# Climate Risk Analysis

DATASCI 203: Lab 02 Assignment

Nicolas Aragon, Daniel Costa, Bella Davies, Andy Main

This study explores whether racial minority communities are more vulnerable to the effects of wildfires. We present an analysis of data from the U.S. Census Bureau and the Council of Economic Quality that finds that, controlling for western/eastern regions of the country, the more white a community is, the less likely it is to be impacted by wildfires. We hope our findings can help public agencies better identify and prepare local communities for the consequences of increased wildfire risk.

# Introduction

Climate change is an ongoing phenomenon that has been studied extensively by climatologists, and there are well-documented impacts all over the globe. From weather impacts like flooding and hurricanes to the disruption of delicate ecosystems like coral reefs, the global ramifications of climate change are something that our society and environment are struggling to adapt to. One observed impact of climate change has been an increase in the frequency and intensity of wildfires, particularly in the American West (Denison). Wildfire seasons have caused over $10 billion in losses in California alone in recent years (Bay Area Council Economic Institute). In this paper, we explore the question: are socially or economically disadvantaged groups more at risk for wildfires? Our findings support what previous literature has found, that there is a 'climate gap' that disproportionately impacts disadvantaged populations, as shown by Morello-Frosch, et al. in their paper "The Climate Gap and the Color Line – Racial Health Inequities and Climate Change". Understanding who is most impacted by wildfires can help local, regional, and national agencies plan for and mitigate those impacts in the interest of the racial justice and the public good.

# Data

Our data source is the Council of Environmental Quality, an executive advisory council that falls under the Office of the President of the United States. In 2021, the CEQ established the Climate and Environmental Justice Screening Tool (Council of Environmental Quality). This publicly available tool aggregates several datasets that track environmental burdens across the nation on the census tract level, with indicators ranging from climate change and legacy pollution to housing and transportation.

Climate and Environmental Justice Screening Tool dataset is sourced from a variety of public sources. Demographic data (feature variables) come from the American Community Survey administered by the U.S. Census Bureau. Fire risk (modeled variable) data included in the CEQ dataset is provided by First Street: a company whose mission is to connect climate change to financial risk (First Street). First Street's fire risk model uses datasets on wildfire incidence and risk estimation models published by the U.S. Forest Service and Department of Interior to determine the percentage of each census tract's population that could likely be exposed to fire damage in the next 30 years. This data is only available for the contiguous 48 states, which will be our region of interest. All demographic and fire-risk data is from 2010, the latest year for which complete data on relevant variables is available.

Core Based Statistical Area data was gathered from the U.S. Census Bureau website and was last updated in July 2023. Although this dataset was updated 13 years after the fire-risk and demography data source, the Census Bureau notes that previous changes in CBSA definitions have found "very small or nonexistent" differences in racial housing patterns due to such re-classifications. Therefore, we consider any similar fire-risk consequences of CBSA re-classifications to be negligible (US Census Bureau).

# Operationalization

## Units of Observation

Demographic and fire risk data are captured on the Census tract level. Using tracts as the units of analysis introduces considerable dependence, as U.S. residents often cluster geographically by race. Additionally, tracts near common sources of wildfires (such as dry forests or emissions from oil fields) will have similar levels of fire risk.

To address this, we created clusters of census tracts to reduce the level of dependence between data points. The units of analysis we chose are Core Based Statistical Areas (CBSAs), which are defined by the Census Bureau as having "at least one urban cluster of at least 10,000 but less than 50,000 population, plus adjacent territory that has a high degree of social and economic integration with the core as measured by commuting ties"(Census Bureau).

Given the standards for delineating CBSAs, we are confident that census tract data would show considerably

more dependence within CBSAs than between them both with regards to demography and environment. Note that in this regression we will be weighting CBSAs equally regardless of population.

Of the 74,000 census tracts, about 1% had missing racial demographic data, and so did not contribute to our aggregated figures. Of the 928 total CBSAs defined by the Census Bureau, only 625 had well-formed county titles that could be cross referenced against our fire risk dataset.

### Variables

After conducting an exploratory data analysis on our holdout dataset, we determined that the only demographic/economic data that would help answer our research question was the percentage of white residents in each region. Thus, our conceptualization of rating CBSAs as "socially or economically disadvantaged" is "percentage white/non-white", and this data is directly available in the data source. Fire risk conceptualization/operationalization is detailed in the "Data Source" section. As we do not have data on the number of residents per property, a modelling assumption of this analysis is that the ratio of residents to properties across CBSAs is equal. We thus scale the percentage of properties at risk of fire damage by the population of each CBSA.

After visualizing the geographic distribution of fire risk across the contiguous U.S., we noticed a clear east/west discontinuity at a longitude near that of the Mississippi river (see Figure 1). We thus conceptualized that east/west fire risk should be modeled in our regression, and we operationalized western CBSAs as those whose counties all lie west of the Mississippi river. All CBSAs in the original Census dataset were defined such that no CBSA exists which crosses the Mississippi river.

## EDA and Visualizations

To explore whether geography is related to the log of the percent fire risk of a county, we analyzed a geographic map of the dispersion of the log of percent fire risk across counties in the US. This map shows a clear relationship between geography and fire risk, as the western region of the country has a much higher risk of fire than the eastern region (Figure 1). Due to this observed association in the exploratory data analysis of the dependent Y variable, we used the binary indicator variable of Region as our $X_2$. This stark difference in fire risk distribution between east/west CBSAs suggests that the Region variable will be effective at addressing concerns of identical distribution with respect to our Y variable.

In alignment with our research question, the next independent variable we explored is the percentage of the population in each county that identifies as white, or percent white as our metric $X_1$. We can see the relationship of X vs. Y in Figure 4, and X vs. Log(Y+1) in Figure 5. These graphs indicate that using Log(Y+1) is an improvement from no transformations as the data points appear slightly more linear since the weighted average line is slightly smoother and closer to 0 with respect to the standard error. We also considered using percent housing burden, percent poverty, and percent no highschool degree, however the exploratory data analysis and visualizations included in the R code showed no clear relationship with our Y variable. This indicates that they would not be very useful in the model, and adding these variables to our model would inflate the variance of the other coefficients without adding much accuracy to our model. Thus, the final variables we chose to model our Y variable of percent fire risk based on our exploratory data analysis are region and percent white.

## Model Specification

To run our models, we continued to use the exploratory set (30%), then validated the final model with the modeling set (70%). Results of our EDA model regressions are displayed in Figure 8.

For our initial simple model, we regressed our Y variable of percent fire risk on our X1 variable of percent_white, as this continuous variable displayed the greatest association with fire risk from our exploratory data analysis. As seen in Figure 8, Model 1 resulted in an $R^2$ value of 0.137, indicating the model does not do very well at capturing the relationship. Model 3 used the log of the Y variable of percent fire risk and resulted

in a 0.047 increase in $R^2$ from Model, which indicates that using the log form of Y better captures the relationship than the normal form of Y (Figure 8). Next we added X2, or region_west, to the simple model without transformations, resulting in an improvement of $R^2$ by 0.209 from Model 1. Building upon this, we transformed the Y variable to obtain the more complex Model 4, as we expect this to result in further improvement. Regressing the log of percent fire risk on region and percent white to obtain Model 4, the $R^2$ is 0.42, an overall improvement of 0.284 from the simplest Model 1. While all coefficients across the 4 models are highly significant with p<0.001, Model 4 captures the relationship best as it has the most improved $R^2$ value, meaning it best explains the variance in fire risk in terms percent_white and region.

Upon choosing our final model, we ran it on the modeling set of 70% of the original dataset, which was held until deciding the final model during our EDA. Running the same regression used in Model 4 now on the modeling dataset, the $R^2$ became 0.462, with coefficients of $B_0 = 2.356$, $B_1 = 2.019$, and $B_2 = -0.033$ with $p < 0.001$ for all 3 coefficients. (Figure 9).

The equation for our final chosen model (Model 4) run on the modeling dataset is as follows:

$$Log(PercentFireRisk + 1) = 3.535 + 2.019 \cdot RegionWest - 0.033 \cdot PercentWhite$$

This model uses non-robust or typical standard errors. When running the Breusch-Pagan test, we observed p-values of 0.003 and 0.002 for the simple Model 1 and complex Model 4, respectively (Figure 7). Although this test is non-dispositive, we believe we have homoskedastic errors and therefore chose to use the typical standard errors.

# Model Assumptions

Given our EDA and full model sample sizes of 187 and 438 each being greater than 100, we choose to use the Large Sample Model assumptions to validate our analysis. Satisfying the assumptions on the EDA set gives us confidence that the full model will also satisfy the Large Sample Model assumptions.

## Assumption #1: I.I.D. Data

Fire risk independence between CBSAs is not fully satisfied, as geographic regions impacting fire risk span areas much larger than the distance between CBSAs (some of which share a border). With regards to independence of racial composition, migratory/residential trends (such as Asian-Americans residing in the west, black Americans in the southeast, Hispanic Americans in the southwest, and white Americans in the north and east) also span large areas, and introduce dependence between clusters of CBSAs. Nonetheless, we believe our conceptualization allows us to analyze data with sufficient independence to perform a meaningful linear regression. Our EDA revealed that there is a different distribution of fire risk in the east and west, and so we modeled this difference to address concerns of identical distribution. Differences in distribution of racial composition is also modeled in our regression. However, fire risk may still be unevenly distributed between areas that have resources to mitigate their fire susceptibility or areas that have run out of timber to fuel these fires.

## Assumption #2: Unique BLP (no perfect collinearity)

Our regression is an estimate of the BLP between the inputs and outcome variables, and so a BLP must exist for our estimate to have a meaningful interpretation. BLP coefficients are directly proportional to the covariances between the Xs and Ys, these quantities must be finite for a BLP to exist. Figure 6 shows that all covariances are indeed finite. In order to be certain that the linear relationship we describe between inputs and outputs estimates their one true relationship, the BLP we must also be unique. This requires that, for inputs $X$, $(X^T X)$ must be invertible. This is true if and only if there exist no features that are linear combinations of other features. As we only have two input variables, checking that they are not perfectly correlated will suffice. Figure 6 shows that they are not perfectly correlated ($|r| < 1$).

## Results and Interpretation

Our interpretation is complicated by our decision to use Log(Percent Fire Risk + 1), adding 1 to address the several 0 values in the dataset. For large input variable values, the difference in practical interpretations of our findings vs. that of a model that only uses Log(Percent Fire Risk) are negligible, but for our $B_2$ coefficient corresponding to our region variable, since the range of the indicator variable is from 0 to 1, the practical interpretation of our findings differs significantly. Due to a conversation with our instructor, we will interpret our constant coefficient with respect to our log + 1 model, and input coefficients with respect to log, while noting that these interpretations approximate the true consequences of our findings.

Our final model regressing percent fire risk on region and percent white has highly significant values for the all coefficients included in the model (see Figure 9). Due to our To interpret the effect of the coefficients on the Y variable, we will convert the values back to the normal form to remove the logs from the interpretations. The $B_0$ of 3.535 represents baseline expected CBSA fire risk of approximately $e^3.535 - 1 = 33.3$ when region_west = 0 (east) and percent_white = 0. The $B_1$ of -0.033 indicates that for each 1 unit increase in percent white, there is a multiplicative effect on percent_fire_risk of $1 - 0.033 = 0.967$. This is equivalent to a 3.3% decrease in CBSA fire risk for each additional 1 unit increase in percent white. From one perspective, this means that CBSAs that have even 10% more racial minorities than another in the same region can expect their fire risk to be 1/3 higher over the next 30 years. The $B_2$ of 2.019 indicates that going from the East to West region (region_west=1) has a multiplicative effect on percent_fire_risk of $1 + 2.019 = 3.146$. This means that, holding percent_white constant, moving from east to west multiplies increases a CBSAs fire risk by 215%. This finding is consistent with the drier climate and higher concentration of oil fields present in the western United States. Figure 13 shows the plots of the final model predictions vs. residuals for the east and west regions which show residuals centered around zero in each respective region, and with a weighted average that never strays more than 1 standard error from 0, indicating at least a moderately strong linear relationship between percent_white/region and log(percent_fire_risk + 1). Residuals for the final transformed model are shown in Figure 10, and the log predictor of the features against the untransformed percent_fire_risk is displayed in Figures 14 and 15.

Adding on to these interpretations, we will interpret some example data points. For an example data point of the CBSA around Clovis, NM in the west region which is 48.23 percent white, our model predicts that this area has a $log(PercentFireRisk + 1) = 3.535 + 2.019 \cdot (1) - 0.033 \cdot (48.23) \approx 3.96$. This converts back to $e^{3.96-1} \approx 51$ percent fire risk for this example CBSA. In comparison, the CBSA around Rapid City, SD in the western region with a larger white population of 81.47% would have $log(PercentFireRisk + 1) = 2.356 + 2.147 \cdot (1) - 0.020 \cdot (81.47) = 2.874$, or $e^{2.874-1} \approx 17$ percent fire risk. Although the variance in the model may allow for limited counterexamples, this lower fire risk among a whiter population within the same region is an example of the practical consequences our regression coefficients.

## Conclusion

Overall, it seems that there are significant disparities in fire risk between metropolitan areas of differing racial demographics. This analysis adds to the conversation about the intersection between racial and climate justice, and could support efforts by lawmakers to advocate for climate justice while highlighting the racial justice implications of pursuing such policies. Hopefully, this will inspire further research into how to bridge the climate risk gap along social and economic lines, and incentivize government agencies to allocate resources towards this end.

# References

Dennison, P. E., Brewer, S. C., Arnold, J. D., & Moritz, M. A. (2014). Large wildfire trends in the Western United States, 1984–2011. Geophysical Research Letters, 41(8), 2928–2933. https://doi.org/10.1002/2014gl059576

The true cost of wildfires. The True Cost of Wildfires | Bay Area Council Economic Institute. (n.d.). https://www.bayareaeconomy.org/report/the-true-cost-of-wildfires/

Council of Environmental Quality. (n.d.). Screeningtool.geoplatform.gov. https://screeningtool.geoplatform.gov/

First Street. (n.d.). First Street Technical Documentation And Faqs On Access. firststreet.org. https://firststreet.org/documentation

Morello-Frosch, R., & Obasogie, O. K. (2023). The Climate Gap And The Color Line — Racial Health Inequities And Climate Change. New England Journal of Medicine, 388(10), 943–949. https://doi.org/10.1056/nejmsb2213250

United States Census Bureau. (2023, June 23). Glossary. Census.gov. https://www.census.gov/programs-surveys/metro-micro/about/glossary.html

Xie, W., & Meng, Q. (2024). Spatial and Temporal Analysis of Vulnerability Disparity of Minorities to Wildfires in California. International Journal of Disaster Risk Reduction, 104949.

# Appendix

Figure 1. Map of Percent Fire Risk

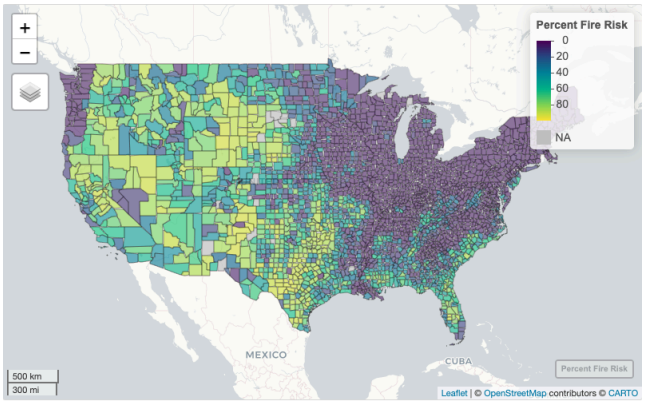


Figure 4. $X_1$ vs. Y

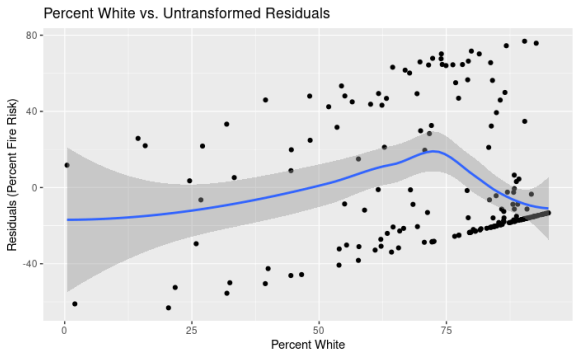


Figure 5. $X_1$ vs. Log(Y+1)

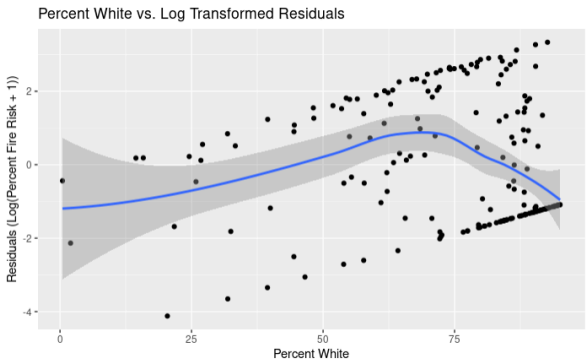


Figure 6. Correlation and Covariance matrices

| | percent_white | region_west | percent_fire_risk |
|---|---|---|---|
| percent_white | 1.0000000 | -0.2407774 | -0.3705060 |
| region_west | -0.2407774 | 1.0000000 | 0.5321467 |
| percent_fire_risk | -0.3705060 | 0.5321467 | 1.0000000 |

| | percent_white | region_west | percent_fire_risk |
|---|---|---|---|
| percent_white | 403.002287 | -2.4229689 | -269.180849 |
| region_west | -2.422969 | 0.2512794 | 9.653949 |
| percent_fire_risk | -269.180849 | 9.6539494 | 1309.755850 |

Figure 7. Breusch-Pagan Test

```
        studentized Breusch-Pagan test

data:  eda_model_1
BP = 8.7083, df = 1, p-value = 0.003168


        studentized Breusch-Pagan test

data:  eda_model_4
BP = 11.712, df = 2, p-value = 0.002863
```

Figure 8. Stargazer Table of Models Attempted

Regression Results

| | Dependent variable: | | | |
|---|---|---|---|---|
| | Percent Fire Risk | | Log(Percent Fire Risk + 1) | |
| | (1) | (2) | (3) | (4) |
| Percent White | -0.668*** | -0.464*** | -0.041*** | -0.029*** |
| | (0.123) | (0.111) | (0.006) | (0.005) |
| Region (West) | | 33.947*** | | 1.904*** |
| | | (4.436) | | (0.219) |
| Constant | 76.786*** | 44.494*** | 4.948*** | 3.137*** |
| | (9.405) | (9.234) | (0.481) | (0.457) |
| Observations | 187 | 187 | 187 | 187 |
| R2 | 0.137 | 0.346 | 0.184 | 0.421 |
| Adjusted R2 | 0.133 | 0.338 | 0.179 | 0.414 |
| Residual Std. Error | 33.706 (df = 185) | 29.436 (df = 184) | 1.724 (df = 185) | 1.456 (df = 184) |
| F Statistic | 29.437*** (df = 1; 185) | 48.574*** (df = 2; 184) | 41.614*** (df = 1; 185) | 66.835*** (df = 2; 184) |
| Note: | | | | *p<0.05; **p<0.01; ***p<0.001 |

Figure 9. Final Model

Regression Results

| | Dependent variable: |
|---|---|
| | Log(Percent Fire Risk + 1) |
| Percent White | -0.033*** |
| | (0.004) |
| Region (West) | 2.019*** |
| | (0.132) |
| Constant | 3.535*** |
| | (0.312) |
| Observations | 438 |
| R2 | 0.462 |
| Adjusted R2 | 0.460 |
| Residual Std. Error | 1.359 (df = 435) |
| F Statistic | 187.060*** (df = 2; 435) |
| Note: | *p<0.05; **p<0.01; ***p<0.001 |

Figure 10. Final Model Data vs. Residuals

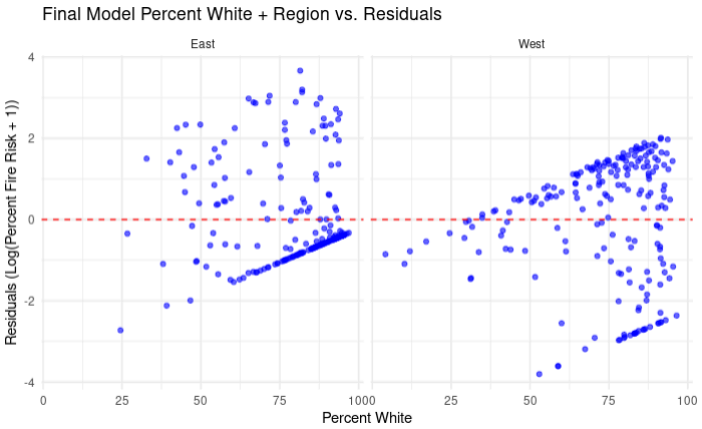


Figure 11. EDA Simple Model Residuals

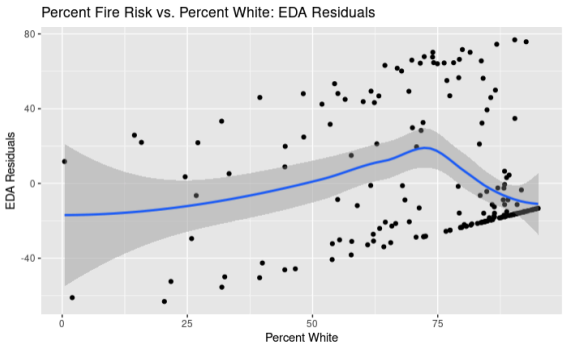


Figure 12. EDA Complex Model Residuals


Figure 13. Final Model 4 Residuals

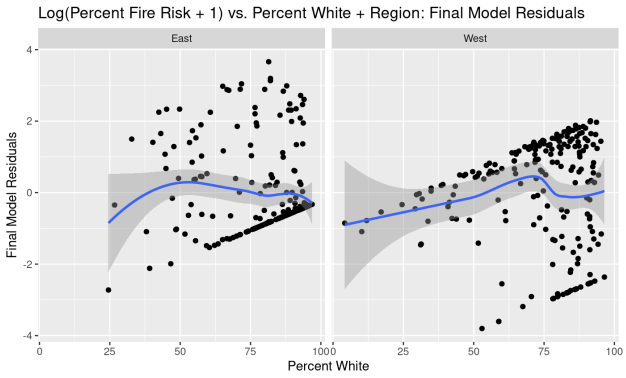


Figure 14. Final Model Predictions (West)

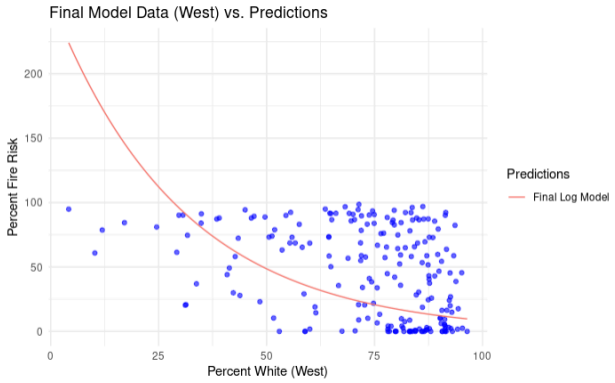


Figure 15. Final Model Predictions (East)