Environmental Statistics Final Project Proposal - Group 3

Matthew David Wallace, Raymond Owino, Alex Salce

November 15th, 2024

Topic and Datasets

Primary topic question: **"What influences the size of a wildfire in Arizona?"**

Our primary research questions will be centered around wildfire data in Arizona, and will be utilizing (if applicable, and not limited to) the following datasets.

- NIFC | Wildland Fire Incident Locations
  - The primary dataset for our research questions will utilize the Wildland Fire Incident Locations dataset. This dataset includes 97 variables including point location data for wildfire incidence in the US from 2014 to the present as recorded from the IRWIN (Integrated Reporting of Wildland Fire Information) system, which aggregates fire data from many data sources. The dataset contains many features, but we will primarily utilize the point data for fire origin, as well as total acreage the fire went on to burn. Some of the assigned tasking will include investigation of other possibly useful features in the dataset.
  - About this dataset
- `tigris` package, `roads()`
  - `tigris` function `roads()` has `sf` data for roads in AZ. From `?roads`: *From the Census Bureau: "The content of the all roads shapefile includes primary roads, secondary roads, local neighborhood roads, rural roads, city streets, vehicular trails (4WD), ramps, service drives, walkways, stairways, alleys, and private roads."*
- NIFC | WFIGS Interagency Fire Perimeters
  - This dataset is the best available fire perimeter data for individual wildfires in the US. Our research may necessitate a deeper dive into some of the individual wildfires and the areas they covered spatially.
- LiDAR data
  - The LiDAR data should give us features of the terrain that can be used as predictor variables for our model fits.
- rFIA data
  - We may incorporate abundance data for standing trees (using `TPA & BAA` data) as predictor variables for our model fits.

Research Questions

Our research questions are derived from our overall topic question: **"What influences the size of a wildfire in Arizona?"**

(1) Does the proximity of a wildfire's origin to a road influence its resulting size?

- *Does proximity to major or smaller roads influence the resulting size of a wildfire? Is proximity from roads a useful predictor in whether a wildfire get "big"?*

(2) Can spatial natural/human factors serve as useful predictors in modeling resulting wildfire size?

- *Are natural factors useful predictors for the size of wildfires? Precipitation, temperature, and season affect the resulting size of a wildfire? Does the terrain (topology, forest abundance, etc.) influence the resulting*

size of a wildfire? Does population density affect resulting size of a wildfire?

(3) Are the patterns of human or non human caused fires spatially CSR, or do they exhibit an inhomogeneous spatial intensity pattern?

- *Do the patterns human-caused wildfires seem to arise according to a homogeneous Poisson Process? Do the non-human cause wildfires seem to arise from a HPP?*

Our research questions cover a lot of detail that we believe we can recover, but may be modified as we implement our data and models.

Statistical Methods

To approach our research questions, our primary dataset we will utilize will be the *Wildland Fire Incident Locations* data, which provides the data that will be the foundation for our modeling efforts. Additional data will be acquired as outlined in Team Responsibilities.

Data for model construction

Wildland Fire Incident Locations

- `IncidentSize`: Size of resulting fire in acres
- The location data for fires:
  - `InitialLatitude` and `InitialLongitude`: Coordinate data for the reported origin of the fire in the IRWIN system
- Other relevant data in dataset:
  - `FireDiscoveryDateTime`: Date and time of fire discovery
  - `IncidentTypeCategory`: Categorized as a Wildfire (WF), Prescribed Fire (RX), or Incident Complex (CX) record
  - `POOCounty`: County in which fire originated

`tigris` package, `roads()`

`roads(state = "AZ")` returns an `sf` object.

- `geometry`: `LINESTRING` geometry for roads in AZ
- `MTFCC`: classification of roads, see here for road class code descriptions

Other predictor data that we seek (natural factors, population density), will be joined on available factors depending upon the data. For example, we hope to implement the terrain slope grade within a 1km radius of coordinates, so that data would be joined on the wildfire coordinates as a predictor.

Models

The statistical methods we plan to use to address answer these questions are the following.

Spatial linear model

Response

- Raw `IncidentSize` data: continuous, fixed spatial data response (Geostatistical).

Predictors

- Closest road (possibly filtered by class) as calculated from something similar to (if not exactly) the geosphere::dist2line function. Ideally, a function similar to `min_dist` of HW4 could be used here [Continuous]
  - What kind of relationship, if any, does distance to nearest roads have with wildfire size?
- `FireDiscoveryDateTime`: time by month, season, or possibly exact time started [Discrete or Continuous]
- Population density data [Continuous]
  - What kind of relationship, if any, does population density have with wildfire size?
- Natural factors [Discrete or continuous]
  - What kind of relationship, if any, do natural factors have with wildfire size?

Example (rough/hypothetical) of a model fit: `wf_splm <- splm(IncidentSize ~ FireDiscoveryDateTime + PopDensity + SlopeGrade + Precip + Abundance + min_dist, data=wf_data, spcov_type = "matern", estmethod = "sv-wls")`

This model will provide insight to questions 1 and 2, as we will be able to investigate the influence of distance from roads as well as human/natural factor influence on size of a wildfire.

Point process model

Response The response data for the point process model will use the `IncidentSize` data to create a binary response of whether the fire was "not large" or "large" (using 0 and 1 response values, respectively). The threshold for a "large" fire will be analyzed and could be modified as part of the analysis. We may also filter data to only human caused fires or non-human caused for analysis.

Predictors The same predictors can be used as the linear model.

Example (rough/hypothetical) of a model fit: `wf_kppm <- kppm(IncidentSize ~ FireDiscoveryDateTime + PopDensity + SlopeGrade + Precip + Abundance + min_dist, data=wf_data, clusters = "LGCP", model = "exponential")`

We can analyze the theoretical $F(r)$ and $G(r)$ against the observed $F(r)$ and $G(r)$ to assess whether the data are CSR, and use the predicted $F(r)$ and $G(r)$ from the model to compare how well our model fits.

Alternative model for binary response data - Binomial response GLM Since in this case our response data is binomial, we can use GLM fit for analysis for a comparable fit to our `splm` model in previous example.

Example (rough/hypothetical) of a model fit: `wf_spglm_binom <- spglm(IncidentSize ~ FireDiscoveryDateTime + PopDensity + SlopeGrade + Precip + Abundance + min_dist, data=wf_data, family = "binomial", spcov_type = "matern")`

This model should give additional insight to road proximity, natural factors, and human influence, with a risk-style model, identifying regions of AZ that have higher probability of large wildfires. Essentially another angle to approach the problem than the spatial lm.

Team Responsibilities

Project Area	Raymond	Matthew	Alex	All
Data	Find natural factors data	Find population density data (possibly R package available)?	Determine methodologies for calculated data like proximity to roads	Determine pertinent factors to our questions and study their background / "Explorations" (histograms, other visuals)
Code	Spatial Linear model fitting & outputs (plots etc)	Point Process model fitting & outputs (plots etc)	Initial model fits / Data combination and wrangling for model inputs	Combining results and model refinements
Collaboration	Meeting scheduling/zoom host	Meeting minutes primary POC for group status	Set up GitHub repo for data/analysis repository	Asynchronous comms on Slack
Deliverables	Final presennation video	Slide Deck for final presentation video	Final Report for submission	Contributions to each member's area (i.e. we will each do slides & give to Matt / record video portions and give to Raymond for editing / write sections of report and give to Alex)

We held a group discussion to deliberate all of the above general tasking, but there is mutual understanding that we will be helping each other for each portion. For example, each of us "owns" a deliverable, but that will only mean that the assignee will be the primary POC for putting together the deliverable. Our plan is to meet at least approximately weekly on Zoom to keep up with each other's progress and make sure we are on task/schedule.

Each member of the group "owns" a part of the analysis, for example Raymond will be collecting and exploring all of our climate factor data, Matt has population density data, Alex has the primary wildfire point data. We will each be responsible for choosing factors we want to include, researching the data to the best extent possible (how it was collected, units, etc.), and do explorations of the data factors we choose to get general ideas of why they will be useful. Each member of the group will own modeling tasks as well. The final presentation sections and report contributions will be assigned according to each member's responsible area. We will be exchanging information and data via GitHub and Slack.

Anticipating Challenges

Challenges that we anticipate:

- Combining data, cleaning data, and filtering data
  - We intend to use data from multiple sources, which will present challenges for combining
- Deciphering available data
  - These datasets have many variables included, some of which will require researching to better understand whether they can be useful in answering any of our research questions or providing useful predictor data for our models.
- Dataset size
  - Can we reasonably answer all of our research questions, given that we are working with fairly large

datasets? Or due to unforseen challenges joining or generating needed data? We may encounter computational challenges getting our models to work, given that we will be working with so much data. If we do run into challenges, we may need to refine our research questions.

- Available data
  - Will we have enough data available to utilize to fit the models we hope to support our research questions with the data that we currently have? We may need to acquire more data, or refine our research questions.