Wildfires in National Parks are an Epidemic

Marquette University

COSC 3570

Kayley Reith

12/5/2023

Literature Review

Wildfires are an epidemic in the United States National Park System. National parks span across 84 million acres of US soil (Geico Living n.d) and this calendar year have had over 44,207 wildfires damaging more than 2 million acres; highlighting the urgency for enhanced resources and wildfire management (National Interagency Fire Center 2023). There are multiple implications of wildfires ranging from environmental factors accounting for 20% of global carbon emissions (CarbonCredits 2023). In addition the federal budget uses over $2.5 billion per fiscal year for wildfire suppression (Congressional Budget Office 2022). These substantial statistics emphasize the large-scale issue of wildfires and underscore the imperative for action in terms of both prevention and detection. The objective of this project is to revolutionize wildfire detection offering the opportunity to prepare for when natural disaster strikes and to control wildfires efficiently reducing environmental damage, fiscal spending, and wildfire occurrence.

Data Collection and Management

As described above national parks are widespread and partitioning the parks based on geographic regions and coordination centers was necessary to further explore wildfires in more depth (Figure 1). Performing exploratory and static visualization analysis highlighted outliers and influential points within the 10 regions (See Appendix A). A comparative analysis between the regions' disparities in fire causation types, human made fires and lightning made fires, showcased how the Southern Area region (blue line) possessed the largest distribution for wildfires in 2022 (Figure 2). This region spanned across 14 different states and after further investigation of wildfire occurrence and severity it was uncovered that Texas was the leading contender for wildfires with over 12,000 (Figure 3).

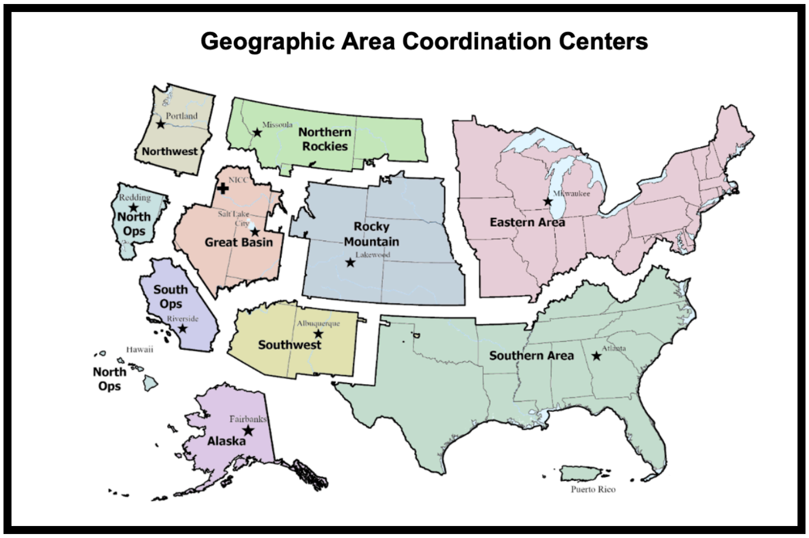


Figure 1: Geographic 10 Region Partition

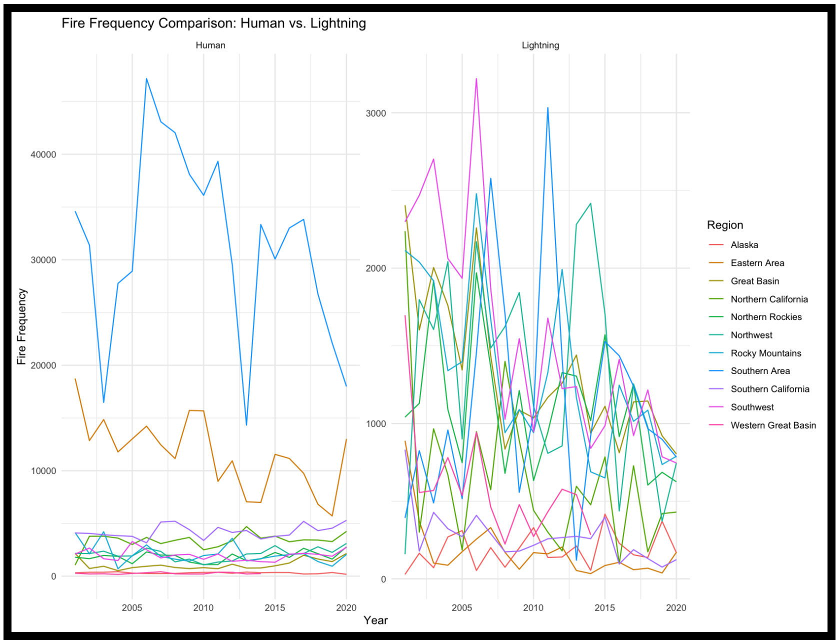


Figure 2: Static Visualization - Fire Frequency Comparison by Region

In this study, we encountered limitations related to the availability and privacy constraints of national park wildfire data. Due to the sensitive nature of the data and privacy concerns, direct access to comprehensive national park wildfire datasets was not feasible. As a result, we explored alternative routes for obtaining relevant information. With an opted focus on uncorrelated and independent variables that were publicly available, which could provide insights into the broader context of wildfire severity and occurrences. In addition to providing a more generalized overview with a region and statewide breakdown of wildfires and independent variables.

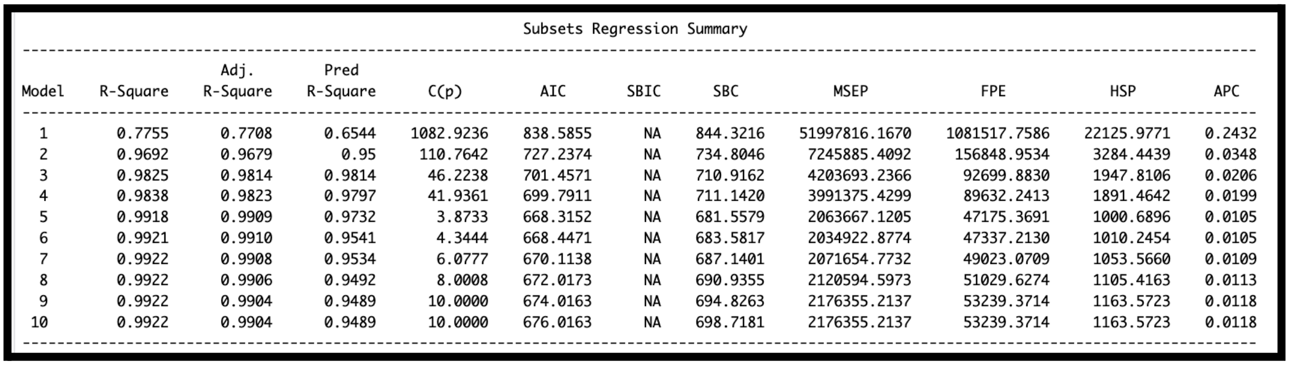
Generalized historical wildfire data broken down by region, ranging from the past 10 years were publicly available through the National Interagency Fire Center, and consisted of differing variables of fire causation type (Human vs Lightning), acres damaged, and the fire frequency for each region. This data was in tabular format and mixed with both nominal and numerical data and underwent missingness accounting for less than 2% of the observed data. As a standard practice, these missing values were omitted from the analysis to maintain the integrity of the dataset. In addition to the historical wildfire information, state data was mutated and merged from several other sources by web scraping. This resulted in the data collection of three independent variables from three meticulously distinct databases. The uncorrelated variables were; unemployment rates obtained from the U.S. Bureau of Labor Statistics , chemical exportation data from Statista, and obesity statistics from Statista, ensuring a robust and diverse representation of the nation's 50 states. The objective was to examine significant relationships between wildfire occurrence in national parks and the key random variables.

Candidate Techniques and Methodology

An array of techniques and methodologies were implemented and developed reflecting the nature of the data and objectives of determining relationships between wildfire occurrences. The selection of appropriate techniques was paramount in being able to extract meaningful insights and draw robust conclusions. Which is why Linear Regression, Logistic Regression, and CART Analysis were selected for their distinctive strengths and capabilities. The power of quantifying the impact of each predictor on the response of wildfires, and effectiveness in predicting categorical outcomes, creating well-suited models in determining which location a wildfire is most likely to occur based on predetermined conditions were impactful.

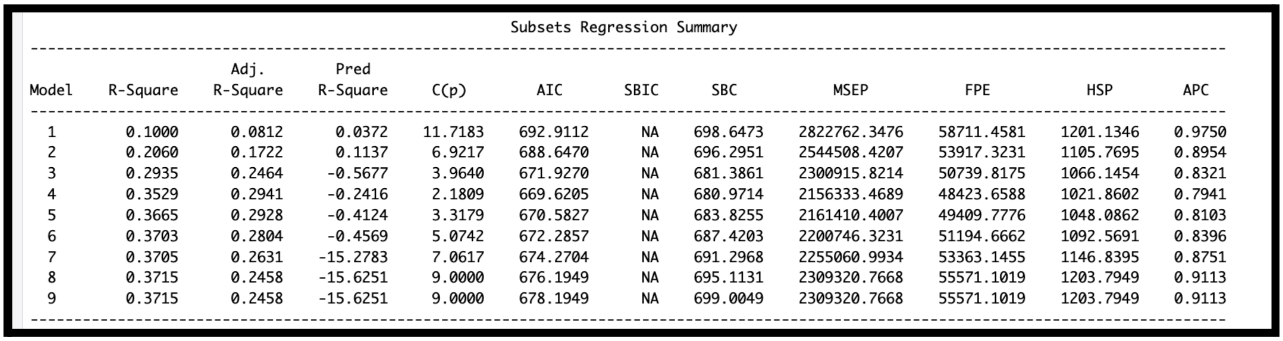
Variable selection was a critical step in developing linear regression models and equations. The technique followed was the Best Subset (All Possible) Selection method (Dr. Yu 2023). Where two full models were created with split focuses on response variables; one was Human Fires (Figure 3) and the other was Lightning Fires (Figure4). The subset regression summary reveals for both full models which model is the strongest candidate. In Figure 3, Model 5 is the strongest candidate for Human Fires as it has a high Adjusted R-Square (0.9909) and relatively low Mallow's

CP (3.8733) compared to other models. This model consists of predictors: Human Acres, Fires Total, Lightning Fires, Lightning Acres, and Acres Total. While Figure 4 displays, Model 5 to be the best balance of these criteria, with a high adjusted R-squared (0.3665) and lower MSEP (Mean Square Error of Prediction) and Mallows CP value compared to other models. Indicating the best suited predictors for predicting Lightning Fires are Region, Fires Total, Acres Total, Chemical Exports(Millions), and Percentage of Obese People. In addition, exploring pairwise distribution and visualization highlights variable correlation and relationships (See Appendix C).



```
                                                  Subsets Regression Summary
--------------------------------------------------------------------------------------------------------------------------------
                   Adj.        Pred
Model  R-Square  R-Square    R-Square      C(p)        AIC       SBIC     SBC          MSEP           FPE           HSP       APC
--------------------------------------------------------------------------------------------------------------------------------
  1     0.7755    0.7708      0.6544   1082.9236   838.5855     NA   844.3216   51997816.1670  1081517.7586  22125.9771   0.2432
  2     0.9692    0.9679        0.95    110.7642   727.2374     NA   734.8046    7245885.4092   156848.9534   3284.4439   0.0348
  3     0.9825    0.9814      0.9814     46.2238   701.4571     NA   710.9162    4203693.2366    92699.8830   1947.8106   0.0206
  4     0.9838    0.9823      0.9797     41.9361   699.7911     NA   711.1420    3991375.4299    89632.2413   1891.4642   0.0199
  5     0.9918    0.9909      0.9732      3.8733   668.3152     NA   681.5579    2063667.1205    47175.3691   1000.6896   0.0105
  6     0.9921    0.9910      0.9541      4.3444   668.4471     NA   683.5817    2034922.8774    47337.2130   1010.2454   0.0105
  7     0.9922    0.9908      0.9534      6.0777   670.1138     NA   687.1401    2071654.7732    49023.0709   1053.5660   0.0109
  8     0.9922    0.9906      0.9492      8.0008   672.0173     NA   690.9355    2120594.5973    51029.6274   1105.4163   0.0113
  9     0.9922    0.9904      0.9489     10.0000   674.0163     NA   694.8263    2176355.2137    53239.3714   1163.5723   0.0118
 10     0.9922    0.9904      0.9489     10.0000   676.0163     NA   698.7181    2176355.2137    53239.3714   1163.5723   0.0118
--------------------------------------------------------------------------------------------------------------------------------
```

Figure 3 - Human Fires Variable Selection Method



```
                                                Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------------------
                   Adj.        Pred
Model  R-Square  R-Square    R-Square     C(p)        AIC       SBIC     SBC         MSEP          FPE          HSP       APC
-----------------------------------------------------------------------------------------------------------------------------
  1     0.1000    0.0812      0.0372   11.7183   692.9112     NA   698.6473   2822762.3476   58711.4581   1201.1346   0.9750
  2     0.2060    0.1722      0.1137    6.9217   688.6470     NA   696.2951   2544508.4207   53917.3231   1105.7695   0.8954
  3     0.2935    0.2464     -0.5677    3.9640   671.9270     NA   681.3861   2300915.8214   50739.8175   1066.1454   0.8321
  4     0.3529    0.2941     -0.2416    2.1809   669.6205     NA   680.9714   2156333.4689   48423.6588   1021.8602   0.7941
  5     0.3665    0.2928     -0.4124    3.3179   670.5827     NA   683.8255   2161410.4007   49409.7776   1048.0862   0.8103
  6     0.3703    0.2804     -0.4569    5.0742   672.2857     NA   687.4203   2200746.3231   51194.6662   1092.5691   0.8396
  7     0.3705    0.2631    -15.2783    7.0617   674.2704     NA   691.2968   2255060.9934   53363.1455   1146.8395   0.8751
  8     0.3715    0.2458    -15.6251    9.0000   676.1949     NA   695.1131   2309320.7668   55571.1019   1203.7949   0.9113
  9     0.3715    0.2458    -15.6251    9.0000   678.1949     NA   699.0049   2309320.7668   55571.1019   1203.7949   0.9113
-----------------------------------------------------------------------------------------------------------------------------
```

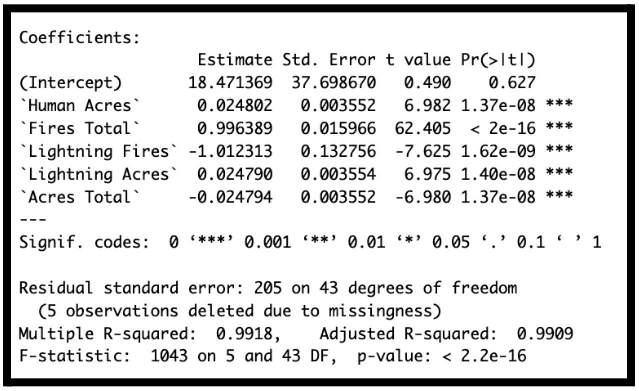
Figure 4 - Lightning Fires Variable Selection Method

The completion of the variable selection process allows for the set up of the linear regression equations. The best model for predicting human fires (EQ1) revealed statistical significance (p-value < 0.05). This signified how the predictors collectively contributed in explaining the variability observed in human fires across all 50 states. Approximately 99.09% (Figure 5) of the variance in human-caused fires is accounted for by the joint influence of the specified predictors. Each predictor—Human Acres, Fires Total, Lightning Fires, Lightning Acres, and Acres Total, was found to have a statistical significance impacting the prediction of human-made fires.

The best model for predicting lightning fires (EQ2) reveals statistically significant results (p-value < 0.05), indicating a meaningful relationship between the response variable and the predictors. The Adjusted R-squared value of 0.2928 indicates that approximately 29.28% of the variance in lightning caused fires is accounted for by the specified predictors (Figure 6). Based on the model's coefficient it emphasizes the influence obese people have on wildfires. For every one-unit increase in percentage of obese people, the model predicts a decrease of 25.41 units in lightning caused fires, holding all other variables constant. This implies that higher percentages of obese people in a state are associated with a significant reduction in the predicted number of lightning-caused fires.

EQ1: lm(Human_Fires ~ Human Acres + Fires Total + Lightning Fires + Lightning Acres + Acres Total, data = Statedata)

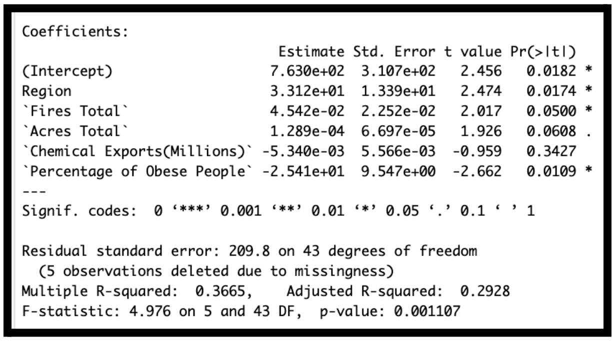
EQ2: lm(Lightning_Fires ~ Region +Fires Total +Acres Total +Chemical Exports(Millions) +Percentage of Obese People, data=Statedata)



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      18.471369  37.698670   0.490    0.627
`Human Acres`     0.024802   0.003552   6.982 1.37e-08 ***
`Fires Total`     0.996389   0.015966  62.405  < 2e-16 ***
`Lightning Fires` -1.012313   0.132756  -7.625 1.62e-09 ***
`Lightning Acres`  0.024790   0.003554   6.975 1.40e-08 ***
`Acres Total`     -0.024794   0.003552  -6.980 1.37e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 205 on 43 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9909
F-statistic:  1043 on 5 and 43 DF,  p-value: < 2.2e-16
```

Figure 5- Human Fires Response Linear Regression Out



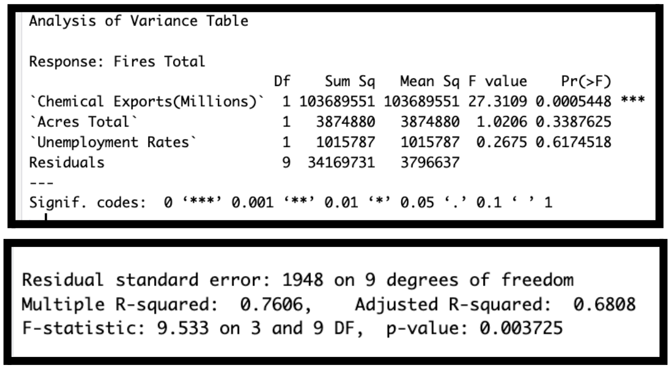
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  7.630e+02  3.107e+02   2.456   0.0182 *
Region                       3.312e+01  1.339e+01   2.474   0.0174 *
`Fires Total`                4.542e-02  2.252e-02   2.017   0.0500 *
`Acres Total`                1.289e-04  6.697e-05   1.926   0.0608 .
`Chemical Exports(Millions)` -5.340e-03  5.566e-03  -0.959   0.3427
`Percentage of Obese People` -2.541e+01  9.547e+00  -2.662   0.0109 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.8 on 43 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.3665,    Adjusted R-squared:  0.2928
F-statistic: 4.976 on 5 and 43 DF,  p-value: 0.001107
```

Figure 6 - Lightning Fires Response Linear Regression Out

When investigating the wildfire occurrence specific to the Southern Area region (EQ3) results yielded statistical significance for the model and individual significance for chemical exportation (Figure 7). The regression analysis depicts for every one-unit increase in chemical exports, the predicted fires total increases by 0.1495 units, holding all other variables constant. Whilst every one-unit increase in unemployment rates predicts the fires total will decrease by 738.2 units, holding all other variables constant. The model explains a substantial portion of the variability in fires total specific to the Southern region as evidenced by the high Multiple R-squared value of 0.7606. While the Adjusted R-squared is 0.6808, which implies that approximately 68.08% of the variability in Fires Total is accounted for by chemical exportation, percentage of obese people, and the amount of acres (See Appendix D). The explanatory power of linear regression techniques were able to justify and explain how uncorrelated and independent factors potentially can and do have some type of influence on the development and progression of wildfires.

EQ3 : lm(Fires Total ~ Chemical Exports(Millions) + Acres Total + Unemployment Rates, data = Southern_Area_data)



```
Analysis of Variance Table

Response: Fires Total
                             Df    Sum Sq   Mean Sq F value    Pr(>F)
`Chemical Exports(Millions)`  1 103689551 103689551 27.3109 0.0005448 ***
`Acres Total`                 1   3874880   3874880  1.0206 0.3387625
`Unemployment Rates`          1   1015787   1015787  0.2675 0.6174518
Residuals                     9  34169731   3796637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

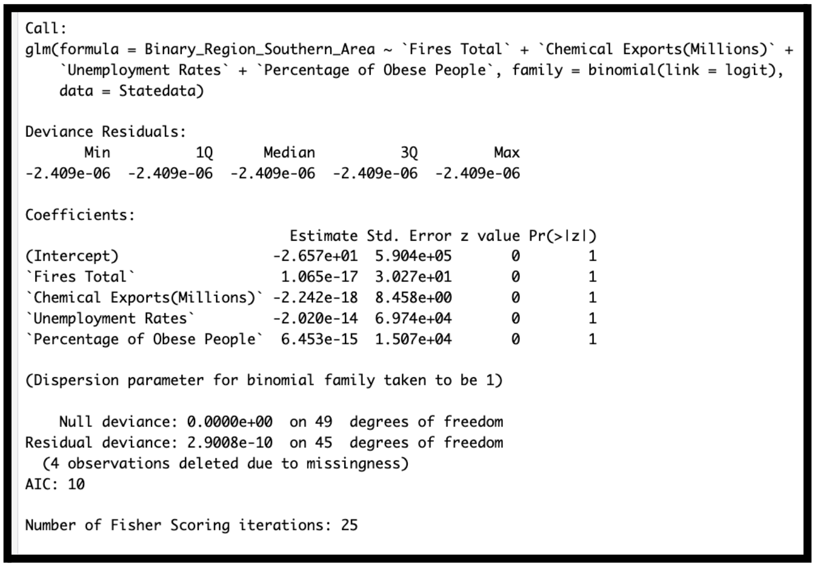
```
Residual standard error: 1948 on 9 degrees of freedom
Multiple R-squared:  0.7606,    Adjusted R-squared:  0.6808
F-statistic: 9.533 on 3 and 9 DF,  p-value: 0.003725
```

Figure 7 - EQ3 ANOVA and Regression Output

The logistic regression aimed to model the probability of a binary outcome as a function of one or more predictor variables. The model tested here was trying to predict the outcome of whether a wildfire observation is in the Southern Area or in the other regions based on the predictors of unemployment, obesity rates, and chemical exports (EQ4).

EQ4: glm(Binary_Region_Southern_Area ~ Fires Total + Chemical Exports(Millions) + Unemployment Rates + Percentage of Obese People, family = binomial(link = logit), data = Statedata)

The main conclusion from the logistic regression analysis (Figure 8) is that the predictors; the percentage of obese people, the total number of fires, chemical exports in millions, and unemployment rates do not seem to have a statistically significant effect on determining whether an observation is in the Southern Area. These predictors' p-values are all equal to 1, and their coefficients are small, close to zero, suggesting that there is insufficient data to reject the null hypothesis that the coefficients are zero. On the other hand, the intercept indicates a significantly lower log-odds of being in the Southern Area. There is a possibility of a data issue and data bias as the independent variables were hand selected. White the results of the logistic regression model were inconclusive, additional research and data are necessary to improve the model.



```
Call:
glm(formula = Binary_Region_Southern_Area ~ `Fires Total` + `Chemical Exports(Millions)` +
    `Unemployment Rates` + `Percentage of Obese People`, family = binomial(link = logit),
    data = Statedata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.657e+01  5.904e+05       0        1
`Fires Total`                1.065e-17  3.027e+01       0        1
`Chemical Exports(Millions)` -2.242e-18  8.458e+00       0        1
`Unemployment Rates`        -2.020e-14  6.974e+04       0        1
`Percentage of Obese People`  6.453e-15  1.507e+04       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 49  degrees of freedom
Residual deviance: 2.9008e-10  on 45  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 10

Number of Fisher Scoring iterations: 25
```

Figure 8 - Logistic Regression Output

The other technique utilized was CART Analysis. Consisting of the primary purpose to break down complex decision making in order to interpret a model and make decisions based on the class predicted. The classification trees developed focused on predicting which state and region were most likely to have a wildfire occur based on the given conditions of thresholds set for unemployment rates, chemical exportation, and percentage of obesity in each state (EQ4). The classification tree forecasts the probability of fire in several states (Figure 9). With a projected loss of 0.98, Node 1, the root, predominantly forecasts Alabama, suggesting a high probability of fire in the state. Node 2 improves the forecast for Alabama, adding to the calculation of fire likelihood with an expected loss of 0.96. At the same time, Node 3 projects a 0.96 anticipated loss for Alaska, indicating a chance of fire in the state's north. These nodes and the predicted losses that go along with them offer vital information on the complicated interactions among variables that affect the likelihood of wildfires in particular states. With an overall accuracy of roughly 93% for both Arkansas and Alaska, these nodes and the expected losses that go along with them offer insights into the detailed prediction of fire likelihood for particular states. This classification tree is an invaluable resource for understanding the complex variables affecting wildfire events and the precision of state-level wildfire forecasts. In essence the classification reveals that these 4 states: Arkansas, Alabama, Alaska, and California are at the highest risk for a wildfire when unemployment and chemical exports spike. This information can be used for state officials to set a plan in motion to help prepare and combat wildfires in their respective areas when conditions worsen to mitigate the spread and minimize negative implications of wildfires.

EQ4: rpart(State ~ `Unemployment Rates` + `Chemical Exports(Millions)` + `Region`, data = trainSet, method = "class")

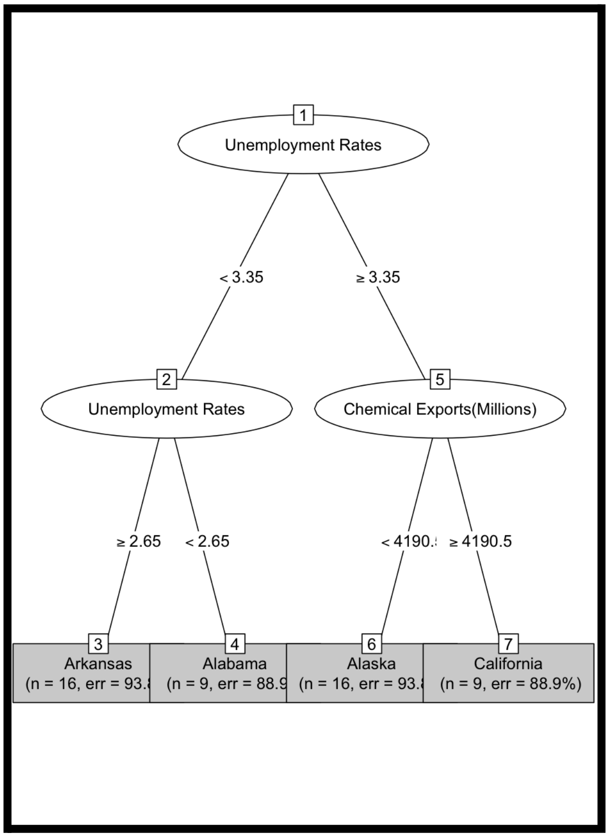


Figure 9 - Classification Tree Predicting State Wildfires

Several classification trees were developed and tested, iterating through conditions and variables thresholds which region is most at risk for a wildfire (EQ5). The classification tree shows the relationship between human acres, chemical exports, and the likelihood of wildfires (Figure 10). The Rocky Mountain region has the highest number of human acres and chemical exports, and is therefore the most likely to have a wildfire. Human acres and chemical exports are both known to be risk factors for wildfires. In this scenario the most meaningful and important factor in deciding where a wildfire will occur is the land capacity and amount of acreage in human-made fires, as depicted as the root node.

EQ5: rpart(Region ~ Human Acres + Lightning Acres + Unemployment Rates + Chemical Exports(Millions),   data = trainSet, method = "class")

The last default classification tree developed was interesting to analyze. With an emphasis and focus on how the percentage of obese people in each state play a role in wildfire frequency (EQ6). It is crucial to note that obesity wasn't the only predictor in the equation but it was the sole root node in the default tree, indicating its importance. The tree reveals how when the percentage of obese people are below 34.75% of a state's population, the Eastern Area region is more at risk for a wildfire(Figure 10). This indicates how less obese people in a given area increase the chances of wildfires happening while more obese people like in the Southern Area region, decrease chances of wildfires occurring.

EQ6: <- rpart(Region ~ Chemical Exports(Millions) + Percentage of Obese People , data = trainSet, method = "class")

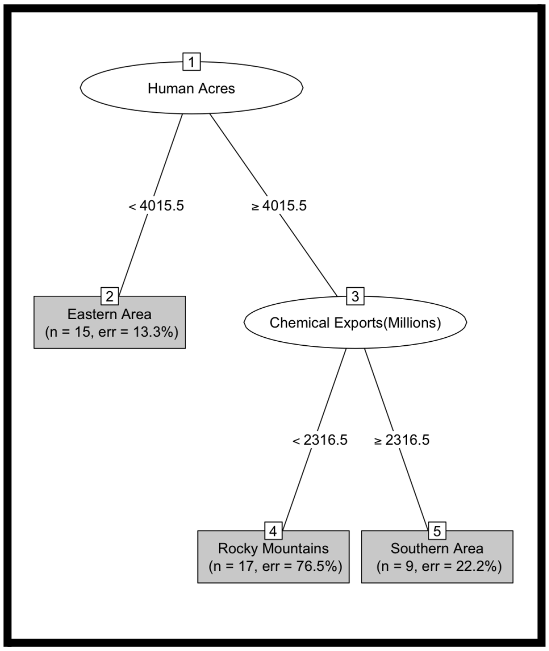


Figure 10 - Classification Tree 2

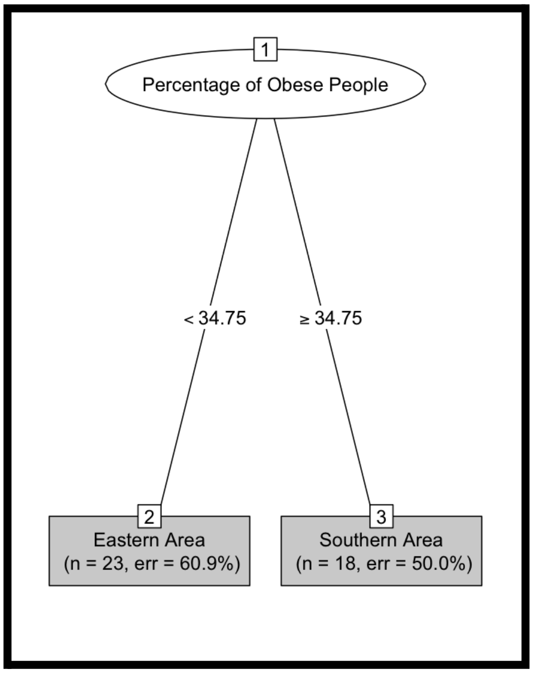


Figure 11 - Classification Tree 3

Results and Findings

The study extracted meaningful insights and fulfilled the gap in knowledge in nuanced relationships between various determinants and the occurrence of wildfires in the Southern Area states. It was discovered that chemical exports have a statistical influence in the severity and occurrence of wildfires (70% explanatory power) in Southern Area states. As well, if Unemployment rates are less than 3.35% and less than 2.65% then the most likely place for a wildfire to occur based on the given conditions is in Arkansas, yielding almost 93.89% accuracy. By detecting which Region or State is most likely to have a wildfire occur next based on these given conditions can help prepare state officials, politicians, and national park executives to reduce environmental and fiscal damage by increasing levels of preparedness.

Limitations and Constraints

Despite the valuable insights gained from this study, it is essential to acknowledge its limitations. There were constraints and challenges encountered during the data collection and management process. Acquiring data specific to national parks including visitor attendance, time series data of live fires, temperature and weather were not publicly accessible. This limitation may have introduced a potential bias in the analysis, as the study relies heavily on alternate data that could be obtained. As well as the computer software utilized, R-Studio could not compute full decision trees as the data combinations were too computationally expensive and the program repeatedly terminated. In addition the study's scope was magnified to fit into regions' and states' wildfires instead of specific national parks, so results may be interpreted in a more generalized sense not specific to certain national parks and their distinctive features.

Future directions

While this study sheds light on a broad overview of critical determinants of wildfires, it also opens avenues for further investigation. Future research needs to refine the focus to specific factors within National Parks. By narrowing the scope, researchers will aim to provide targeted insights into the dynamics between park-specific variables like; visitor attendance, park funding, fire safety personnel, and temperature data to predict and detect the occurrence and severity of wildfires.

To set this plan in motion necessary funding, resources, and experienced personnel are required. Our outlined plan proposes an estimated $10 million in funding, part of the five-year budget to implement this comprehensive wildfire prevention and prediction plan. While this budget seems unrealistic it equates to $2 million a year accounting for only 0.08% of the $2.5 billion federal budget for wildfire suppression. The simple reallocation of funds can be justified to change the practice of wildfire prevention forever. This funding would cover research initiatives, community education programs, and the acquisition of cutting-edge technologies. The plan's execution will be overseen by a committed interdisciplinary team made up of ecologists, meteorologists, fire scientists, software developers, and data analysts. The principal goals are to improve community resilience and response systems, mitigate the risk of wildfires by taking preventative action, and develop a state-of-the-art prediction system that makes use of cutting-edge technologies. The prediction and detection of wildfires are huge technical breakthroughs that employ benefits of resource optimization, early intervention practices, and environmental conservation.

Conclusion

A thorough analysis of this study adds to our knowledge of the variables that fuel wildfires and provides stakeholders with useful information to reduce risks and boost resilience. The multidisciplinary field of wildfire prevention is greatly affected by this research and data. This research enhances the general comprehension of the causes of wildfires and offers useful suggestions for those involved in managing the distinct environments of National Parks. This synthesis advocates for a proactive approach to environmental and wildfire management that takes socio-economic and environmental factors into account, like the increase of chemical exports or unemployment rates, greatly improving preparedness and response tactics. This data advances understanding of wildfire dynamics within the United States and National Parks, paving the way for revolutionary changes in wildfire research that will enable sustainability.

Citations

CarbonCredits. (2023, January 30). Wildfires Cost Over $148B and 30% of Emissions. https://carboncredits.com/wildfires-cost-emissions/#:~:text=Wildfires%20account%20for%20about%206,climate%20change%20can%20get%20worse

Dr. Chen Han Yu. (2023). Marquette University. MATH 4780 Lecture Slides.

National Interagency Fire Center. (2023). Statistics. https://www.nifc.gov/fire-information/statistics

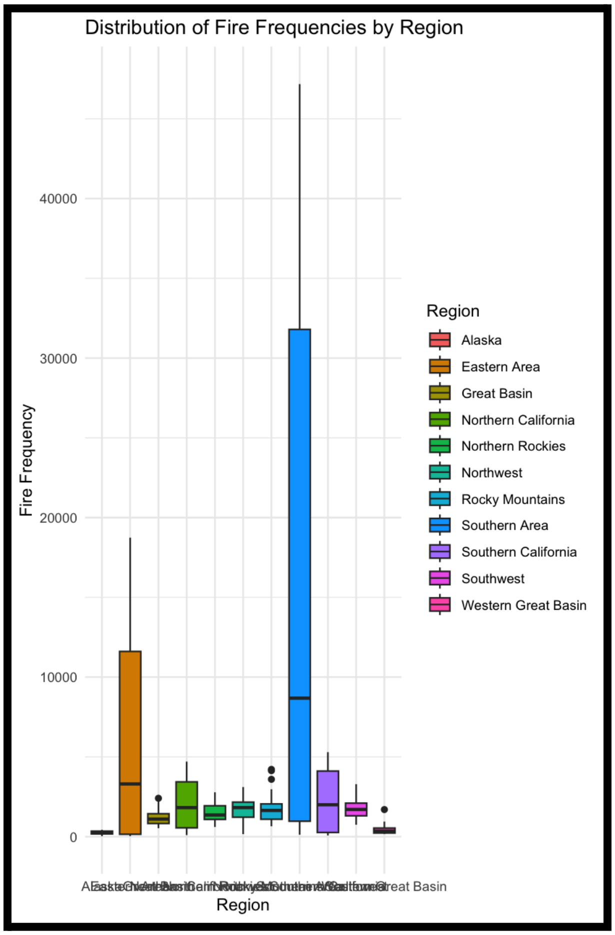
Statista. (2023, September 28). Value of United States chemical exports in 2022, by state (in million U.S. dollars). Retrieved from https://www.statista.com/statistics/297819/chemical-exports-of-us-states/

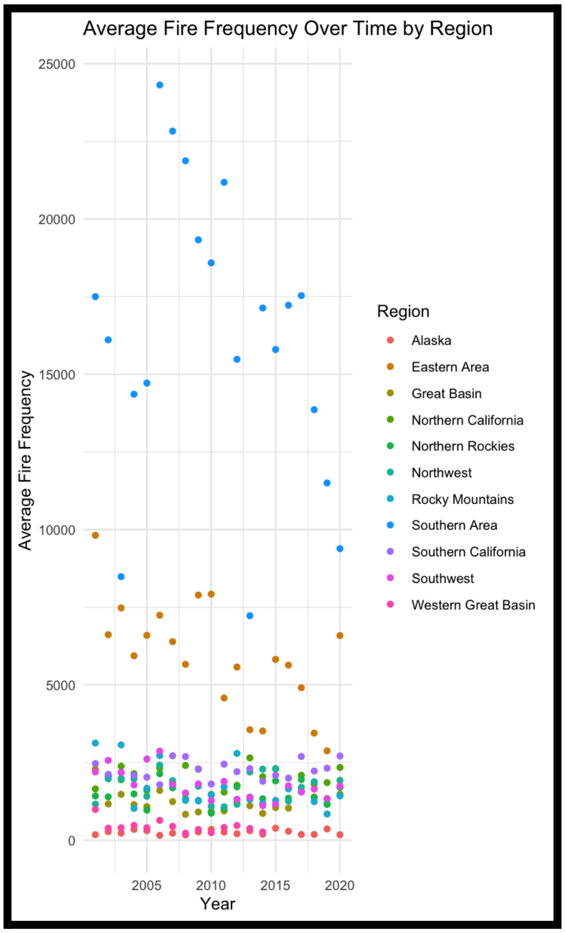
Statista. (2023, October 4). Percentage of obese U.S. adults by state 2022. John Elflein. https://www.statista.com/statistics/378988/us-obesity-rate-by-state/

UChicago News. (October 25, 2022).Wildfires are erasing California's climate gains, research shows. https://news.uchicago.edu/story/wildfires-are-erasing-californias-climate-gains-research-shows#:~:text=What%20is%20often%20ignored%20is,the%20state's%20greenhouse%20gas%20emissions.

U.S. Bureau of Labor Statistics. (2022). Unemployment Rates for States. Local Area Unemployment Statistics. Retrieved from https://www.bls.gov/web/laus/laumstrk.htm

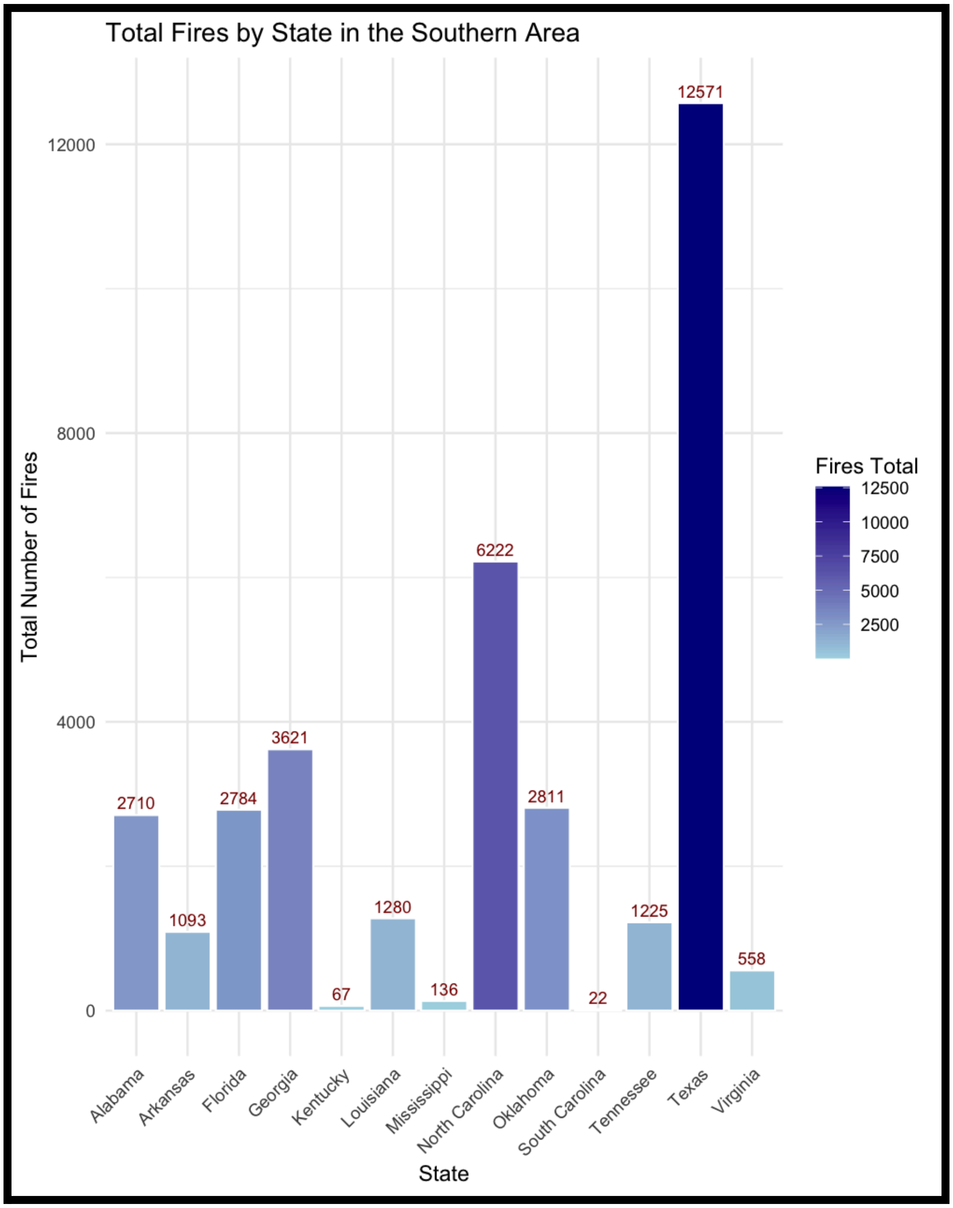
Appendix A





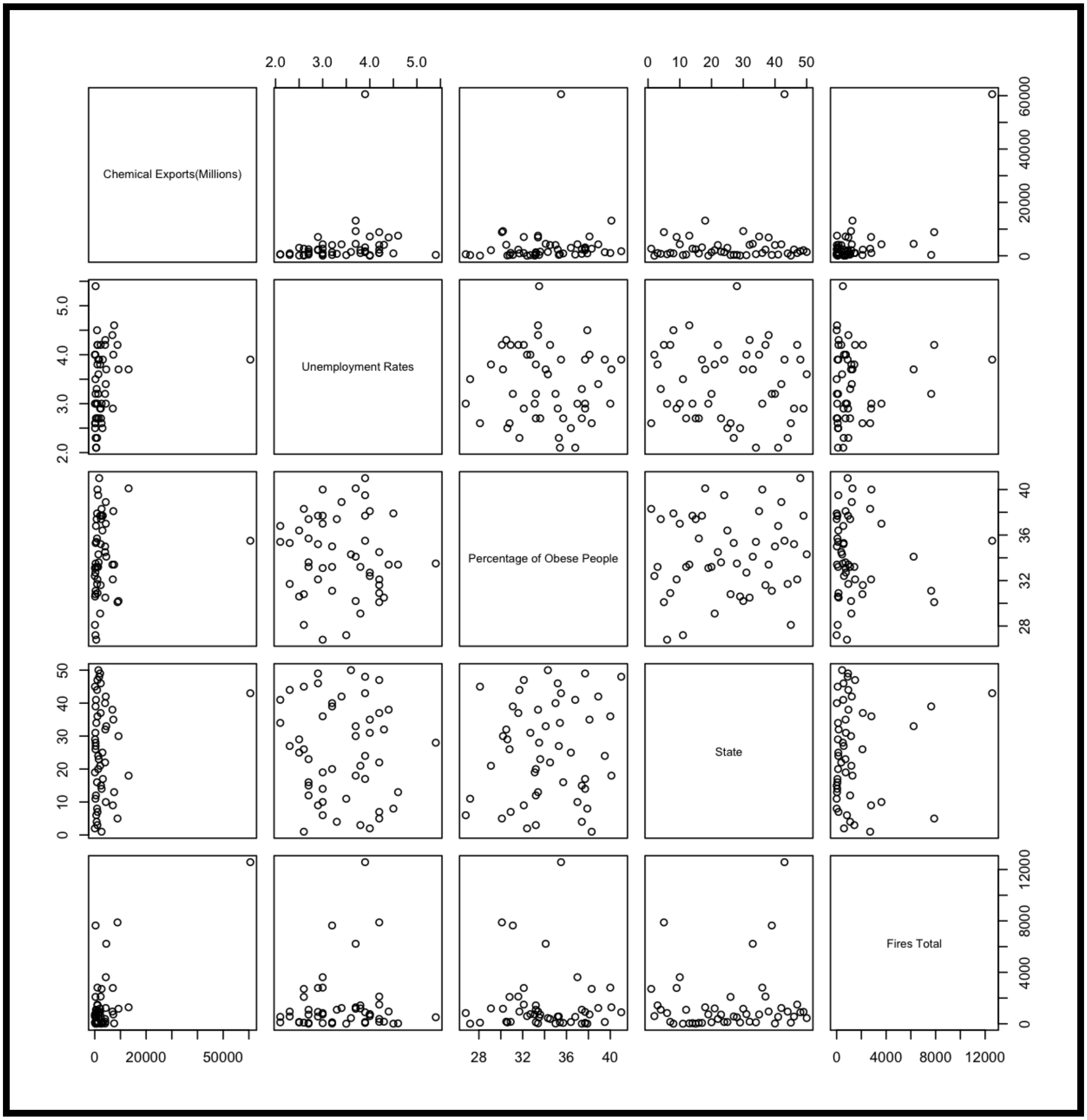
Exploratory visualizations examining the 10 different geographic regions based on coordination centers of national parks and protected lands in the United States and their distributions of fire frequency. The southern area region is spiked the highest and has the largest range of wildfire occurrence with a singular calendar year.

Appendix B



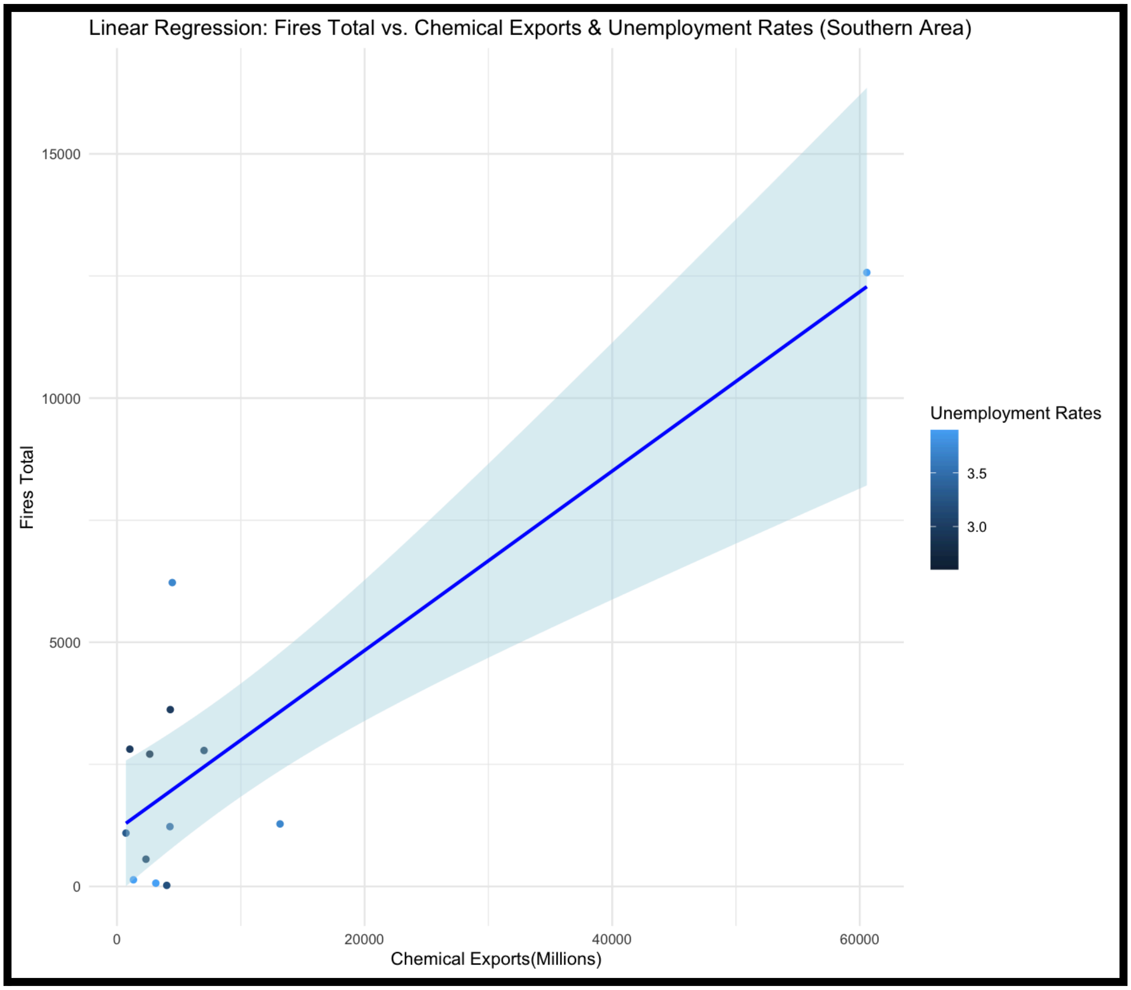
This exploratory visualization was a further analysis on the distribution of wildfires in the Southern Area regions and uncovered how Texas was the largest contender for wildfire occurrence in 2022. This is surprising due to Texas's low national park population, in comparison to the state of California. Which is relatively known for wildfires due to densely populated national parks within the state and dry weather.

Appendix C



This pairwise visualization suggests there is a strong positive correlation between Fires Total and Unemployment Rates. This suggests that states with higher unemployment rates also tend to have more fires. There is a moderate positive correlation between Fires Total and Chemical Exports (Millions). This suggests that states with higher levels of chemical exports also tend to have more fires.

Appendix D



Each data point represents one of the 14 states spanning across the Southern Area region. This visualization depicts how the majority of the observations fall within the 95% confidence interval and are clustered together indicating some type of relationship with chemical exportation and wildfire occurrence.