Effects of Wildfires on Unemployment

Charlotte Hauke
Student ID: 110431854
University of Colorado Boulder
CSCI 4502 - Data Mining
Fall 2024
(719)-244-5825
chha7787@colorado.edu

Quinn Turner
Student ID: 109733139
University of Colorado Boulder
CSCI 4502 - Data Mining
Fall 2024
(719)-238-1451
qutu4904@colorado.edu

Sierra Reschke
Student ID: 111370326
University of Colorado Boulder
CSCI 5502 - Data Mining
Fall 2024
(303)-408-1151
sire7023@colorado.edu

1. ABSTRACT

In several studies, it has been shown that wildfires have a significant impact on local economies and various socioeconomic factors. The goal of our project was to quantify this relationship and explore the interactions between wildfires and unemployment rates in the United States. We hypothesized that larger fires would have a negative impact on unemployment rates in the counties they occurred in, which is supported by previous research. We began with the analysis of the datasets we used to gain insight into how to build a model and to explore interactions between the data we collected. Furthermore, we built a linear regression model to predict future unemployment rates that uses a combination of economic and wildfire data. After experimenting with the inclusion and exclusion of wildfire data in the creation of a model, we concluded that wildfires have a negligible impact on the prediction of unemployment rates on a local level.

2. INTRODUCTION

Wildfires have become an increasingly frequent and severe hazard in the United States, a trend exacerbated to some extent by climate change [1]. According to the National Interagency Fire Center, in the year 2023 alone, there were over 56,000 fires and over 2.6 million acres burned nationally, which incurred costs of over 3 billion dollars for fire suppression [2]. These fires contribute to several short- and long-term effects on the environment and surrounding communities. The short-term impacts and losses that wildfires incur such as soil degradation, air pollution, property damage, and loss of life [3, 4] have been well documented, but the socioeconomic impacts are less understood.

Wildfires can harm nearby communities and affect employment by displacing populations, inducing energy poverty, damaging critical infrastructure, and detrimenting public health [5]. These effects can destabilize the local economy, cause a loss of jobs, and impact local unemployment. For instance, businesses may close due to infrastructure damage, agriculture may experience decreased productivity, and various tourism sectors may struggle to recover. Despite these impacts, there has been limited research quantifying and analyzing the relationship between wildfires and unemployment.

This project uses data mining techniques to analyze the impacts of wildfires on unemployment in various counties in the United States. By using economic and wildfire-related data, we aim to identify key patterns and build a predictive model to estimate unemployment rates in areas affected by wildfires in the county level.

Some potential challenges in this study include the collection of quality datasets, ensuring data completeness, integrating data sources, and addressing potential regional variability in wildfire impacts. Another key challenge is isolating the effects of wildfires from other concurrent socioeconomic factors, such as economic downturns, pandemics, and seasonal changes.

By addressing the challenges outlined above, this project aims to contribute to the growing understanding of disaster economics and to inform strategies for economic recovery and resilience in wildfire-prone areas. Understanding and predicting unemployment from wildfire data is important for disaster response and crucial for supporting sustainable and equitable economic development in wildfire-vulnerable regions. The primary objectives of this project are to investigate the potential relationship between wildfires and unemployment and to build a predictive model that estimates unemployment on a county-level. By exploring these objectives, we aim to develop a more comprehensive understanding of the societal and economic impacts of wildfires.

3. RELATED WORK

3.1 Economic Impacts of Wildfire (John M Diaz)[6]

This factsheet provides insight on the broader economic impacts of wildfires, beyond the immediate costs of suppression and property damage to include long-term economic consequences. It highlights two key case studies: the Florida wildfires of 1998 and the San Diego County wildfires of 2003.

In Florida, the 1998 wildfires burned around 499,000 acres, resulting in an estimated economic cost of $800 million. While the immediate suppression efforts contributed to the overall cost, the primary financial impact came from timber loss and extensive fire management operations. Similarly, the 2003 San Diego County wildfires affected over 376,000 acres and destroyed 3,241 homes. The total economic impact was estimated at $2.45 billion, with a significant amount of costs tied to rebuilding homes, lost business revenue, and infrastructure damage..

State budgets often supply significant short-term financial aid for fire management and recovery, but these costs often surpass state resources, requiring additional federal support. After the 1998 Florida wildfires, the state received $100 million in federal aid to assist in recovery efforts. Infrastructure damage also represents a major cost, which is typically covered by taxpayers. In San Diego, infrastructure repair expenses amounted to $147.3 million, affecting roads, utilities, and other important public services.

While property damage and fire suppression costs are relatively straightforward to quantify, more complex costs related to the loss of ecosystem services are harder to assess. Ecosystem services, such as clean water, air filtration, and biodiversity, are often disrupted by wildfires, and their long-term value can be difficult to measure. This lack of accurate assessment can lead to an underestimation of the true economic impacts, potentially affecting funding decisions and resource allocation for future wildfire management and recovery efforts. Addressing these hidden costs is essential for a more complete understanding of the financial burden wildfires impose on communities and ecosystems.

3.2 The Cost of Forest Fires: A Socioeconomic Analysis (Zoran Poduška & Snežana Stajić)[7]

This chapter analyzes both the role of socioeconomic factors in driving risk of forest fires as well as the role forest fires play in society, the economy, and the environment.

The duality of forest fires is highlighted, noting the negative impacts such as damage to health and property as well as the positive effects including increased biodiversity and ecosystem regeneration. Mainly in response to the Yellowstone fire of 1998 and other significant events, research on the socioeconomic impacts of fires has seen an increase in interest and awareness. Analyzing scientific research indicates that the majority of this research is concentrated in the U.S., Canada, and Australia.

The economic burden resulting from wildfires is substantial, with global assessments indicating billions in damages, however, it is also important to recognize the positive impacts that fire recovery can have on local economies such as increased job opportunities and market stimulation. This analysis emphasizes a shift in perspective on wildfires, emphasizing the need to consider both positive and negative socioeconomic impacts. This understanding can better inform policy decisions and preparedness.

3.3 Associations between Wildfire Risk and Socioeconomic-Demographic Characteristics Using GIS Technology (Seong Nam Hwang, Kayla Meier)[8]

This research explores the link between extreme wildfire incidents in California over the last 30 years and socioeconomic and demographic characteristics.

Employing geographic information systems for spatial analysis, the study analyzes historical wildfire data in relation to demographics from the US Census Bureau, focusing on factors such as race, ethnicity, education, and income.

Findings from this study reveal that higher educational achievement is linked to a preference for lower-risk areas and areas with higher household income and housing prices are less likely to be wildfire-prone. Communities with lower socioeconomic status appear to struggle more and lack appropriate resources to prepare for and recover from wildfires.

The results of this study highlight the need for targeted disaster planning and understanding how socioeconomic and demographic factors influence wildfire vulnerability.

3.4 Contrasting Our Work with Related Studies

Similar to the Florida and San Diego case studies, this project attempts to address the broader economic consequences of wildfires, specifically their effects on unemployment. While the case studies highlighted the economic burden of wildfires in terms of property damage, fire suppression, and infrastructure costs, this project takes a more targeted approach by analyzing how wildfires influence local labor markets. The Florida and San Diego case studies emphasize a focus on the long term economic effects from wildfires versus the more obvious short term effects from immediate suppression and recovery. Our project also takes into consideration the longer term effects, specifically on unemployment, by building predictive models that estimate unemployment up to 12 months after a major wildfire has occurred.

This project was not designed with an inherent bias whether wildfires would have a positive or negative impact on unemployment. Instead, it aims to explore and uncover whether there is any significant relationship at all between wildfires and unemployment. The models used in this project are capable of identifying both positive and negative relationships within the data. This approach highlights the potential duality of wildfires as mentioned in the chapter by Zoran Poduška and Snežana Stajić who stress the importance of considering both the positive and negative socioeconomic impacts from wildfires.

While the research conducted by Seong Nam Hwang and Kayla Meier utilizes a range of socioeconomic and demographic factors to explore wildfire vulnerability in California, this project focuses specifically on the impact of wildfires on unemployment but for all counties in the United States. By concentrating on unemployment as a key economic indicator, this project seeks to identify patterns and correlations that may be overlooked in studies examining a wider array of demographic factors, providing a more focused understanding of how wildfires disrupt local economies across the entire country.

This project uses data mining techniques to analyze the relationship between wildfires and unemployment, a more quantitative approach compared to the qualitative analysis used in the previous studies. This approach allows for the identification of possible patterns through the use of analysis and predictive models, something not greatly explored in the case studies mentioned above.

4. INITIAL METHODOLOGY

4.1 Data Collection

We began this project by conducting an extensive search to find available datasets that could provide valuable insights for analysis and model building. Our primary focus was on locating historical data related to both wildfires and socioeconomic factors, specifically targeting data that was granular at the county and monthly level across the United States. This granularity was essential for our model building and ensuring the relevance of predictions, as it would allow us to closely examine the temporal and geographic patterns associated with wildfires and their economic consequences.

After a thorough search, we were able to locate a comprehensive wildfire dataset on Kaggle, which perfectly aligned with our needs for both historical data and granularity. The dataset, titled *"1.88 Million US Wildfires"*, was uploaded by Rachael Tatman and sourced from the reporting systems of federal, state, and local fire organizations. This dataset covers a wide array of wildfire events across the United States. With over 51 columns, the dataset includes key variables such as state, fire size, discovery date, agency, cause and latitude and longitude. Additionally, it contains over 1.88 million rows spanning from 1992 to 2017.

In terms of socioeconomic data however, finding a suitable dataset that was both historical data, and met our granular needs for months and county was more of a challenge. Many datasets we encountered only covered the last five years, which did not align with the broader timeframe we desired for this project. Eventually we came across another dataset hosted on kaggle *"US Unemployment Rate by County, 1990-2016"*. This dataset, uploaded by Jay Ravaliya, was sourced by scraping data from the United States Department of Labor's Bureau of Labor Statistics. It provides local area unemployment statistics from 1990 to 2016, broken down by state, county, and month. The dataset consists of 885,548 rows and five columns: year, month, state, county, and unemployment rate.

Interested in how climate might be related to wildfires and unemployment we also made use of a classification map using the Köppen-Geiger classification system that provides a detailed and updated understanding of climate zones across the world. This map was generated by MarkusKotteki, Jurgen Grieser, Christioph Beck, Bruno Rudolfm and Franz Rubel after collecting and analyzing climate data, primarily focusing on temperature and precipitation.

The combination of wildfire data and county-level unemployment statistics from both datasets allows us to begin our initial exploration and analysis.

4.2 Data Cleaning/Preprocessing

Although the datasets from Kaggle were relatively clean and well-organized, some preprocessing and cleaning were necessary to prepare them for analysis, merging, and predictive modeling. We used Python's pandas library to perform this.

For the fire dataset, the initial step was to remove columns deemed irrelevant to our analysis, leaving us with the following variables: Agency, Year, Cause, Discovery_Date, Fire_Size, LATITUDE, LONGITUDE, State, and County_Name. We then filtered the data to retain only the years relevant to our analysis. Upon further exploration, we observed that a large proportion of fire records were associated with fire sizes of less than 1 acre. Given that these small fires were unlikely to have a significant impact on unemployment and could introduce noise into our models, we decided to exclude them.

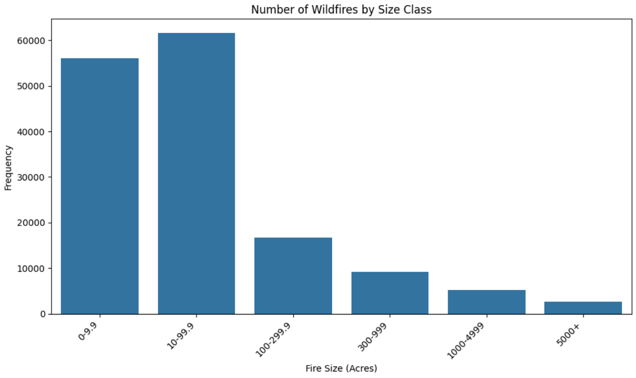
Next, we used geographic data to fill in missing county information based on the provided latitude and longitude coordinates. To do this, we utilized a county shapefile from GitHub, contributed by user sdwfrost, and applied GeoPandas to accurately impute county names for each record. Afterward, we one-hot encoded the categorical variables and aggregated the dataset by month and county. This aggregation involved summing Fire_Size, Cause, and Agency, while calculating the average latitude and longitude for each county, ensuring the data was ready for merging with the climate dataset.

The unemployment dataset required minimal cleaning. We performed basic filtering by year, restructured the columns for consistency, and merged it seamlessly with the fire dataset.

Similarly, the climate dataset required only minor cleaning. We restructured a few columns, applied one-hot encoding to categorical variables, dropped null values, and performed necessary filtering. After these adjustments, we merged the cleaned climate dataset with the fire and unemployment datasets to create a unified dataset for analysis.

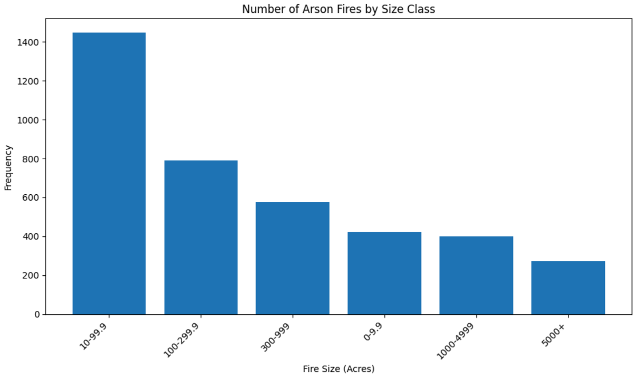
4.3 Data Exploration/Analysis

Our initial data exploration of the Wildfire dataset was focused on examining the frequency and distribution of wildfire sizes. This analysis showed that the majority of wildfires were relatively small.

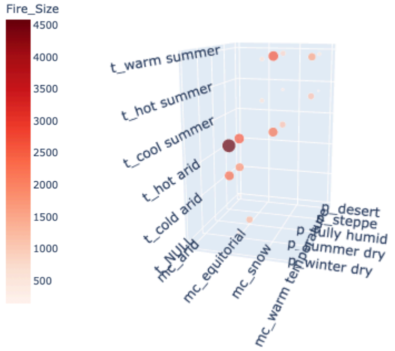


We hypothesized that larger fires would have a more substantial impact on surrounding areas and thus a greater impact on unemployment rates. However, the smaller frequency of large fires indicates a limited number of data points for analyzing their specific impact.

Furthermore, we analyzed the distribution of fire causes for various fire sizes. Our findings indicated that fires smaller than 5000 acres were predominantly caused by burning debris and fires greater than 5000 acres were predominantly caused by lightning. Another insight in the data is that a higher proportion of larger fires were caused by arson, as shown in the figure below. Specifically, for fires that burned 5000 or more acres, 10% of these fires were caused by arson. This is interesting because fires caused by arson are likely to cause more damage to property, which is likely to have a greater impact on socioeconomic factors in the area.



Another factor analyzed was the types of climates that were most susceptible to large wildfires. The visualization below represents the distribution of average wildfire size for each type of climate.



From this figure, it can be observed that regions in the United States with an arid main climate have the highest rate of large fires, and in particular, regions with an arid main climate, a hot-arid temperature, and steppe precipitation have the largest fires on average. This is an important insight because even though these regions do have larger fires on average, the population density of these regions is also lower on average than other regions in the United States[9]. This implies that the largest fires might have a smaller effect on the surrounding areas because of lower population densities, which could give rise to problems in building accurate models given fire size.

4.4 Initial Model Training/Evaluation

After pre-processing of initial fire and economic data was completed, multiple models were implemented on this data to predict the unemployment rates. Basic models were implemented and trained for the following types: Linear Regression, SVM Regression, Stochastic Gradient Descent (SGD) Regression, Nearest Neighbors Regression, and Decision Tree Regression. These models performed mostly satisfactorily and were used as baseline models for our secondary methodology as discussed later. However, it was necessary to further analyze the different variables present within the training data to ensure our models were performing in the expected manner.

A Mutual Information (MI) regression was performed on the independent variables present in the processed fire dataset to determine the degree of dependency between each fire feature and unemployment. The results showed that the current unemployment rate feature had an MI score of 0.946, standing out as having the highest score by far. While this indicates that the current unemployment rate is highly informative about future change in unemployment, the other fire features had MI scores of around only 0.001 to 0.005, suggesting they have very little predictive power and are weakly related to unemployment. Of these other features, fire size had the highest MI score (0.024), and thus had the largest relationship to unemployment. As the goal of this project was to combine multiple features and utilize both fire data and unemployment to form an educated prediction about unemployment, it was determined that this fire dataset was not the most informative nor best choice of independent variables.

5. UPDATED METHODOLOGY

After reviewing the results of our initial models, we decided to enhance our analysis by incorporating additional datasets that we believed were more directly relevant to predicting unemployment. This would allow us to build more comprehensive models and better assess the impact of fire size on model performance. We specifically chose to narrow the scope of our analysis to focus on larger fires, with the hope that this approach would help identify any meaningful relationships between fire size and unemployment trends.

Additionally, we opted to conduct more focused, smaller-scale studies to gain deeper insights into the data and improve our understanding of the results generated by our models. This involved generating visualizations for multiple counties where significant fires had occurred. By plotting unemployment rates before and after these fires, we could compare these trends to the national unemployment rate during the same period. This comparison would help us identify any potential correlations between wildfires and local unemployment, providing a clearer understanding of the fire's impact on the county's economy.

5.1 Data Collection

Given the difficulty in finding datasets that were both granular at the county level and on a monthly basis, we decided to use datasets that provided monthly data for the entire United States and then replicate national values across all counties.

For our unemployment prediction models, we used several datasets that provided key economic indicators. From the Bureau of Labor Statistics (BLS), we incorporated the *Consumer Price Index (CPI)*, *Number of Employees on Nonfarm Payrolls*, and *Work Stoppages* data. These datasets offer insights into inflation, overall employment trends, and labor disruptions, all of which can impact unemployment.

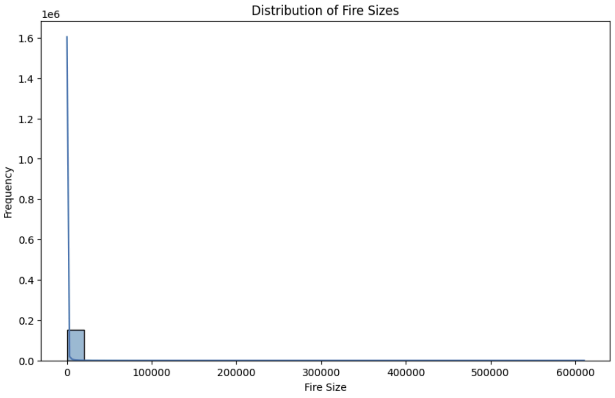
From Kaggle we used *Growth Rate of USA from GDP*, uploaded by Salman Ibne Eunus. This dataset, sourced from open government data, reflects national economic growth, which can influence unemployment trends. We also included the *Federal Reserve Interest Rates, 1954-Present* dataset uploaded by Abigail Larion, which tracks interest rate changes that can affect borrowing costs, inflation, and employment levels.

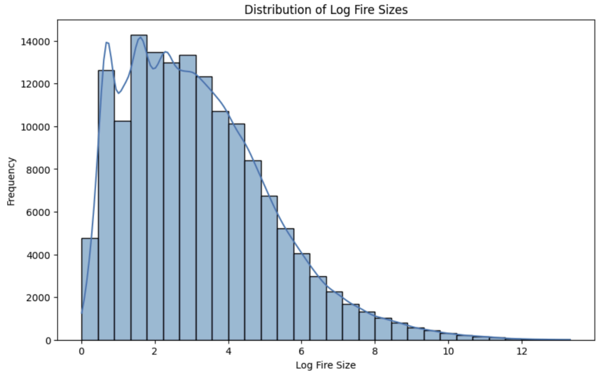
These datasets were chosen because they provide a comprehensive view of national economic conditions, which are crucial for predicting unemployment.

5.2 Data Cleaning/Preprocessing

Much of the new data we incorporated required only basic cleaning, such as filtering and reformatting.

As we refined our methodology, we decided to introduce a new feature in our fire dataset: a fire size bin column. This allowed us to categorize the fires based on their size, enabling us to conduct more targeted analysis and build models for different fire sizes. To determine appropriate fire size bins, we first generated a histogram to explore the distribution of fire sizes. However, the histogram did not provide a clear picture due to the skew in the distribution. To address this, we applied a logarithmic transformation to the fire size variable, which helped normalize the distribution and made it easier to interpret.

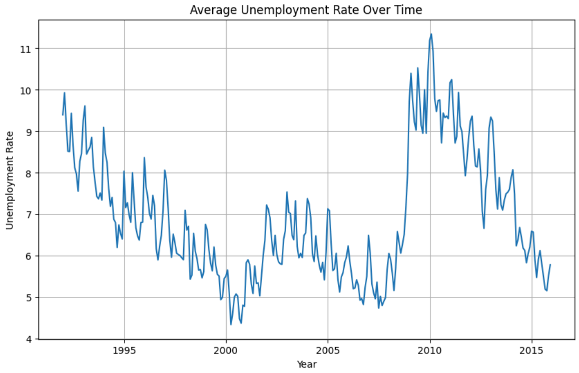




Using this transformed distribution, we then established five distinct bins, with Bin 1 representing the smallest fires and Bin 5 capturing the largest fires. This binning process allowed us to focus on analyzing the largest fires (Bin 5) in our models, in hopes of identifying significant patterns or trends related to fire size and its impact on unemployment.

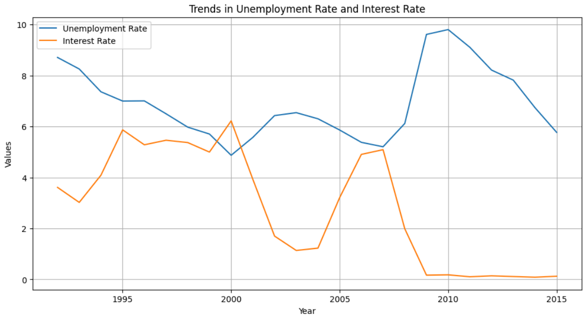
5.3 Data Exploration/Analysis

Building off the analysis performed from our first methodology, we began analysis on the economic datasets and their relation to wildfires and unemployment. We began the analysis by looking at unemployment across time, as shown in the figure below.



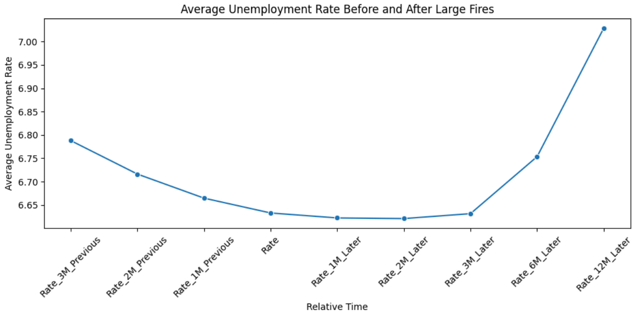
Because of the recession, unemployment rose to a peak during 2009 and 2010. Being able to represent and predict changes like this in a model presents a challenge because dramatic changes in the unemployment rate are hard to predict from other current economic factors. Because of this, we decided to add previous months' unemployment rate as an input to the model to increase the ability of the model to accurately predict trends in the data.

Another feature that we analyzed was the relationship between interest rates and unemployment. We found that they have a weak negative correlation of -0.27. The figure below visualizes this relationship.

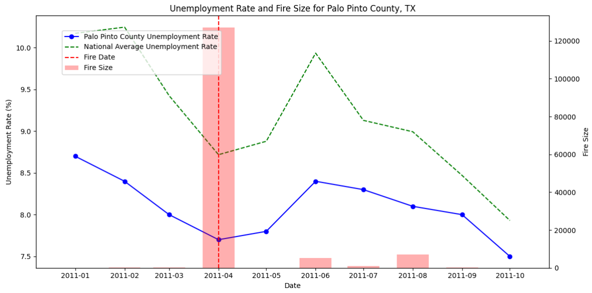


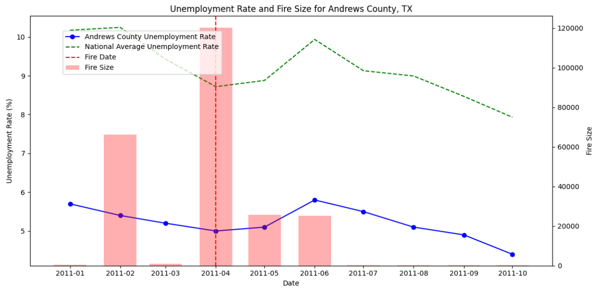
From the visualization, the weak negative correlation is apparent, which is reasonable because the Federal Reserve accounts for unemployment rates to determine adjustments to interest rates, which makes interest rates a valuable feature for predicting the unemployment rate.

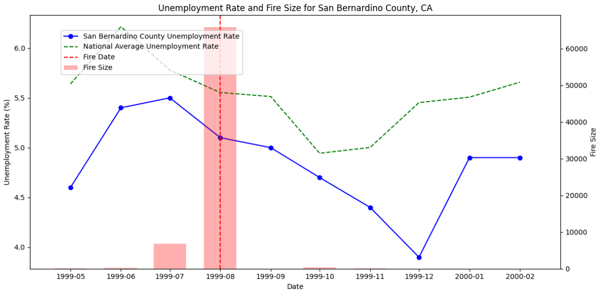
After binning the data, we analyzed the average trend of unemployment rates after the largest fires.

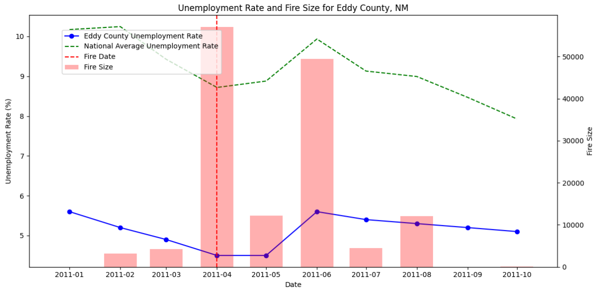


From the figure above, it appears that unemployment rates slightly decrease and then rise again. To explore this further, we performed specific case studies on counties that had very large fires. The following figures show the unemployment rate trends of specific counties compared with the average national unemployment rate.









The figures above show trends in unemployment for Palo Pinto County, Texas, Andrews County, Texas, San Bernardino County, California, and Eddy County, New Mexico. These demonstrate that even after a large fire, the unemployment rate in that county will continue to follow the general national trend despite the effects of large fires.

6. MODELING AND EVALUATION

As with the initial methodology, multiple models were created to analyze and predict the effects of wildfires on unemployment. The largest fires combined with economic data were used to train these models and predict unemployment rates. Based on initial performance, we experimented with each model's hyperparameters to ultimately compare results and select the optimal model.

6.1 Baseline Models

Similarly to the initial methodology, in order to identify the optimal model for predicting the target variable, we built and evaluated several regression models, each with distinct strengths for capturing different data characteristics. The models included linear regression, which assumes a linear relationship between predictors and the target and serves as a baseline; support vector machine (SVM) regression, which maximizes the margin within a hyperplane and can capture non-linear relationships; stochastic gradient descent (SGD) regression, an iterative optimization method efficient for large-scale datasets; nearest neighbors regression, which predicts outcomes based on the average of the k closest data points without assuming a specific data distribution; and decision tree regression, a tree-based approach that segments the dataset into regions with similar outcomes by recursively splitting features. These models were trained and evaluated using *Fire_Size* and economic data features, with the target variable (unemployment) Rate predicted on a testing set.

6.2 Model Evaluation

Each model's performance was assessed using Mean Squared Error (MSE) and R-squared (R^2) metrics to provide insights into the most suitable model for our given data.

Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values. This metric gives greater weight to larger errors, making it particularly helpful for identifying significant differences between predicted values and actual outcomes.

We also considered the R-squared (R^2) metric, which indicates how well the model describes the variability of the socioeconomic outcomes. A higher R^2 value suggests that the model captures a larger portion of the underlying patterns in the data, indicating that the model is more accurate in its predictions.

6.3 Feature Exploration

Given the baseline models previously created, feature exploration was performed with the goal of identifying the most impactful predictors for modeling and forecasting rates under various conditions. Initially, we focused on models trained without any future rate information, ensuring that predictions rely solely on available contemporaneous fire and economic data, such as *Fire_Size, CPI, Interest_Rate, NumWorkStopages, NumWorkersInStoppages,* and *WorkStopopagesTotalDays*. We then expanded the feature set by including the current rate (*Rate*) as a predictor to enhance our ability to train the model on given data and better predict future rates.

To evaluate temporal dependencies, we tested models trained on multiple future months' data without incorporating previous months as features. In contrast, another set of experiments included the rates from the prior three months, which allowed us to capture short-term trends and improve predictions for subsequent months. To test whether a longer historical window could improve predictive accuracy, we expanded the temporal range to include data from the previous 12 months. The rationale for this adjustment was that longer-term trends might capture more nuanced or delayed effects that the 3-month window might miss. While this increased the model's complexity, it also provided the opportunity to leverage a richer set of temporal dynamics.

Additionally, we assessed the impact of excluding *Fire_Size* from the feature set, which revealed that the models performed slightly better without this variable. However, since the focus of our project is to understand the impact of wildfires on unemployment rates, *Fire_Size* must remain a key feature in our models. The decrease in model performance when including fire size is both expected and acceptable, as this variable introduces external, less predictable factors compared to more stable economic data. Economic trends tend to follow consistent patterns, making them easier for models to predict, whereas wildfires represent an external shock that disrupts these trends. Including fire size ensures that our analysis aligns with the project's goal of examining the influence of these unpredictable events on unemployment, even if it adds complexity to the modeling process.

Each configuration was tested across multiple regression models, including Linear Regression, SVM Regression, Stochastic

Gradient Descent (SGD) Regression, Nearest Neighbors Regression, and Decision Tree Regression. This exploration allowed us to evaluate the relative contribution of each feature and feature set combination to the model's performance, providing critical insights into the most effective predictors for rate forecasting under different assumptions and constraints.

6.4 Model Improvements

We selected a linear regression model as the baseline for this analysis because it consistently demonstrated the highest R^2 value and the lowest Mean Squared Error among all tested models during initial feature explorations. This performance indicated that Linear Regression was well-suited to capturing the relationships in our dataset with relatively low complexity. However, recognizing the importance of refining the model further, we implemented several improvement strategies aimed at enhancing its performance and robustness.

To evaluate and improve the model's generalizability, we applied k-fold cross-validation with values of $k = 5$, $k = 10$, and $k = 100$. Cross-validation splits the data into k subsets (folds), training the model on k - 1 folds and validating it on the remaining fold, iteratively. This process reduces the likelihood of overfitting by ensuring the model is tested across diverse subsets of the data. We chose different values for k to examine their impact on the model's performance and identify an optimal balance between bias and variance. $k = 5$, and $k = 10$ are common choices in cross-validation that offer a good trade-off between computational efficiency and model robustness. We also implemented $k = 100$ fold cross validation to explore the effect of a large k value, providing a more granular view of the model's performance but at the cost of increased computational demand.

Feature engineering was another implemented method of model improvement. This process involved creating new variables or transforming existing ones to better represent the underlying relationships in the data. By exploring interactions between features, applying transformations, and normalizing variables where necessary, we aimed to enhance the model's ability to capture complex patterns. Specific features added included *Fire_Size_Rate* (captures interaction between *Fire_Size* and *Rate* to model multiplicative effects; *Fire_Size * Rate)*, *CPI_InterestRate* (captures interaction between *CPI* and *Interest_Rate* for potential economic insights; *CPI * Interest_Rate*), and *Unemployment_MoM_Change* (month-over-month change in *Rate* to capture temporal dynamics). By performing this feature engineering, related predictors were combined into composite indices to reduce noise, variables were scaled to ensure uniform treatment during regression, and lagged variables were created to account for temporal effects in the dataset.

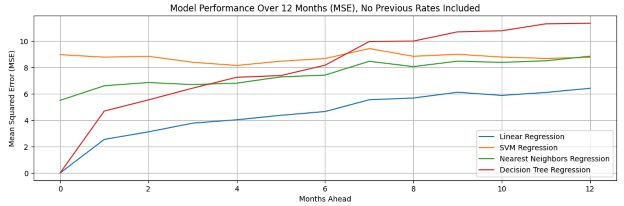
6.5 Model Results

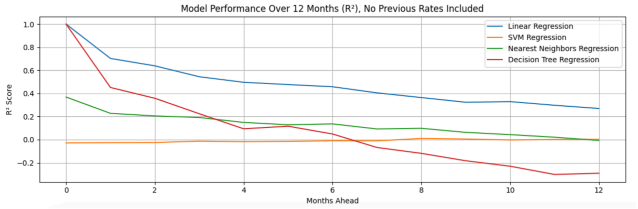
6.5.1 Baseline Model

The performance of the baseline models' performance (multiple models, no future rate, predicting current month's rate) was evaluated using the Mean Squared Error (MSE) metric. Linear Regression achieved the lowest MSE of 9.06, indicating it fits the data more effectively than the other models. The SVM Regression and Nearest Neighbors Regression models followed with MSEs of 10.06 and 9.39, respectively, suggesting comparable but slightly less accurate performance. Decision Tree Regression had the highest MSE at 15.03, indicating potential overfitting to the training data. The Stochastic Gradient Regression model exhibited a significantly higher MSE of 2.53e36, likely due to instability or divergence during training. This suggested that either further tuning of this model's hyperparameters is necessary for this dataset or that this model may not be suitable for our given data. Overall, these initial results identified Linear Regression as the most promising model for further improvement and exploration. However, further exploration and evaluation was needed.

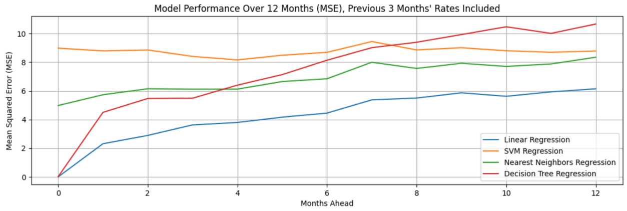
6.5.2 Feature Exploration Models

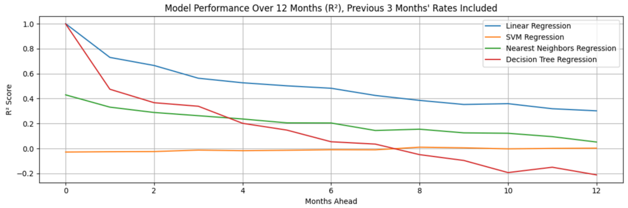
We proceeded to run each model and examine their predictive abilities for future rates (0-12 months in the future). No past month rate data was included. It can be seen in the graphs below that a Linear Regression model continues to have the lowest MSE and high R^2 value for most of the predicted months. It is important to note that as the months progress, it is expected that the MSE will increase and the R^2 values will decrease, potentially to less ideal levels. This trend occurs because predictions further into the future inherently become more challenging due to the increased uncertainty and the diminishing influence of immediate past trends on future outcomes.



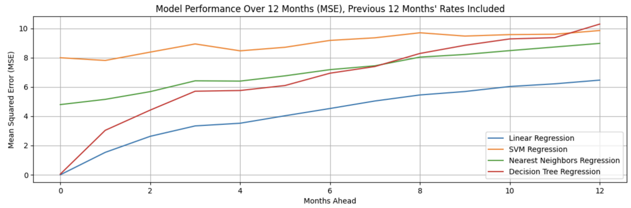


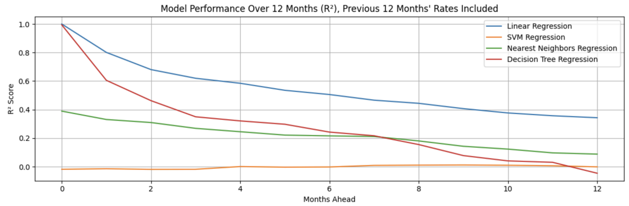
The next step for model exploration was to include the previous 3 months' rates in the features for each model. These columns were added, and the models were trained and evaluated. The results are displayed in the graphs below. The same trend as above continues to hold true, with the Linear Regression model exhibiting the lowest MSE and highest R^2 for each of the predicted months.



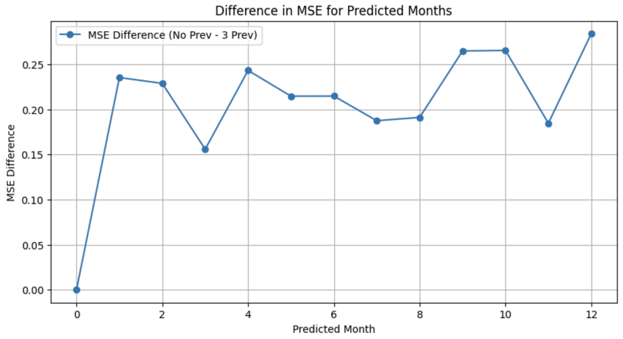


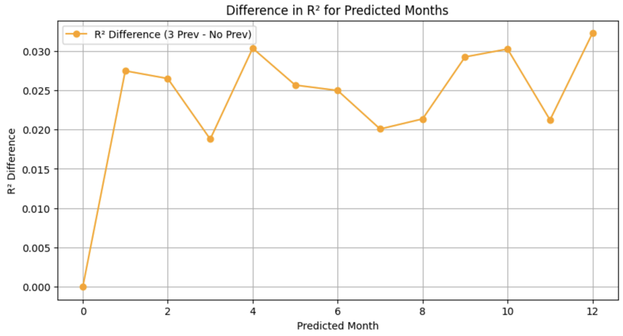
We were also curious to see if including the previous 12 months' rate data resulted in more accurate predictions compared to only including the previous 3 months' data. The results from this comparison are displayed below.

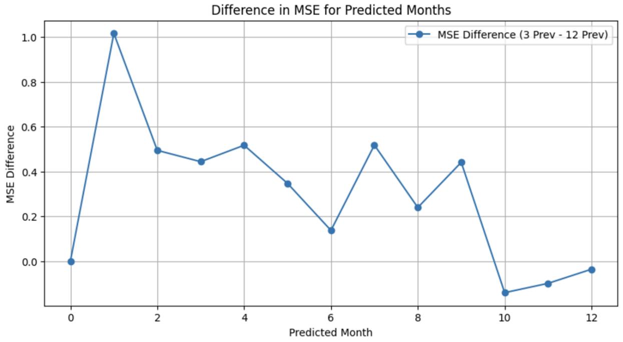


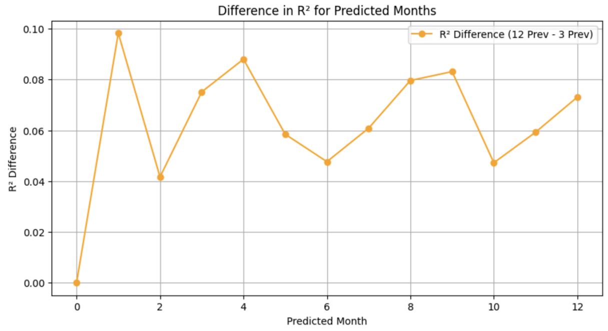


To summarize the above exploration, the two sets of plots below highlight the difference of the above three cases: the first set compares model performance when not including any previous data versus including the previous 3 months' rate data while the second plot compares including 3 months' rate data versus including 12 months' rate data. The overall goal was to decrease MSE values and increase R^2 values with the improved model. For the majority of the predicted months, this desired trend was exhibited. Note that the MSE difference was calculated by subtracting the MSE value of the model excluding previous data from the model utilizing increased previous data while the R^2 value was calculated using the opposite order of subtraction (i.e. 3 previous months - no previous months).



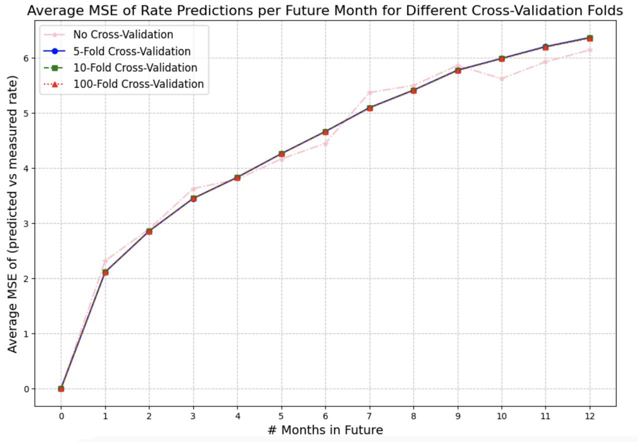


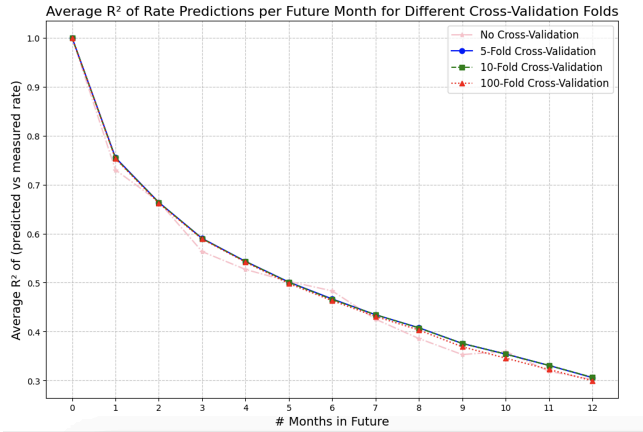




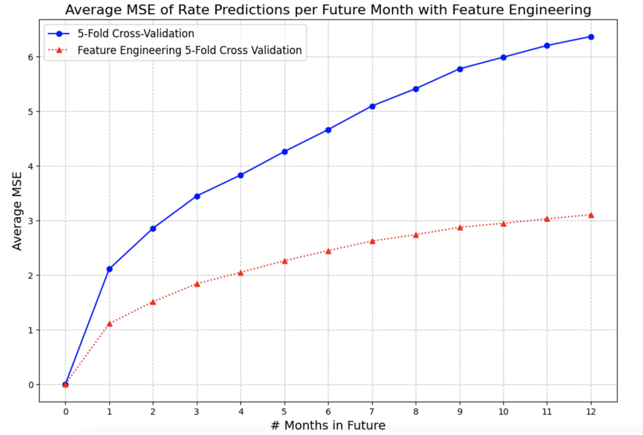
6.5.3 Improved Linear Regression Models

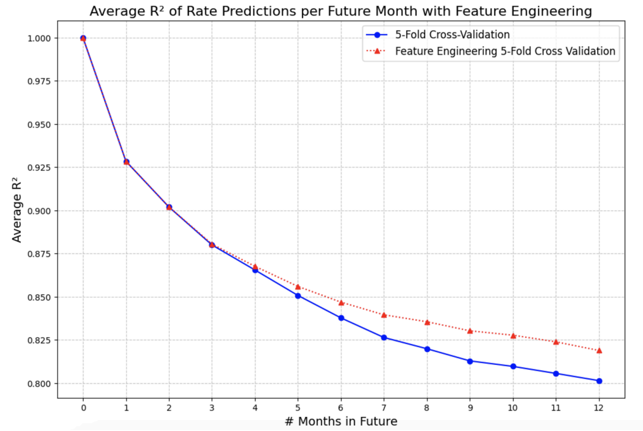
For values of k = 5, k = 10, and k = 100, k-fold cross-validation was performed on the updated Linear Regression model. It was found that each value of k resulted in almost identical values for MSE and R^2. However, implementing k-fold cross-validation led to and improved results compared to when no cross-validation was performed. Thus, it was determined that our current Linear Regression model would be updated to perform 5-fold cross-validation to improve prediction accuracy without detriment to runtime complexity.





The final improvement to the Linear Regression model explored was implementing feature engineering and training the model on additional added data. It can be seen in the graphs below that the inclusion of these features resulted in lower MSE and higher R^2 values for each predicted month.





This refined model represents our final iteration, demonstrating the highest predictive accuracy and most reliable prediction performance across our dataset.

6.5 Interpretation

Throughout the model creation and evaluation process we identified Linear Regression as the most suitable model for predicting both current and future unemployment rates, particularly when combined with past data (3 or 12 months) and improvements like cross-validation and feature engineering. The final iteration of the Linear Regression model, after including feature engineering and 5-fold cross-validation, showed the most promising results with lower MSE and higher R² values.

The improved predictions after incorporating additional economic signifier features demonstrate how these variables are correlated with unemployment, as evidenced by the improved performance after they were included in the training process. Additionally, including the rate data from previous months further improved the model's performance, indicating that historical trends play a significant role in predicting future patterns.

Since *Fire_Size* did not have a significant impact on our model's performance, we can conclude that our model was unable to detect a strong relationship between *Fire_Size* and unemployment. However, this does not imply that no such relationship exists; rather, it suggests that our model was not able to identify one within the scope of the data and features used.

7. DISCUSSION

Throughout this project we learned a lot, particularly the role that data availability plays in successfully addressing research questions. One of the key lessons we learned was that without access to high-quality, relevant data, a project may struggle to progress or even come to a standstill. Securing suitable datasets that meet the specific needs for a given study is a challenge that demands careful attention to both granularity and coverage. We found that datasets often fall short in terms of time range, geographic specificity, or the level of detail required for meaningful analysis. As a result, a significant amount of time and effort was spent identifying, selecting and combining datasets from multiple sources to meet the objectives of this project.

This experience reinforced the importance of thorough data exploration and cleaning. Even when we managed to find datasets that seemed to align with our goals, we encountered inconsistencies and gaps that needed to be addressed before meaningful analysis could be conducted. We quickly realized that the process of locating, evaluating, and preprocessing data is not a mere initial step, but rather an ongoing, iterative process that can significantly impact the overall success of a project.

One of the major factors that contributed to the success of this project was the effective allocation and organization of tasks. We were able to delegate responsibilities related to data cleaning, preprocessing, and analysis in a way that remained structured and efficient, avoiding confusion or disorganization throughout this process. This was achieved through consistent and clear communication about each team member's tasks and the progress they were making.

This level of organization proved to be especially beneficial when it came time to merge the various datasets that different team members had worked on. The clarity in task delegation allowed for a smooth and effective integration of each person's work, ensuring that no important details were overlooked. We also held regular meetings to discuss our progress, brainstorm solutions to any challenges, and troubleshoot any issues that arose. These frequent check-ins allowed us to address problems as they came up and ensured that everyone remained on the same page. It also helped to create a collaborative environment where we could all contribute ideas and feedback.

While wildfires may not have a direct impact on unemployment rates at a national or state level, there could be indirect effects that become evident at more localized levels or within specific industries. Future research could focus on communities or industries that are particularly reliant on natural resources, such as forestry or agriculture, to better understand how these sectors are impacted by wildfires and how this influence correlates with job loss, displacement, or changes in employment patterns. Additionally, we would consider expanding the timeframe for analyzing the effects of wildfires, and examining longer-term impacts to capture potential delays in their influence on employment and economic recovery.

8. CONCLUSION

At the beginning of our project, we hypothesized that wildfires would have a negative impact on unemployment rates at the county level in the United States. Through the use of economic factors like the number of employees, consumer price index, the number of work stoppages, interest rates, and previous months' unemployment rates, we constructed a linear regression model that predicted future unemployment rates in counties with large fires. After analysis of our results, we found that wildfires have little impact on our model's ability to predict future unemployment rates. Even though our results suggest that wildfires have a minimal impact on unemployment rates, many factors at play might not be captured by our model, and the effects of wildfires might be masked. However, these results don't rule out potential impacts on certain industries or locales, or across larger time periods, which would be interesting to explore with future research.

9. APPENDIX

9.1 Honor Code

I have read and am familiar with the provisions of the Honor Code and agree to comply with those provisions while a student at the University of Colorado at Boulder; and further agree that the submission of any academic work shall constitute a representation on my part both that such work has been done, and its submission is being made, in compliance with all applicable provisions of the Code. Furthermore, my responsibility includes taking action when I have witnessed or am aware of another's act of academic dishonesty.

Signatures:

Quinn Turner

Charlotte Hauke

Sierra Reschke

9.2 Individual Contributions

Quinn Turner: I was responsible for getting, cleaning, and preprocessing the climate dataset in the initial methodology. I also analyzed the climate dataset and its relation to the fire data. I was also responsible for one-third of the initial model creation and testing. In the updated methodology, I cleaned and analyzed employee payrolls and consumer price index data. I was also responsible for the binning of the data and looking at specific counties to explore as case studies to look for trends in the data. Along with this, I created and analyzed the data visualizations. Because of this, I also wrote the data analysis sections of the report. In the report, I also wrote the abstract, introduction, and conclusion.

Charlotte Hauke: Together with Sierra, I was responsible for the cleaning and preprocessing of our raw fire dataset. I also conducted the initial analysis on the cleaned data. In addition, I was responsible for one-third of the initial model development and handled the data collection for our updated methodology, along with a portion of the cleaning and preprocessing tasks. My responsibilities also included writing the related work section, a portion of the initial and updated methodology sections, and discussion section of our report.

Sierra Reschke: I was responsible for creating, coding, developing, evaluating, improving, and running all models and modeling tasks for the updated methodology. I also was responsible for a third of the models created, tested, and updated during our initial methodology. Likewise, I wrote the corresponding sections in the report for these sections. Additionally, I worked with Charlotte to perform all importing, pre-processing, cleaning, and analysis of the initial fire dataset, as well as writing the proposed work and evaluation sections in our initial report which I completed for our final report.

10. References

[1] Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch, J. K., & Lettenmaier, D. P. (2019). Observed impacts of anthropogenic climate change on wildfire in California. Earth's Future, 7, 892–910. https://doi.org/10.1029/2019EF001210

[2] Wildfires and acres (2023) Wildfires and Acres | National Interagency Fire Center. Available at: https://www.nifc.gov/fire-information/statistics/wildfires (Accessed: 24 September 2024).

[3] Memoli V, Panico SC, Santorufo L, Barile R, Di Natale G, Di Nunzio A, Toscanesi M, Trifuoggi M, De Marco A, Maisto G. Do Wildfires Cause Changes in Soil Quality in the Short Term? Int J Environ Res Public Health. 2020 Jul 24;17(15):5343. doi: 10.3390/ijerph17155343. PMID: 32722226; PMCID: PMC7432673.

[4] Wildfires (2024) World Health Organization. Available at: https://www.who.int/health-topics/wildfires (Accessed: 24 September 2024).

[5] Chandra Prakash Kala. Environmental and socioeconomic impacts of forest fires: A call for multilateral cooperation and management interventions, Natural Hazards Research, Volume 3, Issue 2. 2023, Pages 286-294. ISSN 2666-5921. https://doi.org/10.1016/j.nhres.2023.04.003.(https://www.sciencedirect.com/science/article/pii/S266659212300032X)

[6] Diaz, J.M. (2012-7) *Economic impacts of wildfire*. Available at: https://fireadaptednetwork.org/wp-content/uploads/2014/03/economic_costs_of_wildfires.pdf (Accessed: 25 September 2024).

[7] Poduska, Z. and Stajic, S. (1970) *The cost of forest fires: A socioeconomic analysis*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-031-50446-4_10 (Accessed: 25 September 2024).

[8] Hwang, S.N. and Meier, K. (2022) *Associations between wildfire risk and socio-economic-demographic characteristics using GIS technology*, *SCIRP*. Available at: https://www.scirp.org/journal/paperinformation?paperid=120257 (Accessed: 25 September 2024).

[9] Ecosystem-Based Approach to Combat Drought and Desertification and Their Relation with Rural Development - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-density-of-human-populations-living-in-arid-areas-in-the-world-persons-km-2-GRID_fig3_330040250 [accessed 8 Dec 2024]

[10] Tatman, R. (2018). 1.88 Million US Wildfires. Kaggle. https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires

[11] Ravaliya, J. (2017). US Unemployment Rate by County, 1990-2016. Kaggle. https://www.kaggle.com/datasets/jayrav13/unemployment-by-county-us

[12] U.S. Bureau of Labor Statistics. (2024) Consumer Price Index (CPI) data. U.S. Bureau of Labor Statistics. https://www.bls.gov/cpi/tables/home

[13] U.S. Bureau of Labor Statistics. (2024) Number of Employees on Non-Farm Payrolls. U.S. Bureau of Labor Statistics. https://www.bls.gov/ces/tables

[14] U.S. Bureau of Labor Statistics. (2024) Work Stoppages. U.S. Bureau of Labor Statistics. https://www.bls.gov/wsp

[15] Enus, (2020) S. US GDP Growth Rate. Kaggle. https://www.kaggle.com/datasets/salmaneunus/us-gdp-growth-rate

[16] Federal Reserve. (2017). Interest Rates. Kaggle. https://www.kaggle.com/datasets/federalreserve/interest-rates

[17] Kotteck, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. 2006. World Map of the Köppen-Geiger Climate Classification Updated. *Meteorologische Zeitschrift*, 15(3), 259-263. DOI: 10.1127/0941-2948/2006/0130.

[18] World Bank. 2019. World Maps of the Köppen-Geiger Climate Classification. The World Bank. Retrieved December 8, 2024 from https://datacatalog.worldbank.org/search/dataset/0042325/World-Maps-of-the-K-ppen-Geiger-Climate-Classification

[19] Frost, S. D. W. 2017. Köppen-Geiger Climate Classification Code. GitHub Gist. Retrieved December 8, 2024 from https://gist.github.com/sdwfrost/d1c73f91dd9d175998ed166eb216994a