Individual Project Preliminary Analyses

Hannah Rodgers

4/27/2022

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). However, there are gaps in our understanding of how soil microbial properties should be interpreted, especially in semiarid systems.

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.

Load Packages

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

Here I merge the datasets, filter out the data I need, and separate it 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')

OREI_all <- OREI_all %>%
#remove plots that are fallow or fertilized
  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 first I'll 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.
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.

1. The samples come from randomized plots, so they are independent.

2. Check for 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
```

The following variables have non-normal residuals (p > 0.05):
OREI_2020: DOC, DON, PHOS
OREI_2016: PHOS

Transform the non-normal variables.

```
#this function performs a box_cox transformation on a variable to make it normal
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
OREI_2016$PHOS <- box_cox_transform(OREI_2016$PHOS)
```

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

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

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

```
#test normality of new data using code from above
```

All the variables are normal now! ($p > 0.05$)

3. Check linearity using an F-test for lack of fit.

```
#This function creates a linear and quadratic model for each variable,
#then tests whether the two models differ significantly using ANOVA
F_test <- function(x) {
  mod <- lm(x ~ compost, data = OREI_2016)
  reduced<-lm(x ~ compost, data = OREI_2016)
  full<-lm(x ~ poly(compost,2), data = OREI_2016)
  return(anova(reduced, full)$"Pr(>F)") }

#run this function on each variable
F.test.2016 <- list()
for (i in colnames(OREI_2016[,7:38])) {
  F.test.2016[[i]] <- F_test(OREI_2016[[i]]) }

F.test.2020 <- list()
for (i in colnames(OREI_2020[,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)
```

4. Check constancy of residuals with the Levene test.

```
#Run the levene test (from car package) on each variable,
#then save the p values
levene.2016 <- list()
for (i in colnames(OREI_2016[,7:38])) {
  result <- leveneTest((OREI_2016[[i]]) ~ as.factor(OREI_2016$compost))
  levene.2016[[i]] <- result$`Pr(>F)`[1] }

levene.2020 <- list()
for (i in colnames(OREI_2020[,7:38])) {
  result <- leveneTest((OREI_2020[[i]]) ~ as.factor(OREI_2020$compost))
  levene.2020[[i]] <- result$`Pr(>F)`[1] }
```

All data passes the levene test! (p $<$ 0.05)

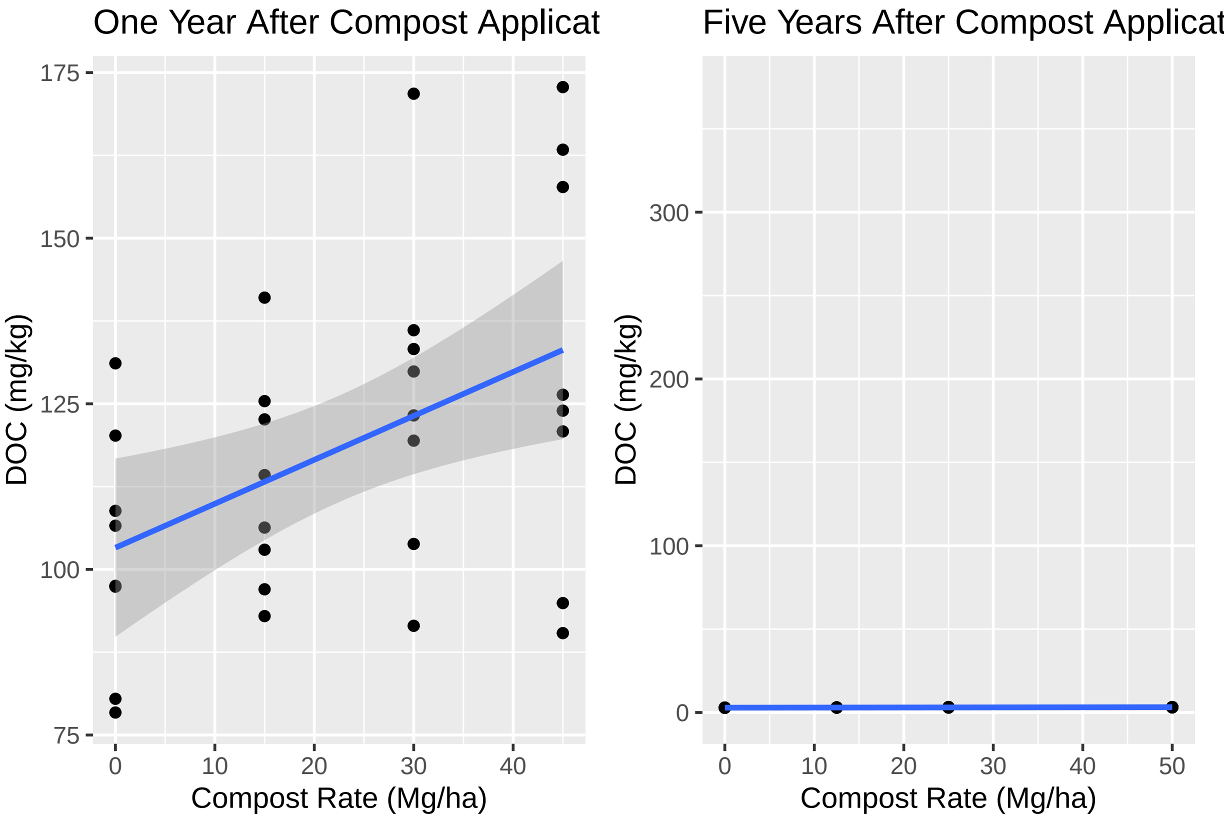
Visualize Data

This section needs some work, but for now I'll just print some plots.

```
p1 <- ggplot(data= OREI_2016, aes(x = compost, y = DOC))+
  geom_point() +
  geom_smooth(method = 'lm') +
  labs (title = 'One Year After Compost Application',
        x = "Compost Rate (Mg/ha)", y = "DOC (mg/kg)")

p2 <- ggplot(data= OREI_2020, aes(compost, DOC)) +
  geom_point() +
  geom_smooth(method='lm') +
  labs (title = 'Five Years After Compost Application',
        x = "Compost Rate (Mg/ha)", y = "DOC (mg/kg)")

library(patchwork)
p1 + p2 + scale_y_continuous(limits = c(0, 375))
```



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.