MARITIME TRANSPORTATION RESEARCH AND EDUCATION CENTER
TIER 1 UNIVERSITY TRANSPORTATION CENTER
U.S. DEPARTMENT OF TRANSPORTATION

MarTREC
Maritime Transportation Research & Education Center

Predicting Soil Type from Non-destructive Geophysical Data using Bayesian Statistical Methods

Project start and end dates: January 1, 2018 – August 15, 2018

Michelle Bernhardt-Barry, Ph.D., P.E.
University of Arkansas
4190 Bell Engineering Center
Fayetteville, AR 72701
479-575-6027

Jyotishka Datta, Ph.D.
Clinton Wood, Ph.D., P.E.
Josh Price

September 14, 2018

FINAL RESEARCH REPORT
Prepared for:
Maritime Transportation Research and Education Center

University of Arkansas
4190 Bell Engineering Center
Fayetteville, AR 72701
479-575-6021

ACKNOWLEDGEMENT

This material is based upon work supported by the U.S. Department of Transportation under Grant Award Number DTRT13-G-UTC50. The work was conducted through the Maritime Transportation Research and Education Center at the University of Arkansas.

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Project Description

Field-based electrical resistivity methods are gaining popularity among geotechnical engineers as an efficient and non-destructive method to collect continuous subsurface data. However, predicting soil-type based on such measurements remains a challenging problem. A more accurate and interpretable predictor of soil type is critically needed in order to assess the many miles of undocumented levees scattered across the United States. The methods assessed herein would allow this information to be gathered non-destructively, saving both time and money.

In a previous MarTREC project, a series of geophysical field trials were conducted to determine the most accurate and efficient methods and the best parameters for detecting various features or defects within levees. Of the available techniques, electrical resistivity measurements and surface wave methods were determined to be the most advantageous in terms of capturing features of interest. While these are the best indicators of a subsurface condition, neither method was able to provide a confident prediction of soil type when used alone. For resistivity in particular, a wide range of predictor values were associated to a given soil type depending on the moisture and density conditions, leading to poor uncertainty quantification. A laboratory study was also conducted to better understand the influence that geotechnical parameters have on a soil's measured electrical resistivity; however, a robust statistical analysis of the data was not carried out.

The goal of this project was to predict soil-type from field geophysical surveys based on the least amount of predictors possible. Ideally, the predictive models would be able to identify soil type based on resistivity or shear wave characteristics alone because these two predictors can be measured relatively rapidly and most importantly, they can be obtained non-destructively. Statistical analysis of the laboratory benchmark data was carried out and three classification procedures were compared on a wide range of soil types. The ability of each method to predict soil type for a given number of predictors was assessed. The resulting accuracy was quantified based on the number of predictors used or provided in the data-set. The classification accuracy of the methods was also assessed using a supervised learning scheme to avoid possible overfitting. The best performing procedure was then applied to a field data-set and the performance along with the predictive power of the variables was assessed.

2. Methodological Approach

2.1 Background

At the Mel Price portion of the Wood River Levee System, a large amount of geotechnical data was provided by the U.S. Army Corps of Engineers (USACE) and consisted of soil type with depth, water content, behavioral classification, and standard penetration test (SPT) blow counts. While this data was used as the "ground truthing" in a previous study, it was used as verification data for the training and test data sets for this study. Measures of electrical resistivity (ER) and shear wave velocity (V_s) at a number of locations across the site were made and were combined with the known soil type and soil conditions to create a data-set which could be used to train and test the developed statistical models.

ER is an intrinsic property defined as a measure of how strongly a given material opposes the flow of electrical current. ER measurements require at least four-electrodes, two current electrodes and two voltage potential electrodes. Multiple measurements are taken at different electrode spacings in field resistivity surveys and a pseudosection (i.e., map of apparent resistivity measurements) is obtained. The measured resistivity represents the resistivity that would have been measured for a uniform subsurface (Everett 2013), and inversion and forward modeling processes are required to obtain the true resistivity distribution for the soil profile (Loke 1997). ER depends on many factors such as the nature and arrangement of solids, shape and size of solids, thickness of diffuse double layer (DDL), ion concentration in pore water and DDL, cation exchange capacity (CEC), water content, pore fluid composition, temperature and even anisotropy due to particle alignment (Parkhomenko 1967; Abu-Hassanein, Benson, and Blotz 1996; Samouelian et al. 2005, Fukue et al. 1999, Friedman 2005). These factors may interact making it difficult or even impossible to isolate their effects in data interpretation.

Seismic geophysical methods (known as stress wave methods) use body or surface stress waves to estimate the subsurface layering and stiffness of earth materials. Stress wave methods result in either compression wave (P-wave, V_p) or shear wave (S-wave, Vs) velocity with depth along a line, providing a 2D profile of V_p or V_s. V_p and V_s are fundamental properties of soil and rock and relate directly to the stiffness of the material for a given density. V_p can vary greatly depending on the saturation of a soil while V_s is not strong influenced by the degree of saturation. Therefore, V_s, was chosen as the most appropriate parameter in this study. V_s depends on the type, density, and stiffness of a given soil and it can be correlated to strength values such as undrained shear strength, SPT blow count, friction angle, or other geotechnical properties and parameters.

Because of the large number of predictors that affect the resistivity value, a laboratory study was carried out which considered the effects of various geotechnical parameters on different benchmark soil types. The details of the full study can be found in Mofarraj Kouchaki et al. (2018); however, some discussion is given here to provide the necessary background for the statistical analysis.

Nine different benchmark soils (Table 1) were made by mixing different portions of commercially available sand, Kaolin clay, Bentonite clay, and red art clay. A summary of the properties of these soils, as well as the range of densities and water contents tested are provided in Table 1. The effects of water quality, water content, degree of saturation, density, and temperature on the measured electrical resistivity of the soils were investigated. Each soil was tested in a laboratory resistivity device and the results were plotted to determine trends in the various properties and parameters. Bulk density and degree of saturation were found to be most effective in the identification of soil type. Results indicated that resistivity values reach a lower threshold at around 60% saturation and density and other parameters become less influential as the saturation increases above this threshold. Temperature was found to greatly influence the resistivity measurements and should be monitored and corrected for when laboratory test results are compared to field data.

Table 1. Material description, index properties, and density and moisture conditions for the soils tested.

	Composition (% mass)				Index Properties							Testing Range	
Soil Type	Sand	Kaolin Clay	Bentonite	Red Art Clay	LL (%)	PL (%)	PI (%)	Gs	D_{90} (µm)	D_{50} (µm)	D_{10} (µm)	Bulk Density (Mg/m^3)	w (%)
SP	100	0	0	0	-	-	-	2.67	850	440	240	1.63-2.00	2-20
SP-SM	90	10	0	0	-	-	-	2.66	800	380	100	1.12-2.11	4-16
SP-SC	90	8.5	1.5	0	-	-	-	2.64	780	375	80	0.98-2.10	3-12
SC	70	25.5	4.5	0	28	15	13	2.70	730	320	-	1.09-2.15	4-18
SM	70	30	0	0	26	15	11	2.64	750	330	-	1.03-2.11	10-15
CL-1	0	0	0	100	38	19	19	2.77	25.4	7.8	0.49	1.13-2.10	2-39
CH	0	85	15	0	72	33	39	2.62	6.6	0.4	-	1.15-1.82	10-60
CL-2	30	70	0	0	48	24	24	2.60	550	1.8	-	1.06-1.94	14-30
MH	0	100	0	0	62	32	30	2.61	5.9	0.4	-	1.06-1.60	6-70

LL = liquid limit

PL = plastic limit

PI = plasticity index

G_s = specific gravity

$D_{\#}$ = diameter of particle by which # % is finer

w = gravimetric water content in percent

2.2 Statistical Approach

Statistical methods were used to further analyze the parameters affecting ER in the laboratory study and determine the most appropriate predictors for determining soil type. In this section, we first compare the performances of three different classification procedures: (1) Linear Discriminant Analysis (LDA), (2) Logistic regression, and (3) Decision tree on the laboratory-based soil data-set, with more predictors, as well as more labels for the categorical response (a higher resolution of soil type classification). This is done by comparing the misclassification errors for the different classifiers in a supervised learning set-up.

Laboratory Resistivity Statistical Analysis

As discussed before, the goal was to predict soil-type based on resistivity characteristics, collected for the following predictors. Table 2 lists the predictors and their descriptions considered in the laboratory data study. Clearly, some of the predictors should exhibit a strong dependence, and this should be incorporated into the classification method built on these variables.

Soil type was the target response variable with nine different categories defined by the main group classifications defined in the Unified Soil Classification System (USCS). It is noted that a low plasticity silt (ML) was not used because a non-processed benchmark soil of this classification was not available. The number of samples for each category is given in Table 3. As discussed below through the analysis, some of these categories have overlapping class boundaries in terms of resistivity properties and as a result, there would be a higher chance of misclassification for them.

Table 2. Description of predictor variables for laboratory soil resistivity data analysis.

Predictor Variable	Description
Dry density	Density of soil WITHOUT water (Mass dry soil/total volume)
Wet density	Density of soil WITH water (Total mass/total volume)
Volumetric water content	The volume of water per a given total volume of soil
Saturation	Portion of void space filled with water (Volume of water/volume of void space)
Resistivity	Measure of how much the soil resists the flow of electrical current
Gravimetric water content	$100 \times \frac{M_W}{M_S}, M_w = mass\ of\ water\ in\ soil, M_s = dry\ mass\ of\ soil.$

Table 3. Soil types considered and number of samples for each. Soil types are labeled using their USCS group classification.

Soil Type	CH	CL-1	CL-2	MH	SC	SM	SP	SP-SC	SP-SM	**Grand Total**
Count	17	17	10	31	13	8	20	12	13	**141**

For this data-set, we applied three different classification tools:

1. **Linear Discriminant Analysis (Fisher's LDA).** Linear Discriminant Analysis is one of the earliest classification methods that calculates the conditional probabilities $P(Y = k \mid X)$ for each category k based on Bayes' rule (James et al., 2013). The LDA assumes linear classification boundaries between the different classes and makes Gaussianity assumptions on X.

2. **Logistic Regression Models:** The logistic regression models a sigmoidal function of the class probabilities as a linear model in Generalized Linear Model framework (Hosmer et al., 2013). The logistic regression / classification model makes weaker assumptions on the predictor variables and is a powerful and popular method because of the interpretability of coefficients through their P-values. This makes model selection feasible and principled for the logistic regression. For a multinomial classification problem, the model can be written as follows:

   $$log\left(\frac{P(Y = k \mid X_1, \dots, X_p)}{P(Y = k_0 | X_1, \dots, X_p)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon;\ \ k = 1, 2, \dots, K.$$

   Here $Y = k_0$ is the baseline category. For a multi-class classification problem, we need to fix a baseline and model the logit of the class-probabilities with respect to the baseline.

3. **Decision Tree:** Decision-tree based methods on the other hand, try to find the best partition of the predictor space to classify data (James et al., 2013). Decision-trees do not assume a

model for the class-probabilities like logistic / LDA which limits their interpretability, but they gain when the true classification boundaries are not linear.

The data-set was split into training data and test data for assessing the classification accuracy of the different methods using a supervised learning scheme. For each soil type, 25% of the samples were designated as test data points, leaving the remaining 75% as the training data. The three classification tools were then analyzed to determine the best performing in terms of prediction accuracy for soil type.

Field Measurement Statistical Analysis

Once the best performing procedure was determined for the laboratory data-set, it was applied to the field data collected at the Mel Price reach of the Wood River Levee System. Only the 'landside' (i.e. dry side) of the levee system was considered here. The field data had fewer soil classification categories compared to the laboratory data (only 'CL' and 'SP') with four continuous predictors for resistivity. These predictors are recorded as follows:

1. R: Electrical Resistivity Value; 'raw' is the raw value measured and 'surfer' is the actual resistivity value determined using the surfer software. Both represent a 10 m average in the horizontal direction (x direction). We denote these columns by R_{ohm} and R_{surf}, respectively.
2. Velocity: Shear Wave Velocity, V_S.
3. SPT: Standard Penetration Test (SPT) blow count (N value).

We applied the logistic regression for classifying the soil types in the field study, as it showed the best classification performance under a more challenging classification task with more soil type categories for the laboratory study. The model is given by:

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \times R_{\text{ohm}} + \beta_2 \times R_{\text{surf}} + \beta_3 \times \text{Velocity} + \beta_4 \times \text{SPT} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

As before, we used a supervised learning approach for validating our model, as well as for preventing overfitting. For testing the predictive accuracy of the model, we divided the entire data into two groups with 80% training and 20% test data: which leaves 196 training data points and 50 test data points. We then fit the model on the training data and predicted soil type for the test data set. Using a single split of training and test data to fit and evaluate a model is called the "validation set" approach for supervised learning. To make sure that the classification accuracy is not an artifact of a particular split of the training and test data, we calculated the misclassification accuracy using the validation set approach on a few simulated data sets and reported the distribution. For this task, we had more sample points per category of soil compared to the laboratory resistivity study which helps in a better estimation and prediction by increasing the statistical power.

3. Results/Findings

3.1 Descriptive Statistics

Figure 1 presents the scatterplot matrix for the resistivity data to show their marginal distributions as well as pairwise dependence. Clearly some of the predictors have strong dependence and therefore, a variable selection step was performed to avoid multicollinearity issues. The extent of the pairwise dependence is shown in Figure 2. Some of the predictors have correlation almost near 1, potentially leading to poor model fit.

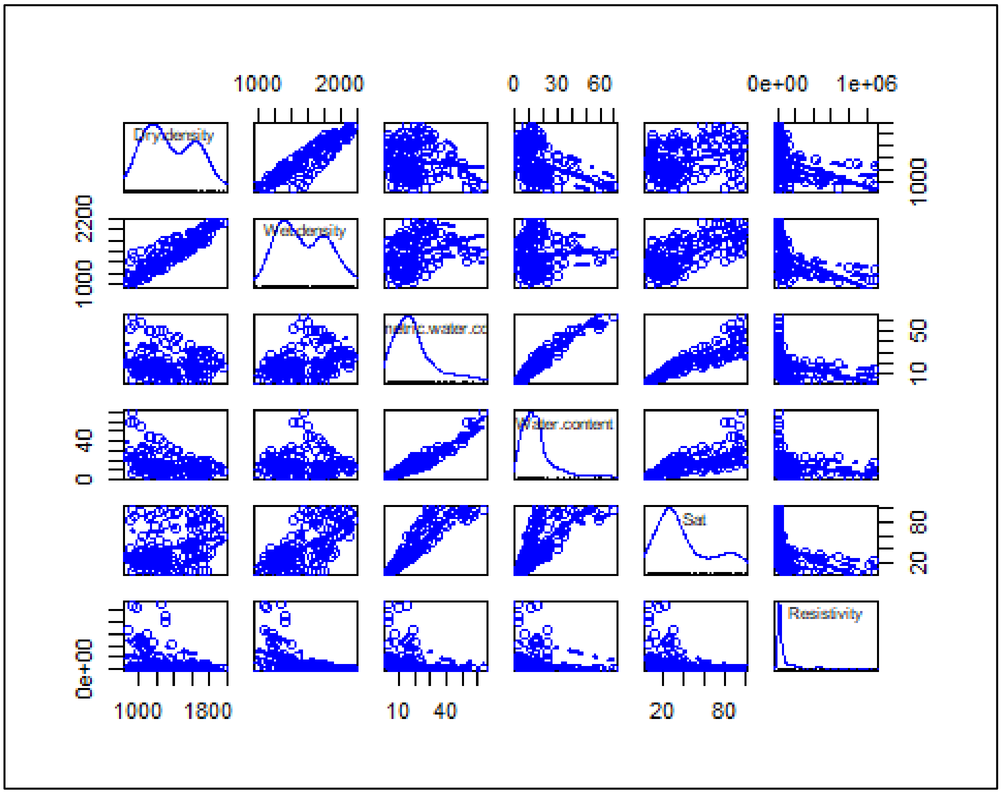

Figure 1. Scatterplot matrix for soil resistivity data.

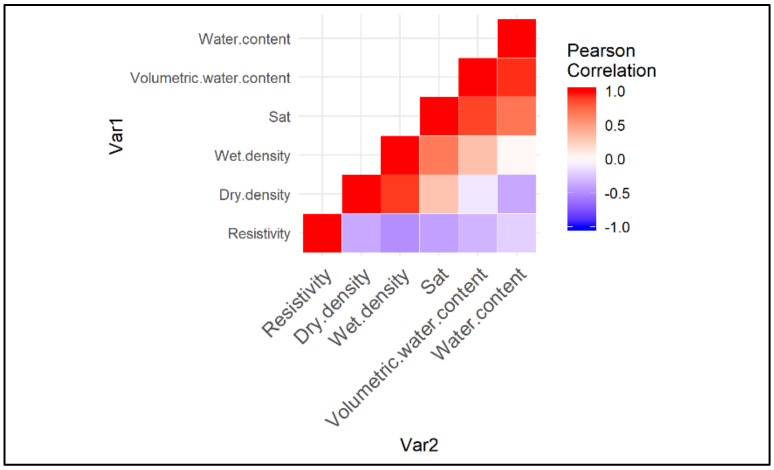

Figure 1. Pairwise dependence for soil resistivity data.

3.2 Supervised Learning: Classification for Laboratory Resistivity Data

The performance of each of the methods applied here is summarized in the next subsections, by reporting their overall classification errors, as well as the classification errors for each soil subtypes, via a “confusion matrix”, where one tabulates the ‘true’ soil-type versus the ‘predicted’ soil types. It should also be noted that these misclassification rates can be compared with a benchmark that randomly assigns each soil sample one of the 9 labels, and results in a true classification rate of only 11.11% (prob = 1/9).

Linear Discriminant Analysis (LDA)

The performance for two different sets of predictors is presented: one with resistivity, dry density, and saturation and the other with resistivity as the only predictor. This shows the relative loss of accuracy by using one predictor and helps the investigator decide which variables to include in a study. As discussed previously, resistivity can be determined non-destructively and it would be advantageous to quantify the accuracy obtained using only this predictor. V_S was not considered in the laboratory study because benchmark specimens were not representative of the field values.

The classification accuracy for the classifier that uses three predictors (resistivity, dry density and saturation) is 0.5151515 and for the classifier that uses only resistivity is 0.3030303. The confusion table for the 3-predictor-classifier is presented in Table 4. The column labels are the true soil type and the row labels are the predicted soil type. The diagonal entries are the number of correctly classified soil samples belonging to each soil type.

Table 4. LDA Confusion matrix with three predictors.

	TRUE LABELS								
PREDICTED LABELS	**CH**	**CL-1**	**CL-2**	**MH**	**SC**	**SM**	**SP**	**SP-SC**	**SP-SM**
CH	**3**	0	0	1	0	0	0	1	0
CL-1	0	**0**	0	0	0	0	0	0	0
CL-2	0	0	**0**	0	0	0	0	0	0
MH	0	2	2	**6**	0	0	0	0	0
SC	1	0	0	0	**2**	1	0	0	1
SM	0	0	0	0	0	**0**	0	0	0
SP	0	1	0	0	1	1	**5**	1	2
SP-SC	0	1	0	0	0	0	0	**1**	0
SP-SM	0	0	0	0	0	0	0	0	**0**

As this table suggests, the LDA does misclassify some of the test samples but it tends to misclassify samples into closely related categories such as SP-SM being misclassified as SP. This suggests that a better classification accuracy by LDA could be achieved if we merge similar categories such as SP, SP-SM and SP-SC to be a single soil type for the purpose of classification. For geotechnical considerations, there is minimal behavioral differences in these group classifications and little harm would likely arise from a misclassification in these categories.

Multi-class Logistic Regression

The **multinom** function from the R package **nnet** (Ripley et al., 2016) was used to fit a multinomial logistic regression model (Hosmer et al., 2013) to the soil data. The misclassification error rate for the multi-class logistic regression with all predictors is 0.6363 and the confusion matrix is given in Table 5. By comparing the diagonal elements on Table 4 and 5, it is clear that the logistic regression improves over the LDA in some categories, e.g., logistic regression correctly classifies all 7 MH samples, whereas LDA misclassified one of them as CH.

Table 5. Confusion matrix for logistic regression.

	TRUE LABELS								
PREDICTED LABELS	CH	CL-1	CL-2	MH	SC	SM	SP	SP-SC	SP-SM
CH	**3**	0	0	0	1	0	0	0	0
CL-1	0	**2**	0	0	0	1	0	0	0
CL-2	0	0	**0**	0	0	0	0	0	0
MH	0	0	2	**7**	0	0	0	0	0
SC	1	0	0	0	**1**	0	0	0	0
SM	0	0	0	0	0	**0**	0	0	0
SP	0	0	0	0	0	0	**5**	0	1
SP-SC	0	1	0	0	0	0	0	**2**	1
SP-SM	0	1	0	0	1	1	0	1	1

Decision Trees

Tree-based methods for classification were also examined. Tree-based methods involve stratifying or segmenting the predictor space into a number of simple regions, and in order to make a prediction for a given observation, the mean or the mode of the training observations is typically used in the region to which it belongs (James et al., 2013). The set of splitting rules used to segment the predictor space can be summarized in a tree, called the "decision tree". Figure 3 shows the decision tree for the soil type classification. The classification error for this decision tree is 0.54545.

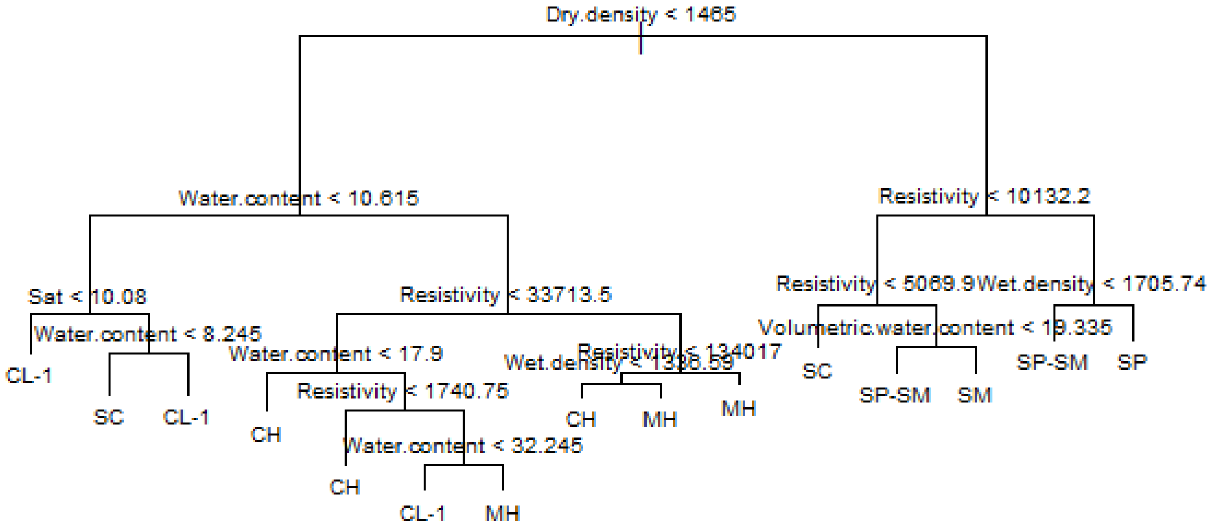

Figure 3. Decision tree for soil prediction.

Although, tree-based methods are simple and useful for interpretation, they are not competitive with the best supervised learning approaches in terms of prediction accuracy. One way to improve a decision tree is to use an Ensemble method, like Random Forest (Breiman, 2001), that involves producing multiple trees which are then combined to yield a single consensus prediction. Combining a large number of trees can often result in improvements in prediction accuracy; however, it is at the expense of some loss in interpretation. To correct for this, one can calculate variable importance for a predictor used in building a random forest. Roughly speaking, the average decrease in residual sum of square can be calculated every time a given predictor is used in a tree to measure its importance in predicting the label of an observation. We used a random forest classifier on the soil data with 1,000 trees; however, the resulting classifier had the same accuracy as a single decision tree. One advantage was that it provided us with the variable importance plot (see Figure 4).

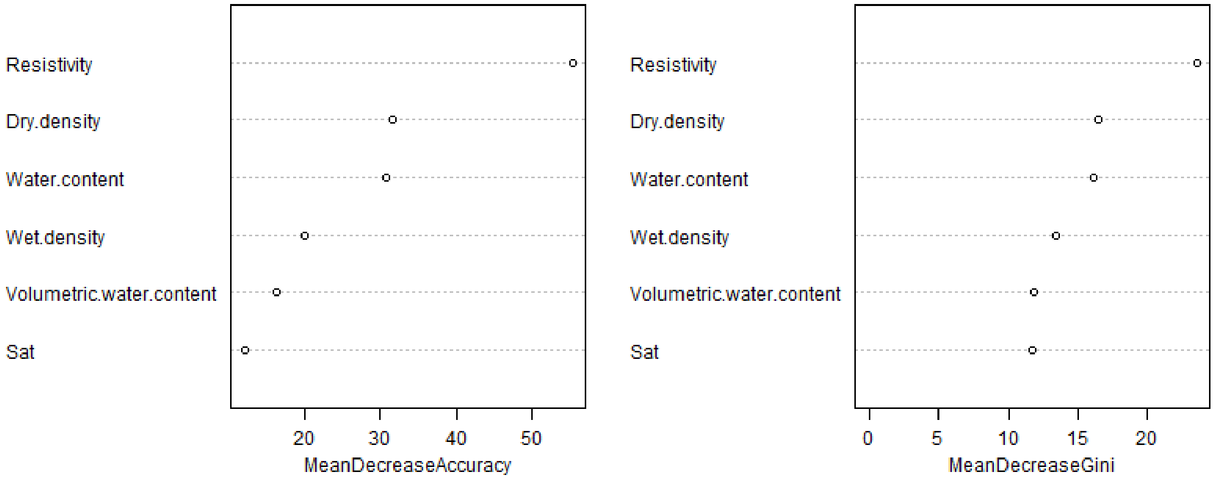

Figure 4. Variable importance plot for random forest classifier with 1,000 trees.

Figure 4 shows that “resistivity” is the most important predictor for soil type prediction, followed by “dry density” and “water content”. Although, we should note that some of these predictors are strongly correlated, and that correlation can affect model selection. For example, an ensemble method can choose one of the two strongly dependent predictors present in a data-set. This variable importance plot is interesting compared with the more qualitative analysis performed in Mofarraj Kouchaki et al. (2018) which showed that bulk density and degree of saturation were the most important variables. Bulk density and degree of saturation are highly correlated through the amount of moisture (water content) in the soil; however, the two identified here are not.

Summary of Performance of Methods for Laboratory Data-set

Table 6 provides the relative performance of all methods used for the laboratory soil data. Based on our experiments and the samples used, logistic regression is the best performer and decision tree/random forest performed second. LDA does not perform as well for classification, perhaps because of the violations of the model assumptions.

Table 6. Relative performance of the methods used

METHOD	CLASSIFICATION ACCURACY
LDA ALL PREDICTORS	0.4545455
LDA RESISTIVITY ONLY	0.3030303
LDA RESISTIVITY, DRY DENSITY, SATURATION	0.5151515
LOGISTIC REGRESSION	0.6363636
DECISION TREE / RANDOM FOREST	0.5454545

3.3 Supervised Learning for Field Data

The logistic regression was applied to the field data to classify the soil types as it showed the best classification performance. Based on boring logs collected at the field, only two soil types (i.e., classifications) were present (CL, and SP). As before, a supervised learning approach was used for validating our model, as well as for preventing overfitting. A single split of training and test data was used to fit and evaluate the model and is called the "validation set". For this validation set approach, only 1 observation out of the total 50 test data points was wrongly classified as "SP" while its true label was "CL". This leads to a probability of correct classification of 98%. Table 6 shows the confusion matrix for this classification task.

Table 7. Confusion data for logistic regression for field data.

	TRUE LABELS	
PREDICTED LABELS	**CL**	**SP**
CL	**14**	0
SP	1	**35**

Significance of predictor variables

One of the biggest advantages of a logistic regression is that it leads to interpretability of the fitted model, that is, we can test which predictors are important in driving this classification accuracy.

The ANOVA table below shows that the predictors R_{ohm} , Velocity (V_s), and SPT were all highly significant with low P-values, confirming our belief that they contain useful information about soil type.

```
## ANOVA in R
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -17.21807   4.24266  -4.058 4.94e-05 ***
## R..ohm.m..raw   0.07355   0.02889   2.546  0.01090 *
## R..ohm.m..Surf  0.03675   0.03528   1.042  0.29753
## Vs..m.s.        0.07345   0.02569   2.860  0.00424 **
## SPT.N           0.40265   0.16197   2.486  0.01292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The receiver operating characteristic (ROC) curve was also plotted and the area under the ROC curve for the fitted logistic regression was measured (Figure 5). We have increased the proportion of observations in the validation data-set from 20% to 50% for a more realistic situation where we have fewer observations for training the logistic model. The resulting ROC curve has AUC = 0.984.

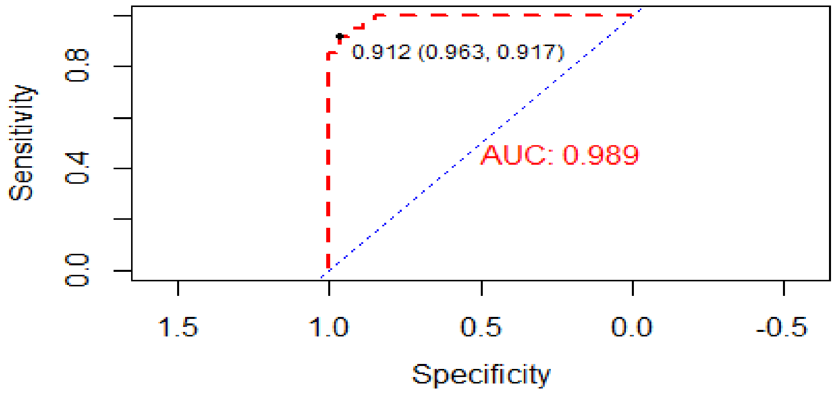

Figure 5. ROC curve.

To ensure that this nearly-perfect classification is not an artifact of the randomly divided training and test sets, we investigated how the performance varied if we chose different random splits of the data into training and validation sets. The accuracy is very high irrespective of the choice of partitioning. The same classification exercise as before was performed, splitting the data 80-20 into training and test sets, respectively. The experiment was replicated 1,000 times and the distribution of the correct classification probability was examined (Figure 6). As expected, the peak of this distribution is around 0.95, substantiating our conclusion that the resistivity variables can predict the basic soil types with a high accuracy.

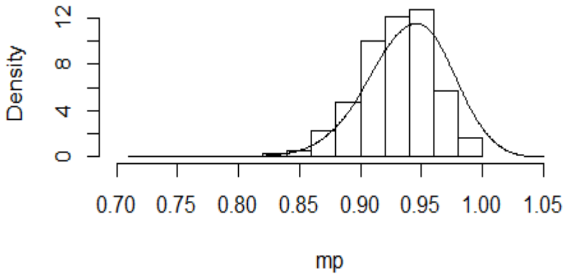

Figure 6. Distribution of probability of correct classification percentages over 1,000 different random splits of training and test data.

Fitting with fewer predictors

The effect of removing some of the predictors from the model on the performance of the logistic regression was also investigated. This will give us insights on how a model with fewer predictors (less cost) will classify the basic soil types. The classification accuracy with just two predictors R_{ohm} and Velocity results in the same accuracy as before (98%), with one SP misclassified as a CL (Table 8). Figure 7 shows the decision boundary, which suggests that this can be expected in this case because the two soil types in the landside data are linearly separable. The ANOVA table below shows that if we use fewer predictors, they all turn out to be significant in predicting the class type, as expected.

Table 8. Confusion data for logistic regression for field data with only two predictors.

	TRUE LABELS	
PREDICTED LABELS	**CL**	**SP**
CL	**15**	1
SP	0	**34**

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -21.28751    5.51046  -3.863 0.000112 ***
## R..ohm.m..raw   0.05780    0.01849   3.127 0.001768 **
## Vs..m.s.        0.12746    0.03379   3.772 0.000162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

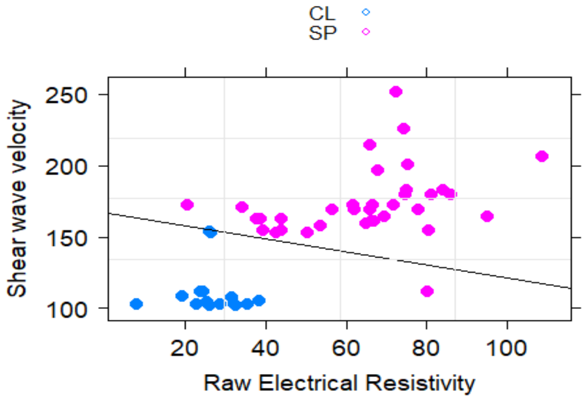

Figure 7. Decision boundary for logistic regression.

The sole purple dot on the wrong side of the boundary in Figure 7 is the misclassified sample point with Soil type = SP that was classified as CL. The solid line separating the two classes is the decision boundary created/calculated by the logistic regression fitted on the data with fewer predictors. While this linear boundary works for the given data set and field site, more analysis is needed for data from additional, more complex, sites in order to determine the appropriateness of this divider. Additional sites should also contain more varied soil stratigraphy so that more than two classifications can be examined.

Impacts/Benefits of Implementation (actual, not anticipated)

This study provides quantitative proof for the first time that shows statistical methods, such as the logistic regression used here, can aid in the classification of soil type using data derived from non-destructive geophysical methods. Although the field site was simple in terms of stratigraphy and number of soil types, the results show that ER and V_s are sufficient (at least for this data set) to capture soil type with an 98% accuracy. The project team is currently searching for an additional test site in order to implement the findings and conduct a "blind study". It is also noted that the models were previously trained with available data. At a new site, this training data would only be available if some destructive traditional drilling and sampling were conducted. The team will also assess the accuracy of the current trained model as it is extended to new sites without the use of prior knowledge. The true benefits and impacts of these findings will be dependent on the success of these future field studies. One additional benefit from this work is the database which was established. This database can be extended during future studies to provide a unique and valuable set of geotechnical data for statistical purposes.

4. RECOMMENDATIONS AND CONCLUSIONS

In this study, we investigated how well several popular statistical methods (e.g., LDA, Logistic Regression and Decision Trees using Random Forest) predicted soil type using ER, V_s, and other geotechnical parameters. A more variable laboratory data-set was used to determine the most accurate method and then the method was applied to a field (landside) data-set. A supervised learning framework was used in both cases.

For the laboratory data, with nine different soil categories, logistic regression showed the highest classification accuracy (~63.63%) on held-out test cases. We applied the logistic regression on the field landside data with 2 soil categories, and 4 different predictors. For the field data, a logistic classifier shows 98% classification accuracy (1 misclassification out of 50 test cases) with only ER and V_s needed as predictors. Although the field data has fewer categories, which makes it an easier classification task, we conjecture that the logistic regression will be able to predict the soil types even with more categories. As stated above, a more variable field site is needed in order to test this conjecture.

There are a number of future research directions for soil-type predictions that are also recommended. A few are summarized as the following:

1. For the laboratory data, the analysis can be repeated for a data-set where similar soil-types are grouped to investigate if the classification accuracy can be improved. The grouped samples would have little differences geotechnically speaking and it would give a much higher sample size per category. A related question is, are there properties of the soil types that lend them to a natural or important pairing which might be outside of the USCS classifications traditionally used? This question arises because ER is a fundamental material property and the measure is based on electrical properties and mineralogy rather than strictly a behavioral classification.

2. A principled variable selection strategy, such as best subset selection or penalized classification (E.g. LASSO), can be applied or a dimension reduction using Principle Components (PCA) can be performed to further develop the analysis.
3. A more variable field site is needed to test the ability of the trained model to be extended to new sites and to provide additional more robust training data for additional models.

5. REFERENCES

Abu-Hassanein, Z. S., Benson, C. H., and Blotz, L. R. (1996). Electrical Resistivity of Compacted Clays. *J. Geotech. Eng.*, 122(5), 397-406.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Everett, M. (2013). *Near-Surface Applied Geophysics.* Cambridge University Press, 1 edition.

Friedman, S.P. (2005). Soil properties influencing apparent electrical conductivity: a review. *Computers and Electronics in Agriculture*, 46(1–3), 45-70.

Fukue, M., Minato, T., Horibe, H., Taya, N. (1999). The micro-structures of clay given by resistivity measurements. *Engineering Geology*, 54(1–2), 43-53.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Loke, M.H. (1997). Electrical Imaging Surveys for Environmental and Engineering Studies. Workshop Held in USM, http://web.archive.org/web/20180114130825/http://pages.mtu.edu/~ctyoung/LOKENOTE.PDF (accessed 09 September 2017).

Mofarraj Kouchaki, B., Bernhardt-Barry, M.L., Wood, C.M., Moody, T. (2018). A laboratory investigation of factors influencing the resistivity of different soil types. *Geotechnical Testing Journal*, doi.org/10.1520/GTJ20170364.

Parkhomenko, E. (1967). Electrical Properties of Rocks. G.V. Keller, translator, ed., Plenum Press, New York, N.Y.

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package ‘nnet’. *R package version*, 7-3.

Samouëlian, A., Cousin, I., Tabbagh, A., Bruand, A., and Richard, G. (2005). Electrical Resistivity Survey in Soil Science: A Review. *Soil Tillage Res.*, 83(2), 173-193.