Prediction of soil properties from PCPT data using model trees

P.U. Kurup
University of Massachusetts, Lowell, Massachusetts, USA
E.P. Griffin
Shaw Environmental & Infrastructure, Stoughton, Massachusetts, USA

ABSTRACT: The existing methods used to infer soil properties from piezocone penetration test (PCPT) data are not always reliable due, in part, to the complexity of cone penetration. In an attempt to overcome some of the limitations of the current PCPT interpretation methods, a data fusion technique was used to develop models for estimating soil properties from measurements of cone tip resistance and pore pressures. In this study, a feature-level data fusion technique, based on a model tree learning algorithm, was used to transform the features extracted from the raw piezocone sensor data into estimates of overconsolidation ratio (OCR), coefficient of lateral earth pressure at rest (K_o), and undrained shear strength (s_u). Overall, the values of OCR, K_o, and s_u predicted by the data fusion models were found to compare very well with the reference values and to be generally more reliable than the results of the current interpretation methods.

1 INTRODUCTION

Data fusion techniques combine data from multiple sensors or sources in order to achieve inferences that may not be feasible from data obtained using just a single sensor (Hall and Llinas 1997). This is because a combination of additional independent and/or redundant data tends to have a synergistic effect, resulting in improved inferences. The brain, which fuses data, including sight, sound, smell, taste, and touch, from multiple sensors (eyes, ears, nose, tongue, skin) and uses its memory, experience, and a priori knowledge to make inferences about the external world, is an excellent example of a data fusion system (Gros 1997). Data fusion is currently being used in numerous applications, including military defense, robotics, medical diagnosis, non-destructive evaluation of equipment, and weather forecasting.

In this study, it is proposed that the process of data fusion be used to estimate soil properties, including overconsolidation ratio (OCR), coefficient of lateral earth pressure at rest (K_o), and undrained shear strength (s_u), from in situ test measurements, and that data fusion algorithms, through training, may be able to overcome some of the limitations of the current piezocone penetration test (PCPT) interpretation methods. A model tree-based feature-level fusion algorithm was employed specifically for its ability to handle missing values. All data fusion model predictions of OCR, K_o,

and s_u were compared with each other, as well as with the estimates obtained using existing interpretation methods, to evaluate the performance of the data fusion algorithm.

2 DATA FUSION FOR INTERPRETING PCPT DATA

2.1 *Feature-Level Identity Fusion*

The most popular area of data fusion is feature-level identity fusion, which is the fusion of parametric data to determine the identity and/or attributes of an observed object. In the feature-level fusion approach, "feature extraction" is performed on the raw data (sensor measurements) to yield a feature vector from each sensor. The feature vectors consist of characteristics, or features, of the data that will aid in the identification of the object. The feature vectors from all sensors are then concatenated together into a single joint feature vector from which an identity declaration is made (Hall and Llinas 1997). Because feature-based pattern recognition techniques are often used for identity fusion, the success of these methods depends on the selection of good, representative features. Figure 1 depicts the feature-level identity approach for extracting features (q_t, u_1, u_2) from raw piezocone sensor data and using a data fusion algorithm to perform a nonlinear transformation between the input feature vector and the output declaration of identity, with the identity declaration being the soil attributes of OCR, K_o, and s_u.

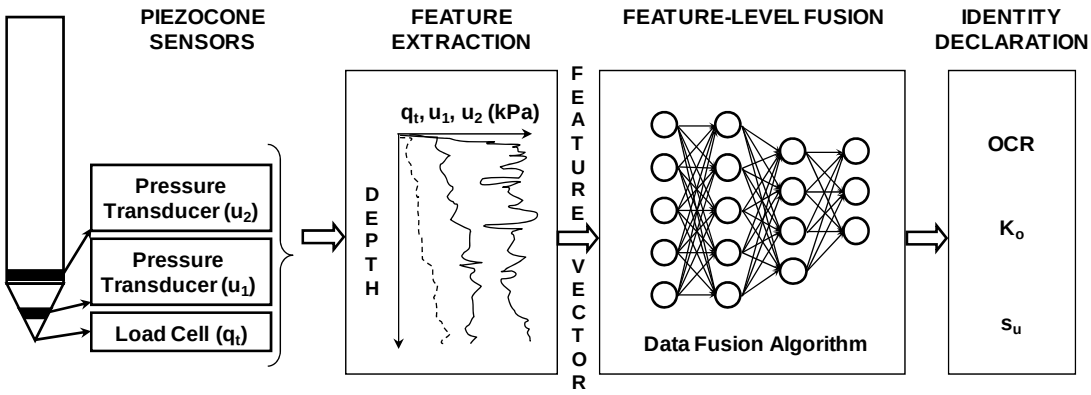


Figure 1 Feature-level data fusion system

2.2 *Tree-Based Data Fusion Techniques*

2.2.1 *Merits and Limitations*

There is no single, perfect data fusion algorithm that is optimal for all problems; each has its own strengths and weaknesses. Although artificial neural networks (ANNs) are widely used for numeric prediction, it is usually very difficult to relate the structure of the resulting model to the predicted output values. Model trees are alternative approaches to ANNs in performing feature-level data fusion for numeric prediction. Like neural networks, model trees are constructed so that they relate the target (output) values of the training cases to the corresponding values of their input variables. The final model may then be tested with a set of previously unseen input cases to de-

termine its ability to predict the unknown output values. Tree induction algorithms tend to learn quickly and the final model is easy to understand. They have the ability to handle numerous input variables and to choose only the most promising variables for inclusion in the final model. Model trees, unlike many neural network algorithms, also have the ability to handle missing data during both training and testing. Tree-based models, however, allow for only a single output, and several tree-based models may be replaced by one ANN model that can allow for multiple outputs (Dwinnell 1998).

2.2.2 *Tree Structure*

A model tree is a flowchart-like structure comprised of nodes and branches that shows how the value of a target variable can be predicted from the corresponding values of the input variables. Model trees are used for prediction of continuous (numerical) variables in estimation problems (Witten and Frank 2005). Each node in a tree represents a subset of cases from the original training set, with the topmost node, representing all the training cases, being called the "root" node. Nodes that have child nodes are called "interior" nodes; those that do not have child nodes are called "exterior" or "leaf" nodes. The linear regression equations of the input variables assigned to the leaf nodes in a model tree represents the model's final prediction.

2.2.3 *Tree Induction*

Model trees are constructed using the basic divide-and-conquer methodology where a complex problem is recursively split into two or simpler problems until the problems may be easily solved directly. Splitting is performed based on the value of one input variable known as the splitting variable. In the case of a binary split at the root node, the training set is partitioned into two subsets, based on the chosen splitting variable, which are as homogeneous as possible relative to the target variable. Branches are created for different values of the splitting variable, and each training case in the root node is sent down one of the branches, depending on its value for the splitting variable, into the corresponding child node. The process is then repeated for each of the child nodes, considering only the cases in that node. When a node cannot be partitioned, it is referred to as a leaf node; when only leaf nodes remain, the recursive partitioning used to construct the tree terminates.

For the training cases that reach a particular node, each input variable is evaluated to find the best split for that node. The training set is partitioned with the goal of obtaining subsets that are more "pure" than the original set, with "purity" referring to how homogenous the subsets are in relation to the values of the target variable. Ideally, each subset consists of training cases that have the same value for the target variable (i.e., completely pure nodes). The best candidate split is chosen based on the expected reduction in impurity resulting from the split, calculated by means of an impurity function, with different tree induction algorithms employing different impurity functions.

2.2.4 *Pruning and Missing Values*

Most tree induction algorithms build the complete tree and then prune it to prevent over fitting of the training data. In fact, many of the branches, which reflect noise or outliers in the training data, are pruned from the final model. The most common pruning operation is subtree replacement, where a subtree is either replaced with a single leaf or left unpruned (Witten and Frank 2005). Because a tree has been constructed

expressly for the training set, it is most desirable to use an independent pruning set to estimate the error rate at each node. This is typically accomplished by holding back a portion of the training data for use during pruning, resulting in less data being used to construct the tree (Witten and Frank 2005).

Datasets typically have missing values because either a variable was not measured or its values were not recorded during data collection. Tree induction algorithms must be amended to handle cases in which one or more of the splitting variables have missing values. This may be accomplished in several ways, depending on the algorithm. For example, the original variable may be replaced with another highly correlated variable, or the missing value may be replaced with the average value for the variable based on the other training cases reaching the node (Witten and Frank 2005).

Model trees may be induced by means of several algorithms, including the M5' algorithm (Witten and Frank 2005), which was used in this study. Details regarding the processes by which the M5' algorithm performs tree induction, accomplishes pruning, and handles missing values are presented in Griffin (2007).

3 PREDICTION OF SOIL PROPERTIES FROM PCPT DATA USING MODEL TREES

3.1 *Overview of Methodology*

In order to predict values of OCR, K_o, and s_u from PCPT data, a database consisting of values of corrected cone tip resistance (q_t), pore pressures measured on the cone face/tip (u_1) and/or just behind the cone base (u_2), vertical total stress (σ_v), and hydrostatic pore pressure (u_o), together with reference values of OCR, s_u, and K_o obtained from laboratory and field test results and empirical correlations, were used to train and test the model tree-based data fusion models (Griffin 2007). Data fusion model predictions were compared with the reference values, as well as with the estimates obtained using existing interpretation methods, to determine if the reliability of inferred soil properties can be improved by using data fusion techniques.

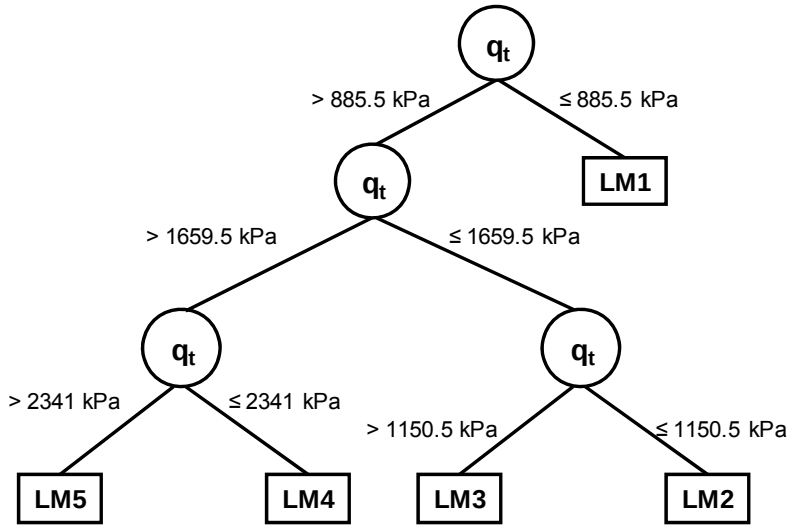
3.2 *Database*

The database, obtained from Sandven (1990) and Chen (1994), included data from 19 intact clay sites located in seven countries, including Norway, Canada, Sweden, the United States, Scotland, Singapore, and Taiwan. Although measurements of u_1 were not obtained from all of these sites, the data were used when available. The reference values of OCR were determined from one-dimensional consolidation tests, while the reference values of s_u were estimated from isotropically and anisotropically consolidated undrained triaxial compression (CIUC and CAUC) tests and in situ field vane shear tests. Since K_o was not measured in the laboratory or in situ, the reference values included for each case in the database were calculated using empirical correlations for normally consolidated and overconsolidated soils (e.g., Jaky 1944; Mayne and Kuhawy 1982). In total, the database contained 153 cases, ranging from soft, normally consolidated cohesive deposits to stiff, overconsolidated clays. OCR values in the database fell between 1.0 and 11.4; values of s_u ranged from 8 to 286 kPa for the laboratory tests (99 cases) and from 12 to 94 kPa for the field tests (54 cases); and computed K_o values ranged from 0.45 to 1.51.

3.3 *Training and Testing Data Fusion Models*

Model trees, induced using the M5' algorithm (Witten and Frank 2005), were used to develop a set of feature-level data fusion models. The PCPT data and in situ stresses (inputs), together with the corresponding values of OCR, K_o, and s_u (targets), were used to develop three tree-based data fusion models, or one data fusion model for each target variable. The feature vector for the models used to predict OCR and K_o consisted of four input variables, including vertical effective stress (σ'_v), q_t, excess pore pressure at the u_1 location (Δu_1), and excess pore pressure at the u_2 location (Δu_2). The feature vector for the models used to predict s_u consisted of five input variables, including σ'_v, q_t, Δu_1, Δu_2, and s_u type. The "s_u type" input was used in order to predict values of both $s_{u(TC)}$ and $s_{u(FV)}$, with an "s_u type" of 1 denoting a triaxial compression (TC) test and an "s_u type" of 2 denoted a field vane shear (FV) test.

Split-sample, or holdout, validation, in which a representative portion of the cases is reserved for testing, was used to estimate generalization error in the data fusion models. The testing set was comprised of 51 cases, representing approximately one third of the available data, while the training set was comprised of the remaining 102 cases. In accordance with standard testing procedures, the testing cases were chosen randomly from each piezocone site and were not used in any way during training. The trained data fusion model for prediction of s_u is shown in Figure 2. Note that interior nodes are represented by circles, leaf nodes are represented by squares, and the values on the branches are the values of the splitting variable in the corresponding parent node. A linear regression equation, or linear model (LM), is assigned to the leaf nodes in the model tree.



LM1: $s_u = 0.0472 * q_t - 6.9771 * s_u\_test - 0.0148 * \sigma_v' + 0.0193 * \Delta u_1 + 0.0109 * \Delta u_2 + 5.8954$

LM2: $s_u = 0.0616 * q_t - 3.0674 * s_u\_test - 0.0494 * \sigma_v' - 0.0072 * \Delta u_1 + 0.0203 * \Delta u_2 + 7.6334$

LM3: $s_u = 0.0610 * q_t - 3.0674 * s_u\_test - 0.0277 * \sigma_v' - 0.0072 * \Delta u_1 + 0.0203 * \Delta u_2 + 10.8457$

LM4: $s_u = 0.0929 * q_t - 3.0674 * s_u\_test - 0.0277 * \sigma_v' - 0.0072 * \Delta u_1 + 0.0203 * \Delta u_2 - 37.7897$

LM5: $s_u = 0.0996 * q_t - 3.0674 * s_u\_test - 0.0277 * \sigma_v' - 0.0072 * \Delta u_1 + 0.0203 * \Delta u_2 - 46.071$

Figure 2 Model Tree 3 for prediction of undrained shear strength

3.4 *Data Fusion Model Results*

After testing, the model tree data fusion model predictions were compared to the reference OCR, K_o, and s_u values to determine the models' success. For each model, the predicted values are plotted against the corresponding reference values and the normalized predicted values are plotted for each testing case in Figure 3. Note that perfect predictions fall on the 1 to 1 (45°) correlation line and have normalized values of unity. Included on all numeric plots are the correlation coefficient (CC), mean absolute error (MAE), and relative absolute error (RAE) of the numeric predictions, which provide a measure of error for the predicted values. Included on all normalized plots are the mean and standard deviation (SD) of the normalized predictions, which give an indication of the central tendency and variability of the normalized values, respectively.

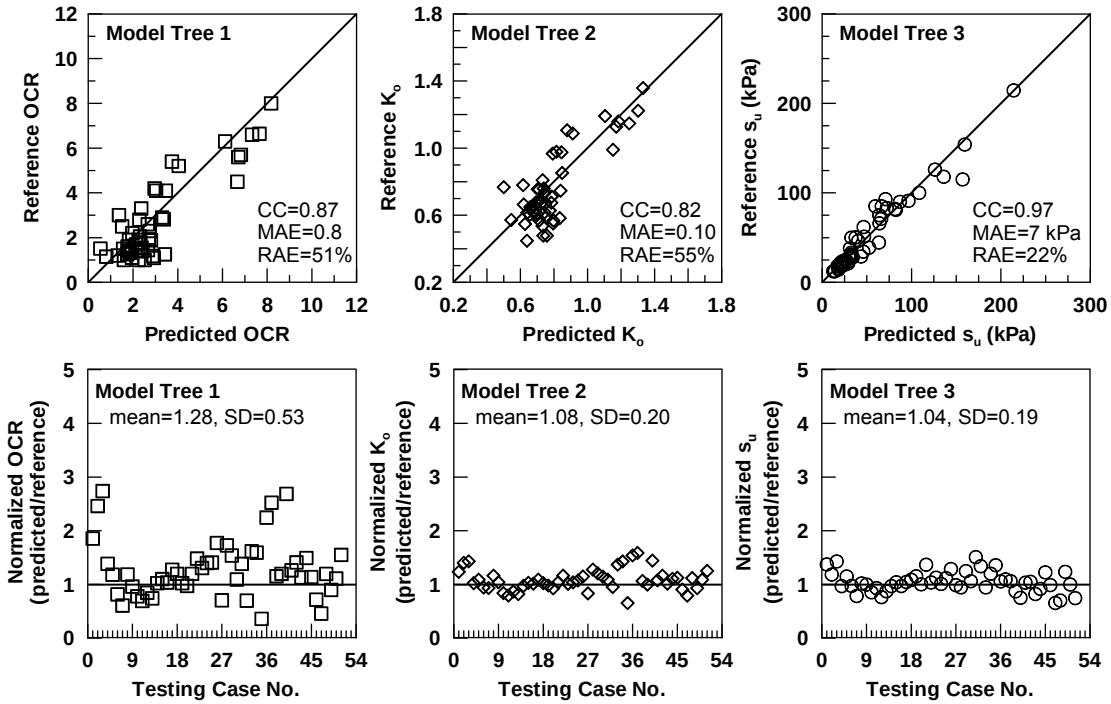


Figure 3 Data fusion model results

3.5 *Interpretation Method Results*

For comparative purposes, values of OCR, K_o, and s_u were each estimated using an existing interpretation method. The values of OCR were estimated using the normalized effective cone resistance correlation proposed by Chen and Mayne (1994) [$OCR=0.46(q_t-u_2)/\sigma'_v$]; values of K_o were estimated using the normalized net cone resistance correlation proposed by Kulhawy and Mayne (1990) [$K_o=0.10(q_t-\sigma_v)/\sigma'_v$]; and values of s_u were estimated using the net cone resistance empirical correlation [$s_u=(q_t-\sigma_v)/N_{kt}$]. In using the net cone resistance relationship to predict s_u, site-specific correlations for the empirical cone factor, N_{kt}, were not developed for each piezocone site; instead, values of N_{kt} were varied until single values which worked well were found for each s_u test type ($s_{u(TC)}$ and $s_{u(FV)}$) in the training set. For this database, the value of cone factor N_{kt} was chosen as 11 for $s_{u(TC)}$ and 17 for $s_{u(FV)}$. The interpreta-

tion method results are plotted in Figure 4. The CC, MAE, and RAE of the numeric predictions are included on the numeric plots, and the mean and SD of the normalized predictions are included on the normalized plots.

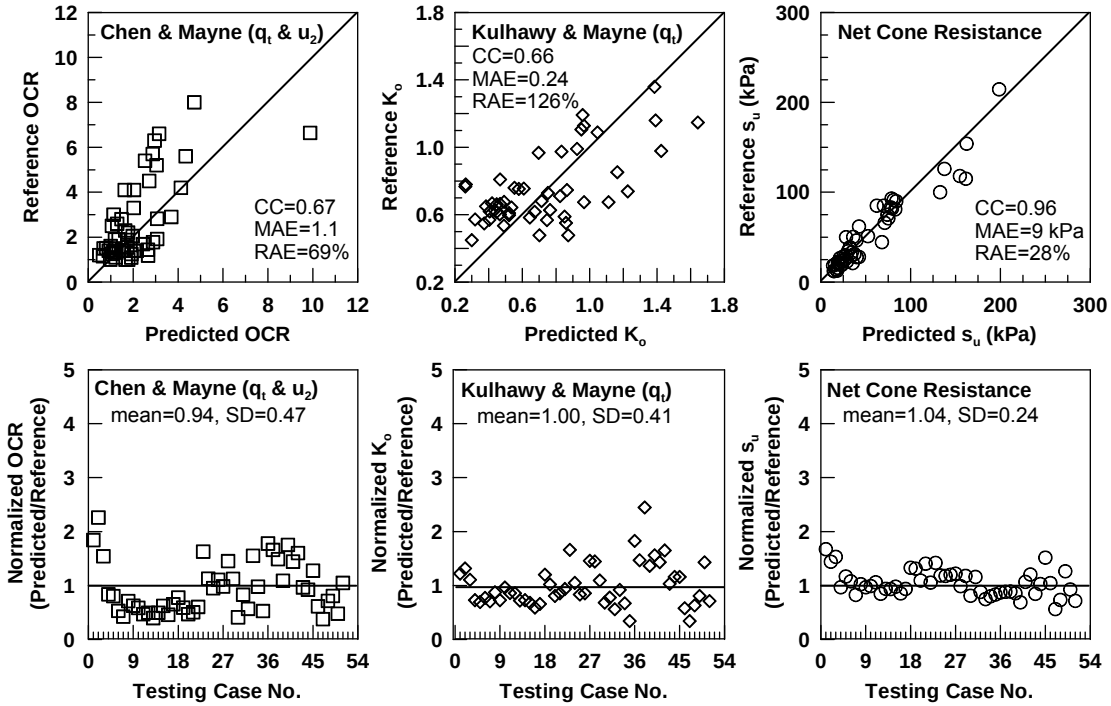


Figure 4 Interpretation method results

4 DISCUSSION OF RESULTS

4.1 *Data Fusion Model Performance*

The model tree feature-level data fusion technique performed well in predicting OCR, K_o, and s_u from the PCPT data and generally performed better than the PCPT interpretation methods. As shown in Figure 3, the data fusion models yielded OCR predictions having a MAE of 0.8, RAE of 51%, and CC of 0.87; K_o predictions having a MAE of 0.10, RAE of 55%, and CC of 0.82; and s_u predictions having a MAE of 7 kPa, RAE of 22%, and CC of 0.97. Compared to the data fusion models, Chen and Mayne's (1994) correlation for estimating OCR, Kulhawy and Mayne's (1990) correlation for estimating K_o, and the net cone resistance empirical correlation for estimating s_u, resulted in predictions having higher values of MAE and RAE and lower values of CC for the numeric predictions, as shown in Figure 4.

Noisy training data, typically caused by errors or anomalies in the laboratory and field test methods and inherent geologic variability, can lead to significant discrepancies between laboratory-measured and PCPT-predicted soil parameters. Laboratory tests are greatly affected by disturbances due to sampling and specimen transportation, storage, and preparation, as well as the type of test, test equipment, testing procedure, and interpretive technique used. In the field, piezocone measurements are greatly impacted by factors such as improper sensor calibration, inadequate filter sa-

turation, penetrometer inclination, temperature changes, damage or excessive wear of the cone, and cross-talk effects. And although the PCPT and sampling locations should be adjacent to each other, there may be discrepancies between the PCPT data recorded at a certain sounding depth and the geotechnical parameters of a sample collected from the same depth at a nearby borehole location.

Because the interpretation methods used herein do not account for such factors as soil fabric, sensitivity, mineralogy, aging, and geologic origin, it is hoped that the data fusion models may be able to "learn" some of these complex nonlinear relationships among sample data through training. However, the success of the data fusion algorithms, measured by their ability to generalize, is highly dependent on the amount, quality, and variability of the data used to train the models.

5 CONCLUSIONS

This study has demonstrated the effectiveness of data fusion and the model tree data fusion algorithm in inferring soil properties from PCPT measurements. The values of OCR, K_o, and s_u predicted by the data fusion models were found to compare well with the reference values and to be generally more reliable than the results of the current interpretation methods. Thus, data fusion may represent an improvement over the methods currently being employed to interpret piezocone penetrometer sensor data.

6 ACKNOWLEDGEMENT

The authors appreciate the financial support of the U.S. National Science Foundation under Grant No. CMS-0409594. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors.

7 REFERENCES

Chen, S. 1994. Profiling stress history of clays using piezocones with dual pore pressure measurements. Ph.D. thesis, Georgia Institute of Technology.

Chen, B.S. & Mayne, P.W. 1994. Profiling the overconsolidation ratio of clays by piezocone tests. Report No. GIT-CEEGEO-94-1, Atlanta, GA: Georgia Institute of Technology.

Dwinnell, W. 1998. Modeling methodology 3: Algorithm selection. *PC AI*, 12(2): 32-38.

Griffin, E.P. 2007. Interpretation of P/CPT data using data fusion techniques. MS thesis, University of Massachusetts Lowell.

Gros, X.E. 1997. *NDT data fusion*. New York, NY: John Wiley & Sons, Inc.

Hall, D.L. & Llinas, J. 1997. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1): 6-23.

Jaky, J. 1944. The coefficient of earth pressure at rest. *Journal of the Society of Hungarian Architects and Engineers* 78(22): 355-358.

Kulhawy, F.H. & Mayne, P.W. 1990. Manual on estimating soil properties for foundation design. Report No. EL-6800. Palo Alto, CA: Electric Power Research Institute.

Mayne, P.W. & Kulhawy, F.H. 1982. K_o-OCR relationships in soil. *Journal of the Geotechnical Engineering Division* 108(GT6): 851-872.

Sandven, R. 1990. Strength and deformation properties of fine grained soils obtained from piezocone tests. Ph.D. thesis, Norwegian Institute of Technology.

Witten, I.H. & Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.