Analyzing Factors Associated with Area Burned by Wildfires in the United States

Info Innovators: Kevin Mao, Arnav Meduri, Ben Trokenheim, Ricardo Urena

2025-03-20

Introduction and Data

Background and Data Description

Wildfires are destructive natural disasters that occur regularly across the United States. According to the EPA, the U.S. has averaged approximately 70,000 wildfires per year since 1983[1]. Although fire is a natural part of many ecosystems, wildfires have important economic and environmental consequences (e.g., property loss, greenhouse gas emissions, and ecosystem degradation). After learning about recent wildfire events in California and western North Carolina, we became interested in better understanding the factors that contribute to variation in the area burned by wildfires. With wildfires affecting many parts of the country, identifying the factors associated with the extent of area burned can help inform decisions by wildfire management agencies. In light of this, we focused on two primary research questions: (a) What factors known before a wildfire has occurred are most strongly associated with the likelihood that a fire burns a greater-than-typical area? and (b) What overall factors (including those available after a wildfire) help explain variability in the continuous size of the burned area? To answer these research questions, we conducted exploratory data analysis and fitted both logistic and linear regression models to examine associations between wildfire characteristics and burned area.

The dataset used in this analysis is an integrated dataset consisting of over 55,000 wildfire records from the United States between 1992 and 2015, compiled from the Fire Program Analysis system. In addition to wildfire-specific attributes recorded in this database, the dataset was supplemented with additional information from the Forest Service Research Data Archive, NOAA Integrated Surface Hourly Database, vegetation and land cover data from Meiyappan and Jain's global land-use dataset, and geographic proximity data from SimpleMap's

[1] United States Environmental Protection Agency, "Climate Change Indicators: Wildfires," 2023, https://www.epa.gov/climate-indicators/climate-change-indicators-wildfires.

World Cities Database. As part of our analysis, we used a subset of these variables, including fire size (measured in acres) as the response variable; cause of fire (categorized as missing/undefined, arson, debris burning, miscellaneous, campfire, fireworks, children, lightning, equipment use, smoking, railroad, structure, or powerline); temperature (°C), wind speed (meters per second), relative humidity (%), and precipitation (millimeters) recorded 30 days prior to the fire; vegetation classification based on land cover (with categories Open Shrubland, Polar Desert/Rock/Ice, Secondary Tropical Evergreen Broadleaf Forest, Temperate Evergreen Needleleaf Forest, C3 Grassland/Steppe, Desert, and Water/Rivers); and remoteness (a unitless value between 0 and 1 representing the scaled distance from the nearest urban center).

Hypotheses

We hypothesize that (1) wildfire event characteristics, environmental conditions prior to discovery, and geographic factors are associated with the likelihood that a wildfire burns a greater-than-typical area, since hotter, drier, and windier conditions, more flammable vegetation, and greater remoteness are expected to increase the chance of larger fires; and (2) information available after a wildfire has occurred, such as underlying cause and environmental conditions during containment, helps explain additional variability in the continuous size of the burned area, as these factors may influence how the fire spread and how difficult it was to control.

Exploratory Data Analysis

Data Cleaning

Before conducting our analysis, we applied many data cleaning steps to prepare our dataset for modeling and interpretation. One of the major decisions we made as part of our data cleaning process was to filter our response variable, acres burned, since the majority of observations in our dataset (over 13,000) recorded fires that burned one acre or less of land (i.e., small-scale wildfires). In our analysis, we decided to focus only on the interquartile range (middle 50%) of wildfires by acres burned because a) the goal of our analysis was to focus on wildfires that can reasonably be addressed during early containment efforts (rather than fires that had already expanded beyond an early intervention phase) and b) to narrow our practical range for analysis (i.e., wildfires representative of "typical" wildfire events). In addition, we created a binary outcome based on fire size (i.e., grouping wildfires as either falling within the lower or upper 50% of fire sizes), which allowed us to focus on the likelihood of a wildfire exceeding the median size later on in our analysis.

Our data cleaning process also involved transforming many of our variables of interest to make downstream analysis and interpretation more manageable. For example, we observed that most precipitation-related variables in our dataset were heavily left-skewed (i.e., the majority of observations corresponded to wildfire events with no recorded precipitation during a given time period before discovery), so we transformed all precipitation variables into binary

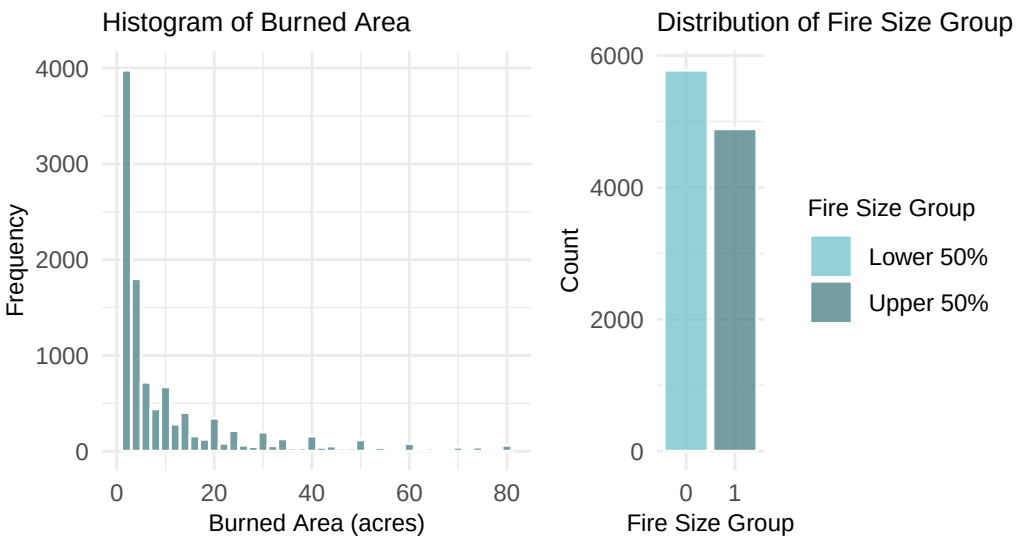
indicators (0 = no precipitation, 1 = precipitation greater than 0). Additionally, we "collapsed" many categorical variables with a large number of levels in our dataset, mainly to help with model interpretability and to reduce model complexity. Specifically, we grouped states into four broader regions based on U.S. Census classifications (Northeast, Midwest, South, and West), vegetation types into broader environmental categories (e.g., Forest, Shrubland, Grassland), and cause descriptions into broader cause categories (e.g., Natural, Recreational, Infrastructure).

Following standard data cleaning practices, we dropped all observations with missing values for any variables of interest in our dataset. Finally, we converted all categorical variables of interest (e.g., vegetation type, fire cause, geographic region) into factor variables to support appropriate handling during modeling and analysis.

Univariate EDA

As part of our analysis, we first conducted univariate EDA to better understand the distribution of variables in our dataset and to identify patterns (e.g., skewness, outliers) and potential data quality issues that could affect subsequent modeling. As mentioned previously, we focused on the middle 50% of wildfires by acres burned because these represent moderate-sized fires that are more likely to be responsive to early containment efforts.

Based on the histogram of burned area (left panel), we can see that the distribution of moderate-sized wildfires is right-skewed and unimodal. Additionally, there is a clear peak at 3,978 observations corresponding to wildfires that burned less than 2 acres of land. According to the summary statistics, the mean burned area within this subset is 11.84 acres, and the typical (median) wildfire size is 5 acres. The middle 50% of burned area values falls between 2.5 acres (first quartile) and 14.7 acres (third quartile), and the minimum and maximum values are 1.5 and 80 acres, respectively. The standard deviation is 15.62 acres, which indicates there is substantial variability in fire size within this range. Additionally, many wildfires have burned areas above approximately 30 acres, extending beyond the typical range of values; these observations could be considered moderate outliers within this subset. In terms of the binary outcome for fire size (right panel), we observe that 54.2% of wildfires fall into the lower 50% of fire sizes (0), while 45.8% of wildfires fall into the upper 50% (1), which is a roughly balanced distribution between the two groups.



Count	Min	Q1	Median	Mean	Q3	Max	SD
10669	1.5	2.5	5	11.83	14.7	80	15.62

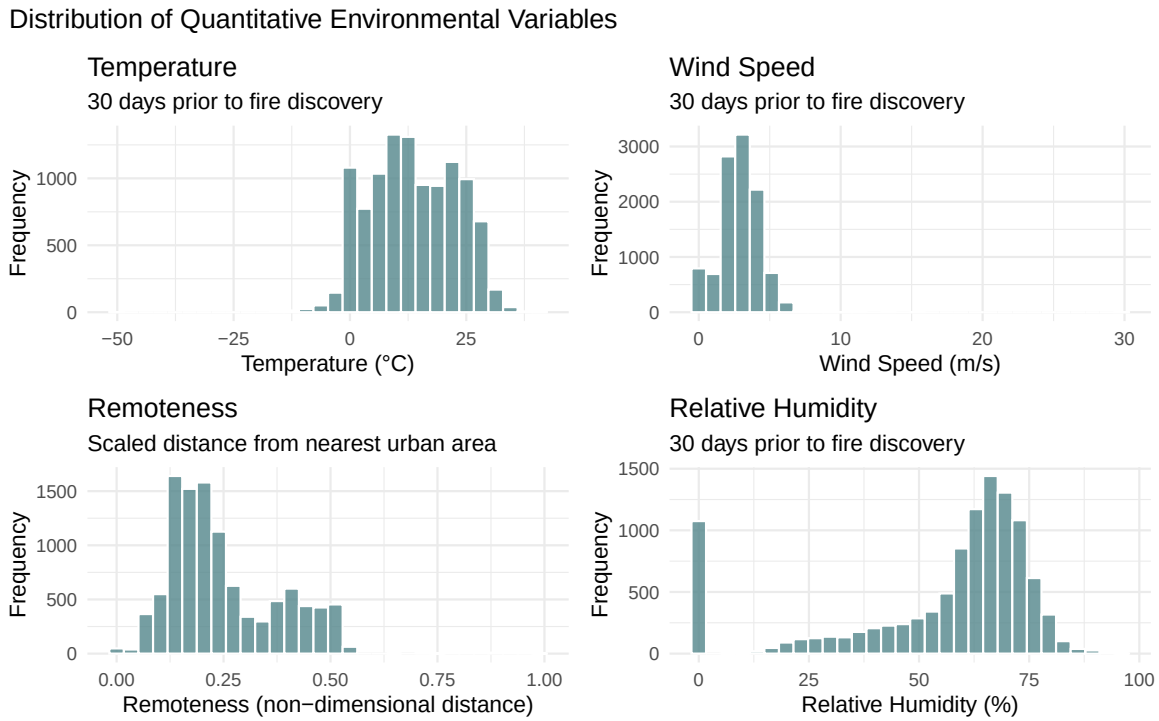
Fire Size Group	Count	Proportion
0	5779	54.2%
1	4890	45.8%

Next, we visualized the distribution of the quantitative variables of interest in our dataset (temperature (°C), relative humidity (%), and wind speed from 30 days prior to fire discovery (m/s), and remoteness), along with summary statistics to better understand their scale and variability. We first examined the distribution of temperature a month before discovery (top left panel), which we observed to be approximately normal and unimodal. (This is consistent with what we would generally expect, since temperature tends to change gradually across different areas rather than clustering at specific values.) The median temperature for wildfire observations in our dataset is 12.97°C and the mean is 13.58°C, with most observations between 6.53°C and 21.36°C.

We then examined the distributions of wind speed thirty days before discovery and remoteness (top right and bottom left panels, respectively). Wind speed appears to be very slightly right-skewed and unimodal, while remoteness appears more strongly right-skewed and unimodal. This indicates that most fires occurred relatively close to populated areas (lower remoteness)

and that most wind speeds were low to moderate. The median remoteness for observations in our dataset is 0.21 and the mean is 0.25, with most observations between 0.15 and 0.35. The median wind speed for observations in our dataset is 2.89 m/s and the mean is 2.87 m/s, with most observations between 2.06 m/s and 3.76 m/s.

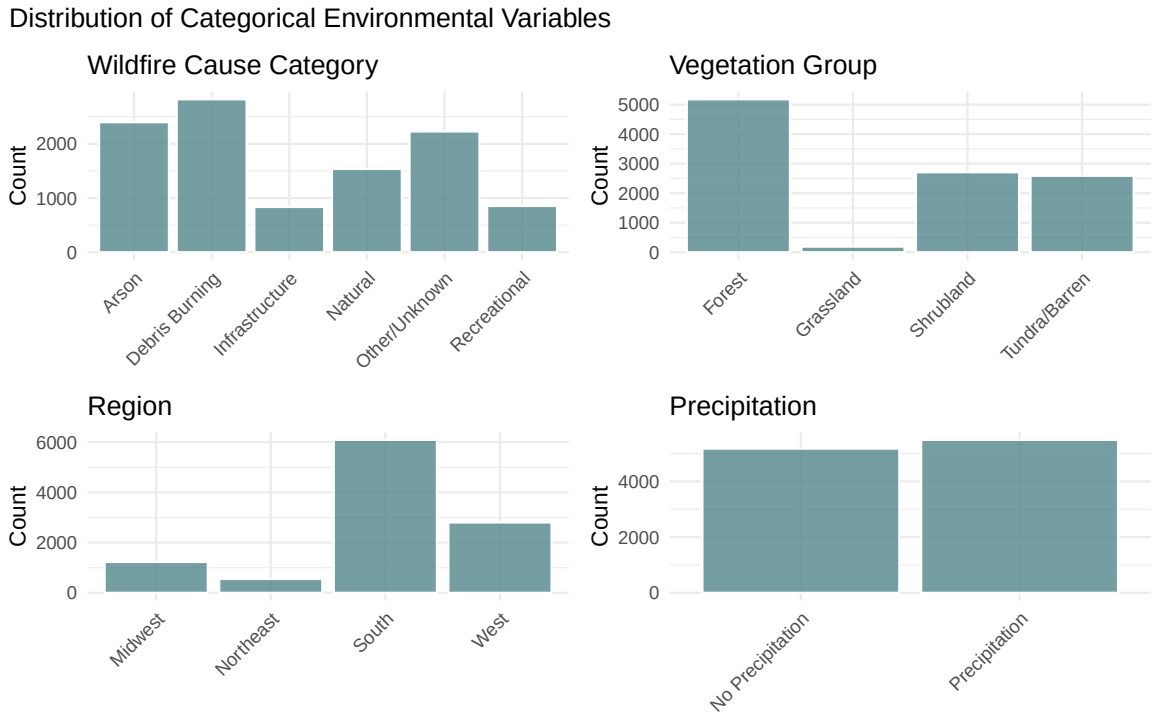
Lastly, we looked at the distribution of relative humidity thirty days before discovery (bottom right panel), which appears to be left-skewed and unimodal. This tells us that lower humidity conditions were more common among wildfires in our data. The median relative humidity for observations in our dataset is 63.53% and the mean is 55.20%, with most observations between 48.80% and 69.97%. However, we noticed that a large number of observations in our data recorded a relative humidity of exactly 0%. Given that 0% relative humidity is unlikely under normal atmospheric conditions (since even very dry air typically contains some moisture), we decided to exclude these observations from our analysis.



Variable	Min	Q1	Median	Mean	Q3	Max	SD
Temperature	-49.21	6.53	12.97	13.58	21.35	41.68	9.21
Humidity	0.00	48.80	63.52	55.20	69.97	96.00	22.76
Wind	0.00	2.06	2.89	2.87	3.75	29.80	1.40
Remoteness	0.00	0.15	0.21	0.25	0.35	0.99	0.13

In addition to examining the distributions of quantitative predictors, we also visualized the

distribution of categorical variables of interest in our dataset (i.e., wildfire cause category, vegetation group, region, and precipitation 30 days prior to wildfire discovery). Starting with precipitation 30 days before discovery, we observed that the majority of wildfire observations (5492 observations or 51.5%) corresponded to fires with no precipitation during that period, while the remaining observations (5179 observations or 48.5%) corresponded to fires where precipitation occurred.



Looking more closely at vegetation group, we can see that most fires occurred in forested areas (5181 observations or 48.6%), followed by shrubland (2711 observations or 25.4%) and tundra or barren areas (2590 observations or 24.3%), with relatively few fires in grassland environments (189 observations or 1.77%). Examining wildfire cause categories, we found that a large portion of fires were attributed to human-related causes rather than natural causes. Debris burning accounted for the highest proportion (2821 observations or 26.4%), followed by arson (2399 observations or 22.5%) and other or unknown causes (2228 observations or 20.9%). Natural causes made up a smaller proportion (1535 observations or 14.4%), followed by recreational causes (853 observations or 8.0%) and infrastructure-related causes (835 observations or 7.82%).

Finally, we examined the regional distribution of fires. Most fires in our dataset occurred in the South (6090 observations or 57.1%), followed by the West (2801 observations or 26.2%), Midwest (1228 observations or 11.5%), and Northeast (550 observations or 5.15%). Only two

fires (2 observations or less than 0.01%) were categorized as occurring in the "Other" region classification.

Bivariate EDA

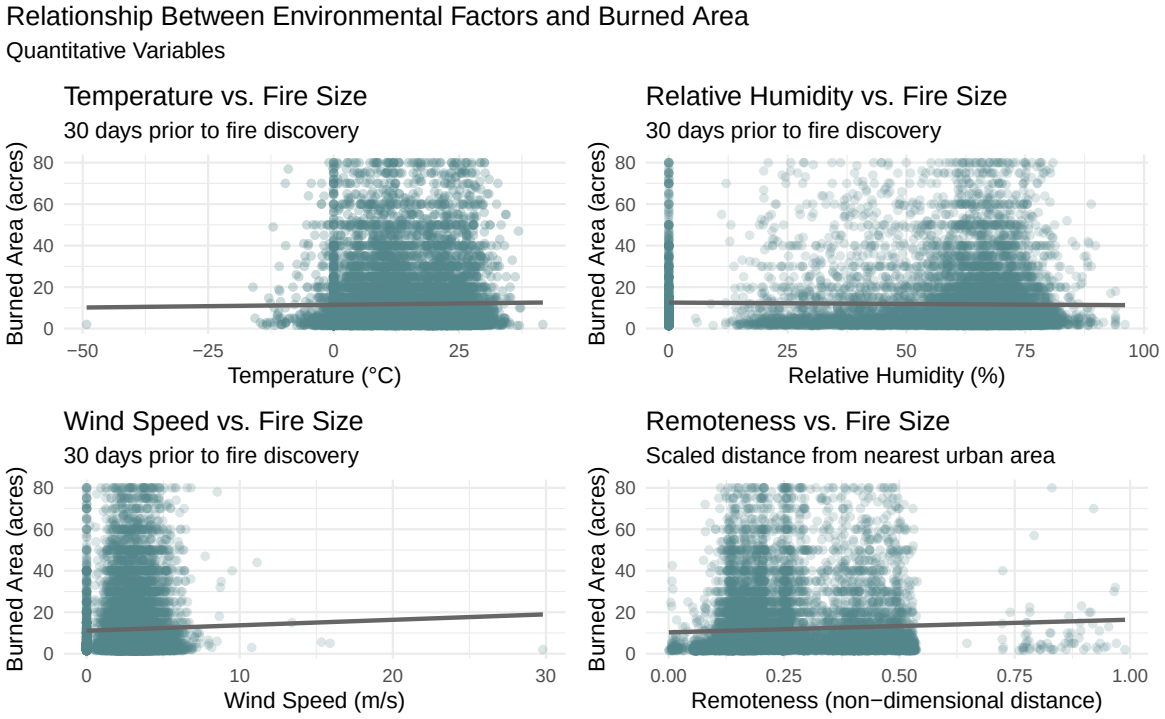
As the next step in our analysis, we examined the relationship between each of our quantitative predictors and fire size to better understand their associations. As we can see in the scatter plot of burned area versus temperature (top left panel), the relationship between these two variables appears to be very weak and slightly positive (i.e., suggesting that larger fires may be slightly more likely at higher temperatures). The relationship also appears to be nonlinear, since most fires in our dataset seem to be pretty small regardless of temperature, and the largest fires (greater than 60 acres in size) most commonly occur between 5°C and 20°C. This tells us that while moderate temperatures may allow for larger fires, temperature alone does not strongly control fire size.

Similarly, the relationship between relative humidity and burned area (top right panel) appears to be very weak and highly nonlinear (with fires of all sizes occurring across almost all humidity levels and the largest fires tending to occur at mid-range humidity values at approximately 25% to 75%), but slightly negative (i.e., suggesting that burned area slightly decreases as humidity increases). As with temperature, the wide vertical spread of data points around the line of best fit tells us that relative humidity may not be a strong predictor of burned area on its own.

Additionally, the relationship between wind speed and burned area (bottom left panel) appears to be slightly positive (i.e., indicating a slight tendency for larger fires to occur at higher wind speeds), very weak, and highly nonlinear. As we can see in the plot, most data points in the plot are concentrated at lower wind speeds, and most larger fires (fires greater than 40 acres in size) occur below 5 m/s, which indicates that higher wind speeds are not consistently associated with larger burned areas.

Lastly, looking at the scatter plot of remoteness versus burned area (bottom right panel), we see a weak, slightly positive relationship between these two variables (i.e., suggesting that fires originating further away from cities may end up burning more acres of land, on average). This relationship also appears to be highly nonlinear, since the largest fires occur at moderate remoteness values between 0.25 and 0.50, and relatively few large fires at very high remoteness values greater than 0.75 (which tells us that burned area does not consistently increase or decrease across the range of remoteness values).

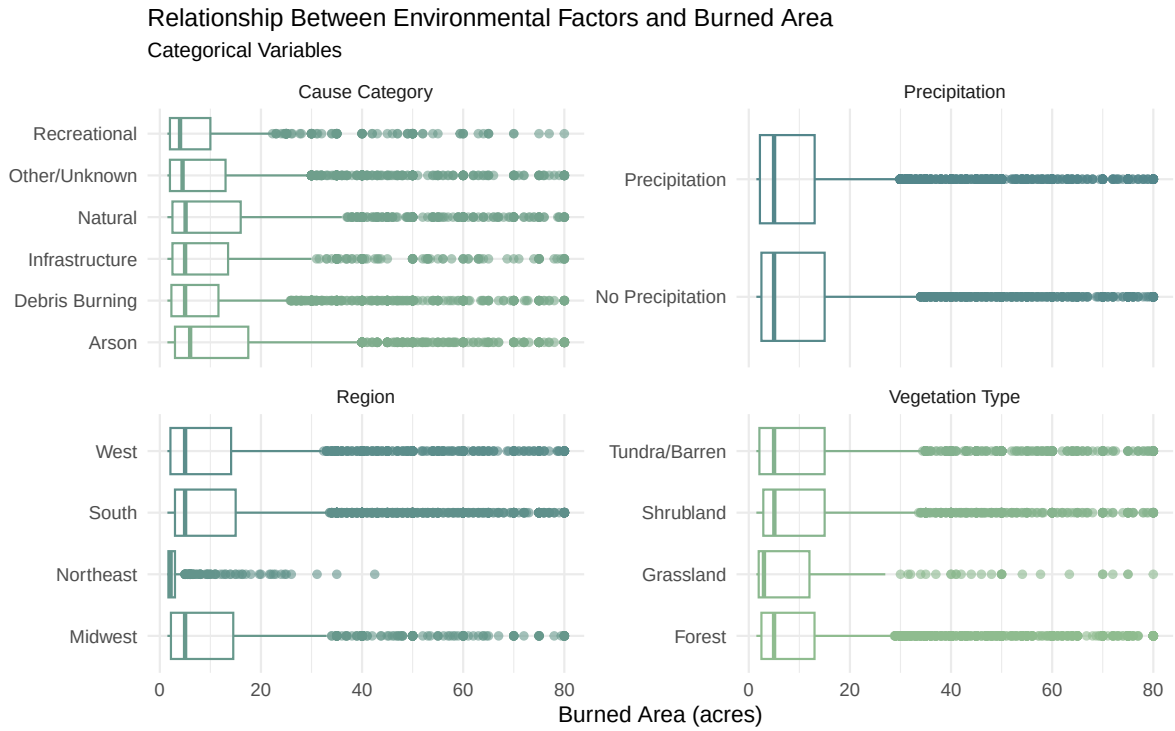
In general, there is a wide vertical spread (i.e., high variability) in the data points around the lines of best fit for all predictors, which suggests that the quantitative predictors in our dataset may have relatively limited predictive power for burned area when considered individually. That being said, this is somewhat expected given that wildfire size can be influenced by many factors beyond the limited set of quantitative predictors available in our dataset.



As part of our bivariate EDA, we also analyzed the relationship between each of our categorical variables of interest and burned area to better understand differences in fire size across groups. We can see from the visualization (top left panel) that the distribution of acres burned is relatively similar for all wildfire cause categories, with medians between 4.00 and 6.00 acres and means between 9.02 and 13.74 acres, and outliers beyond the 30-acre mark across all cause types. This indicates that cause category on its own does not explain much variability in fire size. We observed a similar pattern in terms of the distributions between observations corresponding to no precipitation 30 days before discovery and observations corresponding to precipitation (top right panel). The distributions of fire size appear to be very similar, with medians of 5.00 acres in both groups and means of 12.32 acres for no precipitation and 11.38 acres for precipitation, although there is slightly more spread in the middle 50% of fires with no precipitation compared to fires with precipitation.

We observed more apparent differences in the distribution of acres burned across different regions and vegetation groups (bottom left and bottom right panels, respectively). In particular, the Northeast had a lower median (3.00 acres) and mean (6.04 acres) compared to other regions, with much less variability in the middle 50% of fires and most outliers limited to about the 40-acre mark. In contrast, distributions for the West, South, and Midwest were more similar, with medians between 5.00 and 6.00 acres, means between 12.00 and 13.00 acres, and a large number of outliers beyond the 30-acre mark. For vegetation type, the distribution of fire size was fairly similar across categories, although grasslands had a slightly lower median (4.00 acres) and mean (8.43 acres) compared to other vegetation groups. That being said, the

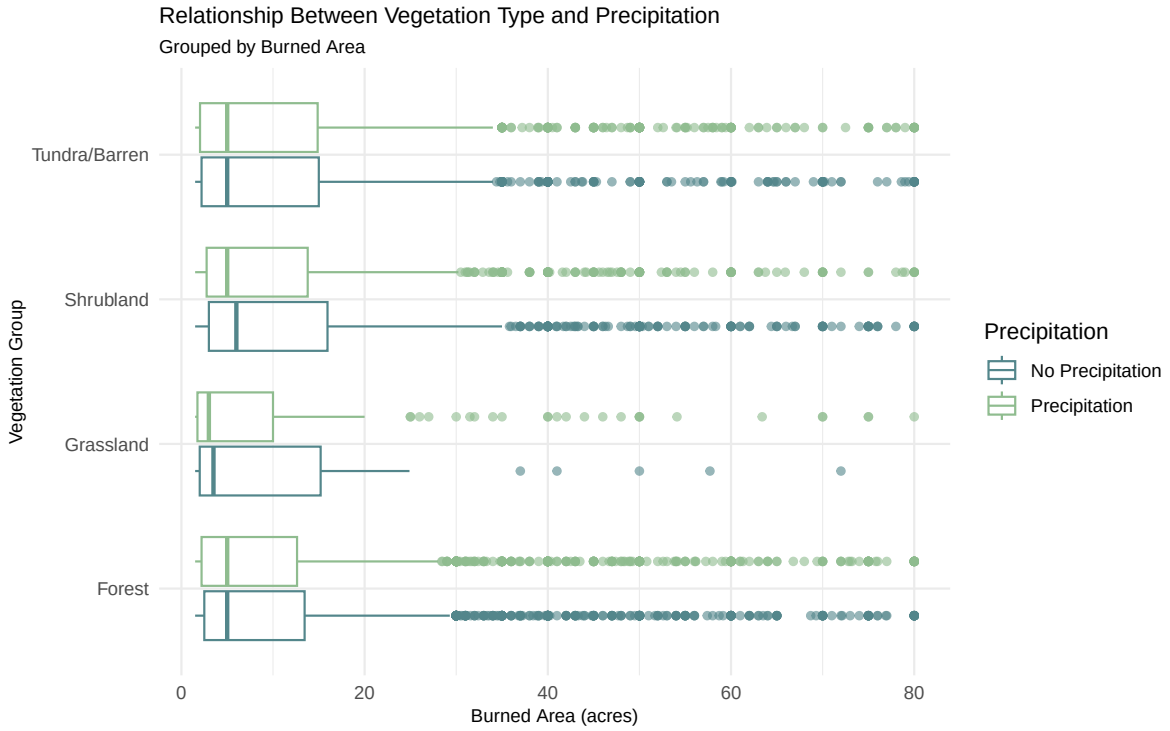
middle 50% of fire sizes and the overall range of outliers were generally similar across vegetation types, with grasslands showing somewhat fewer extreme outliers. Overall, region and vegetation group seem to explain more of the variability in fire size on their own compared to cause category or precipitation, and could potentially be useful predictors in modeling fire size.



Potential Interaction Effects

Since many of our variables of interest relate to environmental conditions and geographic location, there are potential interaction effects among our predictors that could influence acres burned. As part of our analysis, we were interested in a potential interaction between vegetation type and precipitation, since vegetation and fuel moisture could influence burned area together. However, based on the visualization below, the distributions of burned area appear relatively similar between precipitation and no precipitation conditions for each vegetation group. In forest, the median burned area is 5.0 acres for both no precipitation (mean of 11.48 acres) and precipitation (mean of 11.14 acres), while in grassland, the median is 3.5 acres for no precipitation (mean 12.26 acres) and 3.0 acres for precipitation (mean of 11.54 acres). Additionally, in shrubland, the median is 6.0 acres for no precipitation (mean of 13.52 acres) and 5.0 acres for precipitation (mean of11.31 acres), while in tundra/barren, the median is 5.0 acres for both no precipitation (mean of 12.81 acres) and precipitation (mean of 11.86 acres). Even though there does not appear to be a very strong effect, this interaction would be worth

exploring more formally in the modeling process to determine whether it significantly improves model fit.



Methodology

The main goal of our analysis was to examine how various wildfire-related factors are related to burned area, and how these factors can be used to predict fire size. Since these represent two distinct research questions, we chose to fit two separate models: a linear regression model using all variables of interest, including those only available after a fire is contained, to explain variability in burned area (i.e., an explanatory model), and a logistic regression model using only variables that are available at or before fire discovery to guide response efforts (i.e., a predictive model). It is important to note that our response variable, fire size, is continuous and therefore suitable for linear regression. However, logistic regression requires a binary response variable. To address this, we created a binary response variable based on whether a fire exceeded the median fire size, classifying fires greater than the median as 1 and those less than or equal to the median as 0. (In this setup, the model estimates the likelihood that a wildfire falls into the upper half of moderate-sized fires based on available environmental factors.)

Explanatory Modeling

To guide predictor selection for the explanatory model, we used a forward selection approach based on adjusted R^2 (which accounts for model fit while penalizing unnecessary predictors). Ultimately, we included all predictors of interest in our model, since each contributed to an increase in adjusted R^2 in our model. Specifically, we arrived at a linear regression model predicting fire size from region, temperature, wind speed, precipitation, relative humidity (measured 15 days before fire discovery), remoteness, vegetation group, and an interaction between vegetation group and precipitation. We decided to mean-center the quantitative predictors in our model to improve interpretability of the coefficients.

After fitting this initial model, we conducted diagnostics to evaluate whether key assumptions (i.e., linearity, normality, constant variance, and independence) were satisfied. We found that the normality, linearity, and independence conditions were reasonably satisfied, but the constant variance condition was violated, and there was a potential violation in the linearity condition. To address constant variance, we applied a variance-stabilizing (logarithmic) transformation to the fire size variable. For linearity, we examined residuals against predictors to more closely assess if there were any nonlinear relationships between predictors and fire size. However, we found no concerning evidence of non-linearity, so we did not transform our predictors.

Additionally, when examining the distribution of residuals versus fitted values for this initial model, we observed a clustering of observations into two groups: one with fitted values 5, and another with fitted values between 5 and 20. Through further analysis, we realized that this clustering was partially explained by differences in region. (As we saw in our exploratory data analysis, fires in the Northeast generally corresponded to smaller burned areas compared to fires in other regions, which likely contributed to the small separation observed in the residuals versus fitted values plot.) To address this, we explored whether fitting two separate models (one for fires occurring in the Northeast and one for fires occurring in other regions) would improve explanatory power, but we ultimately decided to proceed with a single model including all regions because the improvement in fit was minimal (and for the sake of consistency in interpretation).

We also checked for influential points based on Cook's Distance ($D_i > 0.5$), and found no such observations in our dataset that could substantially affect regression coefficient estimates. Lastly, when examining the correlations between predictors, we found evidence of collinearity between certain region levels (specifically, the similarity between the Southern and Western regions). To address this, we combined these two regions into a single level, which helped reduce collinearity without compromising overall explainability of the model.

Predictive Modeling

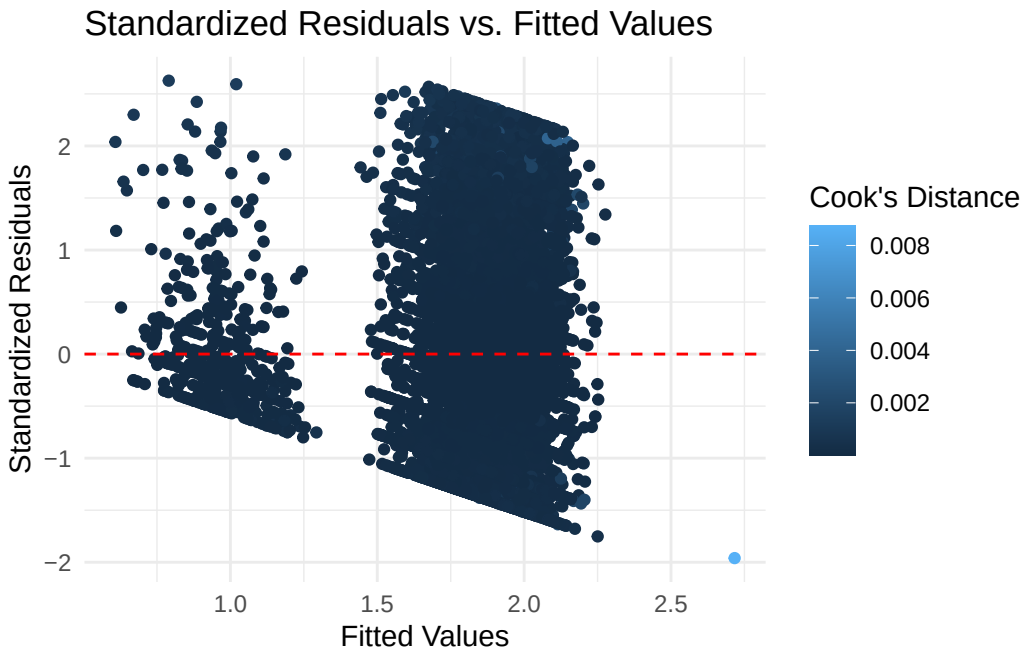
As part of our explanatory modeling process, we retained all of the vairables in oru rpeicigve modeling (rpocodess.idk), but added putout time and cause cagtegory (find mroea aoptiw woridng) to . We consduerd both of thesevarbales at this step since (explain) (The casue of a widlfire is likely only to be known muchbafter the fire and coanot be ued to rpedcigt, and putout time…..) These factros provide

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.035	0.112	0.313	0.755	-0.185	0.255
regionNortheast	-1.939	0.152	-12.785	0.000	-2.243	-1.648
regionSouth_West	0.164	0.070	2.352	0.019	0.028	0.301
Temp_pre_30	-0.007	0.002	-2.741	0.006	-0.012	-0.002
remoteness	-0.641	0.192	-3.338	0.001	-1.019	-0.265
Vegetation_groupGrassland	0.173	0.339	0.512	0.609	-0.501	0.837
Vegetation_groupShrubland	0.213	0.069	3.088	0.002	0.078	0.348
Vegetation_groupTundra/Barren	0.071	0.075	0.947	0.344	-0.076	0.217
Wind_pre_30	0.048	0.016	2.932	0.003	0.016	0.080
Prec_pre_30	-0.046	0.060	-0.776	0.437	-0.163	0.070
Hum_pre_30	0.000	0.001	0.414	0.679	-0.002	0.002
putout_time_num	0.013	0.003	4.042	0.000	0.007	0.020
cause_categoryDebris Burning	-0.412	0.057	-7.247	0.000	-0.523	-0.301
cause_categoryInfrastructure	-0.132	0.083	-1.589	0.112	-0.295	0.031
cause_categoryNatural	-0.046	0.074	-0.623	0.533	-0.191	0.099
cause_categoryOther/Unknown	-0.275	0.062	-4.435	0.000	-0.397	-0.154
cause_categoryRecreational	-0.611	0.084	-7.275	0.000	-0.777	-0.447
Vegetation_groupGrassland:Prec_pre_30	0.124	0.388	0.320	0.749	-0.636	0.892
Vegetation_groupShrubland:Prec_pre_30	-0.140	0.096	-1.454	0.146	-0.329	0.049
Vegetation_groupTundra/Barren:Prec_pre_30	0.080	0.100	0.804	0.421	-0.115	0.276

Results

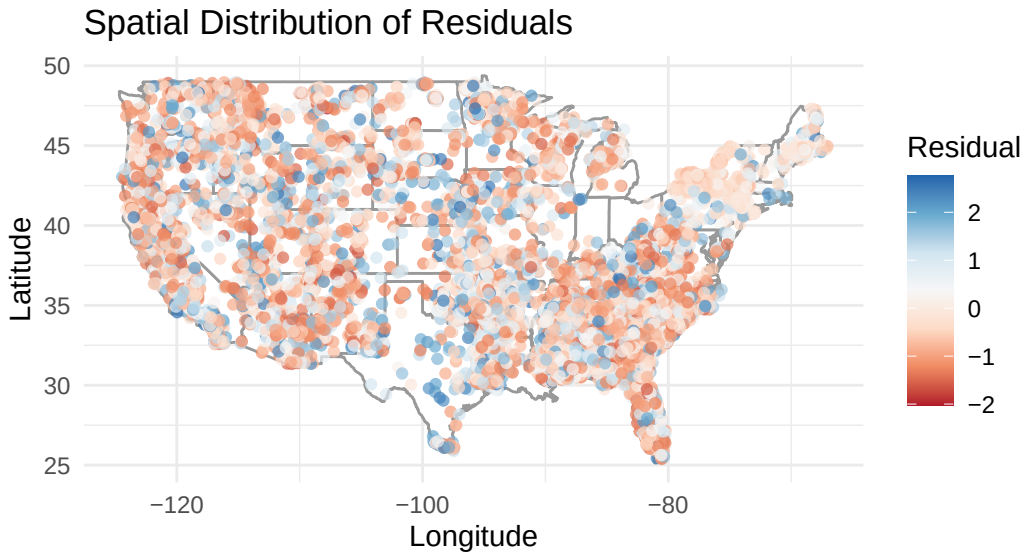
Discussion

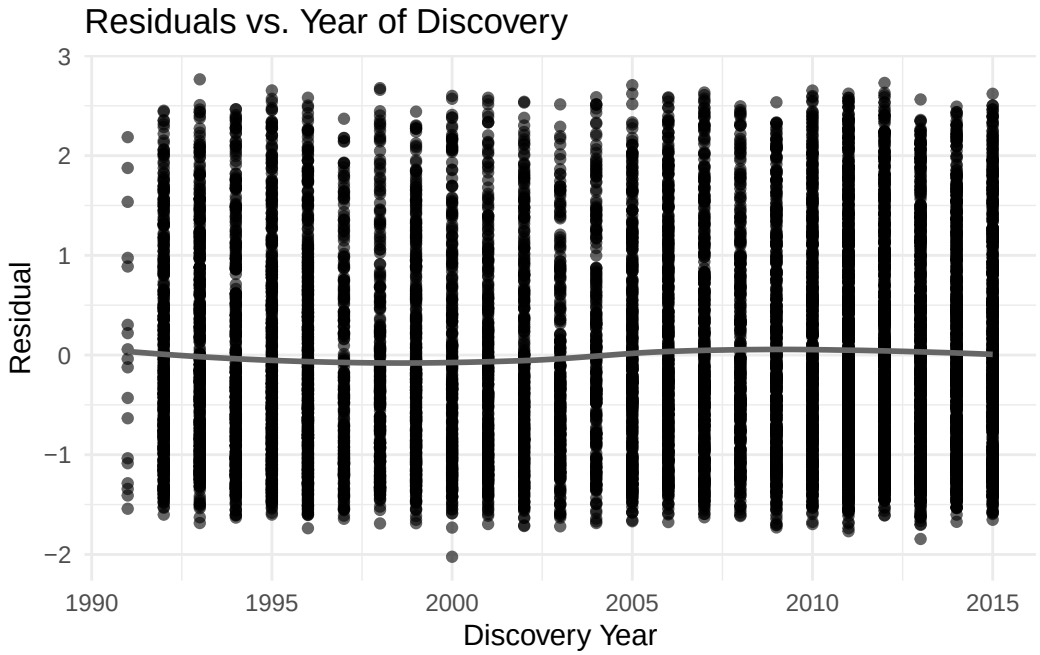
Appendix



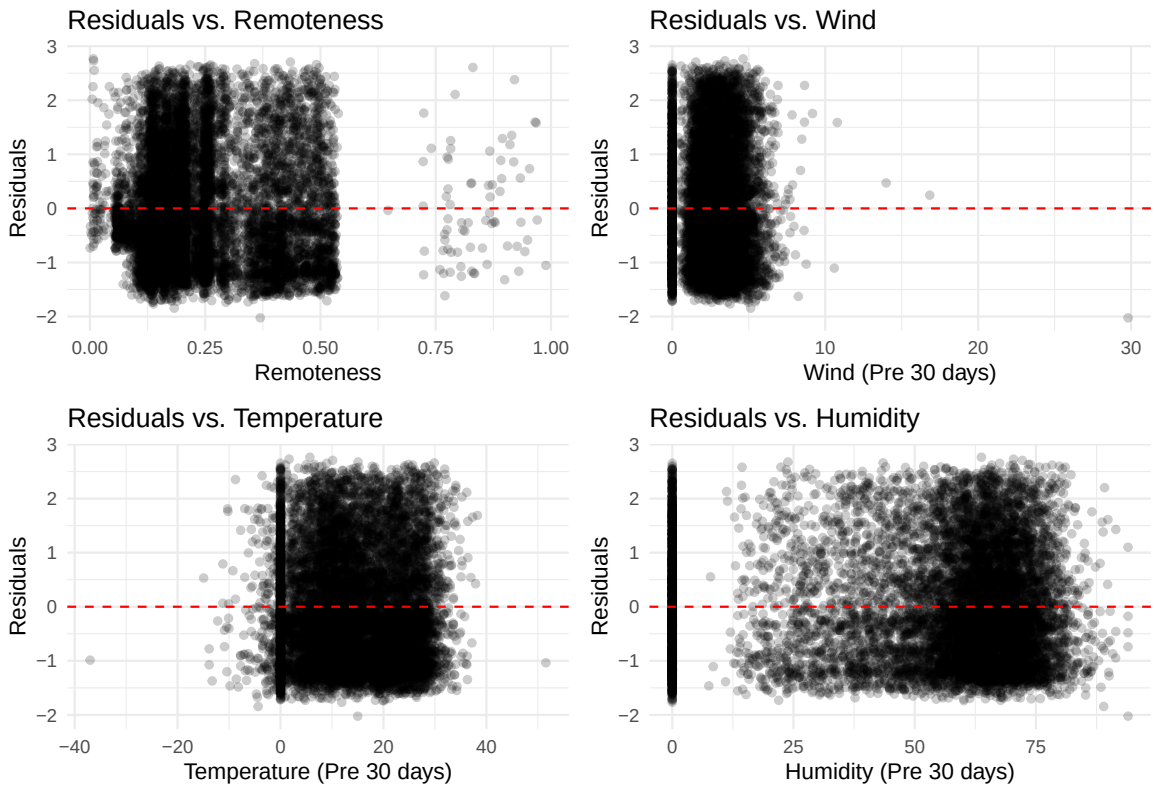

```
# A tibble: 0 x 3
# i 3 variables: .fitted <dbl>, .std.resid <dbl>, .cooksd <dbl>
```

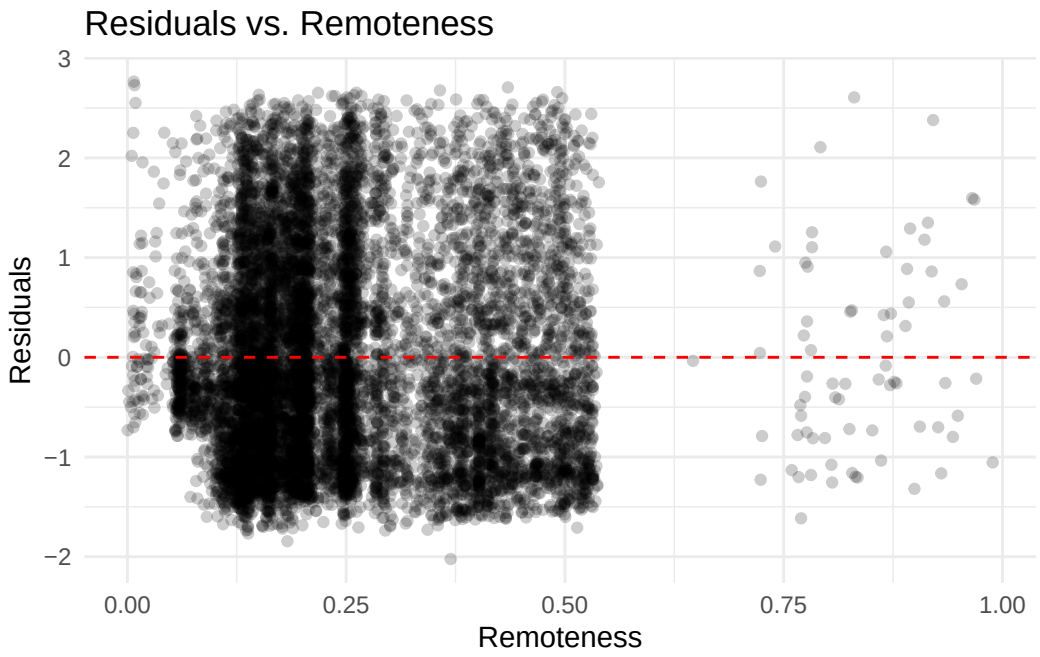
spatial distribution of residuals - independence

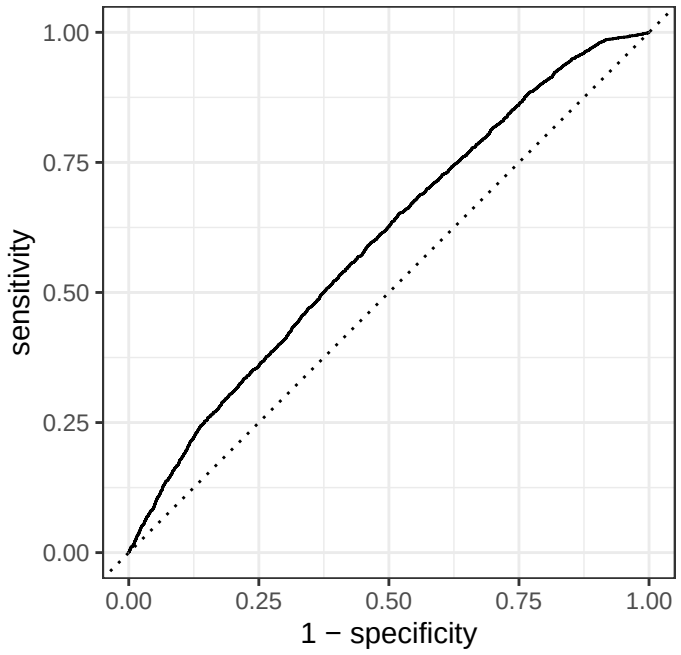

Spatial Distribution of Residuals
Latitude
50
45
40
35
30
25
−120
−100
−80
Longitude
Residual
2
1
0
−1
−2


Residuals vs. Year of Discovery
Residual
3
2
1
0
−1
−2
1990
1995
2000
2005
2010
2015
Discovery Year

Relationship Between Residuals and Predictors








```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.599
```