An Analysis of California Wildfire Damage Area

Chris Nurrenberg, Luke Wukusick

Abstract:

The paper analyzes the data gathered by the National Wildfire Coordinating Group on United States wildfires. It examines the effect the fire cause, location, and year have on the final fire size. The text explores this in the context of approaching fire suppression more intelligently and with a goal of reducing the sizable costs that are associated with fighting large fires. The study described in this text focuses on wildfires in California, and fits a multiple linear regression model to this data. This model includes the variables `fire_year`, `longitude`, `latitude`, and `stat_cause_descr` in the prediction of the variable `fire_size`. The results indicate that fires to the West and the South in California tend to be larger on average, that fire sizes have increased over the years, and that lightning-caused fires and, to a lesser extent, structural fires tend to be larger.

Introduction

Every year out-of-control wildfires damage large amounts of California property and endanger lives. Learning where the most dangerous fires are allows for quicker responses and better preperation. Predicting where and when the largest fires occur lets responders have resources at the ready and even can help prevention efforts. These concerns produce a very clear question to be answered: is there an association between location, time, and cause and the size of California wildfires, and if so, what is this association?

Background

The cost of fighting fires in California was around $1.8 billion in 2017. This cost is stupendous, and any advantage the state can gain against Mother Nature is welcome. Learning a predictive model for latitude and longitude would allow responders to preemptively set up supplies and to spend more on preventative campaigns in high-risk areas. Including year in this model will allow the Forest Service to estimate the trend in wildfires over time; will the next few years require more or less funding than the past few? Finally, including the fire cause in the model will help those interested in preventative measures see what causes should be targeted the most and if any past efforts have helped certain causes taper off.

Data

The wildfire data is sourced from the National Wildfire Coordinating Group. The links to the CSV files can be found in the references section. According to the accompanying material, this data is the aggregation of local fire-reporting systems' observations. This includes data from federal, state, and local organizations. Error-checking was performed and redundant observations were removed as much as possible. The observations span the years from 1992 to 2015. The data set contains observations from the entire United States, including the territory of Puerto Rico, but for this study we only looked at the data gathered in California. While the complete set had approximately 1.9 million observations, the subset of the data collected in California had only 189,550 observations. We also removed observations that had `missing/undefined` or `miscillaneous` causes, further bringing our analyzed number of observations down to 125,002. The predictor variables our analysis considers at are `longitude`, `latitude`, `stat_cause_descr`, and `fire_year`. In regards to the model's response, our analysis considers the `fire_size` variable.

Methodology

We fitted a multiple linear regression model with predictor variables `longitude`, `latitude`, `stat_cause_descr`, and `fire_year`, and with response variable `fire_size`. In order to obtain the sample distribution we used residual bootstrap. We chose this method in lieu of using a t-distribution because it seems extremely unlikely that the errors in the linear model came from a normal distribution. Our complete checking of assumptions can be found in the Appendix. We used the produced sampling distribution to check the significance of the intercepts and slopes of each predictor variable against our significance level of 0.02. We also observed the R^2 of the fit.

Implementation

First, we plot our response and predictor variables to check for non-linear relationships and collinearity.

```
pairs(~fire_size + latitude + longitude + fire_year + stat_cause_descr, data=
cleanedData)
```

The graph of this comparison can be found in the appendix.

As there are no obvious non-linear relationships or collinearity, we proceed with the model fitting as described above. The model summary can be found in the appendix.

```
model <- lm(fire_size ~ longitude + latitude + fire_year + stat_cause_descr,
data=cleanedData)
```

To conduct statistical inference on our dataset, we perform residual bootstrap on the dataset to generate confidence intervals on the predictor parameters.

```
residualModel <- residual.boot(fire_size~longitude + latitude + fire_year + s
tat_cause_descr, data=cleanedData, B=dim(cleanedData)[1], seed = 17)

CI.df <- data.frame(Lower = numeric(), Upper = numeric())
for (i in 1:dim(residualModel$bootEstParam)[2]){
  print(i)
  CI.df[i,]=unlist(BootCI(residualModel$bootEstParam[,i], alpha=0.02))
}

rownames(CI.df) <- colnames(residualModel$bootEstParam)
```

Statistical Interpretation

We can say with 98% confidence that:

- An increase in longitude by 1 degree decreases the average fire size between 13.82 acres and 44.29 acres on average with all other variables held constant.
- An increase in latitude by 1 degree decreases the average fire size between 16.51 acres and 42.34 acres on average with all other variables held constant.
- Each year, average fire sizes in California increase by somewhere between 1.325 acres and 5.55 acres on average with all other variables held constant.
- Fires caused by lightning are between 75.00 and 177.68 acres bigger on average compared to fires started by arson with all other variables held constant.
- Fires with causes labeled as “Structure” are between 450.67 acres and 1426.62 acres larger on average compared to fires started by arson with all other variables held constant.
- Fires caused by Campfire, children, debris burning, equipment use, fireworks, powerlines, railroads, and smoking do not have any statistically significant difference than fires started with arson.

Conclusions

Our model predicts that fire size is generally larger in the southwestern portion of the state. This general trend may not allow for extremely specific pre-allocation of fire-fighting resources, but knowledge and confirmation of the general trend should help influence decisions. The increasing size of fires over the years should serve as a warning and a motivator: fires will likely continue to become more destructive over the years, so increased funding and aid are critical, as well as increased focus on preventative measures. Our results also indicate that fires caused by lighting and fires that are initially structural in nature are larger on average, meaning preventative measures may be focused on these areas in the future. Perhaps effort could be directed toward developing monitoring systems for lightning in remote areas or research into flame retardant building materials could be further incentivized.

More work will likely have to be done in the area, but the information gleaned from this study is useful and directive for those considering how to effectively confront California wildfires.

Limitations and Future Work

There are a few conceptual limitations to our model. One of the most obvious things is condensing the time of the fire to the low granularity of the single year. A more accurate model would be produced by having granularity up to a single day (or even a single hour). We condensed the date into single-year granularity due to a cyclical topography within the data–average fire sizes within a single year had a non-linear relationship with the day of the year, with larger fires happening much more frequently in late spring to summer.

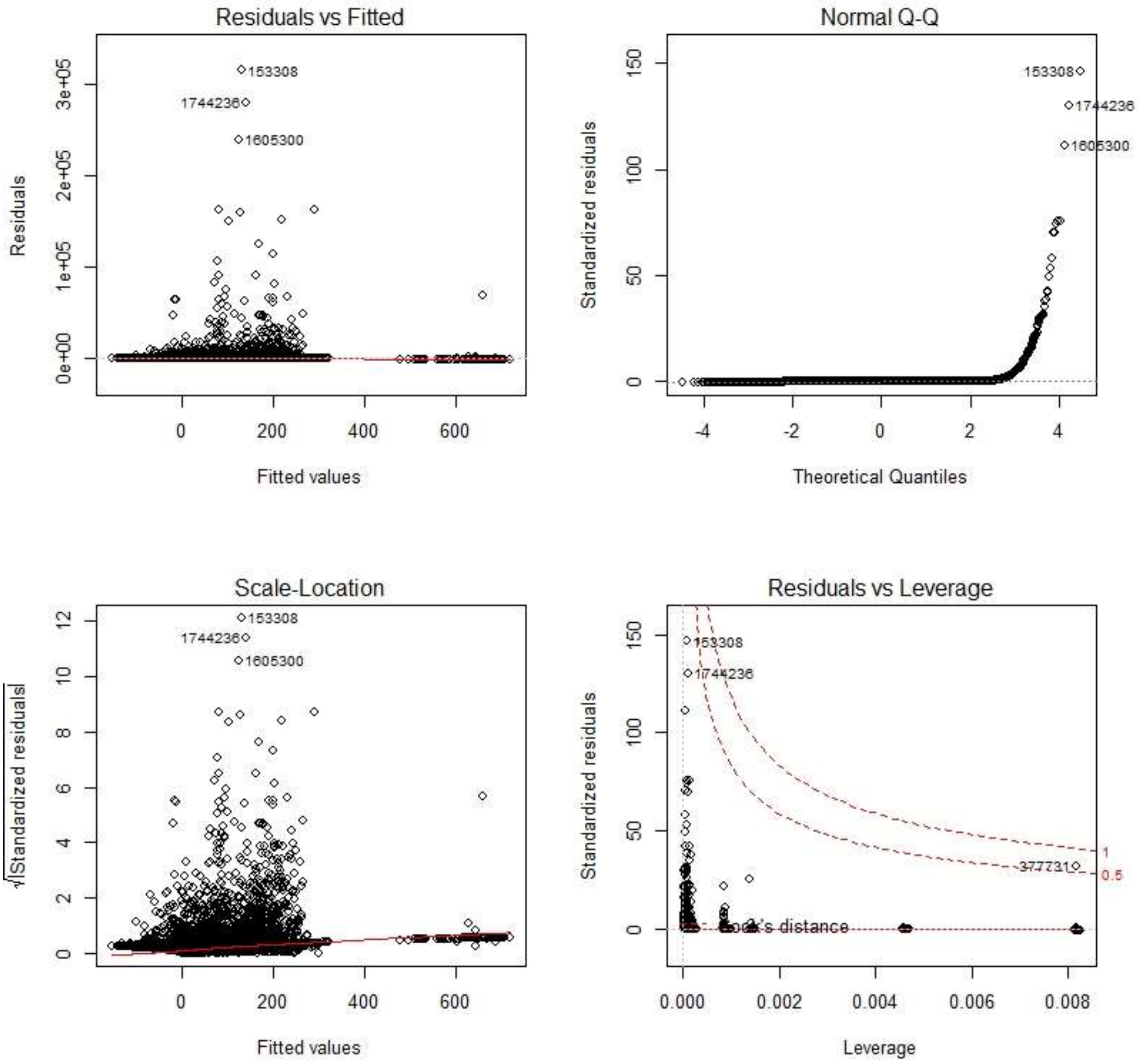
Another problem is the lack of useful information contained within the latitude and longitude of the fire. The model does not (and cannot with this data set) take into account the different geographical differences in certain regions. The data set does not desribe the specific biome in which the fires took place. A particular region may be more or less prone to fires, which may create a non-linear relationship between the expected fire size and coordinates across the entire state of California.

With some of the default functionality in R, the effect of the Arson category in the fire cause was included into the intercept variable. All other categorical confidence intervals were computed with respect to the difference between itself and arson, which isn’t particularly interesting. We would have liked to exclude the arson category from the intercept, but we were unable to do so.

A better approach which could be used in the future is multiplying each quantitative variable by the cause category variable, which would indicate if each variable makes a difference with respect to each cause. This would eliminate most of the previous problems with model comparisons occurring against a base-case of Arson, and it would provide more interesting results.

Appendix

Checking Model Assumptions



As the data is aggregations of local wildfire reporting agencies, we feel it is reasonable to assume errors are independent. We also feel it is reasonable to assume errors have a mean of 0 basd on the residuals vs. fitted plot. There is no particular fan-shape in the residuals vs fitted as well, so we feel constant variance is a reasonable assumption to hold. The normal Q-Q plot shows that it is not reasonable to assume that the errors come from a normal distribution.

Model Summary

```
summary(model)

##
## Call:
## lm(formula = fire_size ~ longitude + latitude + fire_year + stat_cause_des
cr,
##     data = cleanedData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -719   -109    -60    -21 315446
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -9106.7352  1888.2499  -4.823 1.42e-06 ***
## longitude                      -28.9197     6.3859  -4.529 5.94e-06 ***
## latitude                       -29.5188     5.4434  -5.423 5.88e-08 ***
## fire_year                        3.3981     0.9029   3.764 0.000168 ***
## stat_cause_descrCampfire        40.4022    27.0637   1.493 0.135478
## stat_cause_descrChildren       -55.5926    30.2253  -1.839 0.065877 .
## stat_cause_descrDebris Burning -42.5976    23.7509  -1.794 0.072892 .
## stat_cause_descrEquipment Use  -35.7972    18.9020  -1.894 0.058249 .
## stat_cause_descrFireworks      -32.3029   146.3375  -0.221 0.825293
## stat_cause_descrLightning      127.6040    21.6943   5.882 4.07e-09 ***
## stat_cause_descrPowerline       77.1693    64.1553   1.203 0.229036
## stat_cause_descrRailroad        59.6023    81.8249   0.728 0.466362
## stat_cause_descrSmoking        -48.0556    31.9659  -1.503 0.132754
## stat_cause_descrStructure      555.3978   194.7934   2.851 0.004356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2153 on 124988 degrees of freedom
## Multiple R-squared:  0.0009689,  Adjusted R-squared:  0.000865
## F-statistic: 9.324 on 13 and 124988 DF,  p-value: < 2.2e-16
```

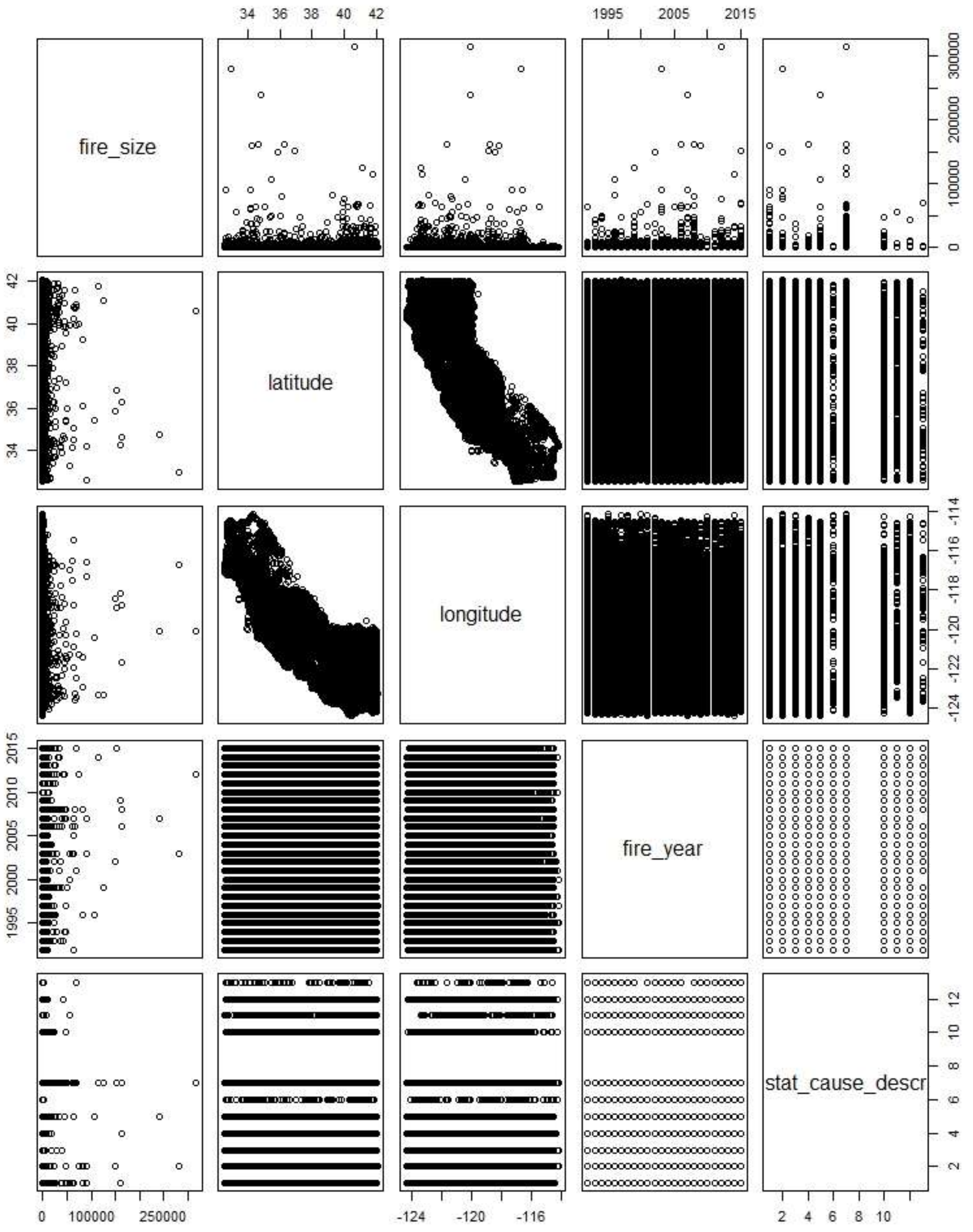
Predictor Coefficient Confidence Intervals

```
knitr::kable(CI.df, caption="Confidence intervals for each predicter coeffici
net, alpha=0.02")
```

Confidence intervals for each predicter coefficinet, alpha=0.02

	Lower	Upper
(Intercept)	-13573.063970	-4754.568440
longitude	-44.292942	-13.822975
latitude	-42.338006	-16.506805
fire_year	1.325271	5.552039
stat_cause_descrCampfire	-17.852299	111.409477
stat_cause_descrChildren	-117.587045	27.719861
stat_cause_descrDebris Burning	-96.784238	15.928474
stat_cause_descrEquipment Use	-83.480377	5.396413
stat_cause_descrFireworks	-131.680947	660.468590
stat_cause_descrLightning	74.999771	177.684974
stat_cause_descrPowerline	-5.446738	320.875154
stat_cause_descrRailroad	-29.674414	425.366294
stat_cause_descrSmoking	-111.320596	42.843372
stat_cause_descrStructure	450.665797	1426.624026

Pairs Plot of Response and Predictor



References

"How A Booming Population And Climate Change Made California's Wildfires Worse Than Ever." BuzzFeed News, BuzzFeed News, 2018, buzzfeednews.github.io/2018-07-wildfire-trends/.

California spent nearly $1.8 billion last year fighting major wildfires. (2018, March 01). Retrieved November 9, 2018, from http://www.latimes.com/local/lanow/la-me-wildfire-costs-20180301-story.html

The dataset can be found in CSV format directly through these links:

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_7.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_6.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_5.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_4.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_3.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_2.csv

https://raw.githubusercontent.com/BuzzFeedNews/2018-07-wildfire-trends/master/data/us_fires/us_fires_1.csv