Estimating soil carbon content using easily obtainable parameters

Francisco Carlos Groppo Filho[1]
ICMC-USP
Paulino Ribeiro Villas Boas[2]
Embrapa Instrumentação

1 Introduction

Among the many strategies adopted to mitigate climate change is the reduction of atmospheric carbon in a process called carbon sequestration, which consists of the transfer of carbon dioxide from the atmosphere to other global pools, such as the soil [1]. Therefore, quantifying the soil carbon is of great importance for successfully measuring the efficiency of carbon sequestration practices and providing accurate reports [2].

The correct measurement of soil carbon is a costly and cumbersome process requiring shipping of samples from the field to laboratories, greatly limiting its applicability [3]. In order to reduce the cost and time required for analyses, several techniques have been developed, such as laser-induced breakdown spectroscopy [4] and online visible and near-infrared spectroscopy with random forests [5]. Even though these new techniques are faster and less expensive, samples still are required to be collected in the field.

Developing a method that could provide estimates of the carbon content of farms, using easily obtained variables such as soil texture and practices, would contribute to understanding the relationship between these variables and soil carbon, facilitating carbon sequestration initiatives [6]. Thus, the aim of this project was to train a model on the data available and verify its validity.

2 Material and Methods

In order to develop the model, data from 53 farms was used, spread across four Brazilian regions: northeast, center-west, southeast, and south (Figure 1). The variables used for training included data from different categories, such as soil characteristics (fine and coarse sand, silt, and

[1] groppofilho@gmail.com
[2] paulino.villas-boas@embrapa.br

clay content, and pH), location (region in Brazil), field type, and average depth. The target variable was defined as the amount of carbon in grams per kilogram of soil.

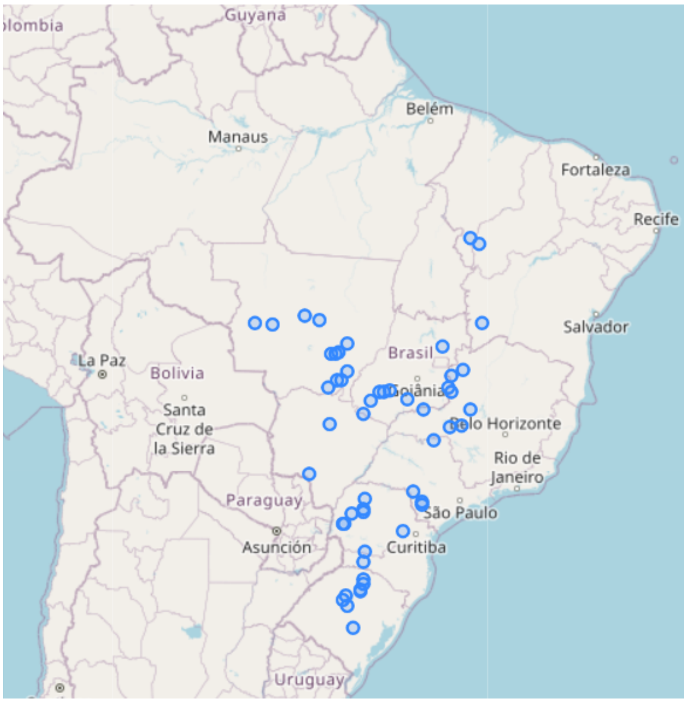

Figure 1: Distribution of farms, where data were collected.

Since the soil samples were obtained in depth ranges, the original data presented a depth start and end, but, for the training, the average depth was obtained by dividing their sum by two. Regarding the field types, four were identified, “FP” (Farmer’s Practice), where the producer employed routinely used methods to plant, such as using a plow; “SP” (Sustainable Practice), where techniques such as direct planting were used; “NV” (Natural Vegetation), usually an area of woods located inside the farm, such as riparian forest; and “OTHERS”, which did not fit in any of the above types.

The next step after defining the variables of interest was removing every row containing nulls, after which 4904 rows remained. They were then split into train, test and validation data frames, with two farms being set aside for the validation and the remainder being divided in a proportion of 0.67:0.33 train:test, culminating in a final ratio of 0.64:0.32:0.04 train:test:validation. The average depths were used as the stratification parameter, so as to retain the proportion and avoid skewing.

Using the split data, a gradient boosted decision trees regression model (XGBoost [7]) was trained, then scored using the percent root mean square error (RMSE), defined by Equation (1) [8], where n is the number of data, R_i is the real, measured value, while E_i is the estimated value and $\bar{R}$ is the mean of real values.

$$RMSE = \frac{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(R_i - E_i)^2}}{\bar{R}} \times 100\% \tag{1}$$

3 Results and Discussions

We were able to find a model that successfully provides an estimate of soil carbon based on the easily obtained properties aforementioned. The model used showed a RMSE = 32.9% on test data. While not optimal, the results demonstrate a model capable of estimating the amount of soil carbon in certain farms. The core difference between the test and validation data used is that the validation set contains all the data for two specific farms, while the test data was not isolated by farm for the training.

Upon inspecting the importance of each feature for the model, we found that the average depth was, by far, the most important attribute (Figure 2). Even though we expected farming practices and soil texture to be more significant, they were less important to the model than the region.

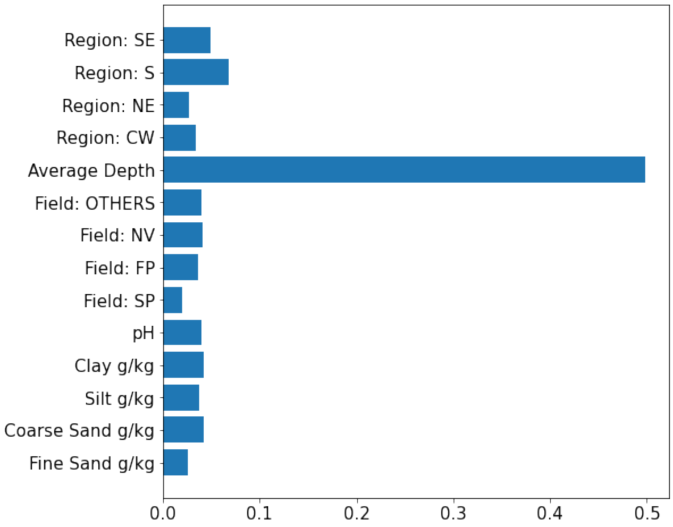

Figure 2: Feature importances.

As mentioned, the RMSE found is not optimal but offers promising results. Initially, the model was trained and tested without the use of validation data and using longitude and latitude. In this scenario, it presented a RMSE of 21.7%, which, at first glance, appears an improvement from the final model selected. However, upon investigating the feature importances and retraining the model with two farms set aside for validation, we found the model was very dependent on latitude and longitude, which meant it was classifying by each farm's attributes before performing the regression itself. In other words, it was as if the model had information about the farm attributes before actually predicting the soil carbon content.

After removing the latitude and longitude and adding the validation set, we obtained the results for the test previously mentioned, along with RMSE of 54.5% and 16.5% for the two farms set aside for validation. The difference between them is noticeable implying that the information regarding the farms is important in this model.

The model showed reasonable results due to the complexity of the problem and the lack of additional data. Previous studies have proven a relationship between climate, soil classes, and plant residues with soil carbon concentration, variables which were not obtained for the present study [9]. Obtaining and adding these parameters to the model would likely improve its accuracy.

4 Conclusion

This study has shown the difficulty of training a model that provides an accurate estimation of soil carbon based on easily obtained parameters. Using soil texture and pH, region, field type and sample depth, the best model presented a RMSE of 32.9%. We believe that if climate, soil class and plant residue data were included in the model, the results would be more accurate.

References

[1] R. Lal. Carbon sequestration, *Phil. Trans. R. Soc.*, B363815–830, 2008. DOI: https://doi.org/10.1098/rstb.2007.2185

[2] K. Paustian et al. Quantifying carbon for agricultural soil management: from the current status toward a global soil information system, *Carbon Management*, 10:6, 567-587, 2019. DOI: 10.1080/17583004.2019.1633231

[3] R. Gehl and C. Rice. Emerging technologies for in situ measurement of soil carbon, *Climatic Change*, 80, 43-54, 2007. DOI: 10.1007/s10584-006-9150-2.

[4] G. Nicolodelli et al. Quantification of total carbon in soil using laser-induced breakdown spectroscopy: a method to correct interference lines, *Appl. Opt.*, 53, 2170-2176, 2014.

[5] S. Nawar and A.M. Mouazen. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning, *Soil and Tillage Research*, Volume 190, Pages 120-127, ISSN 0167-1987, 2019. DOI: https://doi.org/10.1016/j.still.2019.03.006.

[6] P. Smith et al. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal, *Glob Change Biol.*, 26: 219– 241, 2020. DOI: https://doi.org/10.1111/gcb.14815.

[7] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, NY, USA, 2016. DOI: https://doi.org/10.1145/2939672.2939785.

[8] M-F. Li et al. General models for estimating daily global solar radiation for different solar radiation zones in mainland China, *Energy Conversion and Management*, Volume 70, Pages 139-148, ISSN 0196-8904, 2013. DOI: https://doi.org/10.1016/j.enconman.2013.03.004.

[9] Y. Liang et al. Simulating soil organic matter with CQESTR (v. 2.0): Model description and validation against long-term experiments across North America, *Ecological Modelling*, Volume 220, Issue 4, Pages 568-581, ISSN 0304-3800, 2009. DOI: https://doi.org/10.1016/j.ecolmodel.2008.11.012.