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

0.1 Introduction and Data

0.1.0.1 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 destruction, environmental degradation, and human health impacts). After learning about recent wildfire incidents in California and western North Carolina, we became interested in better understanding the factors that contribute to the likelihood that a wildfire burns a greater-than-typical area, and the factors that help explain variability in the continuous outcome of burned area. 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).

0.1.0.2 Hypotheses

We hypothesize that (1) environmental conditions (e.g., temperature, humidity, wind, precipitation), geographic characteristics (e.g., remoteness, region), and vegetation type are associated with the likelihood that a wildfire burns a greater-than-typical area (with hotter, drier, and windier conditions, greater remoteness, and more flammable vegetation expected to increase the likelihood of larger fires), and (2) that both pre-discovery factors (environmental conditions, geographic characteristics, and vegetation type) and fire cause help explain variability in burned area, with human-related causes, hotter, drier, and windier conditions, greater remoteness, and flammable vegetation expected to be associated with larger burned areas.

0.2 Exploratory Data Analysis

0.2.1 Data Cleaning

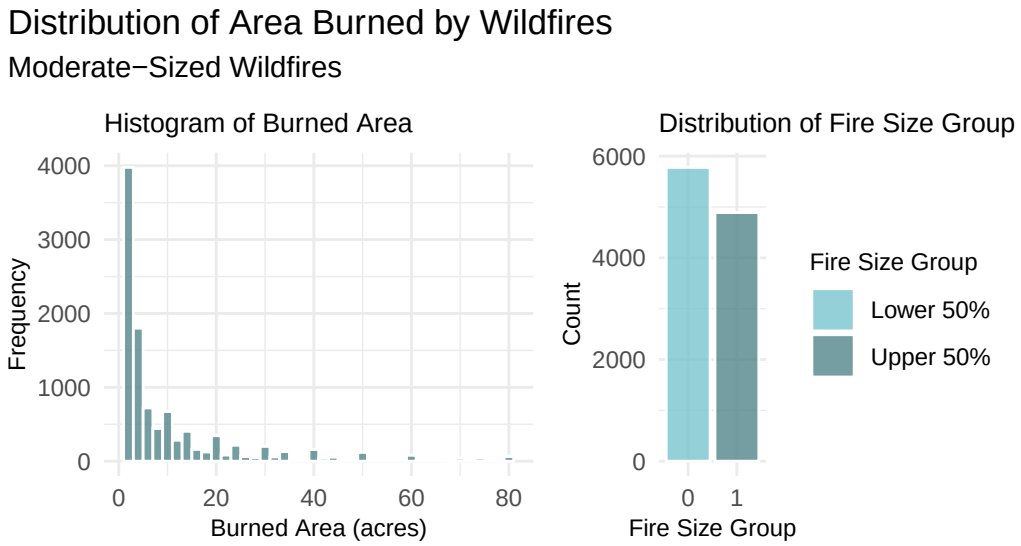
Before conducting our analysis, we applied many data cleaning steps to prepare our dataset for modeling and interpretation. One of the major decisions we made as part of our data cleaning process was to filter our response variable, acres burned, since the majority of observations in our dataset (over 13,000) recorded fires that burned one acre or less of land. Thus, we decided to focus only on the interquartile range (middle 50%) of wildfires by acres burned, since 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 to restrict our analysis to a more practical range of fire sizes. 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 and consolidating variables to improve model interpretability and analysis. Since most precipitation variables were heavily left-skewed (i.e., most fires occurred without recent precipitation), we transformed all precipitation variables into binary indicators (0 = no precipitation, 1 = precipitation > 0). Additionally, we grouped states into four broader regions (Northeast, Midwest, South, and West), vegetation

types into environmental categories (e.g., Forest, Shrubland, Grassland), and fire causes into cause categories (e.g., Natural, Recreational, Infrastructure). We dropped all observations with missing values for variables of interest and converted categorical variables (e.g., vegetation type, fire cause, region) into factor variables to support modeling.

0.2.1.1 Univariate EDA

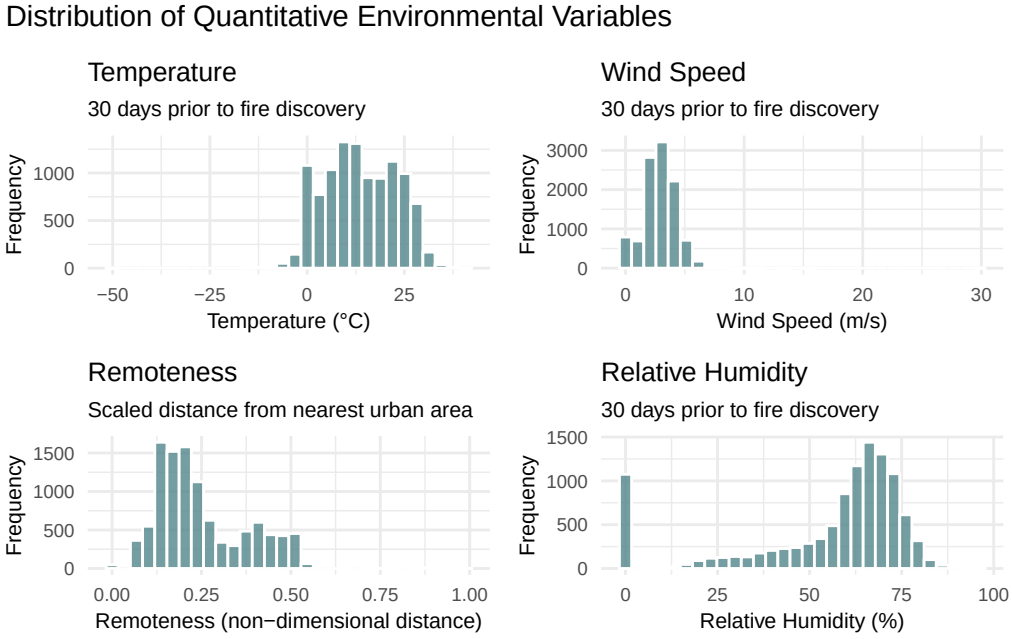
As part of our EDA, we first examined the distribution of the response variable, fire size, to better understand its scale and variability. As mentioned previously, we focused on the middle 50% of wildfires by acres burned, which represent moderate-sized fires that are more likely to be responsive to 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.



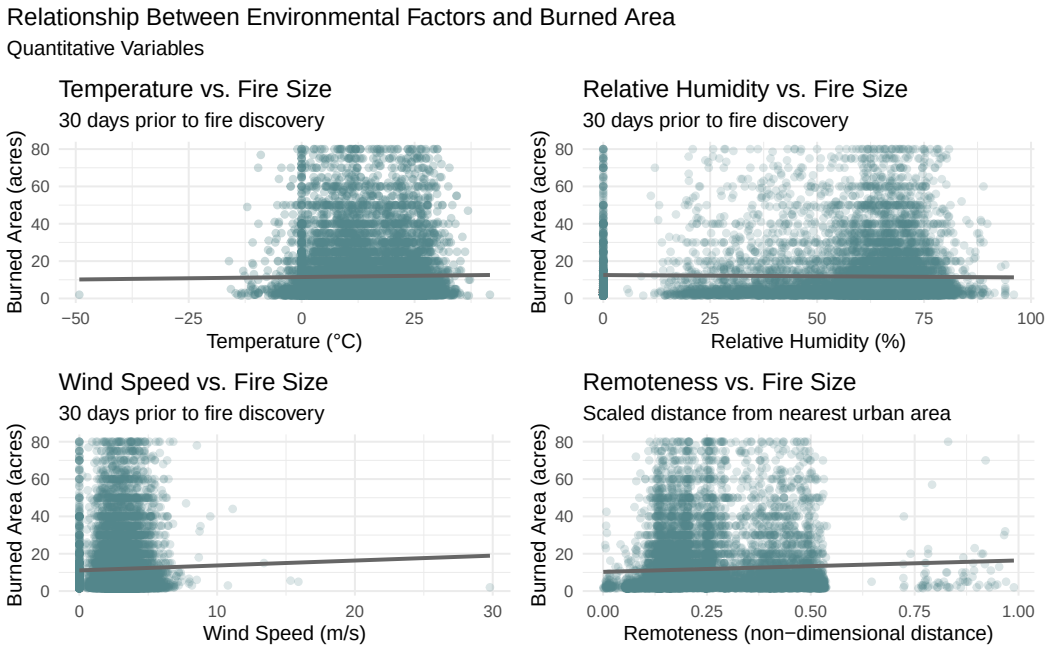
Variable	Min	Q1	Median	Mean	Q3	Max	SD
Temperature	-49.21	6.53	12.97	13.58	21.35	41.68	9.21
Humidity	0.00	48.80	63.52	55.20	69.97	96.00	22.76
Wind	0.00	2.06	2.89	2.87	3.75	29.80	1.40
Remoteness	0.00	0.15	0.21	0.25	0.35	0.99	0.13

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), which can be found in the Appendix.

0.2.1.2 Bivariate EDA

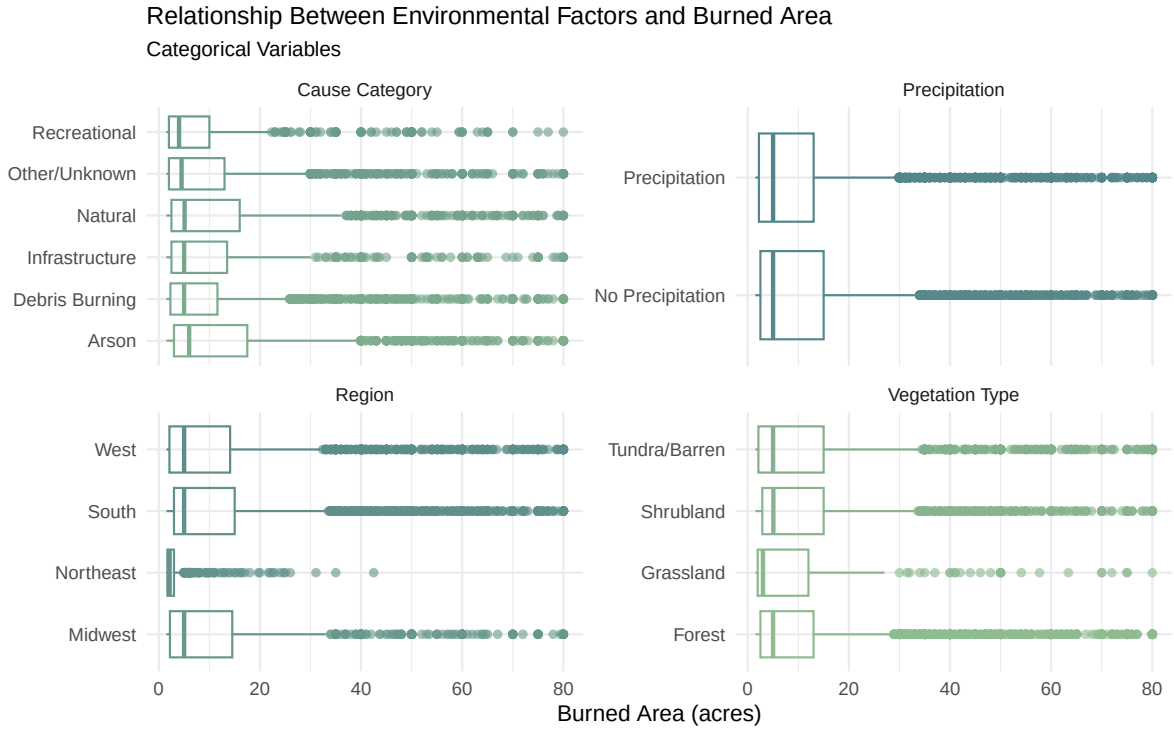
As the next step in our analysis, we examined the relationship between each of our quantitative predictors and fire size to better understand their associations. In the scatterplot of burned area versus temperature (top left panel), the relationship appears very weak and slightly positive (i.e., larger fires may be slightly more likely at higher temperatures), but nonlinear, since most fires were small across the temperature range, and the largest fires (over 60 acres) tended to occur between 5°C and 20°C. The scatterplot of relative humidity versus burned area (top right panel) also shows a very weak, highly nonlinear, and slightly negative relationship, with fires of all sizes occurring across the full humidity range and the largest fires more common between 25% and 75% humidity. For wind speed (bottom left panel), the relationship appears very weak, slightly positive, and nonlinear; most fires occurred at lower wind speeds, and most

larger fires (over 40 acres) were concentrated below 5 m/s, suggesting that higher wind speeds were not consistently associated with larger burned areas. In the scatterplot of remoteness versus burned area (bottom right panel), we observe a weak, slightly positive, and nonlinear relationship, with the largest fires occurring between remoteness values of 0.25 and 0.50 and relatively few large fires at very high remoteness values above 0.75. Overall, the wide vertical spread of data points around the lines of best fit for all predictors suggests that these quantitative predictors have limited predictive power for burned area when considered individually, which is expected given that wildfire size is influenced by many factors beyond the variables 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, 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 between observations corresponding to no precipitation 30 days before discovery and observations corresponding to precipitation (top right panel), with both groups showing medians of 5.00 acres and means of 12.32 acres (no precipitation) and 11.38 acres (precipitation), although there was slightly more spread in the middle 50% of fires with no 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 many outliers beyond the 30-acre mark. For vegetation type, distributions were fairly similar across categories, although grasslands had a slightly lower median (4.00 acres) and mean (8.43 acres) compared to other vegetation groups, with somewhat fewer extreme outliers. Overall, region and vegetation group seem to explain more variability in fire size compared to cause category or precipitation, and could be useful predictors in modeling fire size.



0.2.1.3 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. Through our EDA, we observed that acres burned did not seem to vary much across different precipitation groups within each vegetation type. (A more detailed analysis of this pattern is provided in the Appendix.)

0.3 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 to predict whether a wildfire burns a greater-than-typical area (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. (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.)

0.3.0.1 Multiple Linear Regression

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 including all predictors of interest (each contributed to an increase in adjusted R^2). Our initial multiple linear regression predicted burned area based on region, cause category, remoteness, vegetation group, temperature, precipitation, wind speed, relative humidity (15 days before discovery), containment time, and an interaction between vegetation group and precipitation. We mean-centered the quantitative predictors to improve coefficient interpretability. After fitting this model, we conducted diagnostics to assess linearity, normality, constant variance, and independence. Normality, linearity, and independence appeared reasonably satisfied, but constant variance was violated, and there was some evidence of potential non-linearity. To address constant variance, we applied a variance-stabilizing (logarithmic) transformation to the fire size variable. For linearity, we examined residuals against each predictor and found no clear nonlinear patterns, so we did not transform the predictors. When examining residuals versus fitted values, we observed two distinct clusters (one with fitted values 5 and another between 5 and 20), partly explained by differences across region (specifically, smaller fires in the Northeast). We explored fitting separate models for the Northeast and other regions but ultimately proceeded with a single model because the improvement in fit was minimal. We also checked for influential points using Cook's Distance ($D_i > 0.5$) and found none. Lastly, we found collinearity between levels corresponding to the Southern and Western regions; to address this, we combined these levels into a single group, which helped reduce collinearity without compromising interpretability. A detailed assessment of model diagnostics (including residual plots, VIFs, and checks for constant variance, linearity, and influential points) can be found in the Appendix.

0.3.0.2 Logistic Regression

To guide variable selection for the predictive model, we fit a full model including all variables of interest. This approach was appropriate because our dataset contained a limited number of predictors corresponding to information available prior to fire discovery, making it practical to include all relevant variables rather than selecting a smaller subset. Specifically, our initial logistic regression model predicted the likelihood that a wildfire would burn a greater-than-typical area based on temperature, humidity, wind speed, precipitation, vegetation type, the interaction between vegetation type and precipitation, remoteness, and region. After fitting the model, we assessed whether the key assumptions for logistic regression (i.e., linearity, randomness, and independence) were satisfied. Wildfire records in our dataset can reasonably be treated as independent because, in our assessment of the independence condition for linear regression, we observed no spatial correlations between residuals and geographic location. The randomness assumption was satisfied because the wildfire data were compiled from a large national database covering a wide range of fires across the United States between 1992 and 2015, making it reasonable to assume the dataset is representative of the broader wildfire population. Lastly, the linearity condition appeared reasonably satisfied, as empirical logit plots for the quantitative predictors did not display strong deviations from linearity, and the assumption of linearity in the log-odds was met. A detailed assessment of model diagnostics (empirical logit plots, assessment of logistic regression conditions) can be found in the Appendix.

0.4 Results

0.4.0.1 Multiple Linear Regression

The output of our final multiple linear regression model is displayed below:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.997	0.045	44.329	0.000	1.909	2.085
regionNortheast	-0.942	0.059	-16.003	0.000	-1.057	-0.826
regionSouth_West	0.075	0.036	2.072	0.038	0.004	0.145
cause_categoryDebris Burning	-0.255	0.029	-8.676	0.000	-0.313	-0.198
cause_categoryInfrastructure	-0.113	0.043	-2.630	0.009	-0.198	-0.029
cause_categoryNatural	-0.050	0.038	-1.308	0.191	-0.125	0.025
cause_categoryOther/Unknown	-0.160	0.032	-5.014	0.000	-0.222	-0.097
cause_categoryRecreational	-0.370	0.042	-8.701	0.000	-0.453	-0.286
remoteness	-0.310	0.099	-3.118	0.002	-0.504	-0.115
Vegetation_groupGrassland	0.189	0.089	2.130	0.033	0.015	0.362
Vegetation_groupShrubland	0.065	0.027	2.432	0.015	0.013	0.118
Vegetation_groupTundra/Barren	0.073	0.026	2.767	0.006	0.021	0.125
Prec_pre_15_cent	0.025	0.031	0.793	0.428	-0.036	0.086
Wind_pre_15_cent	0.032	0.008	3.929	0.000	0.016	0.047
Temp_pre_15_cent	-0.003	0.001	-2.643	0.008	-0.006	-0.001
Hum_pre_15_cent	-0.001	0.001	-1.131	0.258	-0.002	0.000
putout_time_num	0.007	0.001	5.183	0.000	0.004	0.010
Vegetation_groupGrassland:Prec_pre_15_cent	-0.203	0.168	-1.206	0.228	-0.532	0.127
Vegetation_groupShrubland:Prec_pre_15_cent	-0.141	0.051	-2.770	0.006	-0.241	-0.041
Vegetation_groupTundra/Barren:Prec_pre_15_cent	0.018	0.051	0.349	0.727	-0.082	0.118

R^2	Adjusted R^2
0.052	0.05

Our final multiple linear regression model predicting log-transformed fire size achieved an R^2 of 0.052 and an adjusted R^2 of 0.050. This indicates that the model explains approximately 5.0% of the variability in log-transformed burned area among moderate-sized wildfires in the dataset. While low, this amount of explained variance is expected given the variability in wildfire behavior and the limited set of predictors available.

We examined several predictors in our model and found that region, cause category, remoteness, vegetation group, wind speed 15 days prior to discovery, temperature 15 days prior to discovery, putout time, and the interaction between shrubland vegetation and precipitation were significant predictors of log-transformed fire size ($p < 0.05$). In contrast, relative humidity 15 days prior ($p > 0.05$), precipitation 15 days prior ($p > 0.05$), and the interaction terms between grassland or tundra/barren vegetation and precipitation ($p > 0.05$) were not significant; this result is consistent with our EDA findings (i.e., differences across region and vegetation type and the effect of wind speed were more apparent, while precipitation and humidity did not show strong relationships with fire size).

The intercept of the model is estimated at 2.00, indicating that the expected fire size is approximately $e^{2.00} \approx 7.39$ acres when all predictors are at their reference or mean-centered values (i.e., for a wildfire in the Midwest region, caused by arson, in a forested area, under average environmental conditions — mean temperature = 13.8°C, mean relative humidity = 52.0%, mean wind speed = 2.77 m/s, mean remoteness = 0.252 — and with putout time equal to 0). The coefficient of `regionNortheast` tells us that fires in the Northeast are expected to burn $e^{-0.942} \approx 0.39$ times the acres compared to fires in the Midwest (reference level), holding all else constant. The coefficient of remoteness tells us that for each one-unit increase in remoteness (on a 0 to 1 scale), the expected fire size decreases by a factor of $e^{-0.310} \approx 0.73$, holding all else constant. The coefficient of wind speed 15 days prior tells us that for each additional meter per second, the expected fire size increases by a factor of $e^{0.032} \approx 1.03$, holding all else constant. The coefficient of the interaction between shrubland vegetation and precipitation tells us that, in shrubland environments, fires following precipitation are expected to burn $e^{-0.141} \approx 0.87$ times the acres compared to fires without prior precipitation, after controlling for other predictors.

0.4.0.2 Logistic Regression

The output of our final logistic regression model is displayed below:

Our final logistic regression model predicting whether a wildfire burns a greater-than-typical area achieved an AUC of 0.569, indicating relatively limited ability to distinguish between smaller and larger moderate-sized fires based on the predictors included (AUC value and corresponding ROC curve are provided in the Appendix). After fitting the full model with

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.220	0.098	-2.257	0.024	-0.412	-0.029
regionNortheast	-1.829	0.148	-12.363	0.000	-2.125	-1.545
regionSouth_West	0.182	0.069	2.642	0.008	0.047	0.316
remoteness	-0.414	0.182	-2.271	0.023	-0.772	-0.057
Vegetation_groupGrassland	0.635	0.273	2.328	0.020	0.104	1.179
Vegetation_groupShrubland	0.209	0.064	3.246	0.001	0.083	0.335
Vegetation_groupTundra/Barren	0.093	0.069	1.344	0.179	-0.043	0.228
Prec_pre_15	0.032	0.060	0.533	0.594	-0.086	0.150
Wind_pre_15	0.036	0.016	2.291	0.022	0.005	0.066
Temp_pre_15	-0.005	0.002	-1.988	0.047	-0.009	0.000
Hum_pre_15	-0.001	0.001	-0.632	0.527	-0.003	0.001
Vegetation_groupGrassland:Prec_pre_15	-0.641	0.347	-1.849	0.064	-1.328	0.034
Vegetation_groupShrubland:Prec_pre_15	-0.185	0.097	-1.897	0.058	-0.375	0.006
Vegetation_groupTundra/Barren:Prec_pre_15	0.012	0.099	0.121	0.904	-0.182	0.206

an interaction between vegetation group and precipitation, we conducted a drop-in-deviance test to assess whether the interaction terms were jointly useful. Since the p-value for the test was 0.013, we retained the interaction between vegetation group and precipitation (full test results are provided in the Appendix); while the individual interaction terms were not significant at the 0.05 level, the combined contribution significantly improved model fit. We examined several predictors related to environmental conditions and geography and found that region, remoteness, vegetation group (grassland and shrubland), wind speed 15 days prior, and temperature 15 days prior were significant predictors of fire size classification ($p < 0.05$). In contrast, relative humidity 15 days prior ($p > 0.05$), precipitation 15 days prior ($p > 0.05$), tundra/barren vegetation group ($p > 0.05$), and the interaction terms between vegetation type and precipitation ($p > 0.05$) were not significant.

The intercept of the model is estimated at -0.220, indicating that the expected odds of a wildfire falling into the larger half of moderate-sized fire sizes is $e^{-0.220} \approx 0.80$ when all predictors are at their reference or mean-centered values (i.e., for a wildfire in the Midwest region, in a forested area, with no precipitation 15 days prior, and under average environmental conditions — mean temperature = 13.8°C, mean relative humidity = 52.0%, mean wind speed = 2.77 m/s, mean remoteness = 0.252). The coefficient of `regionNortheast` tells us that fires in the Northeast are expected to have $e^{-1.829} \approx 0.16$ times the odds of falling into the upper 50% compared to fires in the Midwest (reference level), holding all else constant. The coefficient of `regionSouth_West` tells us that fires in the South or West are expected to have $e^{0.182} \approx 1.20$ times the odds compared to fires in the Midwest, holding all else constant. The coefficient of remoteness tells us that for each one-unit increase in remoteness (on a 0 to 1 scale), the odds of a fire falling into the upper 50% decrease by a factor of $e^{-0.414} \approx 0.66$, holding all else constant. The coefficient of vegetation group (grassland) tells us that fires in grassland areas are expected to have $e^{0.635} \approx 1.89$ times the odds of falling into the upper 50% compared to fires in forested areas, holding all else constant. The coefficient of wind speed 15 days prior tells us that for each additional meter per second, the odds of falling into the upper 50% increase by a factor of $e^{0.036} \approx 1.04$, holding all else constant. The coefficient of temperature 15 days prior tells us that for each one-degree Celsius increase, the odds decrease

by a factor of $e^{-0.005} \approx 0.995$, holding all else constant.

0.5 Discussion + Conclusion

Our analysis used both multiple linear regression and logistic regression to understand factors associated with wildfire size in the United States. The multiple linear regression model, which predicted log-transformed fire size, found that physical factors like vegetation type, wind speed, precipitation, temperature, and humidity and other factors including region, remoteness, fire cause, and containment time were significant predictors, though the model explained only about 5% of the variability in fire size. The logistic regression model, which predicted the likelihood that a wildfire burned a greater-than-typical area, used similar predictors except the cause of the fire, but did not have a very effective classification ability (AUC = 0.569). In both models, fires in shrubland and grassland were associated with larger burned areas, while fires in the Northeast tended to be smaller. Greater wind speeds slightly increased fire size, and greater remoteness was associated with smaller fires.

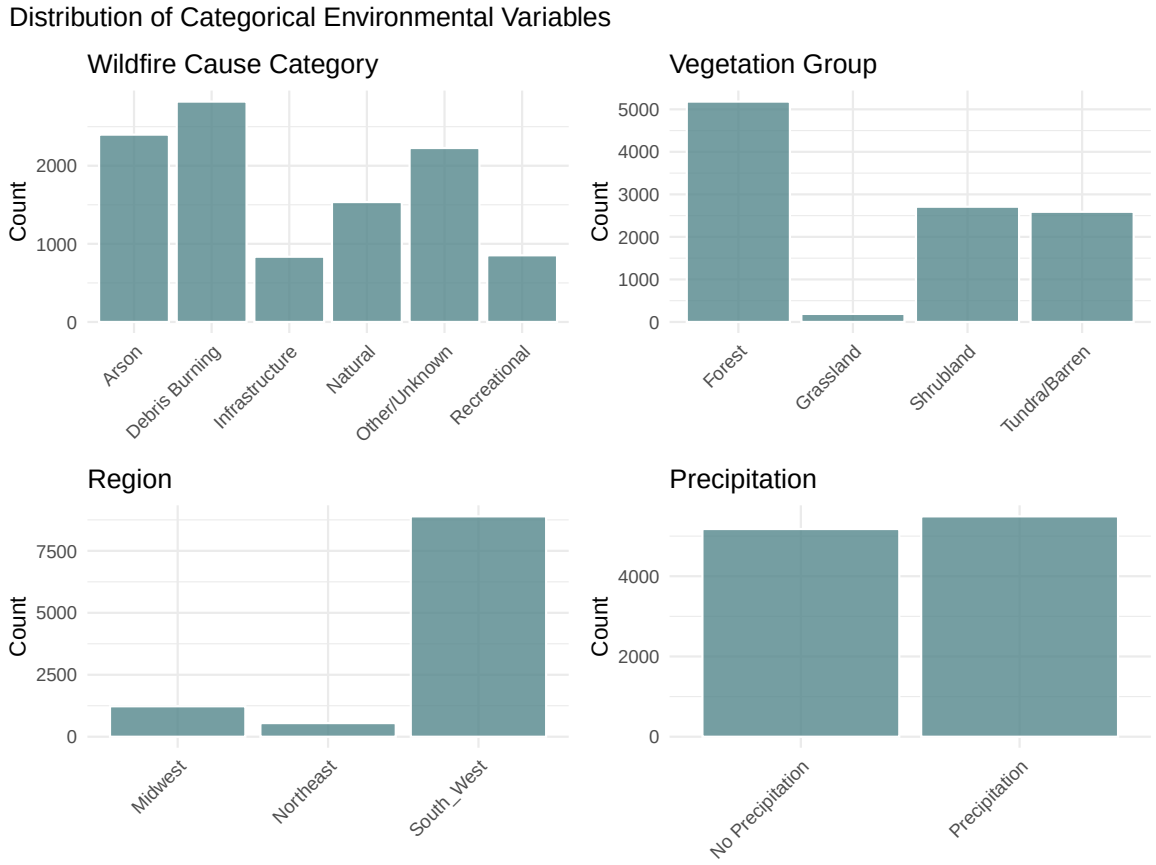
There were several limitations to our analysis. While the number of fires in the dataset was massive, comprehensive data on those fires was often limited, which likely contributed to the low explanatory power of our models. Because the data spanned more than two decades, changes over time in climate patterns, land use, and fire management practices could have introduced variability that our models did not account for. Although most model assumptions were satisfied after transformations, there were issues with non-constant variance that could affect the reliability of our conclusions.

Future work could focus a lot on improving data by collecting more data points for each of the fires. While technology can be limited in this aspect, especially for fires that may be in remote regions, better quality data could be very effective in helping to understand where resources need to go to put out fires and solve this important problem. Despite these limitations, our models did uncover some important factors of wildfire size and confirmed the need for more detailed data to support effective wildfire prediction and management.

0.6 Appendix

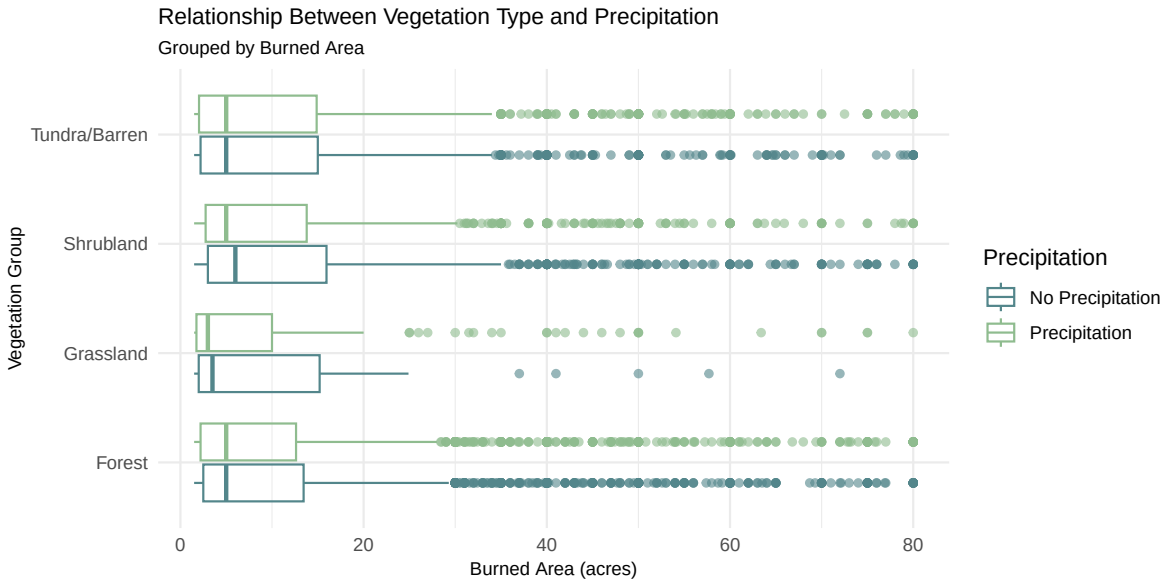
0.6.0.1 Exploratory Data Analysis - Bivariate EDA for Categorical Predictors

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.



0.6.0.2 Exploratory Data Analysis - Potential Interaction Effects

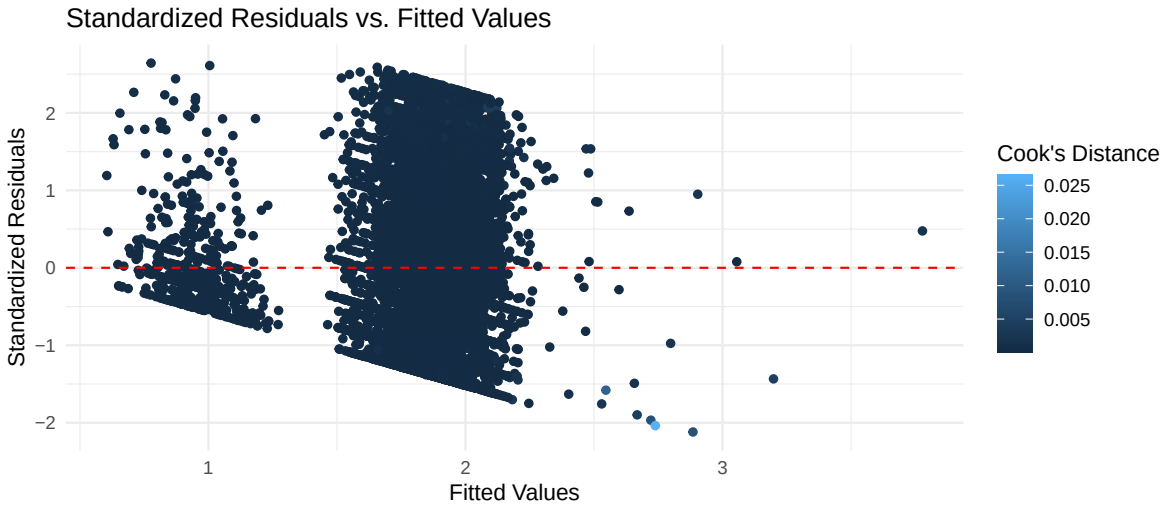
In our analysis, we were interested in a potential interaction between vegetation type and precipitation. Based on the visualization below, the distributions of burned area appearrelatively similar between precipitation and no precipitation conditions for each vegetationgroup. In forest, the median burned area is 5.0 acres for both no precipitation (mean of 11.48acres) and precipitation (mean of 11.14 acres), while in grassland, the median is 3.5 acresfor 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 5.0acres 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 worthexploring more formally in the modeling process to determine whether it significantly improvesmodel fit.

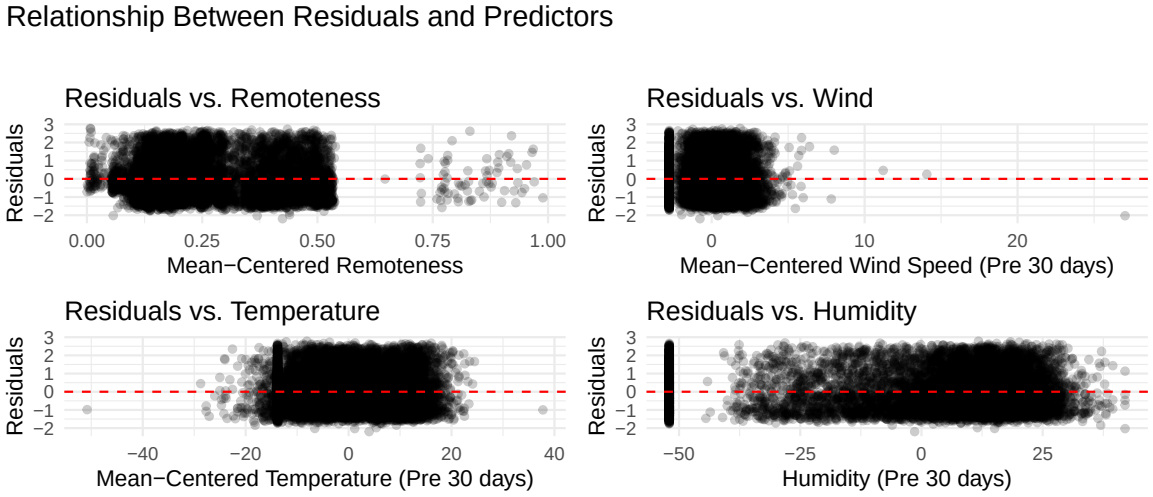


0.6.0.3 Methods - Explanatory Model

A scatterplot of residuals vs. fitted values for our explanatory model is shown below. Given that the initial model we fit had a clear violation of constant variance (an increase in the variability of residuals as fitted values increased), we applied a variance-stabilizing (i.e., logarithmic) transformation to the response variable. We observed an improvement in the spread of residuals (more uniform spread across the range of fitted values) as well as a slight improvement in model fit (an increase in adjusted R^2) after applying this transformation. Because there was also a clear pattern in the residuals (i.e., they were not randomly scattered around $residuals = 0$), we examined scatterplots of fitted values versus each of the predictors and observed no clear nonlinear trends. As a result, we did not transform our predictors. We also did not observe any influential points ($D_i > 0.5$) in our data, so no observations were

removed. Lastly, we observed two distinct groupings in the scatterplot of residuals vs. fitted values (one group with fitted values less than 1.5 and one group with fitted values generally above 1.5). We found that a large portion of this separation could be explained by region, so we fit separate linear regression models for different regional groups. However, this approach did not lead to a meaningful increase in explanatory power, so we retained a single model for all observations.

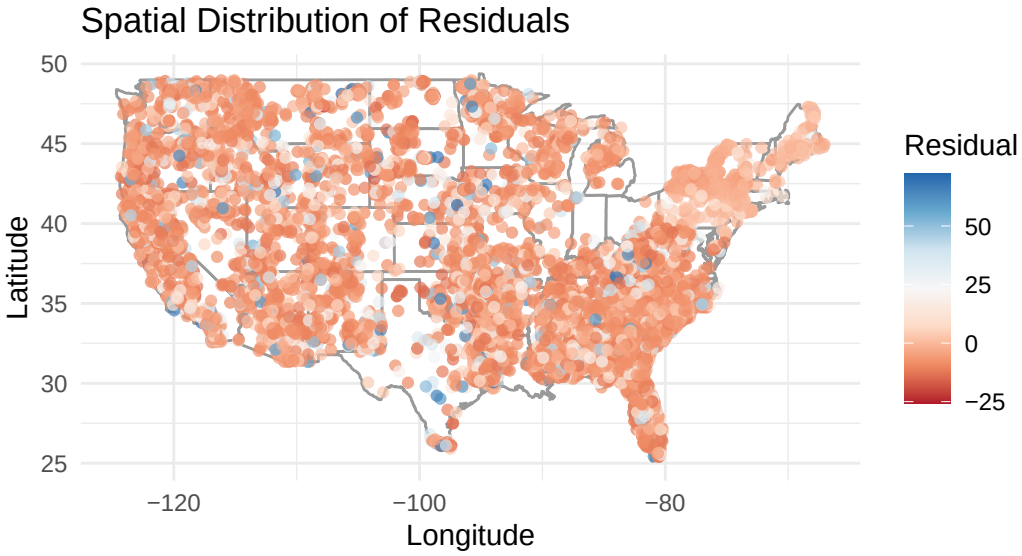


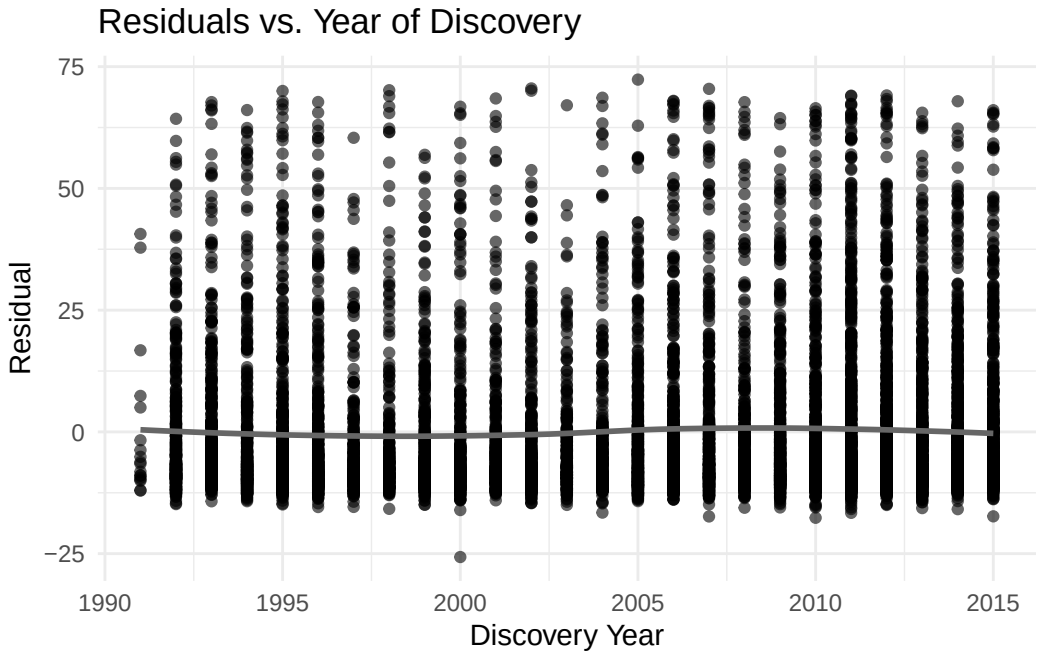


We also screened for potential collinearity between our predictors and, while we initially observed some concerning multicollinearity between levels of our region variable ($VIF > 5$), we addressed this by collapsing the South and West levels of the variable together and refitting the model based on the recoded version. All of the VIFs for the updated model are less than 5, indicating no issues with multicollinearity among the predictors.

Predictor	VIF
Prec_pre_15_cent	2.29
cause_categoryNatural	1.74
regionSouth_West	1.74
remoteness	1.69
regionNortheast	1.63
cause_categoryDebris Burning	1.62
cause_categoryOther/Unknown	1.62
Hum_pre_15_cent	1.58
Vegetation_groupTundra/Barren:Prec_pre_15_cent	1.54
Vegetation_groupShrubland:Prec_pre_15_cent	1.53
Wind_pre_15_cent	1.42
Temp_pre_15_cent	1.38
Vegetation_groupShrubland	1.31
Vegetation_groupGrassland	1.31
Vegetation_groupGrassland:Prec_pre_15_cent	1.31
cause_categoryInfrastructure	1.29
cause_categoryRecreational	1.28
Vegetation_groupTundra/Barren	1.23
putout_time_num	1.04

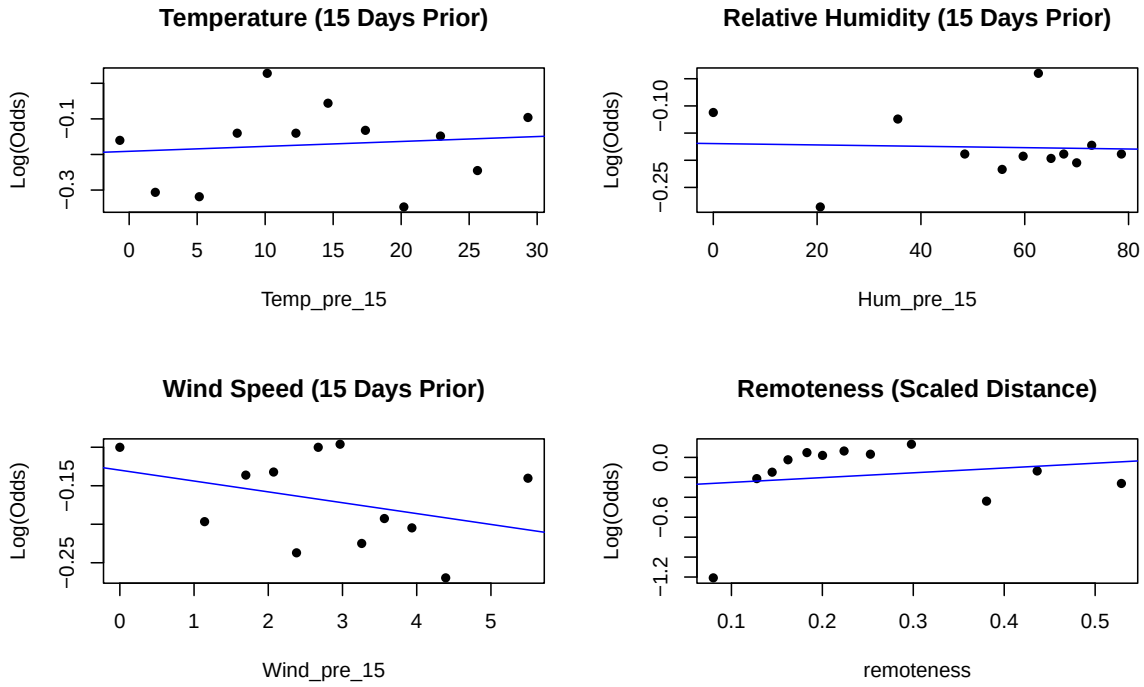
Lastly, we looked for any spatial and temporal correlations in residuals and found no apparent patterns, so independence can be considered reasonably satisfied.





0.6.0.4 Methods - Predictive Model

The empirical logit plots for the quantitative predictors of our logistic regression model are shown below. As we can see, the plots do not display strong deviations from linearity, which tells us that the assumption of linearity in the log-odds appears to be met for these predictors (with no obvious patterns suggesting that non-linear transformations would be necessary, as the relationships do not appear curvilinear, quadratic, or otherwise non-linear in the logit scale). Thus, the linearity condition appears to be reasonably satisfied.



0.6.0.5 Results - Explanatory Model

The (abbreviated) estimated multiple linear regression equation predicting log-transformed fire size is shown below:

$$\begin{aligned}\widehat{\log(\text{Fire Size})} = \; & \beta_0 + \beta_1\,\text{Region} + \beta_2\,\text{Cause Category} + \beta_3\,\text{Remoteness} \\ & + \beta_4\,\text{Vegetation Group} + \beta_5\,\text{Wind} + \beta_6\,\text{Temperature} \\ & + \beta_7\,\text{Putout Time} + \beta_8\,(\text{Vegetation Group} \times \text{Precipitation})\end{aligned}$$

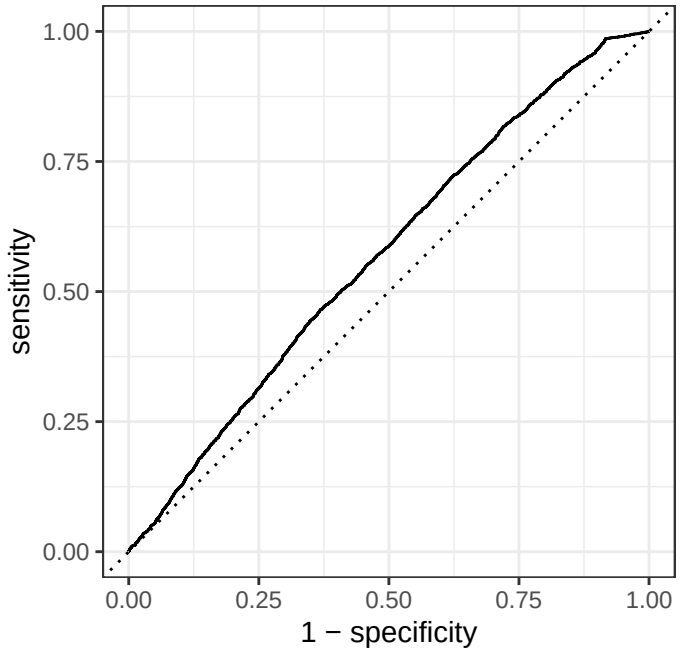
β_0 : 2.00 (Intercept)
β_1 : Region effects (e.g., Northeast: -0.94, South and West: 0.08)
β_2 : Cause effects (e.g., Debris Burning: -0.26, Recreational: -0.37)
β_3 : -0.31 (Remoteness)
β_4 : Vegetation effects (e.g., Grassland: 0.19, Shrubland: 0.07)
β_5 : 0.03 (Wind 15 days prior)
β_6 : -0.00 (Temperature 15 days prior)
β_7 : 0.01 (Containment time)
β_8 : -0.14 (Interaction: Shrubland $\times$ Precipitation)

0.6.0.6 Results - Predictive Model

The estimated logistic regression equation predicting the log-odds that a wildfire falls into the upper 50% of moderate-sized fires is shown below.

$$\begin{aligned}\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \; & \beta_0 + \beta_1 \,\text{Region} + \beta_2 \,\text{Remoteness} + \beta_3 \,\text{Vegetation Group} \\ & + \beta_4 \,\text{Precipitation} + \beta_5 \,\text{Wind} + \beta_6 \,\text{Temperature} + \beta_7 \,\text{Humidity} \\ & + \beta_8 \,(\text{Vegetation Group} \times \text{Precipitation})\end{aligned}$$

Additionally, the ROC curve for this model is presented below. We observed an AUC of 0.569, indicating relatively limited discriminative ability between fires falling into the upper 50% of moderate-sized fires and those falling into the lower 50%. As we can see in the ROC curve, as the true negative rate increases, sensitivity increases only gradually, which indicates limited improvement in correctly identifying larger fires (even as the true negative rate improves).




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.569
```

Lastly, the results of the drop-in-deviance test for our logistic regression are presented below. Since the p-value (0.013) was less than 0.05, we retained the interaction between vegetation group and precipitation in the model.

Model	Residual df	Residual Deviance	Test df	Test Deviance	p-value
mid50_hi region + remoteness + Wind_pre_15 + Temp_pre_15 + Hum_pre_15	10662	14399.02	NA	NA	NA
mid50_hi region + remoteness + Vegetation_group * Prec_pre_15 + Wind_pre_15 + Temp_pre_15 + Hum_pre_15	10655	14381.18	7	17.839	0.013