Impact of Drought on Agriculture

BU-NOAA Capstone Project Mid-Term Summary Report

Team 3: Shamika Kalwe, Selma Sentissi, Jaya Nagesh, and Shuyi Zhu

Date: 29th April 2021

Introduction and Business Problem

In today's era, studying drought and its impact on the agricultural industry is more important than ever before. In simple terms, drought is defined as the absence of water for a prolonged period[1]. As a result, when drought is present, water availability and water quality both reduce, causing a decrease in crop production, an increase in pests/crop diseases, thus all contributing to a negative economic and socioeconomic impact. Within the US, 1,467 counties are experiencing drought currently which translates to 152.6 million acres of crops experiencing drought. In fact, the cost of drought events accounts for 9 billion dollars every year.[2] Hence it is vital for farmers to be notified about drought events so that they can prepare to combat the ill consequences of drought as far in advance as possible.

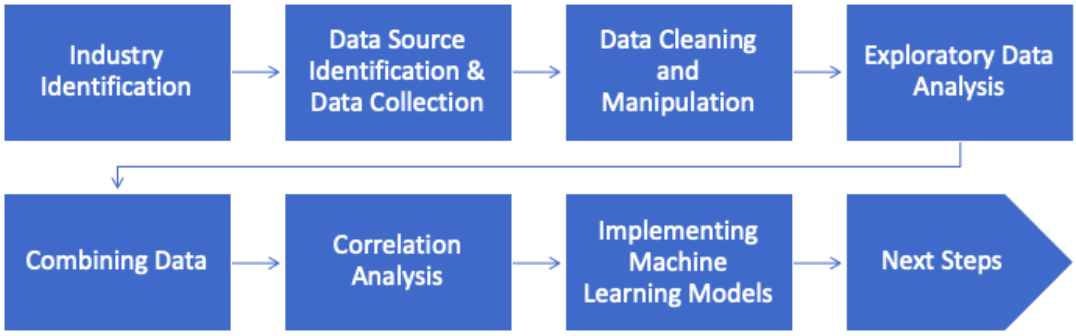
Hence, the purpose of this project is to increase dissemination of drought related information to farmers in a way that allows them to develop a multi-seasonal drought mitigation strategy. Even though there is vast amount of data about drought monitoring already available, we believe a customized front-end may drive the users in the farming industry to use the data more effectively. In this project, we plan to build a database to store the necessary portion of the drought-related data, a backend to update and maintain data, a front end that delivers relevant reports, as well as a plugin that performs predictive analysis based on available data. Technically, the most challenging aspect of this project is predictive analysis plug-in, yet we believe predictive analysis can draw significant attention to the product and it can showcase the capabilities of the current data technologies to the farming industry.

As a final note, for this project, we focus our analysis on New York. New York is known to be a leading agricultural state, producing an astounding 5.75 billion dollars in revenue in 2017. Furthermore, it is critical to note that the crop production has decreased since the year 2012 to 6,866,171 acres in production, down from 7,183,576 in 2012, drought being a possible culprit for this decrease[3]. We will explore the crop yield and drought related effects through the analysis that will be conducted.

---

[1] https://www.ncdc.noaa.gov/monitoring-references/dyk/drought-definition
[2] https://www.drought.gov/sectors/agriculture#key-issues
[3] https://www.nyfb.org/about/about-ny-ag

Methodology



Industry Identification

The goal of this project is to increase drought related information available to end-users who can then use it and develop a drought mitigation strategy. The data is already available publicly, however, the needs vary depending on the user. When considering industries, we thought about Public Education & Public Health, Ecosystem, Water Supply, and Agriculture. We researched all above topics and decided to go with Agriculture. Our choice was based on the fact that we could find a few great datasets that we could use and combine to make predictions. Overall, we would like to correlate drought with Agricultural industry (starting from a crop-state pair and expanding from there, e.g. Corn-New York). Then, we would like to predict parameters related to Agriculture with reference to drought (e.g. yield etc.).

Data Source Identification and Data Collection

We collected the data for different indices in order to start our analysis. Thus far, we have collected and worked with the following: Temperature, Precipitation, U.S. Drought Monitor (USDM), U.S. Department of Agriculture (USDA). We also have collected and are looking forward to working with the following: Evaporative Demand Drought Index (EDDI), Standardized Precipitation Index (SPI), and Standardised Precipitation-Evapotranspiration Index (SPEI).

For each index, we contacted experts in the field whom Sylvia recommended. Mike Hobbins was very helpful in collecting the EDDI, SPI, and SPEI indices. Keith Eggleston assisted us in configuring the API for the temperature and precipitation indices to get daily values for the years 2000 to 2020 and 2002 to 2020 respectively. Brian helped our team retrieve USDM data using USDM REST API. Each and every person's input was very helpful for the project. Here are some more details about our current indices:

Temperature and Precipitation:

- About:

  - *Temperature:* We computed the average, maximum, and minimum daily temperature for every New York county for the years 2000 to 2020 in Fahrenheit.
  - *Precipitation:* We computed the average daily temperature for every New York county for the years 2002 to 2020 in inches.
- Data Source: We used the Applied Climate Information System API services to download and store the data via a Python script as a JSON file [4]. We collected this data for New York state for the period of 2000-2020.

U.S. Drought Monitor (USDM)[5]:

- About: It is a composite index, generated weekly, which gives a measure of drought for a particular region. It uses five classifications: None: No Drought, D0: Abnormally Dry, D1: Moderate Drought, D2: Severe Drought, D3: Extreme Drought, D4: Exceptional Drought. A higher value indicates higher drought.
- Data Source: We used the USDM REST API services to download and store the data via a Python script. We collected this data for New York state for the period of 2000-2020.

Evaporative Demand Drought Index (EDDI)[6]:

EDDI is the measure of how anomalous the atmospheric evaporative demand ("the thirst of the atmosphere"). It is for a given location and across a time period of interest.

Standardized Precipitation Index (SPI)[7]:

SPI is a multiscalar drought index based on precipitation. A negative value is indicative of drought, and positive value is indicative of wet conditions. It can be computed for a given location and across a time period of interest.

Standardised Precipitation-Evapotranspiration Index (SPEI)[8]:

SPEI is a multiscalar drought index based on climatic data that includes both Precipitation and temperature. Similar to SPI, a negative value is indicative of drought, and positive of wet conditions. It also can be computed for a given location and across a time period of interest

Mike Hobbins assisted us in collecting 3-, 5- and 6-month EDDI (ending September 30) for the period of 2000 to 2020. He also helped us collect 3- and 6-month SPI and SPEI (ending September 30) for the period of 2000-2020. We have received this data (files of .nc format) and will incorporate the same in our analysis going forward.

[4] http://www.rcc-acis.org/docs_webservices.html#griddata
[5] https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx
[6] https://psl.noaa.gov/eddi/
[7] https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/indices/spi/div#select-form
[8] https://spei.csic.es/

USDA:

- About**:** Provides us target variables like yield (bushels/acre), acres planted, and other metrics to aid our analysis like commodity, state. We have used the USDA data as our outcome/predictor variable (y-variable).
- Data Source**:** We requested for the USDA API key so that we can make the data downloading process more efficient and automated. After obtaining the API key, which was relatively fast, we worked with the USDA API to extract the relevant data in CSV format. As a last note, the USDA API was fairly well documented which made working with the API much less of a hassle.

Data Cleaning and Manipulation

USDM[9]*:*

To bring the USDM data (multiple values weekly) at a comparable time-scale to USDA (single annual value) we did some manipulation. We computed the Drought Severity And Coverage Index (DSCI) using the formula: *DSCI = 1(D0) + 2(D1) + 3(D2) + 4(D3) + 5(D4)*. This was done to get one value for each county for each week. Further, to aggregate USDM over time we computed Accumulated DSCI (ADSCI) for each year and also for the growing season by using the formula: *ADSCI = Summation of DSCI from i=1 to n weeks*.

Temperature & Precipitation:

We transformed the JSON file onto a data frame by creating a dictionary of the different days, counties, and their values. Then, we aggregated the indices over time to get weekly values for both indices. The Temperature and Precipitation data frames we converted from using ACIS web service record the values on a daily basis. We have 62 New York counties as columns, with 6939 rows of average precipitation value (range from 2002-01-02 to 2020-12-31) in the precipitation data frame. As for the temperature data frame, we have 62 New York counties as columns, and 7670 rows of average, maximum, and minimum temperature values (range from 2000-01-01 to 2020-12-31). We resampled these two data frames to a weekly basis, the same granularity as the USDM and USDA data sets. We also prepare the data frames for both complete year analysis and the typical corn growing season (from May to September) analysis. We will link these data frames with USDM and USDA data for further analysis.

USDA:

The following columns were dropped: "Period", "Week Ending", "Geo Level", “state”, "State ANSI", "Zip Code","Region", "watershed_code", "Watershed". This was mostly because these columns were completely null (NaN) or would not carry much value for our analysis. Next, we casted the remaining columns to either category, float64 or string rather than leaving columns as objects. Finally, two subset data frames were created from the main dataframe, one subsetting the data item to yield in bushels/acre and another subset for acres planted.

[9] https://droughtmonitor.unl.edu/About/AbouttheData/DSCI.aspx

Exploratory Data Analysis

USDM:

We visualised the USDM values for the NY state (Appendix 1) and found that NY saw D0 level of drought frequently across the 2000-2020. D1, D2 level of drought was recorded sparsely across a few years. D3 was recorded only in 2002 and 2016. NY never witnessed a D4 level of drought. The Accumulated USDM (ADSCI) has varied across the years peaking in 2002 and 2016 (Appendix 2).

Temperature & Precipitation:

In the preliminary exploratory data analysis with temperature and precipitation, we worked on the overall trending of these two values among New York counties from 2000s to 2020s. We looked at the temperature and precipitation changes for top 5 counties with the highest value respectively. We can see from our temperature plot that there is an increasing trend in temperature value for the two recent decades. As for the precipitation plot, the precipitation values were pretty spread out, and we did not find a particular pattern showing either an increase or decrease trend.

USDA:

Preliminary EDA was conducted for yield in bushels/acre and acres planted. Firstly, we found that acres planted for corn has steadily decreased throughout the years. Secondly, corn output varies by agricultural district, Western County producing the most output by far, followed by northern, central, and southwest following. Similarly, corn output varies by county – some counties like Cayuga and Livingston produce far greater output than other counties which are low producing like Albany county. Lastly, corn yield over the years has increased, with some ups and downs intermittently (Appendix 3).

Correlations

Region - Value	USDM (complete year)	USDM (growing season)	Temperature (complete year)	Precipitation (complete year)
NY - Yield	-0.054	-0.133	-	-
NY - Acres Planted	-0.006	-0.05	-	-
County - Yield	-0.07	-0.103	0.162	-0.098
County - Acres Planted	0.067	0.142	0.115	-0.303

For USDM we see a slight negative correlation with Corn Yield and a slight positive correlation with Acres Planted under Corn. We can see that the correlation values get amplified when looking at the USDM values in the growing season. For Temperature we see a slight positive correlation which could indicate low temperatures are not ideal for corn production. For Precipitation we see a slight negative correlation which could indicate that extreme wet conditions can negatively impact corn production.

Implementing Machine Learning Models

USDM:

Machine Learning models like Linear Regression, and Random Forest Regressor were applied to predict both Yield and Acres Planted for Corn using the Accumulated DSCI (ADSCI) over 1. Complete year and 2. Growing season (May-September). Overall the models results were not good (low R2, high RMSE) i.e. they had low predictive power. This is probably due to the fact that only one predictor (USDM in this case) may not be effective in predicting Corn metrics.

Temperature:

First we tried Machine Learning models like Linear Regression, Gaussian, and Random Forest Regressor to predict both Yield and Acres Planted using just the Average Annual Precipitation. But the model results were not good (high RMSE, low R2) which indicated low predictive value of the model. Next we tried to add multiple representatives of Precipt (quartiles etc.) but once again, the results were not conclusive.

Precipitation:

Preliminary machine learning was conducted using linear regression and random forest algorithms to predict yield and acres planted. Average mean precipitation (in models with only one X) was used while precipitation in the 25%, 50%, 75% percentiles and the mean were all used as Xs in models with more than one X variable. Out of all models for predicting acres planted, Linear Regression with multiple X variables performed the best with MSE being .16 and for predicting yield in bushels linear regression with multiple variables also performed best with .01 MSE.

Combining Temperature, Precipitation and USDM:

Next, we added all the predictors i.e. accumulated USDM (ADSCI), annual precipitation means and 25%, 50%, 75% percentiles, annual temperature means and 25%, 50%, 75% percentiles. In addition, we also focused on the growing season i.e. we took the predictor values only from the growing season of corn (May-September). The results of the Linear Regression and Random Forest for both are shown below.

Variable: Corn Yield (BU/ acre); Predictors: Temperature, Precipitation and USDM			
Linear Regression		**Random Forest**	
Predictors measured Annually	**Predictors measured over Growing Season**	**Predictors measured Annually**	**Predictors measured over Growing Season**
r2: 0.1885	r2: 0.2357	r2: 0.3202	r2: 0.2645
RMSE: 20.6058	RMSE: 19.9977	RMSE: 18.8602	RMSE: 19.6169
After Standardizing Data			
r2: 0.1885	r2: 0.2357	**r2: 0.3396**	r2: 0.274
RMSE: 0.7876	*RMSE: 0.7643*	*RMSE: 0.7105*	*RMSE: 0.745*

Note: We tried the above analysis for Acres Planted too, however overall the models results were not good, which indicated that Acres planted is probably not a good metric to predict. This could be because there are many other factors that affect the acres being planted under corn.

Challenges Faced

The data sources were spread out. Thus, for the Data identification and collection part, having to identify relevant contacts, contacting them and waiting for them to get back to us slightly slowed down the process. In addition, each dataset had a different process to download (different websites and API processes) and had a different format. Thus the data took a lot of effort in terms of downloading, transforming, cleaning, and finally merging to make it ready to use.

Current Status

Currently, we have largely concluded the data gathering stage, and moved on to perform exploratory data analysis where we correlate Temperature, Precipitation and USDM, EDDI, and SPI indices with USDA data on yield. The most significant challenge we encountered so far has been due to dispersed data with varying formats and the associated learning curve.

Next Steps

Our next step is to conclude the exploratory data analysis, design the sample data distribution for the machine learning and build the predictive models. We will include all the indices we have (temperature, precipitation, EDDI, SPI, etc.) as input values and predict the corn yield. Apart from the linear regression and random forest regressor models, we will also try to work on other machine learning models such as kernel, bagging, boosting, and polynomial regression. We will then continue with dashboard design and create a database to: 1). Store and update relevant data from public repositories, with a backend to update/merge routinely or as needed. 2). Enable dissemination of the above analysis to farmers. We also want to create a Github Readme file where we explain to people how all the code files work together and what are the dependencies among the files, for people interested in our project. Finally, if time permits, we plan to expand our analysis to make a more significant impact - to not only include corn but other crops such as apples or cranberries, which are also economically substantial to new york farmers.

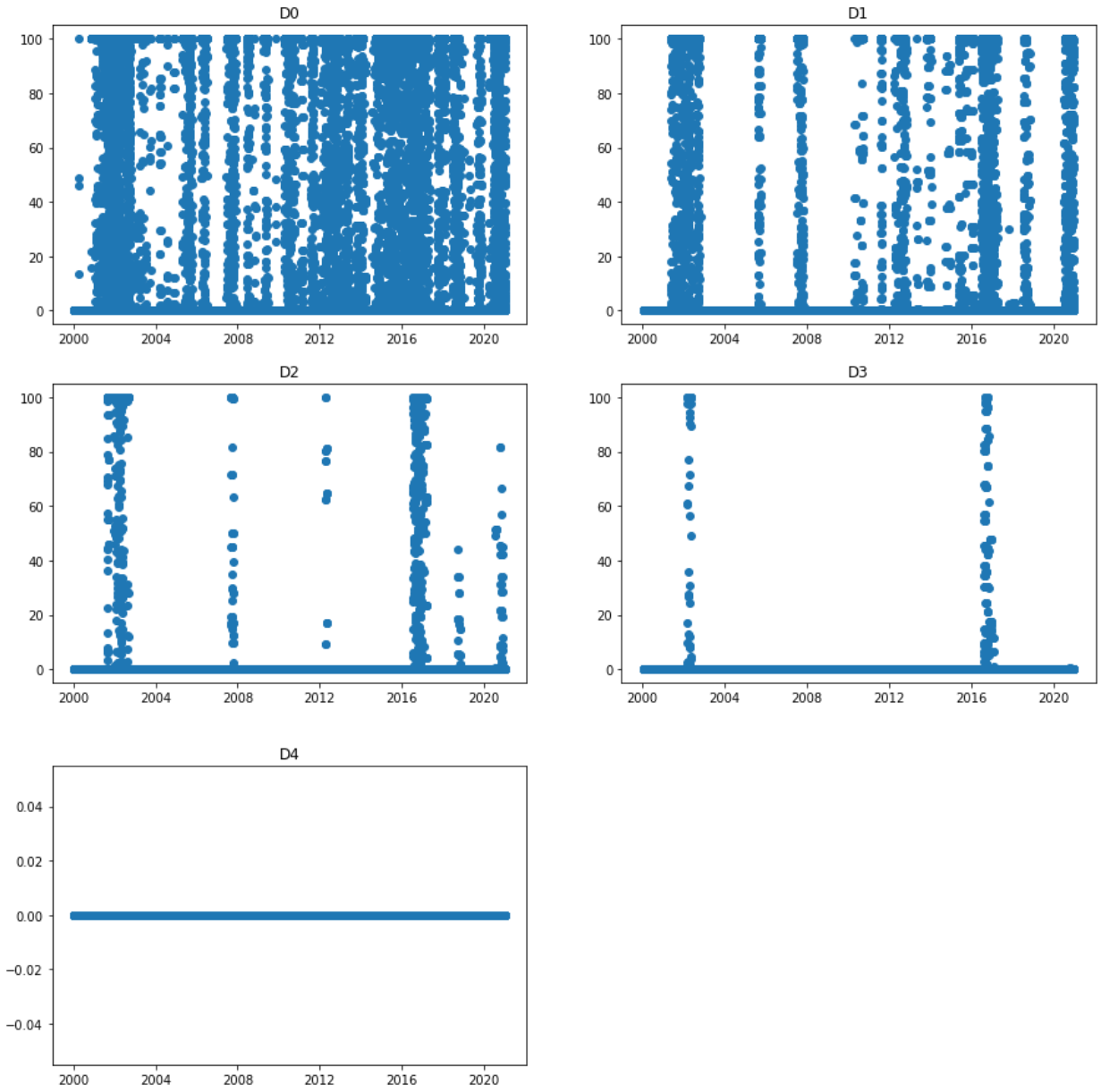
Github Repository Link:

https://github.com/BU-NOAA-Capstone/Capstone

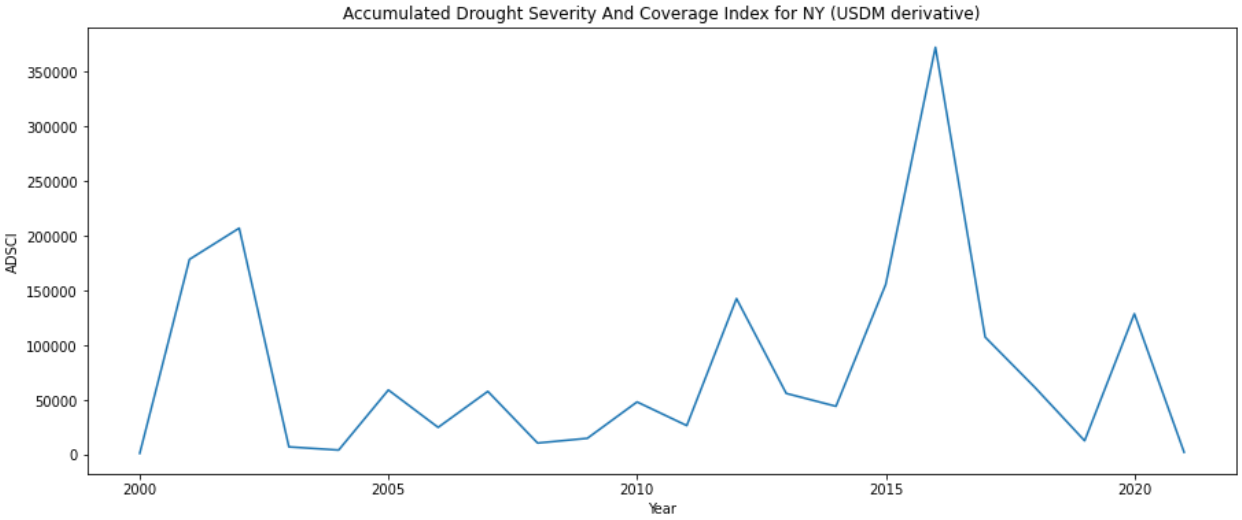
Appendix

1. USDM Drought levels (D0, D1, D2, D3, D4) across the state of New York during 2000-2020

   D0: Abnormally Dry, D1: Moderate Drought, D2: Severe Drought, D3: Extreme Drought, D4: Exceptional Drought.



2. Accumulated Drought Severity and Coverage Index (ADSCI) for NY during 2000-2020



3. Corn Yield (Bushels per Acre) across the years