Predicting Africa Soil Properties Using Machine Learning Techniques

Iretiayo Akinola *, Thomas Dowd †,
Electrical Engineering
Stanford University, Stanford, CA 94305
Email: *iakinola@stanford.edu, †tjdowd@stanford.edu

***Abstract*—Different machine learning algorithms were assessed for estimating five functional soil parameters (SOC content, Calcium content, Phosphorous content, sand content, and pH value). The algorithms used include variants of linear regression and support vector regression. A closer look at the prediction performance for each target revealed that apart from pH, which consistently had worse performance, prediction for the other soil properties was quite satisfactory (RMSE $<$ 0.4). Applying machine learning techniques to soil properties prediction has shown a lot of promising and encouraging results. Getting more data, domain knowledge and intuition, possibly from soil scientist/experts, would surely maximize this potential for accurate soil property prediction.**

I. Introduction

Soil functional properties (such as primary productivity, nutrient and water retention, and resistance to erosion) indicate a location's ability to perform important ecological services. Traditional methods of measuring and characterizing soil properties require expensive and time-consuming scientific procedures. Low cost measurements obtained through diffuse reflectance infrared spectroscopy and remotely collected data have the potential to quickly estimate these same characteristics without the use of costly chemical resources. Inexpensive characterization methods would allow large, data-sparse regions to plan sustainable agricultural development and manage local natural resources.

Our objective was to accurately model the relationship between these inexpensive measurements and five soil characteristics properties: Soil Organic Carbon (SOC) content, Calcium content, Phosphorus content, sand content, and pH value.

II. Literature-review

Since the mid 1900s, pedotransfer functions (PTF), which are predictive functions of certain soil properties using data from soil surveys, have been used to predict soils in temperate regions. However, soil prediction methods can be adapted across climates. According to Minasnya (2011) [1], methods developed in temperate regions can be applied for the soils in the tropical regions albeit with adjustments to calibration and choice of relevant available predictors. Prior to model selection, a literature review was conducted on past work modeling spectroscopy data to soil characteristic qualities.

Pedotransfer functions and most other models used in (predictive) soil science are linear regression based. Rossel et al. (2006) performed an analysis on using visible, near-infrared, and mid-infrared absorbance data to determine soil characteristic properties. Their method used partial least-square regression (PLSR) to model the system, and they were most successful in developing predictions for pH, Organic Carbon, Phosphorous, and sand content using mid-infrared absorbance values. [2]

Rossel and Behrens (2009) experimented with a number of machine learning methods to map full-range spectroscopy data to similar soil characteristics. Support Vector Regression provided the most accurate results for the three target variables in question (SOC content, clay content and pH levels). [3]

More recent prediction of soil properties such as [1] and [4] still embrace the linear model assumption for their research.

In this project, we explore how recent improved machine learning techniques can be applied to the soil science domain to improve the overall prediction performance.

III. Dataset

The dataset consisted of a collection of 1,157 soil sample measures. Soil was collected from a variety of locations in Africa. Each data point contained 3,594 features which represent the following low-cost measurements:

- 3,578 mid-infrared absorbance measurements (wavelengths ranging from 7497.96 cm^{-1} - 599.76 cm^{-1}). Originally obtained via Diffuse Reflectance Infrared Fourier transform Spectroscopy.
- Depth of the soil sample (topsoil or subsoil)
- Remotely-collected data including climate, topographical, vegetation index, surface temperature, and other information about the sample collection site obtained via satellite. All satellite data was mean centered and scaled.

The data set contains monotonically adjusted values for the five target variables (SOC content, Calcium content, Phosphorous content, sand content, and pH value) such that all target variables take both positive and negative values.

The dataset was provided by the "Africa Soil Property" competition hosted on Kaggle.com.

IV. Model Selection

Noting the success of linear fitting in other datasets mapping spectroscopy data to soil content characteristics, a number

of linear regression variants were attempted. In an effort to be thorough, other methods such as support vector regression and clustering were also attempted. The choices of models tried were also guided by the fact that the number of features is far more than the sample size which eliminated options like SGD Regressor (Stochastic Gradient Descent Regressor) that requires relatively large data-points.

Linear Regression:

Determines linear coefficients to map feature set to target variables. Linear coefficients, β, for the feature set data, x, are determined for each target variable, y, across the entirety of the test set (of size m) such that $\sum_{i=1}^{m} \left(y_i - \beta^T x_i\right)^2$ is minimized.

Ridge Regression:

Ridge regression is similar to linear regression in that it determines linear coefficients to map the feature set to target variables. However, ridge regression also attempts to limit the magnitude of its linear coefficients by adding an additional term that penalizes the l_2-norm of the β term. β is chosen as the following:

$$\beta^{ridge} = argmin_\beta \sum_{i=1}^{m} \left(y_i - \beta^T x_i\right)^2 + \gamma \|\beta\|_2^2$$

where γ is a tuning parameter.

Lasso Regression:

Lasso regression takes the same approach as ridge regression only it penalizes the l_1-norm of the β term, which results in the following definition:

$$\beta^{lasso} = argmin_\beta \sum_{i=1}^{m} \left(y_i - \beta^T x_i\right)^2 + \gamma \|\beta\|_1$$

where γ is again a tuning parameter.

Principal Component Regression:

PCR uses principal component analysis (PCA) to first reduce the dimensions of the feature space. Linear regression is then used to determine coefficients mapping the new feature space to the target variables. PCA transforms the feature space by using each sample's feature measurements to determine correlation in the feature space variables, converts the feature data into a set of orthogonal and linearly uncorrelated components, and uses the first n components to limit the original feature space to $\mathbb{R}^n$. PCA work in this project was conducted through packages using the singular value decomposition of $X^T X$, where X is the feature space data.

Partial Least-squares Regression:

PLSR is a similar procedure to PCR, only the objective is to identify and remove cross-correlation between feature set and the target variables. While PCR utilities the singular value decomposition of $X^T X$, PLSR uses the singular value decomposition of the product of the transposed feature space and the target variable space, $X^T Y$, to determine common orthogonal factors and transform both feature and target variable space. Linear regression is then conducted on the transformed sets. [5]

Support Vector Regression:

Developed by Vapnik (1998). [6] For a feature space of $x_i \in \mathbb{R}^n$ and target variable $y_i \in \mathbb{R}$ with a total of m samples, parameters $C > 0$ and $\epsilon > 0$, and a valid kernel function $\phi(x)$, we can develop the following problem that would best fit the model estimating y as $w^T \phi(x) + b$:

$$\begin{array}{rc} \min_{w,b,\xi,\xi^*} & \frac{1}{2} w^T w + C(\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \xi_i^*) \\ \text{subject to} & w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i, \\ & y_i - w^T \phi(x_i) - b \leq \epsilon + \xi_i, \\ & \xi_i, \xi_i^* \geq 0, i = 1, ..., m \end{array}$$

The dual of this problem is the following:

$$\begin{array}{rc} \min_{\alpha,\alpha^*} & \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \\ & \epsilon \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \\ \text{subject to} & e^T(\alpha - \alpha^*) = 0 \\ & 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, ..., m \end{array}$$

where $Q = K(x_i, x)$. The approximate solution of the dual is $\sum_{i=1}^{m} (-\alpha_i + \alpha_i^*) K(x_i, x) + b$. [7] This project attempted to fit the data using the gaussian kernel.

V. Implementation Details

The error metric used in this project was the mean column-wise root mean square error (MCRMSE) between the actual target variables values and predicted target variable values.

$$\text{MCRMSE} = \frac{1}{5} \sum_{j=1}^{5} \sqrt{\frac{1}{n} \sum_{i=1}^{m} (y_i^{(j)} - \hat{y_i}^{(j)})}$$

Root-mean square error (RMSE) was used to compare model-fitting performance of individual target variables.

$$\text{RMSE}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{m} (y_i^{(j)} - \hat{y_i}^{(j)})}$$

K-fold cross validation (K = 10) was used to develop an accurate measurement of these error metrics. Model training and prediction was conducted in Python using the `scikit-learn` package.

VI. Results

Feature Reduction

As the dimension of the feature set exceeds the number of samples, dimensionality reduction through principal component analysis distilled the feature set to its most critical components. Most of the models used did not benefit significantly from the reduced feature space, likely because the other algorithms' constraints performed a similar action of reducing highly related features.

The model with most improved performance was linear regression, which, when conducted after PCA became PCR. To determine the optimal number of features, PCR was conducted a 100 times with a range of feature reduction factors (from 15 to the number of samples available in the training set). The

training and test MCRMSE of these models were calculated for each reduced feature set. The results can been seen in the plot below:

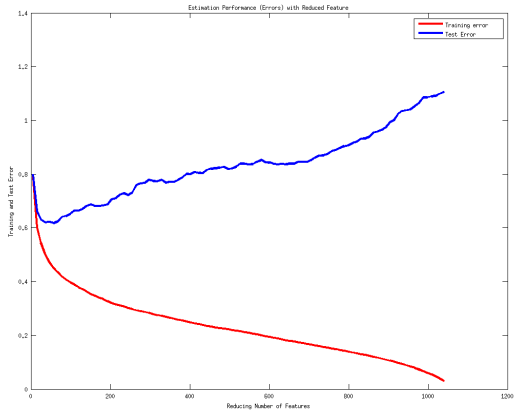

Fig. 1: MCRMSE vs. Total Features

Using these results, the optimal number of features was determined to be 65. To ensure the validity of this conclusion, singular value decomposition was conducted on the feature set, which determined that a reduced feature set in $\mathbb{R}^{65}$ preserved 99.5% of the data.

Model Performance

The figure below shows the prediction performance by six best performing algorithms on the five target variables. The overall prediction performance showed that the linear regression based models had comparable and better performances compared to the others. Ridge regression and PCR, the two top-performing models, both perform the same underlying function of penalizing the non-relevant features. While ridge regression does this by constraining the regression coefficients, PCA transforms the feature set by removing redundancies. Lasso regression also had similar performance.

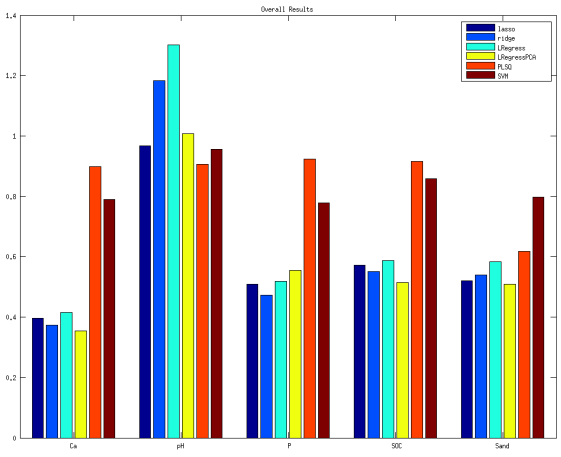

Fig. 2: Target Variable RMSE by Model

The soil data could be divided two classes of similar size: topsoil and subsoil. After getting the performance for the entire dataset, we repeated the whole prediction procedure for each of the classes and found that training and testing errors were very similar to that of the overall performance.

	LR	PCR	Lasso	Ridge	PLSR	SVR
Training Error:	0.363	0.369	0.497	0.335	0.870	0.0951
Test Error:	0.681	0.588	0.593	0.624	0.853	0.8361

Table 1: Model Prediction Performance (MCRMSE)

The table above contains only the results from the most successful models. Some other methods attempted that achieved poorer results are summarized below.

Other Models

One model attempted was a double-layered estimation model was considered where an initial a classification stage on quantized data is followed by a linear regression prediction of data in each of the quanta-bins. The results of the first SVM-based classification step were not as impressive enough. While the linear kernel performed best (compared to gaussian, polynomial, sigmoid kernel) which confirmed the notion that linear models are best for soil prediction, the highest prediction performance achieved for just two class quantization was less 88% for all the target variables. This forms a poor basis for the following regression step on as already misclassified datapoints would get poor estimations results in the end; besides the misclassified datapoints from the first stage might corrupt the regression model learning/ training process.

Bagged Support Vector Regression was also conducted in an effort to prevent overfitting for training data. However, the bagged results were not significantly different from the regular SVR output.

VII. Discussions

For the majority of the target variables, linear regression and other linear variations achieved high performance. The exception was pH, which remained the most difficult to predict for all models. Figure [3] shows that pH was less correlated with other target variables. This could be because the set of features in the data does not capture all the factors that contribute to the pH of soils. While other target variables might have intersection of features that are commonly indicative of them in varying degrees/weights present in the dataset, capturing indicators of some of the many different component minerals that contribute to the pH of a soil might require other types of features.

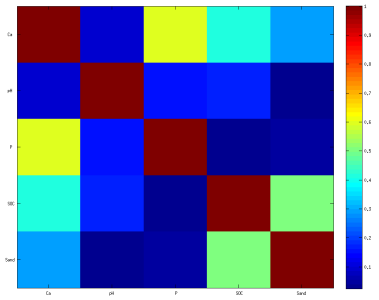

Fig. 3: Target Variable Correlation

Linear Regression

Given that linear regression models seek to minimize least-squares error, linear regression performed very well on the training set. However, the higher test error suggested that the linear relationship between the training features set and target variables could not be directly extended to test data. Variations on linear regression were explored to compensate for this deficiency without adjusting the feature set, while other methods sought to increase accuracy by adapting and increasing the feature set.

Ridge Regression

The estimation performance expectedly depends on the choice of regularizer. Since there are five different target variables, different regularization values optimizes each of the targets. In particular, as the regularization parameter was varied in one direction, performance of the targets (except P) improved; a compromise had to be made on the regularization value chosen. Performing PCA before ridge regression did not improve performance.

Lasso Regression

This is another regularized regression model and the performance was very comparable to that of ridge regression as expected. However, it was observed that Lasso did considerably better than ridge and and slightly better than PCR on the prediction of pH. This suggests that the l_1 norm regularization might be more suited for specific target variables than others.

Principal Component Regression

Based on the MCRMSE metric, the PCR model had the best performance of all models tested. A quick analysis of the results for individual target variables reveal that the PCR method achieved very similar results to lasso regression. PCR served as a significantly better estimator for pH than standard linear regression, demonstrating that removing the highly correlated features contributed to a more generalized model for pH.

Partial Least-squares Regression

PLSR performed significantly worse than the other linear regression variants for almost all the target variables. However, its higher performance for pH estimation was notable. Unlike the other linear methods, the reduction of common correlations between features and target variables allowed for a linear model that achieved similar RMSE across all the target variables.

Support Vector Regression

Support Vector Regression was used in an attempt to obtain a better generalized fit for the data. SVR was conducted with a gaussian kernel. The result had a very low training error metric, but suffered significant drop in performance for the test data, suggesting an overfitting problem. Further attempts were made to improve the gaussian kernel SVR performance through bagging methods, but these did not yield better results.

VIII. CONCLUSION

Asides from pH which performed poorly, the prediction performance of our study shows that some soil functional properties can estimated reasonably well (e.g. Ca with rmse¡.4) using carefully selected cheaper soil characteristics as predictors and smart machine learning algorithms. Our results show that Machine Learning techniques applied soil properties prediction holds a lot of promise. With more data and soil science domain-specific tricks, the potential for applying machine learning to soil property prediction would surely be maximized.

IX. FUTURE WORK

Use of probabilistic graphical models for capturing correlations between target variables. The correlation map above shows that some target variables are quite correlated. Joint probability models might be useful to incorporated these correlation information into the prediction process to enhance the overall prediction performance.

Get access to raw data and try out other transformation techniques to extract features. The dataset obtained for this project had been cured and conditioned by the source. It would be interesting to try out different feature transformation techniques on the raw spectrophotometer measurements. This would help build more accurate models about the data.

Implementation of ensemble methods to combine successful estimation models. In addition to exploring other possible prediction algorithms, ensemble learning technique could be employed to combine the top performing algorithms to improve overall performance.

Cluster analysis of the data might reveal and pull together samples with similar soil characteristics. Target variables can then be separately predicted for each of the clusters.

X. ACKNOWLEDGEMENT

We would like to thank Professor Ng for his excellent instruction, and the CS 229 TAs for their help and advice on this project.

REFERENCES

[1] B. Minasny and A. E. Hartemink, "Predicting soil properties in the tropics," *Earth-Science Reviews*, vol. 106, no. 12, pp. 52 – 62, 2011.

[2] R. V. Rossel, D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma*, vol. 131, no. 1–2, pp. 59 – 75, 2006.

[3] R. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1–2, pp. 46 – 54, 2010. Diffuse reflectance spectroscopy in soil science and land resource assessment.

[4] J. A. C. Medeiros, M. Cooper, J. Dalla Rosa, M. Grimaldi, and Y. Coquet, "Assessment of pedotransfer functions for estimating soil water retention curves for the amazon region," *Revista Brasileira de CiÃ do Solo*, vol. 38, pp. 730 – 743, 06 2014.

[5] H. Abdi, "Partial least squares regression and projection on latent structure regression (pls regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[7] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.