Data 512 - Wildfire Project Part 4 - Final Report

The project focuses on the impact of wildfires on the socio-economic impact of major cities in the US. The smoke caused by these wildfires has affected major fields like agriculture, health, tourism, property, and other societal aspects. The project's goal is to inform policy-makers, city managers, city councils, or other civic institutions, to make an informed plan for how they could or whether they should make plans to mitigate future impacts from wildfires. Split into 4 different stages we are trying to propose to the authorities that bring out the human-centered aspect of the otherwise technical project. The project is a perfect example of the application of Data Science for social change.

Introduction

As mentioned earlier, the project revolves around the human-centered application of utilizing the predictive power of sophisticated Data Science techniques. Wildfires have the potential to wreak havoc on several aspects of our lives. These repercussions were deeply studied by the class and a plethora of socio-economic ramifications were presented. The themes involved education, agriculture, tourism, employment, healthcare, etc. I went down the economic route where I focused on the impact of wildfires on agriculture and the yield is considered an economic indicator as it is responsible for direct cash receipts.

Why is this analysis interesting or important?

Wildfires are coming for our sustenance. Its impact and menace cast shadows beyond scorched lands. They directly impact the economic conditions of the state. We measure the economic impact being proportional to the yield.

- **Sunlight Deprivation**: Wildfire smoke creates a thick haze, obstructing sunlight essential for plant photosynthesis.
- **Harmful Emissions**: Emanating pollutants affect air quality, leading to respiratory health issues and environmental degradation.
- **Crop Impact**: Smoke settling on farmlands reduces crop yield quality due to reduced sunlight and contamination.
- **Barren Lands**: Prolonged exposure to smoke and fire damages fertile soil, rendering acres barren and challenging agricultural sustainability.

From the factors listed above we can see that the wildfires not only impact yield directly but also impact indirectly by coming for the land that the crops are cultivated on.

What motivates asking and answering the question?

Understanding the economic impact of wildfires on agriculture is paramount owing to its multifaceted implications for both the agricultural sector and the broader community. This analysis is not merely helpful in quantifying the relationship between wildfires and agriculture but also is inclined to provide

and assess the damage to the economic stability of communities reliant on agriculture. Wildfires, with their influence, pose a substantial threat to agricultural productivity, soil health, and the livelihoods of farmers, engendering a ripple effect on food security, local economies, and the overall socio-economic fabric. By diving deep to perceive the effect, the analysis seeks to address a pressing concern: how these natural disasters reverberate through the agricultural landscape and the subsequent ramifications for communities reliant on agriculture.

Does it solve a real problem or tackle an unresolved research question?
The analysis asks certain relevant research questions.
Q1: Is the wildfire an impact on the agricultural yield? If yes, is the relationship between the two statistically significant?
Q2: How exactly does the wildfire impact agriculture?

Answers to these questions would likely emphasize the observed correlation between wildfires and agricultural yield, potentially highlighting statistical significance in their relationship. Regarding the impact, the explanations might elaborate on how wildfires affect agriculture, citing instances like diffused sunlight hampering photosynthesis, harmful emissions, and smoke contaminating crops, leading to lower-quality yields, and contributing to land degradation, thereby outlining the broader ecological and economic repercussions.

Why does it matter?

Precise forecasts of daily wildfire expansion rates hold significant importance in light of the escalating frequency of severe wildfires. The determinants influencing these rates are intricate, hinging on various meteorological elements, landscape features, and fuel quantities [1].

Progress so far

The project is nearing completion with a few operational changes and repository changes. The project has been through multiple iterations and a variety of stages where each stage had a particular objective that flowed across each task. The four stages are described below:

- **Part 1 - Common Analysis:** This initial phase lays the foundation for subsequent assignments. In this segment, a fundamental analysis is conducted. Each student explores the same analysis, albeit with a slightly varied dataset.
- **Part 2 - Extension Plan:** This stage involves formulating a human-centered data science query that extends the groundwork from Part 1 - Common Analysis.
- **Part 3 - Presentation:** Here, a condensed PechaKucha presentation based on the completed project is required, showcasing the findings and outcomes.
- **Part 4 - Project Repository:** This phase centers on creating a meticulously documented repository, alongside the submission of a comprehensive written project report.

Part 1 Progress - We focused on examining wildfire data around Bismarck, North Dakota, to estimate smoke impact and evaluate its correlation with the Air Quality Index (AQI). The process involved

filtering relevant fires, estimating smoke impact factors, and developing predictive models for future smoke estimates.

1. Data Extraction: Extracted and pre-processed wildfire attributes such as fire size, duration, proximity to the city, etc.
2. Filtering Relevant Fires: Focused on fires within 1250 miles since 1963, analyzing proximity-related incidents.
3. Smoke Impact Estimation: Developed a formula considering fire size, intensity, and proximity to estimate smoke impact during the fire season.
4. Comparison with AQI: Compared smoke impact estimates against available AQI data, considering variations in estimation methods.
5. Predictive Modeling: Developed predictive models to estimate smoke impact for the next 25 years based on fire data and smoke estimates.

This comprehensive analysis aimed to provide insights into the relationship between wildfires and socio-economic factors, using predictive modeling to forecast smoke impact in the coming years.

Part 2 Progress - This part of the project focuses on moving the attention away from data science and the mathematical implications toward a more human-centered application.

- The project explores the socio-economic effects of wildfires in major US cities (agriculture, health, tourism, property).
- The extension plan focuses on the post-wildfire economic impact on agriculture in North Dakota.
- Aims to analyze crop production, and economic stability in the assigned city (state capital).
- Seeks insights into how wildfires influence local economies, particularly in agriculture.

Part 3 Progress - The extension plan is now implemented and the results are presented to "City Council". This part of the project is all about implementing the idea that was proposed in part 2.

- We first implement the feedback and improvements on the common analysis. This will yield a change in the data we consider and slight alterations in the previous results
- Then the focus is on acquiring agricultural data for project extension.
- This data is then analyzed and an attempt is made to establish the relationship between wildfire and agriculture
- Once the relationship is established and the statistical significance is proven, then we forecast the values sometime into the future to analyze and quantify the effect of wildfire on yield.
- Some recommendations are formulated based on the results of the model.
- All these results are collated into a PechaKucha-style presentation and presented to the city council with recommendations

Work in progress and next steps (Part 4) - The culmination of the efforts of over half a quarter.

- Curated and unified artifacts from all preceding stages into a centralized repository.
- Developed a comprehensive report detailing the project's entire journey, methodologies employed, and the consequential outcomes, articulating them eloquently and succinctly.

Background/Related Work

What other research has been done in this area?
I went through blogs, articles, and some hearings before turning the focus on agriculture, they are summarized below.

Literature review

Blogs and Articles:

1. **NASA Applied Sciences Program** - Wildfire Impact on Agriculture: This resource discusses how NASA uses satellite data and models to monitor wildfires and their effects on agriculture.
2. **UC Davis Center for Spatial Technologies and Remote Sensing (CSTARS) Blog**: The CSTARS team often publishes articles on the use of remote sensing and GIS technologies to analyze wildfire impacts on agriculture.

Academic Papers:

[1] Shmuel, A., & Heifetz, E. (2023). "A Machine-Learning Approach to Predicting Daily Wildfire Expansion Rate." Fire, 6(8), 319. https://doi.org/10.3390/fire6080319 - This research paper delves into machine learning approaches to predict the potential impact of wildfires on crop yields based on historical data.

[2] Meier, S., Elliott, R. J., & Strobl, E. (2023). "The regional economic impact of wildfires: Evidence from Southern Europe." *Journal of Environmental Economics and Management*, *118*, 102787. https://doi.org/10.1016/j.jeem.2023.102787 - This study presents a modeling framework to assess the economic repercussions of wildfires on agricultural production and revenue.

[3] K. C.: "A spatial database of wildfires in the United States, 1992-2011", Earth Syst. Sci. Data, 6, 1–27, https://doi.org/10.5194/essd-6-1-2014, 2014.. - This paper employs spatial analysis techniques to identify and quantify wildfire-affected agricultural areas.

[4] Xi, Dexen & Taylor, S.W. & Woolford, Douglas & Dean, C.B.. (2017). "Statistical Models of Key Components of Wildfire Risk." Annual Review of Statistics and Its Application. 6. 1-26. 10.1146/annurev-statistics-031017-100450. - This research focuses on statistical models to estimate the damage caused by wildfires on crop yields, incorporating various environmental and fire-related factors.

Committee Hearing:

Sumner, D. A. (2020, November 18). Economic Impacts of Recent Wildfires on Agriculture in California [Presentation slides]. California State Assembly Committee on Agriculture hearing.

How does this research inform your hypotheses, your analysis, or your system design?

These articles gave me an introduction to how we could use and relate wildfires and agriculture. Since I was a novice in this field of geospatial data and how to ingest and analyze them, the articles acted as a

guideline. When it was time to provide a human-centered twist, the committee hearing reports and the blogs helped me narrow down my focus on agriculture. One paper discusses the approach of integrating Machine Learning to predict the devastating effects of wildfires which I found very relevant to my approach. I read through the pieces of literature and decided to concentrate on forecasting the impact of wildfires on yield and cultivable land. These numbers would help me drive home the importance of the impact and also recommend suggestions.

What are your hypotheses or research questions?

My research questions are listed below.

Q1: Is the wildfire an impact on the agricultural yield? If yes, is the relationship between the two statistically significant?
Q2: How exactly does the wildfire impact agriculture?

The hypothesis behind these questions was:
Null Hypothesis (H_0): There is no significant impact of wildfires on agricultural yield (economic indicator).

Alternate Hypothesis (H_1): Wildfires have a significant impact on agricultural yield, affecting the economic indicator.

Hypothesis Test

Requirements to Reject the Null Hypothesis:

Statistical Significance: We require statistical evidence suggesting a strong relationship between wildfires and agricultural yield. This evidence should have a low p-value (typically less than 0.05) to support the alternate hypothesis.

Inspiration from Previous Works and Model Selection

Although the papers gave me immense knowledge about the impact of wildfires and how they tie together with the agricultural impact, my focus and the way I wanted to implement deviated from what I read. I wanted to utilize time series forecasting to give an example of how things can look in the future if we don't adopt policies and practice precautions.

Methodology

My approach followed the same flow as the focus and objective of each stage of the project. A basic workflow of my methodology is depicted in the image.

Figure 1 - Flowchart of the methodology

- **Region of Interest - Bismarck, North Dakota**
- **Methodology Summary**
    - Obtain Wildfire data and extract relevant attributes
    - Quantify a wildfire impact by creating a smoke impact
    - Find a focus to establish an economic impact on the city of Bismarck - Agriculture
    - Establish a statistically significant relationship between the area of focus and wildfires
    - Forecasting the future yield
    - Recommendations based on studies

Before diving into how I implemented the project I would like to draw attention to why I made certain decisions and the assumptions I made along the way.

1. *Motivation and Focus:*

   The initial phase, Common Analysis, is the project's starting point, aiming to evaluate wildfires' impact within the allotted cities in the US. This phase involves a detailed analysis of the city's fires to estimate their potential effects, encompassing health, tourism, property, and broader societal aspects. In part 2 - I proposed to extend the focus to agriculture. This analysis is not merely helpful in quantifying the relationship between wildfires and agriculture but also is inclined to provide and assess the damage to the economic stability of communities reliant on agriculture.

2. *Agriculture as an Economic Focus for North Dakota:*

   Agriculture is one of the main revenue-generating industries in the state of North Dakota. Wildfires and Agriculture have a direct impact on each other and agriculture is also directly responsible for the cash receipts and the economic health of the state. Hence, by the transitive property, we can claim that wildfire's effect on agriculture bears a weight on the economy of the

state. In this analysis, we focus on the yield of Oats and use all the corresponding attributes to forecast the yield of oats. The reason we are focusing more on Oats than any other crop is because it is concentrated and grown around the belt of the region where Bismarck lies. Hence there is a direct impact of oat on the economy of the city. The figure below shows the major crops of North Dakota.

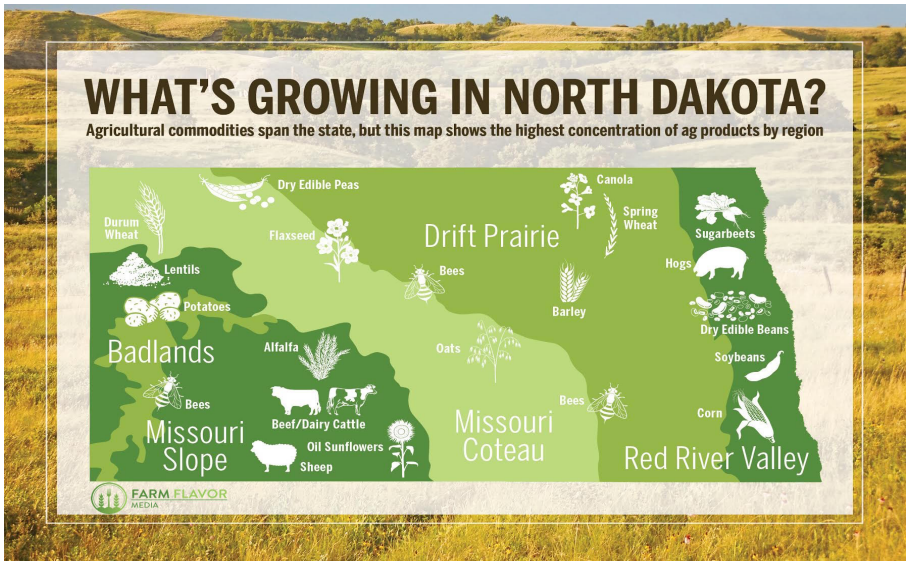


Figure 2 - Crop Distribution in North Dakota

3. *Significance of Yield as an Economic Indicator:*
    a. **Wealth Generation:** Higher yield signifies increased agricultural output, contributing significantly to a state's economic wealth through enhanced revenue from crop sales.
    b. **Cash Receipts and Income:** Improved yield translates to higher cash receipts from agricultural products, directly impacting the income of farmers and stakeholders.
    c. **Employment and Economic Activity:** Robust agricultural yield creates job opportunities and stimulates economic activities in related industries, bolstering the state's economy.
    d. **Market Competitiveness:** States with higher yields often have a competitive edge in regional or global markets, attracting investments and boosting trade activities.
    e. **Food Security and Stability:** Adequate yield ensures food security, stabilizes prices, and reduces dependence on external food sources, fostering economic stability

Obtain Wildfire data and extract relevant attributes

Data Acquisition

- Assigned City - Bismarck, North Dakota [46.825905, -100.778275.]
- Objective - Gather crop data for the assigned city.

- The data resides in a GeoJSON zip file ("Wildland Fire Polygons Fire Feature Data Open Source GeoJSON Files"). The file that contains the data of interest is in USGS_Wildland_Fire_Combined_Dataset.json.

Data Subsetting

Obtain different attributes of the wildfire data.

1. Getting Distance information
2. Get other attributes like Fire_Year, GIS_Acres, Overlap_Within_1_or_2_Flag, Shape_Area, etc to give us factors that can help quantify the impact of the wildfire i.e, smoke estimate.

Filter the wildfire data according to these conditions:

1. The estimate only considers the last 60 years of wildland fires (1963-2023).
2. The estimate only considers fires that are within 1250 miles of your assigned city.
3. An annual fire season will run from May 1st through October 31st.

The wildfire data is processed according to the conditions and stored.

Quantify a wildfire impact by creating a smoke impact

Creating a smoke estimate

Creating a smoke estimate requires us to consider several aspects of the wildfire such as fire size, intensity, distance from the city, and other relevant categories. The approach I have used to create the fire estimate is.

Note - I have considered both prescribed and wildfires since they both attributed to the gaseous and particulate pollution despite a difference in the magnitude.

1. Select relevant variables:

The attributes of the fires were carefully picked to provide the most information about the smoke estimate. The factors I considered are

- Size of the fire
- Distance of the fire from the city
- The intensity of the fire
- Overlap Component

2. Define a function for estimating:

Create a formula or model that combines these variables to estimate smoke impact. This formula assumes that larger fires closer to the city and with higher intensity produce more smoke. The formula I have used is:

Smoke Impact = (Fire Size / Distance) * Fire Intensity * (1+ Overlap Component)

where :

- Fire Size = Area that the fire has covered

- Distance = Shortest distance of the fire from the assigned city
- Fire Intensity = The areas of the shape of the fire and the days the fire lasted
- Overlap component

3. Amortization Over the Fire Season: I am estimating the smoke impact cumulatively throughout the fire season. Since there is no clear information about the duration or time at which the fire was burning.
4. Apply the Estimate: I will then apply my formula to calculate the smoke impact for each fire within 1250 miles of my city during the annual fire season.
5. Aggregate Annual Estimates: I will then average the smoke impact estimates for each year to get an annual estimate.
6. Data Validation: Another problem is trying to understand how good or bad the smoke estimate might be. This estimate will be compared to available AQI (Air Quality Index) data from the US EPA. This will help ensure the reasonableness of the estimates.

The Motivation Behind Direct Contributors of the Smoke Estimate

Size of the fire

The size of a wildfire profoundly influences the volume of smoke it generates. Larger fires, by burning more extensive areas of vegetation, emit substantially greater quantities of smoke particles and gases into the atmosphere.

Distance of the fire

Wildfires closer to cities tend to pose more immediate threats due to the shorter distance for smoke to travel, potentially resulting in poorer air quality and health hazards for residents.

Fire intensity

The intensity of a wildfire profoundly influences smoke production, with higher intensities generating increased amounts of smoke and pollutants.

Estimating Fire Intensity

We estimate the Fire intensity component by the following steps

1. Consider the attribute shape area - This gives us the area of the entire geographical region that was burnt during the wildfire.
2. Listed_Fire_Dates - this gives us the days for which the fire burned. We calculate days for which the fire burned with the field and some regular expression manipulation
3. Then the shape area multiplied by a factor of the number of days the fire burned that will give us the intensity of the fire.
4. For the fires that don't have the number of days, they will just have the area value. The days for which the fire burned are then scaled between 0 and 1 since it is a multiplier
5. The fire intensity is calculated as

Fire Intensity = shape_area * (1+scaled_days_fire_burned)

Overlap with Previous Fires

The contribution of the spatial overlap to the smoke estimate can still be significant.

1. Impact on Environment: Even if wildfires occur at different times but in the same geographical area, their combined impact on the environment can be notable. The land might have experienced reduced vegetation due to earlier fires, making it more susceptible to subsequent fires, altering soil conditions, and impacting recovery processes.
2. Air Quality and Public Health: Although not simultaneous, the spatial overlap can affect air quality over time. The accumulation of smoke-related pollutants, the prolonged exposure of affected areas, and the potential for repeated disruptions to air quality due to recurrent fires can impact public health, particularly for vulnerable populations.
3. Residual Effects: The aftermath of earlier fires, such as increased dryness, altered ecosystems, or changes in vegetation, can influence the behavior and severity of subsequent fires in the same area. This can affect the intensity and duration of smoke production during later fires.

Estimating overlap component

We estimate the overlap component by the following methodology

1. We consider a typical 'Overlap_Within_1_or_2_Flag' value which is of the form 'Caution, this Wildfire in 1963 overlaps with a Wildfire that occurred in 1961 (2 year difference). The overlapping fire overlaps by 30.5% (196.0 acres). Overlapping fire USGS Assigned ID: 13685.'
2. From the string, we will extract 3 important components
    - The years' difference between 2 overlapping flags
    - Area of overlap in acres
    - Percentage of overlap between the wildfires.
3. The area of overlap and percentage of overlap is directly proportional to the factor while the years of difference are inversely proportional
4. We calculate the overlap component by this formula:

Overlap Factor= (1/Time Difference+1) * Percentage of Overlap * Area of Overlap 5

This value is then scaled between 0 and 1 since it is a multiplier

The smoke estimate was then forecasted over 25 years.

Part 1 - Feedback Implementation

The feedback indicated that there were some miscalculations. These were altered before proceeding with the next steps.

1. Issue - An annual fire season will run from May 1st through October 31st. Currently the AQI for a year has erroneously been calculated as the average AQI over all months.

   Solution: This has to be altered to include the AQI only from May to October.

2. Issue - The question is to *Produce a histogram showing the number of fires occurring every 50-mile distance from your assigned city up to the maximum specified distance.*

   Solution: Currently the data used to plot this chart is a yearly fire estimate with AQI. However, the data required for the histogram is supposed to be at a wildfire level. The data is unfortunately at a yearly level which is misrepresented as the number of fires in the histogram. This yields a low count of fire in the histogram. The current order of fire is 50-60; Actual order of fire is in the 90000+.

Find a focus to establish an economic impact

As mentioned above before proceeding to the methodology I narrowed down the focus to agriculture due to its economic Impact.

Data Acquisition

- Assigned City - Bismarck, North Dakota [46.825905, -100.778275.]
- Objective - Analyze wildfire impacts on the city assigned
- In this step, we acquire the data that contains the agricultural data for the state of North Dakota from at least 1963 to 2023. United States Department of Agriculture has records of crops grown in the United States from 1919 till 2022. The initial task is to ingest this data and get the data relevant to the assigned city (Bismarck)
- The data resides in a rar file (https://www.nass.usda.gov/datasets/qs.crops_20231201.txt.gz). The file that contains the data of interest is in qs.crops_20231201.txt after extraction.

Data Subsetting

The data referenced above consists of data for the entire country of the United States. However, the focus of our investigation is Bismarck, North Dakota. So the data by applying subsequent granular funnels at each step is filtered and subsetted. The basic flow of the filtering process is as mentioned below:

1. Filter the data for the state of North Dakota
2. Store the Crop and agriculture information in a CSV file to support further processing.

Data Pre-processing

The code manipulates a dataset by performing various filtering and subsetting operations based on specific constraints. It gradually refines the dataset to meet specific criteria for analysis.

- Data Loading and Initial Exploration
- Duplicate Removal
- Initial Subset Based on Constraints
  - Subsets the data based on constraints like survey data, period (1963 to 2023), and a specific geographical area (Burleigh County).
- Further Refinement Based on Agricultural Practices:
  - Refines the subset data based on production practice, crop class, and utilization practices.
- Filtering Based on Unique Features:
  - Filters crops based on having exactly five unique features.

Incorporating wildfire impact into test relation

- **Identifying Crops with Highest Yield:**
    - Calculates the average yield for each crop over time.
    - Identifies and displays the crop with the highest average year-on-year yield.
- **Data Manipulation for Specific Crop Analysis:**
    - Imputes missing values through group-wise means for enhanced dataset completeness.
    - Converts data from wide format to long format for easier analysis.
- **Incorporating Wildfire Smoke Estimates:**
    - Aggregates the smoke estimates based on 'Fire_Year' to gain insights into smoke occurrences.
- **Final Dataset Assembly:**
    - Selects specific columns from the merged dataset ('final_df') to focus on pertinent factors for analysis.

Establish a statistically significant relationship

Wildfire Feature Selection

The process of selecting wildfire features involved analyzing various attributes to capture essential information. Utilizing correlation analysis and data aggregation facilitated the identification of key features, ultimately leading to the choice of 'smoke estimate' as a representative metric for wildfire impact in comparison to crop data.

1. **Identifying Varied Wildfire Attributes:** Aimed to select attributes capturing maximum variance and information.
2. **Dataset Acquisition and Initial Analysis:** Conducted a correlation heatmap analysis showcasing relationships among features.
3. **Correlation Heatmap Analysis:**
    - Relationships observed:
        - Strong positive correlation between 'smoke_estimate' and 'AQI'.
        - Other features showed a positive correlation except 'shortest_dist' (negatively correlated).

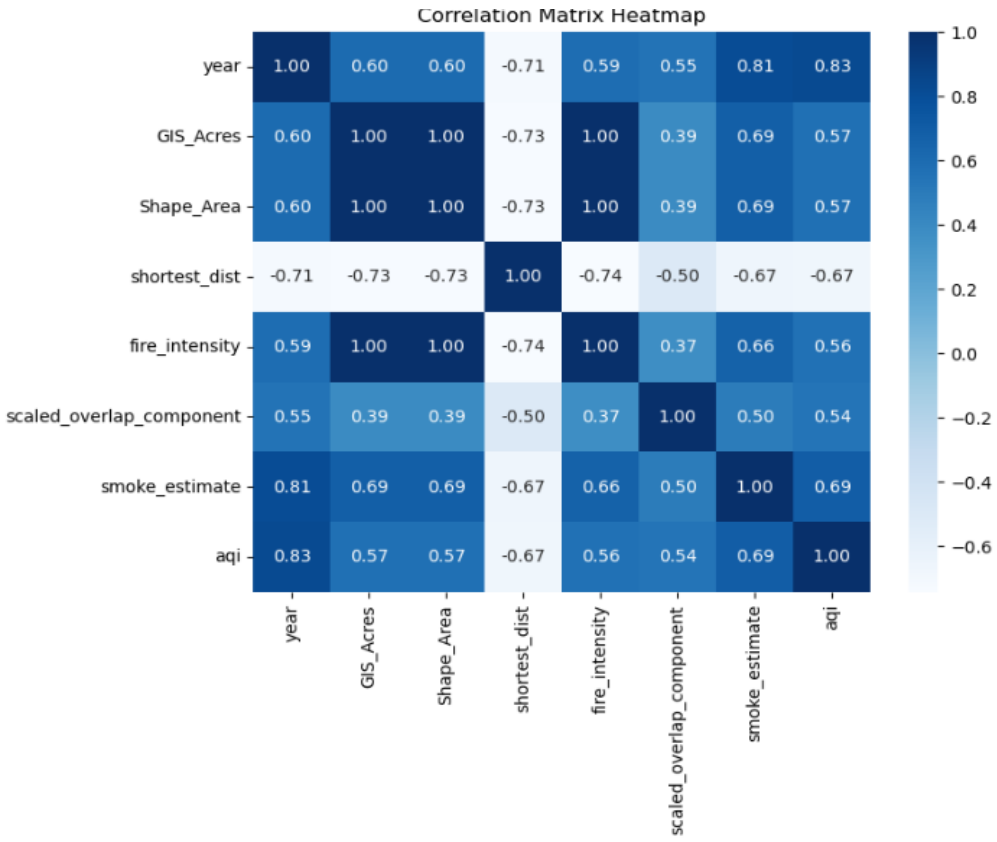


Figure 3 - Correlation Heatmap for wildfire data

4. **Refining Feature Selection:**
    - Noted that 'smoke_estimate' encapsulates information from other factors.
    - A strong positive correlation between 'AQI' and 'smoke_estimate' led to the exclusion of other attributes.
5. **Decision on Feature Selection:** Opted to move forward with 'smoke_estimate' as a proxy for assessing wildfire impact.
6. **Addition of Essential Feature:** Realized the absence of 'number of wildfires' as a crucial feature.
7. **Data Aggregation for Final Selection:**
    - Aggregated data over the year:
        - Counted the number of fires.
        - Calculated the mean of smoke estimate.

Establish a Correlation between Agriculture and Wildfire Data

Generating a correlation matrix and heatmap to visualize relationships between specific attributes related to wildfire impact and crop data. Similar to the steps mentioned above. The correlation heatmap is displayed in the figure below:

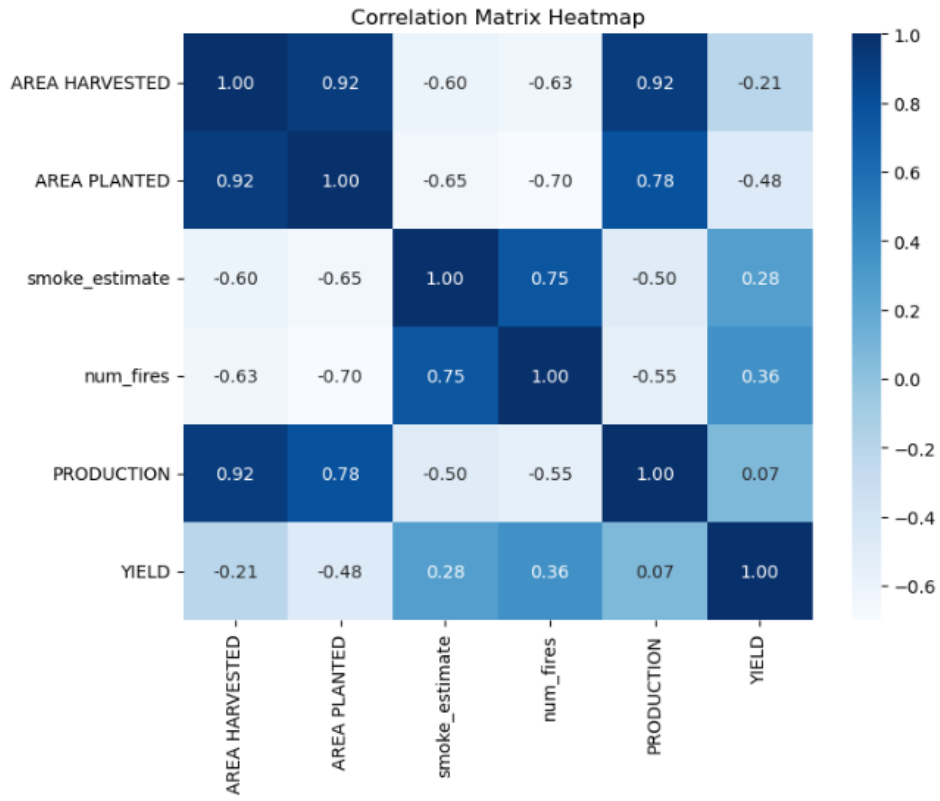


Figure 4 - Correlation Heatmap for crop data and wildfire data

1. **Positive Relationships:**
    - 'AREA HARVESTED' and 'AREA PLANTED' display strong positive correlations, nearly perfect, indicating high consistency between harvested and planted areas.
2. **Negative Correlation:**
    - 'smoke_estimate' shows a moderate negative correlation with 'AREA HARVESTED' and 'AREA PLANTED', suggesting a potential inverse relationship—higher smoke estimates aligning with smaller harvested or planted areas.
3. **Moderate Positive Relationships:**
    - 'num_fires' demonstrates moderate positive correlations with 'smoke_estimate' and 'AREA HARVESTED', indicating a potential connection between the number of fires and smoke estimates or harvested areas.
4. **Robust Positive Correlation:**
    - 'PRODUCTION' exhibits a strong positive correlation with both 'AREA HARVESTED' and 'AREA PLANTED', implying a significant association between production yield and the utilized harvesting or planting areas.
5. **Weak Correlation:**

- 'YIELD' showcases weaker correlations with other attributes, indicating a relatively independent behavior concerning the rest of the variables in this analysis.

Test the statistical significance relation between Agriculture and Wildfire Data

The next analysis conducts a thorough statistical analysis utilizing linear regression and correlation techniques. It defines independent ('AREA HARVESTED', 'AREA PLANTED', 'smoke_estimate', 'num_fires') and dependent ('YIELD') variables, creates a linear regression model, and assesses their significance. The analysis involves examining coefficients, intercepts, Spearman's rank correlation, and overall model significance. This comprehensive evaluation aids in understanding relationships between attributes and their statistical relevance.

1. **Variable Definition:**
   - Identification of independent (X: 'AREA HARVESTED', 'AREA PLANTED', 'smoke_estimate', 'num_fires') and dependent (y: 'YIELD') variables.
   - I am not considering PRODUCTION as an independent variable because YIELD is calculated as PRODUCTION/AREA HARVESTED. This would be redundant information.
2. **Model Creation and training**
3. **Coefficient and Intercept Analysis:** Coefficients and intercept values define the relationship between variables.
4. **Correlation Computation:**
   - Calculation of Spearman's rank correlation coefficient between 'smoke_estimate' and 'PRODUCTION' attributes, indicating the strength and direction of their relationship.
5. **Regression Analysis:**
   - Obtained a detailed summary containing coefficients, significance levels, and model fitness indicators as shown in the image.

OLS Regression Results

Dep. Variable:	YIELD	R-squared (uncentered):	0.928
Model:	OLS	Adj. R-squared (uncentered):	0.922
Method:	Least Squares	F-statistic:	173.0
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	4.21e-30
Time:	01:58:19	Log-Likelihood:	-228.44
No. Observations:	58	AIC:	464.9
Df Residuals:	54	BIC:	473.1
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
AREA HARVESTED	0.0012	0.000	4.752	0.000	0.001	0.002
AREA PLANTED	-0.0012	0.000	-4.398	0.000	-0.002	-0.001
smoke_estimate	5.2889	0.761	6.946	0.000	3.762	6.816
num_fires	-0.0005	0.002	-0.193	0.847	-0.005	0.004

Omnibus:	14.905	Durbin-Watson:	1.488
Prob(Omnibus):	0.001	Jarque-Bera (JB):	20.124
Skew:	0.943	Prob(JB):	4.27e-05
Kurtosis:	5.184	Cond. No.	2.22e+04

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.22e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 5 - Summary of the Liner regression model

Explanation of p-value:

- **F-statistic p-value:** Indicates the overall significance of the regression model.
- **Interpretation:** A small p-value (<0.05) suggests the model is useful in predicting the dependent variable, signifying at least one significant predictor variable.

Forecasting the future yield

The code aims to perform time series analysis using VARMAX modeling, forecast future values, append forecasts to existing data, and visualize the relationship between 'YIELD' and 'smoke_estimate' variables over time.

1. Data Loading and Setup
2. Preparing Endogenous Variables
    - Select columns related to dependent and independent variables for modeling.
    - Constructs an endogenous variable set ('endog') from the data.
3. VARMAX Model Fitting - Fits a VARMAX model to the endogenous variables with a specified order (adjustable).
4. Forecasting Future Values
5. Appending Forecasted Data
6. Data Visualization

Details for Choosing the model:

- The code utilizes VARMAX modeling from the statsmodels library for multivariate time series forecasting.
- Forecasting involves specifying the number of future periods and generating predictions using the VARMAX model.
- Data visualization is achieved using Matplotlib to plot a combo chart displaying the trends of 'YIELD' and 'smoke_estimate' variables over time.

Findings

There were several interesting findings from the analysis.

1. Correlation analysis

From the results of the Correlation study and the linear regression model, we can say that there exists a relationship between yield and wildfire.

1. A negative correlation between yield and the wildfire impact (smoke estimate and number of wildfires) shows that with an increase in the wildfire which in turn is shown as increase in smoke estimate and/or increase in the number of fires, there is a reduction in yield.
2. The statistical significance of the relationship is shown when we regress yield which is a dependent variable, against the wildfire impact (smoke estimate and number of wildfires), the p value of the

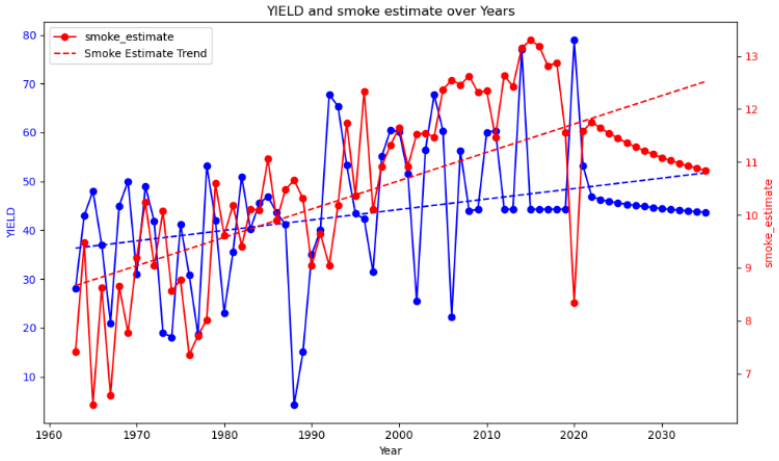
smoke estimate shows that it is significant and we can reject the null hypothesis. (As p-value - 0.00 is lower than the level of significance - 0.05 we can reject the null hypothesis.)

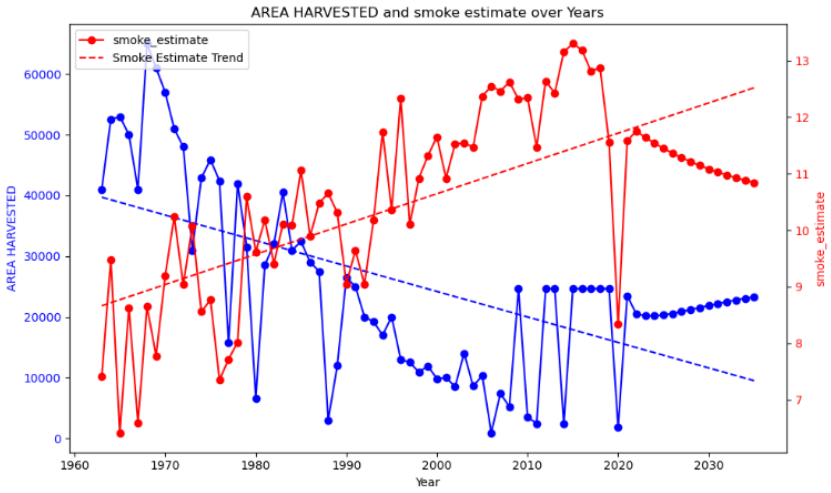
2. Time Series Analysis

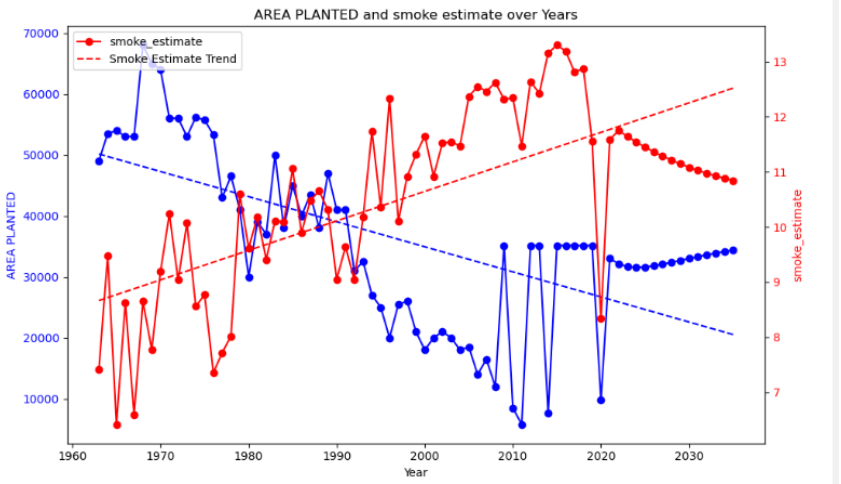
From the time series forecasting I plot the smoke estimate against Yield, Area Harvested and area planted.

Using correlation and linear regression, we proved a relationship and the statistical significance of it. The question that cropped up next: How does wildfire impact the agricultural yield? The time series forecasting has some answers. Although more in-depth analysis is warranted for the question and the approach taken, from the plot we can come up with some possible patterns and deductions

1. The trend for the smoke estimate is going to increase over the next few years - This finding triggers a set of insights. This could mean that the number of wildfires that happen could increase in quantity or it could also mean that the fires that occur as a function of natural be more damaging for the crops and agricultural yield, i.e.; the fires are going to be more intense.







2. While we see an upward projection in the wildfire impact, we see a general negative trend in the yield - There is initially some positive change observed in the yield but the general trend for a majority of time shows that there is a reduction in yield.
3. There is also a negative trend observed in the Area harvested and the area planted. We can hypothesize that wildfire hurts the land available for cultivation.

From these observations, we can infer that the wildfire not only directly impacts the yield, but also indirectly impacts it through reducing the cultivable land which leads to low yield.

Discussion/Implications

Why are your findings important or interesting?

The findings revealing a negative correlation between wildfires and agricultural yield unveil crucial insights into the intricate relationship between environmental factors and crop productivity. This

revelation emphasizes the vulnerability of agricultural lands to natural disasters like wildfires, underscoring the urgency for proactive measures and adaptive strategies. These insights not only highlight the immediate impact on crop yields but also signal broader socioeconomic implications, urging policymakers and stakeholders to adopt holistic approaches that balance environmental resilience with sustainable agricultural practices for long-term stability and food security.

What should the city council, city manager/mayor, and city residents do to address your findings?

City Council, Manager/Mayor, and Residents Action Plan:

- Prevention and Regulation:
    - Enforce wildfire prevention measures and sustainable land use policies in agricultural areas.
- Community Engagement:
    - Educate residents on fire safety, sustainable farming, and early reporting systems.
    - Encourage community groups for fire prevention and emergency response.
- Collaboration and Conservation:
    - Partner with environmental agencies and neighboring communities for coordinated fire management strategies.
    - Support research for resilient farming and conservation efforts.
- Emergency Preparedness:
    - Develop and update emergency response plans, allocate resources for firefighting, and establish evacuation routes.
- Communication and Advocacy:
    - Launch public awareness campaigns on wildfire impact and advocate for regional policies supporting prevention and conservation.
- Early warning Systems:
    - Employ AI and ML algorithms to detect any shift in factors that could lead up to a potential wildfire and warn the farmers early
- Farmer Insurance and Funds
    - Advocate more farmers to insure their lands and provide financial support to the ones in need

How long do they have to make a concrete plan?

The City council can start working on building plans and drafting bills right away which would require a lot of feedback. More enhancements would require in-depth analysis and more data to support the hypothesis. The policy-making and the general awareness-building should be started right away with the first drafts being ready in 6 months. Sophisticated solutions like early warning systems need some specialized labor. This might potentially have a longer turnaround time.

Limitations

Special care and attention need to be given to this section if implementing this project.

- The dataset covers wildfires from 1963 to 2023, but limitations exist in pinpointing exact fire start and end dates
- Availability and reliability of Air Quality Index (AQI) data might vary across different regions

- Incorporating the selected attributes of fires into smoke impact estimation poses challenges due to data constraints
- The EPA's monitoring station coverage might not be exhaustive, affecting AQI estimations for specific cities
- AQI only limited to particulate matter since there was limited information about the gaseous impact
- The missing values for AQI was filled with average values
- AQI API needs credentials and setup before we can request the data
- Runtime is significantly high for the USGS GeoJSON
- Parts 2 and 3 are impelemented with a sole focus in mind and do not consider any additional Impacting factors
- Model Predictions are subject to uncertainty and a certain error margin should be expected
- Yield forecast methodology needs more refinement and statistical proof of result
- Correlation is used to maintain a relationship between variables, however, correlation is not equal to causation. Causation testing is beyond the scope of the project
- The extended analysis is done for the county of Burliegh, of which Bismarck, North Dakota is a part

Data Sources

- **USGS Data -** United States Geological Survey USGS Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) is an open-source US public domain dataset containing the comprehensive data set of fires of polygon and attributes. Link: https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81
- **AQI Data -** The documentation for the API provides definitions of the different call parameter and examples of the various calls that can be made to the API. Some additional information on the Air Quality System can be found in the EPA FAQ on the system. The US Environmental Protection Agency (EPA) Air Quality Service (AQS) API. https://docs.airnowapi.org/ This is a historical API and does not provide real-time air quality data. Page requests were utilized to collect the AQI particle index. This open-source information provided by https://www.airnow.gov/
- **USDA Data -** Most of the information available from this site is within the public domain. Public domain information on the National Agricultural Statistics Service (NASS) Web pages may be freely downloaded and reproduced. However, it is requested that in any subsequent use of this work, USDA-NASS be given appropriate acknowledgment. Link: https://www.nass.usda.gov/datasets/

References

[1] Shmuel, A., & Heifetz, E. (2023). “A Machine-Learning Approach to Predicting Daily Wildfire Expansion Rate.” Fire, 6(8), 319. https://doi.org/10.3390/fire6080319 - This research paper delves into machine learning approaches to predict the potential impact of wildfires on crop yields based on historical data.

[2] Meier, S., Elliott, R. J., & Strobl, E. (2023). “The regional economic impact of wildfires: Evidence from Southern Europe.” *Journal of Environmental Economics and Management*, *118*, 102787.

https://doi.org/10.1016/j.jeem.2023.102787 - This study presents a modeling framework to assess the economic repercussions of wildfires on agricultural production and revenue.

[3] K. C.: “A spatial database of wildfires in the United States, 1992-2011”, Earth Syst. Sci. Data, 6, 1–27, https://doi.org/10.5194/essd-6-1-2014, 2014.. - This paper employs spatial analysis techniques to identify and quantify wildfire-affected agricultural areas.

[4] Xi, Dexen & Taylor, S.W. & Woolford, Douglas & Dean, C.B.. (2017). “Statistical Models of Key Components of Wildfire Risk.” Annual Review of Statistics and Its Application. 6. 1-26. 10.1146/annurev-statistics-031017-100450. - This research focuses on statistical models to estimate the damage caused by wildfires on crop yields, incorporating various environmental and fire-related factors.