Whiterock Project

STATCOM ISU

November 17, 2024

1 Introduction

Soil aggregation, formed by the combination of the soil's physical, chemical, and biological properties, is an important physical property necessary for the movement of air, water, and nutrients in the soil. The soil aggregation is directly linked with important properties such as the water holding capacity of the soil. In the experiment conducted by the Whiterock Conservancy, A Wet Aggregate Stability test is done for analyzing macroaggregates (> 0.25 mm), microaggregates (< 0.25 mm and > 0.053 mm) and calculating total wet aggregate stability as a percentage of total soil. We want to answer the following questions –

- How are we supposed to interpret the aggregate stability test results? Do the percentages of macro, micro, and total aggregates reflect anything meaningful about soil health at each point in time, or can we only look at trends over time?
- What's a good model for data analysis/interpretation that we can replicate going forward?
- Are there any trends over time for each sampling point? Or is it too early to tell?
- Can we compare aggregate stability results between points, or are results going to differ by soil type and site conditions? If we can compare them between points, what are the results?

2 Data Description

The Aggregate Stability dataset is an aggregation of critical data points, organized into two main sections: Metadata and Soil Aggregate Measurements. Metadata provides the backbone of the dataset with 'Sample_Type', 'Lab_No', 'Test_ID' and 'Cust_ID' indicating the unique identifier for the laboratory and test, 'Date_Recd' and 'Date_Rept' capturing the date samples arrived and reported at the lab, 'Name', 'Company' and 'Grower' reflecting the individual or entity responsible for cultivating the land, and 'Field_ID' and 'Sample_ID_1' offering geographic and sample-specific identification, respectively. The Soil Aggregate Measurements delve into the soil's physical condition with 'Latitude', 'Longitude', 'Beginning_Depth', and 'Ending_Depth' giving the location coordinates and depth of the soil sample, 'Macroaggregates' and 'Microaggregates' providing percentages that help gauge soil stability and aeration capacity, while 'Total_Aggregates' gives a holistic view of the soil's aggregate stability. Among them, the selection process for eliminating redundant variables, aimed at refining the dataset for robust analysis, will be elaborated on in the following content. The data, enriched by the Personnel factor of the 'Grower', granularity of the 'Sample_ID_1' and spatial specificity of 'Field_ID', alongside the temporal insights offered by 'Date_Recd', will provide a robust framework for understanding the multifaceted nature of soil health.

3 Exploratory Data Analysis - Data Visualization

Boxplots are invaluable in exploratory data analysis for their succinct representation of distributional characteristics and their ability to highlight outliers. They visually summarize the central tendency, variability, and skewness of the data with their five-number summary: the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. By looking at a boxplot, we can quickly see if the data is leaning more towards higher or lower values and spot any unusual points that don't fit the pattern. This makes it easy to compare different sets of data at a glance. For this data set, we use independent variables 'Date_Recd', 'Grower', 'Sample_ID_1', 'Field_ID_1' and dependent variables Macroaggregates, Microaggregates, Total Aggregates, a total of 12 box plots are constructed to analyze the relation between them.

3.1 Date_Recd vs. Aggregates

The boxplot comparing Aggregates to Date_Recd demonstrates a distinct upward trend in aggregate content as time progresses. The median value of total aggregates visibly increases with each subsequent sampling date. This indicates that, on average, the soil samples taken at later dates contain a higher concentration of total aggregates. Moreover, the interquartile range (IQR) widens over the sampled dates. This suggests that not only is the overall concentration of aggregates increasing, but the variation within the samples is also becoming more pronounced.

3.2 Grower vs. Aggregates

WHITEROCK CONSERVANCY shows a higher median and a narrower IQR of Aggregates compared to COON RAPIDS AG, so the relationship revealed by the Grower vs. Aggregates boxplot suggests that the soil management practices or environmental conditions at WHITEROCK CONSERVANCY may be more conducive to forming soil aggregates than those at COON RAPIDS AG, resulting in a higher and more consistent aggregate content.

3.3 Sample_ID_1 vs. Aggregates

The boxplot for Aggregates versus Sample_ID_1 showcases a variety of distributions across different sample IDs, indicating heterogeneity in soil aggregate content among the samples. Some sample IDs, such as the one corresponding to the pink box, display high median values and a narrow interquartile range, suggesting a consistently higher aggregate content with less variability within those samples. Conversely, other sample IDs, like the one represented by the red box at the bottom, have significantly lower median values and a more compact interquartile range, indicating consistently lower aggregate content. The relationship between Sample_ID_1 and Aggregates is complex, as evidenced by the varying median values, range widths, and outlier presence. This suggests that soil aggregate content can be highly localized and influenced by factors unique to each sample's origin within the field.

3.4 Field_ID_1 vs. Aggregates

The boxplot depicting Aggregates across different Field_IDs indicates varying soil aggregate content among the fields. The highest median aggregate content is found in samples from PATRICK FARM, suggesting richer soil aggregate conditions in this field. The spread of data, as indicated by the interquartile ranges, varies significantly between fields. Some, like HOWE WEST, have a wide range, suggesting a high variability in soil aggregate content within the field. Overall, the boxplot illustrates that soil aggregate content is not only field-specific but also varies within fields, highlighting the complexity of soil structure as influenced by location.

4 Brief Discussion about Redundant Variables

In order to perform analysis based on linear model and ANOVA (Analysis of Variance) tests, it is necessary to identify an injective relation between combination of independent variables (date, company, sample ID, etc) and each observational unit, which means variables containing redundant information should be removed.

From this data, all observational units are labeled with same value of "sample Type", "Beginning Depth", and "Ending Depth", which indicates these variables do not have an influence on the value of dependent variables (Macroaggregates, Microaggregates, and Total Aggregates), thus, should be removed from further analysis. In addition, bijective relations exist among groups of variables, including ('Date Recd', 'Date Rept'), ('Name', 'Company", 'Growers'), and ('Field ID', 'Latitude', 'Longitude'), implying variables in same group contain same information. Thus, choosing one variable from each group is enough for further analysis without losing any information.

Based on these kinds of selection, we generate the conclusion that combinations of variables “Date Recd”, “Field ID”, and “Sample ID” presents an injective relation with the set of all observational units, that is, the characteristics of each observational units can be uniquely represented by a specific value of (“Date Recd”, “Field ID”, and “Sample ID”). Therefore, independent variables “Date Recd”, “Field ID”, and “Sample ID”, together with dependent variables “Macroaggregates”, “Microaggregates”, and “Total Aggregates”, should be selected for further analysis using a linear model and ANOVA.

5 Linear Model and Analysis of Variance (ANOVA)

We use Date, Field, and Sample ID to build a linear model and conduct the ANOVA test. The ANOVA (Analysis of Variance) table is used to analyze the differences among group means and their associated procedures. Based on the tables 13–15, the following conclusions can be drawn:

- The ‘Date_Recd’ has p-values $\Pr(F > F_{\text{obs}})$ close to zero for each of the responses, which suggests that the dates have a statistically significant effect on Macroaggregates, Microaggregates and Total Aggregates. This also means that ‘Grower’ is a significant factor as it has a one-to-one correspondence with ‘Date_Recd’.

- The ‘Field_ID’ (combined effect for all except Dial Farm Baseline) is also close to zero for each of the responses, which suggests that the dates have a statistically significant effect on Macroaggregates, Microaggregates and Total Aggregates.

- The ‘Sample_ID’ (combined effect for all except ‘Sample_ID_1103PS’) has a p-value $\Pr(F > F_{\text{obs}})$ less than 0.05 only for the Microaggregates. So, ’Sample_ID’ is a significant factor only for the Microaggregates.

- The combined effect of ‘Field_ID_Dial_Farm_Baseline’ and ‘Sample_ID_1103PS’ is significant for the Macro and Microagrregates.

6 Decision Tree Model

The decision trees provide insights into the factors that are most predictive of soil health in terms of aggregate stability. They illustrate the importance of specific locations, grower practices, and time points in influencing soil structure. By analyzing the leaves and splits in each tree, we can infer that soil health is not only site-specific but also affected by management practices and temporal changes.

The decision tree model (Figure 13) predicts “Macroaggregates” based on key dependent variables. It identifies “grower” and “date recorded” as significant factors influencing “Macro aggregates” values. Samples attributed to “WHITEROCK CONSERVANCY” generally exhibit lower “Macroaggregates” values. Additionally, samples collected in November show lower “Macroaggregates” than those from May and June in both years.

The decision tree model (Figure 14) was used to predict “Microaggregates” based on previously mentioned dependent variables. It identifies “Field ID,” “Grower,” and “Sample ID” as significant factors influencing “Microaggregates” values. Samples from “BONNEYLAND WEST” tend to exhibit lower “Micro aggregates.” Specifically, among samples from “BONNEYLAND WEST,” those with “Grower” labeled as “WHITEROCK CONSERVANCY” and “Sample ID” labeled as “1101W” show lower “Microaggregates” values compared to other samples.

The decision tree model (Figure 15), tailored to predict “Totalaggregates” based on earlier mentioned dependent variables, highlights “Sample ID” and “Field ID” as significant factors shaping “Totalaggregates” values. Results indicate that samples labeled with “Sample ID” as “1102NW” and “1101NE” tend to have lower “Totalaggregates.” Furthermore, samples associated with “Field ID” such as “FOX FARM,” “HERRON FARM,” and “HOWE EAST 4A” also exhibit relatively lower “Totalaggregates” values compared to other samples.

[1]

7 Conclusion

This report delves into exploring the relations between various factors and the stability of soil aggregation, aiming to address the four proposed questions addressed in the introduction. Following a comprehensive analysis based on graphs and statistical models, the ensuing conclusions are drawn:

- In conclusion, the aggregate stability test results reveal that as time progresses, an increase in macroaggregates is observed, while microaggregates experience a decrease, with the total aggregates showing little to no significant change. These changes are significant for assessing soil health, indicating shifts in soil structure that can impact its functionality.

- A suitable model for ongoing data analysis and interpretation of soil aggregate stability would encompass both linear model (ANOVA) and Decision Tree. ANOVA is adept at quantifying the impact of different factors on soil properties, helping to ascertain changes in macro and microaggregates over time, across various fields, and under different management practices, are statistically significant. On the other hand, Decision Tree analysis excels in elucidating the complex interplay between these factors, offering a visual and intuitive method to decipher the multifaceted influences on soil aggregate stability.

- Due to the limited data available for each sample ID, it is challenging to determine trends over time for each sample point conclusively. However, by examining the overall trends across all samples, utilizing plots that compare "Sample ID" vs. "aggregates" and "Date Recd" vs. "Aggregates," we can glean some understanding of the broader patterns and changes in soil aggregate stability across the sampled area.

- 'PATRICK FARM', 'NIKES & REDS', 'DIAL FARM WEST', 'DIAL FARM EAST', 'DIAL FARM BASELINE', 'BONNEYLAND WEST', 'BONEEYLAND EAST', these seven fields have higher value of total aggregates than others, which indicates that these fields have better soil aggregate stability. Moreover, these seven fields mentioned above and the remaining five fields show a significant difference in 'Total Aggregates', fueled by the difference in the values of "Microaggregates" between these two groups.

[1] Note: * in Table 13-15 means all the levels of the factor except the combined effect of 'Field_ID_Dial_Farm_Baseline' and 'Sample_ID_1103PS'

Figure 1: Box Plot for Macroaggregates vs Date_Recd

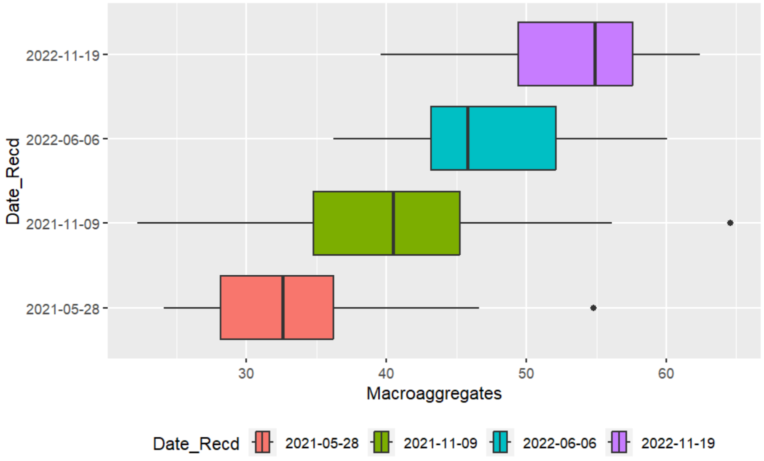


Figure 2: Box Plot for Microaggregates vs Date_Recd

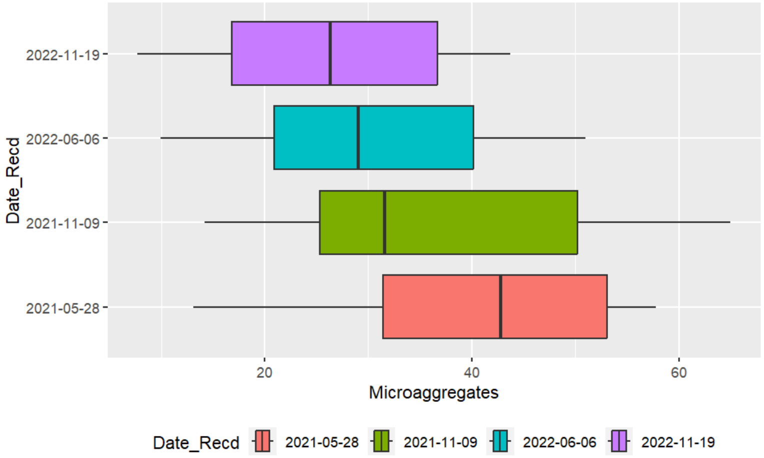


Figure 3: Box Plot for Total Aggregates vs Date_Recd

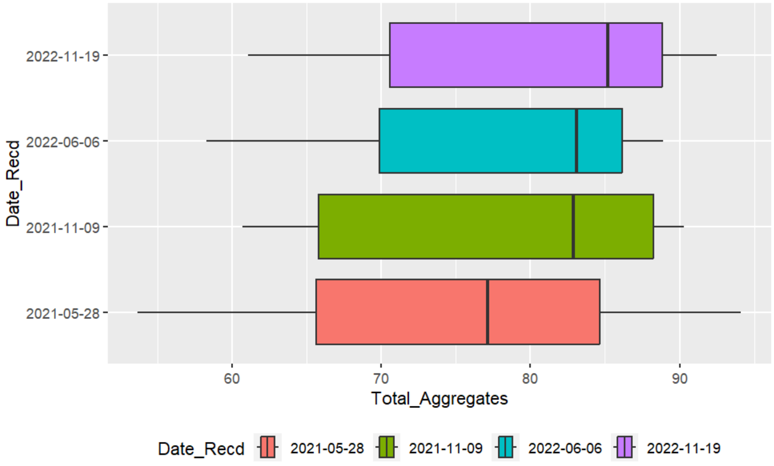


Figure 4: Box Plot for Macroaggregates vs Grower

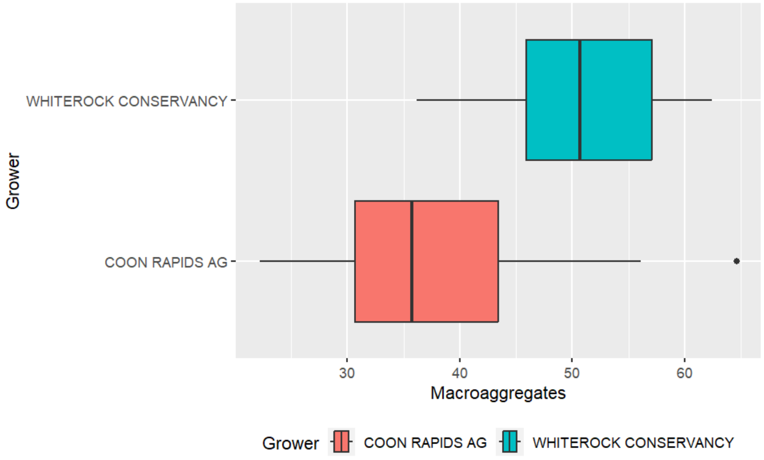


Figure 5: Box Plot for Microaggregates vs Grower

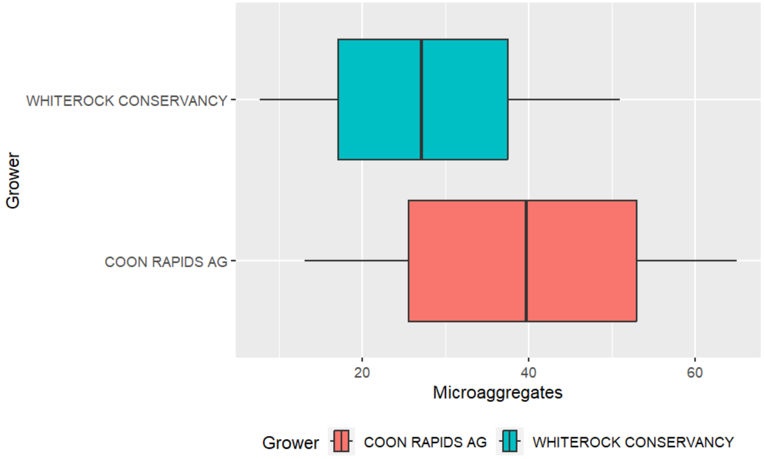


Figure 6: Box Plot for Total Aggregates vs Grower

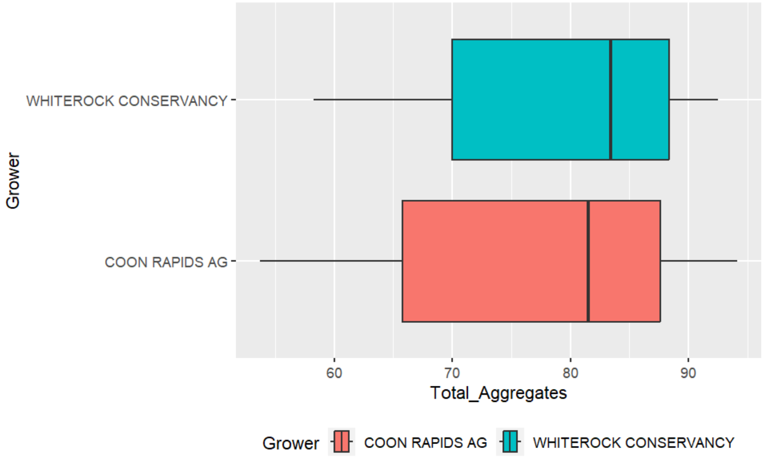


Figure 7: Box Plot for Macroaggregates vs Sample_ID_1

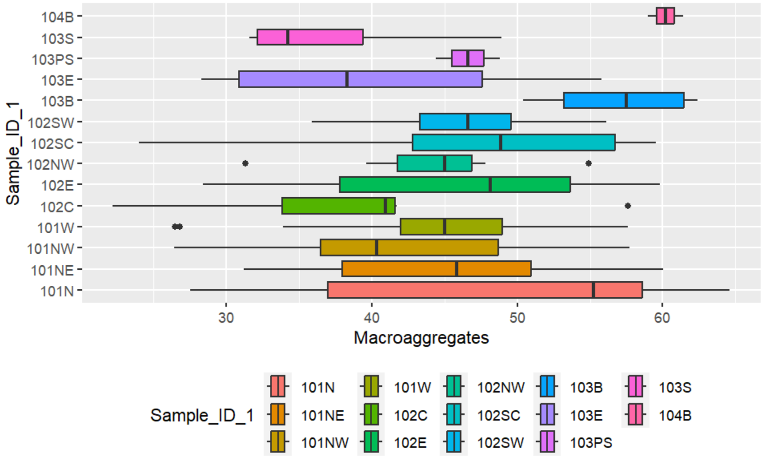


Figure 8: Box Plot for Microaggregates vs Sample_ID_1

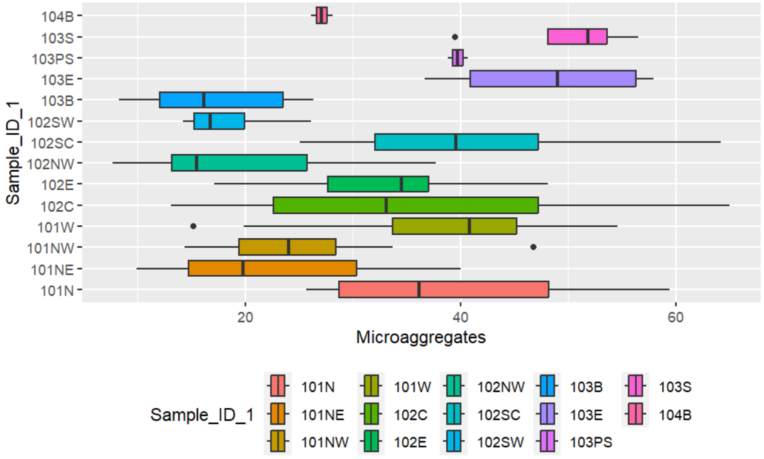


Figure 9: Box Plot for Total Aggregates vs Sample_ID_1

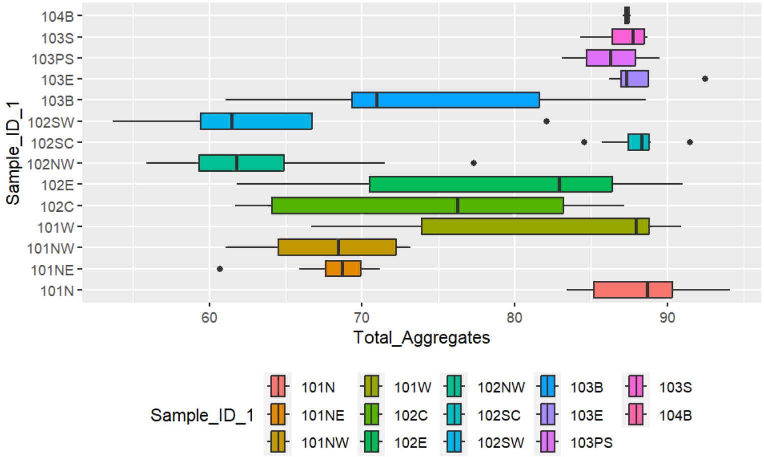


Figure 10: Box Plot for Macroaggregates vs Field_ID_1

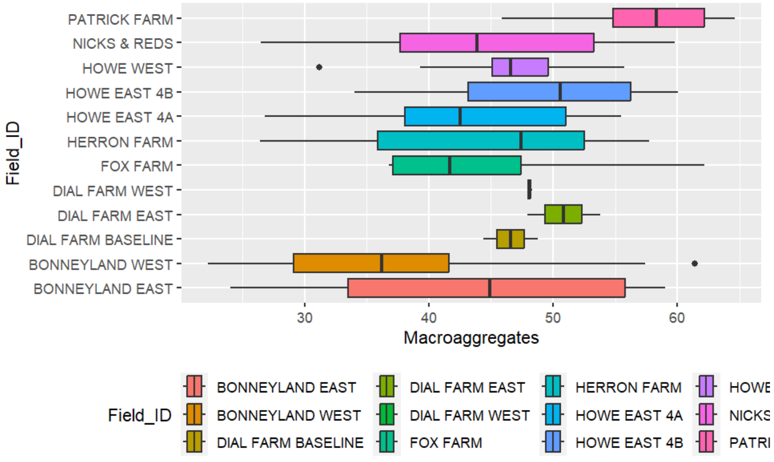


Figure 11: Box Plot for Microaggregates vs Field_ID_1

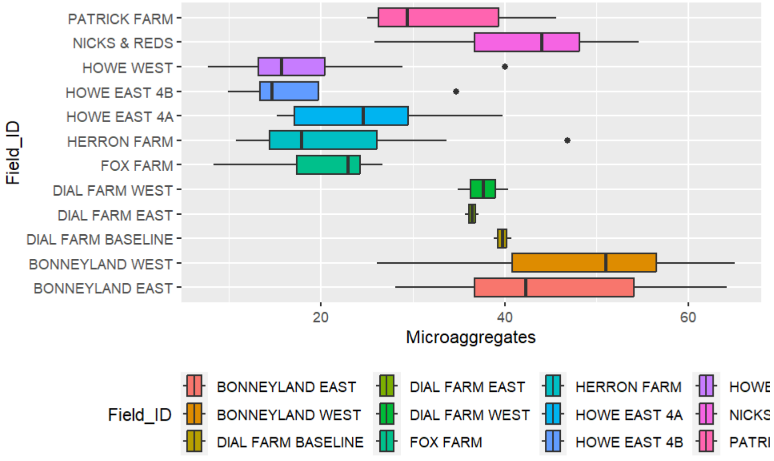


Figure 12: Box Plot for Total Aggregates vs Field_ID_1

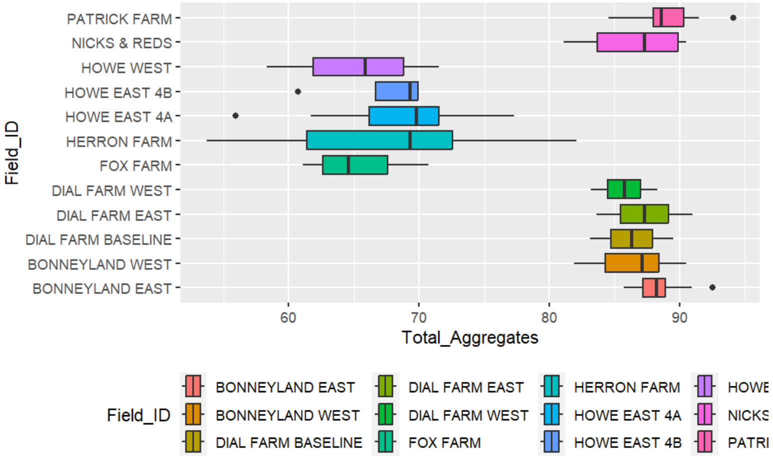
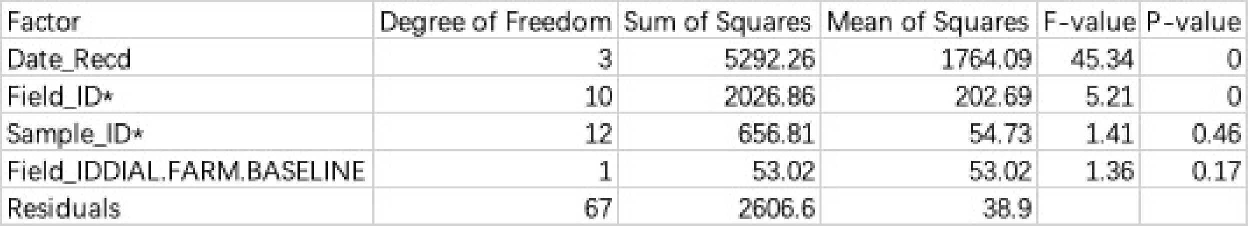
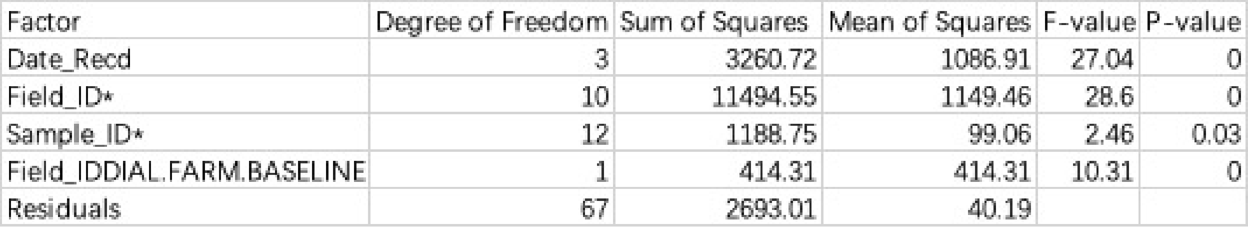


Figure 13: ANOVA Table for Macroaggregates



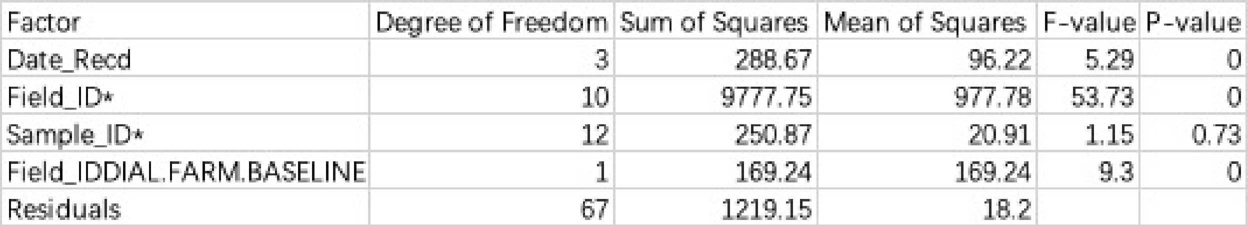
Factor	Degree of Freedom	Sum of Squares	Mean of Squares	F-value	P-value
Date_Recd	3	5292.26	1764.09	45.34	0
Field_ID*	10	2026.86	202.69	5.21	0
Sample_ID*	12	656.81	54.73	1.41	0.46
Field_IDDIAL.FARM.BASELINE	1	53.02	53.02	1.36	0.17
Residuals	67	2606.6	38.9		

Figure 14: ANOVA Table for Microaggregates



Factor	Degree of Freedom	Sum of Squares	Mean of Squares	F-value	P-value
Date_Recd	3	3260.72	1086.91	27.04	0
Field_ID*	10	11494.55	1149.46	28.6	0
Sample_ID*	12	1188.75	99.06	2.46	0.03
Field_IDDIAL.FARM.BASELINE	1	414.31	414.31	10.31	0
Residuals	67	2693.01	40.19		

Figure 15: ANOVA Table for Total Aggregates



Factor	Degree of Freedom	Sum of Squares	Mean of Squares	F-value	P-value
Date_Recd	3	288.67	96.22	5.29	0
Field_ID*	10	9777.75	977.78	53.73	0
Sample_ID*	12	250.87	20.91	1.15	0.73
Field_IDDIAL.FARM.BASELINE	1	169.24	169.24	9.3	0
Residuals	67	1219.15	18.2		

Figure 16: Decision Tree for Macroaggregates

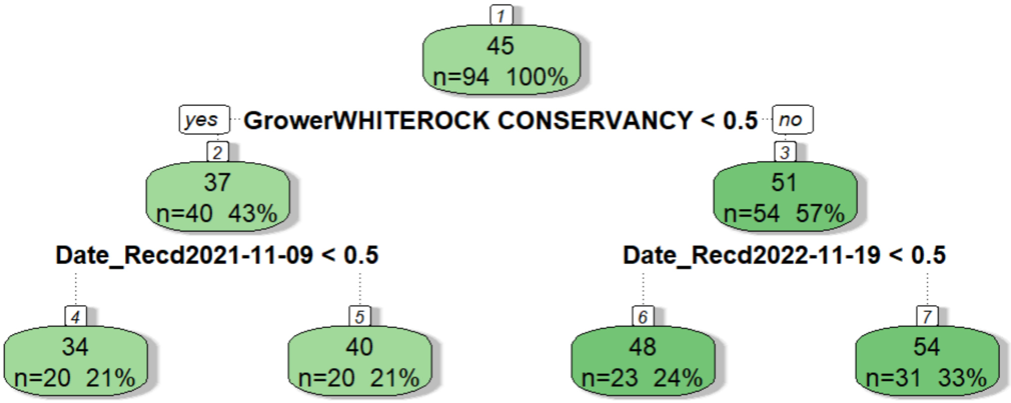

1
45
n=94 100%
yes
GrowerWHITEROCK CONSERVANCY < 0.5
no
2
37
n=40 43%
Date_Recd2021-11-09 < 0.5
3
51
n=54 57%
Date_Recd2022-11-19 < 0.5
4
34
n=20 21%
5
40
n=20 21%
6
48
n=23 24%
7
54
n=31 33%

Figure 17: Decision Tree for Microaggregates

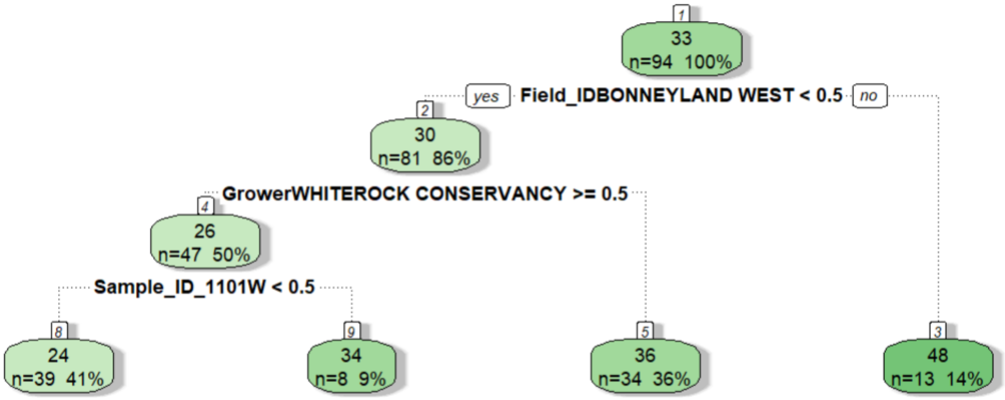
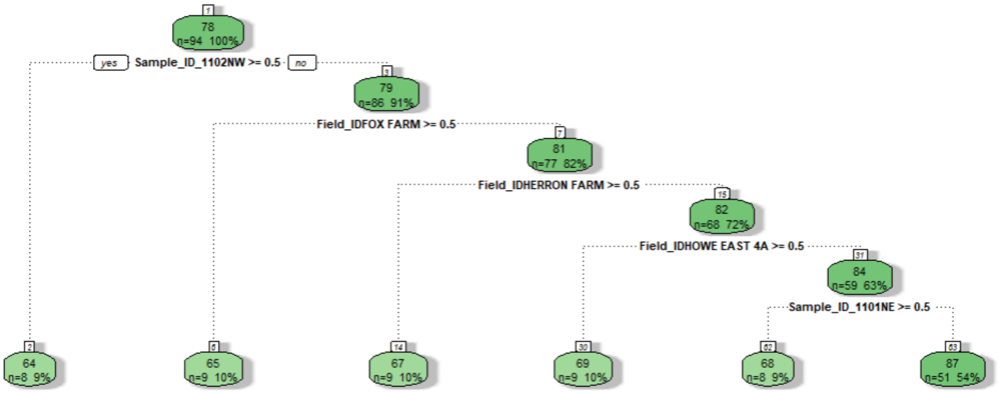


Figure 18: Decision Tree for Total Aggregates


1
78
n=94 100%
yes
Sample_ID_1102NW >= 0.5
no
3
79
n=86 91%
Field_IDFOX FARM >= 0.5
7
81
n=77 82%
Field_IDHERRON FARM >= 0.5
15
82
n=68 72%
Field_IDHOWE EAST 4A >= 0.5
31
84
n=59 63%
Sample_ID_1101NE >= 0.5
2
64
n=8 9%
6
65
n=9 10%
14
67
n=9 10%
30
69
n=9 10%
62
68
n=8 9%
63
87
n=51 54%