# Atmospheric Quality vs. Corn Quality Over Time

STAT 3355.001
Group 20: Evan Meade, Jay Shah, Enrique Cardenas
Due Date: 11/17/2020

## 1. INTRODUCTION

Our purpose in investigating air pollution and corn crop quality was primarily to determine if there exists a correlation between the amount of trace gases (including pollutants) within the atmosphere and the quality of corn over time. We hypothesized that a positive relationship would exist between the two variables. In other words, a general increase in air quality, indicated by a decrease in the amount of trace gases would correlate with an increase in the quality of corn; alternatively, a general decrease in air quality would correlate with a decrease in the quality of corn. Other questions, elaborated on in the "Analysis and Findings" section, were also looked into.

Specific gases analyzed include ammonium ($NH_4$), sulfur dioxide ($SO_2$), sulfuric acid ($SO_4$), nitrate ($NO_3$), total nitrates ($TNO_3$), and nitric acid ($HNO_3$). These gases were measured as micrograms per cubic meter of air, dating from 1990 to 2020. A variety of sites for gas measurements was used. For one particular question, these sites were later filtered into four main divisions based on the site region: agricultural, coastal, natural, and urban.

The sites for the location of corn crops analyzed also greatly varied. The time frame researched was the same as the time frame for trace gases (1990-2020). The quality of corn was determined by our data source as being in one of five categories, including excellent, good, average, and so on. Our data set also indicated what percent of any given sample had a positive crop quality.

This group utilized data sets from both the United States Environmental Protection Agency (EPA) [1] and the United States Department of Agriculture (USDA) [2]. Using sources authorized by the U.S. government indicated the high level of reliability of our data sources. It was up to us to correctly apply the information provided to us and create reasonable answers to our questions.

## 2. DATA CLEANING

### 2.1 - **Atmospheric Data**, EPA CASTNET

The dataset used for Atmospheric Quality Analysis was from the EPA CASTNET database. The CASTNET data is sorted by site, as data is collected from various research and monitoring stations across the U.S.$_3$. Aggregate atmospheric trace data was used in this analysis for the last 30 years. Of the atmospheric data in the dataset, the following 6 air quality parameters were studied: Nitric Acid (HNO3), Ammonium (NH4), Nitrate (NO3), Total Nitrates (TNO3), Sulfur Dioxide (SO2) and Total Sulfates (SO4), as they were the most relevant air pollutants in the dataset to answer our questions [3]. In addition, the atmospheric dataset contained information on each site's location, id, date of the measurements starting and stopping. Additionally, the data was further filtered to only include data from 54 reporting locations in the dataset (referred to as

sites) to match agricultural data and provide a balanced variation of locations types. The locations were sorted into 4 subcategories, based on the geographical location of the reporting site, which were: Agricultural, Natural, Coastal, Urban. The determination was made upon the land use classification given by the EPA for each site as well as proximity to urban areas or oceans. These classifications were made to analyze the effects of geography and human activity. Ultimately this resulted in a dataframe using 11 variables and 12782 data points.

**2.2 - Corn Crop Quality Data**, USDA NASS

A defining challenge of our project was the fact that we sourced all of our data directly from primary sources, rather than relying on an existing clean dataset from a site like Kaggle. As a result, we had to spend a lot of time acquiring and wrangling our data into a convenient form, but ultimately gained a better understanding of its contents than we otherwise would have.

We began our analysis of the corn crop quality data by downloading it from the USDA NASS data portal. One challenge was the fact that data requests are limited to 50,000 observations, so we had to split up our request into three separate downloads: corn_quality__excellent_fair.csv, corn_quality__good_poor.csv, and corn_quality__verypoor.csv. In total, these files contained the historical data for all 5 crop quality categories for all sampling regions. Together, the files contained 84,023 observations of 21 variables. We united them by reading the files individually and then using a simple rbind function to create a master data frame of the corn crop quality observations.

Next, we dropped any column which had only 1 unique value, whether that value was null or simply unchanged across all observations. Such columns were useless to our analysis because they give us no differentiating information about our observations, so we dropped them to simplify our analysis. This left us with only 8 variables in a data frame we named corn.

At this point, the data was united and trimmed, but was still not in a totally optimal form since each row represented the percentage for only 1 of the 5 quality categories in a given sample. We preferred to have each row represent all data relating to a given sample, which would include geographic and temporal identifiers, along with all 5 quality category percentages. So we wrote a small loop which ran over the corn dataframe and placed each value into its appropriate position in a new data frame called corn_flat. Since older sampling records simply didn't include entries for values measuring 0, null values were filled with 0 in corn_flat. After this, we just did a few small polishing tasks, such as renaming the columns to more convenient names and recasting the week column as integers.

To summarize, at this point we had a data frame where each row corresponds to the crop quality data for a particular corn sample taken by the USDA. Our final cleaning task was to select which states' observations to use in our analysis. We had to find states which had a lot of samples in both our atmospheric and crop quality datasets. Here is a chart showing the number of observations for each state.

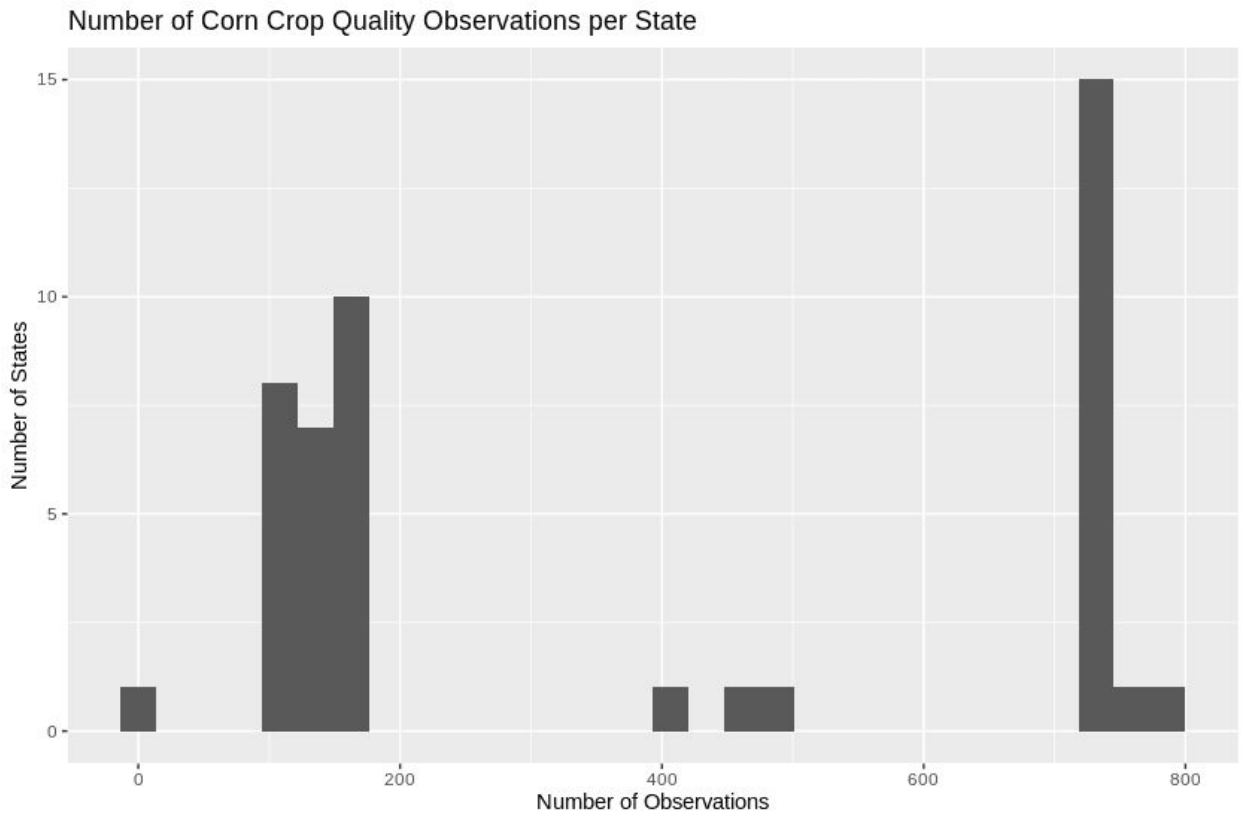


Figure 1: Number of corn crop quality observations per state

There is a clear division between the number of samples in the top 17 states and the number of samples in the rest of the states. To maximize the resolution of our time baseline, we took this set of the most frequently sampled states and manually cross referenced it with the available EPA CASTNET sites. By selecting the intersection of states with large representations in both data sources, we ended up with a set of 13 states, along with national totals for the corn data. This subsetting allowed us to focus our analysis on states with the most complete atmospheric and corn crop quality data, which would make any correlations between the two (or lack thereof) more apparent.

## 3. ANALYSIS AND FINDINGS

### 3.1 - Atmospheric Quality

1. **How has the concentration of atmospheric pollutants changed over time?**
   To answer this question, the 6 most prominent air pollutants (Nitric Acid (HNO3), Ammonium (NH4), Nitrate (NO3), Total Nitrates (TNO3), Sulfur Dioxide (SO2) and Total Sulfates (SO4)) in the EPA CASTNET database were plotted over time for the entire U.S. from 1990 [3]. The results were plotted as shown below.

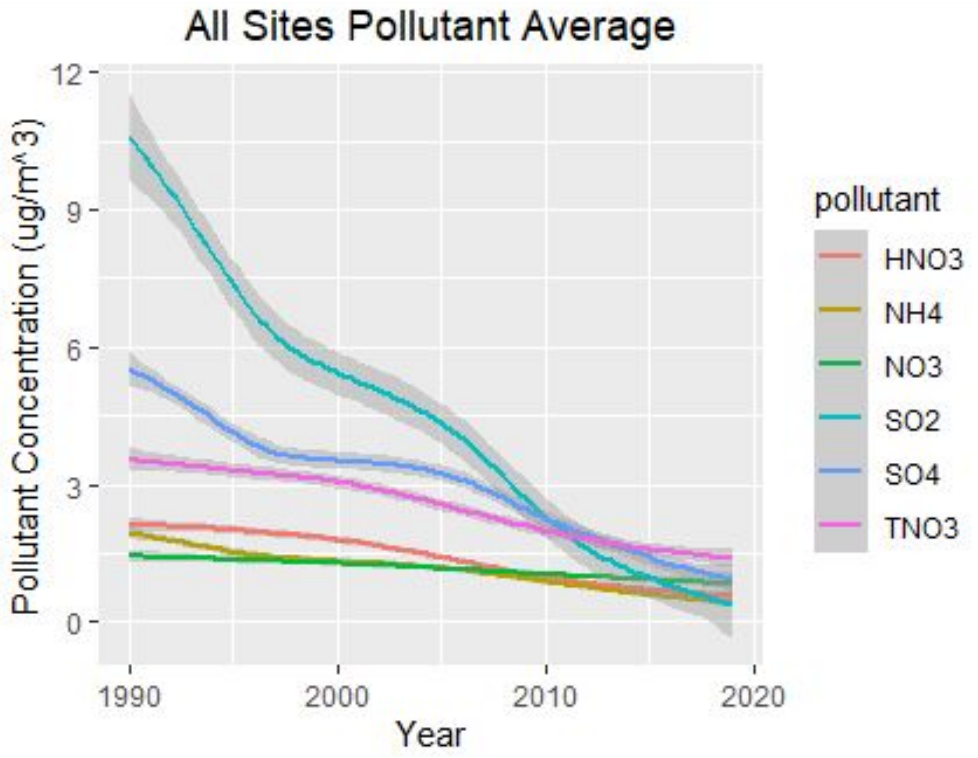


Figure 2: Overall pollutant concentrations in the U.S. since 1990

The average results for the U.S. show that all of these pollutants have reduced in quantity over time, likely attributed to environmental and emission regulations as well as green technology.

**2. How can geography and human activity affect concentrations of atmospheric pollutants?**

To study the effect of human activity and geography across the U.S., the data set was narrowed to look at just 2019 data to filter out the effects of new human settlement and population growth. Additionally, another variable was added to our data set, which was based on the type of the location the data was collected from. The reporting sites were classified into 4 categories: Agricultural, Coastal, Urban, and Natural. These classifications were made based on the primary land use surrounding the reporting site according to E.P.A. site info (for example, if the primary land use for a site was classified as Forest, Desert, Natural park, etc. the site was classified as natural, indicating lack of human settlement in the area.), as well as locational proximity to Oceans and Urban areas [4]. Sites with primary land use denoted as Agricultural or Range were classified as Agricultural. Sites with a metropolitan or urban area in the county were classified as urban, and sites in states bordering an ocean were classified as coastal. If multiple of these conditions were met, the site would be classified with the following priority order: Coastal, Urban, Natural, Agricultural. The reasoning for these classifications is that certain geographical factors are known to impact various atmospheric gas and particulate matter concentrations in the region. For example, oceans dissolve some gasses and pollutants better than other terrain types, and wind patterns following ocean currents are

known to carry pollutants from other continents with them [5]. Natural regions are generally lower population density as well as forests which affect air quality in the area. Urban areas have the highest human density, so this classification was created to see how urban areas differ in air pollution. Finally, sites that were classified as agricultural regions comprised a large number of the total sites, and showed the impact of another type of human activity on pollutants. The following 4 man made pollutants: Nitric Acid (HNO3), Ammonium (NH4), Sulfur Dioxide (SO2) and Total Sulfates (SO4) were then plotted based on the location type of where the data was collected.

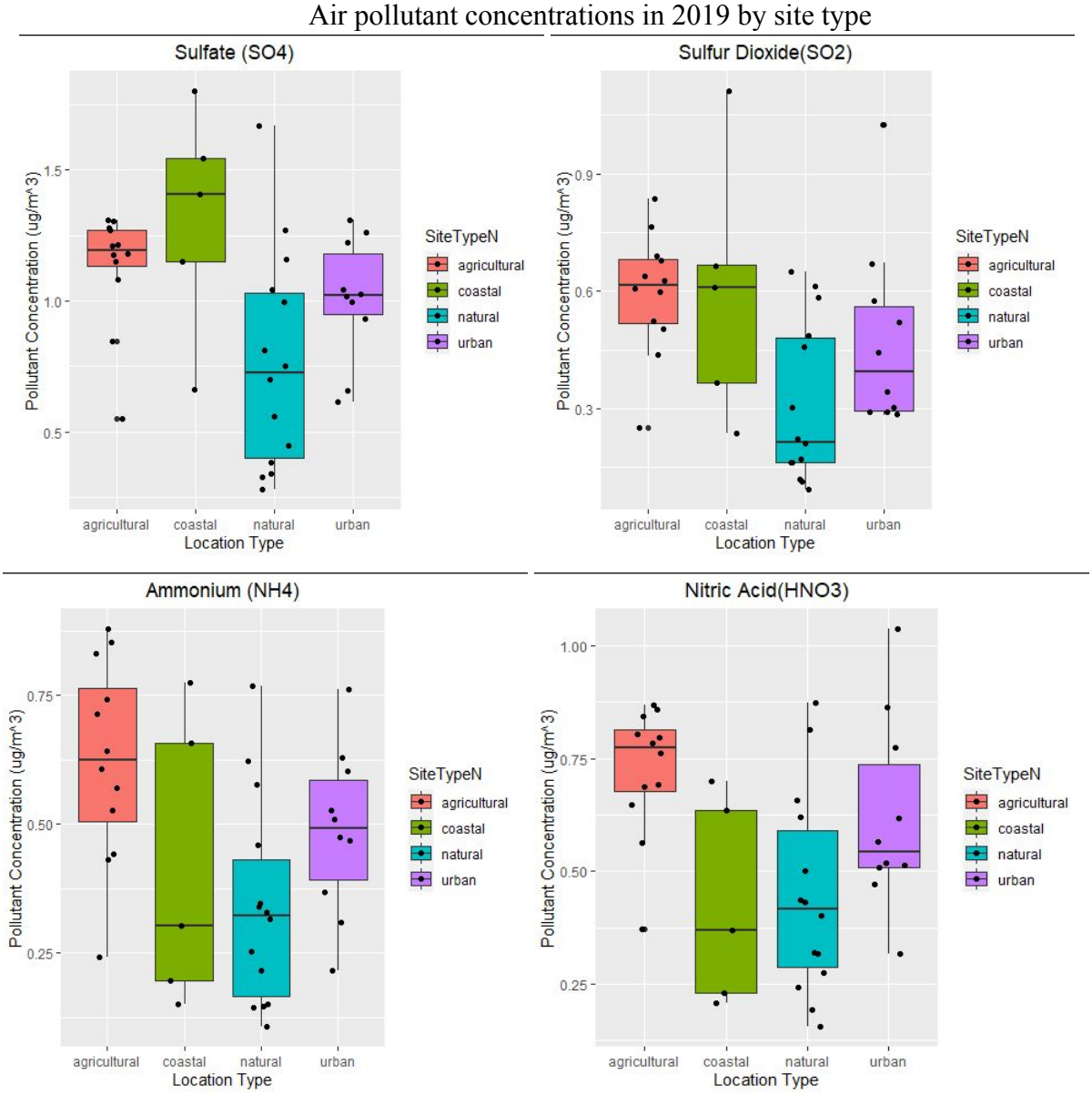


Figure 3: Pollutant concentrations by site classification in the U.S. in 2019

The figures above show the average pollutant concentrations based on the site type for each of the 4 pollutants described above. Overall, this shows that there is an impact of

geographical factors and human activity on pollutant concentrations. For Sulfates (SO4), an automobile pollutant, natural sites appear to have the lowest concentrations, which could suggest lower human populations in those areas could be attributed to this. A similar but less pronounced trend can also be seen for Sulfur Dioxide (SO2), another automobile and industrial pollutant, showing the impacts of geography [6]. For Ammonium (NH4) concentrations, Agricultural sites have the highest median amount of this pollutant, which is likely due to the fact that ammonium is a compound present in fertilizers and used in many agricultural processing applications, making it a significant component of pollution and runoff in these areas. The higher ammonium concentrations in agricultural sites indicates that this type of activity does indeed have an impact on pollution levels. Nitric Acid (HNO3) is also seen to be higher in agricultural areas. As a component of acid rain (along with sulfates), this could be adversely affecting crop quality, as described later in the report [7]. Overall, coastal regions showed the greatest variance in pollution levels, which is likely because the different coasts the sites are in are subjected to different climate and wind patterns as well as having varying population density in those areas. Agricultural areas have the least variance in pollutant concentrations which can be explained by the fact that they often occupy similar terrain and have similar types of human activity in them resulting in more consistent pollution levels in those areas. Additionally, a large proportion of the agricultural sites were in neighboring midwestern states.

### 3.2 - Corn Crop Quality

3. **How does national corn crop quality change over time?**

To complement our time series analysis of the atmospheric data, we wanted to begin our analysis of the corn crop quality dataset by plotting its changes over time. While our trimmed dataset included samples from 13 different states, we first wanted to investigate the national samples also included in the dataset. With this, we aimed to reveal any obvious national trends or events which may have impacted at an interstate level.

In figure 4 we see a scatter plot of every national sample, colored by quality level. As a reminder, these samples are sorted into 5 ordered quality levels as a percentage. In order of decreasing quality, these are: excellent, good, fair, poor, and very poor.

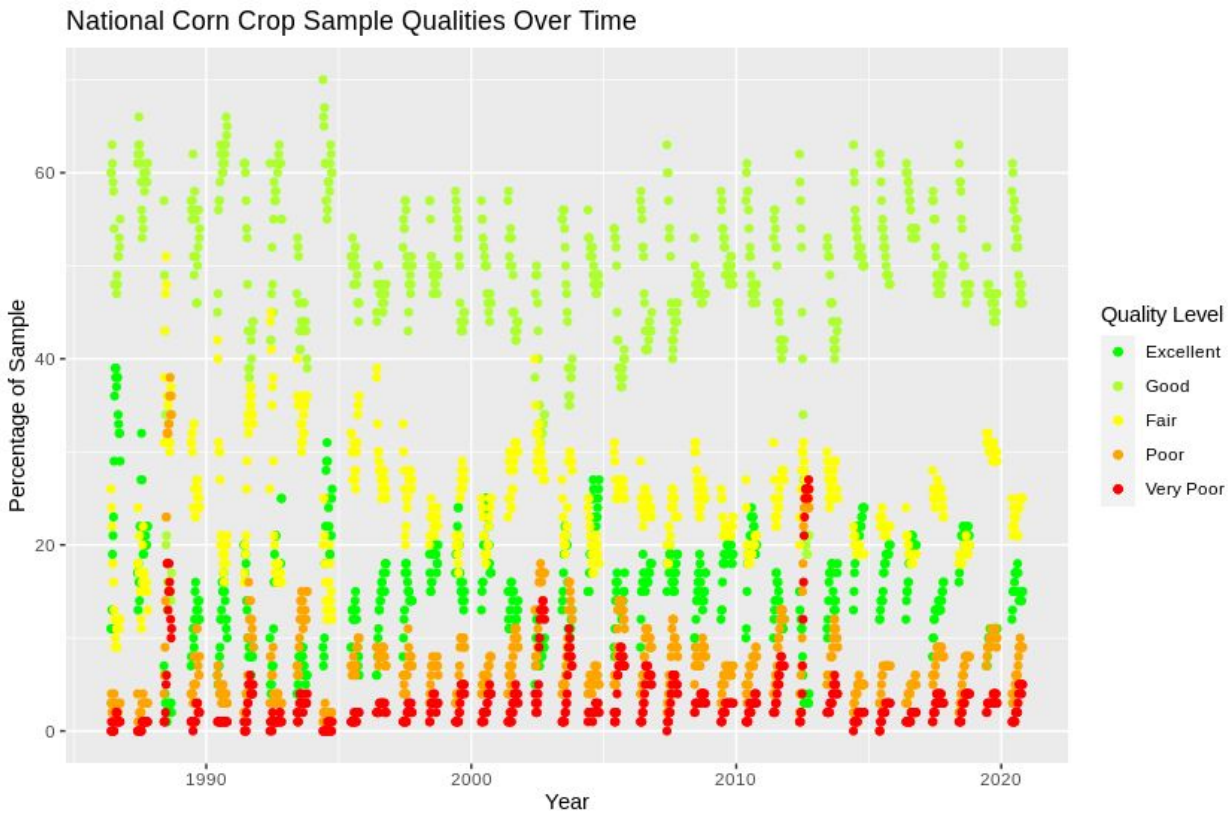


Figure 4: National corn crop sample qualities over time

As we can see in figure 4 above, while there is certainly noise in the sample values for each quality category over time, values are still largely banded and do not change much over time. For instance, the "good" quality values are largely confined to the 40-60% range, while the "very poor" quality values are generally below 10%. Of course, this is not a perfect characterization, and we can definitely see a number of spikes, particularly in lower quality categories. Two such spikes can be observed around 1988 and 2012, where the "very poor" and "poor" quality values reach the 20% mark. The 2012 spike is likely due to historic droughts caused by back-to-back La Niña episodes [8] . Such droughts would likely correlate with lower crop growth, leading to overall poorer quality than in a normal year. The 1988 spike is also likely due to widespread droughts decreasing corn growth and overall quality [9]. The fact that these significant ecological events are reflected in our data makes us confident that we have selected a useful metric for analyzing the impact of environmental conditions on crop health.

After considering outlier spikes, we wanted to turn our attention back to long term national trends in corn crop quality. To simplify analysis in our plots and remove noise from each quality level's data, we performed a locally weighted regression on each quality level in the national sample subset. The resulting smooth curves are shown below in figure 5, where they are plotted with 95% confidence intervals.

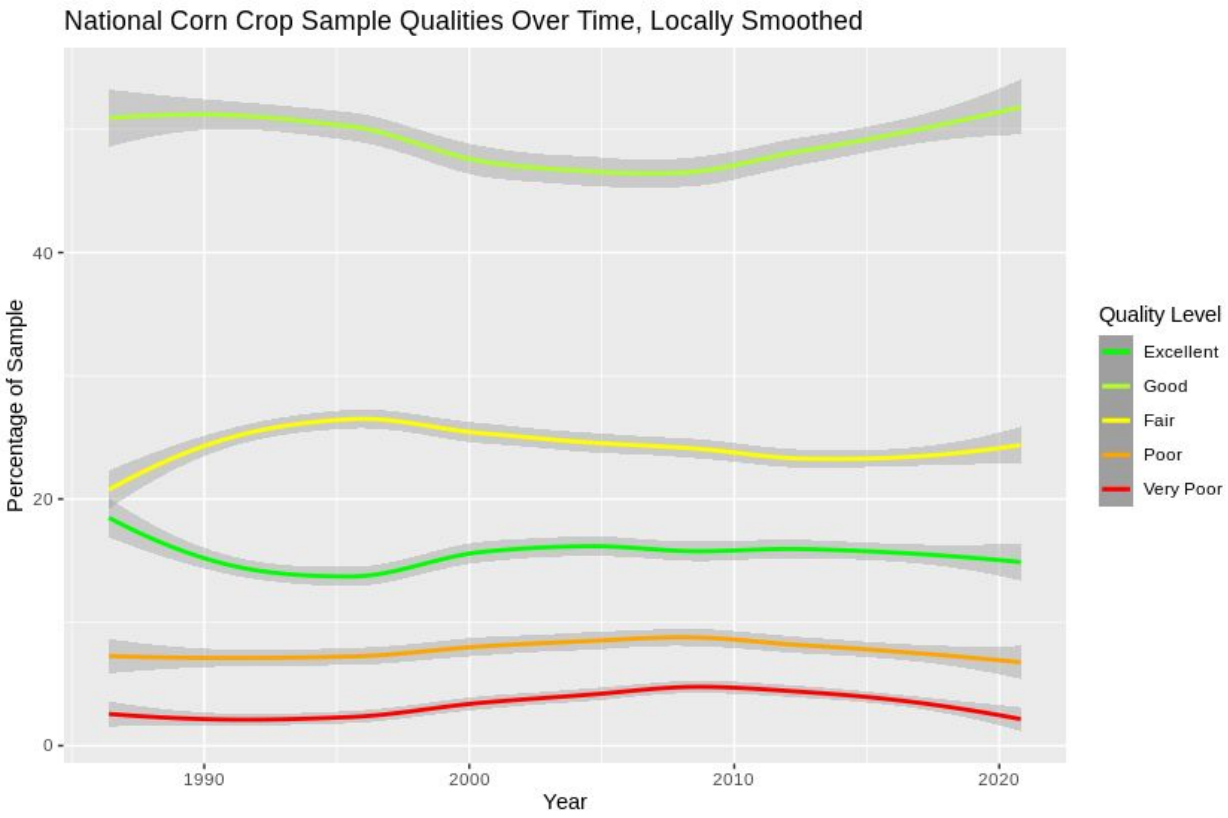


Figure 5: National corn crop sample qualities over time, locally smoothed

In this plot, it is much easier to see that these sampling averages for each quality level are roughly constant over time. Aside from a divergence in the "good" and "fair" curves near 1986, these regressions are approximately flat over time. This may be a result either of updates to USDA rubrics which artificially keep these statistics flat, or it may simply reflect the natural conditions which arise when growing corn in the United States. However, the documentation on the USDA's quality definitions over time is a bit dense and unclear, so further research is required to attribute these flat curves to a particular source.

**4. Can a metric be engineered to summarize all 5 crop quality levels in a sample?**

After examining this plot, we realized that a large hurdle in continuing our analysis of the corn crop quality dataset with all 13 states was going to be the fact that each state has 5 distinct quality curves, which would make for a visually overwhelming plot and a numerically difficult analysis. We wondered if there was any way to simplify these sample values to a single metric which reflected corn crop quality over time. Additionally, we wanted to keep any engineered metrics as simple as possible to minimize the introduction of personal biases to the dataset. In order to determine if this was possible, we plotted a correlation matrix showing the r values between each pair of quality levels. This is shown below in figure 6, where cells are colored based on this correlation coefficient.

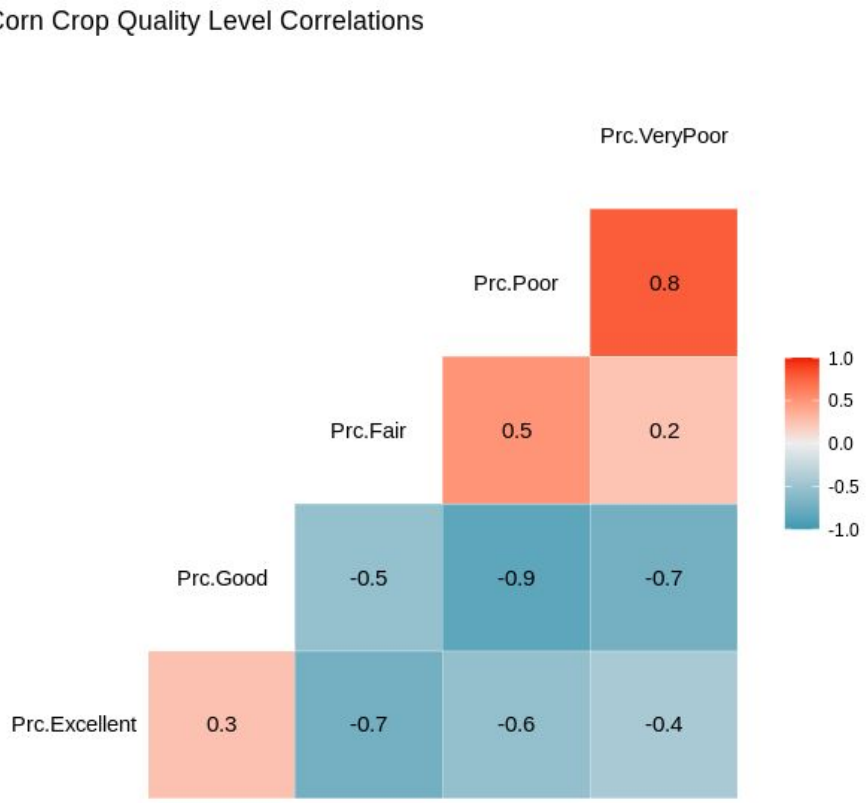


Figure 6: Corn crop quality level correlations

At this point, we were excited to see these correlations because they imply that the "excellent" and "good" categories can be combined in a way which doesn't necessarily contradict the inherent relationships in the dataset. What we mean by that is best reflected by the block of blue cells in the correlation matrix above. These values of -0.4 to -0.9 represent moderate to strong anti-correlations between the two groups it intersects. For instance, the "good" and "poor" categories have a correlation of -0.9, which means that an increase in one of these metrics is strongly correlated with a decrease in the other metric. We can see that both the "good" and "excellent" categories have fairly strong negative correlations with the "fair", "poor", and "very poor" categories. Thus, we can summarize all 5 metrics by splitting the values into these two groupings. We will define the "positive" quality category as the sum of the "excellent" and "good" categories. This engineered metric meets our initial goals since it is simple and reflects natural relationships in the data.

**5. How do states under- or over-perform relative to the national average over time?**

Using the metric engineered above, we can summarize each state using a single class of points or a single smoothed curve over time. In figure 7 below, we plot the locally smoothed curves for each state, this time following our "positive" quality metric. The national level sampling curve is highlighted in red to emphasize the national average. This acts as a baseline against which to compare the states' performance.

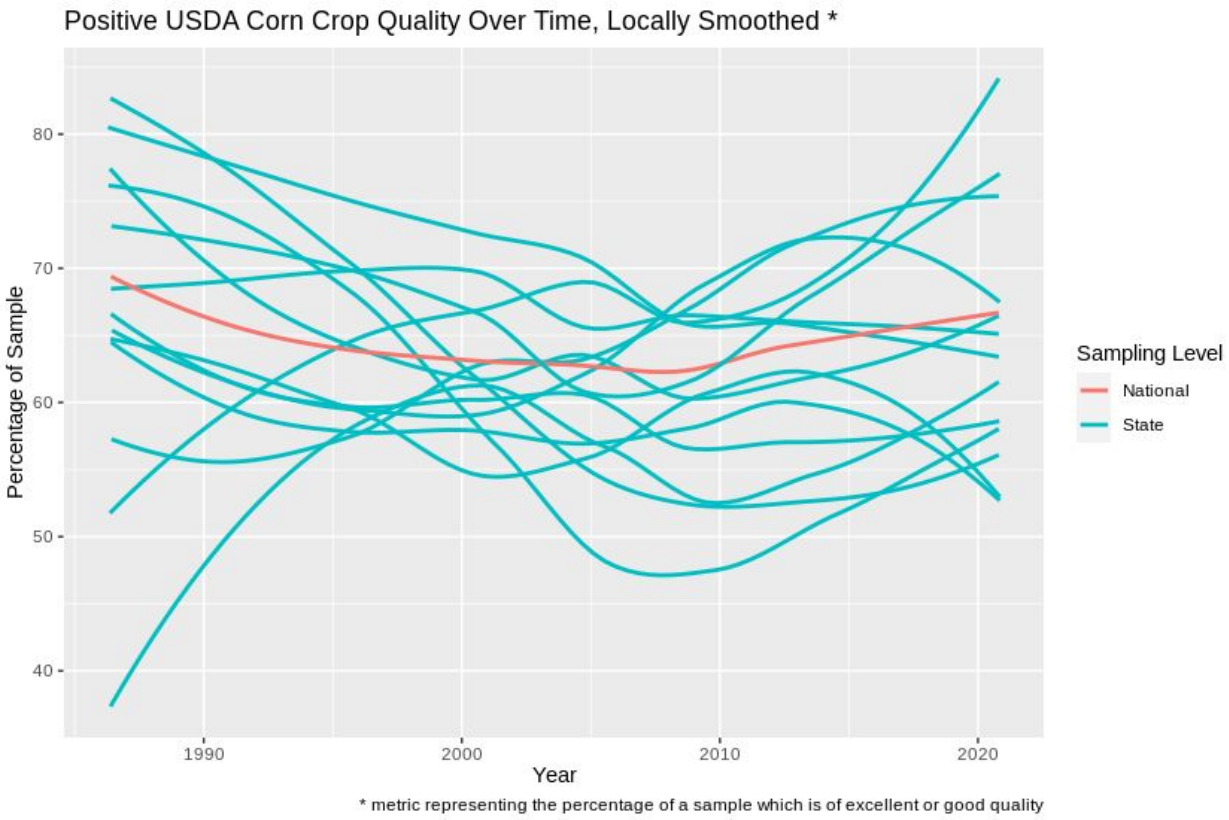


Figure 7: Positive USDA corn crop quality over time, locally smoothed

This plot was exciting to us because it shows that these metrics actually change over time in a significant way. Though the national average is relatively constant, we can plainly see that some of the state curves change considerably over time. One thing to note is that the regression curves appear to be possibly ill-fitting in early years, as demonstrated by a few stray curves reaching much further down than others. This may simply be a side effect of sparse data near the ends, or the pull of a few outliers. Overall though, the grouping of the curves within the band from 50% to 75% demonstrates reassuring similarity in these metrics. While they don't vary wildly from state to state, they still have the natural variation one would expect from a diverse set of data sources.

At this point, we were very close to the kind of plot we had imagined at the outset of this question. Namely, one which shows each state's performance over time relative to a national average. The only thing we really need to do to the previous figure is normalize by subtracting the national regression curve from each curve. This gives us figure 8 below.

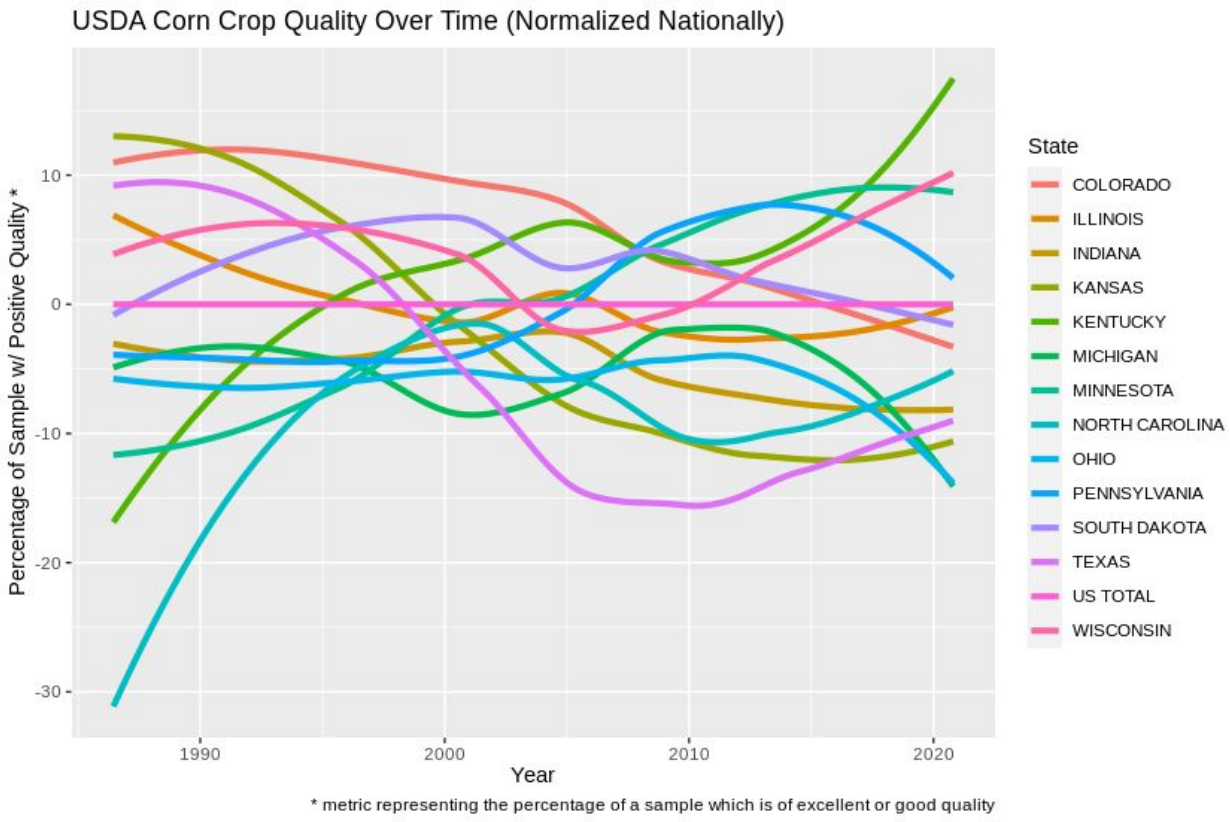


Figure 8: USDA corn crop quality over time (normalized nationally)

Here, we can finally see each state's positive quality curve normalized against the national average. While there are a few more colors than we would like in this plot, it is still illustrative and informative. We can easily see a few interesting stories in this plot. For instance, while Texas used to have an over-performing metric back in the 1990s, it fell sharply around the turn of the millennium and is now a consistently under-performing state. By contrast, Kentucky has seen a sharp rise from one of the worst under-performers in the 1990s to the best over-performer today. Of course, some states are more constant over time, as reflected by South Dakota's consistently slight overperformance.

So what exactly is this plot telling us? Primarily, it shows us which states grow better corn on average than the rest of the country over time. In the next section, we compare these under- and over-performers against our atmospheric quality data to explore any correlations between corn crop health and atmospheric health. This engineered regression metric simplifies that analysis by providing a single statistic to compare against.

### 3.3 - Atmospheric Quality vs. Corn Crop Quality

**6. What correlations exist between trace gas concentrations and corn crop quality?**

Now that we have successfully engineered a metric for corn crop quality which describes performance against a national average, we investigate correlations with our atmospheric data. This is initially done using a correlation matrix plot similar to the one found in the corn crop analysis.

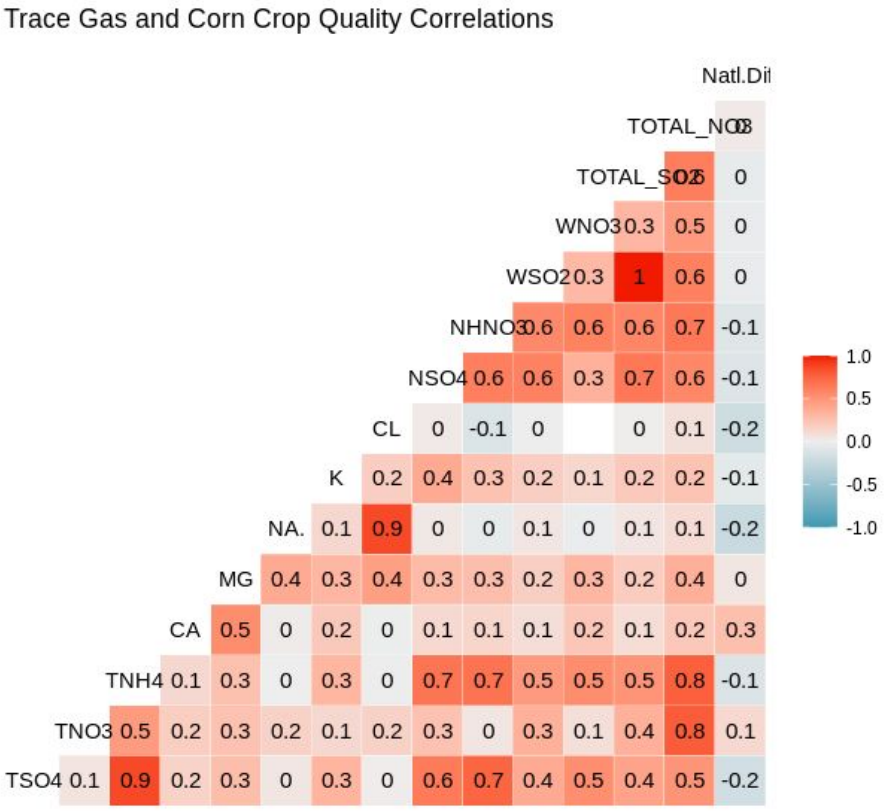


Figure 9: Trace gas and corn crop quality correlations

By looking at the rightmost column, we can see that the correlation between our crop quality metric and the gas concentrations is relatively weak. However, even weak correlations can be indicative of a relationship in some systems. For something as complex as the atmosphere or general environment, some of these correlation values are non-negligible. We consider two of the most strongly correlated examples by plotting our crop metric against calcium and sulfate concentrations for each sampling point. This is shown in the two figures below.

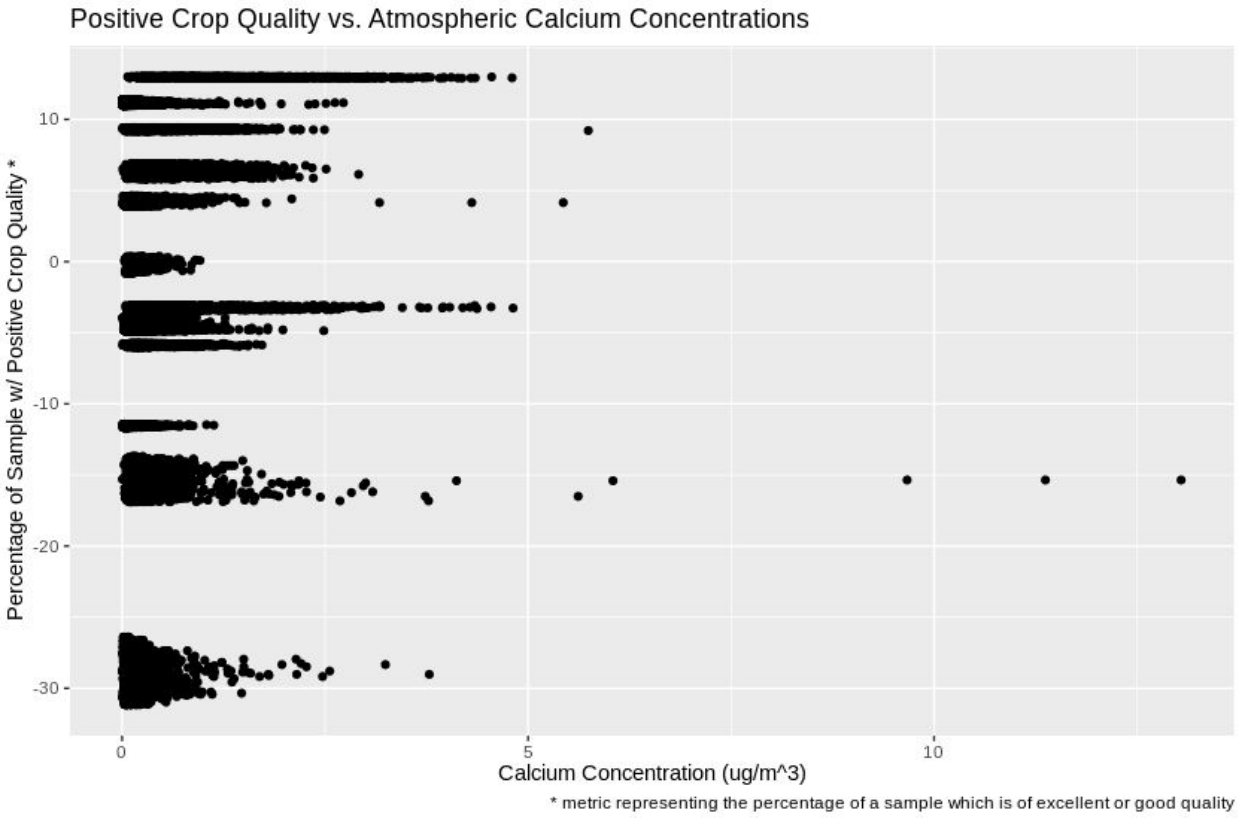


Figure 10: Positive crop quality vs. atmospheric calcium concentrations

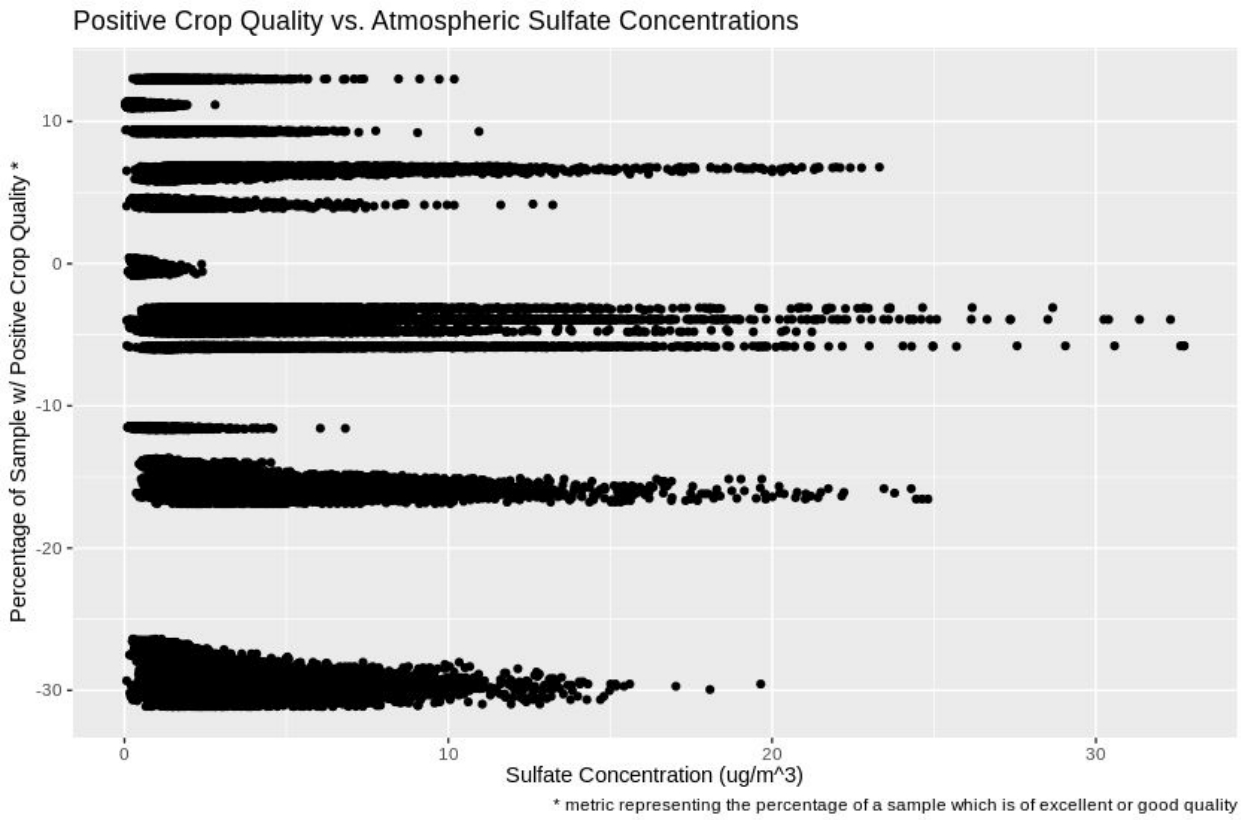


Figure 11: Positive crop quality vs. atmospheric sulfate concentrations

While it is natural to begin a correlation analysis with a scatterplot, these plots aren't very useful because of the high clustering of points along our corn crop quality metric axis. There is simply too much overlap to notice any useful relationships. Even with that being the case, the correlation with calcium is not very apparent. Below, we try a different plot for sulfate correlations, this time binning by our crop metric and doing a series of box plots.

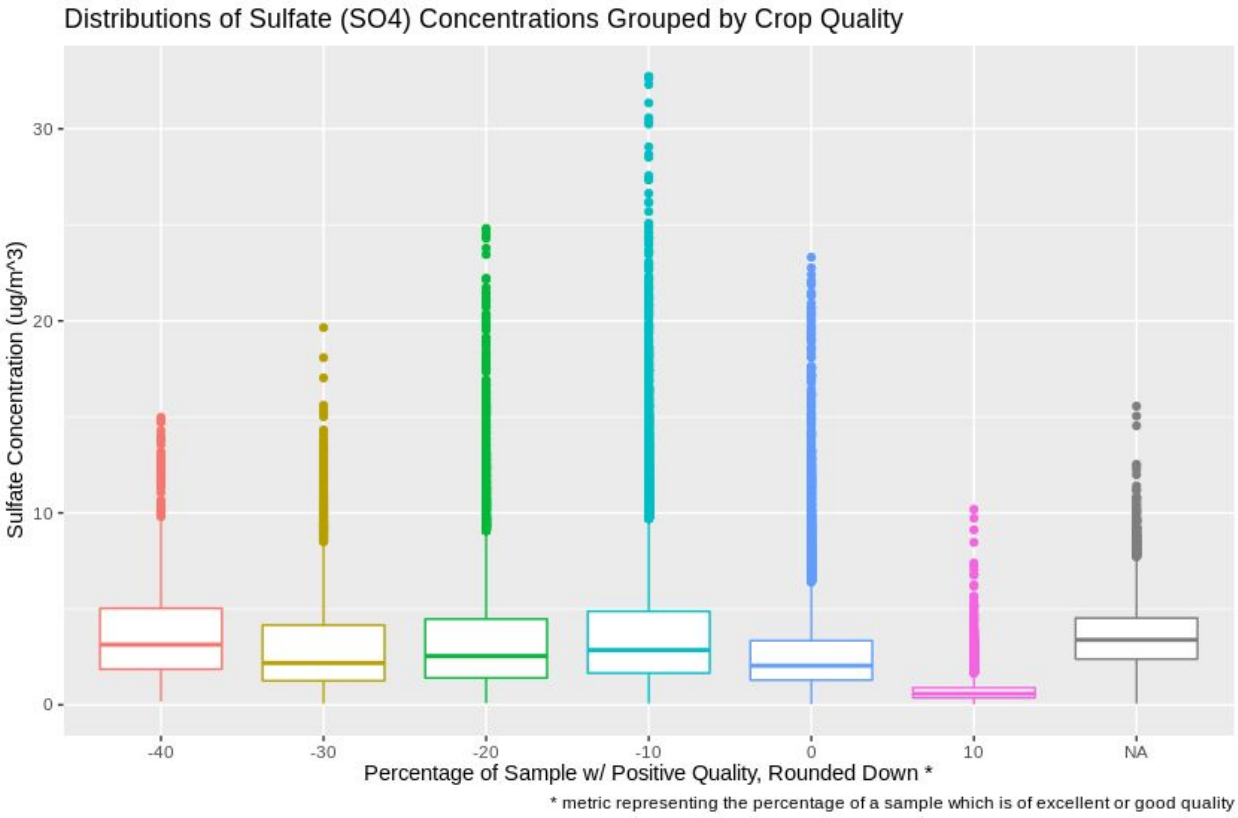


Figure 12: Distributions of sulfate (SO4) concentrations grouped by crop quality

If we had to select one figure as the smoking gun of this project, it would be this one. Here, we clearly see that the states with the best crop quality (according to our metric) have sulfate concentrations which are visibly much lower than they are for average or under-performing states. As a note, this isn't some anomaly from small sample sizes. Each box plot here corresponds to roughly the same number of observations, ~6,000. Even with these large sample sizes, we still see this obviously lower distribution for over-performing states. This is likely due to the fact that lower sulfate concentrations may lead to lower levels of acid rain. Since acid rain is harmful to crops, crop quality would naturally improve if there was less of it.

In short, this figure demonstrates at least one specific impact that atmospheric quality has on crop health. This is exactly what this project set out to do, so a result like this is exciting. Of course, this relationship would have to be followed up with other data sources and analysis to solidify the effect, but the acid rain hypothesis is incredibly reasonable given the data and basic knowledge of plant biology.

## 4. CONCLUSIONS

Analyzing trace gases in locations throughout the U.S. over time provided that the average level of air pollution due to these gases has generally decreased with time. We conclude that this decrease is the effect of several new regulations placed by the U.S. EPA upon the large corporations emitting these gases. These regulations have become more strict with passing time,

likely due to the fact that a climate crisis arrives closer with each year. We also partially attribute this overall decrease in air pollution to more earth-friendly technology being developed. Specific trace gases can be observed at higher concentrations in some regions, based on the geography and human activity at location, with higher levels of agricultural pollutants in agricultural areas and lower levels of automobile pollutants in natural areas with less roads.

Corn crop quality remained, for the most part, consistent throughout the past 30 years. Notable variations in a few specific years can be observed due to natural causes, such as droughts. There have been a great many fluctuations, however, in terms of individual state performance for corn crop quality.

We found that there is indeed a correlation between atmospheric pollution levels and the quality of corn crops, albeit a somewhat weak correlation. For the sites which held a lower overall concentration of sulfates in the atmosphere, it was evident that they frequently performed better in terms of corn crop quality. This relationship indicates that the two variables may be inversely related. One reason we found for this potential correlation was the fact that these sites often had lower levels of acid rain, caused by atmospheric sulfates and nitrates.

## 5. REFERENCES


1. https://java.epa.gov/castnet/clearsession.do
2. https://quickstats.nass.usda.gov/
3. https://www.epa.gov/criteria-air-pollutants
4. https://www.epa.gov/castnet/castnet-site-locations
5. https://news.psu.edu/story/329095/2014/10/06/research/air-pollution-and-ocean
6. https://ww2.arb.ca.gov/resources/sulfur-dioxide-and-health
7. https://www.usgs.gov/special-topic/water-science-school/science/acid-rain-and-water?qt-science_center_objects=0#qt-science_center_objects
8. https://www.sciencedirect.com/science/article/pii/S2212094715300360
9. https://www.purdue.edu/newsroom/outreach/2012/120705HurtDrought.html


## 6. APPENDIX

### 1. Code for atmospheric data analysis

**AllDataSites.R**

```
library(ggplot2)
library(tidyr)

#classifying data based on location and site type
site_state <- c("TEXAS", "ILLINOIS", "ILLINOIS", "MICHIGAN", "ILLINOIS",
"PENNSYLVANIA",
```

```
                "PENNSYLVANIA", "TEXAS", "NORTH CAROLINA", "ILLINOIS",
"KENTUCKY", "KENTUCKY",
                "NORTH CAROLINA", "NORTH CAROLINA", "NORTH CAROLINA",
"OHIO", "OHIO", "NORTH CAROLINA",
                "COLORADO", "MICHIGAN", "PENNSYLVANIA", "KANSAS", "KANSAS",
"KENTUCKY",
                "PENNSYLVANIA", "OHIO", "KENTUCKY", "KENTUCKY",
"KENTUCKY", "COLORADO",
                "PENNSYLVANIA", "OHIO", "TEXAS", "KENTUCKY", "NORTH
CAROLINA", "WISCONSIN",
                "PENNSYLVANIA", "OHIO", "MICHIGAN", "COLORADO", "COLORADO",
"NORTH CAROLINA",
                "INDIANA", "ILLINOIS", "MICHIGAN", "INDIANA", "MINNESOTA",
"MICHIGAN",
                "SOUTH DAKOTA", "CALIFORNIA", "CALIFORNIA", "FLORIDA",
"CALIFORNIA", "FLORIDA")

site_id <- c("ALC188", "ALH157", "ALH257", "ANA115", "ANL146", "ARE128",
        "ARE228", "BBE401", "BFT142", "BVL130", "CDZ171", "CKT136",
        "CND125", "COW005", "COW137", "DCP114", "DCP214", "DUK008",
        "GTH161", "HOX148", "KEF112", "KIC003", "KNZ184", "LCW121",
        "LRL117", "LYK123", "MAC426", "MCK131", "MCK231", "MEV405",
        "MKG113", "OXF122", "PAL190", "PBF129", "PNF126", "PRK134",
        "PSU106", "QAK172", "RED004", "ROM206", "ROM406", "RTP101",
        "SAL133", "STK138", "UVL124", "VIN140", "VOY413", "WEL149",
        "WNC429", "PIN414", "JOT403", "IRL141", "DEV412", "EVE419")
# Site types: 1 - natural, a less habited area, away from major population centers
#         2 - urban, near a major city or high population area
#         3 - coastal, near an ocean
#         4 - agricultural

site_type <- c(1, 4, 4, 1, 2, 4,
        4, 1, 3, 4, 4, 4,
        2, 1, 1, 1, 4, 1,
        1, 1, 2, 2, 2, 4,
        2, 4, 4, 2, 4, 1,
        1, 4, 4, 4, 1, 2,
        3, 2, 1, 1, 1, 2,
        4, 2, 4, 3, 1, 2,
        4, 3, 2, 3, 2, 3)
#Reading data from file
site_info <- data.frame(site_state, site_id, site_type)
AllSiteData <- read.csv("AnnualConcentrationsNew.csv")

#Filtering data based on selected sites
Site1 <- AllSiteData[(AllSiteData$SITE_ID %in% site_info$site_id),]
```

```
Site1 <- Site1[order(Site1$YEAR),]
names(Site1) <- c("ID", "Year", "StartDate", "EndDate", "SO2", "SO4", "NO3", "HNO3",
"TNO3", "NH4")
Site1_melt <- tidyr::gather(Site1, pollutant, value, SO2:NH4)

#Creating plot
Site1plot <- ggplot(Site1_melt, aes(x = Year, y = value, color = pollutant)) +
  geom_smooth(na.rm = TRUE) + xlab('Year') + ylab('Pollutant Concentration (ug/m^3)') +
  theme(plot.title = element_text(hjust = 0.5))
Site1plot <- Site1plot + labs(title = "All Sites Pollutant Average")
Site1plot
```

---

**ClassifiedAirData.R**

```
library(ggplot2)
library(tidyr)

#classifying data based on location and site type
site_state <- c("TEXAS", "ILLINOIS", "ILLINOIS", "MICHIGAN", "ILLINOIS",
"PENNSYLVANIA",
                "PENNSYLVANIA", "TEXAS", "NORTH CAROLINA", "ILLINOIS",
"KENTUCKY", "KENTUCKY",
                "NORTH CAROLINA", "NORTH CAROLINA", "NORTH CAROLINA",
"OHIO", "OHIO", "NORTH CAROLINA",
                "COLORADO", "MICHIGAN", "PENNSYLVANIA", "KANSAS", "KANSAS",
"KENTUCKY",
                "PENNSYLVANIA", "OHIO", "KENTUCKY", "KENTUCKY",
"KENTUCKY", "COLORADO",
                "PENNSYLVANIA", "OHIO", "TEXAS", "KENTUCKY", "NORTH
CAROLINA", "WISCONSIN",
                "PENNSYLVANIA", "OHIO", "MICHIGAN", "COLORADO", "COLORADO",
"NORTH CAROLINA",
                "INDIANA", "ILLINOIS", "MICHIGAN", "INDIANA", "MINNESOTA",
"MICHIGAN",
                "SOUTH DAKOTA", "CALIFORNIA", "CALIFORNIA", "FLORIDA",
"CALIFORNIA", "FLORIDA")

site_id <- c("ALC188", "ALH157", "ALH257", "ANA115", "ANL146", "ARE128",
         "ARE228", "BBE401", "BFT142", "BVL130", "CDZ171", "CKT136",
         "CND125", "COW005", "COW137", "DCP114", "DCP214", "DUK008",
         "GTH161", "HOX148", "KEF112", "KIC003", "KNZ184", "LCW121",
         "LRL117", "LYK123", "MAC426", "MCK131", "MCK231", "MEV405",
         "MKG113", "OXF122", "PAL190", "PBF129", "PNF126", "PRK134",
         "PSU106", "QAK172", "RED004", "ROM206", "ROM406", "RTP101",
         "SAL133", "STK138", "UVL124", "VIN140", "VOY413", "WEL149",
```

```
        "WNC429", "PIN414", "JOT403", "IRL141", "DEV412", "EVE419")

# Site types: 1 - natural, a less habited area, away from major population centers
#         2 - urban, near a major city or high population area
#         3 - coastal, near an ocean
#         4 - agricultural

site_type <- c(1, 4, 4, 1, 2, 4,
        4, 1, 3, 4, 4, 4,
        2, 1, 1, 1, 4, 1,
        1, 1, 2, 2, 2, 4,
        2, 4, 4, 2, 4, 1,
        1, 4, 4, 4, 1, 2,
        3, 2, 1, 1, 1, 2,
        4, 2, 4, 3, 1, 2,
        4, 3, 2, 3, 2, 3)

#reading data from file
SiteClassifcation <- data.frame(site_state, site_id, site_type)
AllSiteData <- read.csv("AnnualConcentrationsNew.csv")

#filtering data to selected sites
FilteredSiteData <- AllSiteData[(AllSiteData$SITE_ID %in% SiteClassifcation$site_id),]
names(FilteredSiteData) <- c("ID", "Year", "StartDate", "EndDate", "SO2", "SO4",
                             "NO3", "HNO3", "TNO3", "NH4")

#Creating vector of site type classification values
TempSiteType <- integer()
SiteTypeName <- factor()
for(i in 1:nrow(FilteredSiteData)) {
  for(j in 1:nrow(SiteClassifcation)){
        if(FilteredSiteData[i,1] == SiteClassifcation[j,2]){
        TempSiteType <- c(TempSiteType, SiteClassifcation[j,3])
        }
  }
}

#assigning each site's classification based on its type
for(i in TempSiteType) {
  if(i == 1) {
        SiteTypeName <- c(SiteTypeName, "natural")
  }
  if(i == 2) {
        SiteTypeName <- c(SiteTypeName, "urban")
  }
  if(i == 3) {
```

```
        SiteTypeName <- c(SiteTypeName, "coastal")
  }
  if(i == 4) {
        SiteTypeName <- c(SiteTypeName, "agricultural")
  }
}

#adding the type data to the data frame
FilteredSiteData$SiteTypeN <- factor(SiteTypeName)
FilteredSiteData$SiteType <- TempSiteType
FilteredSiteData2019 <- FilteredSiteData[FilteredSiteData$Year == 2019,]
str(FilteredSiteData)

#creating a plot for each gas
HNO3Plot <- ggplot(FilteredSiteData2019, aes(SiteTypeN, HNO3, fill = SiteTypeN)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  labs(x = "Location Type", y = "Pollutant Concentration (ug/m^3)", title = "Nitric
Acid(HNO3)") +
  theme(plot.title = element_text(hjust = 0.5))
HNO3Plot

NH4Plot <- ggplot(FilteredSiteData2019, aes(SiteTypeN, NH4, fill = SiteTypeN)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  labs(x = "Location Type", y = "Pollutant Concentration (ug/m^3)", title = "Ammonium (NH4)")
+
  theme(plot.title = element_text(hjust = 0.5))
NH4Plot

SO2Plot <- ggplot(FilteredSiteData2019, aes(SiteTypeN, SO2, fill = SiteTypeN)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  labs(x = "Location Type", y = "Pollutant Concentration (ug/m^3)", title = "Sulfur
Dioxide(SO2)") +
  theme(plot.title = element_text(hjust = 0.5))
SO2Plot

SO4Plot <- ggplot(FilteredSiteData2019, aes(SiteTypeN, SO4, fill = SiteTypeN)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  labs(x = "Location Type", y = "Pollutant Concentration (ug/m^3)", title = "Sulfate (SO4)") +
  theme(plot.title = element_text(hjust = 0.5))
SO4Plot
```

---

**2. Code for corn data analysis**

**corn_cleaning.r**

```
# corn_cleaning.r
#
# Evan Meade, 2020
#
# This script reads the raw corn crop quality .csv files and combines
# them into a single synthesized .csv. As noted in the README, the NASS
# data portal only allows for 50,000 results to be returned at once,
# which means a dataset of this size requires multiple downloads.
#

# Library imports
library(ggplot2)


# First we read in each of the files as their own dataframes
corn_1 <- read.csv("data/raw/corn_quality__excellent_fair.csv")
corn_2 <- read.csv("data/raw/corn_quality__good_poor.csv")
corn_3 <- read.csv("data/raw/corn_quality__verypoor.csv")

# Now we combine the individual dataframes into a master dataframe
corn <- rbind(corn_1, corn_2, corn_3)


#
# If you look at the dataframe at this point, you will see a number of
# columns which only have 1 value. In other words, they provide no
# information. They are likely vestigial variables from other USDA
# analysis.
#
# Most of these columns simply have all null values, likely due to
# being irrelevant to the dataset at hand. Some simply have only one
# value. Either way, we will remove them because they are cluttering
# up our analysis and not providing any value.
#

# Finding all columns with only one value
drop_cols <- c()
for (var in colnames(corn)) {
  unique_count <- length(unique(corn[, var]))
  if (unique_count == 1) {
          drop_cols <- c(drop_cols, var)
  }
```

```
}

# Dropping all columns with only one value
corn <- corn[, setdiff(colnames(corn), drop_cols)]


#
# Now, we would like to combine all 5 quality variables for each sample.
# We can do this by noting that each sample has the same time and location
# data. Then, we split on this for the 5 values we seek.
#

# Creating new row names for more convenient indexing
obs_id <- c()
for (i in 1:nrow(corn)) {
  new_id <- paste0(corn[i, "State"], "-", corn[i, "Week.Ending"])
  obs_id <- c(obs_id, new_id)
}
obs_id <- unique(obs_id)

# Creating new dataframe for flattened corn data, one row per observation
corn_flat <- unique(data.frame(corn[1:6]))
rownames(corn_flat) <- obs_id

# Fill with values form original corn dataframe, replace data NA with 0
for (i in 1:nrow(corn)) {
  new_id <- paste0(corn[i, "State"], "-", corn[i, "Week.Ending"])
  data_var <- corn[i, "Data.Item"]
  data_value <- corn[i, "Value"]
  corn_flat[new_id, data_var] <- data_value
}
corn_flat[, 7:11][is.na(corn_flat[, 7:11])] <- 0


#
# Now we just polish the dataframe by renaming, recasting, and
# reordering the columns. Then, we save it to a new .csv file for
# easy reading in the future.
#

# Renaming and rearranging data columns in corn_flat for readability
colnames(corn_flat) <- c("Year", "Week", "Week.Ending",
                         "Geo.Level", "State", "State.ANSI",
                         "Prc.Excellent", "Prc.Fair", "Prc.Good",
                         "Prc.Poor", "Prc.VeryPoor")
col_order <- c(1, 2, 3, 4, 5, 6, 7, 9, 8, 10, 11)
```

```
corn_flat <- corn_flat[col_order]

# Substituting week with the actual number
for (i in 1:nrow(corn_flat)) {
  corn_flat[i, "Week"] <- substr(corn_flat[i, "Week"], 7, 8)
}
corn_flat$Week <- as.integer(corn_flat$Week)

# Creating time variable to help plot all of the data on the same scale
corn_flat$Week.Total <- (corn_flat$Year - 1986) * 52 + corn_flat$Week

# Creating time variable for more clear plotting of year
corn_flat$Year.Frac <- corn_flat$Week.Total / 52 + 1986

# Sort rows by row names
corn_flat <- corn_flat[order(rownames(corn_flat)), ]


# FIGURE: corn_obs_by_state.png
#
# Plots the distribution of the number of observations per state.
corn_obs_by_state <- ggplot() +
  geom_histogram(mapping = aes(x = table(corn_flat$State))) +
  labs(title = "Number of Corn Crop Quality Observations per State",
          x = "Number of Observations",
          y = "Number of States")
print(corn_obs_by_state)


# Here, we subset corn_flat to include selected states, which have a
# high number of observations in both the EPA CASTNET dataset and
# the USDA NASS dataset
our_states <- c("COLORADO", "ILLINOIS", "INDIANA", "KANSAS", "KENTUCKY",
                  "MICHIGAN", "MINNESOTA", "NORTH CAROLINA", "OHIO",
                  "PENNSYLVANIA", "SOUTH DAKOTA", "TEXAS", "WISCONSIN",
                  "US TOTAL")
corn_used <- corn_flat[which(corn_flat$State %in% our_states), ]


# Save final corn_flat to synthesized .csv file
write.csv(corn_flat, file = "data/synthesized/corn_flat.csv")

# Save final corn_used to synthesized .csv file
write.csv(corn_used, file = "data/synthesized/corn_used.csv")
```

---

**corn_analysis.r**

```
# corn_analysis.r
#
# Evan Meade, 2020
#
# This script generates the figures I used to explore, characterize,
# and analyze the corn crop quality data.
#

# Library imports
library(ggplot2)
library(GGally)

# Reading in the corn_used.csv data
corn_used <- read.csv("data/synthesized/corn_used.csv", row.names = 1)



#
# First, I wanted to explore just the national level data to see if
# there are any obvious nationwide trends over time. This can also
# serve as a baseline to compare states against.
#

# Creating National data subset
natl_corn <- corn_used[which(corn_used$Geo.Level == "NATIONAL"), ]

# FIGURE: natl_corn_point_plot.png
#
# Scatter plot of national corn crop sample qualities over time.
quality_classes <- c("green", "greenyellow", "yellow", "orange", "red")
names(quality_classes) <- c("Excellent", "Good", "Fair", "Poor", "Very Poor")

natl_corn_point_plot <- ggplot(data = natl_corn) +
  geom_point(mapping = aes(x = Year.Frac, y = Prc.Excellent, color = "green")) +
  geom_point(mapping = aes(x = Year.Frac, y = Prc.Good, color = "greenyellow")) +
  geom_point(mapping = aes(x = Year.Frac, y = Prc.Fair, color = "yellow")) +
  geom_point(mapping = aes(x = Year.Frac, y = Prc.Poor, color = "orange")) +
  geom_point(mapping = aes(x = Year.Frac, y = Prc.VeryPoor, color = "red")) +
  labs(title = "National Corn Crop Sample Qualities Over Time",
       x = "Year",
       y = "Percentage of Sample") +
  scale_color_identity(name = "Quality Level",
                       guide = "legend",
                       breaks = quality_classes,
                       labels = names(quality_classes))
```

```
print(natl_corn_point_plot)

# FIGURE: natl_corn_smooth_plot.png
#
# Locally regressed curves providing a smooth fit of each crop quality class
# over time. Helps expose long term trends and removes some noise.
natl_corn_smooth_plot <- ggplot(data = natl_corn) +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.Excellent, color = "green"),
          method = "loess", formula = "y ~ x") +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.Good, color = "greenyellow"),
          method = "loess", formula = "y ~ x") +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.Fair, color = "yellow"),
          method = "loess", formula = "y ~ x") +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.Poor, color = "orange"),
          method = "loess", formula = "y ~ x") +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.VeryPoor, color = "red"),
          method = "loess", formula = "y ~ x") +
  labs(title = "National Corn Crop Sample Qualities Over Time, Locally Smoothed",
          x = "Year",
          y = "Percentage of Sample") +
  scale_color_identity(name = "Quality Level",
                       guide = "legend",
                       breaks = quality_classes,
                       labels = names(quality_classes))
print(natl_corn_smooth_plot)


# FIGURE: corn_corr_plot.png
#
# Correlations between each pair of corn crop quality levels.
corn_corr_plot <- ggcorr(corn_used[, 7:11], label = TRUE) +
  labs(title = "Corn Crop Quality Level Correlations")
print(corn_corr_plot)


#
# Creating a new metric which summarizes the corn crop quality data
# by combining the "Excellent" and "Good" categories since they have
# negative correlations with all the others.
#

# Defining new metric as Prc.Positive
corn_used$Prc.Positive <- corn_used$Prc.Excellent + corn_used$Prc.Good


# FIGURE: all_states_pos_smooth.png
```

```
#
# Plotting smoothed Prc.Positive curves for each state/national average.
all_states_pos_smooth <- ggplot(data = corn_used) +
  geom_smooth(mapping = aes(x = Year.Frac, y = Prc.Positive,
                            group = State, color = Geo.Level),
              method = "loess",
              formula = "y ~ x",
              se = FALSE) +
  labs(title = "Positive USDA Corn Crop Quality Over Time, Locally Smoothed *",
              x = "Year",
              y = "Percentage of Sample",
              caption = "* metric representing the percentage of a sample which is of excellent or good
quality") +
  scale_color_discrete("Sampling Level",
                       labels = c("National", "State"))
print(all_states_pos_smooth)


#
# I want to create a metric representing the gap between each state's smoothed
# Prc.Positive and the national smoothed Prc.Positive. This creates a single
# metric which represents a state's performance relative to the country.
#

# First, subsetting the data to have endpoint weeks containing all states
# Allows the regressions to be totally interpolative
x <- table(corn_used$Week.Total)
num_states <- length(unique(corn_used$State))
week_min <- min(as.integer(names(x[which(x == num_states)])))
week_max <- max(as.integer(names(x[which(x == num_states)])))
corn_used_trimmed <- corn_used[which(corn_used$Week.Total %in% week_min:week_max), ]

#
# Now to create loess values for each state, interpolate on all weeks
# contained here, and calculate differences from national.
#

corn_loess <- data.frame()

natl_loess <- loess(Prc.Positive ~ Week.Total,
                    corn_used_trimmed[which(corn_used_trimmed$State == "US TOTAL"), ])

for (state in unique(corn_used_trimmed$State)) {
  state_loess <- loess(Prc.Positive ~ Week.Total,
                       corn_used_trimmed[which(corn_used_trimmed$State == state), ])
```

```
  state_col <- rep(state, (week_max - week_min + 1))
  week_col <- week_min:week_max
  if (state == "US TOTAL") {
        level_col <- rep("NATIONAL", (week_max - week_min + 1))
  } else {
        level_col <- rep("STATE", (week_max - week_min + 1))
  }
  diff_col <- predict(state_loess, week_col) - predict(natl_loess, week_col)

  state_df <- data.frame(state_col, week_col, level_col, diff_col)
  corn_loess <- rbind(corn_loess, state_df)
}
colnames(corn_loess) <- c("State", "Week.Total", "Geo.Level", "Natl.Diff")
corn_loess$Year.Frac <- corn_loess$Week.Total / 52 + 1986


# FIGURE: states_pos_norm_smooth_plot.png
#
# Each state's performance in the Prc.Positive metric, locally smoothed
# with loess to reduce noise. Normalized nationally by subtracting the
# national smoothed loess curve.
states_pos_norm_smooth_plot <- ggplot(data = corn_loess) +
  geom_line(mapping = aes(x = Year.Frac, y = Natl.Diff, group = State, color = State),
        size = 1.5) +
  labs(title = "USDA Corn Crop Quality Over Time (Normalized Nationally)",
        x = "Year",
        y = "Percentage of Sample w/ Positive Quality *",
        caption = "* metric representing the percentage of a sample which is of excellent or good
quality")
print(states_pos_norm_smooth_plot)
```

---

**3. Code for atmospheric and corn joint analysis**

**corn_epa_cleaning.r**

```
# corn_epa_cleaning.r
#
# Evan Meade, 2020
#
# This script reads in the weekly filter pack data from the EPA
# and restructures it into a form that works well with the corn
# crop quality data. Namely, it calculates weeks and states for each
# entry, so that a master dataset can be constructed for weekly
# corn and atmospheric data.
#
```

```
# First we read in the weekly EPA filter pack data
epa <- read.csv("data/raw/epa_weekly_filter_packs.csv")

# Then we add a state column using our knowledge of SITE_ID
site_states <- c("TEXAS", "ILLINOIS", "ILLINOIS", "MICHIGAN", "ILLINOIS",
"PENNSYLVANIA",
                 "PENNSYLVANIA", "TEXAS", "NORTH CAROLINA", "ILLINOIS",
"KENTUCKY", "KENTUCKY",
                 "NORTH CAROLINA", "NORTH CAROLINA", "NORTH CAROLINA",
"OHIO", "OHIO", "NORTH CAROLINA",
                 "COLORADO", "MICHIGAN", "PENNSYLVANIA", "KANSAS", "KANSAS",
"KENTUCKY",
                 "PENNSYLVANIA", "OHIO", "KENTUCKY", "KENTUCKY",
"KENTUCKY", "COLORADO",
                 "PENNSYLVANIA", "OHIO", "TEXAS", "KENTUCKY", "NORTH
CAROLINA", "WISCONSIN",
                 "PENNSYLVANIA", "OHIO", "MICHIGAN", "COLORADO", "COLORADO",
"NORTH CAROLINA",
                 "INDIANA", "ILLINOIS", "MICHIGAN", "INDIANA", "MINNESOTA",
"MICHIGAN",
                 "SOUTH DAKOTA")
names(site_states) <- c("ALC188", "ALH157", "ALH257", "ANA115", "ANL146", "ARE128",
                 "ARE228", "BBE401", "BFT142", "BVL130", "CDZ171", "CKT136",
                 "CND125", "COW005", "COW137", "DCP114", "DCP214", "DUK008",
                 "GTH161", "HOX148", "KEF112", "KIC003", "KNZ184", "LCW121",
                 "LRL117", "LYK123", "MAC426", "MCK131", "MCK231", "MEV405",
                 "MKG113", "OXF122", "PAL190", "PBF129", "PNF126", "PRK134",
                 "PSU106", "QAK172", "RED004", "ROM206", "ROM406", "RTP101",
                 "SAL133", "STK138", "UVL124", "VIN140", "VOY413", "WEL149",
                 "WNC429")
epa$State <- site_states[epa$SITE_ID]

# Now we extract the year and week number for each observation from the DATEOFF
epa_time_obj <- strptime(epa$DATEOFF, format = "%m/%d/%Y %H:%M:%S")
epa$Year <- as.integer(strftime(epa_time_obj, format = "%Y"))
epa$Week <- as.integer(strftime(epa_time_obj, format = "%V"))

# To reduce matching difficulties for fractional 53rd weeks, we group them with 52nd weeks
epa[which(epa$Week == 53), "Week"] <- 52

# We can save this version of epa which leaves the sites separate
# After this, we will link EPA and USDA data
write.csv(epa, "data/synthesized/epa_weekly_filter_packs__all_sites.csv")
```

```
# Making a copy of epa measurements trimmed down to useful columns
col_order <- c(26, 27, 28, 1, 5:18)
epa_trimmed <- data.frame(epa[col_order])
epa_trimmed$Count <- rep(1, nrow(epa_trimmed))

# Labeling rows in the regression dataframe by state and time for matching
for (i in 1:nrow(corn_loess)) {
  obs_label <- paste0(corn_loess[i, "State"], "-", corn_loess[i, "Week.Total"])
  corn_loess[i, "Obs.Label"] <- obs_label
}

# Matching EPA and USDA data by state and time
for (i in 1:nrow(epa_trimmed)) {
  week_total <- (epa_trimmed[i, "Year"] - 1986) + epa_trimmed[i, "Week"]
  epa_trimmed[i, "Week.Total"] <- week_total
  obs_label <- paste0(epa_trimmed[i, "State"], "-", week_total)
  epa_trimmed[i, "Obs.Label"] <- obs_label
  if (obs_label %in% corn_loess$Obs.Label) {
          epa_trimmed[i, "Natl.Diff"] <- corn_loess[which(corn_loess$Obs.Label == obs_label),
"Natl.Diff"]
  }
}

# Saving matched data to file
write.csv(epa_trimmed, "data/synthesized/epa_usda_matched.csv")
```

---

**corn_and_atmosphere_analysis.r**

```
# corn_and_atmosphere_analysis.r
#
# Evan Meade, 2020
#
# Here, we unite the analyses of corn and atmospheric quality over
# time to investigate any possible correlations between the two.
#

# Library imports
library(ggplot2)
library(GGally)

# Reading in the epa_usda_matched.csv data
epa_trimmed <- read.csv("data/synthesized/epa_usda_matched.csv", row.names = 1)
```

```
#
# First, we explore the possible correlations between trace gas
# concentrations and our positive crop quality metric from the corn
# analysis. Then, we investigate specific examples to find


# FIGURE: corn_atmo_corr.png
#
# Correlation matrix for each of the trace gas concentrations and the
# positive corn quality metric constructed in the corn crop analysis
# (normalized nationally).
corn_atmo_corr <- ggcorr(epa_trimmed[, c(5:18, 22)], label = TRUE) +
  labs(title = "Trace Gas and Corn Crop Quality Correlations")
print(corn_atmo_corr)


# FIGURE: natl_diff_ca.png
#
# Scatter plot of samples' positive crop quality plotted against
# atmospheric calcium concentration.
natl_diff_ca <- ggplot(data = epa_trimmed) +
  geom_point(mapping = aes(x = CA, y = Natl.Diff)) +
  labs(title = "Positive Crop Quality vs. Atmospheric Calcium Concentrations",
          x = "Calcium Concentration (ug/m^3)",
          y = "Percentage of Sample w/ Positive Crop Quality *",
          caption = "* metric representing the percentage of a sample which is of excellent or good
quality")
print(natl_diff_ca)


# FIGURE: natl_difff_so4.png
#
# Scatter plot of samples' positive crop quality plotted against
# atmospheric sulfate concentration.
natl_diff_so4 <- ggplot(data = epa_trimmed) +
  geom_point(mapping = aes(x = TSO4, y = Natl.Diff)) +
  labs(title = "Positive Crop Quality vs. Atmospheric Sulfate Concentrations",
          x = "Sulfate Concentration (ug/m^3)",
          y = "Percentage of Sample w/ Positive Crop Quality *",
          caption = "* metric representing the percentage of a sample which is of excellent or good
quality")
print(natl_diff_so4)


#
# The scatter plots are a bit hard to read because there is so
```

```
# much overlap between the points. So a box plot is probably better
# for examining the relationship between the distributions.
#

# Binning crop quality metric due to high grouping
epa_trimmed$Natl.Diff.Rounded <- as.factor(floor(epa_trimmed$Natl.Diff / 10) * 10)

# FIGURE: natl_diff_so4_box_plot.png
#
# Box plot of binned corn crop quality metric against sulfate
# concentrations.
natl_diff_so4_box_plot <- ggplot(data = epa_trimmed) +
  geom_boxplot(mapping = aes(x = Natl.Diff.Rounded, y = TSO4, color = Natl.Diff.Rounded),
           show.legend = FALSE) +
  labs(title = "Distributions of Sulfate (SO4) Concentrations Grouped by Crop Quality",
           x = "Percentage of Sample w/ Positive Quality, Rounded Down *",
           y = "Sulfate Concentration (ug/m^3)",
           caption = "* metric representing the percentage of a sample which is of excellent or good
quality")
print(natl_diff_so4_box_plot)
```