Final Project: Impacts to Corn Yield based on Weather Variation

Group 1: Rohan Dalmia, Isabel Fudali, Youngseok Hahm, Hyun Jae Lee, Bharath Raghavan

Table of Contents

I. Introduction

A. Background

John Deere is a company that manufactures agricultural, construction, and forestry machinery. Their product sales are dependent on changes in the various industries they produce machines for. With changing global climates, we are interested in the effects of weather variation on crop yields in order to better help their agriculture-based clients. Crop production in Illinois used to have irrigation solely based on precipitation, but due to changes in weather, we have seen an increased need for supplemental irrigation systems in order to maintain crop production during the growing season. Precipitation is a large contributing factor to crop yields across Illinois, but there are also many other factors that influence crop yield as well including seed genetics, land management, and other weather patterns. We have been given data to try to identify other influential weather patterns that may impact crop yields in the future.

The datasets we were provided were retrieved from the National Agricultural Statistics Service. It provides us with data from 20 weather stations across Illinois. There are 20 different datasets with many years of data. There was one weather station, SFM, which does not have data from the year 2017 so we will not be using it in our analysis. We are only concerned with the data from 2017 for our analysis which left us with 19 weather stations and lots of information. There was a supplementary file of crop yield total per site for 2017.

B. Objectives

Our task is to find out significant biophysical properties that affect crop yields in various climate locations in Illinois in 2017. This analysis will allow us to offer insights to customers about weather patterns to influence land management decisions. We initially had 25 continuous and 29 categorical variables related to weather details in Illinois. All these numerical contributes were used in our testing models as predictors. Some of the information these variables provided included time, wind speeds, air temperature, precipitation, potential

evapotranspiration, soil temperatures, soil radiation, and more. The description of variables is listed below (more details in Appendix 1).

- **year** = year
- **month** = month
- **day** = day
- **max_wind_gust** = maximum daily wind gust (miles per hour)
- **xwser** = error flag for maximum daily wind gust
- **avg_wind_speed** = average daily wind speed(miles per hour)
- **awser** = error flag for average daily wind speed
- **avg_wind_dir** = average daily wind direction (degrees, clockwise from north)
- **awder** = error flag for average daily wind direction
- **sol_rad** = total daily solar radiation (megaJoules per square meter)
- **soler** = error flag for total daily solar radiation
- **max_air_temp** = daily maximum air temperature (degrees Fahrenheit)

The distribution of our variables and some key descriptive statistics are shown in Fig. 1 and Table I, respectively

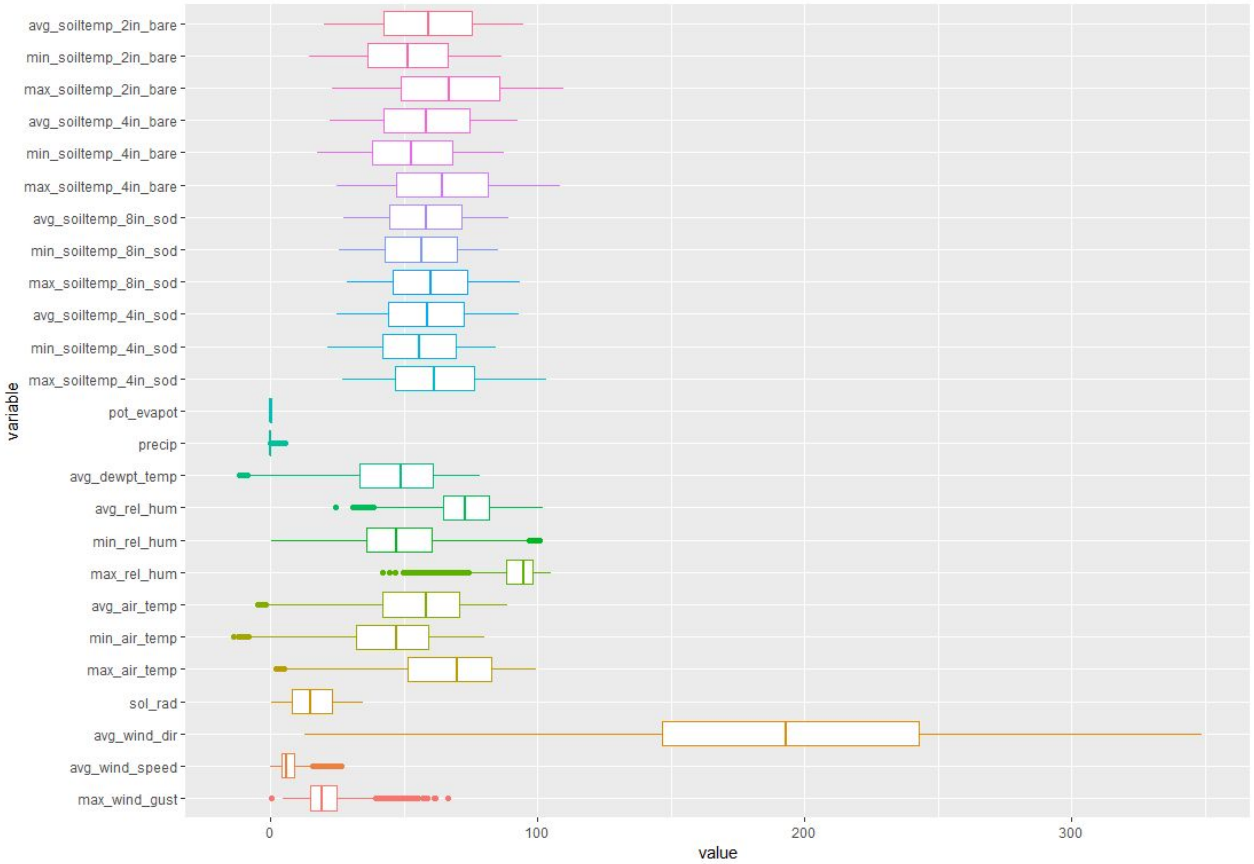


Fig. 1: Box plots showing the distribution of the important variables

Table I: Descriptive statistics of the most important weather variables

	avg_wind_speed	avg_wind_dir	sol_rad	avg_air_temp	avg_rel_hum	avg_dewpt_temp
Min.	0	13	0.4	-4.8	24.2	-11.8
1st Qu.	4.1	146.6	7.9	41.9	64.7	33.2
Median	6.1	193.15	14.8	58.4	72.8	48.8
Mean	6.876523506	190.7269588	15.50502031	55.59467499	72.93048462	46.00339524
3rd Qu.	8.8	242.8	22.9	70.8	81.7	61
Max.	26.5	348.9	34.9	88.7	101.9	78.6
	precip	pot_evapot	avg_soiltemp_4in_sod	avg_soiltemp_8in_sod	avg_soiltemp_4in_bare	avg_soiltemp_2in_bare
Min.	0	0	24.7	27.4	22.2	20.3
1st Qu.	0	0.05	44.2	44.4	42.3	42.4
Median	0	0.11	58.6	58.1	58.2	59
Mean	0.103826175	0.12196895	58.07087928	57.69634359	58.04583575	58.61813697
3rd Qu.	0.03	0.19	72.6	71.8	74.5	75.5
Max.	5.37	0.33	93.1	89.2	92.8	94.7

II. Method and Results

A. Data cleaning

In terms of data cleaning and exploratory data analysis, the first thing we did was append, subset, and clean our datasets to include only data from the years 2017 from the 19 weather stations. We merged all of the 19 datasets together, then subsetted the data to include only observations within the growing season as well - April 1st to September 30th. We cleaned the data by removing variables with missing values and unnecessary columns with metadata about the variables. We also removed some values that were hyphenated and having metadata indicative of Missing Data (M) and Estimated Data (E). The final cleaned dataset contains 3474 observations with 29 variables.

B. Understanding multicollinearity for variable selection

Once our data was cleaned, we proceeded to our analysis. The first thing we did was look at a correlation matrix of all the variables in order to identify highly correlated variables and potential multicollinearity.

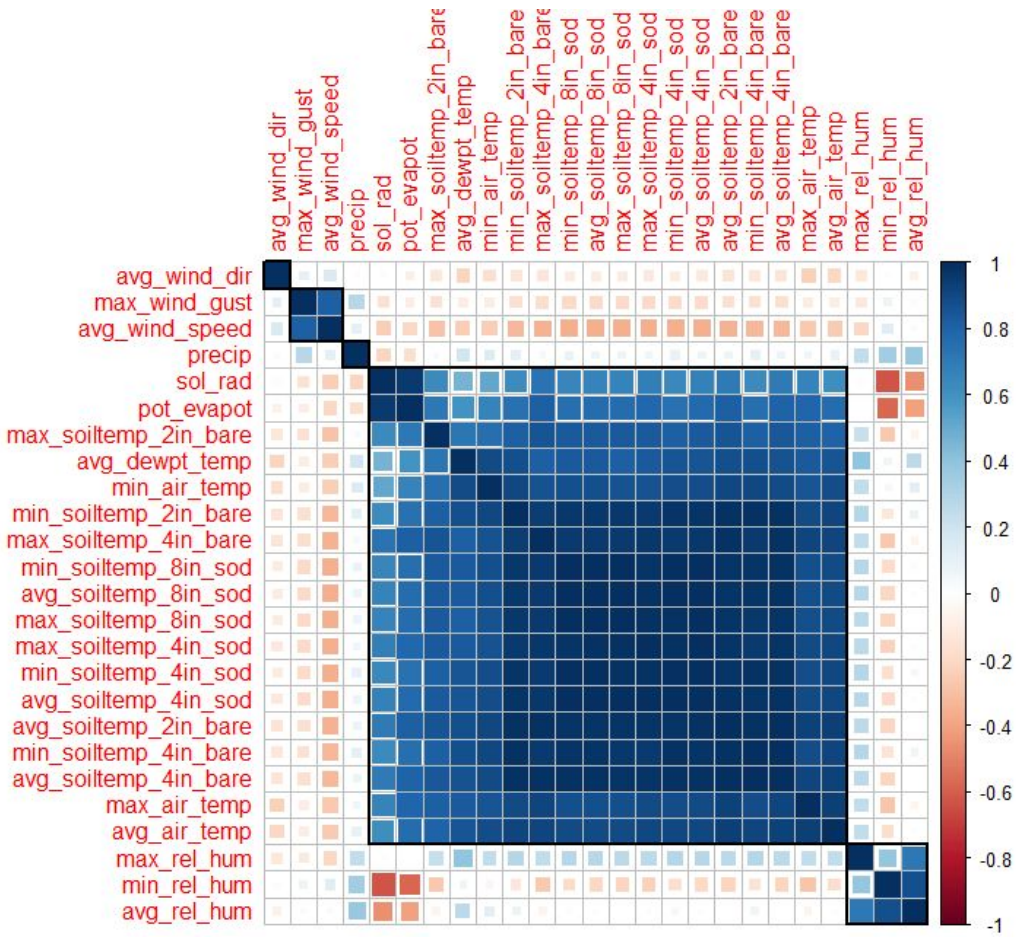


Fig. 2: Correlation matrix showing the level of correlation between weather variables.

The correlation matrix (Fig. 2) shows correlations between all of the variables excluding time variables - day, month, year. We used this to determine the subset of predictors to use for our model. Using this plot, we decided to subset our variable list using a threshold of 0.8 to remove highly correlated variables. Highly correlated variables make inference difficult and the potential model prone to overfitting. Therefore, we have decided to proceed with the following variables:

- sol_rad
- precip
- pot_evapot
- avg_rel_hum
- avg_air_temp
- avg_wind_dir

- avg_wind_speed
- avg_dewpt_temp
- avg_soiltemp_4in_sod

Instead of using the minimum/maximum extreme values, we decided to proceed with the average values. The reason why we chose the average values for many of these metrics is because they were highly correlated. We also chose to include some varied weather parameters. When choosing the type of soil measurement, we had to do some extra research and came to the conclusion that we would use the 4 inch sod soil temperature. We found out that sod is an insulator, so it is less susceptible to changes in temperature which would lead to less variability within our data.

When looking at our original variables, we noticed that they were distributed differently, so when performing principal component analysis, we normalized and scaled our data to allow for comparison and effective analysis (shown in Appendix 2). As an overview for our data, we identified three different types of analysis to do. This included doing principal component analysis, cluster analysis, and penalized regression modeling for corn yield.

C. Principal Component Analysis and clustering

The first method we used to analyze our data was principal component analysis. Our approach was to first visualize the data to get a sense of how it can be clustered. Since, it is impossible to plot more than 3 dimensions, we chose PCA to reduce the dimensions and facilitate the process of visualization.

In order to better visualize our data, we decided to do an initial principal component analysis. This method of analysis reduces the dimension of large datasets and allows us to distinguish the impacts of certain variables on a model. Based on the scree plot (Fig. 3) we see that the first two components explain the most variation in the data.

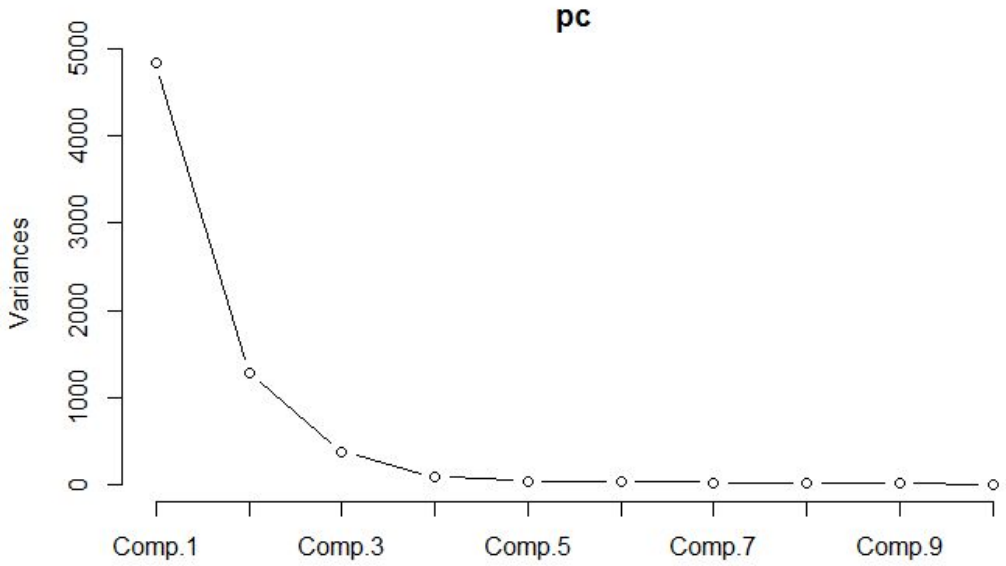


Fig. 3: Scree plot showing amount of variance captured based on principal components utilized

Using this dimensional reduction technique, we were able to visualize the clusters and do more in depth exploration of our results. Based on the scree plot and scatter plot (Fig. 4), we see that there is an elbow at the third principal component indicating that the first two principal components explain the most variation in the data. This method was also useful in identifying variable importance.

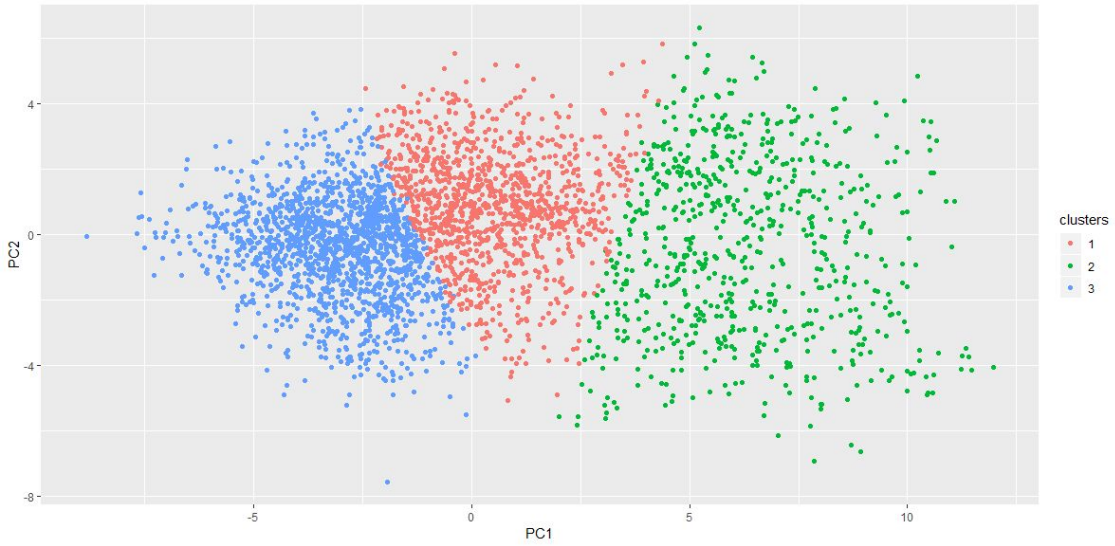


Fig. 4: Cluster analysis on 2 principle components

Looking at the coefficients of the principal components (Appendix 3), we can see that principal component one contrasts all of the temperature values along with some humidity values with precipitation, wind speed, and wind gust. There seems to be a distinction between those weather features. Similarly, the second principal component has different distinctions, but it is not clear as to what factors they are separating out.

After running the principal component analysis, we performed k-means and then drew an elbow plot in order to choose the optimal k value to proceed with. Based on the elbow plot (Fig. 5), we have decided that we will proceed using 3 clusters as it is the optimal point. At k = 3, the addition of another cluster does not give much better modeling of the data. The clusters are split amongst the variation in weather patterns and the clustering technique that we were using was based on k-means.

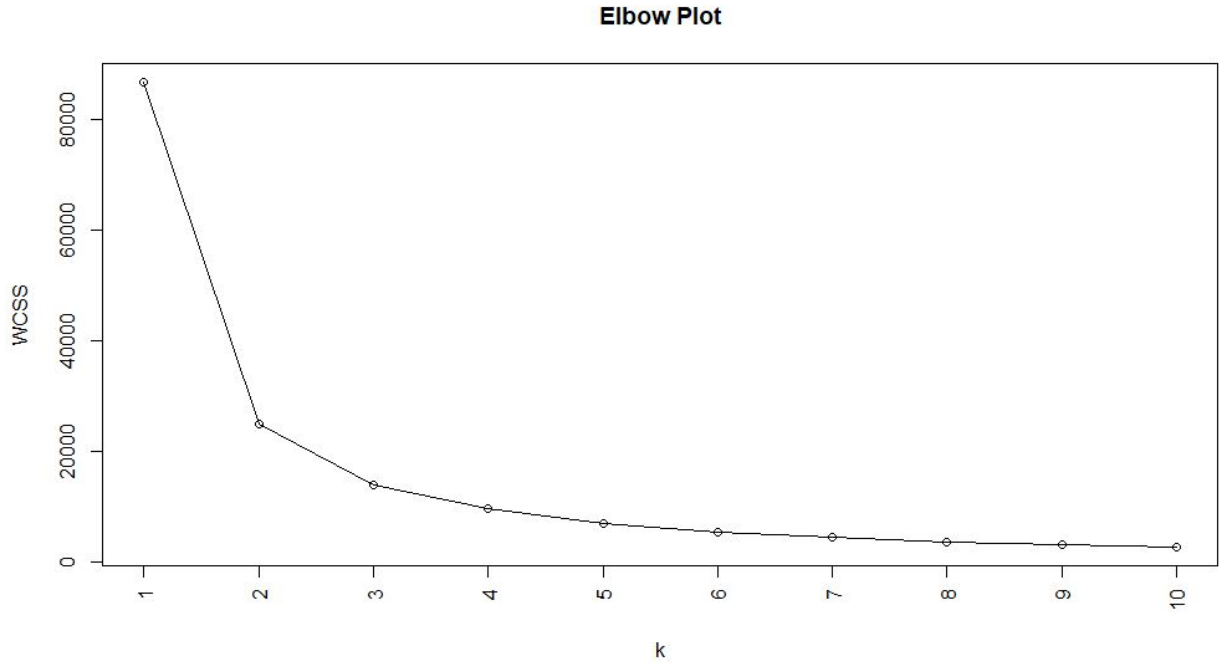


Fig. 5: Elbow plot illustrating the number of clusters (k) required.

Analyzing the elbow plot, we can see that 3 is the optimal value for amount of clusters. We created a visualization that shows the three cluster groups (shown above). After this, we did a more thorough analysis of each cluster. Each cluster is separated into commonalities amongst weather patterns. A more in depth analysis of this can help us determine trends and influences of these weather patterns on total crop yield for each weather station.

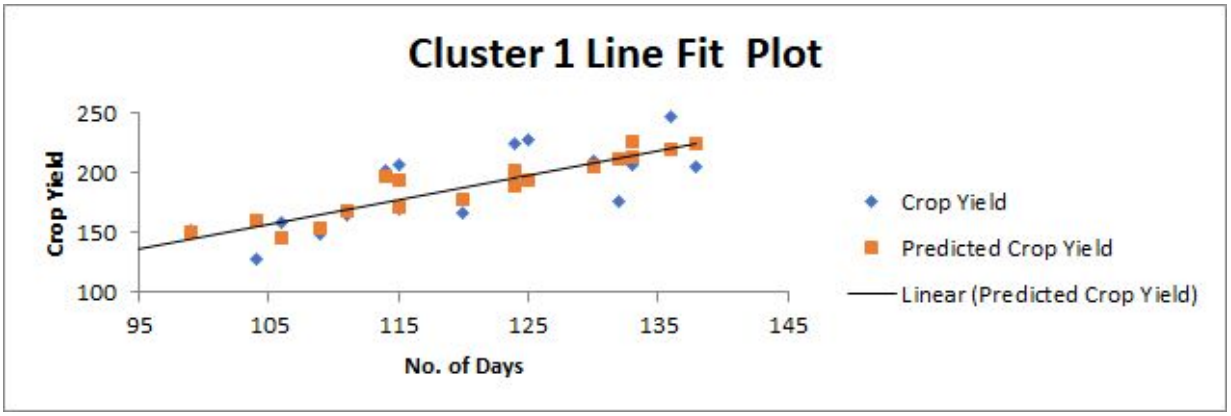
We used this data to try to gather insights within each cluster about correlations between frequency of a site in a cluster of weather trends and how that may potentially affect crop yield. With this we can determine what the ideal weather patterns are in order to maximize crop yield. Differences in weather patterns (based on the combination of conditions) may help determine which cluster of conditions has the largest influence on yield. The frequency of a site in a cluster may help identify the types of weather patterns that influences yield at that site. Our method of thinking is that if a cluster has a large number of observations for a specific site and a large crop yield, it may be indicative of better growing conditions. Below is the table that we used for analysis:

Table II. Cluster Group Frequency of Site and Crop Yield

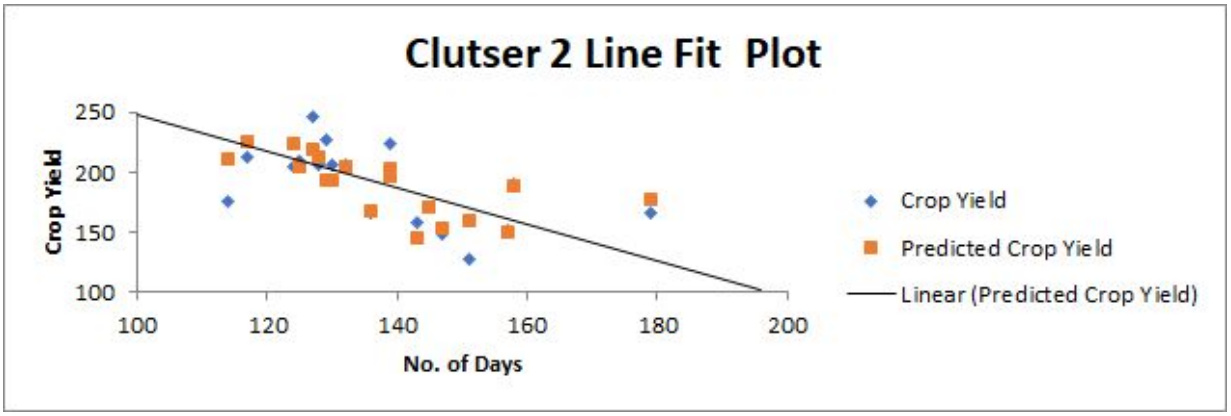
	Cluster 1	Cluster 2	Cluster 3	Crop Yield
bbc	130	125	110	209.4
brw	115	145	104	170.3
bvl	130	132	103	206.8
cmi	115	130	112	206.8
dek	138	124	103	204.1
dxs	99	157	102	151.9
fai	106	143	116	157.6
fre	133	128	104	206
frm	111	136	115	165.4
icc	125	129	111	227.9
llc	124	139	98	224.4
mon	136	127	102	246.7
oln	109	147	109	148
orr	124	158	83	191
rnd	104	151	104	127.9
siu	120	179	66	166.7
sni	114	139	102	201.8
ste	133	117	111	212.4
stc	132	114	119	176.4

The clusters are segregated by weather conditions. Initially, we did a simple linear regression on the frequency of the weather station for each cluster (Fig. 6) . The cluster group frequency of site and crop yield data that from above was used for analysis. We used ANOVA modeling to try to see if differences in frequency in each cluster affect crop yield. Performing a basic regression analysis on the frequency of days a site is in a cluster versus crop yield shows us the effects of weather conditions on crop yield.

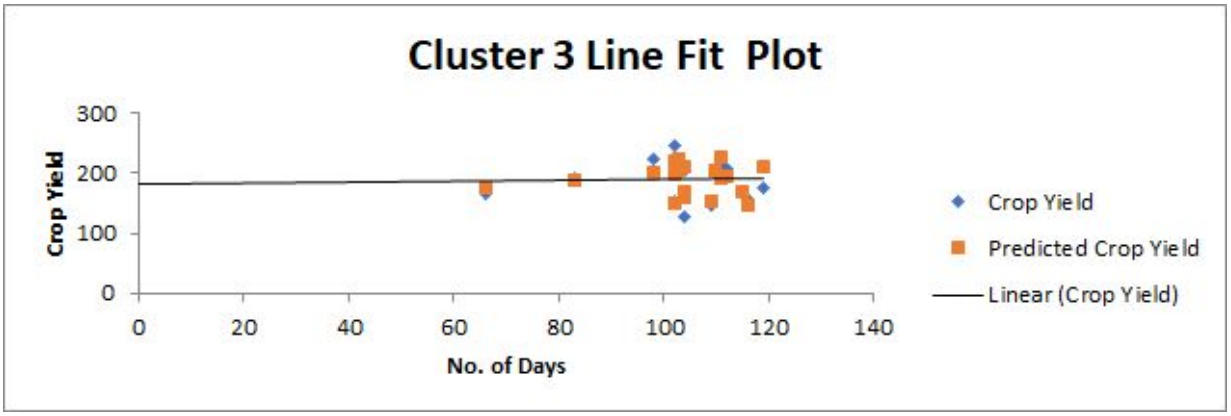
The graphs corresponding to the linear regressions are shown in Fig. 6. These may indicate that if a site has more days with weather conditions similar to that in cluster 1, the total crop yield may be higher. This shows that certain patterns that positively and negatively affect total crop yield. The frequency of days prevalent in each cluster per county has an affect on crop yield.



(a)



(b)



(c)

Fig. 6: Linear regression models illustrating the effect of the number of days a county in Illinois experiences weather patterns corresponding to each cluster on corn yield. (a) Number of days in cluster 1, (b) number of days in cluster 2, and (c) number of days in cluster 3.

D. Feature Engineering to eliminate the temporal nature of weather patterns

However, clustering our data by weather conditions was not the best approach for modeling crop yield. We encountered challenges distinguishing amongst the variations in weather. Therefore, we decided to proceed with a more time-based approach to modeling crop yield.

To tackle the time series aspect of the data, and bearing in mind the mismatch in timescales (i.e. weather data was daily, but corn yield data was yearly), we converted the 9 features that spanned approximately 183 days (April-September) into individual features. In other words, after the transformation we obtained 9*183 = 1647 features that described the final end-of-year crop yield for each county. This removed the temporal dependency and made it easier to measure individual daily influence on the total crop yield. That way we have rates for daily variables and can model the crop yield for each county based on just the one variable of time. We did this for all 19 counties and then we will be modeling the corn yield based on these features. An illustrative snapshot of the transformed features is shown in Table III.

Table III. Snapshot of the feature transformation used to eliminate temporal dependency. Every weather variable on each day of the month is transformed into an independent feature.

site	avg_wind_speed_4-1-2017	avg_wind_speed_4-10-2017	avg_wind_speed_4-11-2017	avg_wind_speed_4-12-2017	avg_wind_speed_4-13-2017	avg_wind_speed_4-14-2017	avg_wind_speed_4-15-2017	avg_wind_speed_4-16-2017	avg_wind_speed_4-17-2017	avg_wind_speed_4-18-2017
bbc	3.2	10.7	8.4	7.5	10.3	9.3	13.4	8.5	3	10.3
brw	4.8	10.9	6.4	3.4	4.9	5.5	10.6	6.4	4.1	5.3
bvl	7.1	17.2	7.9	5.4	9.6	10.5	18.7	13.4	7	13
cmi	4.1	7.3	4.3	2.1	3.8	4	8	5.8	2.8	4
dek	3.5	14	6.1	6.9	11.2	10.8	16.8	6.4	5.3	12.8

E. Penalized Regression

We then proceeded to use a penalized regression to model corn yield based on these features. Since the number of features is much greater than the number of observations we have, there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all. By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training.

To approach this, we considered three different penalized regression models. All three approaches account for the high amount of variables. These methods help with multicollinearity and overfitting. We used 80 and 20 percent training and test sets respectively using our data. We used k-fold cross validation on the tuning parameter to attempt to get the best R^2 parameter. The regression models for Lasso, Ridge, and Elastic Net are visualized in Figs. 7, 8, and 9, respectively.

After using these three types of regressions, we compared their coefficients of determination in order to decide which model is the best (Table IV). Based on the results below, we see that elastic net performs the best. This makes sense as it combines both aspects of lasso and ridge which control for multicollinearity and regularization.

Table IV. Performance summary of Lasso, Ridge, and Elastic Net models

	Lasso	**Ridge**	**Elastic Net**
R^2	0.49	0.50	0.61
α	0.001	10	0.01

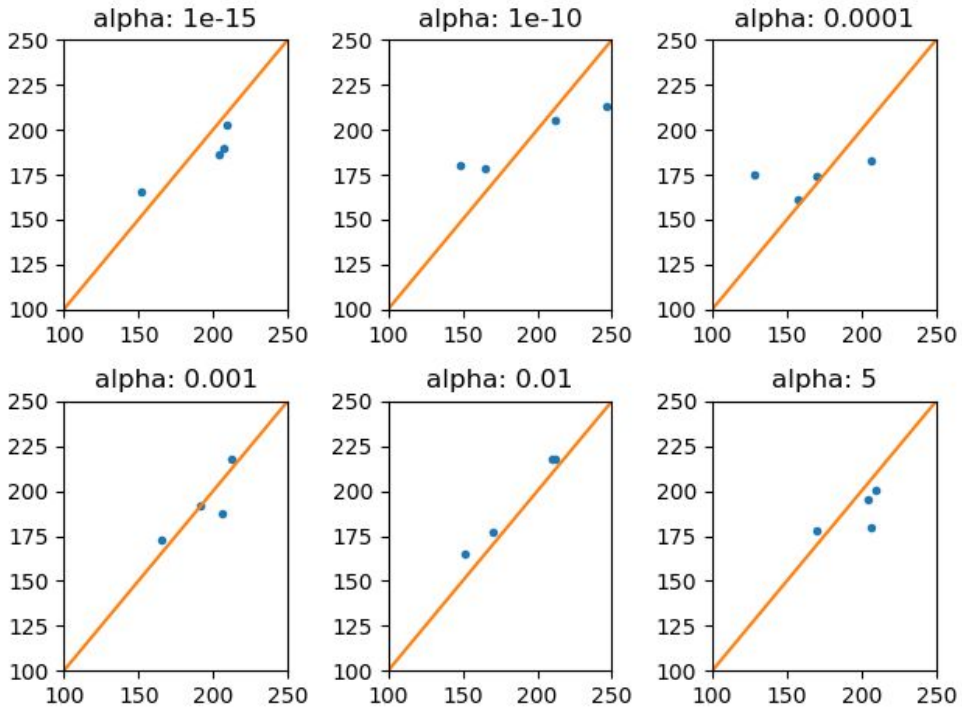


Fig. 7: Visualization of fit for Lasso regression for different values of the hyperparameter α. The plots how one instance of the K-Fold cross-validation.

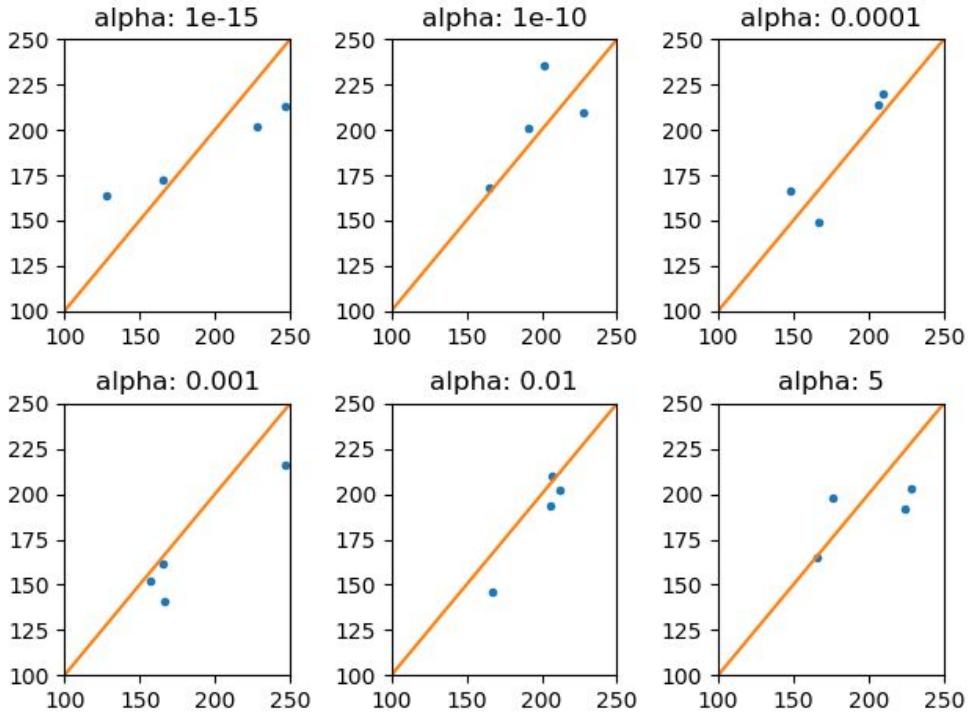


Fig. 8: Visualization of fit for Ridge regression for different values of the hyperparameter α. The plots how one instance of the K-Fold cross-validation.

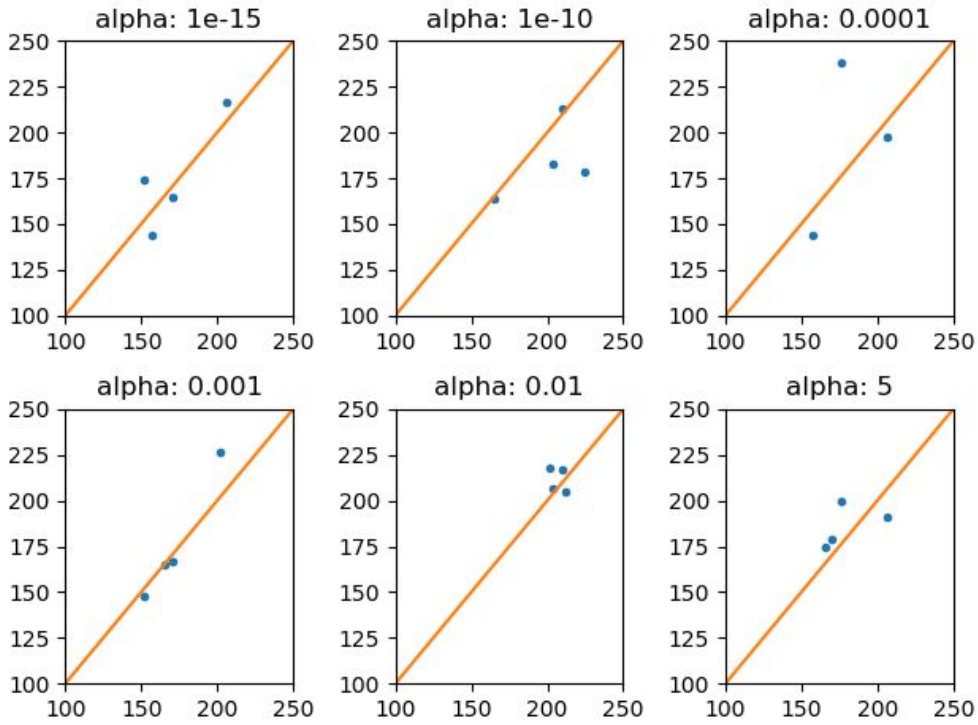


Fig. 9: Visualization of fit for ELastic Net regression for different values of the hyperparameter α. The plots how one instance of the K-Fold cross-validation.

Based on this daily analysis, we decided to average the values and analyze the results by week and month as well. Table V shows two representative weeks within the growing season (first week of April and last full week of September). This information provides us insight into the importance of each variable on total crop yield during a specific week. Analysis by week helps us analyze each variable's significant impact on a weekly level. We can address changes in weather patterns amongst each week.

Table V. Representative snapshot of feature weights for Elastic Net regression on a weekly basis.

Date	Avg air_temp	Avg dewpt temp	Avg rel_hum	Avg Soiltemp 4in_sod	Avg wind_dir	Avg wind_speed	Pot evapot	precip	sol_rad
4/1/2017 - 4/7/2017	-0.04157	0	0.09353	0	0.06649	0	-0.02876	0.03391	0.00641
9/23/2017 - 9/29/2017	0.03911	-0.00117	0	0	0.03548	0	0.01041	0.06274	0.21716

III. Discussion and conclusion

After modeling our data using various penalized regression processes, we are able to conclude that there are differences in weather patterns varying on time. Based on our findings, we decided to look into some recommendations we could potentially give customers. Collectively, we decided that making recommendations for a monthly basis would be more realistic. We initially analyzed variable impacts on a weekly basis, but practical implementation would probably be on a monthly basis. Also, weeks are highly variable so based on that and the fact that managing so much land would be difficult on such a frequent basis, we proceeded with making recommendations by month for practicality purposes.

Table VI. Monthly summary of recommendations for land management

Month	Influential Variable	Recommendation
April	Precipitation (-)	Decrease artificial irrigation
June	• Precipitation (+) • Wind direction (-)	• Increase artificial irrigation • Wind screens
August	Air temperature (-)	Cooling mist

Some of our recommendations include weather-mediating suggestions per month. One example is that for the month of April, the coefficient for precipitation was negative, meaning that it had a negative impact on crop yield, so during this month (despite it being in the growing season), we would advise to decrease (or not use) artificial irrigation. In our opinion, controlling these factors would influence crop yield.

For future improvements, we would have liked to analyze the data geospatially. Unfortunately, none of us are familiar with geographic information systems or similar software.

Next steps include fine tuning the model in order to make it better for predictive modeling. We mostly gathered insight from the values of the coefficients of the model, but making a better predictive model would be a future improvement. Other improvements would

have been to have looked at the weather data from the other years we were provided to compare and see how weather trends have changed annually. Working with an agronomist could have helped identify correct methodology and reasoning behind variable selection and interpretation.

Another thing we could have done was look at USDA/Climate center data and compare the growing season data with potential yield data to find out whether our model accurately predicts ideal yield. We could also use this information to see if we lose yield based on weather patterns within stages of the growing season.

Some challenges we faced was that recommending solutions based on only one year of data is difficult to do as there is no basis of comparison. Trying to provide time/weather based solutions without a good basis of annual comparison is quite difficult as we focused on the significance of the variables themselves. Another limitation is that some weather factors are beyond our control and don't have practical artificial methods to control them.

Despite all of this, the insights from our overall data analysis could be used to help individualize field management to maximize crop yields in various areas of weather patterns. Customers can use this information to decide whether to introduce artificial irrigation or to adopt new technologies in Illinois if weather patterns continue to change. Weather is constantly changing, so continuing this sort of analysis on a more granular level could offer better insight in the future.

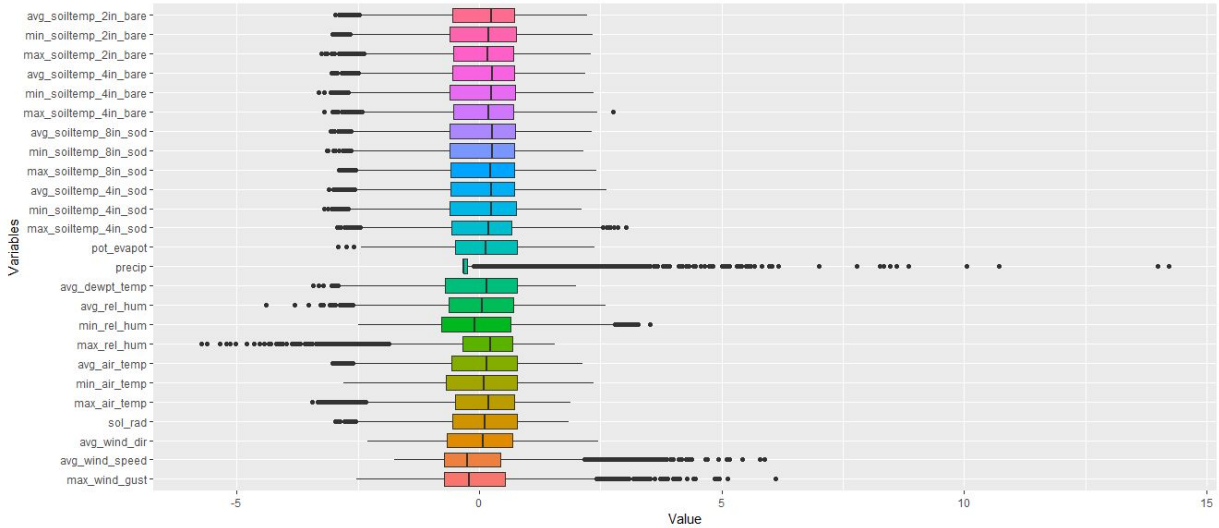
Appendices

I. Extending descriptions of initial variables

xater = error flag for daily maximum air temperature
min_air_temp = daily minimum air temperature (degrees Fahrenheit)
nater = error flag for daily minimum air temperature
avg_air_temp = average daily air temperature (degrees Fahrenheit)
aater = error flag for average daily air temperature
max_rel_hum = daily maximum relative humidity (percent)
xrher = error flag for daily maximum relative humidity
min_rel_hum = daily minimum relative humidity (percent)
nrher = error flag for daily minimum relative humidity
avg_rel_hum = average daily relative humidity (percent)
arher = error flag for average daily relative humidity
avg_dewpt_temp = average daily dew point temperature (degrees Fahrenheit)
adper = error flag for average daily dew point temperature
precip = total daily precipitation (inches)
pcer = error flag for total daily precipitation
pot_evapot = total potential evapotranspiration (inches)
pevaper = error flag for total potential evapotranspiration
max_soiltemp_4in = daily maximum 4-inch soil temperature under sod (degrees Fahrenheit)
xst4er = error flag for daily maximum 4-inch soil temperature under sod
min_soiltemp_4in = daily minimum 4-inch soil temperature under sod (degrees Fahrenheit)
nst4er = error flag for daily minimum 4-inch soil temperature under sod
avg_soiltemp_4in = average daily 4-inch soil temperature under sod (degrees Fahrenheit)
ast4er = error flag for error flag for average daily 4-inch soil temperature under sod
max_soiltemp_8in = daily maximum 8-inch soil temperature under sod (degrees Fahrenheit)
xst8er = error flag for error flag for daily maximum 8-inch soil temperature under sod
min_soiltemp_8in = daily minimum 8-inch soil temperature under sod (degrees Fahrenheit)
nst8er = error flag for daily minimum 8-inch soil temperature under sod
avg_soiltemp_8in = average daily 8-inch soil temperature under sod (degrees Fahrenheit)
ast8er = error flag for error flag for average daily 8-inch soil temperature under sod
max_soiltemp_4in_bare = daily maximum 4-inch soil temperature under bare soil (degrees Fahrenheit)
xst4bareer = error flag for daily maximum 4-inch soil temperature under bare soil

min_soiltemp_4in_bare = daily minimum 4-inch soil temperature under bare soil (degrees Fahrenheit)
nst4bareer = error flag for daily minimum 4-inch soil temperature under bare soil
avg_soiltemp_4in_bare = average daily 4-inch soil temperature under bare soil (degrees Fahrenheit)
ast4bareer = error flag for error flag for average daily 4-inch soil temperature under bare soil
max_soiltemp_2in_bare = daily maximum 2-inch soil temperature under bare soil (degrees Fahrenheit)
xst2bareer = error flag for daily maximum 2-inch soil temperature under bare soil
min_soiltemp_2in_bare = daily minimum 2-inch soil temperature under bare soil (degrees Fahrenheit)
nst2bareer = error flag for daily minimum 2-inch soil temperature under bare soil
avg_soiltemp_2in_bare = average daily 2-inch soil temperature under bare soil (degrees Fahrenheit)
ast2bareer = error flag for error flag for average daily 2-inch soil temperature under bare soil
site = station name

II. Remedies for unevenly distributed variables



When performing principal component analysis, we decided to scale the data to remedy the high variability in distribution of our variables. This boxplot shows the scaled and normalized distributions of our variables. One thing we notice is that the interquartile range of precipitation is smaller than other variables. This newly scaled and normalized data will help us find out which variables impact corn yield the most.

III. PCA Coefficient Analysis

	max_wind_gust	avg_wind_speed	avg_wind_dir	sol_rad	max_air_temp	min_air_temp	avg_air_temp	max_rel_hum	min_rel_hum	avg_rel_hum	avg_dewpt_temp	precip	pot_evapot	max_soiltemp_4in_sod	min_soiltemp_4in_sod	avg_soiltemp_4in_sod
PC1	0.077	0.111	-0.02	-0.10	-0.23	-0.22	-0.23	-0.04	0.025	-0.01	-0.21	0.040	-0.13	-0.24	-0.24	-0.25
PC2	-0.03	0.009	0.053	0.393	0.082	-0.10	0.016	-0.29	-0.45	-0.47	-0.21	-0.25	0.375	-0.00	-0.08	-0.04

These are the variable weights of the first two principal components. They are analyzed in the methods section above.

IV. References & Resources

1. Midwest Regional Climate Center, https://mrcc.illinois.edu/
2. Illinois Office of the State Climatologist, https://www.isws.illinois.edu/statecli/General/Illinois-climate-narrative.htm
3. USDA National Agricultural Statistics Service, https://www.nass.usda.gov/
4. Illinois Boundary shapefile, with country boundaries: http://clearinghouse.isgs.illinois.edu/data/reference/illinois-county-boundaries-polygons-and-lines

V. Code & Data

The supporting code and data are in the following files:

- Final_code_upto_modeling.docx
- cleaned_unprocessed_dataset.csv
- ml_code_2.py
- coeff.xlsx

VI. Individual contributions

Rohan Dalmia:

- Automated the process of reading data by writing code in RStudio
- EDA: Visualized PCA and correlation matrix, built box plot and elbow plot to assist data analysis
- Interpreted the results of penalized regression to understand the effects of coefficients on crop yield.
- Used excel to clean the data from Elastic net to convert data for weekly and monthly analysis

Isabel Fudali:

- Task delegation and time management
- Creation, organization, and writing of presentation (slides) and written reports (both midpoint check and final) in order to communicate all of our findings effectively

- Interpretation of Principal Component Analysis
- Results interpretation and land management recommendation suggestions

Youngseok Hahm:

- Cleaned the dataset merged from raw data by removing missing values or invalid values
- Worked on visualization for EDA and PCA with Rohan
- Generated additional dataset required for modeling

Hyun Jae Lee:

- Data description and exploratory data analysis
- Initial variable selection from correlation matrix and their coefficients
- Results interpretation

Bharath Raghavan:

- **Feature Engineering:** Section II D.
- **Regression:** Section II E. I wrote a python script for the penalized regression models in order to gauge model performances between Lasso, Ridge, and Elastic Net regression.
- **General Brainstorming:** The idea for the feature transformation, which is the crux of our project, would not have come to fruition where it not for the stimulating discussions we held as a group. My immense thanks to Angela Bowman, Dr. Glosemeyer, and all my group members for this tremendous learning opportunity.