Assessing Wildfire Vulnerability in California: A Spatial Analysis of Socioeconomic and Demographic Factors

Eva Cao
Carleton College

Abstract:
Wildfires pose significant threats to California driven by several factors. This report examines the relationship between socioeconomic and demographic factors and wildfire vulnerability in California, with a special focus on the influence of the number of CAL FIRE facilities at a county level. Using spatial analysis techniques, the study investigates if the location and number of these facilities correspond to areas with high wildfire occurrence and if they mitigate the extent of wildfires.

Findings indicate that areas with higher percentages of white population tend to experience more severe wildfires, and the number of facilities has a positive effect on wildfire intensity with other conditions controlled, suggesting potential underlying factors related to fire risk or suppression resources. The study also reveals spatial clustering of wildfire incidents, highlighting areas such as southern California and parts of the central region as hotspots. These findings underscore the complex interplay of socioeconomic, demographic, and environmental factors in shaping wildfire vulnerability. The insights from this study can inform policymakers and stakeholders in developing more effective wildfire management and community resilience strategies.

I. Introduction

Wildfires present a significant threat to communities, ecosystems, and economies, particularly in regions prone to these disasters. California, with its history of devastating wildfires, serves as a stark example of the challenges posed by these natural hazards. Recent research has highlighted the health risks associated with exposure to wildfire smoke, particularly for the elderly population in the Western United States [1]. Various factors, including climate change, have been identified as key contributors to wildfire risk [2]. On top of that, the vulnerability of a region to wildfires is influenced by its socioeconomic and demographic characteristics, as well as characteristics of the population itself, including aging, economic deprivation, and high unemployment rates [3]. While climatic factors such as seasonal temperature or precipitation significantly influence annual fire activity in some regions in the United States, human presence also plays a crucial role. Regions more affected by human activities show less impact of climate on fire activity, suggesting that human influence can override the effects of climate on wildfires. This highlights the need to consider geographical context and human influence alongside climate in wildfire policy and management decisions [4].

Despite significant insights gained from previous studies, there is still a pressing need for a deeper understanding of the complex interactions and influences shaping wildfire vulnerability. Several studies have utilized regression analysis to investigate wildfire patterns and vulnerabilities, modeling wildfire susceptibility based on the relative influence of human, vegetation, climate, and topographic variables on fire occurrence in the United States. They found that the relationship between climatic variation and fire activity is complex and varies geographically, emphasizing the importance of spatial analysis in understanding wildfire dynamics and informing management strategies [5] [4].

However, there is a relative scarcity of spatial analysis focused on California, a state highly susceptible to wildfires due to its unique geographical and environmental characteristics. This study seeks to bridge these gaps by conducting a detailed spatial statistical analysis. It aims to investigate the relationship between socioeconomic and demographic factors and wildfire vulnerability in California, with a particular focus on understanding how CAL FIRE facilities, responsible for fire protection in the state, influence these vulnerabilities and mitigate risks. The study will assess if the location of these facilities corresponds to areas where wildfires frequently occur, and whether the number of facilities has a mitigating effect on the extent to which the county is affected by wildfires. The findings from this study are expected to provide valuable insights for policymakers and stakeholders involved in wildfire management and community resilience planning.

II. Data

The study leverages four key datasets. The first dataset, sourced from CAL FIRE, provides detailed information on wildfire incidents in California, including dates and severity levels [6]. The second dataset, retrieved from Open Data California, offers information on fire suppression facilities, including their locations and operational statuses [7]. The third dataset gathers interpolated climate data for California during 2021 from the WorldClim database [8] [9]. Additionally, data from the American Community Survey (ACS) are utilized to gather information such as total population, median household income, ethnicity/race distribution, and poverty levels at the

county level [10]. County-level data were aggregated to match the spatial resolution of wildfire incident data and were manipulated to derive new variables such as population density, percentage of elderly population, and poverty rate. As Table 1 shown, these variables provide insights into the socioeconomic and demographic landscape of California, which are critical for addressing the research question.

Variable	Description
Population Density	Concentration of population in a County
Median Income	Median household income in the past 12 months (in 2019 inflation-adjusted dollars)
White Population	Total Population Percentage of White alone
Poverty Population	Total Population Percentage whose Income in the past 12 months below the poverty level
Old population	Total population percentage of 85 years and over

TABLE I
VARIABLES AND DESCRIPTIONS

The primary response variable indicating wildfire event severity is the extent of land burned, calculated as the percentage of acres burned by wildfires in each county. Missing values were filled with a meaningful 0, indicating no wildfires occurred during 2021 [6]. To address skewness, this variable was cube-transformed. In cases where a wildfire spanned multiple counties, the total acres burned by that wildfire were evenly divided among the affected counties when calculating the total acres burned for each county. Since most of the wildfires in California occurred starting from July 2021, the temperature variable is recorded based on the average maximum temperature in July. Furthermore, the presence and distribution of CAL FIRE facilities are considered as variables indicating the level of fire protection available in different areas, incorporated as a count per county.

The analysis focused on data from the year 2021, thus the temperature, facility and wildfire datasets were subset to include only data from that year, and the ACS 5-year data was updated for 2021. In total, the dataset includes records for 160 wildfires and 1498 facilities across the entire state of California.

III. METHODS

A. Areal Analysis

An Ordinary Least Squares (OLS) linear regression was fit for complete spatial randomness testing, with the formula [11]:

$$Y = X\beta + \epsilon$$

where Y is the response variable, representing the cube-root transformed percentage of acres burned per county. X is the matrix of predictor variables, including population density, median income, number of fire facilities, etc. β is the vector of unknown coefficients to be estimated. ϵ is the error term, assumed to be independently and identically distributed (i.i.d) with a normal distribution with mean 0 and standard deviation σ.

Moran's I test [12] was employed to assess whether the residuals of the OLS model experience clustering, dispersion, or randomness in its spatial pattern. The test's null hypothesis is that the attribute being analyzed is randomly distributed in the study area, while the alternative hypothesis is the presence of spatial autocorrelation. The p-value associated with Moran's I test serves as the indicator of the strength of evidence against the null hypothesis. With the p-value being statistically significant, we would reject the null hypothesis, indicating the presence of spatial autocorrelation in the residuals.

Therefore, a Spatial Lag Model (SLM) was fit to account for spatial dependencies in the wildfire vulnerability data [13]. The SLM extends traditional regression models by incorporating a spatially lagged term of the response variable, which captures the influence of neighboring areas on the target area. The SLM is expressed as:

$$Y_i = \rho \sum_{j=1}^{n} W_{ij} Y_j + X_i \beta + \epsilon_i$$

where Y_i is the cube-root of percentage of acres burned in county i, ρ is the spatial autoregressive coefficient, W_{ij} is the spatial weight between counties i and j, Y_j is the cube-root of percentage of acres burned in county j, X_i is a vector of socioeconomic variables for county i, β is a vector of coefficients, and ϵ_i is the error term for county i.

The estimation process included the specification of a spatial weights matrix based on contiguity, specifically using a row-standardized Queen matrix. This choice was made to account for the varying numbers of neighboring counties each county has, as opposed to a binary matrix which would treat all neighboring counties equally regardless of the number of shared boundaries, or a Rook contiguity which may underestimate the influence of a potential neighboring county due to it's rigorous on the length of the border shared to be included as a neighbor. After specifying the spatial weights matrix, we employed maximum likelihood estimation to estimate the model parameters, including the spatial autoregressive coefficient ρ. To assess the significance of ρ, we conducted a likelihood ratio test, which compares the fit of the model with and without the spatial autoregressive term. This test provides a formal statistical evaluation of the importance of spatial dependence in the model, helping understand the role of spatial relationships in explaining the variability in acres burned by wildfires per capita across California counties.

The fitted models were evaluated using the Akaike Information Criterion (AIC), a metric for model comparison that balances model fit and complexity [14]. AIC considers both how well the model fits the data and the number of parameters in the model, aiming to find a balance where the model is parsimonious yet adequately explains the data. A lower AIC indicates a better balance between goodness of fit and simplicity. In our analysis, we sought to find the model that minimizes the AIC, as it indicates a better trade-off between model complexity and explanatory power.

B. Spatial Point Pattern Analysis

The wildfire incident data were transformed into a spatial point pattern based on the longitude and latitude coordinates of

the incidents. We conducted exploratory spatial data analysis on it, calculating the G-function and the F-function to assess the spatial clustering and dispersion of the wildfire incidents [15]. The G-function summarizes the distribution of distances from an arbitrary event to its nearest neighbor event. It provides a cumulative distribution function, indicating the proportion of events that have a nearest neighbor at a distance less than a given threshold. On the other hand, the F-function, also known as the Empty Space Function, serves as a measure of the average space between events, with more empty space indicating longer distances between a point and its nearest event. We repeatedly simulate the wildfire dataset under CSR for 500 times in the window and with the same intensity $\hat{\lambda}$, then compare these simulated datasets to the distribution of the original data, to understand whether there are spatial clustering or dispersion patterns of wildfire incidents.

Kernel Density Estimates were utilized to describe wildfire data in continuous space rather than aggregating [16]. The choice of kernel was the Gaussian kernel, as it offers a bandwidth (σ) parameter to regulate the spread. The log-likelihood function for an Inhomogeneous Poisson process (IPP) where the intensity varies spatially was then conducted for better understanding the relationship [17] [18], with the general expression as:

$$\log(L(\lambda)) \backsimeq \sum_{i=1}^{n} \log(\lambda(s_i)) - \int_W \lambda(s)\, ds$$

$$where \lambda(s) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

For parameter estimation, we aimed to utilize maximum likelihood estimation. The likelihood function involves the logarithm of the intensity function $\lambda(s_i)$ as an exponential function of β_0, β_1, and β_2, where X_1 and X_2 are covariates for each wildfire incident, which is dependent on the spatial location s_i. The integral of $\lambda(s)$ over a region W provides the expected number of IPP cases within that region.

Various combinations of spatial covariates were incorporated into several models applied to the wildfire incident data. The goodness of fit for each model was assessed by AIC to select the best-fitting one.

IV. Results

A. Explanatory Data Analysis

The spatial distribution of wildfire incidents and fire facilities in California is critical for understanding wildfire management. Figure 1(a) maps the locations of wildfire incidents and fire facilities in California. Wildfire incidents are represented by red points, with the size of the point corresponding to the acres burned in each wildfire. In contrast, fire facilities are represented by blue points. The spatial comparison between areas affected by wildfires and the proximity of fire facilities highlights potential gaps in coverage and areas where additional resources may be needed. Figure 1(b) further explores this by focusing on the spatial distribution of fire facilities in California counties, with color representing the number of fire facilities in each tract. This visualization aids in understanding the availability and distribution of fire suppression resources across the state, which is crucial for effective wildfire management.

Figure 1(c) illustrates the distribution of the percentage of acres burned in California counties for the year 2021. Using a cube-root scale due to the skewness of the data, the map showcases varying intensities of wildfire impact across different regions, highlighting regions such as Alpine and Amador with higher wildfire intensity relative to their size. Figure 1(d) illustrates the distribution of residuals after fitting a linear regression model. This maps demonstrate the potential spatial autocorrelation in the residuals, emphasizing the need for further analysis to understand whether there are spatial patterns of percentage of acres burned by wildfires in California.

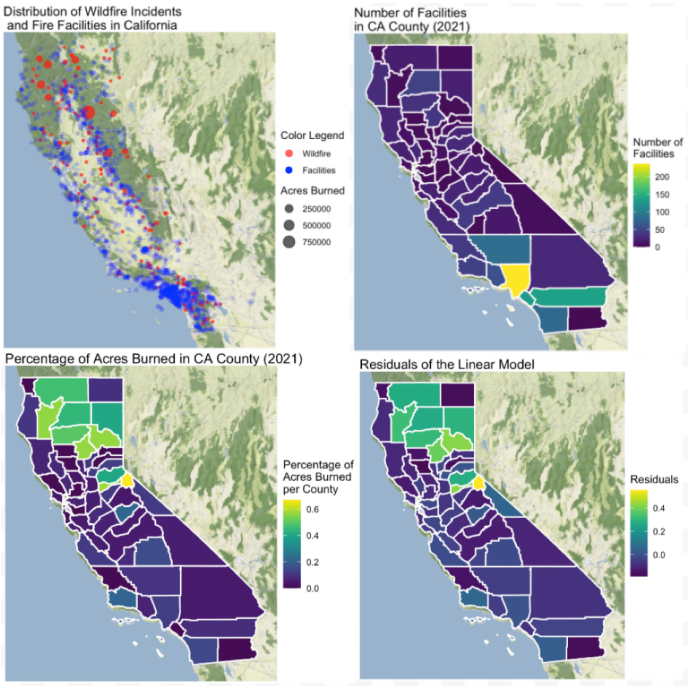

Fig. 1. (a) point plot that maps the locations of wildfire incidents and fire facilities;(b) spatial distribution of fire facilities; (c)distribution of percentage acres burned per county; (d) distribution of residuals of OLR

B. Areal Analysis

We fit an OLR model to the dataset:

acres burned percentage cube-root $\sim$ population density + percentage of poverty + percentage of old population + number of facilities + median income

and test for spatial structure in the residuals of the linear regression using Moran's I test. The Moran's I test statistic was calculated to be 0.23839, with a corresponding p-value of 0.002799, which indicates that there is evidence of spatial autocorrelation in the residuals of the OLS regression model, and the OLS model may not fully account for the spatial dependence in the data.

Two SAR models were fitted and compared:

Lag model1: acres burned percentage cube-root $\sim$ population density + percentage of poverty + percentage of old population + number of facilities + median income

Lag model2: acres burned percentage cube-root $\sim$ population density + percentage of poverty + percentage of old population + number of facilities + median income + percentage of white population

In this case, *Lag model1* has an AIC of -33.213, while *Lag model2* has an AIC of -35.109. Since the goal is to minimize the AIC, *Lag model2* is considered to be the better fit, suggesting that the inclusion of the percentage of white population variable improves the model's ability to explain the variation in percentage of acres burned by wildfires across California counties.

Variable	Estimate	Std. Error	z value	Pr(—z—)
(Intercept)	-3.9676e-01	2.9906e-01	-1.3267	0.18462
pop_density	-1.4615e-03	6.0268e-03	-0.2425	0.80840
perc_poverty	8.4071e-01	7.9672e-01	1.0552	0.29132
perc_old	-2.2299e+00	3.8650e+00	-0.5769	0.56398
num_facilities	4.2264e-04	5.1041e-04	0.8280	0.40765
med_income	1.5170e-06	1.7242e-06	0.8798	0.37896
white_perc	4.2601e-01	2.0586e-01	2.0694	0.03851

TABLE II
SPATIAL LAG MODEL RESULT

As Table 2 shown, the coefficients in the SAR model provide insight into the relationship between percentage of acres burned by wildfires and the predictor variables across California counties. The statistically significant predictor is the percentage of white population, which has a positive coefficient of 4.2601, suggesting that a higher percentage of white population is associated with an increase in percentage of acres burned.

This unexpected finding may suggest underlying factors related to fire risk or suppression resources that are not accounted for by the other variables in the model. It's possible that areas with a higher percentage of white population have different land use patterns, housing types, or fire management practices that contribute to increased wildfire risk. Yet this result aligns with our expectations to some extent. By examining Figure 1(c), we can see that the area near Yuba City is a potential hotspot for wildfires. In 2021, Yuba City, CA had 2.04 times more White (Non-Hispanic) residents (27.7k people) than any other race or ethnicity [19]. Additionally, areas with a higher white population tend to be relatively affluent, and having a large wooded lot is a luxury few can afford. Therefore, the wealthier areas have more woods, which increases the likelihood of wildfires. Further investigation into the specific mechanisms behind this association is necessary to better understand the complex interactions between demographic factors and wildfire risk.

C. Point Process Analysis

In this analysis, we examine how wildfires are distributed across California, first using statistical techniques to understand if these incidents are randomly scattered or exhibit clustering. As shown in Figure 2, the observed G function is above the upper simulation envelope, which suggests clustering, meaning that wildfire incidents tend to be closer to each other than expected under a complete random distribution. The observed F function falls below the lower envelope, which suggests that there is a greater distance from one wildfire incident to its nearest neighbor event. This aligns with the G function that there is potential clustering of the locations of wildfire incidents in California. The Kernel Density Estimation(Figure 3) likewise shows that a high number of wildfires occurs in southern California, such as Los Angeles and Palm Springs, with a significant number of incidents also occurring in the northern central and central regions spanning the entire state where seems to be the Eastside of Sacramento Valley and the San Joaquin Valley.

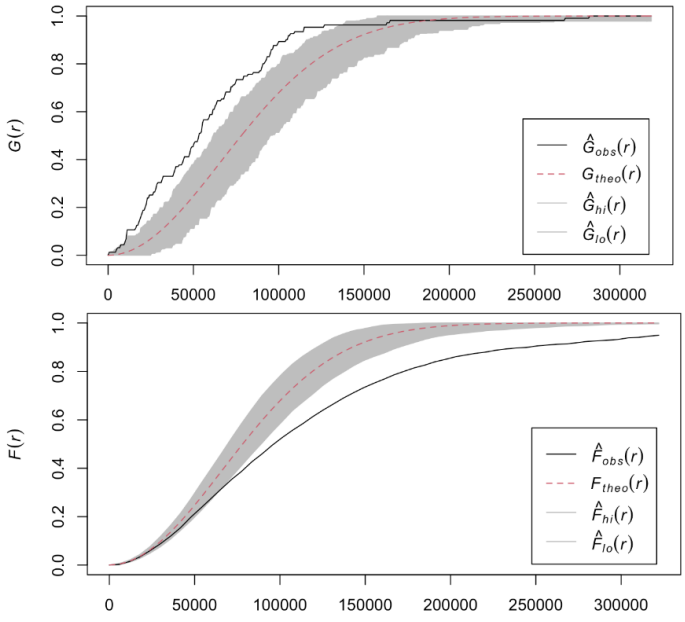

Fig. 2. G and F Function Envelope

2 IPP models were fitted and compared:

Model1: trend = temperature + number of facilities + total population

Model2: trend = temperature + number of facilities + total population + percentage of white population

Model1 has an AIC of 8003.6, while *Model2* has an AIC of 8005.2. Since the goal is to minimize the AIC, *Model1* is considered to be the slightly better fit.

By comparing the KDE with the fitted trends in Figure 3, we can also observe that *Model1* better represents the trend in KDE, while *Model2* seems to overestimate the intensity of wildfire incidents in regions other than the South Coast. Specifically, the SE plot displays the standard errors associated with the estimated intensity function at different spatial locations. For areas like the North interior, *Model2* showcases higher standard errors compared to *Model1*, indicating greater uncertainty in the estimated intensity, which suggests less

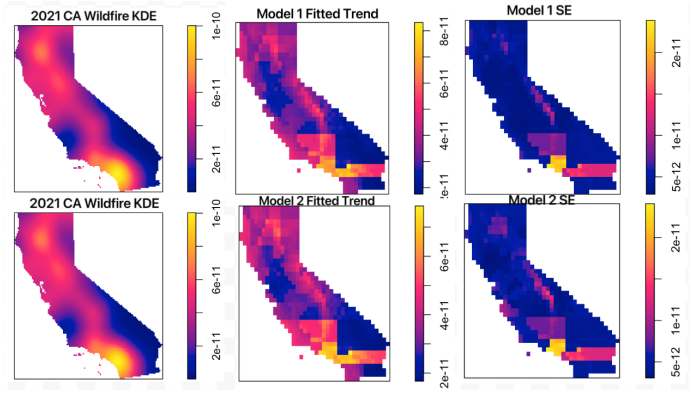

Fig. 3. California fire incidents KDE, Model 1 fitted trend and SE(top row); California fire incidents KDE, Model 2 fitted trend and SE(bottom row)

certainty about the expected number of wildfire incidents in those locations.

The results in Table 3 show the estimated coefficients for the variables in the model. The temperature of the county has a statistically discernible negative effect on the spatial intensity of wildfire incidents in California after controlling for population and number of fire facilities. As the temperature of the county increases, it's less likely that wildfire incidents occur. This result seems counterintuitive, as higher temperatures are generally associated with increased wildfire risk. However, despite areas that are near the east side of the valleys or counties such as Los Angeles experience relatively lower temperatures compared to the central of the valley, the weather conditions are still severe enough to make them conducive to wildfire spread. Additionally, most of Silicon Valley where the temperature is the most extreme is dense with roads and houses, and is primarily grassland, so a forest fire couldn't sweep through the way it does in a forest. But there are plenty of areas on the east fringes of the Valley that are heavily wooded with oaks, therefore more prone to wildfire [20]. Other wooded areas like the South Coast, the North Interior, and the central part of Great Basin, are also at risk. Therefore, temperature alone may not fully capture the complex interplay of factors that contribute to wildfire risk, such as city construction patterns and vegetation type.

Variable	Estimate	Std. Error	z value	Pr(—z—)
(Intercept)	-2.3524e+01	3.4784e-01	-2.4206e+01	-2.284e+01
temp	-3.4209e-02	1.5570e-02	-6.4727e-02	-3.6913e-03
num_facilities	1.0106e-02	2.9097e-03	4.4034e-03	1.5809e-02
total_pop	-1.4914e-07	7.3441e-08	-2.9308e-07	-5.2042e-09

TABLE III
IPP MODEL RESULT

The number of facilities of the county has a statistically discernible positive effect on the spatial intensity of wildfire incidents in California after controlling for population and temperatures. As there are more facilities in the county, it's more likely that wildfire incidents occur. One would commonly expect that more facilities dedicated to fire suppression would lead to a decrease in the intensity of wildfires. However, this positive effect could be explained by several factors. Firstly, the presence of more facilities might indicate areas with a higher historical risk of wildfires, leading to more facilities being established in those locations. Furthermore, the number of facilities alone may not capture the effectiveness or capacity of fire suppression efforts. It's possible that areas with more facilities still face challenges such as inadequate resources, difficult terrain, or high fire danger conditions, leading to a higher intensity of wildfires despite the presence of these facilities.

The total population of the county has a statistically discernible negative effect on the spatial intensity of wildfire incidents in California after controlling for temperature and number of fire facilities. As the county has a higher population, it's less likely that wildfire incidents occur. This finding aligns with expectations. Higher population density tends to correspond to more urbanized areas, where land use is less likely to involve activities that can trigger wildfires, such as agricultural practices or outdoor burning. Besides, areas with higher population densities often have more resources and infrastructure dedicated to fire prevention, detection, and suppression.

Overall, the result underscores the importance of considering the human-environment interaction in wildfire risk assessment. While natural factors like temperature play a role in fire behavior, human activities and infrastructure significantly influence the spatial distribution of wildfires. Policy and planning efforts aimed at mitigating wildfire risk should take into account not only environmental factors but also demographic and land-use patterns to effectively target prevention and response measures.

V. DISCUSSION

Based on the findings of our analysis, which sought to understand the factors influencing wildfire vulnerability in California, several key insights emerge. Firstly, we identified significant correlations between socioeconomic and demographic factors and wildfire vulnerability, that areas with a higher percentage of white residents tend to experience a higher percentage of acres burned. This relationship could be attributed to wealthier areas with a higher percentage of white population having more wooded lots, increasing the likelihood of wildfires. However, further investigation is needed to fully understand these complex interactions and the underlying mechanisms driving this relationship. Additionally, we found that counties with a higher number of CAL FIRE facilities located, lower total population, or cooler temperatures are prone to experiencing a greater number of wildfires. However, it is important to interpret these results with caution, taking into consideration factors such as land use and regional wildfire history.

Despite these insights, several limitations of our analysis should be acknowledged. The correlational nature of the data limits our ability to draw causal inferences, and there may be unmeasured variables or external factors that could influence the results. Furthermore, our analysis focused solely on the

influence of the number of facilities per county and did not consider other factors such as staffing levels or equipment availability, which could impact the effectiveness of CAL FIRE facilities.

Looking ahead, future research could explore the impact of additional climate variables, such as precipitation, wind speed, and vegetation types, on wildfire vulnerability. Understanding the interplay between natural factors and human activities is crucial for informing effective policy and planning efforts to mitigate wildfire risks in California.

REFERENCES

[1] Liu, J. C., Wilson, A., Mickley, L. J., Dominici, F., Ebisu, K., Wang, Y., Sulprizio, M. P., Peng, R. D., Yue, X., Son, J. Y., Anderson, G. B., & Bell, M. L. (2017). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. *Epidemiology (Cambridge, Mass.)*, 28(1), 77–85. https://doi.org/10.1097/EDE.0000000000000556

[2] Abatzoglou, J. T., Williams, A. P., & Barbero, R. (2019). Global emergence of anthropogenic climate change in fire weather indices. *Geophysical Research Letters*, 46, 326–336. https://doi.org/10.1029/2018GL080959

[3] de Diego, J., Fernández, M., Rúa, A., et al. (2023). Examining socioeconomic factors associated with wildfire occurrence and burned area in Galicia (Spain) using spatial and temporal data. *Fire Ecology, 19*, 18. https://doi.org/10.1186/s42408-023-00173-8

[4] Syphard, A. D., Keeley, J. E., Pfaff, A. H., & Ferschweiler, K. (2017). Human presence diminishes the importance of climate in driving fire activity across the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 114(52), 13750–13755. https://doi.org/10.1073/pnas.1713885114

[5] Hawbaker, T. J., Radeloff, V. C., Stewart, S. I., Hammer, R. B., Keuler, N. S., & Clayton, M. K. (2013). Human and biophysical influences on fire occurrence in the United States. *Ecological Applications: a publication of the Ecological Society of America*, 23(3), 565–582. https://doi.org/10.1890/12-1816.1

[6] CAL FIRE. (2021). Title of dataset [CAL FIRE Facilities for Wildland Fire Protection]. CAL FIRE. Retrieved from https://www.fire.ca.gov/incidents

[7] Open Data California. (2021). Title of dataset [Incident Archive]. Open Data California. Retrieved from https://data.ca.gov/dataset/cal-fire-facilities-for-wildland-fire-protection/resource/684b8b5b-31b6-4430-830a-a93895b8ffd1

[8] WorldClim. (n.d.). Retrieved from https://worldclim.org/data/index.html

[9] Hijmans, R. J. (2022). Raster: Geographic data analysis and modeling. Retrieved from https://cran.r-project.org/web/packages/raster/raster.pdf

[10] United States Census Bureau. (2021). American Community Survey (ACS) API [API]. Retrieved from https://api.census.gov/data/2022/acs/acs5/variables.html

[11] Gujarati, D. N., & Porter, D. C. (2009). Basic econometrics. McGraw-Hill Education.

[12] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.

[13] Anselin, L. (1988). Spatial econometrics: Methods and models. Kluwer Academic Publishers.

[14] Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press.

[15] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

[16] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Møller, J., & Waagepetersen, R. (2004). Statistical inference and simulation for spatial point processes. Chapman and Hall/CRC.

[18] Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Applied spatial data analysis with R. Springer Science & Business Media.

[19] Data USA. (n.d.). Yuba City, CA. Retrieved from https://datausa.io/profile/geo/yuba-city-ca/

[20] Fenn, M. E., Allen, E. B., Weiss, S. B., Jovan, S., Geiser, L. H., Tonnesen, G. S., ... & Bytnerowicz, A. (2010). Nitrogen critical loads and management alternatives for N-impacted ecosystems in California. *Journal of Environmental Management*, 91(12), 2404-2423. https://doi.org/10.1016/j.jenvman.2010.07.034