**Analysis and Prediction on Corn Yield in Illinois**

**Consultants: Shihao Duan, Smruthi Iyengar, Kagen Quiballo, Doris Wang, Lucy Zhao**

**Client: Mark Schleusener, National Agricultural Statistics Service &**

**Darren Glosemeyer, University of Illinois at Urbana-Champaign**

**University of Illinois at Urbana-Champaign**

**STAT 443**

1. **INTRODUCTION**

**1.1 Background & Motivation**

Corn is a crop that is widely planted in the Midwest. In Illinois, there are nine agricultural districts, all of which plant a vast amount of corn. However, due to temperature and precipitation disparities, corn yields in different areas and different years may have huge differences.

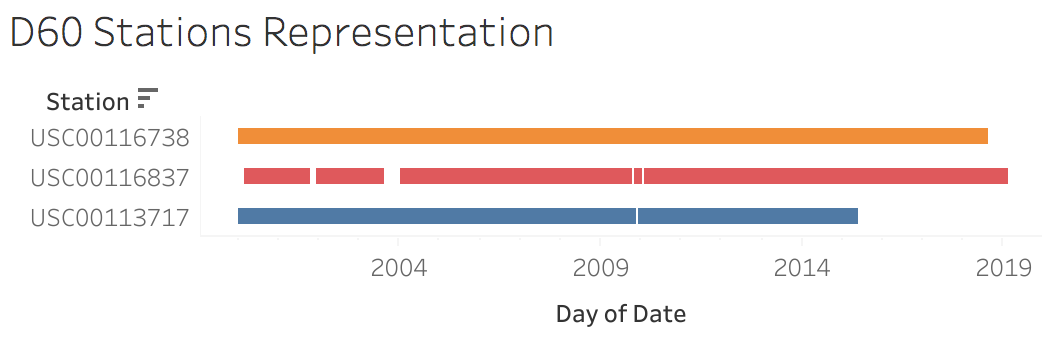
According to the requirements of the client, our goals are to tell if there is a difference between the corn yield in district 20 and district 60 over the past two decades, point out the reasons behind the difference, predict the corn yield by building statistical models, and figure out if innovative technology have an impact on the corn yield. We got the crop yield and progress data from Mark Schleusener from the National Agricultural Statistics Service (NASS). The crop yield data contains the information of the corn yields in different agricultural districts, so we believe this dataset can be used to see if there is a difference in corn yields between the districts 20 and 60. The weather and climate data were provided by the National Oceanic and Atmospheric Administration (NOAA), containing the detailed meteorological data in different weather stations in district 20 and 60 from 2000 to 2018, and we think these data are important because they provide important factors that may contribute to the difference of corn yields in different areas and years.

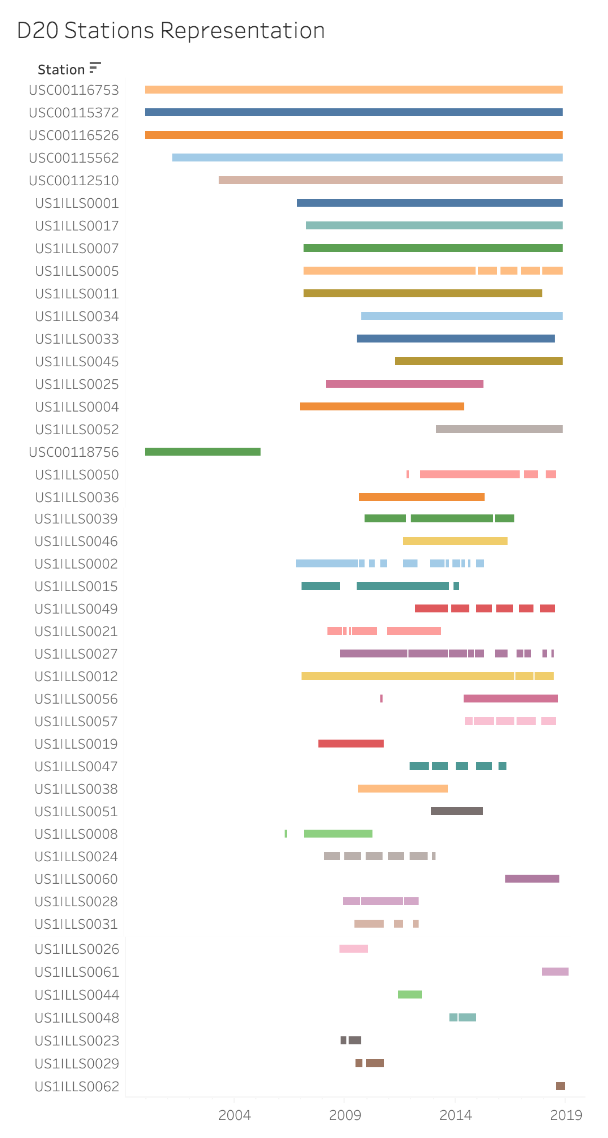
**1.2 Data Preparation**

The four datasets given are *Illinois Corn Progress Data*, *Illinois Corn Yield Data*, *LaSalle County*, *District 20*, and *Pike County, District 60*. The *Illinois Corn Progress Data* includes 13 variables and 2868 observations, recording corn’s growth status in 9 districts in Illinois weekly from April to November. The *Illinois Corn yield Data* has 11 variables and 190 observations. It records the corn grain yields per acre and acres harvested of 9 districts and the state yearly from 2000 to 2018. The *LaSalle County, District 20* data has weather information in district 20 including 7 variables and 95949 observations recorded by 45 weather stations. The *Pike County, District 60* data has weather information in district 60 including 7 variables and 18405 observations recorded by 3 weather stations. After reading these datasets into R, the primary analytical tool we are using for building predictive models, we performed data examination and cleaning. One of the issues that we noticed was that for the datasets *LaSalle County, District 20* and *Pike County, District 60*, there were lots of missing values that were useless for our analysis, so we subsetted these two datasets by removing the rows that contained missing values. For example, in *LaSalle County, District 20*, there were over 90k observations, after cleaning, only 19603 of them are found useful. The cleaned datasets are used for building regression models. The second problem existing in the datasets is that the variables in the datasets were recorded in different time scale. For example, the variables regarding precipitation and temperature were recorded daily in the datasets *LaSalle County, District 20* and *Pike County, District 60*, however, the data of the percentages of different corn growth stages were measured by week in *Illinois Corn Progress*, and the data of corn yield per acre and acres harvested in *Illinois Corn Yield Data* were measured by year in *Illinois Corn Yield Data*. How to utilize and reconcile, if necessary, these different measures when doing analysis is definitely what we should think about. We will address this issue later with the description of the specific analysis performed. Last but not least, to get better analysis, we added several variables that might be helpful in the variable analysis part.

1. **VARIABLE ANALYSIS AND VISUALIZATION**

**2.1 Precipitation**

When choosing to clean the precipitation data, we were faced with the issue of predicting annual yield with daily temperature data from multiple stations across districts. Converting the metrics of the two posed a problem as well as missing values on many dates, but the thought process of our methodology is the following. District 20 had 45 stations and District 60 had 3 stations that each recorded data from different days, some of them overlapping. The following is a graph representing which days each station recorded precipitation 



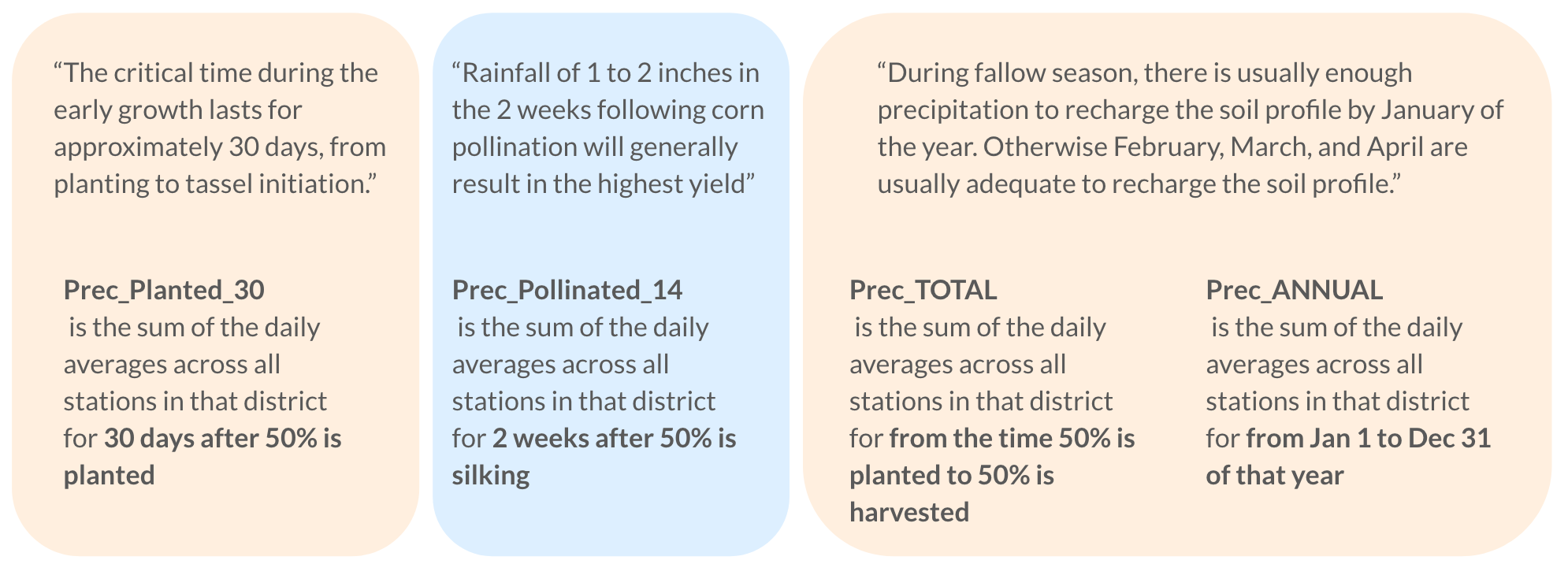
The first method of creating a cumulative annual precipitation variable introduces the bias of over-representation from dates that have more data points. For example, if all the stations happen to record the day of a flood, the year would reflect precipitation that is too high, given that there are missing values from other dates that year from those given stations.

The next method of only choosing stations that have data points for all days from 2000-2018 would also be biased by not representing other areas in the district.

Therefore, the method we chose was to query the data by selecting daily averages across all stations in that district such that each day was equally represented despite missing values. From there, we take the sum of select days each year.

Looking at some simple statistics in relation to NASS standards, we see that dates with over 4 inches of precipitation may affect crop yields. This occurs once in 2013 for district 20 and twice in 2003 and 2014 from district 60. This may be taken note of when looking at the time vs. yield graph. Additionally according to NASS standards, approximately one inch of rain per week is good, and district 20 and 60 have an average weekly precipitation of 0.736 inches and 0.761 inches respectively. According to the Illinois State Water Survey, “Generally, annual rainfall exceeds the water requirement of Illinois crops. The average for southern Illinois is 45 inches, the rest of the state is 37 inches,” and is reflected by the average annual rainfall of 38.45 inches and 39.77 inches of district 20 and district 60 respectively.

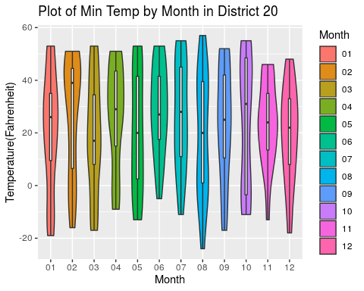
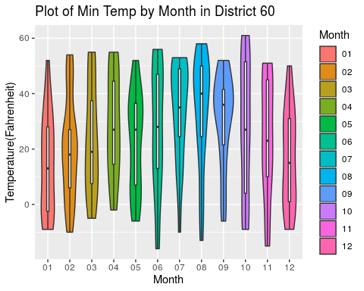
4 new precipitation variables were created as follows according to quotes from the Illinois State Water Survey. We used these quotes to create the variables because the time periods selected were indicated as critical time periods that may affect crop yield. Querying this data took a lot of time, and we were unable to integrate them into the predictive models. They have the potential to add more accuracy to the models and may be used in future directions. The 4 new queried variables are defined below and are available for viewing with [this link](http://tinyurl.com/STAT443grp6-prcp-data): *tinyurl.com/STAT443grp6-prcp-data.*

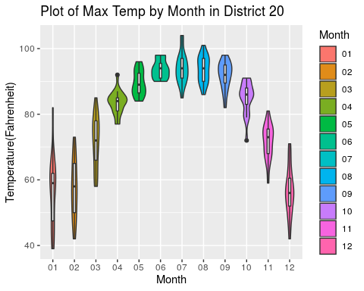
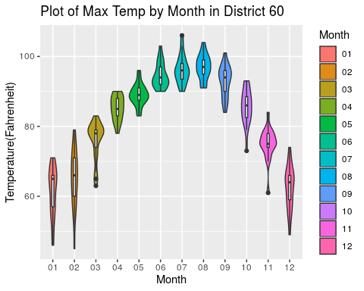


One last factor to consider is droughts. According to the Palmer Drought Severity Index on isws.illinois.edu, 2005 was considered drought and 2012 was considered extreme drought. The left graph indicates the years of these droughts have significantly lower annual precipitation than adjacent years. The right graph indicates that these droughts also impacted annual yield as shown by the significantly lower yields in 2005 and 2012.

|  |  |
| --- | --- |

**2.2 Temperature**

****

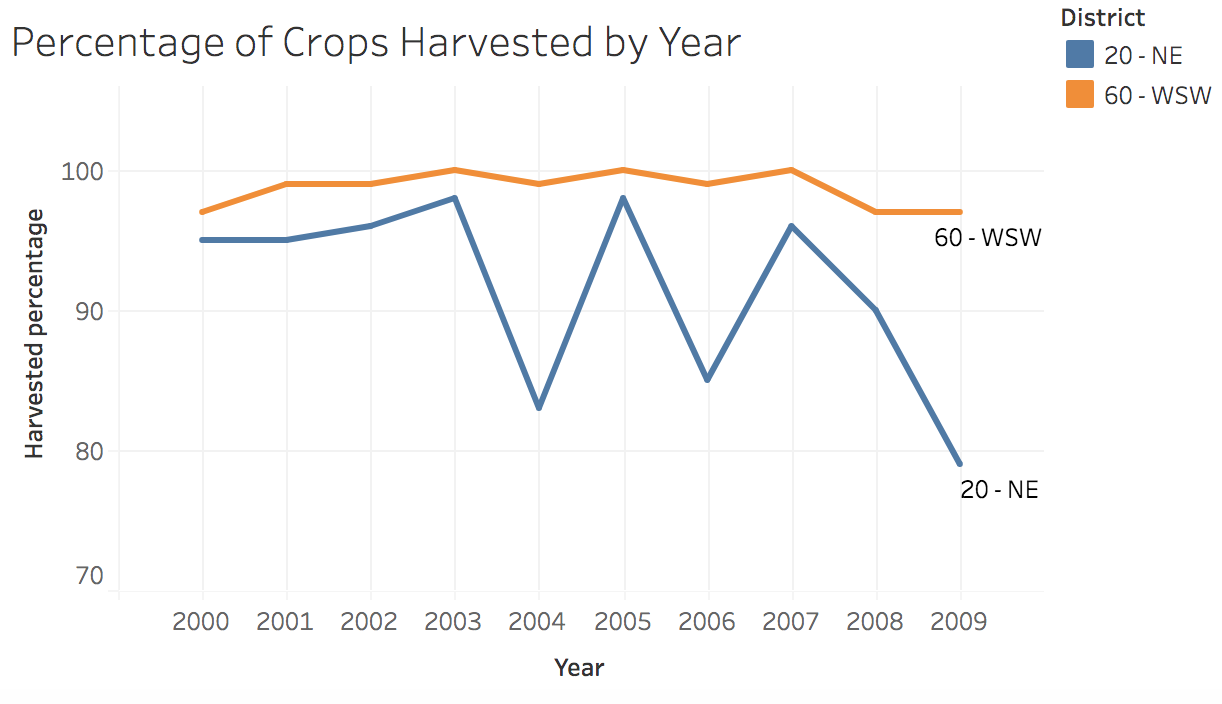
****

In order to examine the temperature variable, we want to see how temperature changes on average throughout the years. The dataset contains daily minimum and maximum temperatures of each district from the years 2000 to 2018. We created a monthly average of minimum temperature and maximum temperature throughout the years. This monthly average was made by picking the highest temperature value and the lowest temperature value of each month between January 2000 and December 2018 and creating a violin plot to determine if there are any major differences in temperature between the two districts. For both districts, the average temperatures were somewhat similar. But, District 20 had slightly lower averages than district 60 in both maximum temperature and minimum temperature. For both districts, the range in average the range of the minimum temperature is much more than the range of the maximum temperature. For average maximum temperature, the summer months have the smallest range in both districts.

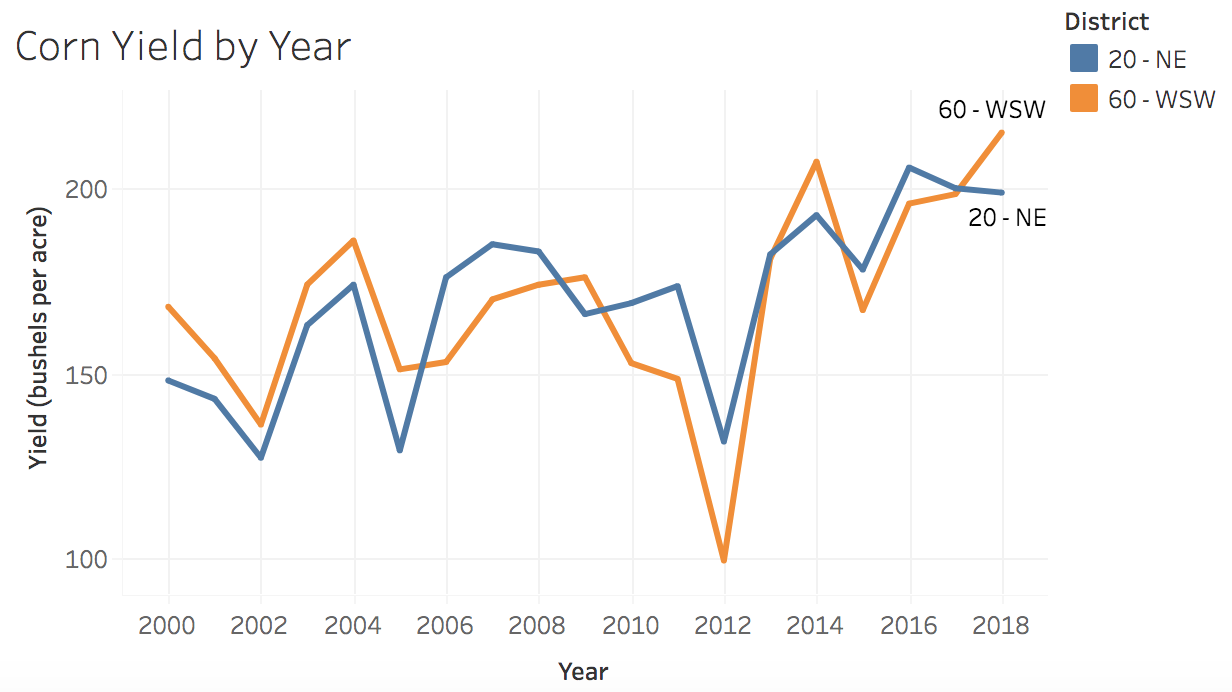
Although we can visualize that there are some differences in temperature between the two districts, we need to determine if the difference is statistically significant. The overall average maximum temperature in district 60 is 81.71°F and the average maximum temperature in district 20 is 78.72°F. After running a two-sample paired t-test for the maximum temperature variable we can determine that this difference in temperature is statistically different because the p-value is 0.03 and the difference has a 95 percent confidence interval of (-5.62, -0.37). The average minimum temperature in district 60 is 25.33°F and the average minimum temperature in district 20 is 23.24°F. The p-value from the paired two-sample t-test was .30 and the 95 percent confidence interval for the difference in minimum temperatures included 0 (-6.10, 1.91). So, the difference in minimum temperature between the two districts are not statistically significant.

1. **DISTRICT VARIANCE ANALYSIS AND VISUALIZATION**

As mentioned before, the datasets are given in different time scale, so it is difficult for us to compare the difference in yield using data with different time scales, as we notice that the dataset *Illinois Corn Yield Data* gives corn yield for all regions in Illinois, we made two subsets that contain the data of district 20 and district 60, respectively. We performed a paired t-test to compare the difference in means.



The result shows that when looking at the percentage of crops harvested after each season, there is a significant differencebetween the harvested percentage in district 20 and district 60. District 60 typically has a higher percentage harvested of about 98.7% compared to district 20’s 91.5% average harvested. One may also note that data spans from 2000-2009. Our p-value for the test was 0.00573, so at a 0.05 significance level, we conclude that the average harvest percentage of district 20 is less than the average harvest percentage of district 60.

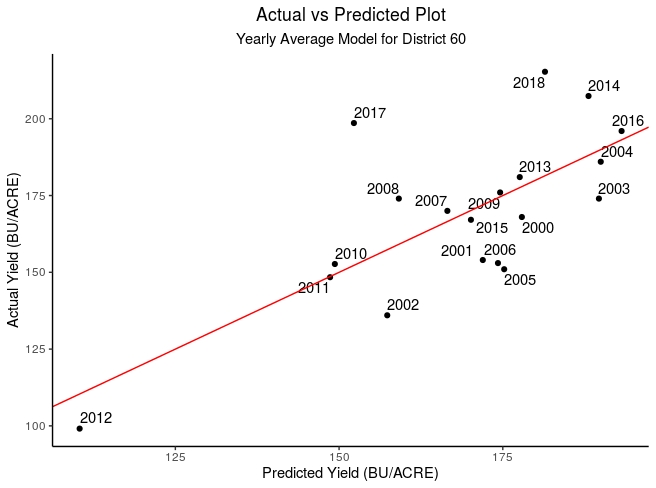
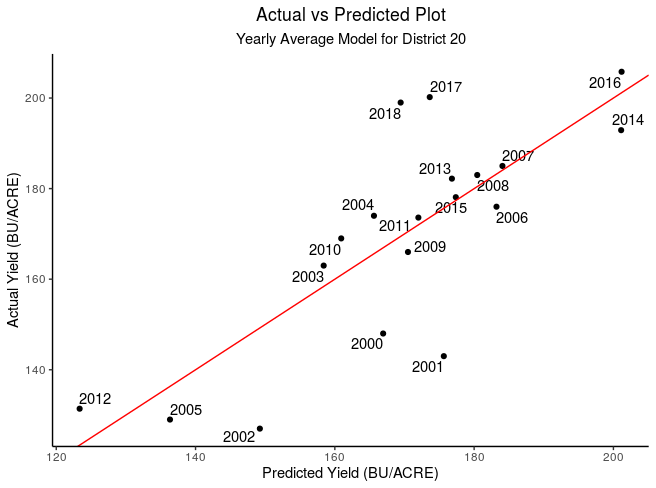


However, when running the same paired t-test to look at the yield measured in BU/acres for each year, there appears to be no significant difference between district 20 and district 60. They both have similar yields of about 169.8 BU/acre for district 20 and 168.82 BU/acre for district 60, which are relatively similar with large variations throughout each year ranging from 99.1 - 215.0. Our test resulted in a p-value of 0.799, so at a 0.05 significance level, we conclude that there is no significant difference between the average yield of district 20 and district 60 when paired by year.

1. **PREDICTIVE MODEL ANALYSIS**

**4.1 Predictive Model - Yearly Averages**

In order to predict future corn yield, we built linear regression models for each district using variables regarding temperature and precipitation from the datasets *LaSalle County, District 20* and *Pike County, District 60* as the predictors, and the corn yield variable measured in bu/acre from the dataset *Illinois Corn Yield Data* as the response variable. Since temperature and precipitation were measured daily while corn yield was measured yearly, we calculated the yearly averages of precipitation, minimum temperature, and maximum temperature variables to make the predictors in the same measurement as the response variable. The summary datasets we used for building the models are in Appendix 7.1.1, and they can also be viewed [here](https://docs.google.com/spreadsheets/d/1dKJVqDbOq5sNdhQ_Jc9mnRFAxwOuPwaq15sEW2fsvaY/edit?usp=sharing). We started from fitting a full linear regression model with three yearly-averaged variables (precipitation, maximum temperature, and minimum temperature) and all their interaction terms, and then selected the variables in both directions based on the Akaike Information Criterion (AIC), which is an estimator of the relative quality of statistical models for a given set of data. The selection procedure showed that the best linear model for District 20 is *Yield (BU/ACRE)* = 11084.501 - 25371.63 \* *Precipitation* - 182.966 \* *Maximum Temperature* - 199.392 \* *Minimum Temperature* + 427.621 \* *Precipitation* \* *Maximum Temperature* + 3.338 \* *Maximum Temperature* \* *Minimum Temperature*, and for District 60 is *Yield* = 933.316 - 820.936 \* *Precipitation* - 24.251 \* *Maximum Temperature* + 20.646 \* *Minimum Temperature*. We proved that all the variables in both models are significant under 0.1 significant level (Appendix 7.1.2), so they have significant impacts on corn yield; the values of Multiple R-Squared, which measures how much variability of the respondent variable can be explained by the predictors, are 0.623 and 0.528, respectively, which means that more than half of the variability in corn yield can be predicted by precipitation, temperature, and their interaction effects. The values of Adjusted R-Squared, which adds a penalty for the number of coefficients to prevent using lots of predictors to boost the Multiple R-Squared values, for both models are 0.478 and 0.434, respectively. The actual vs. predicted corn yield graphs of the linear models of the two districts are presented as follows:



As we can see, the points are falling around the red line of the actual equal to the predicted, which means that most of our predictions are pretty close to the actual corn yields. However, there are also some year’s corn yields that are obviously over- or under-estimated, such as the years 2001, 2002, 2017, and 2018 of District 20 as well as the years 2017 and 2018 of District 60. The Mean Absolute Errors (MAE), which are the means of the absolute differences between the actual and predicted values, of District 20 and 60 using our models are 10.639 and 13.565, respectively. These discrepancies in actual and predicted values may result from other unconsidered factors that may also affect corn yield, or from the fact that this model using yearly averages is too general to make accurate predictions. Therefore, we decided to move on to build a model based on averages by growth stages to see if the accuracy of predictions can be improved.

**4.2 Predictive Model - Growth Stages**

As discussed above, if a farmer possessing sufficiently detailed climate data wants to learn more about how each growing stage affects total corn yield of the year and makes more accurate predictions, he/she may want to use our predictive models for corn yield which was developed based on the averages by growth stage. Based on the beginning and ending dates of each growth stage of each year given in the dataset *Illinois Corn Progress Data* and the daily climate data from *LaSalle Country, District 20* and *Pike Country, District 60*, we calculated the average precipitation, maximum temperature, and minimum temperature by each growth stage. Since the climate daily data contains information from 2000 up to 2018 but *Illinois Corn Progress Data* only gives the dates of each stage up to 2009, so we assumed the time periods for the growth stages in the years 2010 to 2018:

| **Plant** | **Emerge** | **Silk** | **Dough** | **Dent** | **Mature** | **Harvest** |
| --- | --- | --- | --- | --- | --- | --- |
| 04/20 - 06/01 | 05/01 - 06/15 | 07/01 - 08/10 | 07/15 - 09/01 | 08/10 - 09/20 | 09/01 - 10/20 | 09/10 - 11/15 |

So we got the summary datasets (Appendix 7.1.3) for [District 20](https://docs.google.com/spreadsheets/d/1PnhKqbSM8OEyW6lUPftYN0RDmOyJk2Q1sBcE5MNCVME/edit?usp=sharing) and [District 60](https://docs.google.com/spreadsheets/d/1uS8TYtmsGBNVbWKBjMXC3WpTQo6nujszjv7k6eUhxlQ/edit?usp=sharing) about the average precipitation and temperature by each growth stage for each year from 2000 to 2018, which have 19 observations and 21 climate properties, and built our predictive models based on these datasets. As the averages of maximum temperature and minimum temperature are significantly correlated, we build models using average precipitation and average maximum temperature or average minimum temperature separately for each of the districts. We still use the corn yield (BU/ACRE) for each year as our response variable and implement the technique of AIC selection for both directions, that is, the algorithm can either add or drop variables from the original full model to optimize AIC.

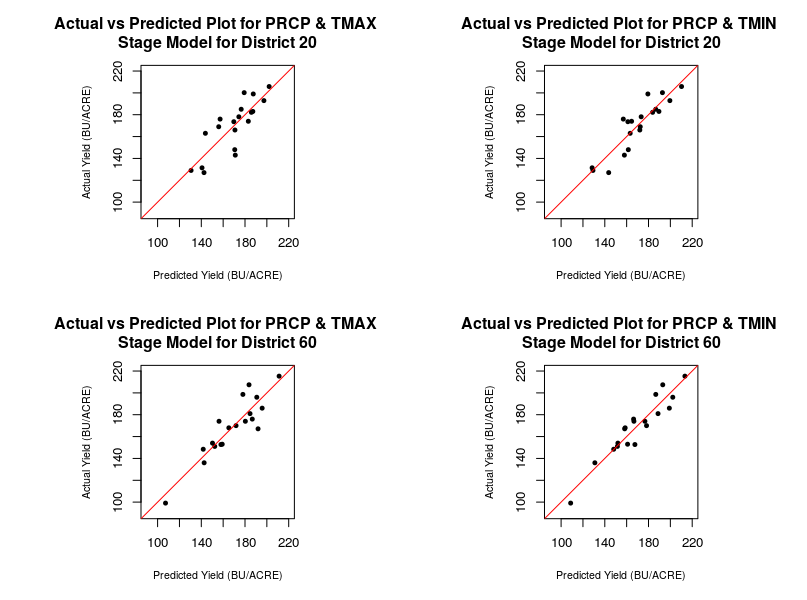
For District 20, the model we built based on precipitation and maximum temperature is: *Yield (BU/ACRE)* = 495. 053 + 119.327 \* *Emerge PRCP* + 303.12 \* *Dough PRCP* - 7.861 \* *Dough TMAX* + 5.227 \* *Dent TMAX* - 2.125 \* *Harvest TMAX*, in which *PRCP* stands for the average precipitation, and *TMAX* stands for maximum temperature of that growth stage. All the coefficients except for that of *Harvest TMAX* are significant at p = 0.1 significance level (Appendix 7.1.4) with *Dough TMAX* being the most significant factor with a p-value of 0.001262, and the p-value of the F-test which compares this model with the model that only uses the intercept as its predictor is 0.005981, which means the model is also overall significant. The Multiple R-Squared is 0.681, meaning that 68.1% of the variability in the corn yield can be explained by the predictors, and the Adjusted R-Squared is 0.558, indicating the model is effective. The significant factors that have a positive effect on the corn yield of this district include the precipitation of emerging and dough and the maximum temperature of denting, and those have a negative effect is the maximum temperature of dough. The overall most influential factor that is significant is the precipitation of dough.

For the same district, the model we built based on precipitation and minimum temperature is: *Yield (BU/ACRE)* = 526.285 - 66.224 \* *Plant PRCP* + 142.258 \* *Emerge PRCP* + 353.579 \* *Silk PRCP* + 504.496 \* *Dough PRCP* - 452.319 \* *Dent PRCP* + 201. 151 \* *Mature PRCP* + 3.515 \* *Planted TMIN* - 5.616 \* *Emerge TMIN* - 6.956 \* *Silk TMIN* + 8.063 \* *Dented TMIN* - 7.359 \* *Harvest TMIN*, in which *TMIN* stands for the average minimum temperature of that growth stage. The coefficients for *Silk PRCP*, *Dough PRCP*, *Silk TMIN*, *Dent TMIN*, and *Harvest TMIN* are significant under p = 0.1 significance level (Appendix 7.1.4) with *Silk TMIN* being the most significant factor with a p-value of 0.00585, and the p-value of the F-test is 0.08716 which shows that the model is overall significant. The Multiple R-Squared is 0.818, meaning that 81.76% of the variability in the corn yield can be explained by the predictors, and the Adjusted R-Squared is 0.531, from which we can see that although the Multiple R-Squared of this model is greater than that of the model with precipitation and maximum temperature, there is no obvious difference in the values of Adjusted R-Squared of these two models, so the greater Multiple R-Squared may mainly result from the increased predictors and thus we do not recommend one of the models over another. The significant factors that have a positive effect on the corn yield of this district include the precipitation of silking and dough and the minimum temperature of denting, those have a negative effect include the minimum temperature of silking and harvest. The most influential factor that is significant is the precipitation of dough.

For District 60, the model we built based on precipitation and maximum temperature is: *Yield (BU/ACRE)* = 490.624 + 65.249 \* *Emerge PRCP* + 252.249 \* *Dent PRCP* - 198.729 \* *Mature PRCP* - 142.653 \* *Harvest PRCP* + 2.325 \* *Emerge TMAX* - 2.487 \* *Silk TMAX* - 7.414 \* *Dough TMAX* + 4.165 \* *Dent TMAX*. The coefficients for *Dent PRCP*, *Mature PRCP*, *Harvest PRCP*, *Dough TMAX*, and *Dent TMAX* are significant under p = 0.1 significance level (Appendix 7.1.5) with *Dent TMAX* being the most significant factor with a p-value of 0.023, and the p-value of the F-test is 0.007641, indicating that the model is overall significant. The Multiple R-Squared is 0.814, meaning that 81.4% of the variability of the response variable can be explained by the predictors, and the Adjusted R-Squared is 0.664. The significant factors that have a positive effect on the corn yield of this district include the precipitation of denting and the maximum temperature of denting, and those have a negative effect include precipitation of mature and harvest, and the maximum temperature of dough. The most influential factor that is significant is the precipitation of denting.

For the same district, the model we built based on precipitation and minimum temperature is: *Yield (BU/ACRE)* = 600.784 + 151.055 \* *Emerge PRCP* + 383.585 \* *Dough PRCP* - 90.994 \* *Dent PRCP* + 4.38 \* *Plant TMIN* - 2.222 \* *Emerge TMIN* - 5.815 \* *Silk TMIN* - 6.242 \* *Dent TMIN* + 8.747 \* *Mature TMIN* - 6.657 \* *Harvest TMIN*. All coefficients except for that of *Emerge TMIN* are significant under p = 0.1 significance level (Appendix 7.1.5) with *Dough PRCP* being the most significant factor with a p-value of 0.000473, and the p-value of the F-test is 0.002069, indicating that the model is overall significant. The Multiple R-Squared is 0.893, meaning that 89.3% of the variability of the response variable can be explained by the predictors, and the Adjusted R-Squared is 0.786. Both Multiple R-Squared and Adjusted R-Squared of this model are better than those of the model with precipitation and maximum temperature, so the predictors of this model can better account for the changes in the corn yield. The significant factors that have a positive effect on the corn yield of this district include the precipitation of emerging and dough as well as the minimum temperature of planting and mature, and those have a negative effect include the precipitation of denting and the minimum temperature of planting, silking, and denting. The most influential factor that is significant is the precipitation of dough.

The following graphs show the Actual vs. Predicted corn yields for each of the four models discussed above:

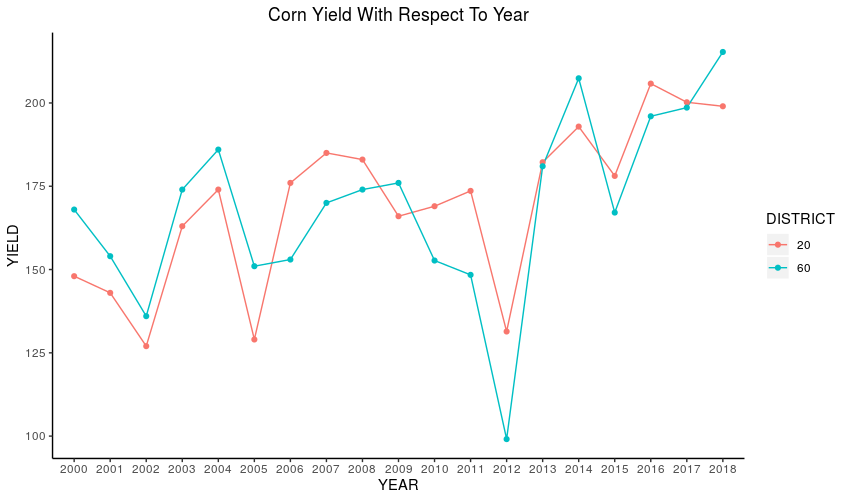


We can see that almost all the points are falling around the red line, which is the line of the actual equal to the predicted value, without any obvious outliers. The MAEs of the *PRCP & TMAX* and *PRCP & TMIN* models for District 20 are 10.863 and 7.982, respectively, and those of the *PRCP & TMAX* and *PRCP & TMIN* models for District 60 are 8.890 and 7.460, respectively, so according to the MAE, the models based on the average precipitation and minimum temperature perform better than those based on the average precipitation and maximum temperature. Looking at the values of Adjusted R-Squared and MAEs, the models incorporating growth stage information have higher values of Adjusted R-Squared and lower MAEs than the yearly-averages models, so we can conclude that the growth-stages models perform better.

1. **INNOVATIVE TECHNOLOGY ANALYSIS**

**5.1 Overall Corn Yield Trend Analysis**

In order to figure out whether technology affects the yield of corn in general, we assume that technology is growing year by year. From the yield vs. year graph below, we find that corn yield has an increasing trend with respect to time. Based on common sense, we made the assumption that technology has been growing year by year. By looking at the corn yield vs year graph, we find the graph has a general positive trend depending on the year, so that we think that technology has a general positive impact on corn yield. Yet we also find that there are some points on the graph showing a relatively low corn yield, as is analyzed previously, drought is a cause for low corn yield, but other reasons such as floods and frozen weather can also damage corn during different stages.

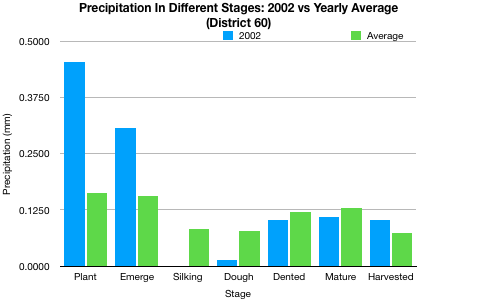
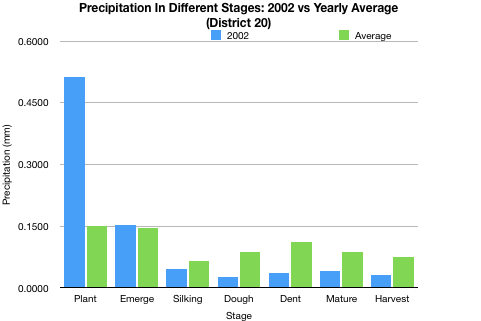


In order to show the pure effect of technology on corn yield, we want to pull out other factors that we already know to have an effect on corn yield. The yield vs. year graph shows that yield valleys occur in the years 2002, 2005, 2012 and 2015. So we decide to take a closer look at the temperature and precipitation for different stages in these four years.

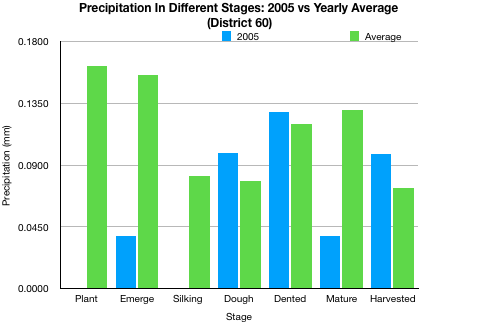
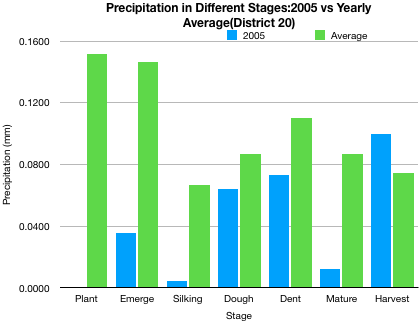
We use the same dataset as we used to fit the growth stage predictive model to plot the precipitation vs stage histogram graph for district 20 and district 60. In order to decide if the temperature or precipitation at a particular stage can be considered as abnormal, we made comparisons between the actual temperature/precipitation variable in each stage to its 19-year average. The following shows the unusual precipitation we have detected:

**5.2 Pointing Out Climate Factors for Abnormal Years**

Looking at the two histograms for 2002’s precipitation, we believe that unusual high precipitation in the plant stage is the main attribution for low corn yield in 2002. For District 20, the average precipitation in the plating stage is 0.1516mm, however, in 2002, the precipitation in the plating stage is 0.5122 mm, which is more than 3 times higher than usual, from the stage predictive model, the coefficient for *Plant PRCP* is -66.224, which means precipitation at plant stage is negatively correlated with yield. For District 60, we noticed that floods occurred during both the plant stage and the emerging stage, for the plating stage the average daily precipitation is 0.4550 mm, and 0.3075 mm for the emerging stage, which is much higher than the average precipitation in this region, thou the predictive model does not include the planting stage precipitation, we still think that a three times higher precipitation would be harmful to corn yield according to NASS’s recommendation.

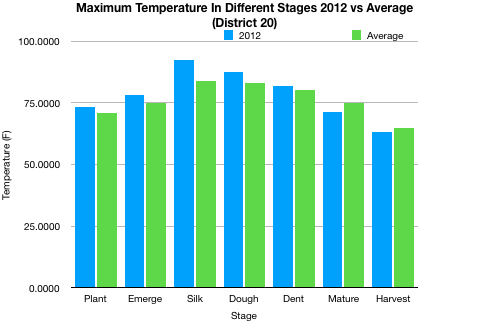
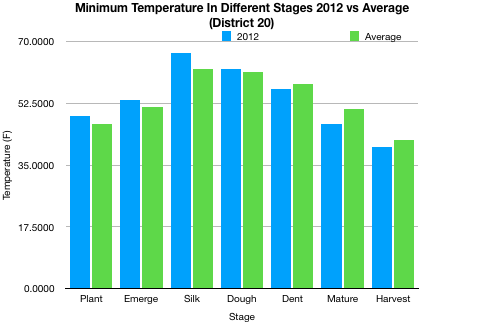


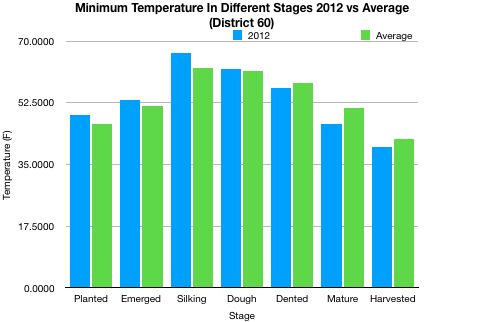
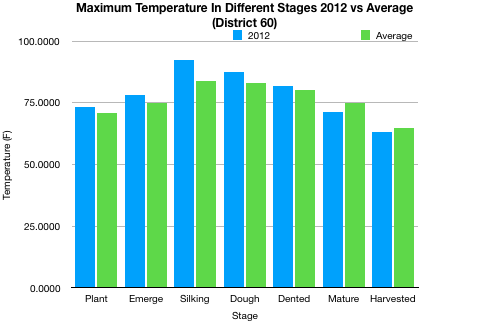
Just as we mentioned previously, droughts occurred during the year 2005, we find droughts happened in multiple stages during the year 2005. By looking at the two histograms below, we find that low precipitation during multiple stages is the main cause of droughts in 2005.



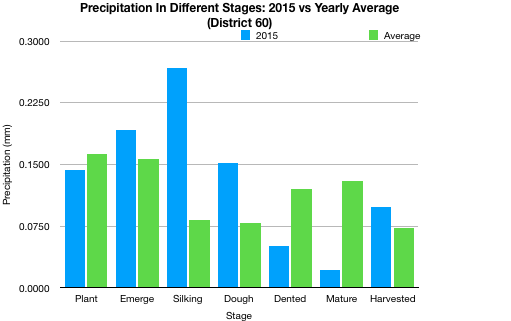
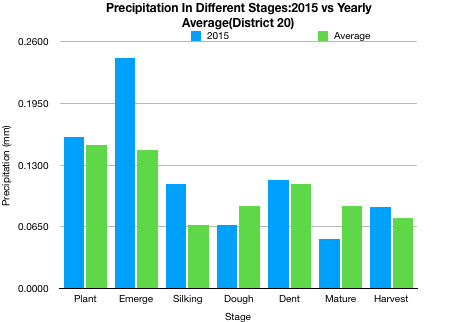
Another severe drought happened in 2012, however, from the analysis, we find that precipitation is actually normal if we compare the rainfall in 2012 to the average precipitation. However, comparing to the 19-year average, temperatures were higher overall during the first half of 2012 and colder during the second half of 2012. During the plant, emerge, silk and dough stage, maximum and minimum temperature is 4~8 degrees higher, especially in the silk stage, the maximum temperature is 8.36 degrees Fahrenheit higher than the average and min temperature is 4.34 degrees Fahrenheit higher than the average.

District 60 has just the same situation as district 20, by looking at the two histograms for district 60, for the silking stage, district 60’s maximum temperature is 8.3615 degrees higher than usual, and the minimum temperature is 4.3635 degrees higher than the average.





The two histograms below show the precipitation in different stages during the year 2015. 2015’s corn yield is a low for the period 2013-2018. We find that unusual high precipitation is the main reason for relatively low corn yield. For District 20, daily precipitation is 0.2432mm, which is 1.7 times more than the average precipitation. Floods seem to be more severe during the silking stage, daily precipitation for the silking stage is 0.2674mm, which is 3 times more than the average precipitation.



1. **CONCLUSION**

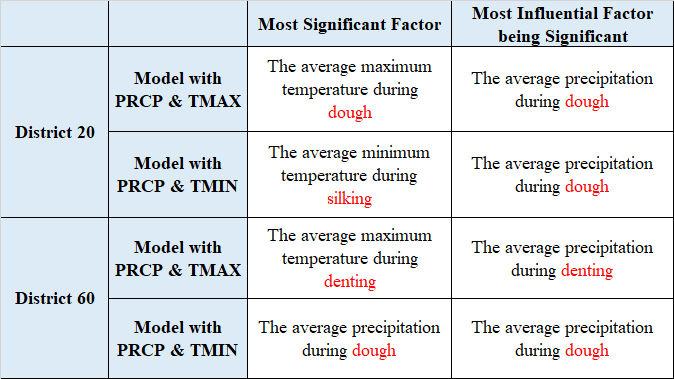
According to the requirements of the client, our goals are to tell if there is a difference between the corn yield in district 20 and district 60 over the past two decades, point out the reasons behind the difference, predict the corn yield by building statistical models, and figure out if innovative technology have an impact on the corn yield.

From the ‘District Variance Analysis’ section, although there is no significant difference between the yield between the districts, we see a statistically and significantly higher harvested percentage in district 60 than district 20.

From the analysis of the precipitation variable, we see that district 60 also has a higher average weekly precipitation and higher annual precipitation which may be the reason for the higher harvested percentage.

From the analysis of the temperature variable, we can determine that the difference in temperatures between the two districts is only significant for the average maximum temperature. The average maximum temperature is higher in district 60 which can explain why district 60 usually has a higher yield. Higher temperatures lead to higher yields.

From the analysis of the predictive models, we define the most significant factor of a model as the one with the smallest p-value and the most influential factor as the one with the largest absolute coefficient. We summarized the most significant and influential factors given by each of the growth-stage models into the following table:



We can see that the precipitation is highly influential in the corn growth as it is the most influential factor of all the models. For the growth stages, we believe that the stages of dough, silking, and denting play a relatively significant role in the growing processes which the farmers should pay special attention to.

From the analysis of the innovative technology, we found that the corn yield has a general increasing trend over the years except for several valleys at the years 2002, 2005, 2012, and 2015. We found that all of these yield valleys were caused by the abnormal climate of those years, so it is reasonable to believe that increasingly innovative technology is the push behind the ascending yield.

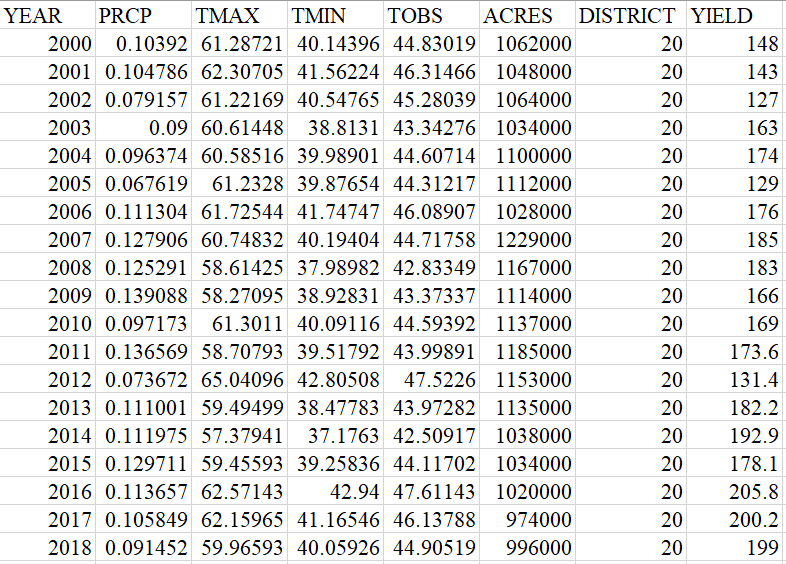
Because higher corn yield is linked with a lower selling price according to the economic supply and demand curve as the demand for corn yield is relatively stable across the years. We recommend farmers to pay attention to weather conditions, especially in the dough stage. We can also advise farmers to be sensitive to any abnormal rainfall because we find droughts or floods can be destructive. If we notice that a flood or drought happens in an early growth stage, we would recommend farmers to plan ahead for expected low corn yield.

**APPENDIX**

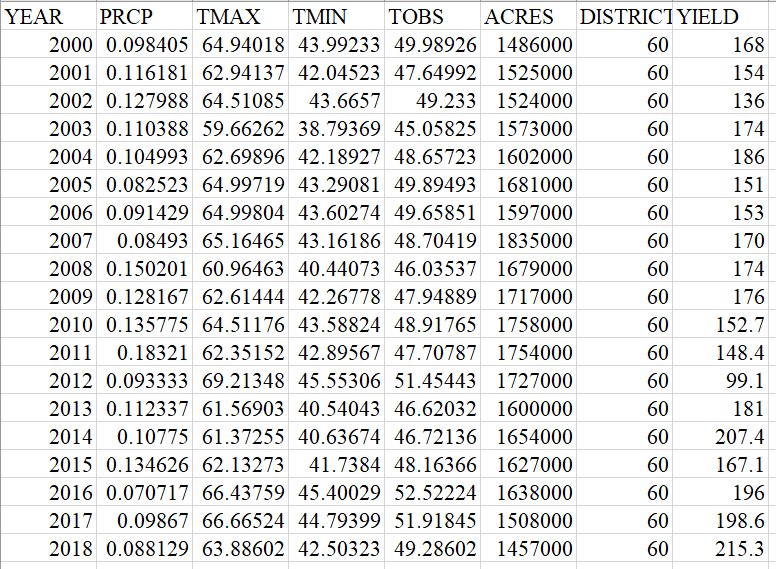
**7.1 Relevant Tables**

**7.1.1 Datasets for Building the Yearly Averages Models**

**District 20**

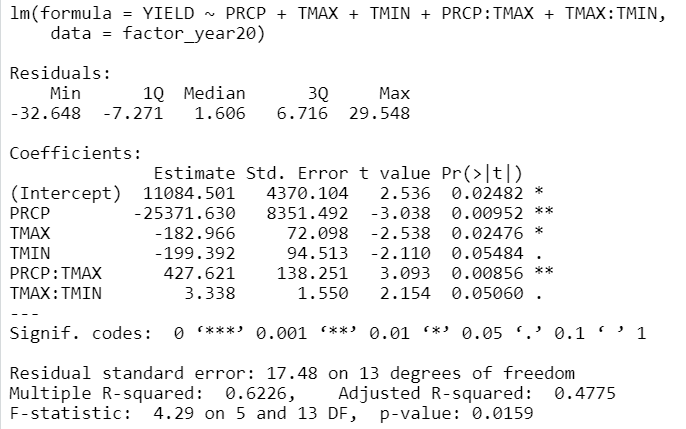
****

**District 60**

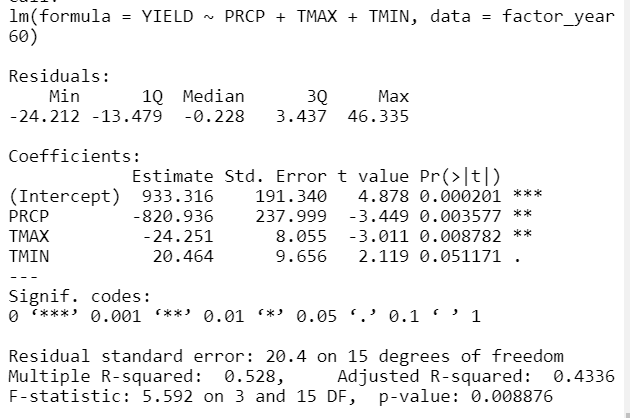
****

**7.1.2 R Outputs of the Yearly Averages Models**

**District 20**

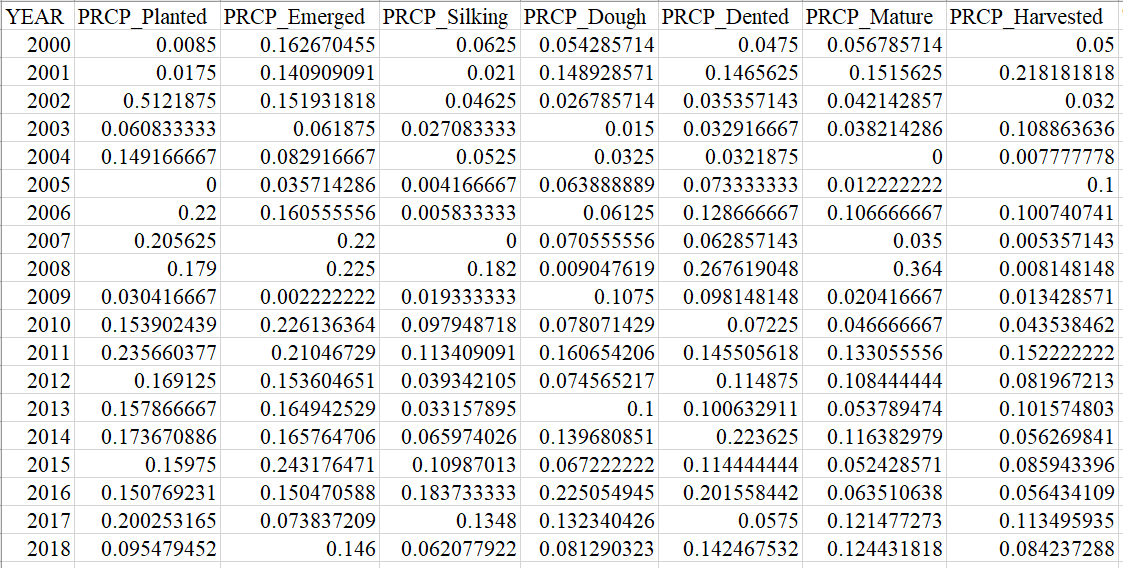


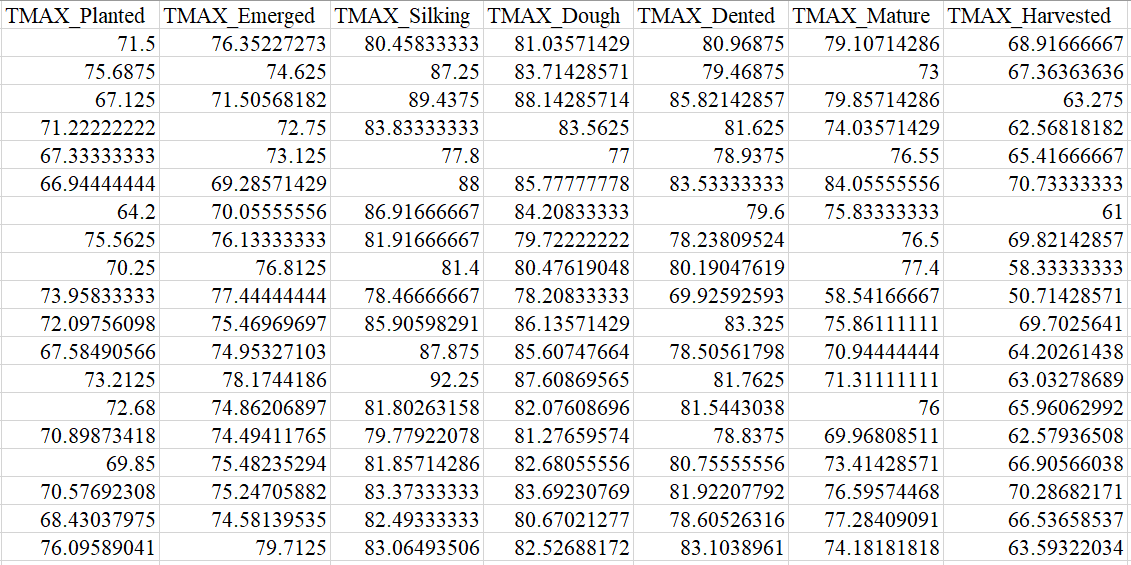
**District 60**

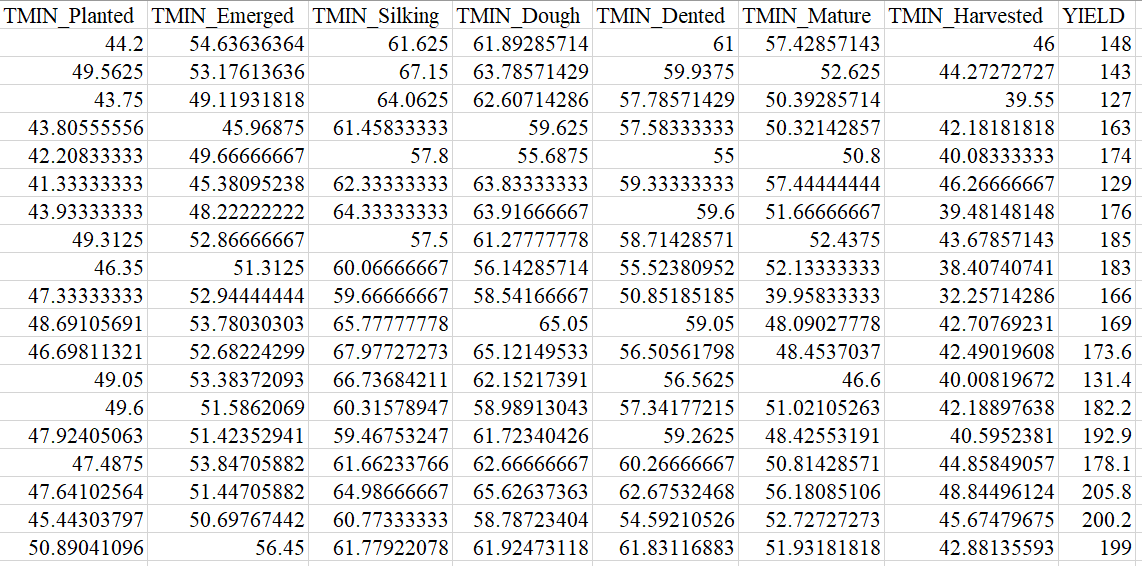


**7.1.3 Datasets for Building the Growth Stage Models**

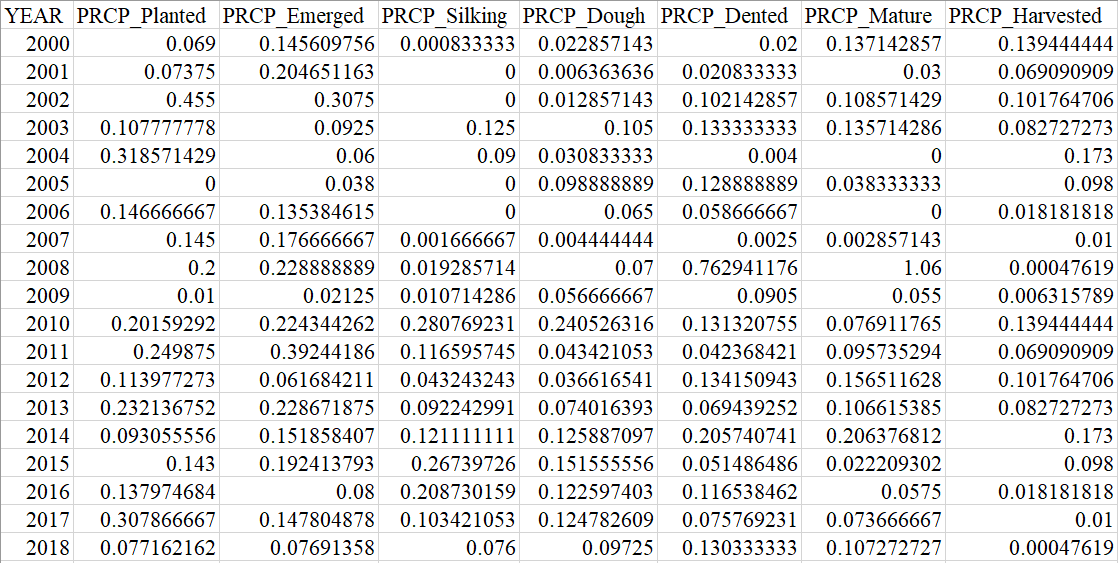
**District 20**

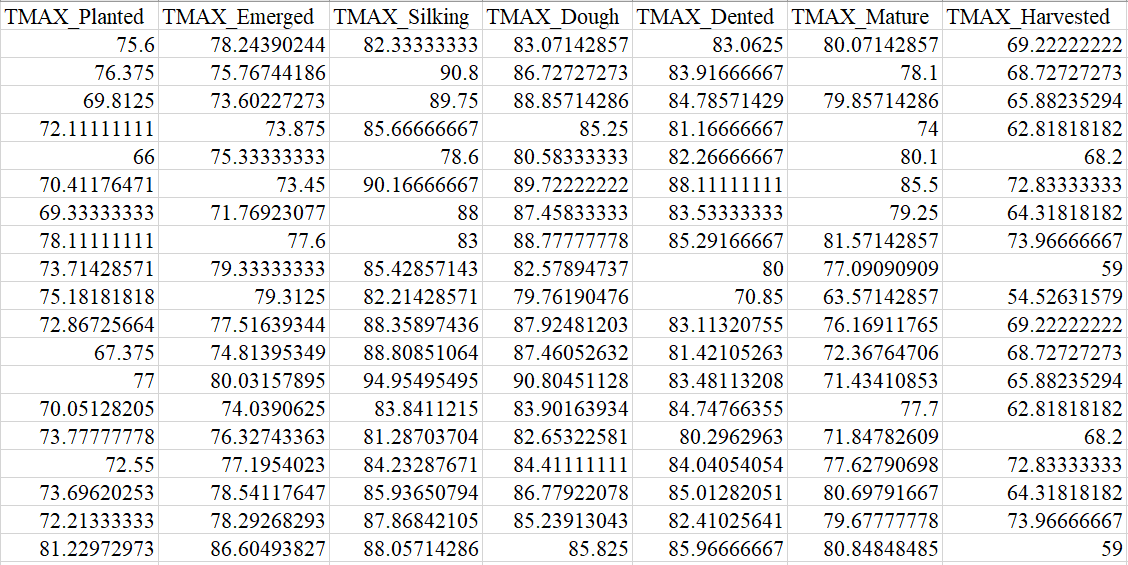


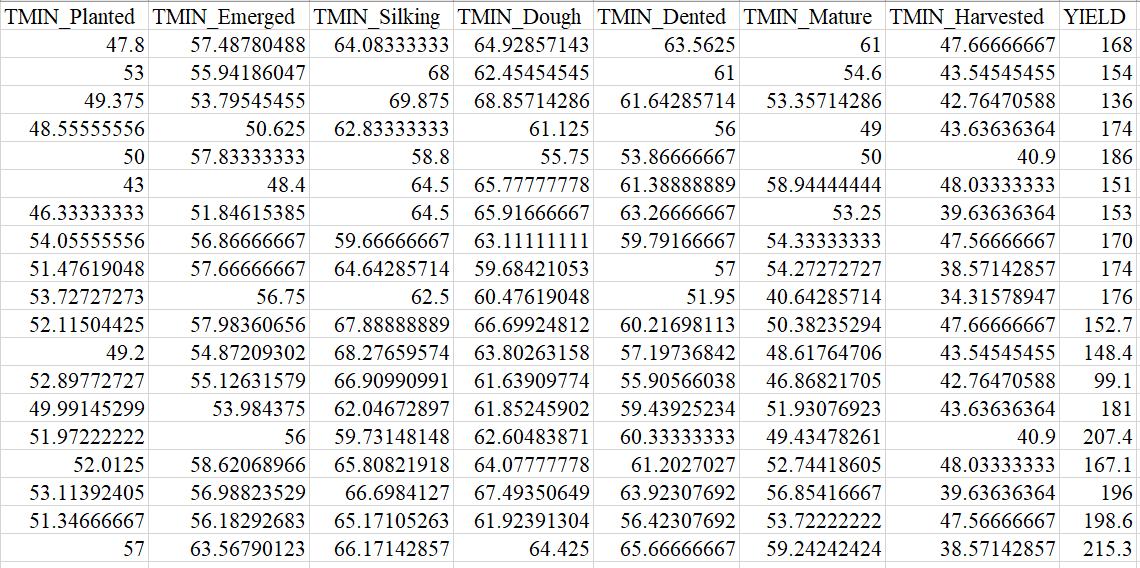




**District 60**

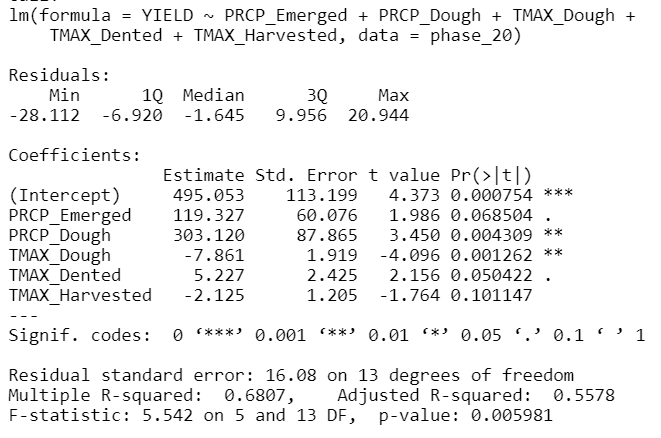




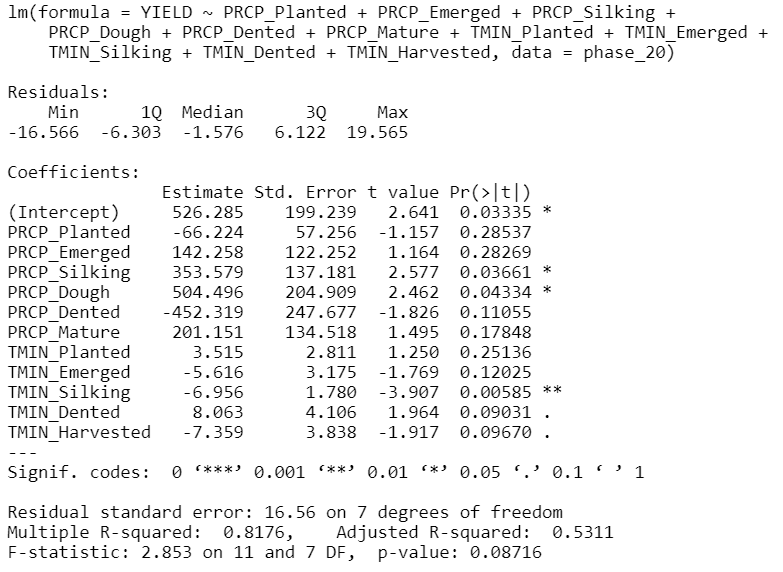


**7.1.4 R Outputs of the Growth Stage Models - District 20**

**The Model with PRCP & TMAX**

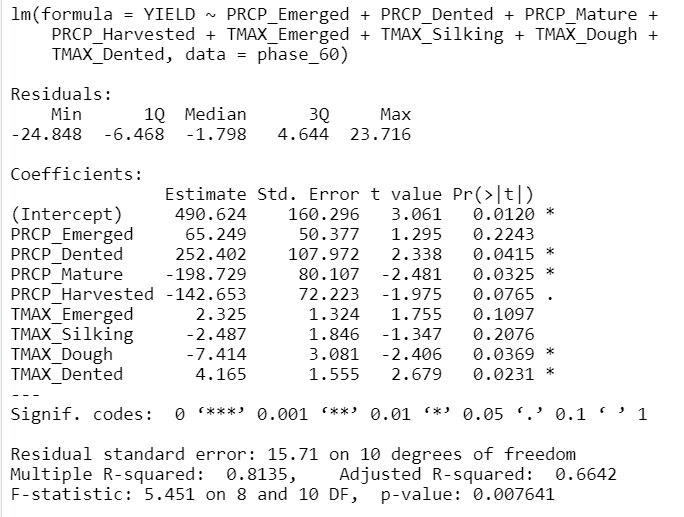


**The Model with PRCP & TMIN**

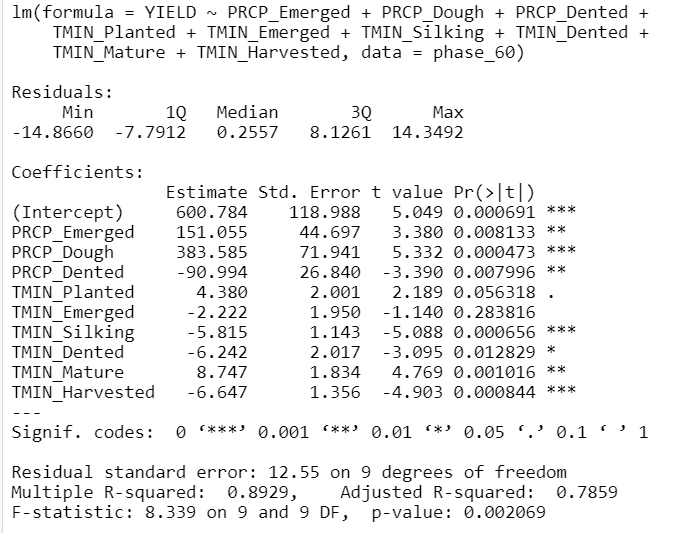


**7.1.5 R Outputs of the Growth Stage Models - District 60**

**The Model with PRCP & TMAX**



**The Model with PRCP & TMIN**



**7.2 Code**

The code for this report will be submitted in the zip file. Each file has commented explanation in the code, a description in the title of the file, and/or an indication of what it is used for in the report. Important data files will also be included in the zip file containing this report.

**7.3 Future Directions**

From the analysis of the precipitation variable section, incorporating some of the 4 created variables may help improve the accuracy of predictive models because it is less biased and focuses on critical times indicated by the Illinois State Water Survey

We realized that many variables are highly correlated in this study, which means that the multicollinearity problem occurs in our predictive model. In a regression model, we hope that we can learn how a one-unit change of independent variables can cause variability in the dependent variable. However, in our case, due to multicollinearity, the coefficient estimates can swing wildly based on which variables to include in the model. This is why we find that the value of coefficient for a variable varies greatly in two different models we estimated to predict the same corn yield ( For example, in the model that uses stage precipitation and maximum temperatures to estimate district 20’s corn yield, the coefficient for emerge PRCP is 119.327 and the coefficient for Dough PRCP is 303.12, but in the model that uses stage precipitation and minimum temperature to estimate district 20’s corn yield, the coefficient for *Emerge PRCP* is 142.258 and the coefficient for Dough PRCP is 504.496), so that it is difficult for us to know the real effect of each attribute on corn yield. Some of the estimated coefficients we got from the variable are contradicted with the point-out analysis for abnormal corn yields. Even if a temperature variable has a positive coefficient, this does not mean that higher temperature is always the better. In the pull-out factor analysis, we realize that high temperature shall cause drought, which is actually detrimental towards corn yield.

Another limitation we had with the model is that we only had 19 years’ observations, which is too few for us to predict a good model, if we want to better understand how temperature and precipitation are impact corn yield, more observations are needed in order to predict a more accurate model.

Moreover, for the innovative technology, though we already find that innovative technology plays a role in the increasing corn yields throughout the years, figuring out how to quantify the technology factor and putting it into the predictive models to further explore its impact on corn yield still require future relevant analyses.

**7.4 References**

*Hollinger and Angel - ‘Weather and Crops’*

<http://extension.cropsciences.illinois.edu/handbook/pdfs/chapter01.pdf>

This source was used to indicate critical stages of corn growth to focus on for the precipitation variable.

*State Climatologist Office for Illinois - ‘Drought Trends in Illinois’*

<https://www.isws.illinois.edu/statecli/climate-change/ildrought.htm>

This source was used to indicate that 2005 and 2012 were droughts using the Palmer Severity Index.

**7.5 Group Member Contributions**

Shihao Duan worked on the Introduction part, which includes Background & Motivation and Data Preparation.

Smruthi Iyengar worked comparing the temperatures in both district 60 and district 20. She created four violin plots in R to visualize the differences in temperature. She also conducted the T-test to determine if there were any statistical differences in temperature. She worked on section 2.2 and the code in the temperature data file.

Kagen Quiballo worked on the analysis of the precipitation variable under the ‘Variable Analysis and Visualization’ section which discusses methods for creating 4 different precipitation variables at the following website (*tinyurl.com/STAT443grp6-prcp-data*) in parallel with Illinois State Water Survey article, as well as visualizing 2005 and 2012 drought effects on annual corn yield. He also worked on the ‘District Variance Analysis and Visualization’ section which uses 2 sample t-tests to compare the average yield and average harvested percentage differences across districts 20 and 60. Lastly, he wrote the majority of all sections in the appendix.

Doris Wang worked on processing and cleaning datasets, calculating averages, and creating summary tables to prepare the variables used in the predictive models. Building, analyzing, and evaluating all the predictive models using R (the code file about predictive models), and making the slides and writing the report for the predictive model part. In addition, she wrote the majority of the Data Preparation part, formatted, and created a title page for the report. She created the visualization of Corn Yield with respect to Year used in the section of Innovative Technology Analysis. She also prepared and presented all the tables in the Appendix.

Lucy Zhao worked on analyzing “How technology has affected corn yield” and abnormal precipitation and temperatures each year to figure out the cause of low yield in specific years. She created 10 histograms to examine the effects of the abnormal temperature and precipitation on corn yield. She is also responsible for the future direction part.