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; USGCRP 2018). 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). Sustaining agriculture in the High Plains will require alternative farming systems that can rebuild soil health while remaining profitable (**hansen2012?**).

Soil health is critical for sustainable and resilient food production, yet evaluating soil health in semiarid climates remains challenging. Agriculture impacts soil health through practices such as crop rotation, fallowing, tillage, and fertilization, though these effects vary depending on environmental factors such as climate and soil type. 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.

This project uses soil data 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 was applied to 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 (**larney2006?**). This includes data on soil physical properties (such as aggregate stability and bulk density), soil chemical properties (such as total organic matter, labile carbon and nitrogen pools, and pH) and soil microbiology (enzyme activity, microbial biomass, and microbial community composition). I aim to use regression analysis to evaluate the relationships between compost rate and soil properties. However, many of the variables do not meet the regression assumptions. Therefore, this project focuses on cleaning and tidying my dataset, 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 help growers work towards building the long-term health of their soil, and could guide efforts to protect soils and create an agricultural system resistant in the face of climate change and land degradation.

#Load Packages

#Import Data

#Clean Up Data

```r
#merge the sheets
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 fallow phase and chemical fertilized plots
  filter(Rotation == "wheat", Treatment != "fertilizer") %>%

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

#separate into the two groups
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.

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

#OREI_2016, removed outliers from: yield, NO3, PMN, DON, MBC, CBH, PHOS, NAG, actino,  gram_pos, AMF, s
#OREI_2020, removed outliers from: NO3, protein, MBC, MBN, SOC, N, NAG, BX, AG, SUL

#save no_outlier file to disk
#write_xlsx(OREI_2020, "OREI_2020_no_outliers.xlsx")
#write_xlsx(OREI_2016, "OREI_2016_no_outliers.xlsx")

#read in no_outlier datasets 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"))
```

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.

```r
#create an empty list
p.vals <- list()

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

#any variable with p>0.05 has non-normal residuals
p.vals
```

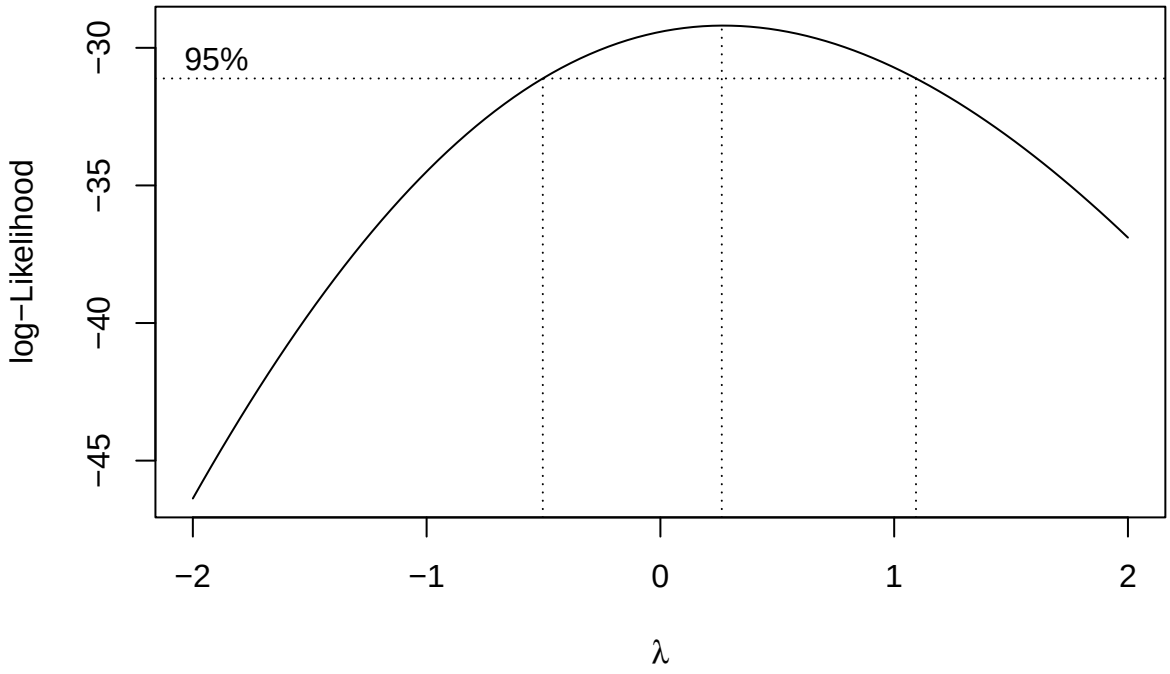
```
## $Yield
## [1] 0.9840273
##
## $Tillers
## [1] 0.4959564
##
## $Heads
## [1] 0.7077767
##
## $WFPS
## [1] 0.5932977
##
## $NO3
## [1] 0.6432678
##
## $PMN
## [1] 0.4940133
##
## $Porosity
## [1] 0.5394321
##
## $Protein
## [1] 0.09652962
##
## $POXC
## [1] 0.6199739
##
## $DOC
## [1] 0.9843709
##
## $DON
## [1] 0.9873995
##
## $PMC
## [1] 0.2809314
##
## $MBC_fumigated
## [1] 0.8234193
##
## $MBN_fumigated
## [1] 0.9491492
##
```

```
## $SOC
## [1] 0.648513
##
## $N
## [1] 0.09972005
##
## $BG
## [1] 0.8934904
##
## $CBH
## [1] 0.3173866
##
## $PHOS
## [1] 0.02234545
##
## $NAG
## [1] 0.5288161
##
## $BX
## [1] 0.2330646
##
## $AG
## [1] 0.5735598
##
## $SUL
## [1] 0.6220938
##
## $LAP
## [1] 0.653469
##
## $actinomycetes
## [1] 0.613477
##
## $gram_neg
## [1] 0.1433333
##
## $gram_pos
## [1] 0.7731213
##
## $AMF
## [1] 0.6186391
##
## $sapro_fungi
## [1] 0.5985867
##
## $total_MB
## [1] 0.7721015
##
## $total_fungi
## [1] 0.9505337
##
## $total_bacteria
## [1] 0.1436672
```

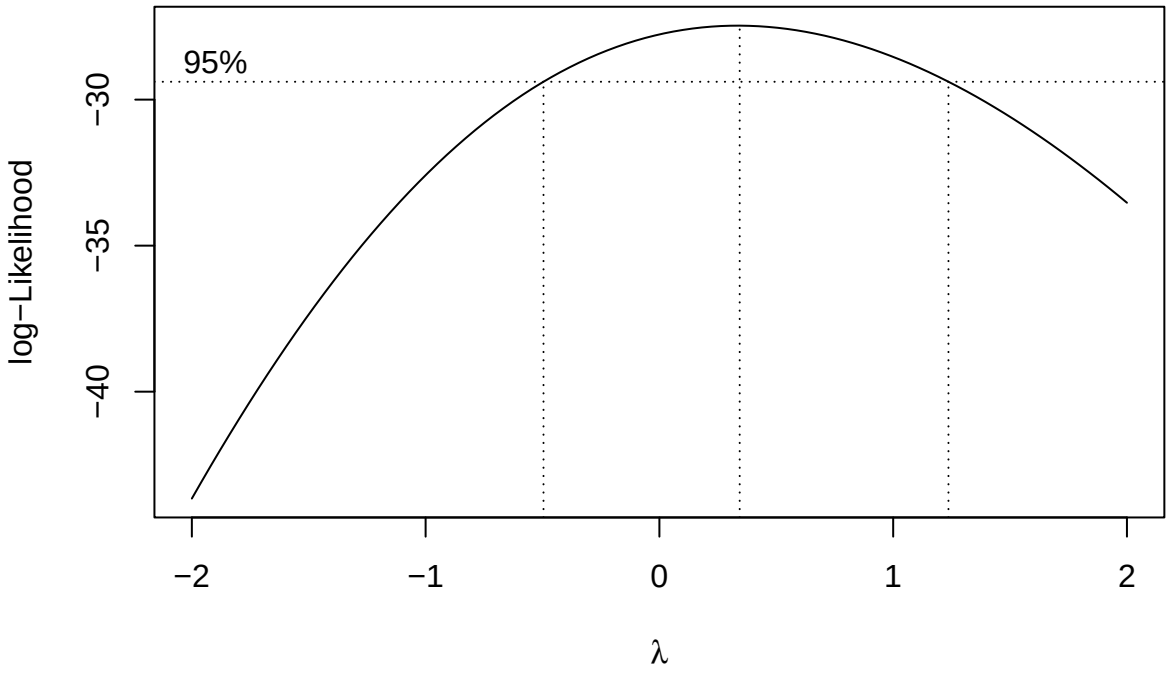
```
#OREI_2020 with non-normal: DOC, DON, PHOS
#OREI_2016: only PHOS
```

#Transform any non-normal variables.

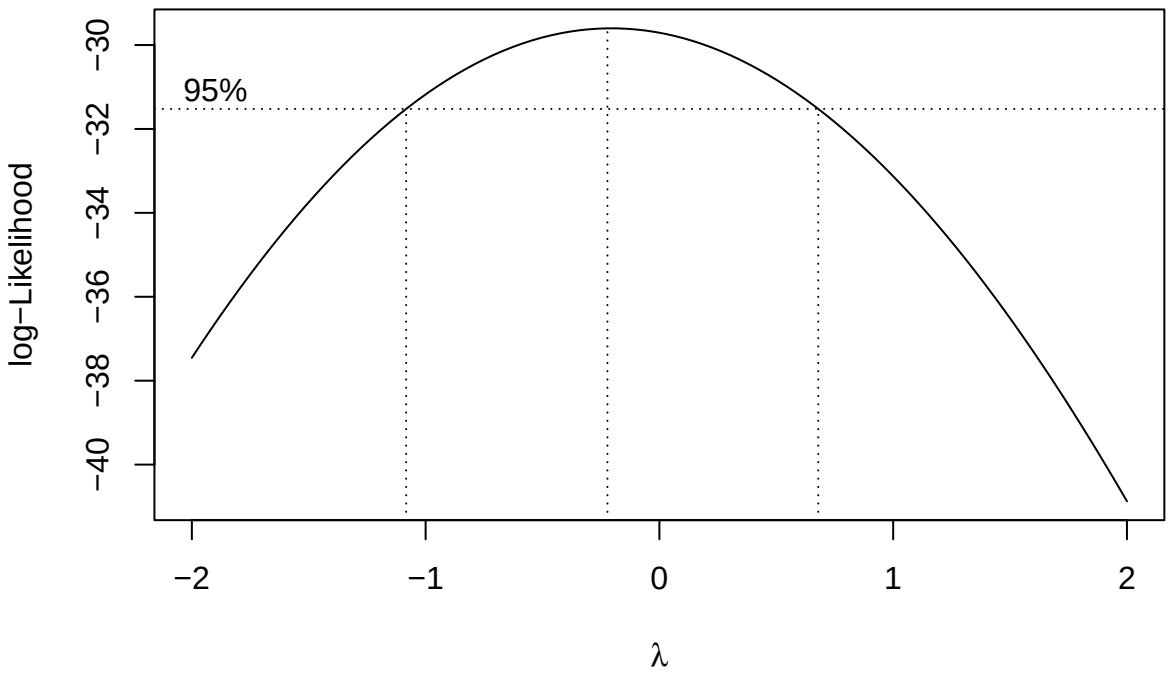
```
#this function performs the optimal box_cox transformation on a variable to make it normal
box_cox_transform <- function(v) {

#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)
```



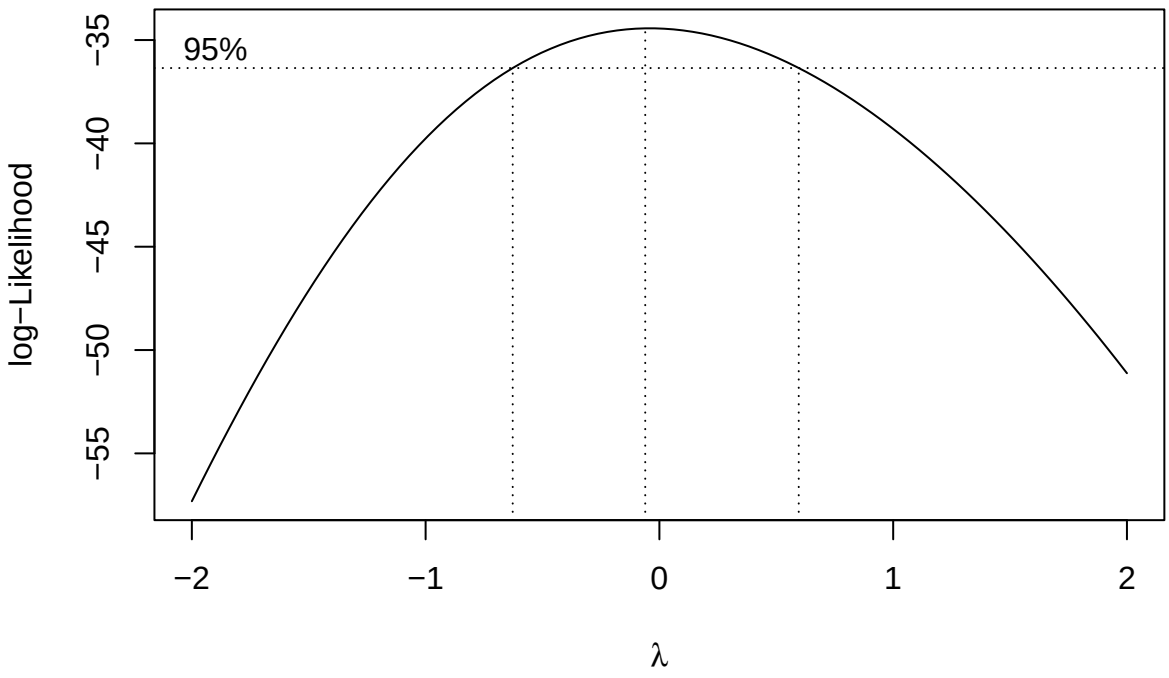
```
OREI_2020$PHOS <- box_cox_transform(OREI_2020$PHOS)
```




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




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




```
#test normality of new data using same loop from above
p.vals.trans <- list()

for (i in names(OREI_2016[,7:38])) {
  mod <- lm(get(i) ~ compost, data = OREI_2016)
  p.vals.trans[[i]] <- (shapiro.test(mod$residuals))$p.value }

#everything is normal now! All p values >0.05
p.vals.trans
```

```
## $Yield
## [1] 0.9840273
##
## $Tillers
## [1] 0.4959564
##
## $Heads
## [1] 0.7077767
##
## $WFPS
## [1] 0.5932977
##
## $NO3
## [1] 0.6432678
##
## $PMN
```

```
## [1] 0.4940133
##
## $Porosity
## [1] 0.5394321
##
## $Protein
## [1] 0.09652962
##
## $POXC
## [1] 0.6199739
##
## $DOC
## [1] 0.9843709
##
## $DON
## [1] 0.9873995
##
## $PMC
## [1] 0.2809314
##
## $MBC_fumigated
## [1] 0.8234193
##
## $MBN_fumigated
## [1] 0.9491492
##
## $SOC
## [1] 0.648513
##
## $N
## [1] 0.09972005
##
## $BG
## [1] 0.8934904
##
## $CBH
## [1] 0.3173866
##
## $PHOS
## [1] 0.1282244
##
## $NAG
## [1] 0.5288161
##
## $BX
## [1] 0.2330646
##
## $AG
## [1] 0.5735598
##
## $SUL
## [1] 0.6220938
##
## $LAP
```

```
## [1] 0.653469
##
## $actinomycetes
## [1] 0.613477
##
## $gram_neg
## [1] 0.1433333
##
## $gram_pos
## [1] 0.7731213
##
## $AMF
## [1] 0.6186391
##
## $sapro_fungi
## [1] 0.5985867
##
## $total_MB
## [1] 0.7721015
##
## $total_fungi
## [1] 0.9505337
##
## $total_bacteria
## [1] 0.1436672
```

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

```
#this function creates a linear and quadratic model for each variable, and then tests whether the two m
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)")
}

#use sapply to 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]]) }

#The regression differs significantly from linear if p>0.05
#For OREI_2020: DOC, protein, porosity, WFPS differ
#For OREI_2016: all good!
```

#####4. check constancy of residuals (levene test? brown forsythe?) look at Liana's code ####

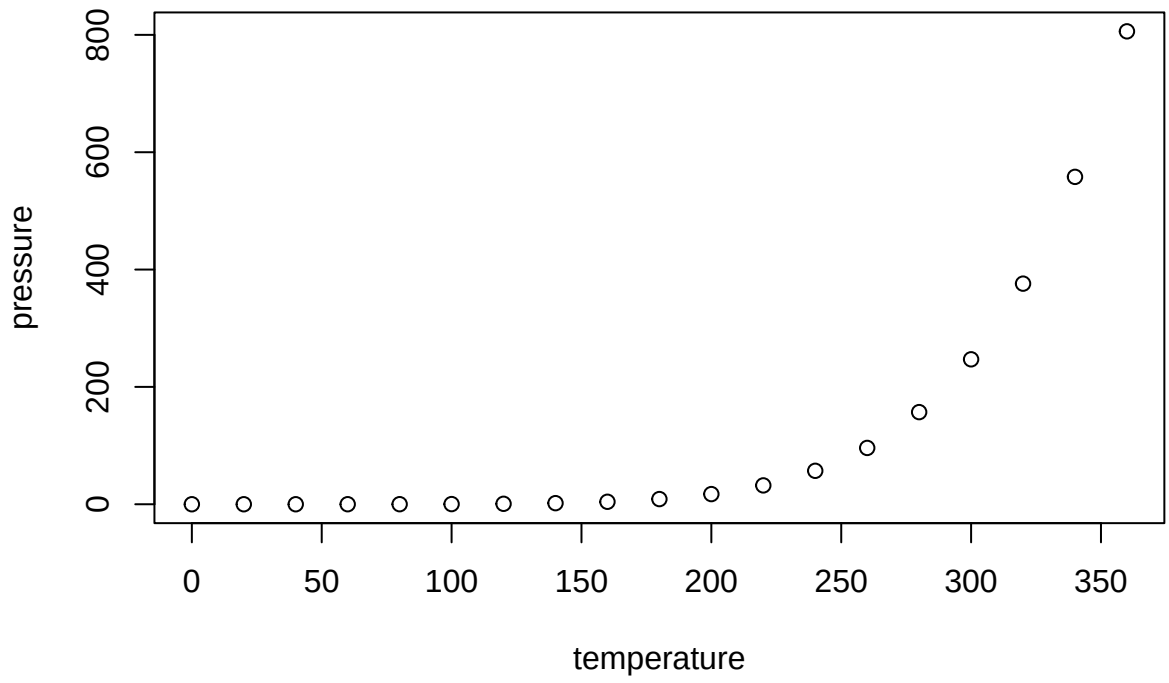
```
#library(car)
#leveneTest(mod, compost)
```

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

#p1 + p2 #& scale_y_continuous(limits = c(0, 375))
```

VISUALIZE DATA

Including Plots

You can also embed plots, for example:



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.

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.

USGCRP, Climate. 2018. "Fourth National Climate Assessment."