical model [16] is defined for concrete material, it was tested for JG laboratory formulation due to the existence of similarities these materials. According to this model, the evolution of the UCS over time is given by the following expression:

$$q(t) = e^{a(1 - (\frac{t}{70})^{b})}f_{cm}$$

where \(q\) is the strength at age \(t\) (MPa); \(f_{cm}\) is the 28 day strength of each studied formulation (MPa); \(a = 0.2\) to cement CEM I 42.5R and CEM II 42.5R and \(s = 0.25\) to cement CEM IV/A (V) 32.5R and \(t\) is the age of the formulation in days.
The quality of the predictions obtained by applying the EC2 model is presented in Fig. 3, which shows that the EC2 normative model is able to predict the UCS of JG laboratory formulation with good accuracy. This performance is confirmed by the MAD, RMSE and \( R^2 \) values (see Fig. 4). However, to use EC2, it is necessary to know the UCS of each formulation over the 28 day time-frame of the cure \( (f_{cm}) \). This implies waiting 28 days to carry out the experimental tests to measure the respectively formulation strengths. An inadequate value of \( f_{cm} \) leads to significant errors in the predictions. Thus, to solve this problem, it is possible to introduce other parameters that affect the 28 day strength. An alternative solution is to define a set of reference formulations with known strength at 28 days of curing time, and interpolate the value of \( f_{cm} \) for other formulations.

In this study, DM techniques were applied to define a predictive of UCS without carrying out laboratory formulations. With this approach, we can consider all parameters that affect the UCS of laboratory formulations.

### 3.2. Data mining techniques

Under the Leave-On-Out evaluation scheme, three DM techniques were trained using the dataset previously described in order to predict the respective strengths.

Among the three DM algorithms tested, the best results were obtained by the NN and SVM model, which had similar performances. SVM was adopted as the reference model because it had a lower level of variation among the several runs executed (over a total of 20 runs).

Analyzing Figs. 4 and 5 (which show the relationships between measured versus predicted values by the SVM model) we can conclude that, for JG laboratory formulations, the UCS control can be achieved using the parameters studied, i.e. W/C, TC, RC, kg/m³, \( t \), \( w \), %sand, %silt, %clay and %OM. However, for strength values above 6 MPa, the predictions were less accurate. This dispersion can be justified by the fact that the database included just a few records in this range of strengths (only 25% of the records), leading to weaker predictions.

The better performance of NN and SVM models over MR shows that the relationships between the variables are predominantly nonlinear. Still, the satisfactory performance obtained by the MR technique (see Fig. 4) allows us to state that it is possible to establish a linear relation between the parameters studied with reasonable accuracy. This aspect is particularly interesting because the MR results are more easily interpreted.

Fig. 6 shows the importance attributed by SVM to each input parameter that was obtained by applying a sensitivity analysis. Based on this figure, we can say that the most influential parameters are the age of the mixture \( (t) \), the water/cement ratio \( (W/C) \) and the amount of cement by cubic meter of soil \( (kg/m^3) \). The percentage of organic matter also had some impact on the UCS prediction. These results corroborate the empirical knowledge that considers the age of the mixture, the amount of cement by cubic meter and the percentage of organic matter of soil \( (%OM) \) as the most-relevant parameters.
Despite the high performance presented by the SVM, it should be stressed that this model is only valid for conditions found in the dataset used in this study (e.g., fine soil type).

When comparing SVM models and EC2, the former is better, as shown in Fig. 4. Unlike EC2, the SVM model can be used without carrying out laboratory tests to obtain the UCS of each formulation at 28 days. As a result, SVM facilitates mixture design at early stages of a project.

Despite its high predictive accuracy, the SVM model is complex, requiring eleven input parameters. Thus, in order to reduce its complexity, we tested a similar model (SVM-6) where the six most-relevant parameters (see Fig. 6) were considered including: the age of the sample (\(t\)), the water/cement ratio (W/C), the amount of cement by cubic meter of soil (kg/m\(^3\)), the percentage of organic matter (%OM), the type of cement (TC) and the water content of the soil (\(w\)). The prediction obtained by this new, simpler model is illustrated in Fig. 7, and its performance, as assessed by the MAD, RMSE and \(R^2\) measures, is presented in Fig. 4. The errors have increased, although the values are still more accurate than the ones given by the EC2. Thus, in this case, the performance of the DM model decreases when some parameters are discarded, i.e., there is a trade-off between simplicity and performance.

For the MR model, another exercise was carried on. First, the relationship between all doubly parameters was determined. Then, one of the doubly parameters was removed when the \(R^2\) value is equal or greater than 0.70 (see Table 3). As a result, the strength class of the cement (RC), the water content (\(w\)) and the percentage of clay (%clay) were removed. After running this new model (MR-8) without these three variables, we observed a poor performance when compared to the model trained with all variables (see Fig. 4). Thus, despite the strong correlation between some variables, all should be included in the model.

### 3.3. DM techniques and EC2 combined model

The SVM model displayed excellent performance, yet it is very difficult to fully understand its behavior (i.e., to make an analytical formulation). On the other hand, the EC2 analytical model is simple and widely used but depends on experimental tests to determine the 28-day strengths. Therefore, we propose a novel approach (SVM-EC2) that combines EC2 with SVM, allowing the use of the EC2 model without the need to carry out experimental tests.

The novelty of this model consists of first using the SVM model to determine the 28 days strength of each laboratory formulation. Then, the coefficient \(f_{cm}\) in the expression of the EC2 model (see Eq. (7)) is set to the predicted value. The strength estimated by this hybrid model is illustrated in Fig. 8.

The results obtained reveal a decrease in the provisional capacity, but one that can still be considered acceptable. In Fig. 4 we can compare the performance of the different individual and hybrid models.

### Table 3

\(R^2\) values between all doubly parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>WC</th>
<th>TC</th>
<th>RC</th>
<th>kg/m³</th>
<th>t</th>
<th>ro</th>
<th>w</th>
<th>%sand</th>
<th>%silt</th>
<th>%clay</th>
<th>%OM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>0.05</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td>0.10</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kg/m³</td>
<td>0.16</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ro</td>
<td>0.02</td>
<td>0.14</td>
<td>0.31</td>
<td>0.03</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>0.01</td>
<td>0.10</td>
<td>0.26</td>
<td>0.02</td>
<td>0.00</td>
<td>0.97</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%sand</td>
<td>0.02</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>0.00</td>
<td>0.46</td>
<td>0.38</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%silt</td>
<td>0.06</td>
<td>0.10</td>
<td>0.08</td>
<td>0.07</td>
<td>0.00</td>
<td>0.24</td>
<td>0.20</td>
<td>0.68</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>%clay</td>
<td>0.04</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.08</td>
<td>0.00</td>
<td>0.14</td>
<td>0.07</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Tinoco J et al. Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time. Constr Build Mater (2010), doi:10.1016/j.conbuildmat.2010.09.027
approaches, and it is shown that SVM achieved the best predictive performance.

4. Conclusions and future development

Jet grouting (JG) is one of the most popular ground improvement techniques. However, most JG works are planned from empirical rules that are often too conservative, frequently resulting in less-than-optimal technical and economic efficiencies. In this work, we propose several models to predict the UCS of JG laboratory formulations over time.

The known EC2 analytical model is simple and displays good predictive capacity, but requires knowing the 28 days strengths of each formulation. As a result, it is time consuming to apply this model in mixture design at an early stage of a project.

By applying DM techniques and, in particular, a SVM, we obtained a model with excellent predictive capacity of the UCS over time. Moreover, the SVM model can be applied without performing experimental tests to quantify the 28 days strength of each formulation. The proposed model also allows for the identification of the key (input) parameters that control the behavior of JG material including: the sample age (\(t\)), the amount of cement by cubical meter of soil (kg/m\(^3\)) and the water/cement ratio (W/C). The best SVM model included a total of eleven input parameters, and its performance decreased when only the six most relevant inputs were used; showing that the prediction of uniaxial compressive strength of JG material is a complex phenomenon. Furthermore, we also showed that doubly parameters with a strong interrelationship (\(R^2 \geq 0.70\)), both should be considered in the model.

Given the high quality results obtained by the application of the DM techniques, we intend to expand our field of application of the DM tools. In particular, we propose to develop models to predict the final diameter of the JG columns, as well as to quantify the Young modulus of JG materials.

Acknowledgments

The authors would like to thank Tecnasol-FGE, especially its Engineer manager, João Falcão, for the interest and financial support. We would also like to thank Tiago Valente for providing laboratory data used in the analysis. The work of the first author is also supported by a PhD scholarship Grant of FCT (SFRH/BD/45781/2008).

References


Fig. 8. Predicted versus measured strength of JG laboratory formulations using EC2 analytical model considering \(f_{cm}\) equal to 28 days strength estimated by SVM model.