Data Analysis of Environmental Parameters of Soil Rough Draft

Data-151: Intro to Data Science

11/13/2024

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

Introduction

This project analyzed data from the previous research projects: *Environmental Restoration Target Estimation Around Engquist Nature Preserve* and *Assessing Carbon Sequestration Potential of Cropland Conversion in Porter County, IN*. These research projects collected data at 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), Doc Janowiak (recent alumnus), and Justin Self (recent alumnus). This research is also supported by Dr. Jon-Paul McCool of the Geography Department. Also of note is that this research is still ongoing this semester. These past projects collected soil samples from different environment types at properties owned by SHLT and the samples were subsequently analyzed for various soil parameters. The different environment types included prairie, forested wetland, forest, floodplain forested wetland, and agriculture. Environmental parameters that were specifically analyzed were bulk density, gravel mass, rock volume, soil volume, wet weight, dry weight, mineral sample, organic matter percent, soil organic carbon (SOC) percent, calcium carbonate percent, estimated carbon in horizon per sq. m. of soil, and estimated carbon in 10 cm depth per sq m. of soil.

This project for DATA-151 explored healthy ranges for environmental variables for agriculture and then compared that to our existing dataset. It is important to develop a prediction for one or more variables within this dataset due to the cost and time to run all the tests. 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 soil organic carbon (SOC) content in Indiana soils varies depending on factors such as land management, soil type, and climate. 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, like a linear regression model and ANOVA.

The project is important because of SOC's necessity to agriculture and climate change, specifically global warming. In terms of agriculture, SOC helps farmers make informed decisions about their crops and soil management. The more knowledgeable farmers are about their soil health, the faster they can identify nutrient deficiencies or imbalances, which can be addressed before planting crops. 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 2023, 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 Benjamin J. Cutting et la. has also used models like a random forest to try and address 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.

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 1.

Figure 1.

	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%

Fig. 1: 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 2.

Figure 2.

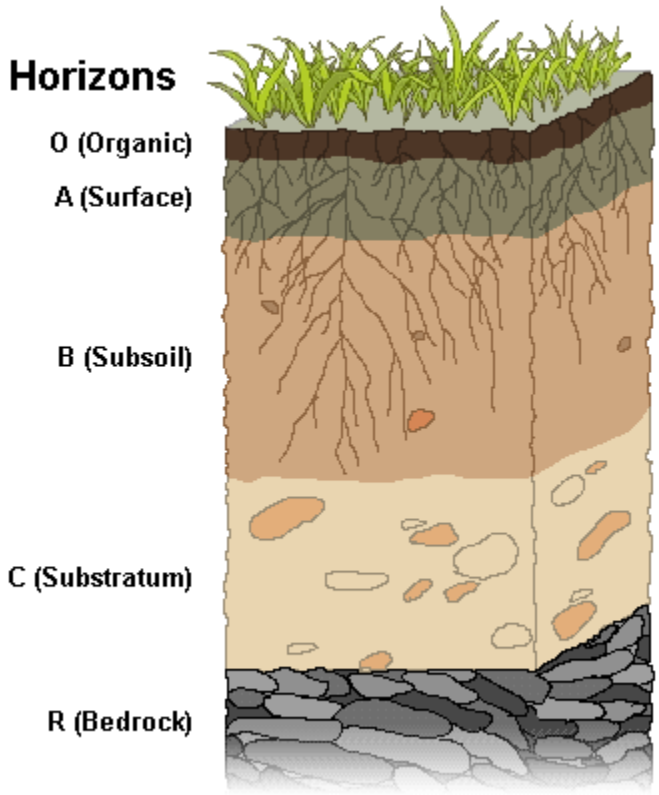

Fig. 2: The A Horizon is the top most layer of the soil profile (top 10 cm). The subsurface is anything below the A Horizon. Image from Courses.lumenlearning.com.

Therefore, we broke the data into two subsets for analysis. One containing the sampling points (only A horizon measurements) with bulk density measurement, and the other containing the remaining data. 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.

The five number summary has shown that the *bulk density for dry soil under 2 mm* has a large range, while most of the other variables are closer in range of 22 to 36 and 0 to 15 as a max value. Looking at the first quartile to the third quartile for many of the variables, the numbers are close together. This shows that the spread of the data is not very large so there is less variation and there is not a skewed distribution.

From the correlation matrix, there are several negative correlations, meaning that as one variable increases, the other decreases, indicating these variables move in inverse directions. Along the diagonal, there is a high number of 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 dataset including bulk density, and the dataset that does not. ANOVA analysis was conducted with the highly correlated variables to see which specific categories impact the soil on being suitable for agriculture.

For this project, the effectiveness of the models were not clear. We decided to take an exploratory approach and create and test as many methods as possible with the tools we learned in Python. Having the MSE (Mean Squared Error) and the R squared (percentage of variation in a dependent variable that can be explained by an independent variable) as our main indicators to measure if the model was effective or not with our data.

All of the models tested were targeted to predict the estimated soil organic carbon mass ['LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m'], based on the other variables that gave better results in the correlation matrix.
['LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)',

'LOSS-ON-IGNITION_Wet_Sample_(g)', 'LOSS-ON-IGNITION_Mineral_Sample_(g)']. Those variables are the most adequate for the modeling because they were the ones that had the closest absolute correlated percentages with the estimated soil organic carbon mass. Consequently, we pick the 3 variables whose abs() is closer to 1. But not counting the ones that have 0.98< because those relations that are very close to or are 1 are variations of the same variable. Or variables that were created based on the other one, which is why their variability would always be 100%. variables with direct correlation are useful for our modeling purposes.

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. Its performance was surprisingly adequate, as it resulted in:

MSE: 0.29663

R^2: 0.76398

The linear regression model explains 76% of the variance in the estimated soil organic carbon mass. This result was satisfactory in terms of statistical significance, but we knew there is a big margin of improvement. So we decided to try with other models as well.

The second model tested was a decision tree model, we tried the two main approaches for it, as we tried without max depth, and with a controlled maximum depth. Both models did not give any satisfactory results.

Results

The best predictors for SOC content ($Kg/10cm/m^2$) (e.g. they have the highest correlation) are:

Predictors	**R^2 Value**
Loss-on-ignition Mineral Sample (g)	-0.6
Loss-on-ignition Crucible 105 C	-0.33
Loss-on-ignition Crucible 550 C	-0.33

These R^2 values were collected from our heatmap displayed in Figure 3.

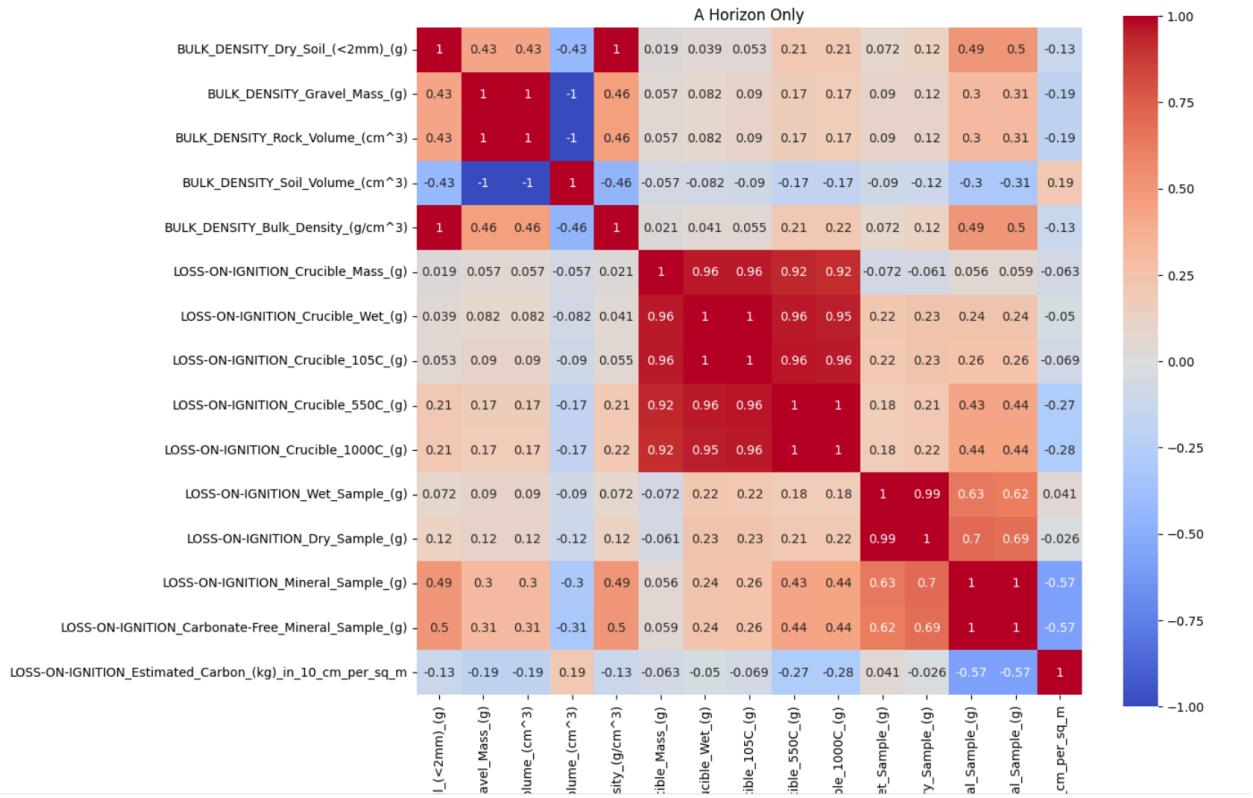

Figure 3: A correlation heatmap of all of the variables that are present in our data spreadsheet, but only from soil types that were collected from the A horizon.

These correlation results are not great, but using the best subsets selection method, we were able to find some combinations of these variables that have a better correlation. Figure 4

shows that 68.79% of the variances in SOC content can be explained by the Loss-on-ignition Crucible 105 C and Loss-on-ignition Crucible 550 C variables because the adjusted R^2 was .6879. Additionally, 8.64% of the variance in SOC content can be explained by the Loss-on-ignition Mineral Sample, Loss-on-ignition Crucible 105 C, and the Loss-on-ignition Crucible 550 C because the adjusted R^2 was .6864.

Regression Statistics	
Multiple R	0.831302
R Square	0.691063
Adjusted R	0.687926
Standard E	0.962723
Observatic	200

Regression Statistics	
Multiple R	0.831325
R Square	0.691101
Adjusted R	0.686373
Standard E	0.965116
Observatic	200

Figure 4: Various results of R^2 categories from the regression tests performed in Excel.

By doing best subsets selections, we learned what combinations of the three best predictors for SOC content have the highest correlation coefficient. After running best subsets selection in excel, we are able to determine that the combination of Loss-on-ignition Crucible 105 C and Loss-on-ignition Crucible 550 C, as well as the combination of Loss-on-ignition Mineral Sample, Loss-on-ignition Crucible 105 C, and the Loss-on-ignition Crucible 550 C gave by far the best correlation values (adj R^2).

To continue, from Figure 3 we learned that very few categories of soil in the A horizon are highly correlated with each other when comparing them one by one. The few that do have a strong correlation, are mostly because it is just comparing two of the same or similar categories. Most of the categories have weak correlations with one another, which means that in order to find combinations with stronger correlations, we need to run tests like the best subsets selection

that are capable of doing regression analysis with multiple explanatory variables at once. This can result in a possibility of having completely different correlation results than all of the explanatory variables tested individuals versus all together.

Furthermore, from Figure 5 we learned that a lot of the categories of soil located in the SubSurface are very similar to one another, and have very strong correlations that have an R^2 of either 1 or very close to it.

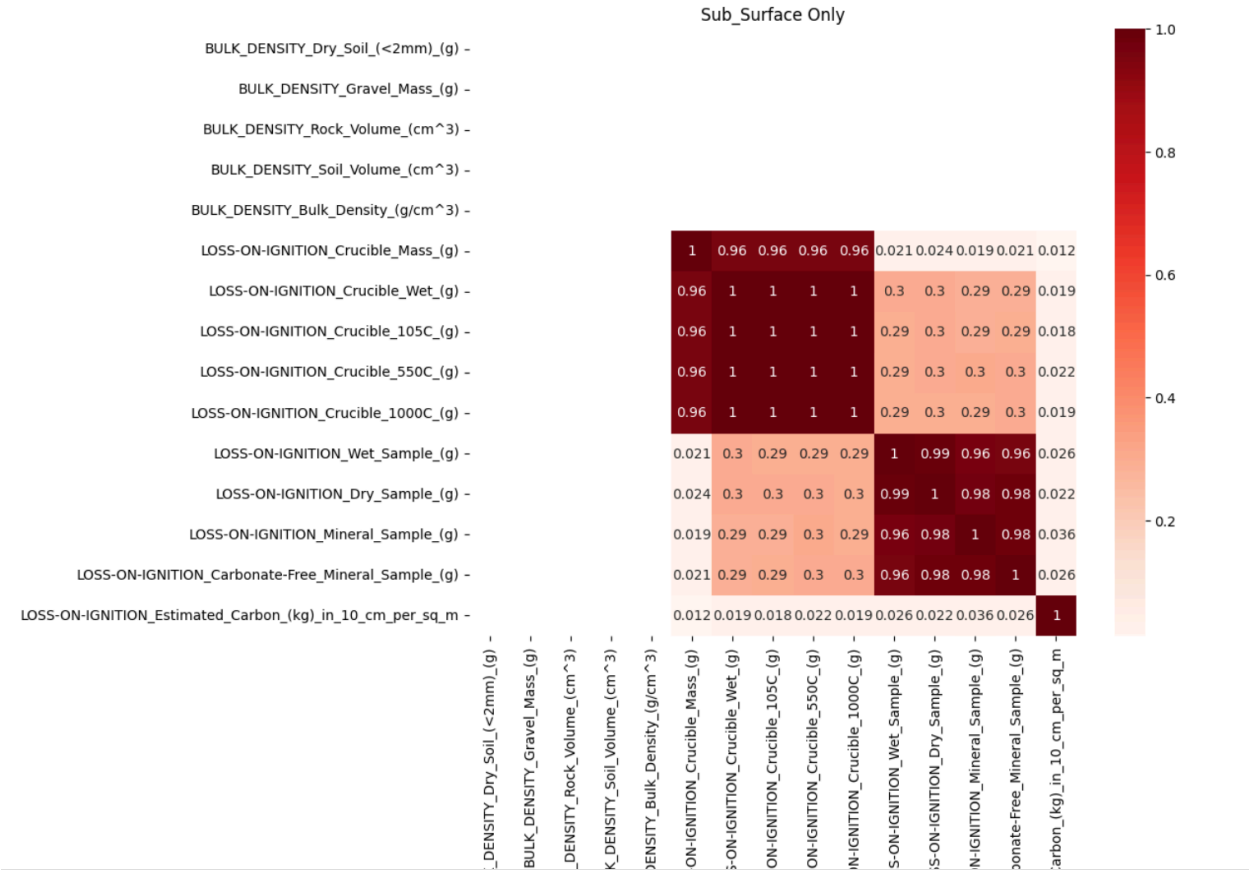

Figure 5: Correlation heatmap of all of the variables that are present in our data spreadsheet, but only from soil types that were collected from the Sub Surface.

Because there are multiple individual variables that have very strong correlations, no further tests are needed for most variables to see if we can find additional combinations that have

high R^2 values. However, SOC content (Kg/10cm/m^2) is the variables that we want to predict, and there are no explanatory variables that have a higher R^2 value than 0.03 for soil located in the SubSurface. As a result, we can conclude that there is no correlation between any of the explanatory variables and SOC content (Kg/10cm/m^2), and that there is no need to a best subsets selection test because that requires the variables to have at least some correlation on the individual level.

Conclusion

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.