Data Analysis of Soil Parameters at a Local Land Preserve

Data-151: Intro to Data Science

12/8/2024

By: Jack Colwell, Owen Galicia, Santiago Gutiérrez-Morales, and Lilu Smith

Executive Summary

Data Analysis of Soil Parameters at a Local Land Preserve used a data analysis approach to analyze data previously collected in a study titled *Environmental Restoration Target Estimation Around Engquist Nature Preserve* by Jack Colwell and other students at Valparaiso University. This project's objective is to use our various soil parameters as predictors for SOC in the previously collected soil samples. The SOC content in soil varies, influencing the ability to manage healthy agriculture and the global carbon cycle. Understanding these SOC levels helps farmers make more informed decisions on how to maintain healthy soil levels and mitigate climate change.

The data was cleaned and errors were removed using Python in a Google Colab environment. Then a correlation matrix was created to narrow down the highest correlated variables to use in our prediction for SOC content. From this matrix, it was found that two variables from this dataset are the best predictors to use: Carbonate-Free Mineral Sample (g) and Mineral Sample (g). By using these two variables, several regression models were run to output the RMSE (Root Mean Square Error) and R^2 value of each regression model, as these results influence the decision on which model is best to apply.

Key findings indicate that the random forest regressor, which tends to resist overfitting, is the best model for this situation and data. That predictive model outputs are as follows:

≈ 0.760 RMSE
≈ 0.914 R^2

This means that approximately 91.4% of the variance in the data can be explained by our model. That is to say, using these two variables, the SOC content can be reliably predicted, without a direct measure of SOC content itself. We validated that the random forest regressor is not overfitting as we checked the R^2 value in the training section and it resulted in 0.8731. When dealing with an overfitted model the R^2 value is extremely close to 1.

These findings demonstrate that accurate SOC predictions can be achieved without measuring all variables, significantly reducing costs and time. By optimizing soil testing practices, farmers can improve crop management and enhance resilience to climate change. Moreover, this research aligns with broader goals in climate-smart agriculture by supporting efforts to increase soil carbon stocks and reduce greenhouse gas emissions.

Introduction

This project analyzed data from a previous research project titled: *Environmental Restoration Target Estimation Around Engquist Nature Preserve*. This research project collected data at the Dale B. Engquist Preserve operated by Shirley Heinze Land Trust (SHLT) by students of the Geography Department at Valparaiso University, including Jack Colwell (member of this group), Korbin Opfer (current student), and Doc Janowiak (recent alumnus). This research was also supported by Dr. Jon-Paul McCool of the same department. These past projects collected soil samples from different environment types at Engquist and the samples were subsequently analyzed for various soil parameters. The different environment types are shown in Figure 1 and soil parameters are shown in Figure 2 respectively.

Environment Types:				
Prairie	Forested Wetland	Forest	Floodplain Forested Wetland	Agriculture

Figure 1: The different environment types that data was collected from in the previous projects.

Soil Parameters:					
Bulk density	Gravel mass	Rock volume	Soil volume	Wet weight	Dry weight
Mineral sample	Organic matter percent	SOC Percent	Calcium carbonate percent	Estimated carbon in horizon per sq. m. of soil	Estimated carbon in 10 cm depth per sq m. of soil.

Figure 2: The different soil parameters that were measured from the previous projects.

This project explored healthy ranges for environmental variables for agriculture and then compared that to our existing dataset. Furthermore, it is important to develop a prediction for one or more variables within this dataset due to the cost and time to run tests on these parameters. By modeling we showed that one soil parameter can accurately predict another, or at least the magnitude and direction of another parameter. When these relationships are identified it can lead to accurate ways to take proxy measurements of soil parameters, rather than measure every single variable.

The average SOC content in Indiana soils varies depending on factors such as land management, soil type, and climate. In general, the SOM levels across Indiana soils can be anywhere from less than 1 percent to over 10 percent in the muck soils of Northwestern Indiana (Purdue University, Indiana State Climate Center, 2021). In general, the SOC levels across Indiana agricultural soils are often between 0.67% and 2% in the topsoil layer (for our project, this means the top 10cm of a soil profile). Though some areas with organic-rich soils may exceed this range the estimated SOC range.

In order to develop our model, we ran different analyzation techniques, like a correlation matrix of the different parameters that were measured. Then these outcomes informed our predictive models (linear regression, decision tree, and LASSO were used).

The project is important because of SOC's necessity to agriculture and climate change, specifically global warming. *Elements of the Nature and Properties of Soils* by Weil and Brady mention, "One year after plant residues are added to the soil, most of the carbon in them has returned to the atmosphere as CO_2." This dynamic underscores SOC's nature and the importance of sustainable soil management to retain carbon in the soil. In terms of agriculture, SOC helps farmers make informed decisions about their crops and soil management. By understanding SOC levels, farmers can quickly identify nutrient deficiencies or imbalances, ensuring that these issues are resolved before planting, leading to more productive and sustainable farming practices. In terms of climate change, soils hold a substantial amount of organic carbon, which can be released into the atmosphere as carbon dioxide if soils degrade, or prevented if managed sustainably, thus playing a significant role in the global carbon cycle.

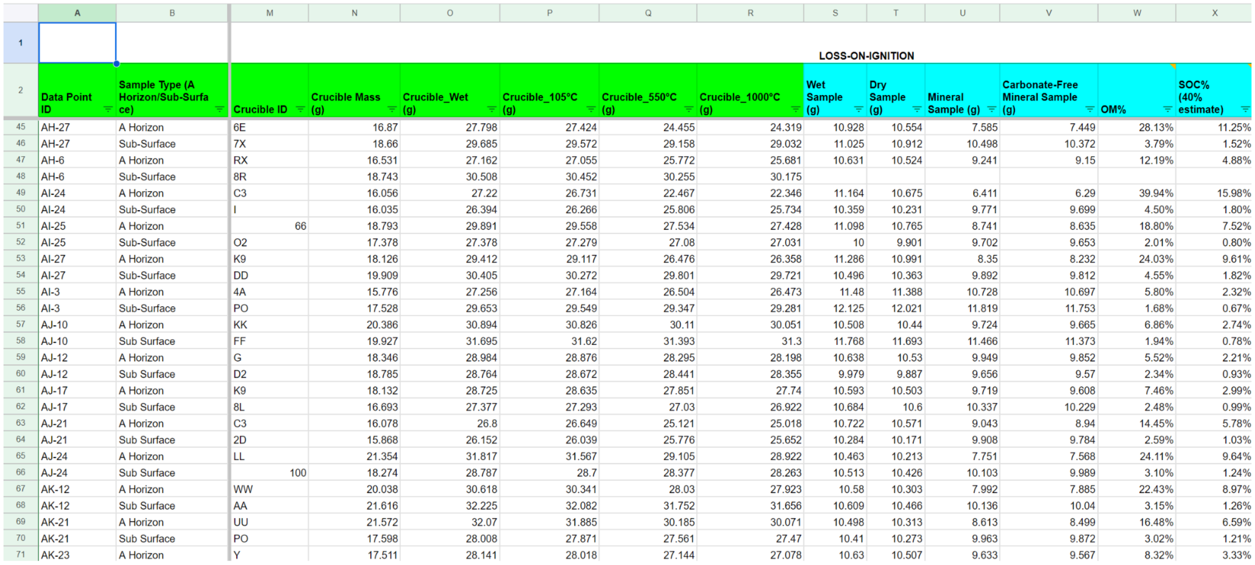
As a process based model, this research builds on previous work by Elizabeth Ellis and Keith Paustian whose study in “Importance of On-Farm Research for Validating Process-Based Models of Climate-Smart Agriculture” underscores the need for refining such models to help with climate smart agriculture. Climate-smart agriculture aims to boost soil carbon stocks, reduce greenhouse gas (GHG) emissions, and improve crop resilience under climate change. Ellis and Paustian’s research concludes that these models often lack representation of real-world conditions, leading to a gap in their accuracy when applied to places like commercial farms. However, measuring and modeling soil organic carbon (SOC) stocks continuously, instead of just once, through on-farm research will capture the diversity of agricultural practices and local conditions. Our approach is modeling data from 2024, which will improve model precision, aiding farmers, ranchers, and forest landowners in adopting practices that not only enhance productivity but also support climate change mitigation.

A “Remote Quantification of Soil Organic Carbon: Role of Topography in the Intra-Field Distribution” article by Cutting et al. has also used models like a random forest to try and model SOC levels. This study has used different models to try and map spatial distribution of SOC since that distribution provides important insights into the biophysicochemical processes involved in the retention of SOC and climate change. Similar to our approach, they took certain parameters of soil and tested their importance to SOC levels. Cutting found that topographic wetness index (TWI) has an impact on soil depending on the climate. From this, knowing the causes of SOC levels will help farmers manage their soil and mitigate climate change.

Therefore, our research question is what variables from our existing dataset can be used to model SOC? Our H0 (null) is: There are no variables that can accurately predict SOC. Ha: (alternative): At least two variables are significant predictors of SOC.

Data and Methods

The data that our team has been analyzing has been collected by Jack Colwell, Korbin Opfer, and Doc Janowiak. The data in its original form is shown in Figure 3.



	A	B	M	N	O	P	Q	R	S	T	U	V	W	X
1									LOSS-ON-IGNITION					
2	Data Point ID	Sample Type (A Horizon/Sub-Surface)	Crucible ID	Crucible Mass (g)	Crucible_Wet (g)	Crucible_105°C (g)	Crucible_550°C (g)	Crucible_1000°C (g)	Wet Sample (g)	Dry Sample (g)	Mineral Sample (g)	Carbonate-Free Mineral Sample (g)	OM%	SOC% (40% estimate)
45	AH-27	A Horizon	6E	16.87	27.798	27.424	24.455	24.319	10.928	10.554	7.585	7.449	28.13%	11.25%
46	AH-27	Sub-Surface	7X	18.66	29.685	29.572	29.158	29.032	11.025	10.912	10.498	10.372	3.79%	1.52%
47	AH-6	A Horizon	RX	16.531	27.162	27.055	25.772	25.681	10.631	10.524	9.241	9.15	12.19%	4.88%
48	AH-6	Sub-Surface	8R	18.743	30.508	30.452	30.255	30.175						
49	AI-24	A Horizon	C3	16.056	27.22	26.731	22.467	22.346	11.164	10.675	6.411	6.29	39.94%	15.98%
50	AI-24	Sub-Surface	I	16.035	26.394	26.266	25.806	25.734	10.359	10.231	9.771	9.699	4.50%	1.80%
51	AI-25	A Horizon	66	18.793	29.891	29.558	27.534	27.428	11.098	10.765	8.741	8.635	18.80%	7.52%
52	AI-25	Sub-Surface	O2	17.378	27.378	27.279	27.08	27.031	10	9.901	9.702	9.653	2.01%	0.80%
53	AI-27	A Horizon	K9	18.126	29.412	29.117	26.476	26.358	11.286	10.991	8.35	8.232	24.03%	9.61%
54	AI-27	Sub-Surface	DD	19.909	30.405	30.272	29.801	29.721	10.496	10.363	9.892	9.812	4.55%	1.82%
55	AI-3	A Horizon	4A	15.776	27.256	27.164	26.504	26.473	11.48	11.388	10.728	10.697	5.80%	2.32%
56	AI-3	Sub-Surface	PO	17.528	29.653	29.549	29.347	29.281	12.125	12.021	11.819	11.753	1.68%	0.67%
57	AJ-10	A Horizon	KK	20.386	30.894	30.826	30.11	30.051	10.508	10.44	9.724	9.665	6.86%	2.74%
58	AJ-10	Sub Surface	FF	19.927	31.695	31.62	31.393	31.3	11.768	11.693	11.466	11.373	1.94%	0.78%
59	AJ-12	A Horizon	G	18.346	28.984	28.876	28.295	28.198	10.638	10.53	9.949	9.852	5.52%	2.21%
60	AJ-12	Sub Surface	D2	18.785	28.764	28.672	28.441	28.355	9.979	9.887	9.656	9.57	2.34%	0.93%
61	AJ-17	A Horizon	K9	18.132	28.725	28.635	27.851	27.74	10.593	10.503	9.719	9.608	7.46%	2.99%
62	AJ-17	Sub Surface	8L	16.693	27.377	27.293	27.03	26.922	10.684	10.6	10.337	10.229	2.48%	0.99%
63	AJ-21	A Horizon	C3	16.078	26.8	26.649	25.121	25.018	10.722	10.571	9.043	8.94	14.45%	5.78%
64	AJ-21	Sub Surface	2D	15.868	26.152	26.039	25.776	25.652	10.284	10.171	9.908	9.784	2.59%	1.03%
65	AJ-24	A Horizon	LL	21.354	31.817	31.567	29.105	28.922	10.463	10.213	7.751	7.568	24.11%	9.64%
66	AJ-24	Sub Surface	100	18.274	28.787	28.7	28.377	28.263	10.513	10.426	10.103	9.989	3.10%	1.24%
67	AK-12	A Horizon	WW	20.038	30.618	30.341	28.03	27.923	10.58	10.303	7.992	7.885	22.43%	8.97%
68	AK-12	Sub Surface	AA	21.616	32.225	32.082	31.752	31.656	10.609	10.466	10.136	10.04	3.15%	1.26%
69	AK-21	A Horizon	UU	21.572	32.07	31.885	30.185	30.071	10.498	10.313	8.613	8.499	16.48%	6.59%
70	AK-21	Sub Surface	PO	17.598	28.008	27.871	27.561	27.47	10.41	10.273	9.963	9.872	3.02%	1.21%
71	AK-23	A Horizon	Y	17.511	28.141	28.018	27.144	27.078	10.63	10.507	9.633	9.567	8.32%	3.33%

Figure 3: A snippet of the data that was collected and analyzed from the aforementioned research projects.

The dataset was brought into a Google Colab environment using Python as the scripting language. Then the data was cleaned and columns were removed that were simply markers of tasks to be done on the samples, as they are not important to analysis of the data. For example, one column header read “A Horizon/Sub-surface Dried”, denoting simply if the sample had been dried or not. Furthermore, in the data cleaning process it was found that in the original Google spreadsheet, a few cell formulas had not been applied to certain rows. This error was rectified to eliminate the missing values. Finally, three rows that had complete missing values were dropped, as those samples were not able to be collected correctly in the field. In doing so, all sample points that contained missing values were eliminated.

Furthermore, it should be noted that bulk density is a measurement that can only be taken in the A horizon of a soil profile and therefore there is missing data for those bulk density related variables in the subsurface horizon in the complete dataset. An example soil profile is shown in figure 4.

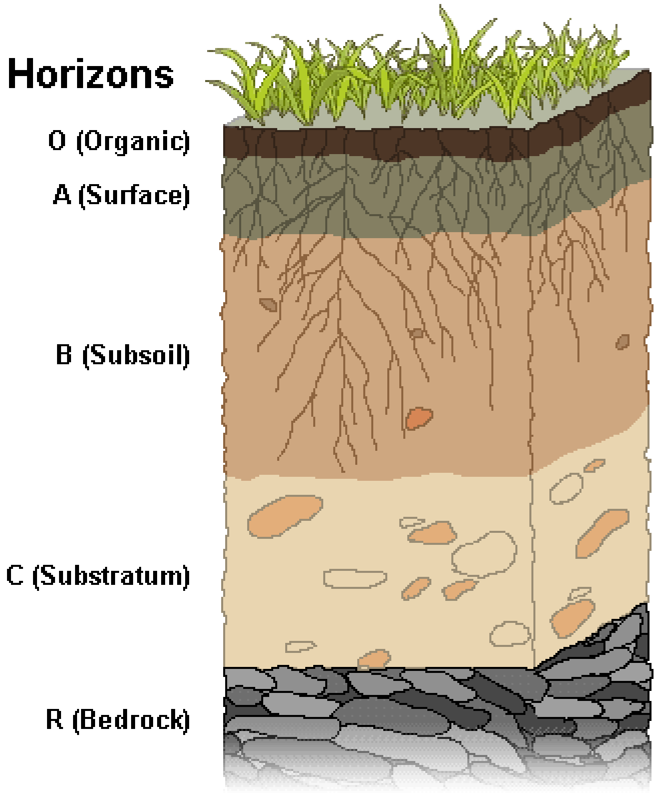


Figure 4: The A Horizon is the top most layer of the soil profile (top 10 cm). The subsurface is anything below the A Horizon. Image from Courses.lumenlearning.com.

Therefore, we broke the data into two subsets for analysis. One containing the sampling points with bulk density measurement (only A horizon measurements), and the other containing the remaining data (only subsurface measurements). The analysis conducted included a five number summary creation for all variables, as well as initial correlation matrix creation. This was done for both of the datasets, the bulk density dataset with values and the other dataset without any bulk density data. Furthermore, histograms were created of the SOC content in order to examine the distribution of the data. The resulting histograms are shown in Figure 5 and Figure 6 below.

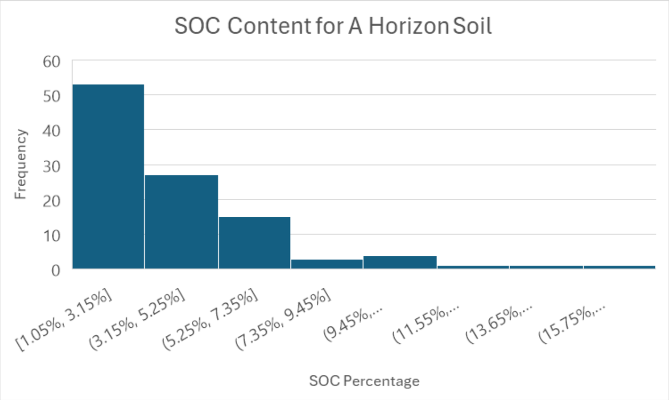


Figure 5: The SOC content distribution throughout the A-horizon.

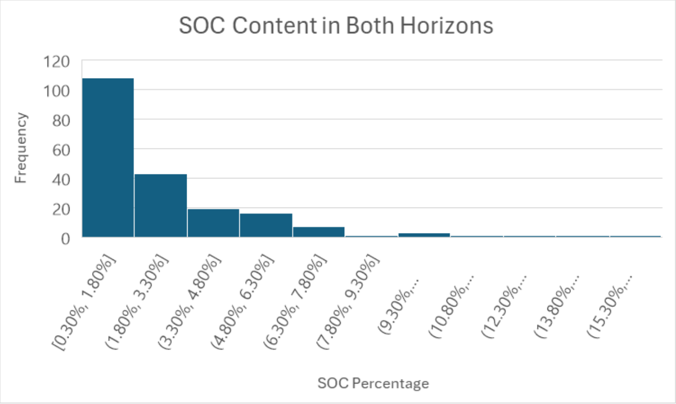
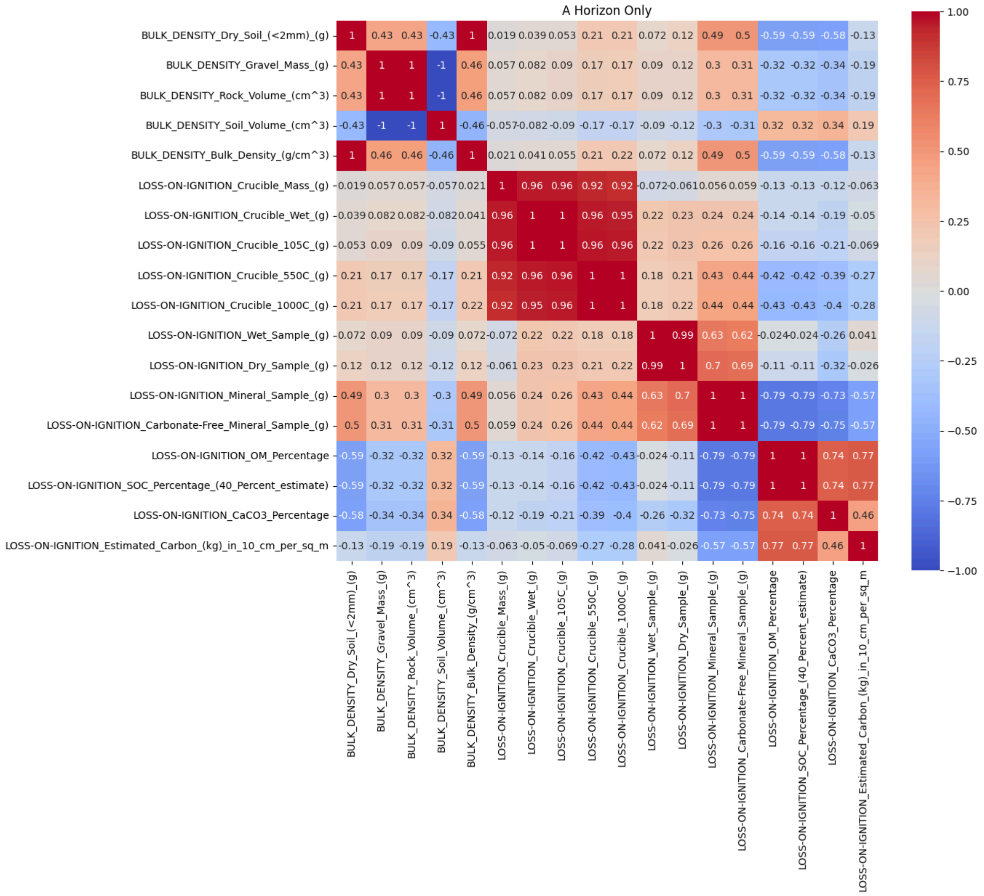


Figure 6: SOC content distribution throughout both horizons.

Our data is heavily skewed to the right, which presents challenges when modeling. Our team attempted to take the logarithm of our data to make a relatively normal distribution. However, that transformation showed negative values, which does not make logical sense when it comes to SOC content. As a result, we decided to use the data that is heavily skewed.

Correlation matrices were made for both of the data subsets. The correlation matrix for the A-horizon in Figure 7, showed several negative correlations, indicating that as one variable increases, the other decreases. This suggests an inverse relationship between these variables. Along the diagonal, there is a high number of large positive correlations. These are in fact red herrings, and are inherently correlated. This is because some of the variables are mathematical conversions of each other. Therefore, since those parameters are simply being mathematically transformed, there will be autocorrelation that should be ignored. The correlation matrix of variables was then analyzed for the subsurface dataset, where similar patterns appeared, shown in Figure 8. Both correlation matrices were used to determine which features should be used in our predictive model. The best predictors for SOC content (Kg/10cm/m^2) are shown in Figure 9.



	BULK_DENSITY_Dry_Soil_(<2mm)_(g)	BULK_DENSITY_Gravel_Mass_(g)	BULK_DENSITY_Rock_Volume_(cm^3)	BULK_DENSITY_Soil_Volume_(cm^3)	BULK_DENSITY_Bulk_Density_(g/cm^3)	LOSS-ON-IGNITION_Crucible_Mass_(g)	LOSS-ON-IGNITION_Crucible_Wet_(g)	LOSS-ON-IGNITION_Crucible_105C_(g)	LOSS-ON-IGNITION_Crucible_550C_(g)	LOSS-ON-IGNITION_Crucible_1000C_(g)	LOSS-ON-IGNITION_Wet_Sample_(g)	LOSS-ON-IGNITION_Dry_Sample_(g)	LOSS-ON-IGNITION_Mineral_Sample_(g)	LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)	LOSS-ON-IGNITION_OM_Percentage	LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)	LOSS-ON-IGNITION_CaCO3_Percentage	LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m
BULK_DENSITY_Dry_Soil_(<2mm)_(g)	1	0.43	0.43	-0.43	1	0.019	0.039	0.053	0.21	0.21	0.072	0.12	0.49	0.5	-0.59	-0.59	-0.58	-0.13
BULK_DENSITY_Gravel_Mass_(g)	0.43	1	1	-1	0.46	0.057	0.082	0.09	0.17	0.17	0.09	0.12	0.3	0.31	-0.32	-0.32	-0.34	-0.19
BULK_DENSITY_Rock_Volume_(cm^3)	0.43	1	1	-1	0.46	0.057	0.082	0.09	0.17	0.17	0.09	0.12	0.3	0.31	-0.32	-0.32	-0.34	-0.19
BULK_DENSITY_Soil_Volume_(cm^3)	-0.43	-1	-1	1	-0.46	-0.057	-0.082	-0.09	-0.17	-0.17	-0.09	-0.12	-0.3	-0.31	0.32	0.32	0.34	0.19
BULK_DENSITY_Bulk_Density_(g/cm^3)	1	0.46	0.46	-0.46	1	0.021	0.041	0.055	0.21	0.22	0.072	0.12	0.49	0.5	-0.59	-0.59	-0.58	-0.13
LOSS-ON-IGNITION_Crucible_Mass_(g)	0.019	0.057	0.057	-0.057	0.021	1	0.96	0.96	0.92	0.92	-0.072	-0.061	0.056	0.059	-0.13	-0.13	-0.12	-0.063
LOSS-ON-IGNITION_Crucible_Wet_(g)	0.039	0.082	0.082	-0.082	0.041	0.96	1	1	0.96	0.95	0.22	0.23	0.24	0.24	-0.14	-0.14	-0.19	-0.05
LOSS-ON-IGNITION_Crucible_105C_(g)	0.053	0.09	0.09	-0.09	0.055	0.96	1	1	0.96	0.96	0.22	0.23	0.26	0.26	-0.16	-0.16	-0.21	-0.069
LOSS-ON-IGNITION_Crucible_550C_(g)	0.21	0.17	0.17	-0.17	0.21	0.92	0.96	0.96	1	1	0.18	0.21	0.43	0.44	-0.42	-0.42	-0.39	-0.27
LOSS-ON-IGNITION_Crucible_1000C_(g)	0.21	0.17	0.17	-0.17	0.22	0.92	0.95	0.96	1	1	0.18	0.22	0.44	0.44	-0.43	-0.43	-0.4	-0.28
LOSS-ON-IGNITION_Wet_Sample_(g)	0.072	0.09	0.09	-0.09	0.072	-0.072	0.22	0.22	0.18	0.18	1	0.99	0.63	0.62	-0.024	-0.024	-0.26	0.041
LOSS-ON-IGNITION_Dry_Sample_(g)	0.12	0.12	0.12	-0.12	0.12	-0.061	0.23	0.23	0.21	0.22	0.99	1	0.7	0.69	-0.11	-0.11	-0.32	-0.026
LOSS-ON-IGNITION_Mineral_Sample_(g)	0.49	0.3	0.3	-0.3	0.49	0.056	0.24	0.26	0.43	0.44	0.63	0.7	1	1	-0.79	-0.79	-0.73	-0.57
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)	0.5	0.31	0.31	-0.31	0.5	0.059	0.24	0.26	0.44	0.44	0.62	0.69	1	1	-0.79	-0.79	-0.75	-0.57
LOSS-ON-IGNITION_OM_Percentage	-0.59	-0.32	-0.32	0.32	-0.59	-0.13	-0.14	-0.16	-0.42	-0.43	-0.024	-0.11	-0.79	-0.79	1	1	0.74	0.77
LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)	-0.59	-0.32	-0.32	0.32	-0.59	-0.13	-0.14	-0.16	-0.42	-0.43	-0.024	-0.11	-0.79	-0.79	1	1	0.74	0.77
LOSS-ON-IGNITION_CaCO3_Percentage	-0.58	-0.34	-0.34	0.34	-0.58	-0.12	-0.19	-0.21	-0.39	-0.4	-0.26	-0.32	-0.73	-0.75	0.74	0.74	1	0.46
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m	-0.13	-0.19	-0.19	0.19	-0.13	-0.063	-0.05	-0.069	-0.27	-0.28	0.041	-0.026	-0.57	-0.57	0.77	0.77	0.46	1

Figure 7: A correlation matrix of all of the variables that are present in our data spreadsheet, but only from soil types that were collected from the A horizon.

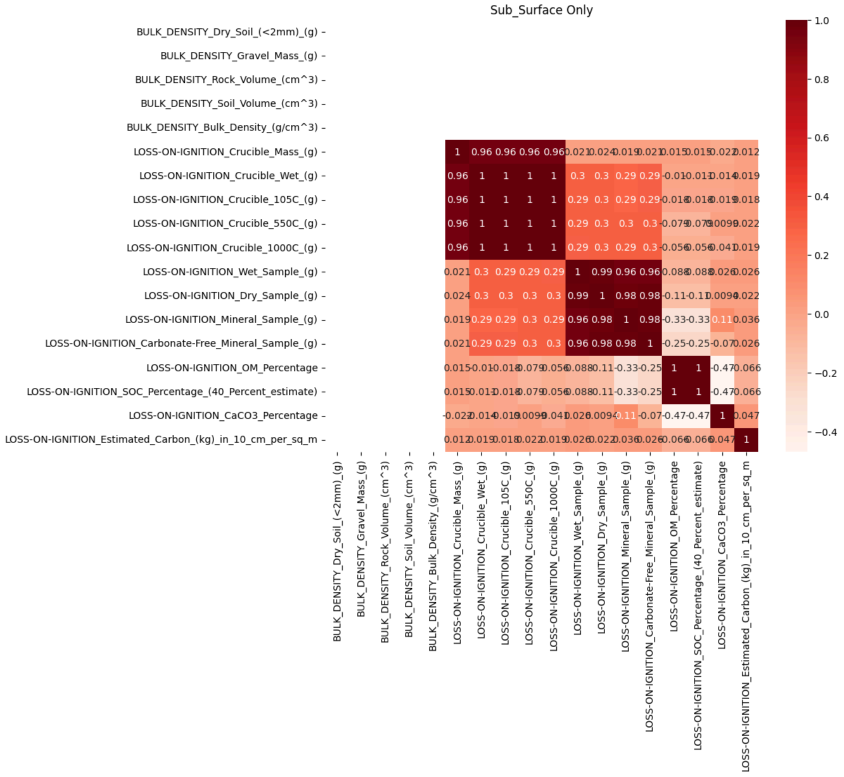


Figure 8: Correlation heatmap of all of the variables but only from subsurface soil samples.

Predictors	**Correlation Coefficient**
Loss-on-ignition Dry Sample (g)	-0.026
Loss-on-ignition Mineral Sample (g)	-0.57
Loss-on-ignition Carbonate Free Mineral Sample (g)	-0.57
Loss-on-ignition OM%	0.77
Loss-on-ignition SOC%	0.77
Loss-on-ignition CaCO3%	0.46

Figure 9: The correlation coefficients for our best predictors.

For this project, it was unclear which models would be most effective, therefore we created and tested as many methods as possible with the tools we learned in Python. Having the

RMSE (Root Mean Squared Error) and the R^2 as our main indicators to measure if the model was effective or not with our data. We originally used estimated carbon in kg/10cm depth/m^2 as our dependent variable. However, after further data analysis, it was found that SOC content percentage was a better variable to predict.

The libraries used for this project were some of the most used in typical data science practices. The main tools used were:

Library	Pandas	Numpy	Math	Seaborn/ Matplotlib	Sklearn
Purpose	Data management	Data cleaning/analysis	Mathematical operations	Graphics	Machine learning creation/testing

Figure 10: Python libraries used in the Google Colab.

At first, we used the linear regression model, as it has a relatively straightforward interpretation and can identify variables that might have predictive powers for soil characteristics. Then we tried the LASSO Regression model with multiple variations of alpha values. We used a range from 0.01 to 0.3 to find what variation works the best. The lower alpha values seemed to give better results in our data. This model is closely related to linear regression, resulting in outputs that are similar. Consequently, both models are similarly limited when the data does not exhibit a linear relationship.

Afterward, we used models that fit better for more complex patterns, like the Decision Tree and the random forest. The key difference between linear and non-linear relationships lies in how they handle data patterns. Linear modeling performs well only when the data follows a straight-line pattern. In contrast, non-linear models are better suited for handling complex data with diverse patterns. They achieve this by making a series of sequential decisions based on feature values, offering greater flexibility in identifying complex relationships between variables.

Results

SOC percentage modeling results:

Model	**RMSE**	**R^2**
Linear Regression	1.2794	0.7559
LASSO Regression	1.3014	0.7474
Decision Tree	0.8222	0.8991
Random Forest	0.9138	0.8731

Figure 11: The different modeling techniques used and the resulting RMSE and R^2.

The linear regression model explains 75.5% of the variance in the estimated soil organic carbon percentage, with a root mean squared error (RMSE) of 1.2794. About 74.7% of the variance in the estimated soil organic carbon percentage is explained by the LASSO regression model, with an RMSE of 1.3014. The decision tree model explained 89.9% of the variance, showing a significant improvement in R^2 compared to linear and LASSO regression. However, this came with a higher risk of overfitting, as evidenced by its RMSE of 0.8222. The random forest model emerged as the best-performing approach, explaining 87.3% of the variance in the estimated soil organic carbon mass with an RMSE of 0.9138.

Following the best model results, we wanted to visualize the residuals of the Random Forest model. The following scatter plot and histogram show the patterns of the residuals.

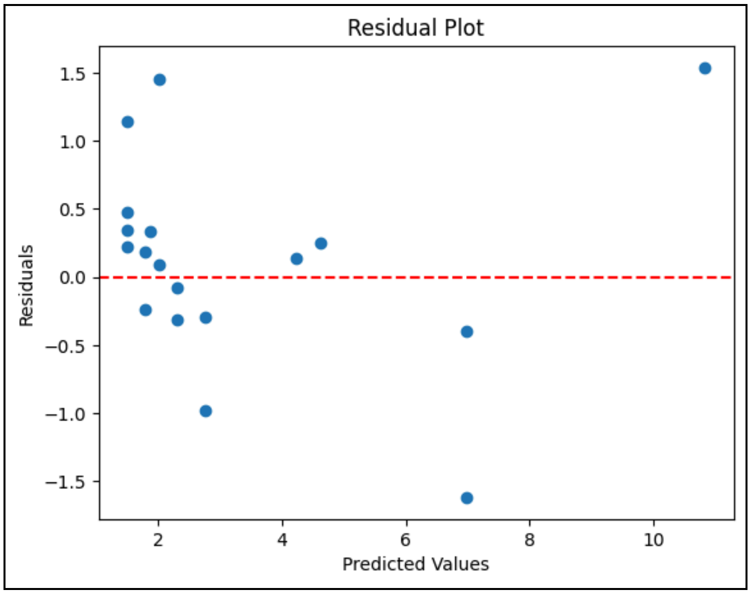


Figure 12: Plotted residuals of the random forest model plotted.

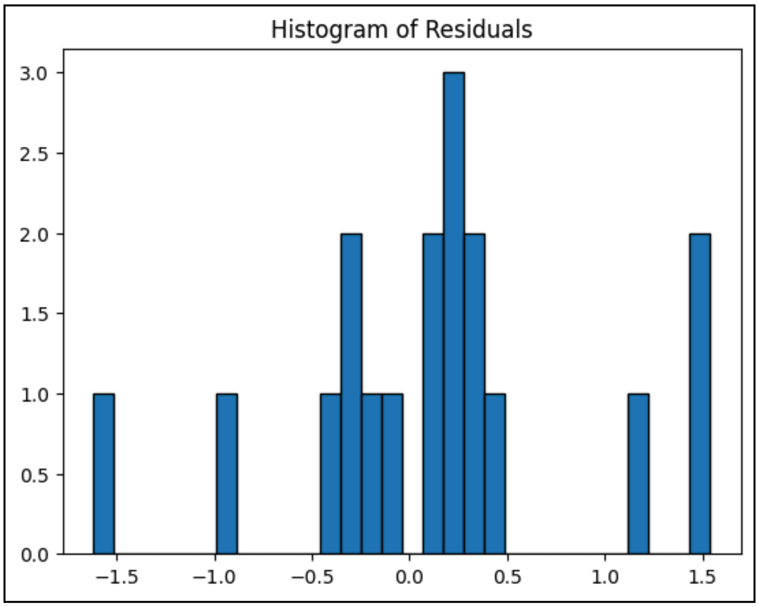


Figure 13: Histogram showing the distribution of the residuals of the Random Forest model.

Since our model is machine learning, our team also created a histogram showing the success and failure of our predictions within the testing subset.

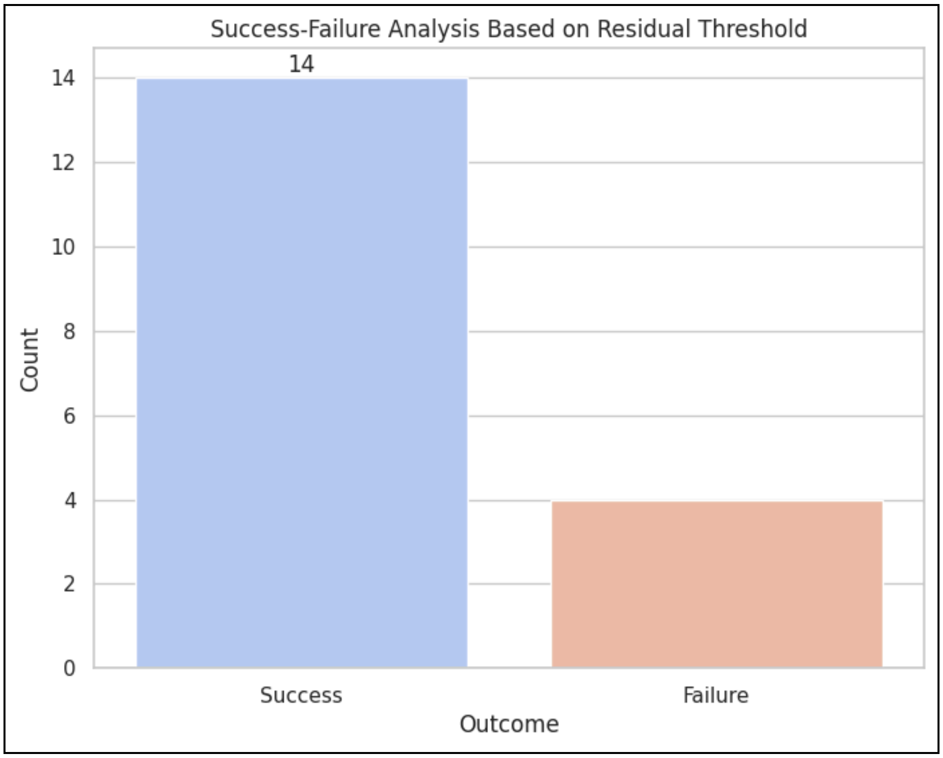


Figure 14: Histogram showing the success/failure performance of the model predictions in the test data set.

Discussion

The linear regression model provided a baseline for understanding the relationship between the predictor variables and the target variable. While the result was significant, in that it suggested room for improvement in terms of predictive power. The LASSO regression result was comparable to the linear regression model, as LASSO adds regularization to prevent overfitting and simplify the model. While LASSO showed promise in maintaining a balance between complexity and interpretability, its performance was slightly less accurate compared to linear regression.

The decision trees can capture nonlinear patterns effectively, their tendency to overfit without sufficient depth cut off meant they struggled to generalize well to new data though so that is not ideal to use as a predictive model. By averaging the results of multiple decision trees, the random forest model mitigated the overfitting tendency of single decision trees and captured complex relationships in the data. This ensemble method effectively balanced variance reduction and generalization, outperforming all other models in terms of both explanatory power and error reduction.

While all models contributed to understanding the relationship between soil composition indicators and soil organic carbon percentage, this project shifted more towards the random forest model to utilize after observing its ability to mitigate overfitting. This change improved the balance between predictive accuracy and model strength. Its ability to handle complex interactions and reduce overfitting made it the most reliable choice for modeling this dataset.

Conclusion

This project analyzed soil samples from the Dale B. Engquist Nature Preserve to investigate the predictors of soil organic carbon content. We used various parameters such as multiple regression models, linear regression, LASSO, decision trees, and random forest models. These were all applied to determine how effective they are in predicting the SOC levels. The random forest model was the most effective, because it explained 87.31% of the variances in the SOC and has an RMSE of 0.9138. This model was among the most accurate and demonstrated superior handling of overfitting compared to some of the simpler models. The performance of the random forest model can help farmers and conservationists in properly managing soil health and understanding carbon levels. Accurate predictions of SOC help land management agencies be able to contribute to carbon solutions with climate change mitigation strategies. Special consideration should be taken in future projects to include a wide range of soil data. Care must also be taken when selecting samples and interpreting models. If our dataset lacks diversity, the predictors may not be generalized. Future researchers must make sure that a large range of soil data is used to minimize bias and make the data more reliable.

Works Cited

9.1 Soil Profiles & Processes | Environmental Biology. n.d. Courses.lumenlearning.com. https://courses.lumenlearning.com/suny-environmentalbiology/chapter/9-1-soil-profiles-processes/.

Cutting, Benjamin J., Clement Atzberger, Asa Gholizadeh, David A. Robinson, Jorge Mendoza-Ulloa, and Belen Marti-Cardona. "Remote Quantification of Soil Organic Carbon: Role of Topography in the Intra-Field Distribution." Remote Sensing 16, no. 9 (May 1, 2024): 1510. doi:10.3390/rs16091510.

Ellis, Elizabeth, and Keith Paustian. "Importance of On-Farm Research for Validating Process-Based Models of Climate-Smart Agriculture." *Carbon Balance & Management* 19, no. 1 (May 29, 2024): 1–12. doi:10.1186/s13021-024-00260-6.

Purdue University Agricultural Communication. Soil organic matter matters. Indiana State Climate Office [Internet]. c2021 [cited 2024 Dec 2]. Available from: https://ag.purdue.edu/indiana-state-climate/research/farming-for-a-better-climate/f4abc-soil-organic-matter/

Weil, Ray R, and Nyle C Brady. Elements of the Nature and Properties of Soils. New York, Ny Pearson. 2019.