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 area burned? and (b) What overall factors (including those available after a wildfire) help explain variability in burned area? To answer these research questions, we conducted exploratory data analysis and fitted multiple linear regression models to examine associations between wildfire characteristics and area burned.

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 (such as fire size, cause, discovery and containment dates, and location), 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 World Cities Database. As part of our analysis, we used a subset of these variables, including fire size (measured in

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

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 area burned, since hotter, drier, and windier conditions are likely to be positively associated with burned area, more flammable vegetation and greater remoteness are also likely to be positively associated, and certain fire characteristics may influence fire size, and that (2) variables available after a wildfire occurs help explain additional variability in burned area, since information available after discovery (e.g., underlying cause of the fire and environmental or geographic conditions observed during containment) may be related to 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).

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

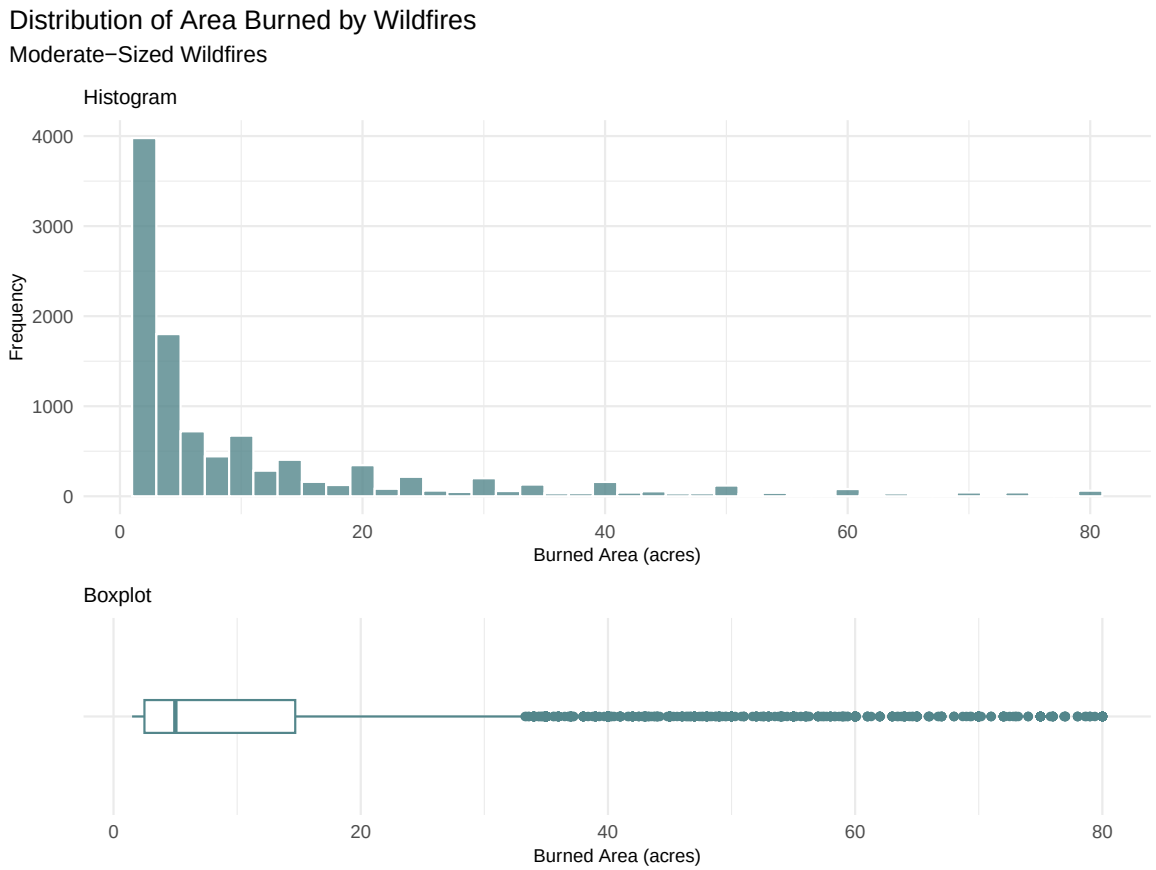
into four broader regions based on U.S. Census classifications (Northeast, Midwest, South, and West), vegetation types into broader environmental categories (e.g., Forest, Shrubland, Grassland), and cause descriptions into broader cause categories (e.g., Natural, Recreational, Infrastructure).

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 (top panel), we can see that the distribution of moderate-sized wildfires is right-skewed and unimodal. Additionally, we can see that 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 in our dataset 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 tells us there is substantial variability in fire size within this range. Additionally, based on the boxplot visualization (bottom panel), we can see that there are many wildfires with burned areas above 30 acres that extend beyond the typical range of values. These wildfires are moderate outliers within this subset.



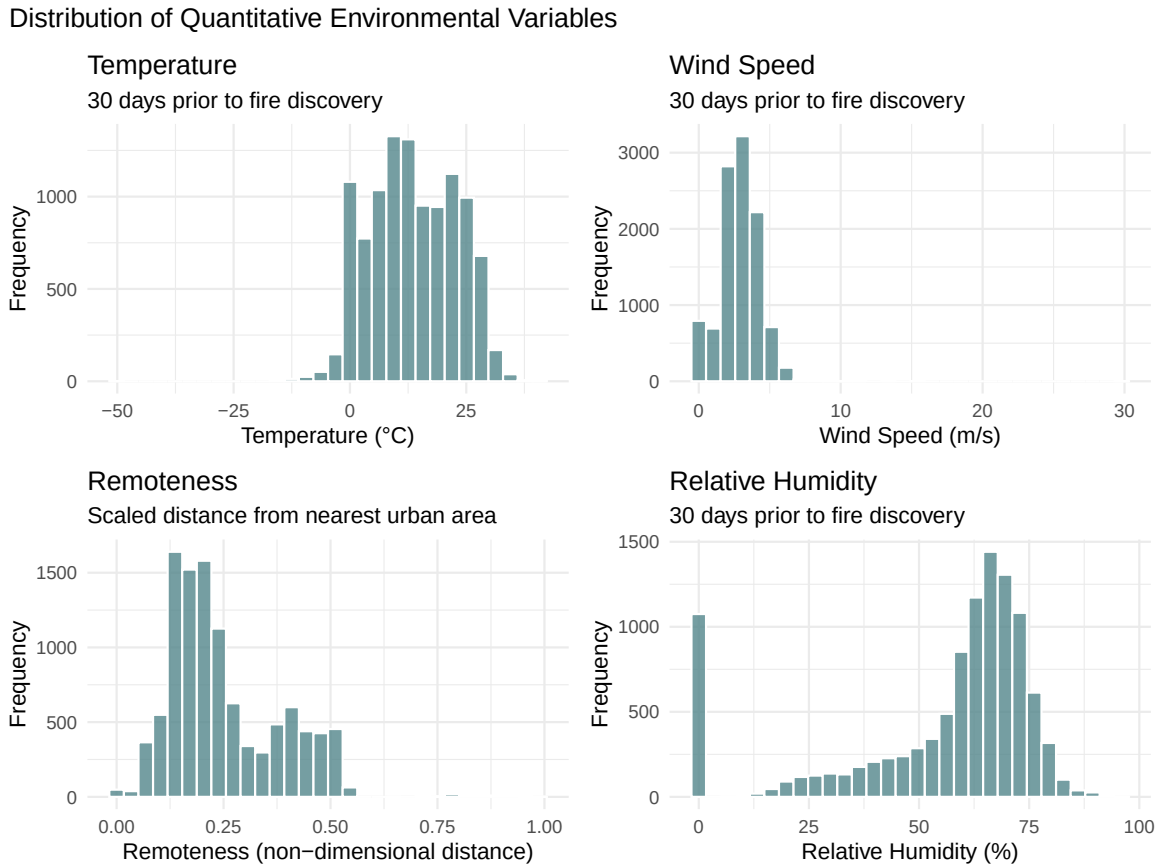
Count	Min	Q1	Median	Mean	Q3	Max	SD
10669	1.5	2.5	5	11.83	14.7	80	15.62

Next, we visualized the distribution of the quantitative variables of interest in our dataset (temperature, relative humidity, and wind speed from 30 days prior to fire discovery, as well as containment time remoteness), along with summary statistics to better understand their scale and variability. Among these predictors, 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, and a range from -49.21°C to 41.68°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, and values ranging from 0 to 0.99. 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, and a range from 0 m/s to 29.80 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%, and values from 0% to 96%. 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.

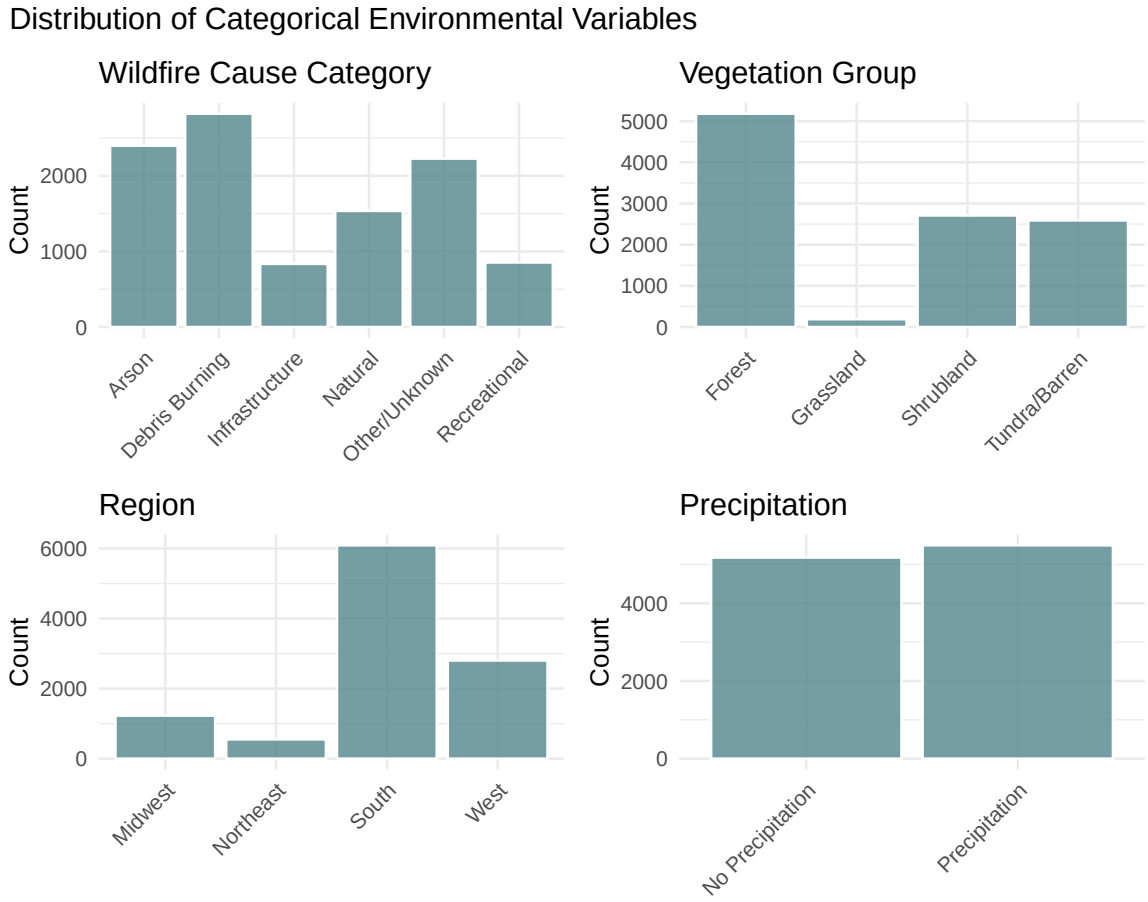


Variable	Min	Q1	Median	Mean	Q3	Max	SD
Temperature	-49.21	6.53	12.97	13.58	21.35	41.68	9.21
Humidity	0.00	48.80	63.52	55.20	69.97	96.00	22.76
Wind	0.00	2.06	2.89	2.87	3.75	29.80	1.40
Remoteness	0.00	0.15	0.21	0.25	0.35	0.99	0.13

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.

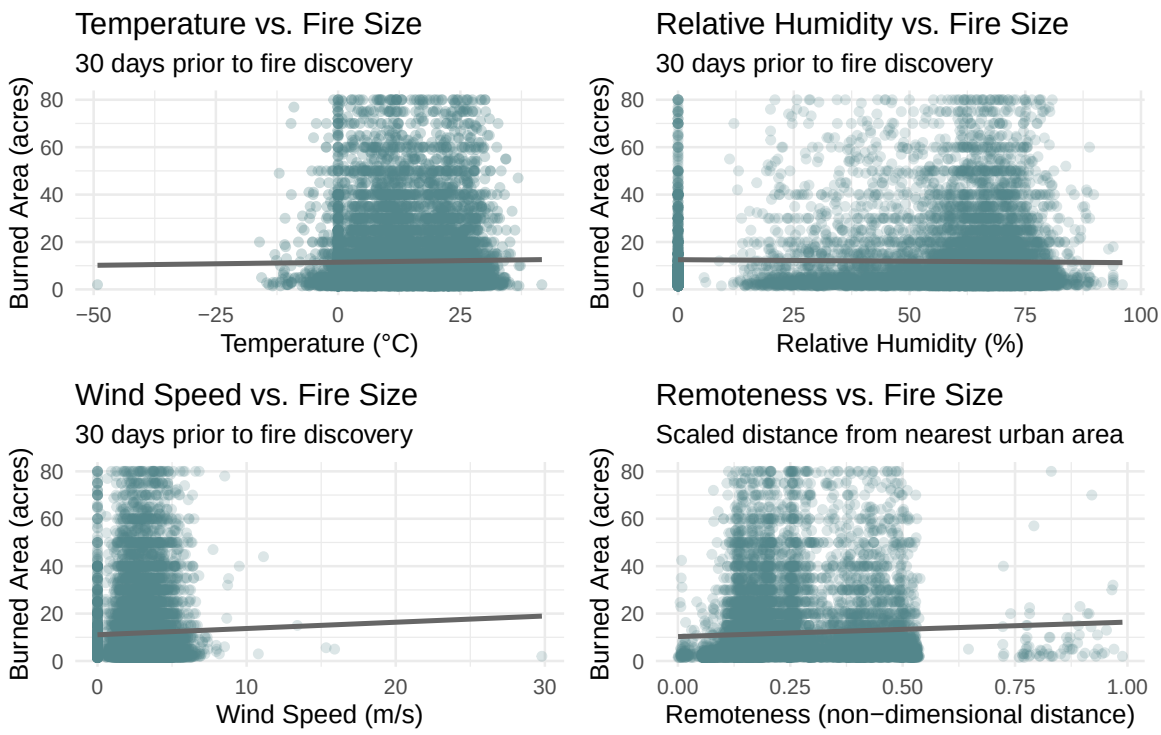
Similarly, the relationship between relative humidity and burned area (top right panel) appears to be very weak and highly nonlinear (with fires of all sizes occurring across almost all humidity levels and the largest fires tending to occur at mid-range humidity values at approximately 25% to 75%), but 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).

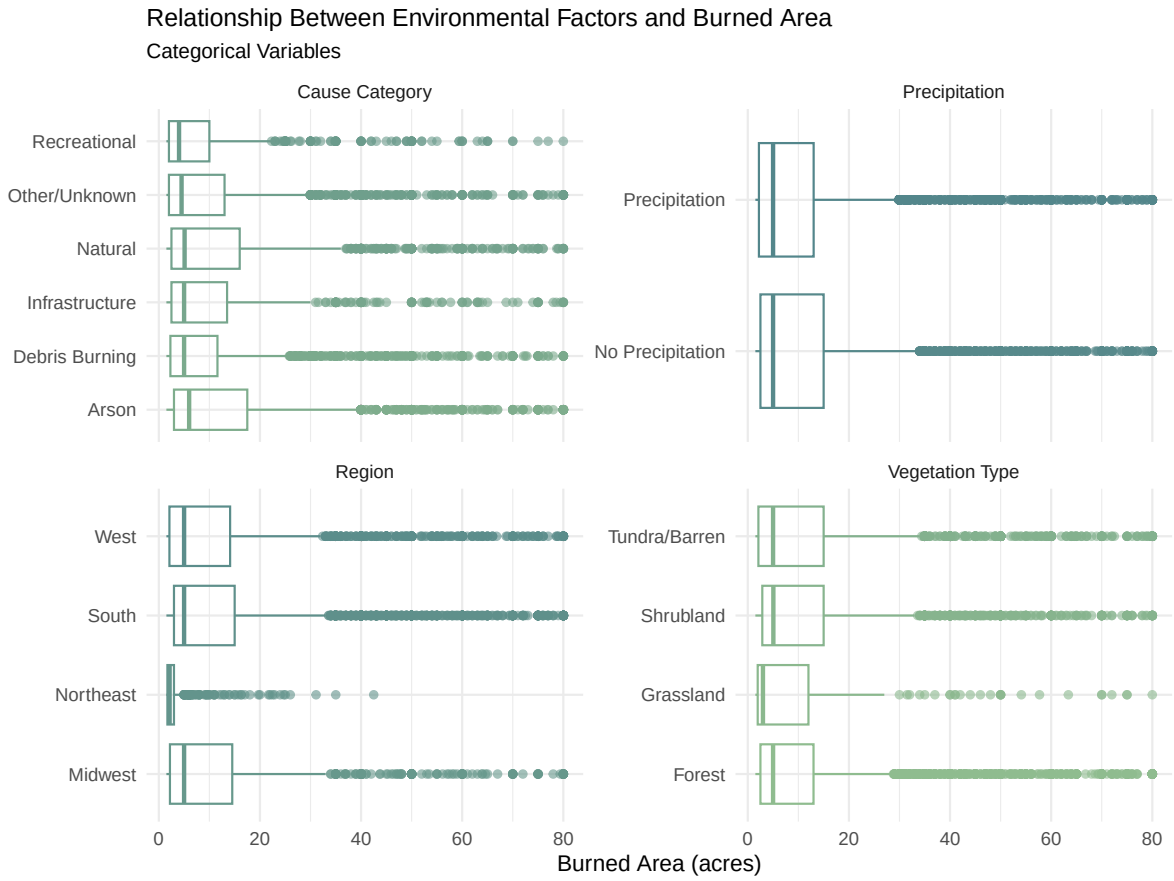
In general, there is a wide vertical spread (i.e., high variability) in the data points around the lines of best fit for all predictors, which suggests that the quantitative predictors in our dataset may have relatively limited predictive power for burned area when considered individually. That being 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,

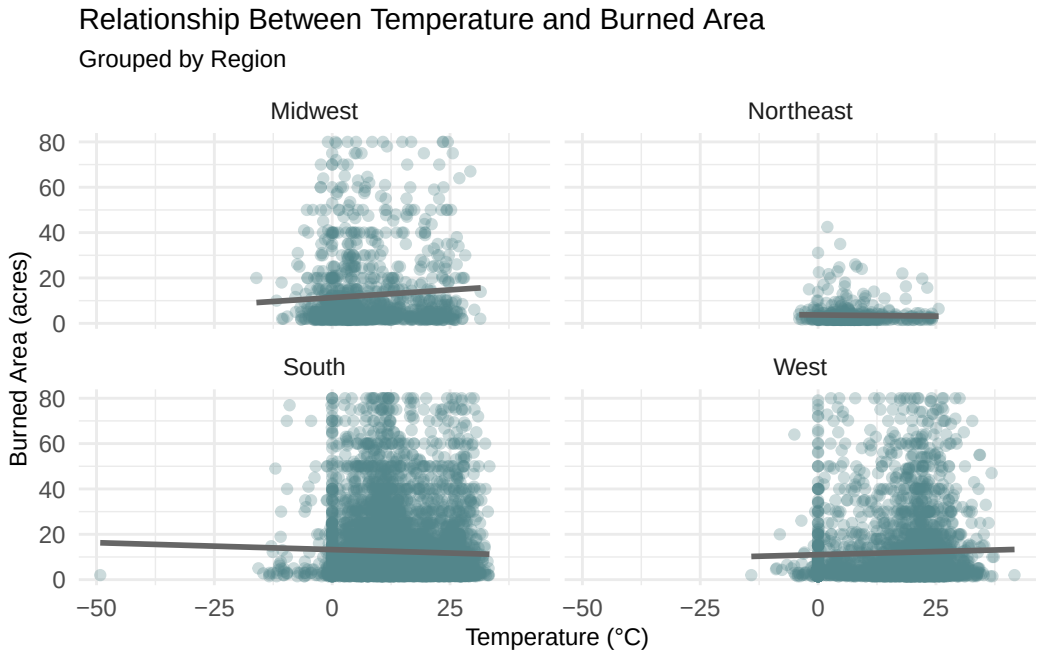
and a large number of outliers beyond the 30-acre mark. For vegetation type, the distribution of fire size was fairly similar across categories, although grasslands had a slightly lower median (4.00 acres) and mean (8.43 acres) compared to other vegetation groups. That being said, the middle 50% of fire 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.



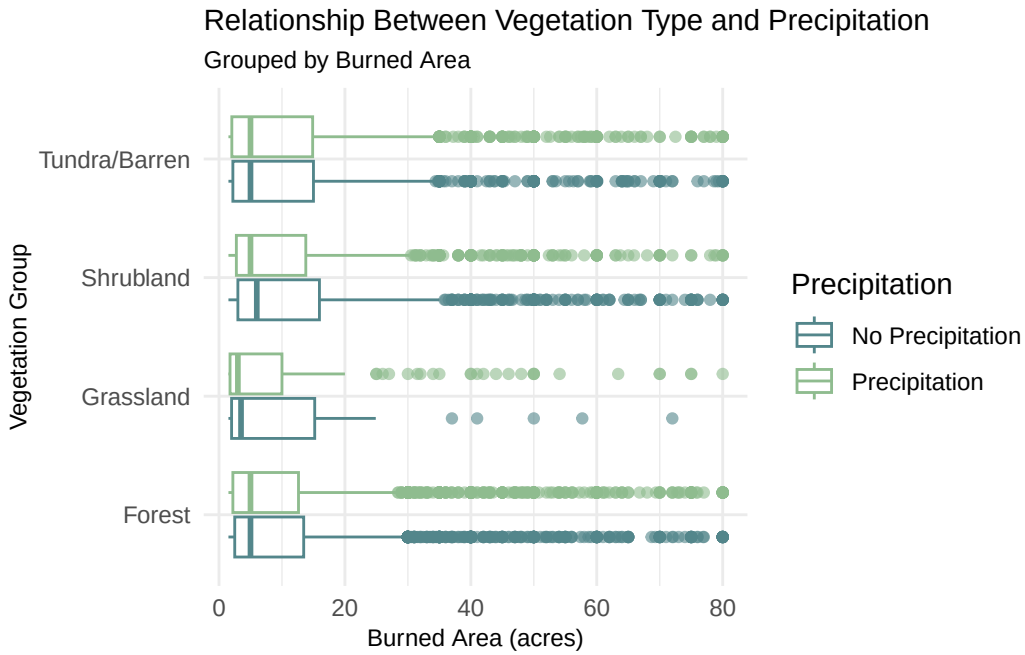
Potential Interaction Effects

Since many of our variables of interest relate to environmental conditions and geographic location, there are many possible interaction effects among our predictors that could help explain variability in acres burned. As part of our analysis, we looked at two potential interaction effects that we thought might be the most meaningful. First, we analyzed whether there was a potential interaction between temperature and region in explaining variability in burned area, since temperature may influence wildfire size differently across regions (e.g., regions like

the Northeast generally experience cooler temperatures, which could lead to different fire behavior compared to hotter, drier regions like the West). As we can see in the visualization below, there does appear to be an interaction effect, since the magnitude and direction of the relationship between temperature and burned area vary across regions. (Specifically, the relationship between temperature and burned area appears to be slightly positive for regions like the Midwest and West, while the trend appears slightly negative for the South and Northeast.) A possible explanation for this could be that, in cooler regions, increases in temperature may support fire spread, while in already hot and humid regions, temperature may not have the same effect due to other limiting factors like vegetation or moisture levels. This interaction would be interesting to look into more formally when modeling fire size to determine whether it significantly improves model performance.



We were also 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 seem to be a very strong interaction effect, and we will further examine this during modeling.



Methodology

Since the main goal of our analysis was to understand how different wildfire-related factors are associated with burned area (a continuous numerical response variable) and how these factors could be used to predict burned area, we decided to use multiple linear regression as the primary modeling approach for this project. Specifically, since our analysis was guided by two different research questions, we fit two different models as part of our analysis: one model using only variables that would be available before or at the time of fire discovery to inform early response efforts (i.e., a predictive model), and another model using all variables of interest, including those that would only be available after a fire was contained (i.e., an explanatory model) to explain variability in burned area that would not be captured by predictive variables alone.

To guide variable selection for the predictive model, we fit a full model including all of the explanatory variables we were interested in. This approach made sense because our dataset contained a limited number of predictors corresponding to information that would already be available prior to fire discovery, so it was practical to include all relevant variables rather than selecting a smaller subset. For the explanatory model, we used a forward selection-based approach based on adjusted R^2 (since adjusted R^2 accounts for model fit while penalizing the inclusion of unnecessary predictors) to build a model that added predictors of interest only when they improved adjusted R^2. The final explanatory model included many of the predictors from the predictive model, along with additional variables that could help explain variability in fire size.

Predictive Modeling

To begin our predictive modeling process, we first fit a model predicting fire size from region, temperature, wind speed, precipitation, and relative humidity (all measured thirty days before fire discovery), along with remoteness, vegetation group, and interaction terms between region and temperature and between vegetation group and precipitation. After fitting this model, we conducted model diagnostics to evaluate whether key assumptions of linear regression were reasonably satisfied (i.e., linearity, normality, constant variance, and independence). Based on our analysis, we found that the linearity, normality, and independence conditions were reasonably satisfied, but the constant variance condition was not satisfied. As a result, we applied a variance-stabilizing (i.e., logarithmic) transformation to the response variable to address non-constant variance.

Additionally, when examining the distribution of residuals versus fitted values for this initial model, we observed a clustering of observations into two groups (one group with fitted values less than or equal to 5, and another with fitted values between 5 and 20). Through further investigation, we realized that this clustering was in part 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 model fit, but we ultimately decided to proceed with a single model including all regions because the improvement in fit was very limited (and for the sake of consistency in model 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, in this model, we found evidence of collinearity between region and temperature (with the main effects and interaction terms for these variables having VIFs greater than 10). To address this, instead of including an explicit interaction term, we centered temperature within each region and used this centered variable in the model to account for within-region temperature variation while reducing collinearity. To note, we did not find concerning evidence of collinearity between vegetation group and precipitation, so the interaction between these variables was retained in the model.

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df  logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>   <dbl>  <dbl>  <dbl>
1    0.0445        0.0436  1.06      45.2 3.34e-97    11 -15723. 31472. 31567.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
regionNortheast                   regionSouth
       1.903576                      3.676861
```

```
                  regionWest                    Temp_pre_30
                    5.254582                       1.229149
                  remoteness     Vegetation_groupGrassland
                    5.159586                       1.076820
Vegetation_groupShrubland Vegetation_groupTundra/Barren
                    1.337842                       1.217749
                Wind_pre_30                    Prec_pre_30
                    1.346729                       1.203015
                 Hum_pre_30
                    1.563948
```

Explanatory 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

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
1    0.0587        0.0570  1.05      33.2 8.41e-124    20 -15643. 31331. 31491.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
                        regionNortheast
                               1.943275
                            regionSouth
                               3.645250
                             regionWest
                               5.202818
                Temp_pre_30_region_centered
                               1.177316
                             remoteness
                               5.215893
                  Vegetation_groupGrassland
                               4.753863
                  Vegetation_groupShrubland
                               2.382362
              Vegetation_groupTundra/Barren
                               2.647680
```

```
                                  Wind_pre_30
                                     1.355327
                                  Prec_pre_30
                                     2.254819
                                   Hum_pre_30
                                     1.594550
                               putout_time_num
                                     1.036432
                  cause_categoryDebris Burning
                                     1.633041
                  cause_categoryInfrastructure
                                     1.295402
                         cause_categoryNatural
                                     1.830561
                   cause_categoryOther/Unknown
                                     1.643019
                    cause_categoryRecreational
                                     1.293491
      Vegetation_groupGrassland:Prec_pre_30
                                     4.794561
      Vegetation_groupShrubland:Prec_pre_30
                                     2.607795
Vegetation_groupTundra/Barren:Prec_pre_30
                                     3.045045
```

Results

We fit four linear regression models to predict `log(fire_size)` using combinations of environmental and fire-related predictors. Two models included precipitation as a continuous predictor (`log(Prec_pre_15 + 0.001)`), and two models used a binary indicator variable representing the presence of precipitation. Each modeling approach included a main-effects model and an interaction model with `Vegetation`.

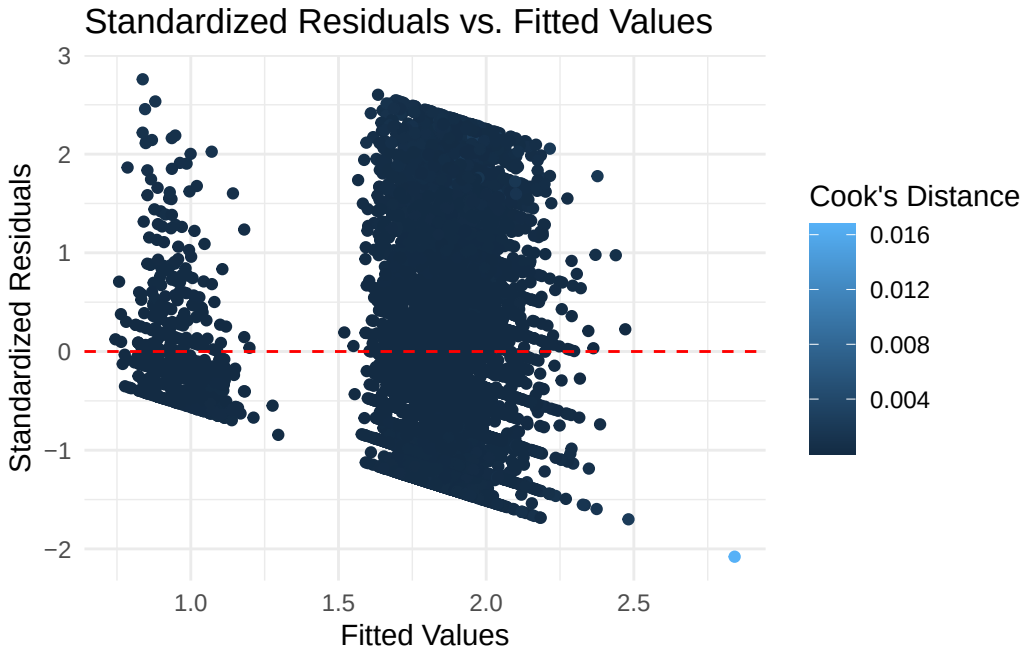
The adjusted R^2 values were low across all models. The main-effects model with continuous precipitation (`fire_main_fit`) had an adjusted R^2 of 0.037, and the interaction model (`fire_int_fit`) had an adjusted R^2 of 0.041. The models using the binary precipitation indicator produced similar results: 0.037 for the main-effects model (`fire_main_bin`) and 0.039 for the interaction model (`fire_int_bin`).

We evaluated prediction error using root mean squared error (RMSE) on the log scale. The RMSE was 1.059 for both `fire_main_fit` and `fire_main_bin`, and slightly lower for the interaction models: 1.057 for `fire_int_fit` and 1.058 for `fire_int_bin`. These values indicate that the average prediction error on the log-transformed scale was around 1.06. On the

original scale, this corresponds to a multiplicative prediction error of approximately 2.88 (i.e., $\exp(1.06) \approx 2.88$).

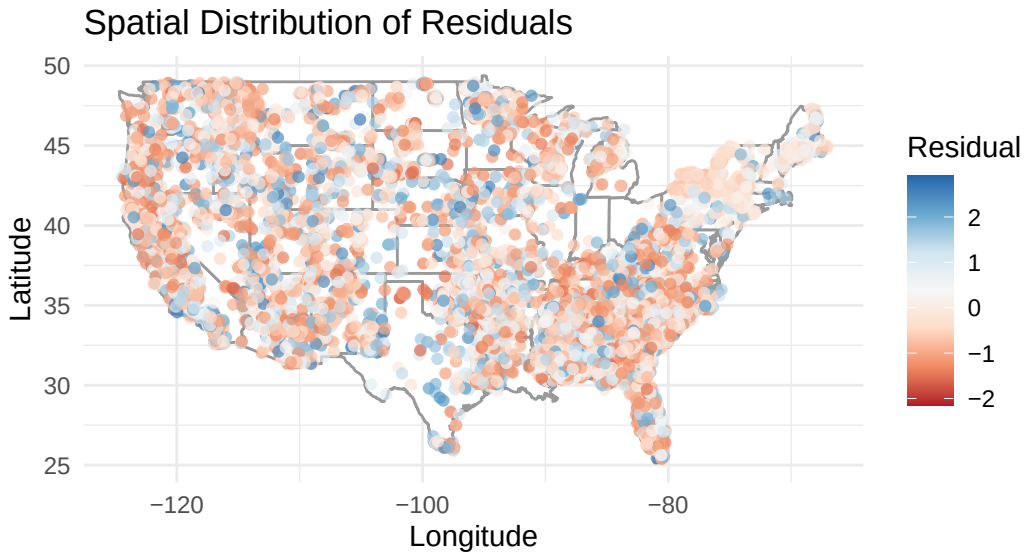
Overall, there were minimal differences in adjusted R^2 and RMSE between the models using continuous precipitation and those using a binary indicator. The addition of interaction terms with `Vegetation` slightly increased model complexity without meaningfully improving predictive performance.

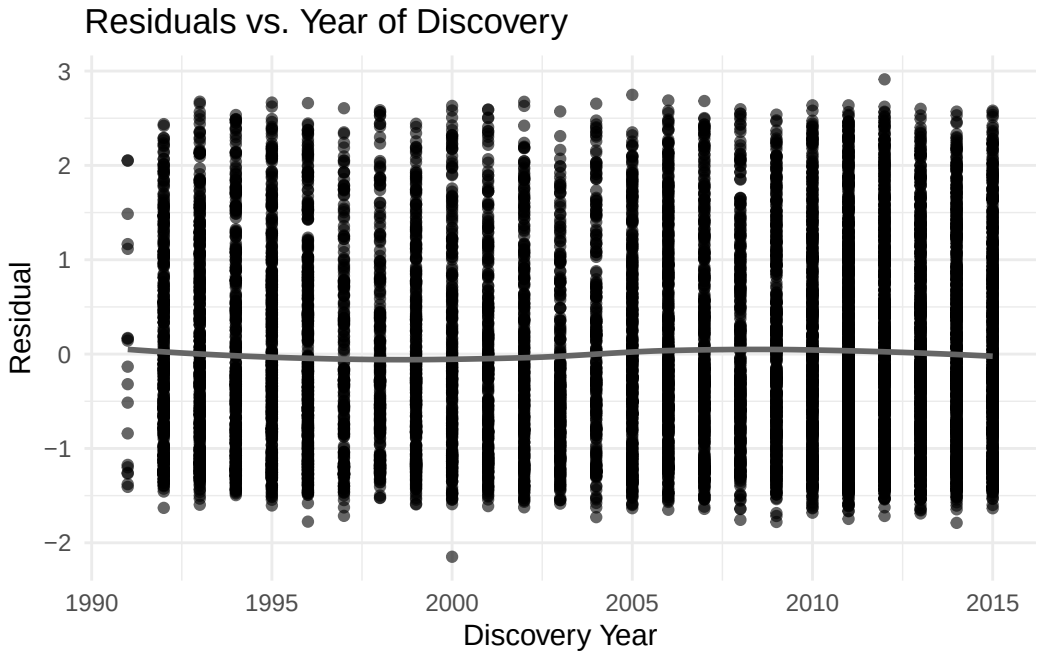
Appendix




```
# 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
Longitude
Residual
50
45
40
35
30
25
−120
−100
−80
2
1
0
−1
−2


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

Relationship Between Residuals and Predictors