Statistical Analysis of Biomass Yield: A Case Study

Maria F. Aranguren[1], Mylinh Nguyen[2], Roxanne Salinas[3]

[1] Texas Sustainable Energy Research Institute and Mechanical Engineering Departs, The University of Texas at San Antonio. One UTSA Circle, San Antonio, TX 78253, USA; aranguren.maria@outlook.com

[2] The University of Texas at San Antonio. One UTSA Circle, San Antoni, TX 78253, USA; mylinh.nguyen@my.utsa.edu[2], r.b.salinas88@gmail.com[3]

Abstract: Co-firing biomass with coal is one of the popular choices for renewable energy alternatives. Advantages of co-firing biomass include: the creation of jobs, the efficient use of current power plant infrastructure, the growth of electricity generation from renewable sources, and the reduction of greenhouse gas (GHG) emissions. Our project is based on the available data from a U.S Department of Agriculture (USDA) sponsored simulation case study in Texas. The chosen biomass, switchgrass, is grown in two counties near a power plant. A USDA simulation software, ALMANAC, will be used to determine the biomass yield. The goal is to help the researcher in charge of this USDA project to determine which factors are required in the study and if the granularity currently used is necessary or not. By aiding in this decision, we allow our client to cut in computational time and simplify the model used. The results show that the granularity regarding the years should be maintained, and therefore the collection of weather data and analysis should be continued and improved in future studies. On the other hand, the results show that many of the soils in the area analyzed are not significantly different and the granularity can be scaled back allowing our client's model to be simplified.

Submission Date: May 4th, 2018

Keyword: Biomass, Switchgrass, ANOVA

1. Introduction

Coal-fired power plants generate one-third of the electricity in the United States [1] (Figure 1). The emissions produced from these power plants can be reduced by co-firing with biomass. Biomass co-firing entails replacing a portion of the coal used in the power plant with biomass to reduce net emissions. Biomass provide substantial environmental benefits such as net zero emissions when burned for energy production [2]. Advantages of co-firing biomass include: the creation of jobs, the efficient use of current power plant infrastructure, the growth of electricity generation from renewable sources, and the reduction of greenhouse gas (GHG) emissions. Thus, co-firing biomass with coal can reduce emissions without incurring in major plant infrastructure changes/investments along with benefiting the local economy by creating agricultural jobs. The co-firing rate must be between 10-25% to avoid affecting the thermal efficiency of the power plant and keep plant infrastructure unchanged [3].

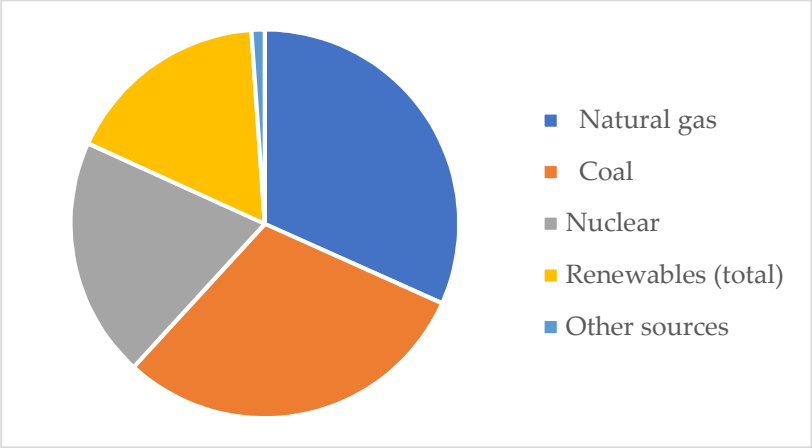

Figure 1: US Energy Consumption by Energy Source [1]

Considering all the benefits biomass has to offer the renewable energy industry, the Texas Sustainable Energy Research Institute (TSERI) at the University of Texas at San Antonio has taken interest in the topic and is working to make its growth and collection more efficient. Our client works at TSERI researching this topic by implementing mathematical optimization models to the biomass supply chain. The case study used in our client's latest research is located at the south of San Antonio. In this location, the highest yielding biomass is determined to be switchgrass through previous analyses [4]. In the case study, high levels of granularity are implemented, which entails high computational capacity and complexity. In our study, we wish to determine if this level of granularity is required for the study by implementing statistical methods. Further detail regarding the methodology, results, and discussion to follow.

2. Data Collection

The following section details how the data utilized in this project was collected. The data is mostly collected through databases, the most important of which: SSURGO and NCDC NOAA.

The USDA's *Soil Survey Geographic Database (SSURGO)* [5] is a database which contains information about soil as collected by the National Cooperative Soil Survey over the course of a century. The information available within SSURGO comes in various formats, once of which allows for the mapping of the soils in the county in question (Figure 2). SSURGO does not offer the maps for soil types as there are too many to make a meaningful map, but it maps the soil in the area by ranks. The ranks represent production potential of the area, in this case Range crops, which are comparable to switchgrass. In previous research, our client divided the region in small enough parcels (or farms) to be able to represent soil information accurately throughout the analysis; this lead to over 2,000 locations.

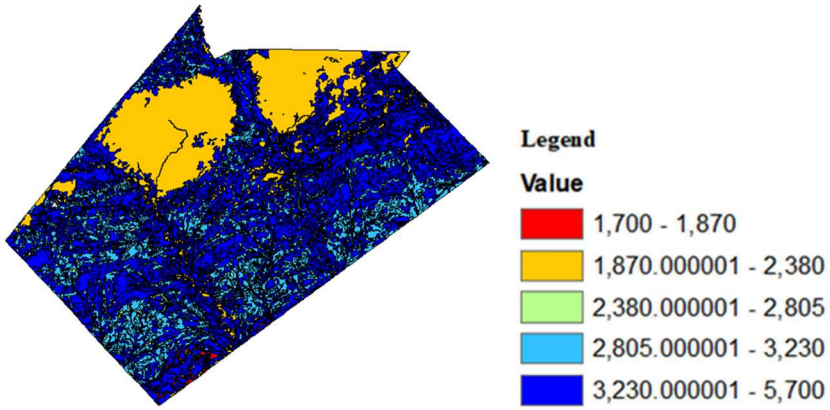

Figure 2: SSURGO Input

The weather information was obtained through National Oceanic and Atmospheric Administration's National Climatic Data Center (NOAA's NCDC) [6]. The closest weather station that data was drawn from is located in Floresville, TX, USA. All historical weather information, regarding minimum and maximum temperature, and precipitation, were obtained for years 1950-2000. But due to missing data, some years were excluded from the statistical analysis presented in this paper. The weather data was used in our client's research with the predetermined notion that the wide range would increase the model's accuracy. By doing this, our client is also increasing computational time and model complexity. In further analysis, we will determine if this is in fact necessary and whether it should be pursued further, or if the years are not as significant as predetermined by our client.

The simulations to determine biomass yield in the area are done through the United States Department of Agriculture's *Agricultural Land Management with Numerical Assessment Criteria* (ALMANAC) [7]. ALMANAC is a comprehensive simulation model that is able to predict plant growth in various locations. ALMANAC simulations use weather, soil, and location data to determine growth. Using the *BatchRun* capabilities of ALMANAC, we can simulate crop growth in the randomly selected coordinates for each potential planting location; which are obtained through means of geographical information systems (GIS). Even though there are other parameters that ALMANAC is able to account for, such as solar radiation, we were unable to collect this data and therefore used the built-in ALMANAC prediction feature; these other parameters were not account for in the statistical analysis.

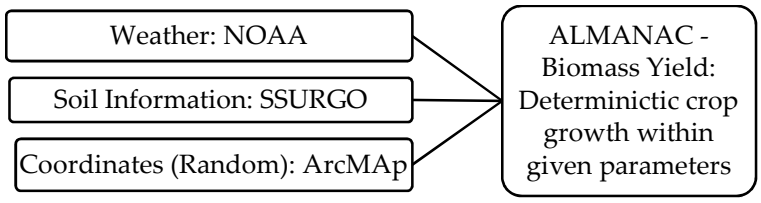

Figure 3: Data Collection Summary

3. Methodology

The case study on biomass cofiring contained over 50,000 data points with several variables related to our purpose of factorial experimental design study. From the data, we selected five factors that had possible contribution to the yield of biomass: maximum temperature, minimum temperature, soil type, precipitation, and year. Due to the nature of large data set, missing and incomplete data problems occur frequently. Because of the consistency in this time frame, we chose to work on the historical data from the years 1962 to 1999. The analysis for the project is based on SAS application and the complete codes can be found on the supporting information section at the end of the paper (Appendix A) [8]. In our project, we focused on soil type since it was the factor that we believed had some significant impact on the yield of biomass. Before proceeding with any test, Levene's and Barlett (1) test were conducted on the data to test the homogeneity of variance using factor soil type and yield of biomass. Even though we had almost 7,000 data points, we still wanted to check whether the assumption of normality is applicable.

$$\mathrm{H_0}: {\sigma^2}_1 = {\sigma^2}_2 = .. = {\sigma^2}_k \qquad (1)$$

$$\mathrm{H_1}: \text{at least one pair(ij)} {\sigma^2}_i = {\sigma^2}_j$$

Next, a randomized complete block (RCB) design was carried out with soil type as treatment and year as block. The factors in the study varied from year to year. In order to reduce the error in studying the effect of soil type on the mean yield of biomass we blocked our data by year. All 30 soil types were present in each year from 1962 to 1999, the conditions in each year regarding temperatures and rainfall levels are said to be homogeneous in this case. We want to test on the hypothesis that soil type treatment effects are present via an F-test. The model used is presented by equation (2).

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (2)$$

Where μ is the overall mean of the yield of biomass, α_i is the deviation from the mean of the i^{th} treatment group (soil types), β_j is the deviation from the mean of j^{th} block (year), and ε_{ij} is the random deviation association with each observation. After the test on soil type treatment through RCB design, we continued with multiple pairwise comparisons for the different types of soils. Tukey's HSD, Fisher LSD, and Duncan's test were chosen to see which soil types are different from the rest (3).

$$\mathrm{H_0}: \mu_i - \mu_j = 0 \ \ vs. \ \ H_1: \mu_i - \mu_j \neq 0 \qquad (3)$$

Finally, we ran the analysis of variance for all five factors of the study to determine the effect of each factor on the yield of biomass. Since we had a mixed model, we tested on two different hypotheses

For fixed factors: maximum temperature, minimum temperature, year, and precipitation

$$H_0: \text{all } \alpha_i = 0 \quad (4)$$

For random factor: soil type

$$H_0: \text{all } \sigma_\alpha^2 = 0 \quad (5)$$

4. Results & Discussion

In order to find the most reasonable methods of growing this biomass, we started by analyzing how soil influences the mean yield of switchgrass. We were attempting to discern two things: whether any of the soils could be determined to be better than the others, and which types of soil could be reasonable substitutes if the best soil is unavailable. Before we could proceed with our analysis, we had to test the normality assumption in order to analyze the nature of the case study data. Levene's test was used to ascertain the validity of our hypothesis that all variances are equal.

Levene's Test for Homogeneity of Yield Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Soil	29	1057.3	36.4582	3.22	<.0001
Error	6430	72845.6	11.3290		

Figure 4: Levene's Test

The test produced a significant result, which means that there is sufficient evidence to suggest that not all the variances are equal. Since our sample size is large, we proceeded carefully under the assumption of normality.

Next, we sought to determine whether soil type had any effect on the mean yield of switchgrass. If soil type has no effect, pursuing which soil benefits biomass growth the most would be irrelevant and further analysis would not be necessary. A one factor analysis of variance was done to determine if the main effect, soil, was present.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	29	2923.0994	100.7965	3.55	<.0001
Error	6430	182580.2687	28.3951		
Corrected Total	6459	185503.3681			

R-Square	Coeff Var	Root MSE	Yield Mean
0.015758	52.34801	5.328702	10.17938

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Soil	29	2923.099364	100.796530	3.55	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Soil	29	2923.099364	100.796530	3.55	<.0001

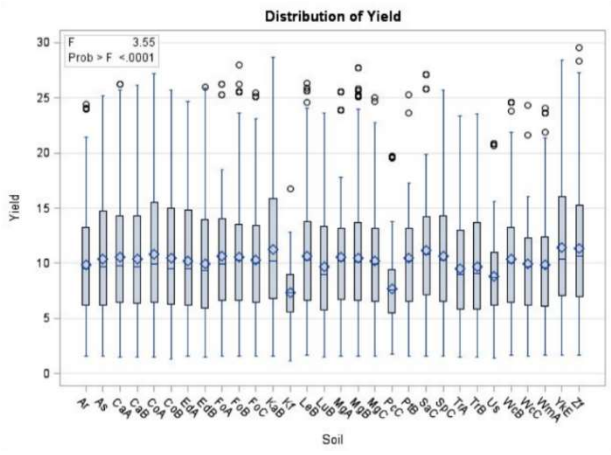

Figure 5: SAS ANOVA Output

The tables and graphs in Figure 5 tests the hypothesis of whether all mean yields of switchgrass by soil type are equal. Since the p-value is less than 0.0001, we can reject the null hypothesis and conclude that at least one of the soil types is not exhibiting the same effect on biomass growth as the others. Subsequently, we aimed to determine which soil is best for the growth of switchgrass. Three comparison tests were used: Fishers, Duncan's, and Tukey's (Appendix A). We were in search of soil types that were worse than the rest, or better than the rest. The results of the tests indicate there is not a clear soil that is better than the others. Between all three multiple comparison tests, the soils labeled PcC and Kf were not grouped with the soils that had the highest mean yield. Therefore, it is recommended to avoid these soil types or use cost effective methods of improving the current soil as needed. The highest mean yield soils, YKE, Zf, and KaB, do not show a significant difference with the rest of the soils except for the two mentioned previously. This allows for flexibility soil selection when that option is available.

The previous ANOVA table, Figure 5, with soil type only, showed a coefficient of determination of 0.015; this suggests that only 1.5% of the variation in mean yield of switchgrass is determined by soil alone. This means soil type may not be the most influential factor for the mean yield of switchgrass. Other factors were investigated by first blocking our soil by year to determine if there was a difference in the results by year.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1138	184622.9817	162.2346	980.54	<.0001
Error	5321	880.3864	0.1655		
Corrected Total	6459	185503.3681			

R-Square	Coeff Var	Root MSE	Yield Mean
0.995254	3.995939	0.406762	10.17938

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	37	178711.1006	4830.0297	29192.4	<.0001
Soil	29	3035.2839	104.6650	632.59	<.0001
Year*Soil	1072	2876.5971	2.6834	16.22	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Year	37	73745.98425	1993.13471	12046.4	<.0001
Soil	29	2984.00903	102.89686	621.90	<.0001
Year*Soil	1072	2876.59711	2.68339	16.22	<.0001

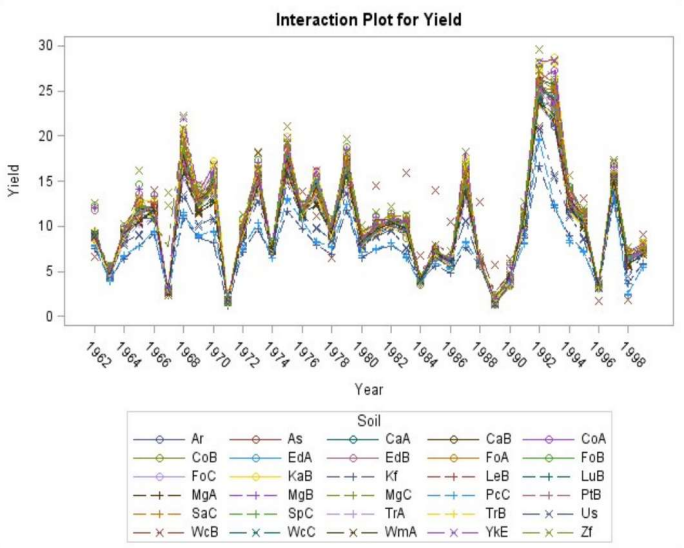

Figure 6: SAS ANOVA output for soil blocked by years

The new model, shown in Figure 6, has a coefficient of determination of 0.995, which indicates it explains 99.5% of the variation in mean yield of switchgrass. Soil effects are still significant as well as our blocking effects. This implies the year the switchgrass was harvested had an impact on the quantity of the biomass collected. These results coincide with our hypothesis since blocking by year would account for common weather patterns like extreme droughts, extreme heat, wind patterns, and even insect and animal migration through the area. Since the yield is significantly different as time progresses, we would advise against using the average for

biomass yield and recommend our client to change their deterministic study to a stochastic study to create a more meaningful analysis in future research.

The interaction effects between our blocking variable and soil were also significant. This suggests there is something happening within our blocking variable we need to investigate. With the available data, more variables are added into the model despite the current model already explaining 99.5% of the variation. Our goal is to figure out what other factors are influencing the mean yield of switchgrass. The variables precipitation, maximum temperature, and minimum temperature were added to our model (Table 1).

Table 1: SAS Results of Mixed Model Analysis of Variance

Source	DF	Type III SS	Mean Square	F Value	Pr > F
max	91	98.461500	1.081995	2.30	<.0001
min	78	49.905143	0.639810	1.36	0.0192
year	37	370471	10013	21293.1	<.0001
Soil	29	4688.043431	161.656670	343.78	<.0001
precipitation	199	95.443840	0.479617	1.02	0.4096
Error: MS(Error)	11716	5509.272975	0.470235		

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	434	399567.0528	920.6614	1957.88	<.0001
Error	11716	5509.2730	0.4702		
Corrected Total	12150	405076.3257			

R-Square	Coeff Var	Root MSE	Yield Mean
0.986399	6.457859	0.685737	10.61864

Source	DF	Type I SS	Mean Square	F Value	Pr > F
max	91	8582.3307	94.3113	200.56	<.0001
min	78	5158.1790	66.1305	140.63	<.0001
year	37	380924.5035	10295.2569	21893.9	<.0001
Soil	29	4806.5957	165.7447	352.47	<.0001
precipitation	199	95.4438	0.4796	1.02	0.4096

Source	DF	Type III SS	Mean Square	F Value	Pr > F
max	91	98.4615	1.0820	2.30	<.0001
min	78	49.9051	0.6398	1.36	0.0192
year	37	370471.3059	10012.7380	21293.1	<.0001
Soil	29	4688.0434	161.6567	343.78	<.0001
precipitation	199	95.4438	0.4796	1.02	0.4096

The ANOVA table with the new variables explains 98.6% of the variation without any interaction variables present. All of the new variables were significant except for precipitation. This could be due to the fact that other variables in the model explain the mean yield better. This does not mean the amount of precipitation plays no effect in mean biomass yield. In fact, when our blocking variable is removed, precipitation becomes significant, but our coefficient of determination is reduced. The year variable itself still accounts for most of the model, which means there are other factors within each year that influence the mean yield of switchgrass.

Data on other potential variables is needed to determine what may be more influential in achieving a greater biomass output.

5. Conclusion

Considering all the results and the analysis done on the data provided regarding the case study, there are plenty of recommendations we can give our client for future research regarding biomass yield in the area in questions. The first recommendation is to decrease the granularity level regarding the soil in the area. In recent research, our client divided the county into sections which lead to a total of over 2,000 locations for their model. We would advise that this is not necessary in this county, and that a larger scale can still capture most of the growth behavior in the area accurately and efficiently. Aside from what can be eliminated, we would recommend our client to broaden the data collection and analysis on the years examined. In previous research, the yield in this case study used deterministic data. We advise to move into a stochastic model that accounts for the uncertainty seen in this study, as the biomass yield drastically changes through the years.

To further improve this analysis, the interaction effects of the five-factor model could be considered. This was not a possibility for this study as we lacked the computer power to run such a large dataset for the desired interactions. Additional factors can also be added if the data is available in the area, such as solar radiation and wind.

6. Acknowledgements

We would like to give our gratitude to Dr. Keying Ye, Ph. D. for the academic and moral support during this semester's journey. We appreciate your work and passion for statistics.

7. References

1. Pankaj; Alok Short-Term Energy Outlook (STEO). **2017**.
2. Maung, T. A.; Mccarl, B. A. Economics of Biomass Fuels for Electricity Production: A Case Study with Crop Residues. **2008**.
3. Tillman, D. A. Cofiring benefits for coal and biomass. *Biomass and Bioenergy* **2000**, *19*, 363–364, doi:10.1016/S0961-9534(00)00048-9.
4. Hart, J. Biomass Supply Chain Logistics For Co-firing Coal Power Plants, San Antonio, 2016.
5. SSURGO USDA Web Soil Survey. Available online: https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx (accessed on Sep 1, 2017).
6. NOAA NCDC Available online: https://www.ncdc.noaa.gov/climate-information.
7. Kiniry, J. R.; Sanderson, M. A.; Williams, J. R.; Tischler, C. R.; Hussey, M. A.; Ocumpaugh, W. R.; Read, J. C.; Esbroeck, G. Van; Reed, R. L. Simulation Alamo Switchgrass with the ALMANAC Model. *Agron. J.* **1996**, *88*.
8. SAS Available online: https://www.sas.com/en_us/software/stat.html.

APPENDIX A

Means with the same letter are not significantly different.								
t Grouping						Mean	N	Soil
			A			11.4268	38	YkE
			A					
	B		A			11.3437	190	Zf
	B		A					
	B		A	C		11.2729	114	KaB
	B		A	C				
	B	D	A	C		11.1516	152	SaC
	B	D	A	C				
E	B	D	A	C		10.8539	38	CoA
E	B	D	A	C				
E	B	D	A	C		10.6892	38	SpC
E	B	D	A	C	F	9.8788	190	WmA
E	B	D		C	F			
E	B	D		C	F	9.8519	190	Ar
E		D		C	F			
E		D		C	F	9.7208	950	TrB
E		D			F			
E		D			F	9.6607	190	LuB
E					F			
E					F	9.5247	38	TrA
					F			
			G		F	8.7953	190	Us
			G					
			G			7.6638	152	PcC
			G					
			G			7.3605	38	Kf

Means with the same letter are not significantly different.				
Duncan Grouping		Mean	N	Soil
	A	11.4268	38	YkE
	A			
	A	11.3437	190	Zf
	A			
	A	11.2729	114	KaB
	A			
	A	11.1516	152	SaC
	A			
	A	10.8539	38	CoA
	A			
B	A	10.6892	38	SpC
B	A	9.8788	190	WmA
B	A			
B	A	9.8519	190	Ar
B	A			
B	A	9.7208	950	TrB
B	A			
B	A	9.6607	190	LuB
B	A			
B	A	9.5247	38	TrA
B				
B	C	8.7953	190	Us
	C			
	C	7.6638	152	PcC
	C			
	C	7.3605	38	Kf

Means with the same letter are not significantly different.					
Tukey Grouping			Mean	N	Soil
	A		11.4268	38	YkE
	A				
	A		11.3437	190	Zf
	A				
	A		11.2729	114	KaB
	A				
	A		11.1516	152	SaC
	A				
	A		10.8539	38	CoA
	A				
	A		10.6892	38	SpC
B	A	C	9.8788	190	WmA
B	A	C			
B	A	C	9.8519	190	Ar
B	A	C			
B	A	C	9.7208	950	TrB
B	A	C			
B	A	C	9.6607	190	LuB
B	A	C			
B	A	C	9.5247	38	TrA
B	A	C			
B	A	C	8.7953	190	Us
B		C			
B		C	7.6638	152	PcC
		C			
		C	7.3605	38	Kf

Note: The middle 16 soils were left out since they follow the same grouping as before and after the ..

Supporting Information: SAS code

```
libname project "E:\";
*merging data with yield of biomass and five factors;
data rox;
set project.rox;
if year>=1962 and year<=1999 then output rox;
run;
data raw;
Iinfile 'E:\project2.txt'  dlmstr='  ' dsd;
input year month day max min precipitation;
run;
data good;
```

```
set raw;
if max ne 999 and min ne 999 and precipitation ne 999 then output good;
keep year max min precipitation;
run;
proc means data=good noprint;
class year;
output out=g mean=/autoname;run;
data g2;
set g (rename=(max_mean=max min_mean=min precipitation_mean=precipitation));
if _n_=1 then delete;
drop _type_ _freq_;
run;

proc sort data=rox;
by year;
run;
proc sort data=g2;
by year;
run;
data combine;
merge rox(in=x) g2(in=y);
by year;
if x*y;
run;
*testing for normality assumption;
proc glm data=project.comb;
class soil;
model yield =  soil;
```

```
means soil /hovtest=levene(type=abs);
means soil /hovtest=bartlett;
run; quit;
*multiple comparision tests for soil;
proc glm data=project.comb;
class  soil;
model yield = soil;
means soil / Bon Tukey LSD Duncan;
run; quit;

*RCB design for soil type and year
proc glm data=project.comb;
class year soil;
model yield =year|soil;
run; quit;
*ANOVA with multi-factors
proc glm data=project.comb;
class max min precipitation soil;
model yield = max min precipitation soil ;
random soil/test;
run; quit;
```