Individual Project Preliminary Analyses

Hannah Rodgers

4/27/2022

Contents

This R script is used to clean up and transform soil data to perform regression analysis.

Overview of Goals

My research focuses on how soil health and soil microbiology influence long-term sustainability in semiarid wheat systems. Wheat is one of the most important crops worldwide, yet the semiarid landscapes where it is primarily grown are particularly vulnerable to climate change and land degradation (Asseng et al. 2015). In particular, the traditional wheat-fallow system practiced throughout the High Plains ecoregion inefficiently stores water, depletes soil fertility, and has a high potential for erosion, pushing farmers and researchers to search for economical ways to build long-term soil health (Norton, Mukhwana, and Norton

2012; Kaur et al. 2015). Soil health is critical for sustainable and resilient food production, yet evaluating soil health in semiarid climates remains challenging. In particular, soil microorganisms are sensitive indicators of changing soil health, and can help evaluate the long-term sustainability of land management practices (Rodgers, Norton, and Diepen 2021).

Soil data was collected from an experiment on the long-term impacts of compost application on soil health and soil microbiology in Wyoming High Plains organic wheat systems. Specifically, four rates of compost were applied to 128 small plots in a randomized complete block design in either 2016 or 2021. In contrast to fresh manure or chemical fertilizers, composted manure is made of concentrated, stabilized nutrients and carbon that release slowly over time and improve soil physical properties (Larney et al. 2006). Data was collected on soil physical and chemical properties (such as aggregate stability and bulk density), total and labile organic matter pools, and soil microbiology (enzyme activity, microbial biomass, and microbial community composition). Regression analysis will be used to evaluate the relationships between compost rate and soil properties. **Therefore, this project focuses on cleaning the data, testing the regression assumptions on each variable, transforming the data to meet those assumptions, and finally visualizing the data. Ultimately, quantifying soil health benefits over many years could guide efforts to protect soils and create agricultural systems resiliant in the face of climate change and land degradation.**

Peer Review Comments

In the overview, I included a table of contents and bolded my goals statement to make them stand out. In the workflow, I copied code wherever I ran it a second time, and made sure to print the p-values I was interested in to demonstrate my decision-making process. I also greatly improved my data visualization section. I overlayed the regressions for the 2016 and 2020 data on the same plots to make them easier to compare, and used faceting to display the how several different variables all responded to compost. I cleaned up my graph layout using patchwork, combining the legends, titles, and x axis labels into one. Finally, I included a description of the variables I plotted, and a short conclusion of my findings.

Data Tidying

Load Packages

```r
library(MASS)
library(readxl)
library(writexl)
library(car)
library(patchwork)
library(tidyverse)
```

Import Data

```r
#read in data
OREI_soil <- read_excel("OREI_05.2021.xlsx", sheet = "Soil_Data")
OREI_treatments <- read_excel("OREI_05.2021.xlsx", sheet = "Treatment_Data_2021")
OREI_enzymes <- read_excel("OREI_05.2021.xlsx", sheet = "Enzymes")
OREI_PLFA <- read_excel("OREI_05.2021.xlsx", sheet = "PLFAs")
```

Clean Up Data

Merge the datasets, select the data I need, and separate into two groups by year of compost application.

```
#merge the data
OREI_all <- OREI_treatments %>%
  left_join(OREI_soil, by = 'Sample_ID') %>%
  left_join(OREI_enzymes, by = 'Sample_ID') %>%
  left_join(OREI_PLFA, by = 'Sample_ID')

#remove plots that are fallow or fertilized
OREI_all <- OREI_all %>%
  filter(Rotation == "wheat", Treatment != "fertilizer") %>%

#remove unwanted variables
 dplyr::select(-PER1, -ID, -InorganicC, -Treatment, -Compost_Rate,
         -Crop, -Rotation, -H2O, -BulkDensity)

#separate into two groups by year of compost application
OREI_2016 <- subset(OREI_all, Compost_Year != "2020")
OREI_2020 <- subset(OREI_all, Compost_Year != "2016")
```

Remove Outliers

Outliers can skew a regression, so remove any outliers from each variable using the outlierKB function.

```
#I ran this function on each variable in both OREI_2016 and OREI_2020 to
#remove outliers.
source("http://goo.gl/UUyEzD")
  #outlierKD(OREI_2020, total_bacteria)

#save no_outlier data
  #write_xlsx(OREI_2020, "OREI_2020_no_outliers.xlsx")
  #write_xlsx(OREI_2016, "OREI_2016_no_outliers.xlsx")

#read in no_outlier data to continue
OREI_2020 <- as.data.frame(read_excel("OREI_2020_no_outliers.xlsx"))
OREI_2016 <- as.data.frame(read_excel("OREI_2016_no_outliers.xlsx"))
```

I removed outliers from:
OREI_2016: *yield, NO3, PMN, DON, MBC, CBH, PHOS, NAG, actino, gram_pos, AMF, sapro_fungi, total_MB, total_fungi* OREI_2020: *NO3, protein, MBC, MBN, SOC, N, NAG, BX, AG, SUL*

Check Regression Assumptions

The four main assumptions of regression analysis are:

1. Observations are independent. (The samples come from randomized plots, so they are already independent.)
2. Normality: The residuals are normally distributed.
3. Linearity: The relationship between X and Y is linear.
4. Homoscedasticity: The residuals have constant variance for all values of X.

2. Check normality of residuals.

```
#Create an empty list
Shapiro.pvals.2016 <- list()

#This loop runs a linear model on compost ~ each variable,
#evaluates normality with the Shapiro test, and saves the p value.
for (i in names(OREI_2016[,7:38])) {
  mod <- lm(get(i) ~ compost, data = OREI_2016)
  Shapiro.pvals.2016[[i]] <- (shapiro.test(mod$residuals))$p.value }

#now same thing for OREI_2020
Shapiro.pvals.2020 <- list()
for (i in names(OREI_2020[,7:38])) {
  mod <- lm(get(i) ~ compost, data = OREI_2020)
  Shapiro.pvals.2020[[i]] <- (shapiro.test(mod$residuals))$p.value }

#print variables with p < 0.05
as.data.frame(t(as.data.frame(Shapiro.pvals.2016))) %>%
  filter(V1 < 0.05)
```

```
##              V1
## PHOS 0.02234545
```

```
as.data.frame(t(as.data.frame(Shapiro.pvals.2020))) %>%
  filter(V1 < 0.05)
```

```
##              V1
## DOC  0.03051935
## DON  0.03714962
## PHOS 0.03320653
```

Any variables with $p < 0.05$ have non-normal residuals and need to be transformed!

Transform the non-normal variables.

```
#This function performs a box_cox transformation on a variable
box_cox_transform <- function(v) {

#calculates the boxcox plot and pulls out lambda
  bc <- boxcox(v ~ compost, data = OREI_2016)
  lambda <- bc$x[which.max(bc$y)]

#transforms the data using lambda and saves it
  v <- (v^lambda-1)/lambda
  return(v) }

#run box_cox_transform on all non-normal variables and save in OREI_year_t
OREI_2016_t <- OREI_2016
OREI_2020_t <- OREI_2020
```

```
OREI_2016_t$PHOS <- box_cox_transform(OREI_2016$PHOS)
```

```
OREI_2020_t$PHOS <- box_cox_transform(OREI_2020$PHOS)
```

```
OREI_2020_t$DOC <- box_cox_transform(OREI_2020$DOC)
```

```
OREI_2020_t$DON <- box_cox_transform(OREI_2020$DON)
```

```
#test normality of new data using the same code from above
Shapiro.pvals.2016 <- list()
for (i in names(OREI_2016_t[,7:38])) {
  mod <- lm(get(i) ~ compost, data = OREI_2016_t)
  Shapiro.pvals.2016[[i]] <- (shapiro.test(mod$residuals))$p.value }

Shapiro.pvals.2020 <- list()
for (i in names(OREI_2020_t[,7:38])) {
  mod <- lm(get(i) ~ compost, data = OREI_2020_t)
  Shapiro.pvals.2020[[i]] <- (shapiro.test(mod$residuals))$p.value }

as.data.frame(t(as.data.frame(Shapiro.pvals.2016))) %>%
  filter(V1 < 0.05)
```

```
## [1] V1
## <0 rows> (or 0-length row.names)
```

```
as.data.frame(t(as.data.frame(Shapiro.pvals.2020))) %>%
  filter(V1 < 0.05)
```

```
##             V1
## DON 0.02258343
```

All the variables are normal except 2020 DON- I'll have to figure that one out later.

3. Check linearity with the F-test.

```
#This function creates a linear and quadratic model for each variable,
#tests whether the two models differ using ANOVA, and saves the p.value
F_test <- function(x) {
  mod <- lm(x ~ compost, data = OREI_2016_t)
  reduced<-lm(x ~ compost, data = OREI_2016_t)
  full<-lm(x ~ poly(compost,2), data = OREI_2016_t)
  return(anova(reduced, full)$"Pr(>F)") }

#run this function on each variable and save the p.values in a list
F.test.2016 <- list()
for (i in colnames(OREI_2016_t[,7:38])) {
  F.test.2016[[i]] <- F_test(OREI_2016_t[[i]]) }
```

```
F.test.2020 <- list()
for (i in colnames(OREI_2020_t[,7:38])) {
  F.test.2020[[i]] <- F_test(OREI_2020[[i]]) }

#print variables with p < 0.05
as.data.frame(t(as.data.frame(F.test.2016))) %>%
  filter(V1 < 0.05)
```

```
## [1] V1 V2
## <0 rows> (or 0-length row.names)
```

```
as.data.frame(t(as.data.frame(F.test.2020))) %>%
  filter(V1 < 0.05)
```

```
## [1] V1 V2
## <0 rows> (or 0-length row.names)
```

All variables have a linear relationship with compost! (no p values < 0.05)

4. Check constancy of residuals with the Levene test.

```
#Run the levene test (from car package) on each variable and save the p values
levene.2016 <- list()
for (i in colnames(OREI_2016_t[,7:38])) {
  result <- leveneTest((OREI_2016_t[[i]]) ~ as.factor(OREI_2016_t$compost))
  levene.2016[[i]] <- result$`Pr(>F)`[1] }

levene.2020 <- list()
for (i in colnames(OREI_2020_t[,7:38])) {
  result <- leveneTest((OREI_2020_t[[i]]) ~ as.factor(OREI_2020_t$compost))
  levene.2020[[i]] <- result$`Pr(>F)`[1] }

#print variables with p < 0.05
as.data.frame(t(as.data.frame(levene.2016))) %>%
  filter(V1 < 0.05)
```

```
## [1] V1
## <0 rows> (or 0-length row.names)
```

```
as.data.frame(t(as.data.frame(levene.2020))) %>%
  filter(V1 < 0.05)
```

```
## [1] V1
## <0 rows> (or 0-length row.names)
```

All data passes the levene test! (no p values < 0.05)

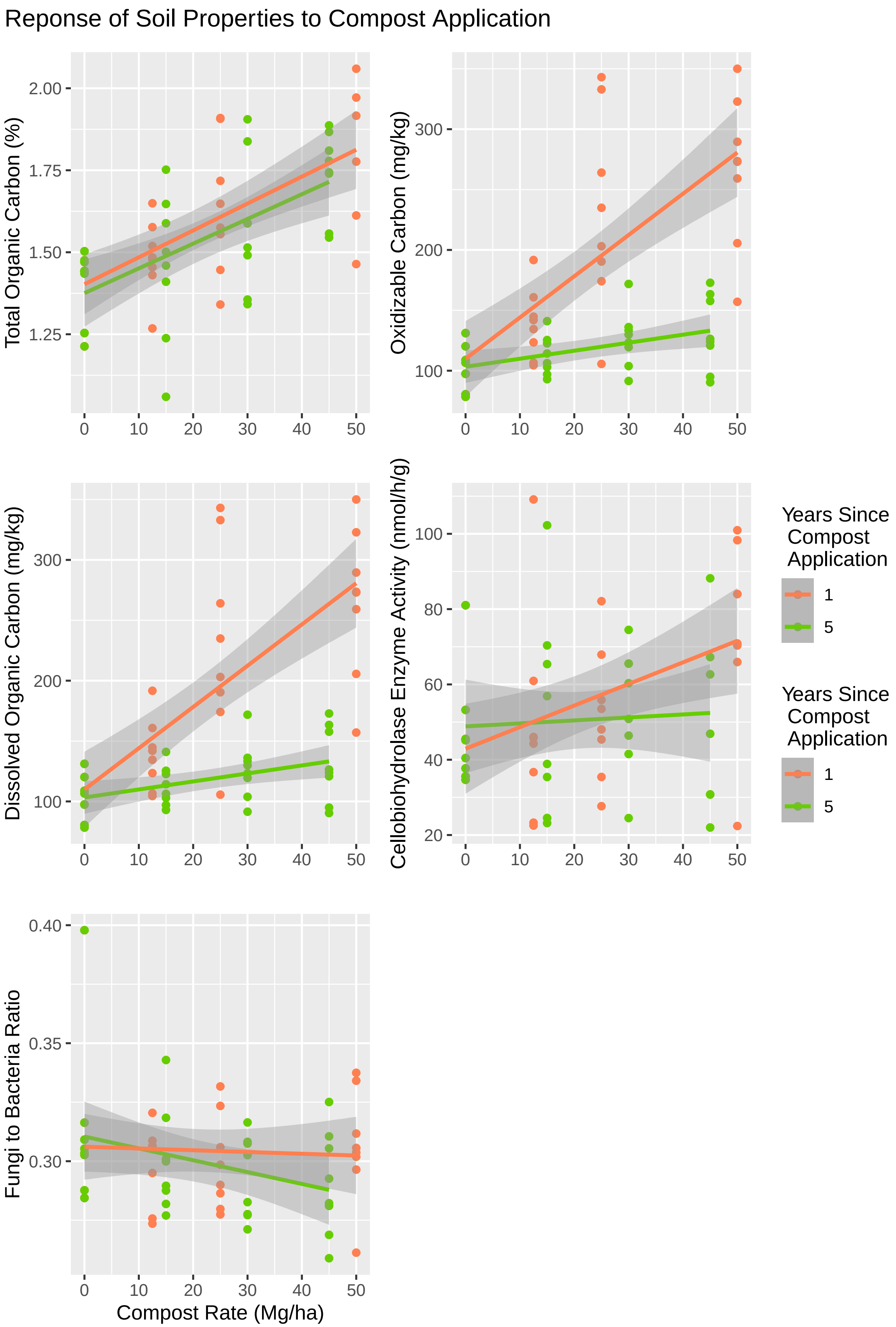
Visualize Data

Here, I visualize relationships between compost rate and a few different soil health variables. I use the untransformed data for plotting, but when I later add in statistics (p values and R^2), I'll use the transformed data that meets the assumptions of linear regression.

I'm visualizing a few key variables that will provide a picture of carbon cycling and microbial activity in these soils:

1. Total organic carbon
2. Dissolved organic carbon and permanganate oxidizable carbon. Both are labile carbon pools that serve as food sources for microorganisms in the soil.
3. Cellobiohydrolase activity, an enzyme produced by microorganisms that decomposes organic carbon (specifically, cellulos)e.
4. Fungi to bacteria ratio, which indicates changes in microbial community composition.

```
#I set the figure size in the chunk options

#Bind both dataframes, and select the data I want to graph.
#Include an ID column that tells me which dataframe each row came from
OREI_graphing <- OREI_2020 %>%
  bind_rows(OREI_2016, .id = "ID") %>%
  mutate(ID = replace(ID, ID == 2, 5)) %>%
  select(compost2016, compost2020, DOC, POXC, SOC, CBH, total_MB, ID, F_to_B)

#Create graphs for each variable, with 2016 and 2020 compost graphed separately
DOC <- ggplot(data = OREI_graphing)+
  geom_point(aes(x = compost2016, y = DOC, color = ID)) +
  geom_point(aes(x = compost2020, y = DOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2016, y = DOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2020, y = DOC, color = ID)) +
  labs (x = "", y = "Dissolved Organic Carbon (mg/kg)", color = "Legend") +
  scale_color_manual(name = "Years Since \n Compost \n Application", breaks = c("1", "5"), values = c("

POXC <- ggplot(data= OREI_graphing)+
  geom_point(aes(x = compost2016, y = DOC, color = ID)) +
  geom_point(aes(x = compost2020, y = DOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2016, y = DOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2020, y = DOC, color = ID)) +
  labs (x = "", y = "Oxidizable Carbon (mg/kg)", color = "Legend") +
  scale_color_manual(name = "Years Since \n Compost  \n Application", breaks = c("1", "5"), values = c(

SOC <- ggplot(data= OREI_graphing)+
  geom_point(aes(x = compost2016, y = SOC, color = ID)) +
  geom_point(aes(x = compost2020, y = SOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2016, y = SOC, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2020, y = SOC, color = ID)) +
  labs ( x = "", y = "Total Organic Carbon (%)", color = "Legend") +
  scale_color_manual(name = "Years Since \n Compost \n Application", breaks = c("1", "5"), values = c("

CBH <- ggplot(data= OREI_graphing)+
  geom_point(aes(x = compost2016, y = CBH, color = ID)) +
  geom_point(aes(x = compost2020, y = CBH, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2016, y = CBH, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2020, y = CBH, color = ID)) +
  labs (x = "", y = "Cellobiohydrolase Enzyme Activity (nmol/h/g)", color = "Legend") +
  scale_color_manual(name = "Years Since \n Compost \n Application", breaks = c("1", "5"), values = c("

F_to_B <- ggplot(data= OREI_graphing)+
  geom_point(aes(x = compost2016, y = F_to_B, color = ID)) +
  geom_point(aes(x = compost2020, y = F_to_B, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2016, y = F_to_B, color = ID)) +
  geom_smooth(method = 'lm', aes(x = compost2020, y = F_to_B, color = ID)) +
  labs ( x = "Compost Rate (Mg/ha)", y = "Fungi to Bacteria Ratio", color = "Legend") +
  scale_color_manual(name = "Years Since \n Compost \n Application", breaks = c("1", "5"), values = c("

#Use patchwork to lay out plots
SOC + POXC + DOC + CBH + F_to_B +
  plot_annotation (title = "Reponse of Soil Properties to Compost Application") +

#combine the legends
  plot_layout (guides = "collect", ncol = 2)
```


Reponse of Soil Properties to Compost Application
Total Organic Carbon (%)
2.00
1.75
1.50
1.25
0
10
20
30
40
50
Oxidizable Carbon (mg/kg)
300
200
100
0
10
20
30
40
50
Dissolved Organic Carbon (mg/kg)
300
200
100
0
10
20
30
40
50
Cellobiohydrolase Enzyme Activity (nmol/h/g)
100
80
60
40
20
0
10
20
30
40
50
Years Since Compost Application
1
5
Years Since Compost Application
1
5
Fungi to Bacteria Ratio
0.40
0.35
0.30
0
10
20
30
40
50
Compost Rate (Mg/ha)

Conclusions

While total organic carbon remains high even five years after compost application, the labile carbon pools increase one year after compost application but decrease almost to baseline after five years. Enzyme activity shows a similar (though less pronounced) trend, and fungi to bacteria ratio decreases only slightly with compost application. Since labile organic matter pools serve as food for microorganisms, we might expect to see a larger microbial response to compost application. However, microbial activity and decomposition may be limited by drought in this ecosystem. This would help explain why total organic carbon can persists for five years: it is not being transformed into a labile, microbial-available pool. Going forwards, I will also look at soil physical properties, which help soils resist erosion. I also hope to incubate some of these soil samples under wet conditions, to see how available water would impact the microbial response to compost application and the long-term stability of this compost.

References

Asseng, Senthold, Frank Ewert, Pierre Martre, Reimund P Rötter, David B Lobell, Davide Cammarano, Bruce A Kimball, et al. 2015. "Rising Temperatures Reduce Global Wheat Production." *Nature Climate Change* 5 (2): 143–47.

Kaur, Gurpreet, Axel Garcia y Garcia, Urszula Norton, Tomas Persson, and Thijs Kelleners. 2015. "Effects of Cropping Practices on Water-Use and Water Productivity of Dryland Winter Wheat in the High Plains Ecoregion of Wyoming." *Journal of Crop Improvement* 29 (5): 491–517.

Larney, Francis J, Dan M Sullivan, Katherine E Buckley, and Bahman Eghball. 2006. "The Role of Composting in Recycling Manure Nutrients." *Canadian Journal of Soil Science* 86 (4): 597–611.

Norton, Jay B, Eusebius J Mukhwana, and Urszula Norton. 2012. "Loss and Recovery of Soil Organic Carbon and Nitrogen in a Semiarid Agroecosystem." *Soil Science Society of America Journal* 76 (2): 505–14.

Rodgers, Hannah R, Jay B Norton, and Linda TA van Diepen. 2021. "Effects of Semiarid Wheat Agriculture Management Practices on Soil Microbial Properties: A Review." *Agronomy* 11 (5): 852.