Spatial Statistical Modeling of Wildfire Incidence in Arizona

STAT574E | Group 3

Raymond Owino, Alex Salce, Matthew Wallace

December 2024

Executive Summary

The *Wildland Fire Incident Locations* dataset we selected for study (provided by the National Interagency Fire Center) contains spatial coordinates of a wildfire's origin and its resulting size in acreage (among several other useful attributes) for all wildfire incidence in the US as archived in the IRWIN system since 2014. Our research goals included the study of factors that may influence the resulting size of a wildfire, and to build spatial models that may aid as a tool for wildfire risk assessment and resource allocation, forecasting, and other possible real-world applications. We were also interested whether wildfire incidence, when treated as a point pattern, occurs as a completely spatially random process, or if there really is inhomogeneity in spatial intensity for incidence and thus truly higher risk in some spatial regions than others.

Attributes contained in the dataset allowed us to take three different approaches to spatial statistical modeling. Using the incident size as a continuous response, we built a spatial linear model that predicts wildfire size based on spatial and selected covariate inputs. Filtering the data using incident size as well as other incident attributes also allowed us to treat wildfire incidence data as a point process. We fit a Log-Gaussian Cox Process (LGCP) spatial model to study spatial intensity, as well as a binary response spatial logistic regression model to map spatially the regions of highest probability of large wildfire (≥ 1000 acres) incidence in the region of study.

Our modeling efforts studied Coconino County (CC) in Northern Arizona, which has the highest overall wildfire incidence as compared to the rest of the state. We gathered data for incidence proximity to roads, environmental, and census data corresponding to spatial coordinates to use as covariates for our models to support our research questions. These covariates improved performance of our models by AIC metrics. Our LGCP model indicated some interesting patterns of CSR/non-CSR for all types of wildfire incidence, notably non-CSR at the AZ state level, and possible evidence for CSR in only CC. All models built will require some further refinements, for which we have identified several potential remedies, but each serves as a solid foundation and proof of concept for our real-world application goals.

Introduction

Wildfire incidence data describes the spatial origin of a wildfire, and is one of many ways that wildfires can be studied; it is also a type of data that lends itself well to be studied by methods in spatial statistical analyses. Spatial statistical models can be valuable in real applications for wildfire prevention, giving inferential insights into wildfire risk as well as tools to aid in resource allocation for mitigation efforts and planning. Our project seeks to build models that could be useful for these kinds of efforts, and our research questions aim to provide insight to what kinds of environmental and human factors can influence the size of a resulting wildfire based upon where it originates spatially.

Wildfire Incidence Data

Accompanying this paper is `wfigs_az_sf_EPSG32612.RData` which imports `wfigs_az_sf` to the R environment, an `sf` object with 18089 observations of wildfire incidence in the state of Arizona as well as corresponding covariate data for each entry (to be discussed later). This data was originally acquired via Wildland Fire Incident Locations from the National Interagency Fire Center. It contains spatial point data indicating the origin of each wildfire

recorded in the IRWIN database (since 2014), and includes many useful data attributes for each entry. The attributes from this dataset that we used were

- `x` and `y` | Spatial coordinates in lat/lon
- `IncidentSize` | Size of the resulting wildfire in acres
- `FireCause` | Human, Natural, Unknown, Undetermined
- `FireDiscoveryDateTime` | Date & time of incident reporting
- `IncidentTypeCategory` | WF (wildfire) or RX (prescribed burn)

Other covariate data that was captured for each incidence point will be discussed later in the report.

Research Questions

Our research questions were driven by our curiosities about the ways in which we could approach modeling the data. We were also motivated in part by our findings during our explorations, and we wanted to evaluate whether real data we felt could influence wildfire size could serve as useful covariate information for our models. Additionally, we wanted to evaluate spatial randomness for different types of incidence to be sure what we are investigating is more than a completely spatially random process. In general, we found this to be a particularly interesting dataset because it seemed as though there was more than one way to approach its analysis with tools we have learned in this course.

In what ways can we approach spatial modeling of this data to produce useful insights? [RQ1]

Our overall goal is to find models that could support wildfire risk assessment, tools for resource allocation, and prediction for wildfire size for forecasting. This got us thinking about how we may approach our data in different ways. During our data explorations, we rationalized different ways we could model the dataset. We determined that we could use the data to build the following models that we will detail throughout the report.

- Spatial Linear Model - used to predict wildfire size for input coordinates and conditions found to affect size.
- Log-Gaussian Cox Process (LGCP) - used to determine whether a type of wildfire (e.g. large human caused fires) have characteristics of spatial randomness, or if they are not random, and we can study where higher-risk areas of incidence are spatially.
- Binary response GLM logistic regression - model to be used to assess risk in terms of the model's chosen risk factors, with outputs expressed as probabilities of the incidence in question spatially.

Can we find useful covariate data that can improve our models? [RQ2]

Based on combinations of findings from data explorations and our intuition about what could influence wildfire size, we chose to collect covariate data for the following factors.

- Proximity to roads - Is a fire's size influenced by how accessible it is via roads? Do more remote fires tend to get bigger?
- Environmental factors - can factors like temperature and precipitation influence wildfire size? Can terrain factors like forest and grass coverage, or local steepness influence?
- Population density - do areas of more concentrated human settlement influence wildfire incidence?

Are the patterns of human or non human caused fires spatially CSR, or do they exhibit an inhomogeneous spatial intensity? [RQ3]

An extension of the LGCP model, we wanted to investigate whether certain types of fires occur spatially randomly, or if they exhibit characteristics to the contrary? For example, we may find large human caused fires are not spatially random, and a risk model from an LGCP fit could help determine higher-risk areas of such incidence spatially.

Exploratory Data Analysis

Note: all spatial data was projected to the UTM Zone 12N coordinate reference system (EPSG:32612) for analysis. Plots shown in latitude and longitude representations are for visual reference only. Please refer to supplemental diagrams and data explorations in each appendix for complementary data and visuals to this discussion; Appendix A for environmental data, B for population density, C for roads.

Wildfire Incidence Types, Coconino County

The full set of incidence data for Arizona contained 18089 total observations, 10174 of which human caused, 5409 naturally occurring, and 2210 unknown (using the `FireCause` attribute in the dataset). We opted to discard unknown cause data to avoid studying uncharacterized incidents. The `sf` function `st_intersection` allowed us to subset incidence data `sf` objects spatially to Coconino County (CC) for more focused analysis, which has the highest count of wildfire incidence in the state with 3824 incidents (1692 Human, 2128 Natural).

The selection of covariates was, in part, influenced by some further subsetting that could be done using the `IncidentSize` attribute. We were interested in studying differences between "large wildfires" and "small wildfires", using `IncidentSize` to threshold our data. The chosen threshold for large wildfires was `IncidentSize`≥ 1000 acres (66 natural, 6 human caused in CC), or class F and G wildfires (the largest) as defined by Short (2014). Small wildfires were all others (2062 natural, 1686 human caused).

Proximity to roads and Population density

During visual explorations of the data, some distinct patterns emerged that appeared related to human activity. Human caused wildfire incidence showed clear concentration around more densely populated areas, and the outlines of roads are clearly visible. Both of these features are not present for naturally caused fires.

We opted to include the distance in meters of an incident location to the nearest road as covariate data for each point. We wanted to explore whether this distance from a road had any predictive power for the resulting fire size. Intuitively, we felt that it may be a good measure of "accessibility" to a fire; if a fire starts and is far away from a major road, valuable time to contain the fire may be lost and the fire may grow out of control.

The `roads()` function in the `tigris` package was used to generate `sf` objects for AZ roads, and `st_distance()` function in the `sf` package helped us generate this data to be used as a covariate. Each `roads` geometry has a corresponding classification code according to the Census Bureau, and we chose to include "Primary Roads" (`S1100`) and "Secondary Roads" (`S1200`) to be representative of what we will call "major" roads. Separately, we used "Vehicular Trail (4WD)" (`S1500`) roads to represent "remote" roads. The covariate data measures an incident to its nearest major road and nearest remote road, and the histograms in Appendix C bin incidence by which type of road is nearest (and how far away it is), to give a visual measure of remoteness of the incident type in question. At both the full state level and the Coconino County level, human fires tend to be less remote (more concentrated near major roads) and natural fires tend to be much more remote. The distribution of distance from *any* roads for large wildfires indicates that while there is a small concentration near major roads, many are far away from major and remote roads, indicating that there may be a relationship between distance from roads and large wildfire incidence.

Environmental factors

.

Statistical Analyses

Fixed, Continuous Response

Spatial Linear Model

$$\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{e} \quad \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$$

Point Process

Log-Gaussian Cox Process

$$\log(\lambda(u)) = \boldsymbol{Z}(u)\boldsymbol{\beta} + e(u), \quad e(u) \sim \mathrm{N}(0, C(\boldsymbol{\theta})), \quad C(u, u') = \sigma^2 e^{-||u-u'||/h}$$

Binary Response GLM Logistic Regression

$$\mathrm{logit}(\lambda_1(\boldsymbol{s})) = \mathbf{x}(\boldsymbol{s})^T\boldsymbol{\beta} + e(\boldsymbol{s}) + \log(\lambda_0), \quad Y(\boldsymbol{s}) \sim \mathrm{Bern}(p(\boldsymbol{s})), \quad \mathrm{E}[Y(\boldsymbol{s})] = p(\boldsymbol{s}) = \frac{\lambda_1(\boldsymbol{s})}{\lambda_0(\boldsymbol{s}) + \lambda_1(\boldsymbol{s})}$$

Conclusions

References

Dumelle, Matt AND Ver Hoef, Michael AND Higham. 2023. "Spmodel: Spatial Statistical Modeling and Prediction in r." *PLOS ONE* 18 (3): 1–32. https://doi.org/10.1371/journal.pone.0282524.

Short, K. C. 2014. "A Spatial Database of Wildfires in the United States, 1992-2011." *Earth System Science Data* 6 (1): 1–27. https://doi.org/10.5194/essd-6-1-2014.

Zimmerman, & Ver Hoef, D. L. 2024. *Spatial Linear Models for Environmental Data*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429060878.