Wildfire Incidence in Arizona

Group 3 | STAT574E Final Project

Raymond Owino, Alex Salce, Matthew Wallace

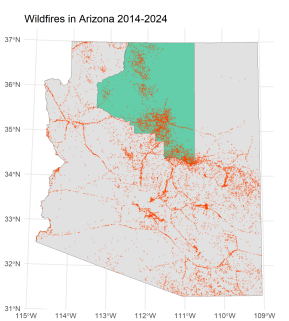
2024-12-11

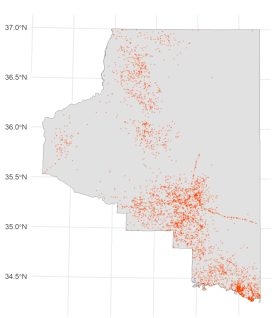
Wildfire Incidence in Arizona

Our project investigates different approaches to spatial statistical modeling of wildfire incidence data in Arizona, with a focus on Coconino County (Northern AZ).

The “Wildfire Incidence” data we will be studying specifies coordinates for a fire’s origin, and its resulting size in acres.

Overall practical goal: modeling wildfire incidence to aid in prediction/assessment of wildfire risks in AZ based on relevant information.





Wildfire Incidence Dataset

Our incidence data uses the **National Interagency Fire Center** Wildland Fire Incident Locations dataset.

This dataset includes point locations and corresponding data for all wildland fires in the United States reported to the IRWIN system, which aggregates wildfire incidence data from around the country.

The dataset has all IRWIN data entries since 2014 (it is still updated daily).

There are over 300K records in this dataset, and just over 18K in Arizona.

Wildfire Incidence Dataset

Refinement of this dataset utilized the following data attributes (of 96 columns) for incidence filtering & covariates.

- `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)

Research Question 1

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

Wildfire Incidence Data - Continuous, Fixed Spatial Index

This dataset offered us a unique opportunity to take multiple approaches to spatial statistical modeling. Using the available data attributes, we can approach our data from different angles.

- **Continuous, fixed** - Treating `x` and `y` as coordinates as fixed and using `IncidentSize` as a continuous response, our data is geostatistical.

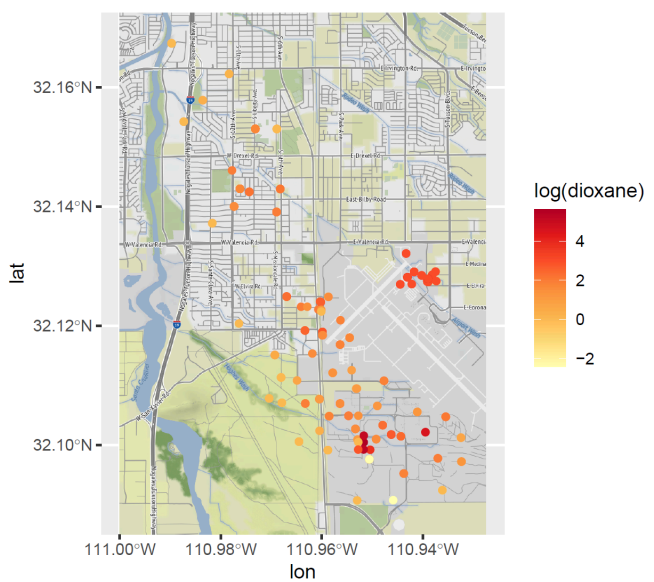
Modeling approach

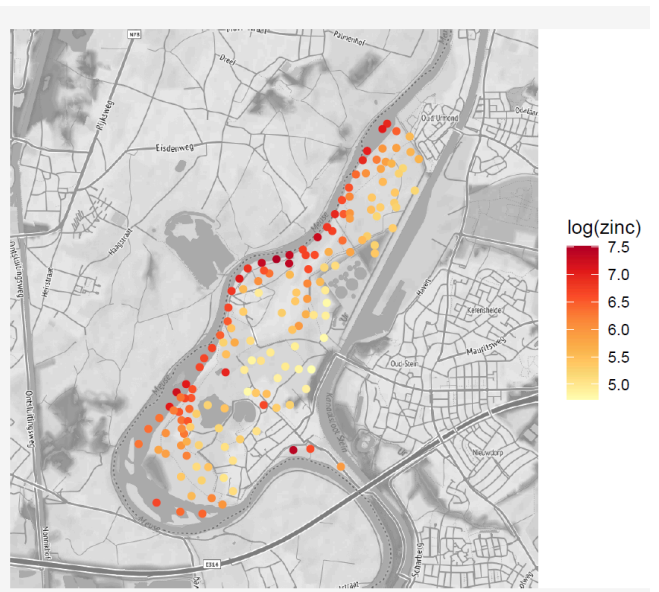
- **Spatial Linear Model** - Module 2, similar to Holland example in *Continuously indexed spatial data (geostatistics)*, or the `dioxane` analysis in HW2.

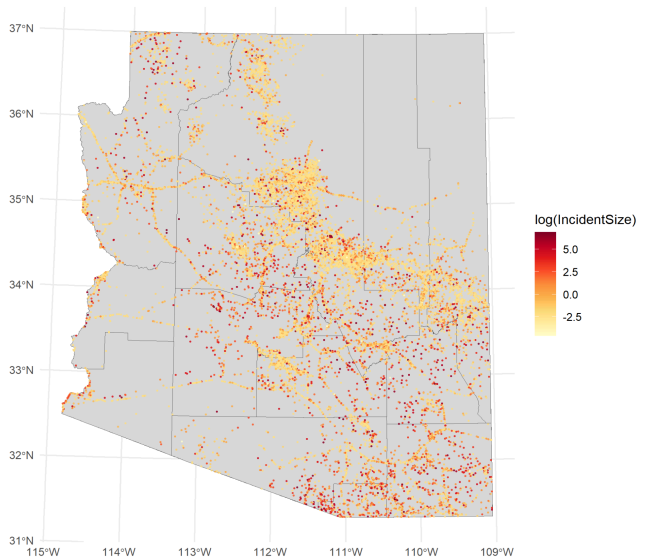
$$\mathbf{y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{e} \quad \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$$

Wildfire Incidence Data - Continuous, Fixed Spatial Index

Spatial Linear Model Data







Wildfire Incidence Data - Point Process 1

This dataset offered us a unique opportunity to take multiple approaches to spatial statistical modeling. Using the available data attributes, we can approach our data from different angles.

- **Point process** - Using `x` and `y` as coordinates and `IncidentSize` as a *threshold*, we can study "large wildfires" (`IncidentSize` ≥ 1000 acres) as point process data.

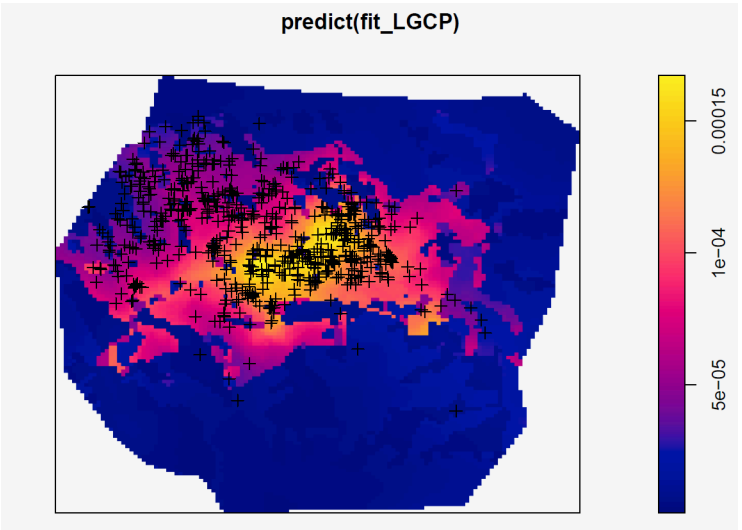
Modeling approach

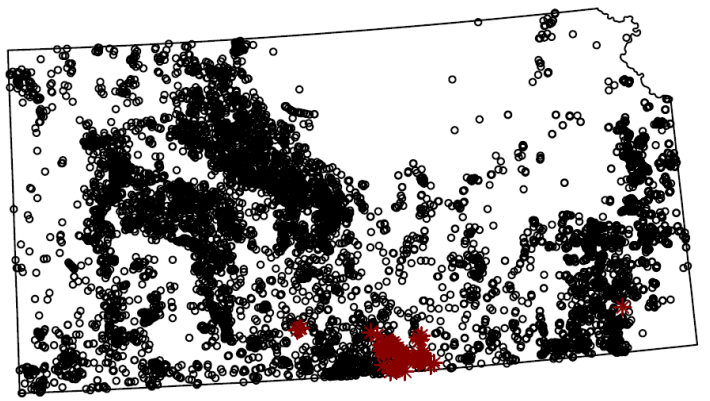
- **Log-Gaussian Cox Process** Module 4, similar to Gorillas LGCP example in *Random spatial index (point pattern)*, or the `earthquakes_um` analysis of Kansas in HW4.

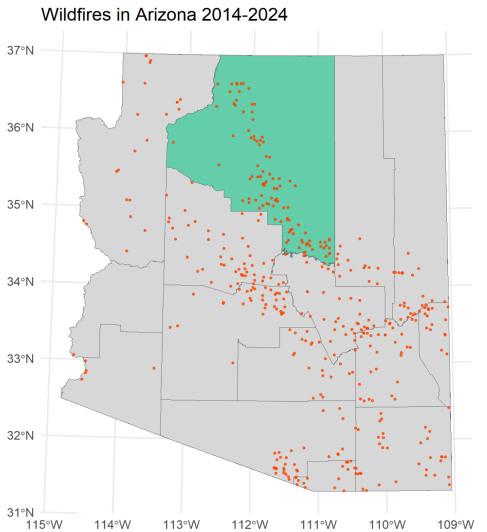
$$\log(\lambda(u)) = \boldsymbol{Z}(u)\boldsymbol{\beta} + e(u), \quad e(u) \sim N(0, C(\boldsymbol{\theta})), \quad C(u, u') = \sigma^2 e^{-||u-u'||/h}$$

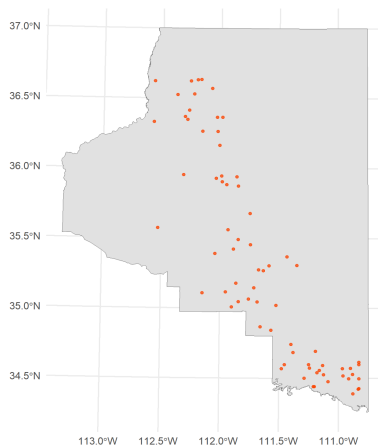
Wildfire Incidence Data - Point Process 1

Log-Gaussian Point Process Model Data









Wildfire Incidence Data - Point Process 2

This dataset offered us a unique opportunity to take multiple approaches to spatial statistical modeling. Using the available data attributes, we can approach our data from different angles.

- **Point process** - Using `x` and `y` as coordinates and `IncidentSize` as a *threshold*, we can study "large wildfires" (`IncidentSize` ≥ 1000 acres) as point process data.

Modeling approach

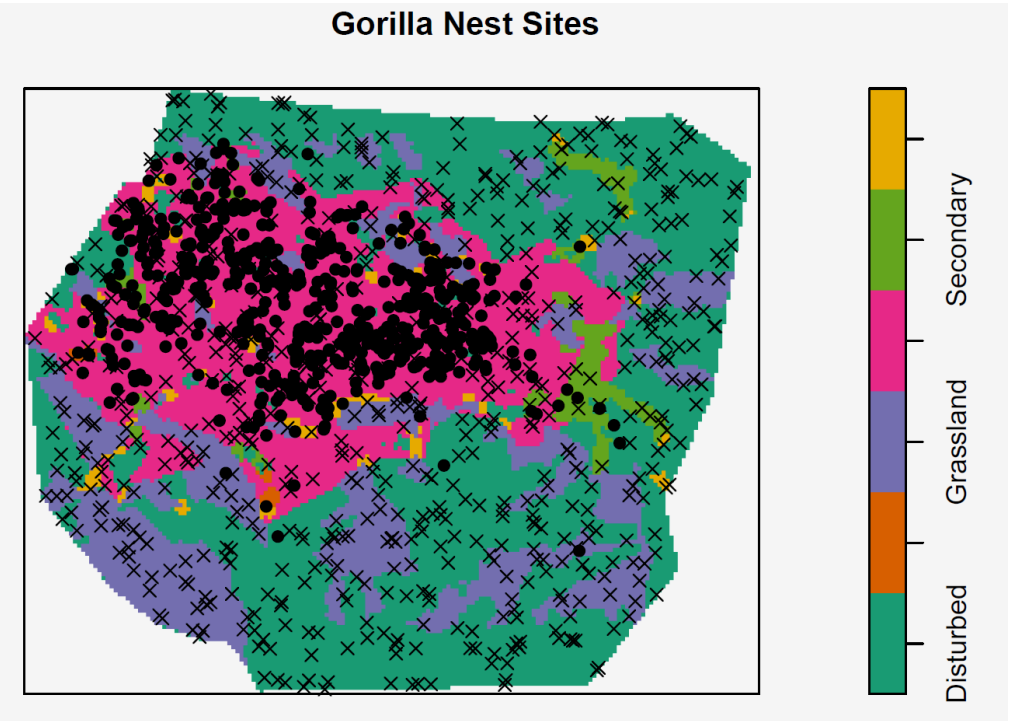
- **Binary Response Spatial Logistic Regression** - Module 5, similar to Gorillas Logistic Regression GLM example in *Non-Gaussian spatial data*.

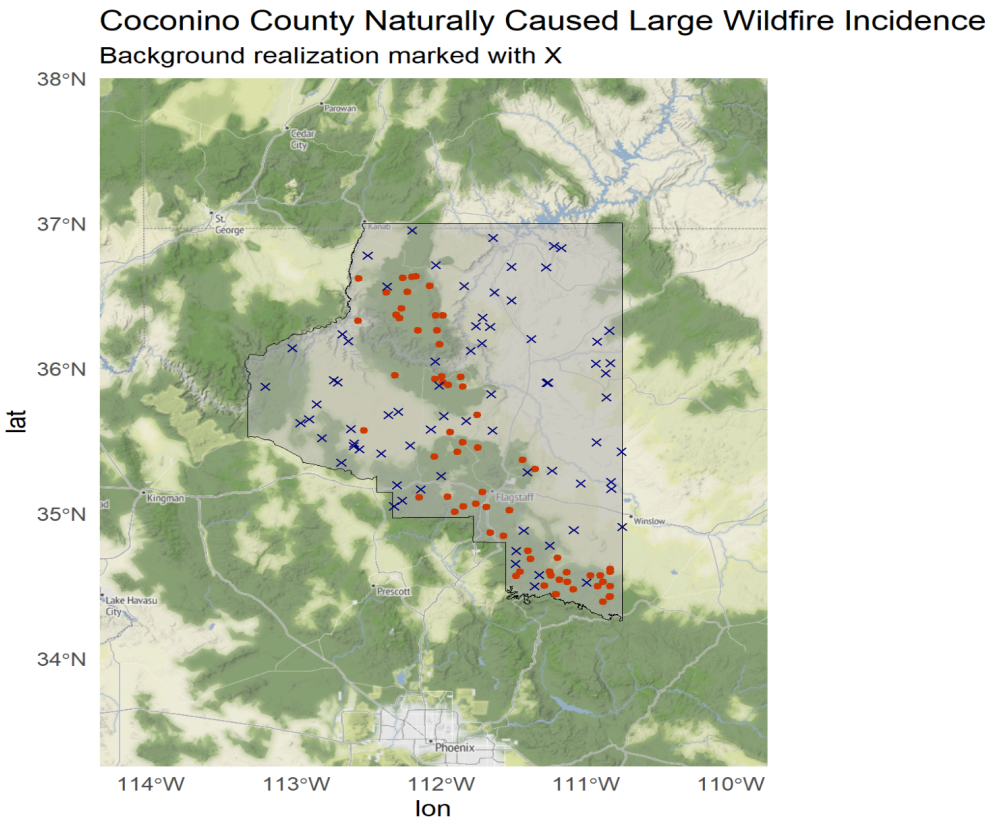
$$\text{logit}(\lambda_1(\boldsymbol{s})) = \mathbf{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \mathrm{e}(\boldsymbol{s}) + \log(\lambda_0),$$

$$\mathrm{Y}(\boldsymbol{s}) \sim \text{Bern}(\mathrm{p}(\boldsymbol{s})), \quad \mathrm{E}[\mathrm{Y}(\boldsymbol{s})] = \mathrm{p}(\boldsymbol{s}) = \frac{\lambda_1(\boldsymbol{s})}{\lambda_0(\boldsymbol{s}) + \lambda_1(\boldsymbol{s})}$$

Wildfire Incidence Data - Point Process 2

Binary GLM Spatial Logistic Regression Model Data



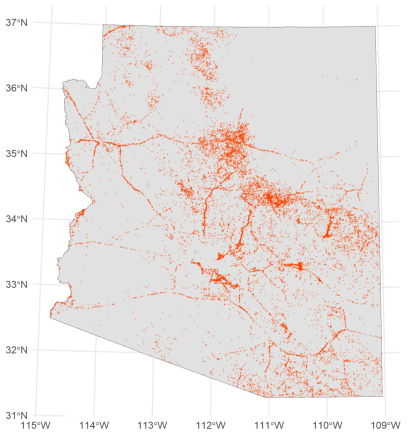


Research Questions (2)

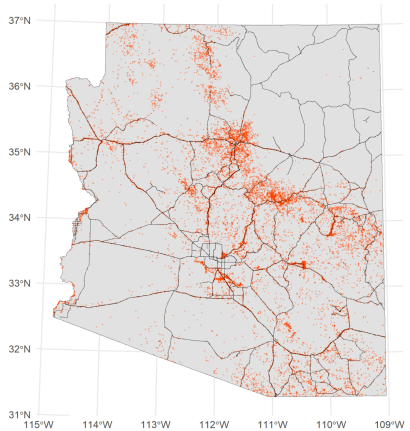
Can we find useful covariate data that can improve our models?

Research Questions (2)

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



(a) All wildfires



(b) Proximity to major roads

Figure 1: Roads and wildfires in AZ

Data explorations showed some clear wildfire patterns near roads (the outlines are visible). We wanted to include the distance in meters to the nearest major roads (and "remote" roads) as usable predictors. For each point, the `roads()` function in the `tigris` package 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.

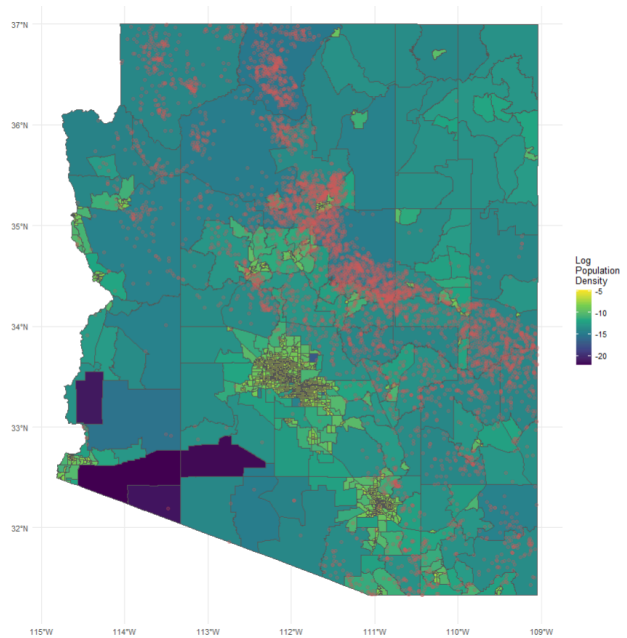
Research Questions (2)

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

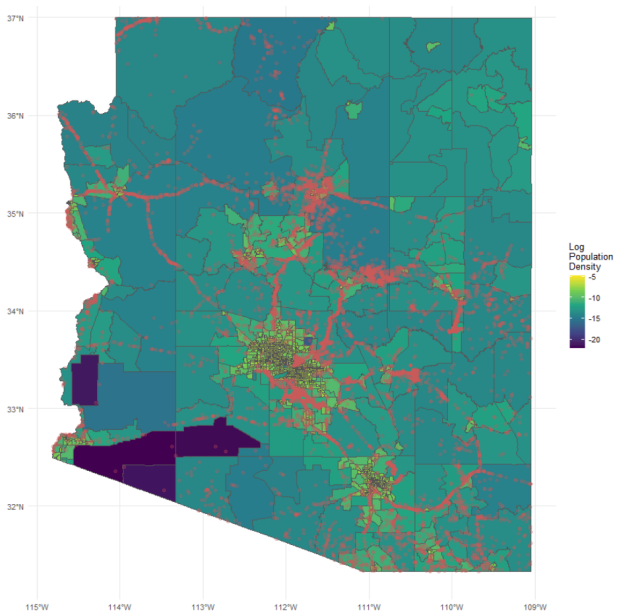
Raymond and Matt collected data for environmental and human factors to address research question 2 (RQ2) in our modeling efforts.

Research Questions (2)

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



Natural Wildfires



Human Caused Wildfires

Figure 2: Data retrieved from the `tidycensus` R package with `get_decennial` function. Plots show log population density for each tract in Arizona from 2010.

Research Questions (3)

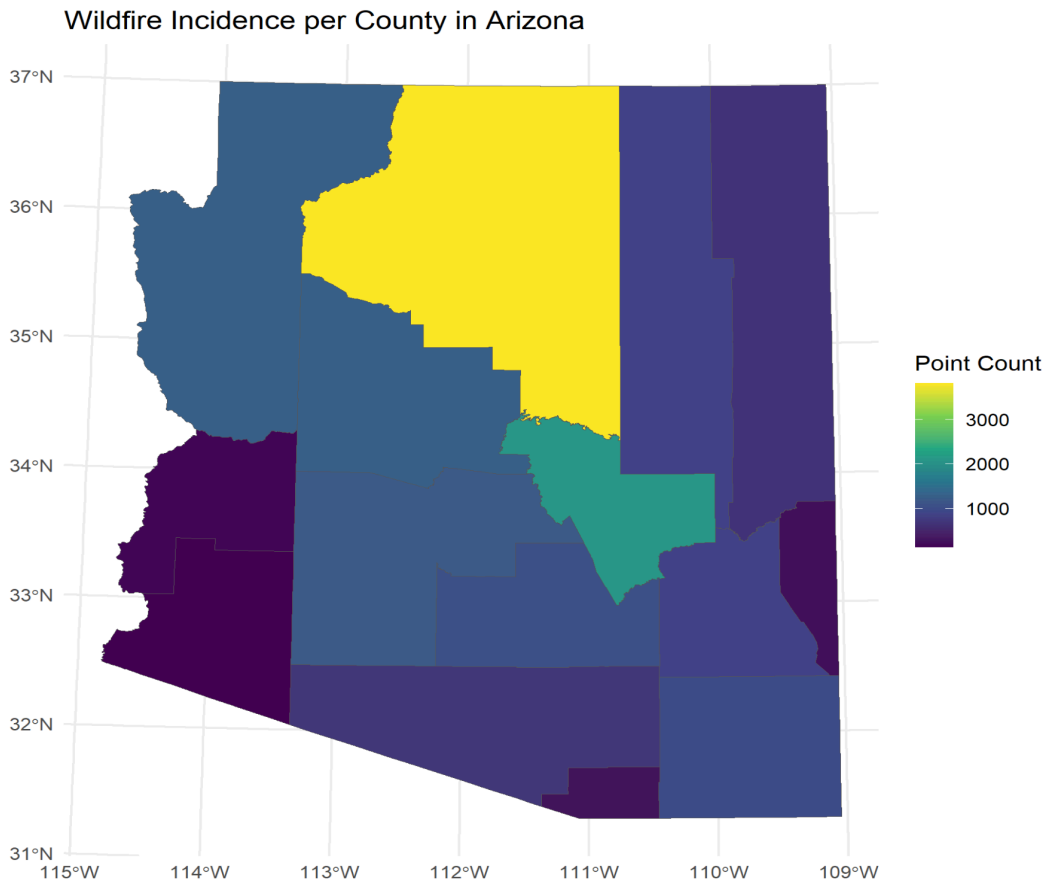
Are the patterns of human or non human caused fires spatially CSR, or do they exhibit an inhomogeneous spatial intensity?

Matt's LGCP models address research question 3.

Modeling - Coconino County

We opted to focus our modeling efforts on Coconino County based on some inital model fits, and overall size of the dataset.

Coconino County is home to national forests like Coconino, Kaibab, and Apache-Sitgreaves, and has the most wildfire incidence in the state by a good margin.



Spatial Linear Model approach

- **Continuous, fixed** - Treating `x` and `y` as coordinates as fixed and using `IncidentSize` as a continuous response, our data is geostatistical.

Modeling approach

- **Spatial Linear Model** - Module 2, similar to Holland example in *Continuously indexed spatial data (geostatistics)*, or the `dioxane` analysis in HW2.

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

Wrangling Natural Factors data

Daymetr package and Raster images

Variables of Interest

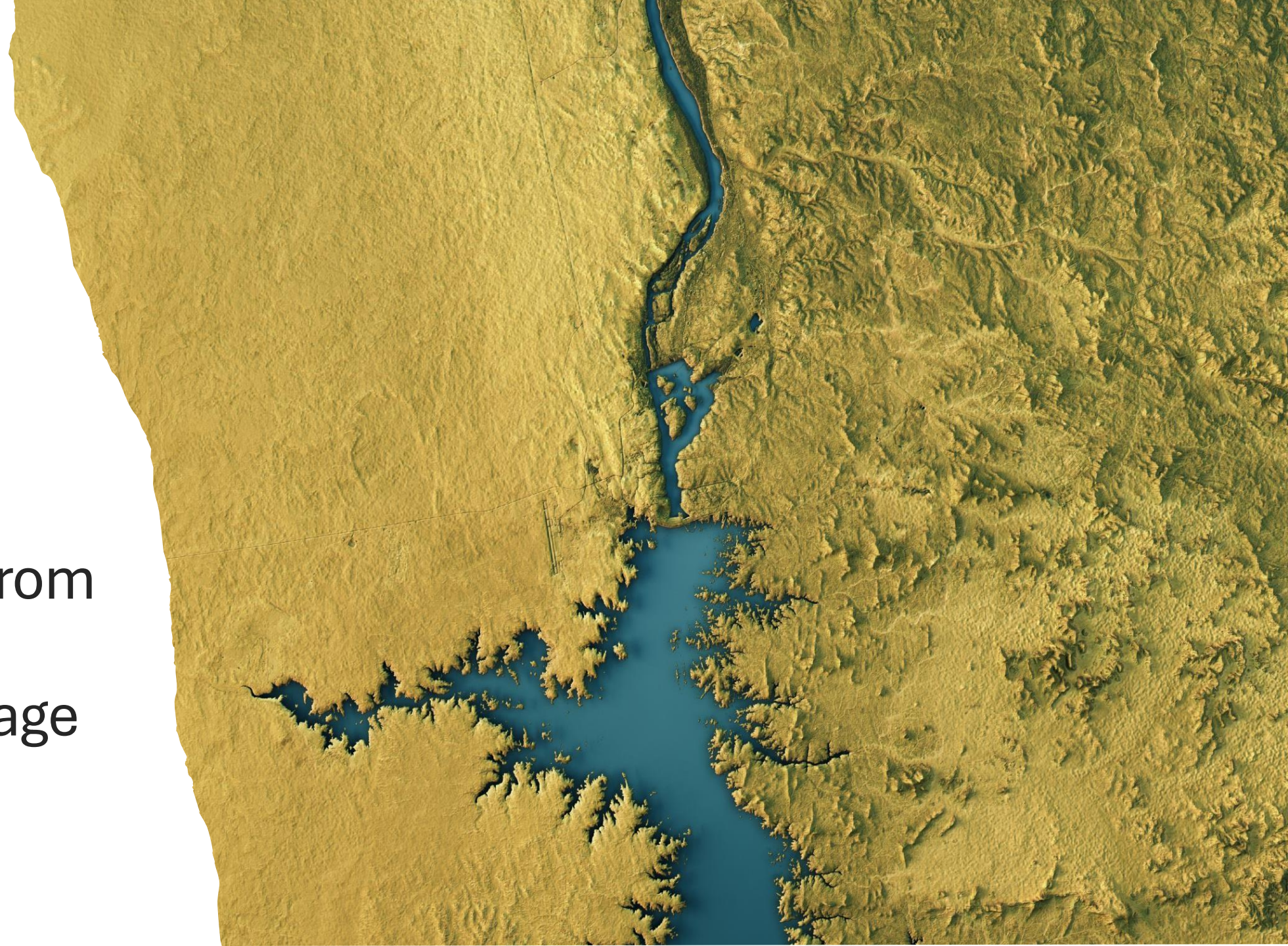
PRECIPITATION

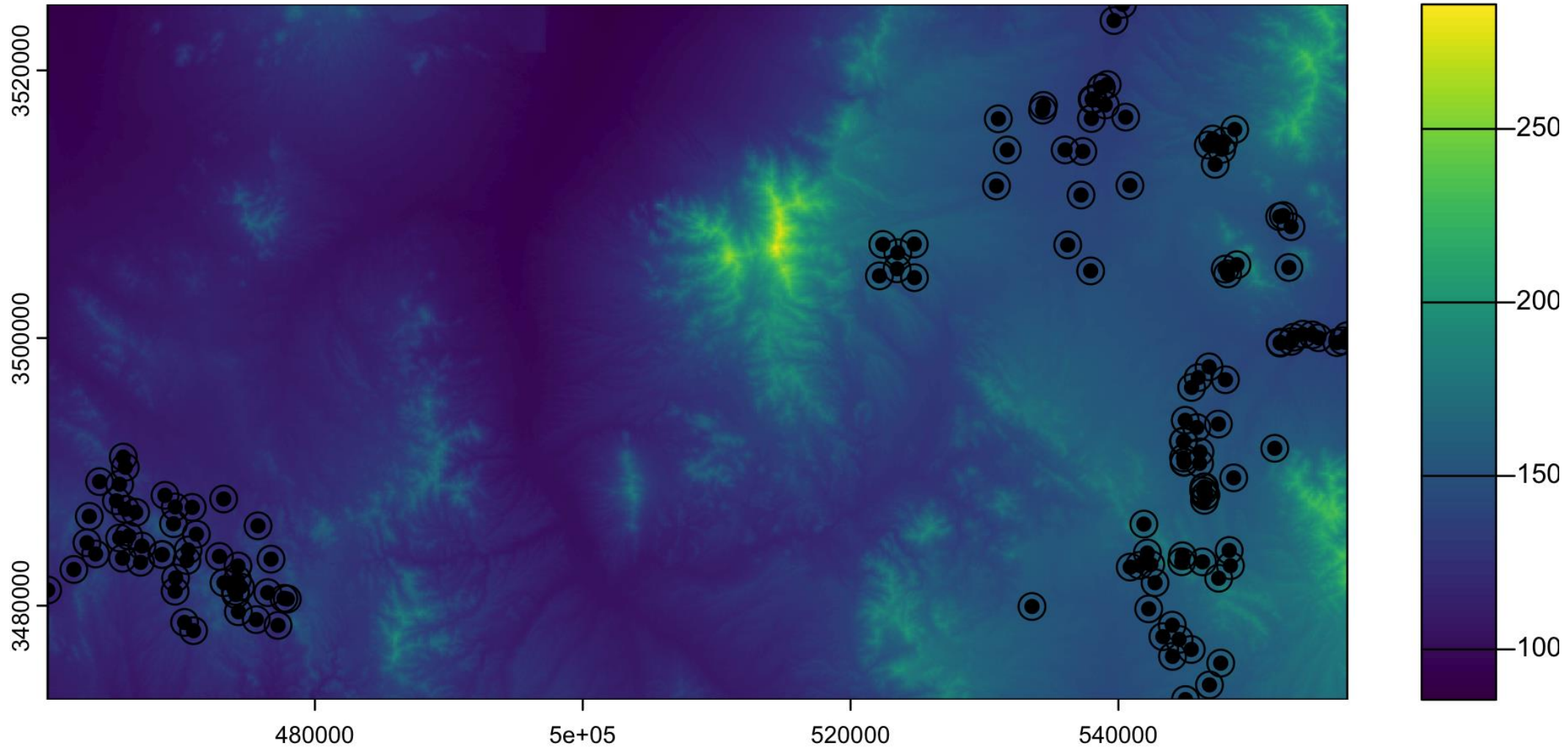
TEMPERATURE

TOPOGRAPHY

VEGETATION

Extracting from local raster satellite image

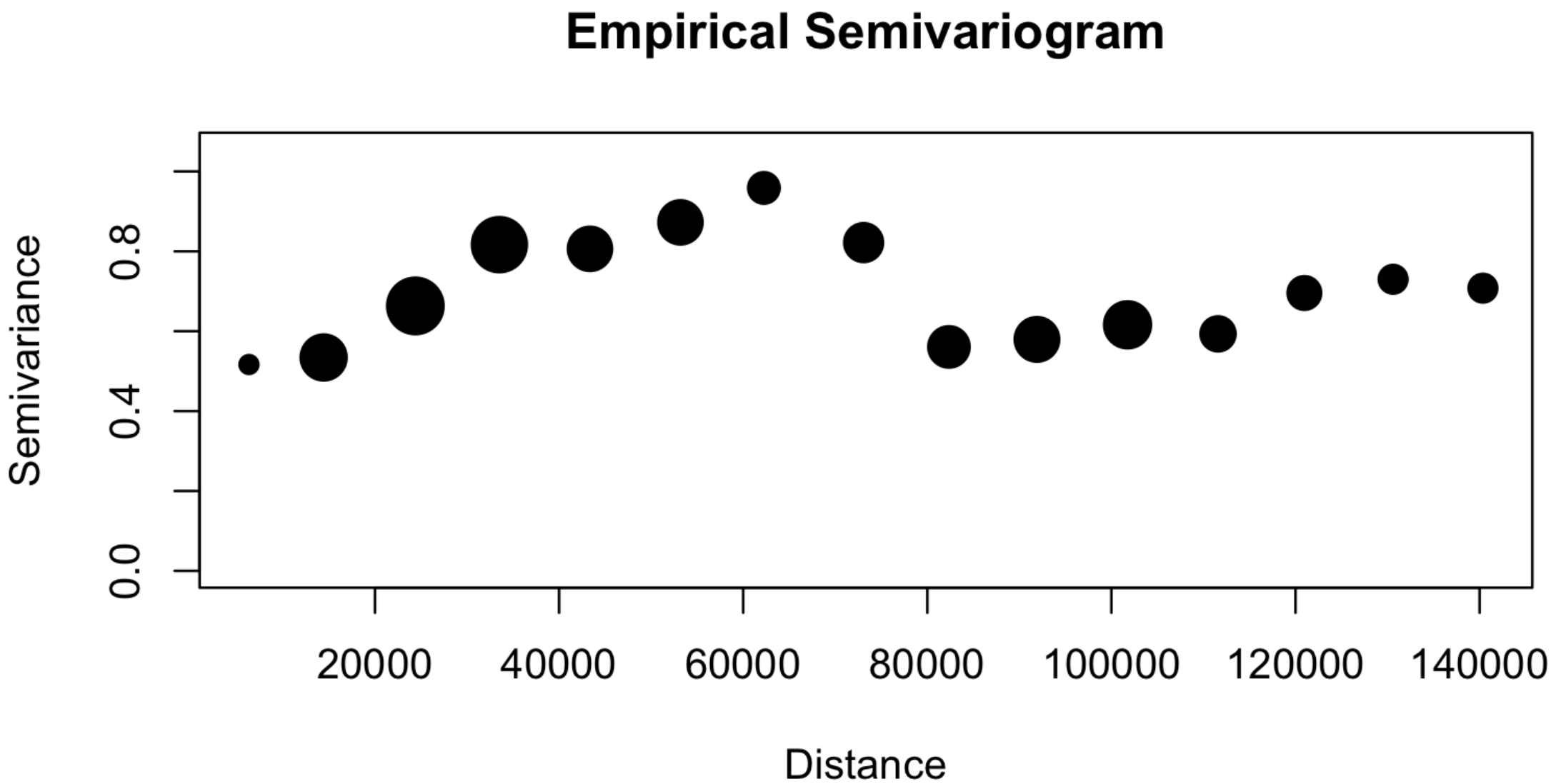



3520000
3500000
3480000
480000
5e+05
520000
540000
250
200
150
100

Daymetr package for precipitation and temperature

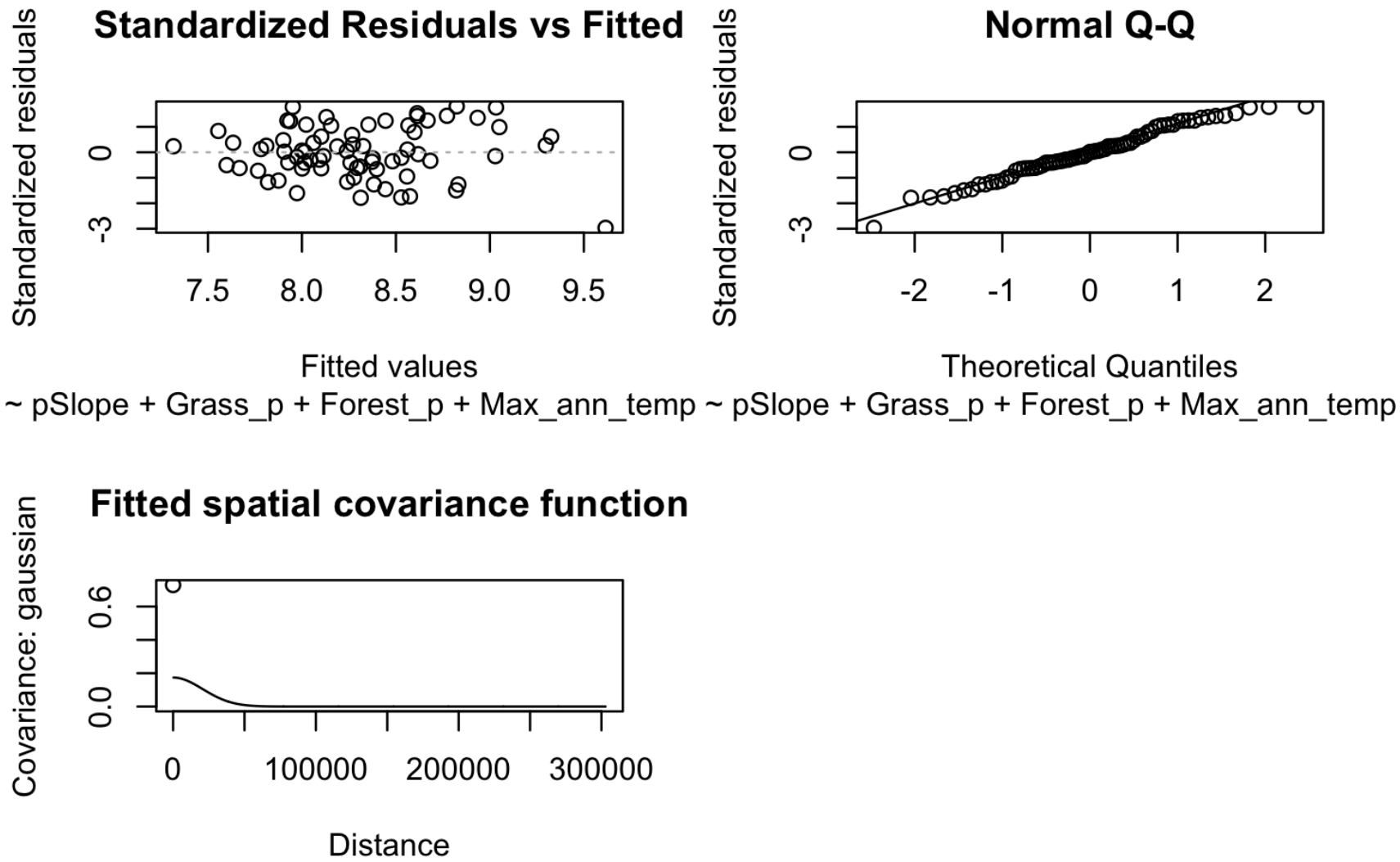
Choosing best model AIC

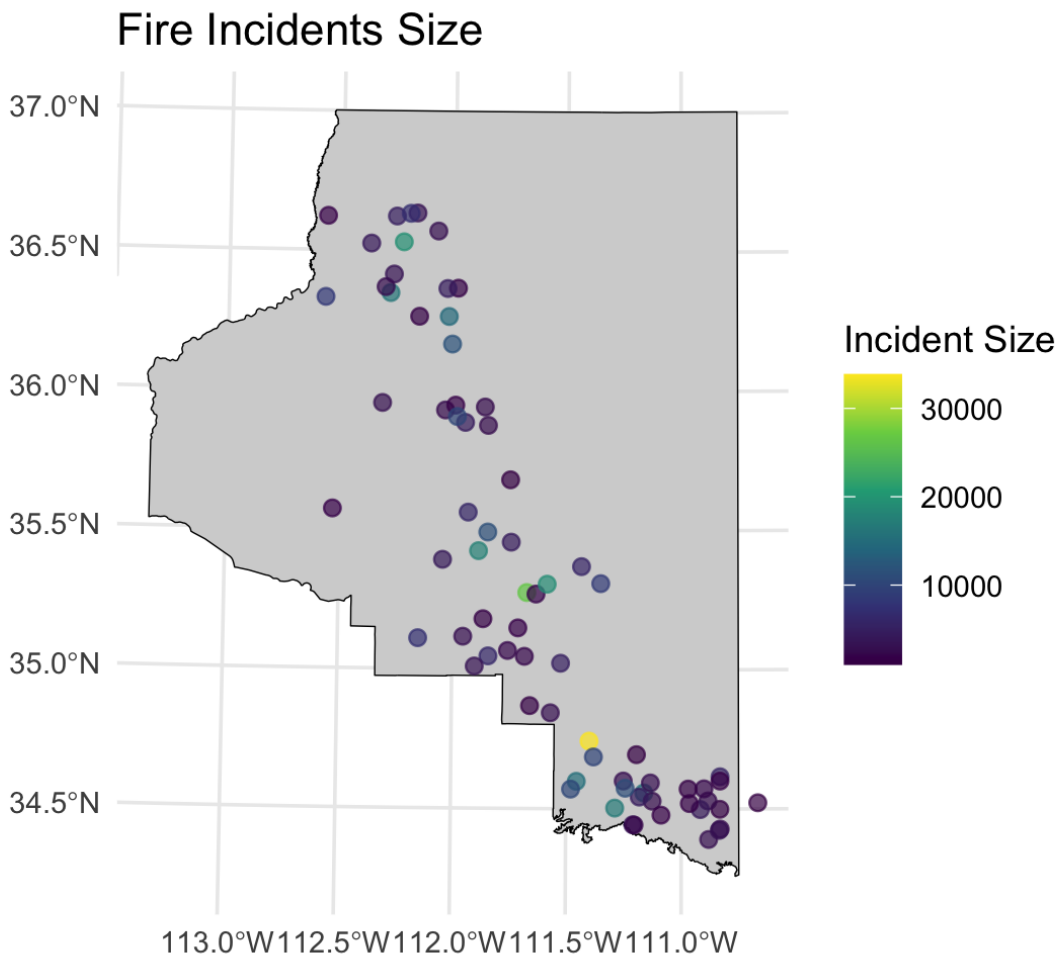
- spmod <- `splm(log(Size) ~ pSlope + Grass_p + Forest_p + Max_ann_temp + Min_ann_temp + Prcp_ann + pop_density + Population + Pri_rd, data = coconino_sf, spcov_type = "gaussian")`

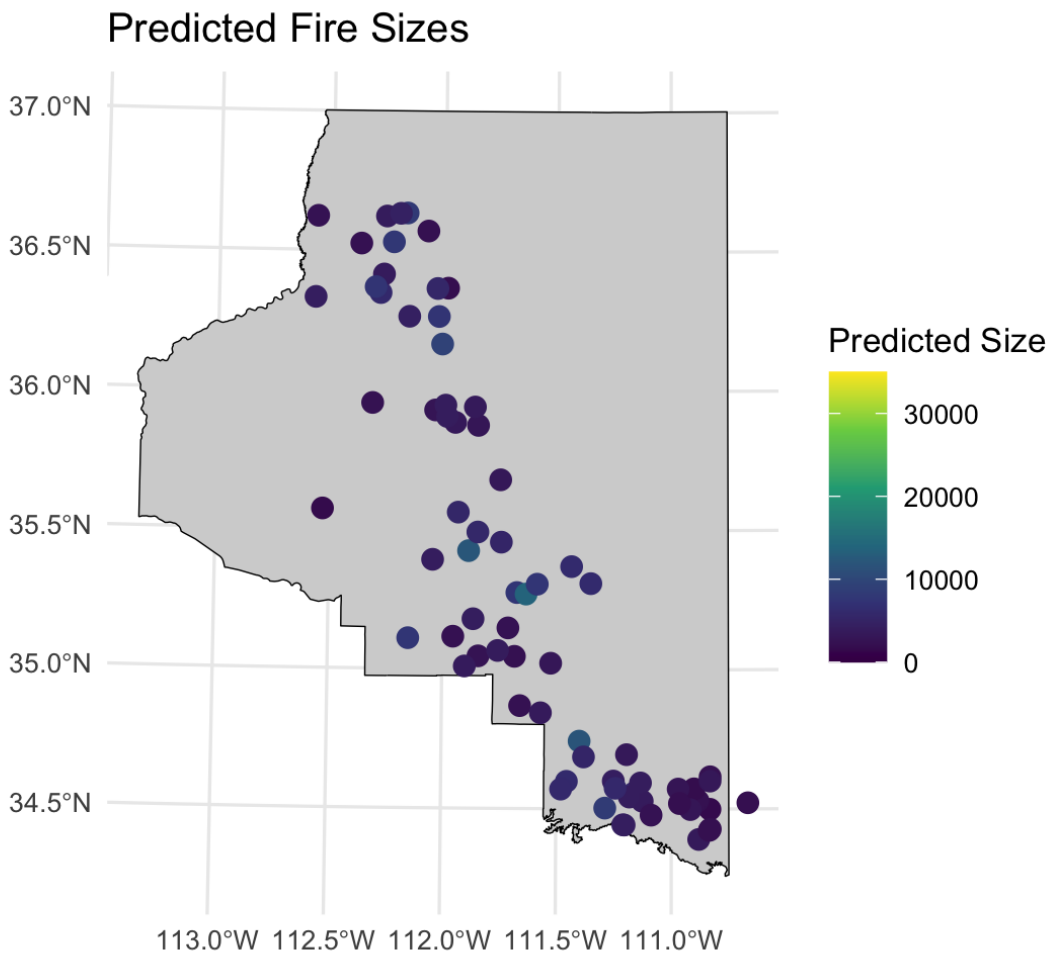

Empirical Semivariogram
Semivariance
0.0
0.4
0.8
20000
40000
60000
80000
100000
120000
140000
Distance

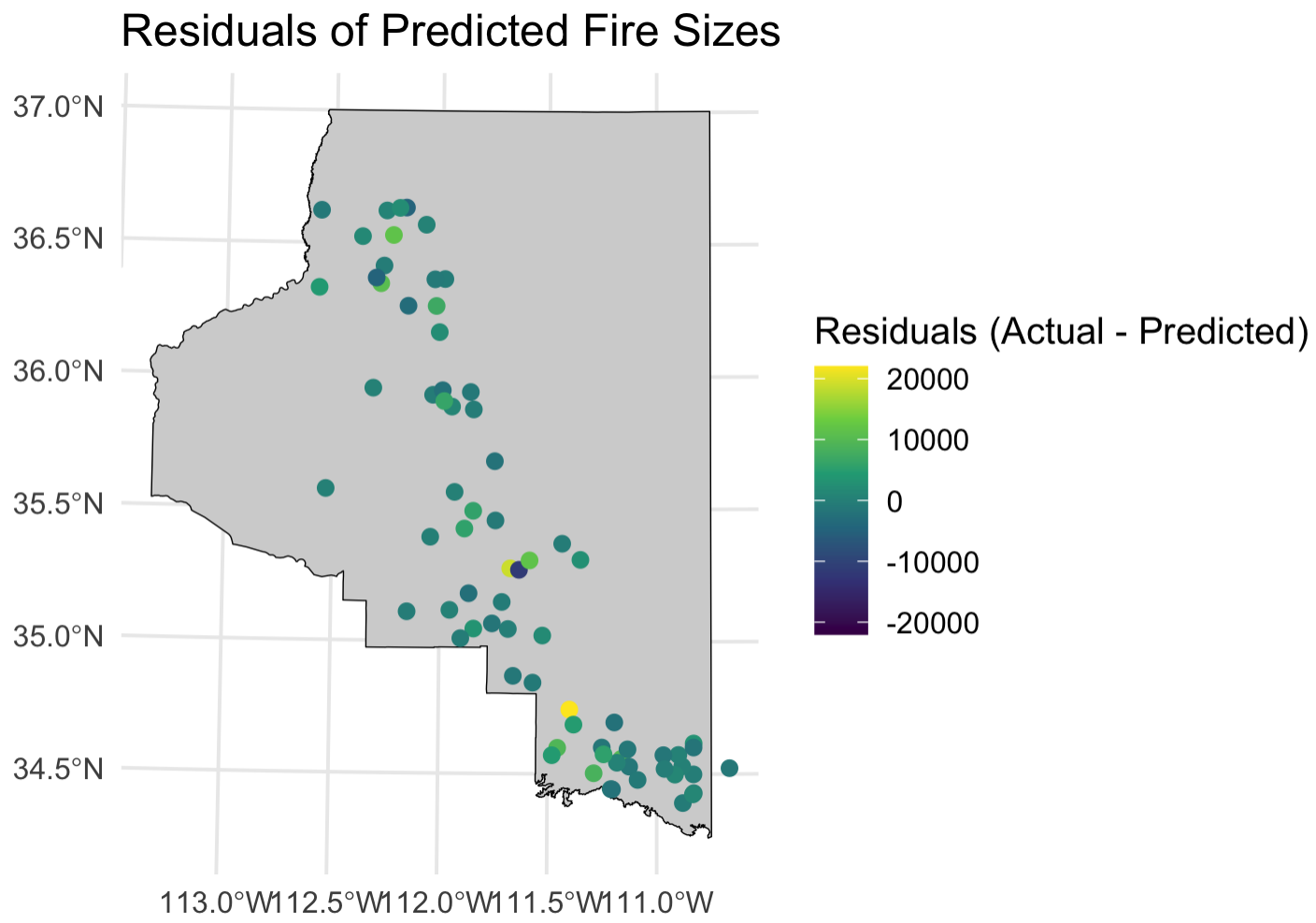
Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	18.26	3.39	5.39	0.0000001
pSlope	0.009	0.012	0.76	0.4500
Grass_p	0.277	0.582	0.48	0.6338
Forest_p	-0.534	0.448	-1.19	0.2328
Max_ann_temp	**-0.516**	**0.177**	**-2.91**	**0.0036**
Min_ann_temp	0.226	0.165	1.37	0.1695
Prcp_ann	**-0.810**	**0.388**	**-2.09**	**0.0370**
pop_density	16400	108300	0.15	0.8796
Population	0.00007	0.0001	0.75	0.4542
Pri_rd	**-0.00001**	**0.000005**	**-1.91**	**0.0561**

de =0.174, ie = 0.555, range = 29119.613


Standardized Residuals vs Fitted
Standardized residuals
-3
0
7.5
8.0
8.5
9.0
9.5
Fitted values
~ pSlope + Grass_p + Forest_p + Max_ann_temp
Normal Q-Q
Standardized residuals
-3
0
-2
-1
0
1
2
Theoretical Quantiles
~ pSlope + Grass_p + Forest_p + Max_ann_temp
Fitted spatial covariance function
Covariance: gaussian
0.0
0.6
0
100000
200000
300000
Distance


Fire Incidents Size
37.0°N
36.5°N
36.0°N
35.5°N
35.0°N
34.5°N
113.0°W
112.5°W
112.0°W
111.5°W
111.0°W
Incident Size
30000
20000
10000


Predicted Fire Sizes
37.0°N
36.5°N
36.0°N
35.5°N
35.0°N
34.5°N
113.0°W
112.5°W
112.0°W
111.5°W
111.0°W
Predicted Size
30000
20000
10000
0


Residuals of Predicted Fire Sizes
37.0°N
36.5°N
36.0°N
35.5°N
35.0°N
34.5°N
113.0°W
112.5°W
112.0°W
111.5°W
111.0°W
Residuals (Actual - Predicted)
20000
10000
0
-10000
-20000

Log-Gaussian Cox Process Approach

- **Point process** - Using `x` and `y` as coordinates and `IncidentSize` as a *threshold*, we can study "large wildfires" (`IncidentSize` ≥ 1000 acres) as point process data.

Modeling approach

- **Log-Gaussian Cox Process** Module 4, similar to Gorillas LGCP example in *Random spatial index (point pattern)*, or the `earthquakes_um` analysis of Kansas in HW4.

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

Log-Gaussian Cox Process Approach

Prediction Surfaces

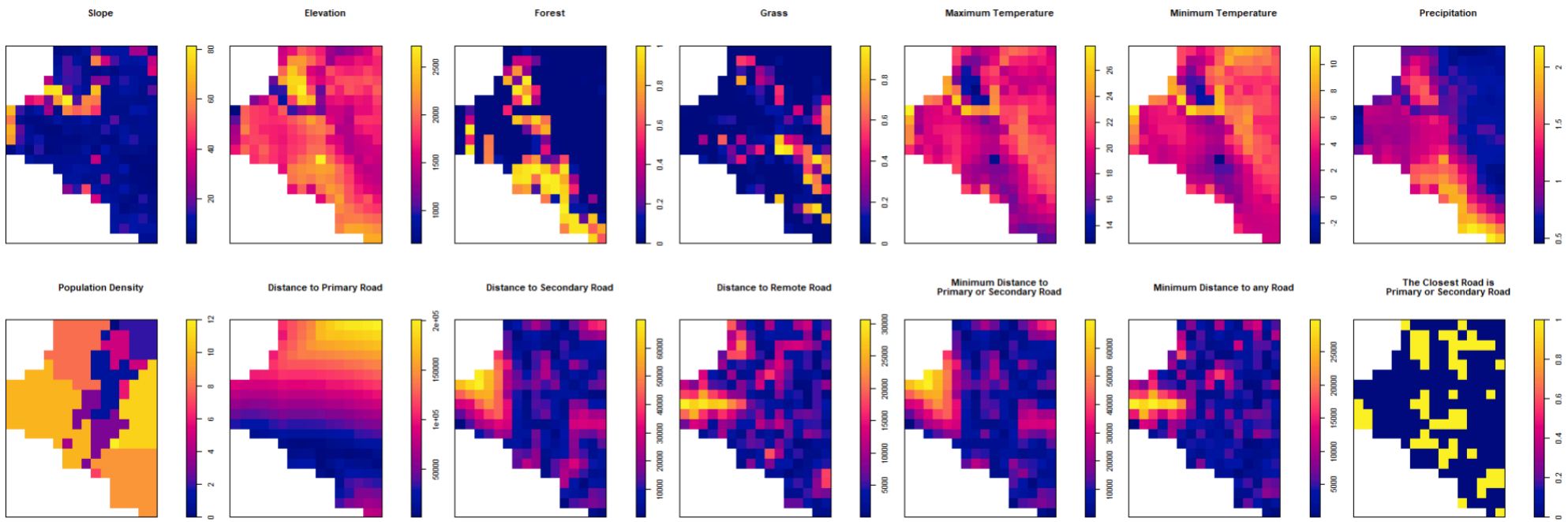


Figure 3: Heat map surfaces of each predictor used in the LGCP model.

Log-Gaussian Cox Process Approach

Response Surfaces

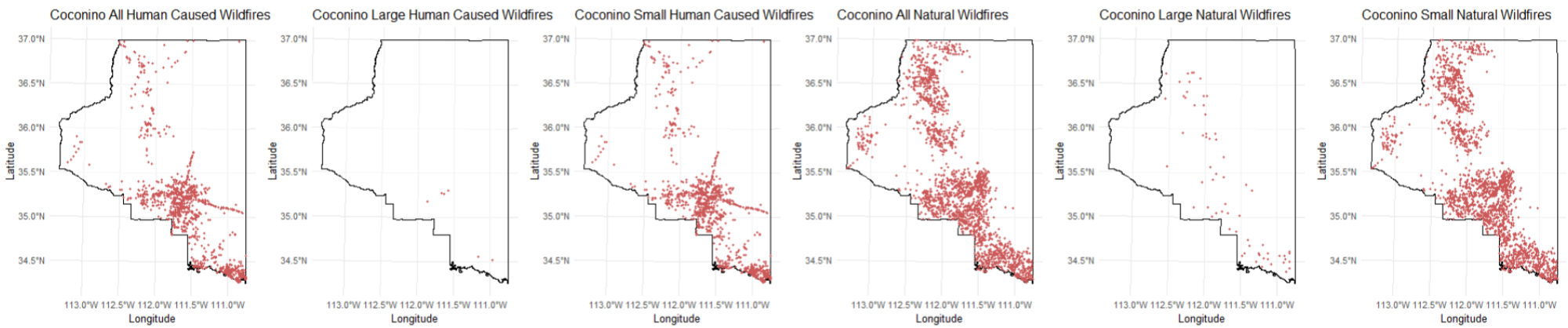


Figure 4: Locations of wildfires in Coconino. Separated by cause and size.

Goal: Fit an intensity surface for each of these plots using our predictors in a LGCP model.

Log-Gaussian Cox Process Approach

All Human Caused Wildfires

```
coco_wf_hum.kppm <- kppm(unmark(coco_wf_hum.ppp)~.,
                         data=predictors.im,
                         clusters = "LGCP",
                         model = "exponential")
plot(predict(coco_wf_hum.kppm, eps = 15000))
```

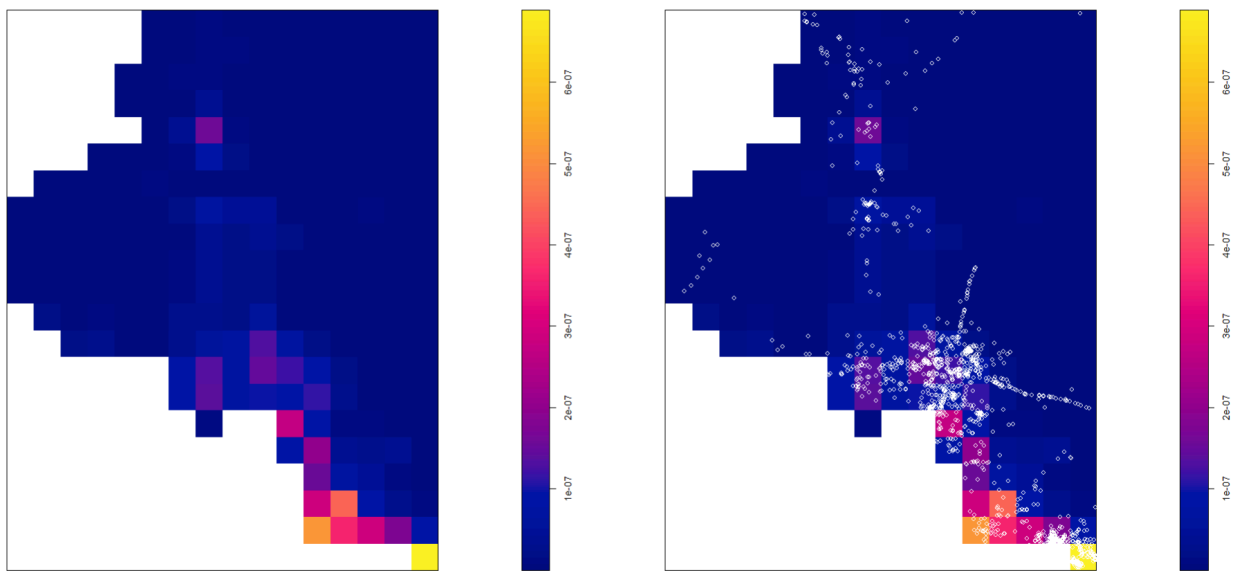


Figure 5: All human caused wildfire intensity surface.

Log-Gaussian Cox Process Approach

Large Human Caused Wildfires

```
coco_wf_hum_lg.kppm <- kppm(unmark(coco_wf_hum_lg.ppp)~.,
                            data=predictors.im,
                            clusters = "LGCP",
                            model = "exponential")
plot(predict(coco_wf_hum_lg.kppm, eps = 15000))
```

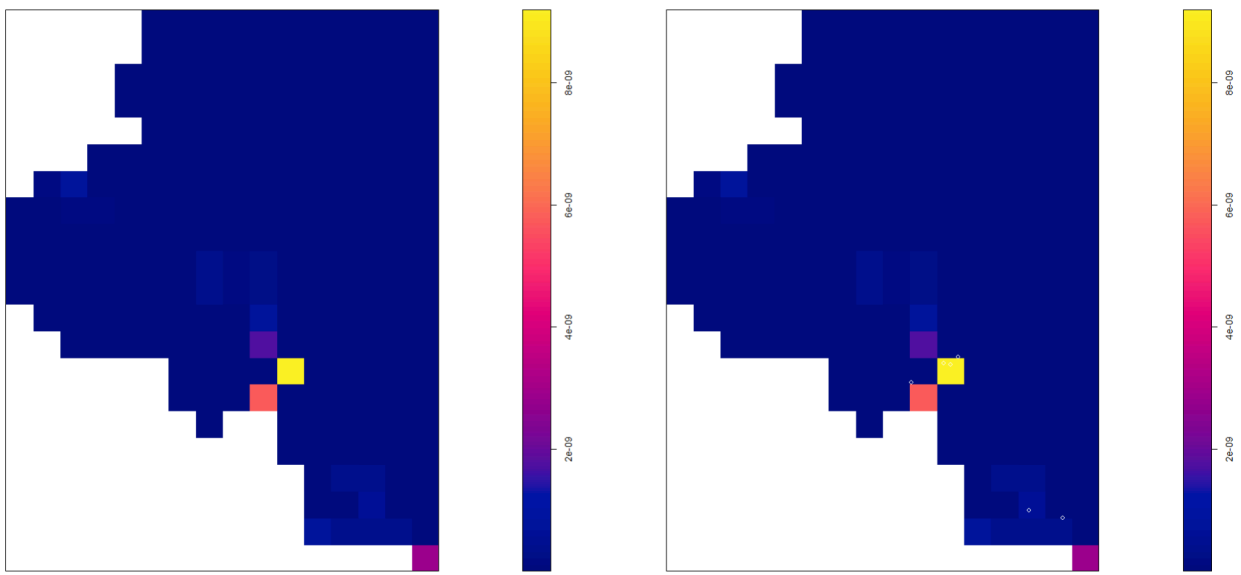


Figure 6: Large human caused wildfire intensity surface.

Log-Gaussian Cox Process Approach

Small Human Caused Wildfires

```
coco_wf_hum_sm.kppm <- kppm(unmark(coco_wf_hum_sm.ppp)~.,
                            data=predictors.im,
                            clusters = "LGCP",
                            model = "exponential")
plot(predict(coco_wf_hum_sm.kppm, eps = 15000))
```

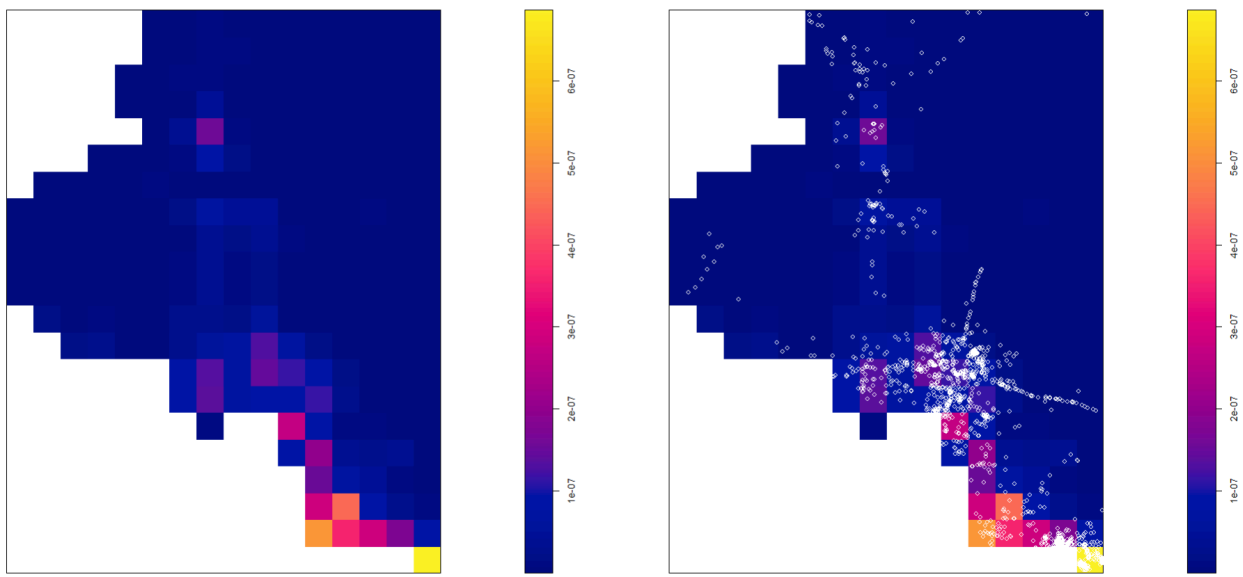


Figure 7: Small human caused wildfire intensity surface.

Log-Gaussian Cox Process Approach

All Natural Wildfires

```
coco_wf_nat.kppm <- kppm(unmark(coco_wf_nat.ppp)~.,
                         data=predictors.im,
                         clusters = "LGCP",
                         model = "exponential")
plot(predict(coco_wf_nat.kppm, eps = 15000))
```

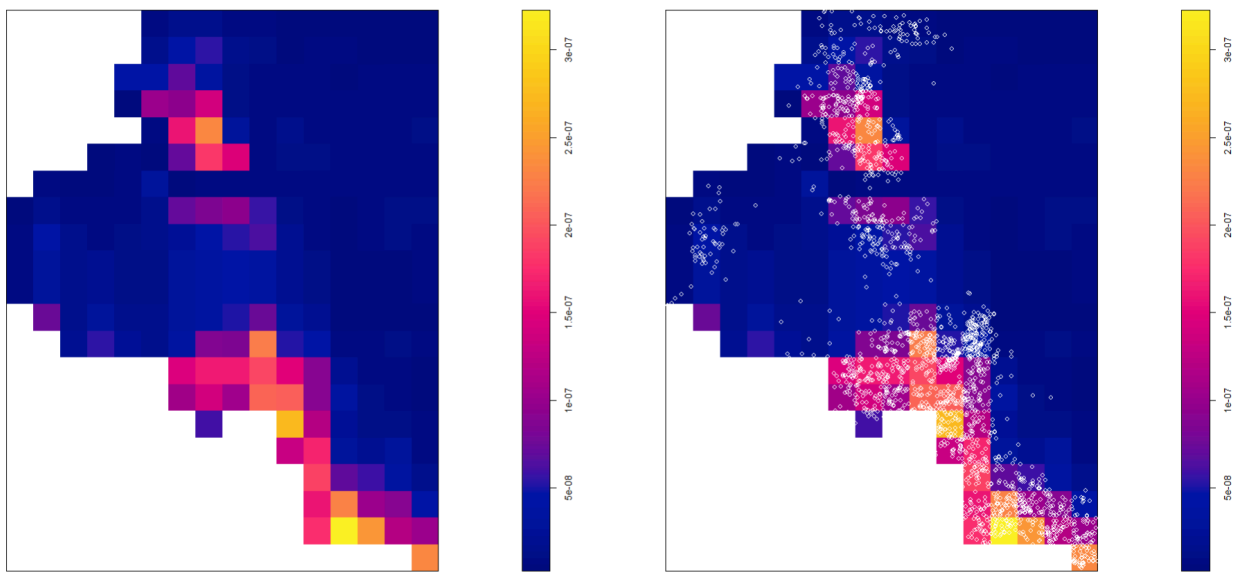


Figure 8: All natural wildfire intensity surface.

Log-Gaussian Cox Process Approach

Large Natural Wildfires

```
coco_wf_nat_lg.kppm <- kppm(unmark(coco_wf_nat_lg.ppp)~.,
                            data=predictors.im,
                            clusters = "LGCP",
                            model = "exponential")
plot(predict(coco_wf_nat_lg.kppm, eps = 15000))
```

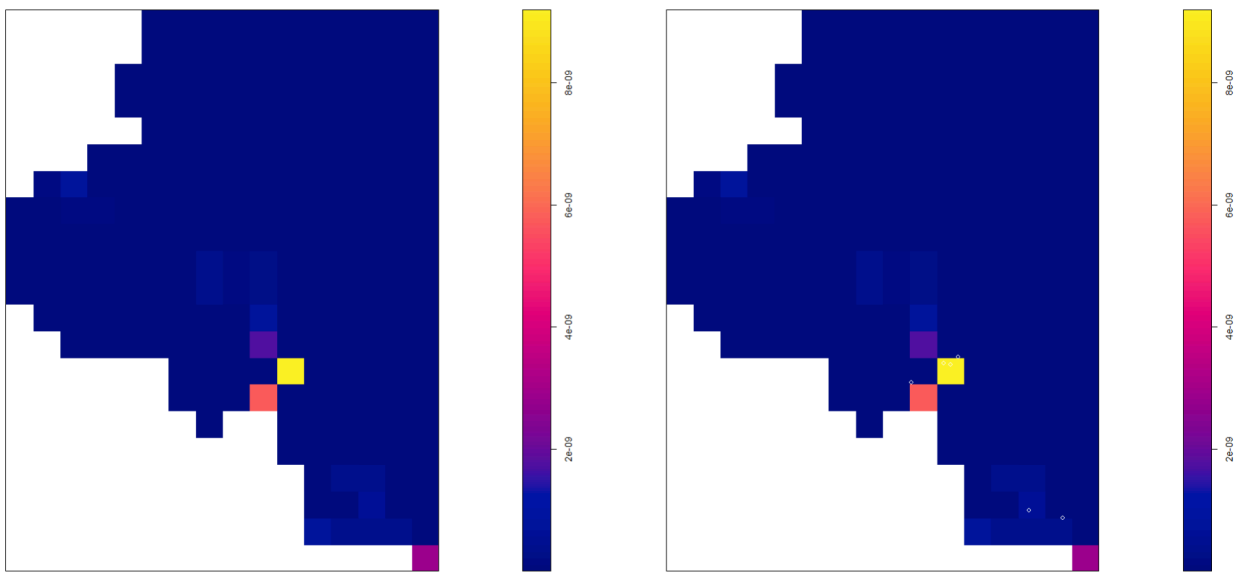


Figure 9: Large natural wildfire intensity surface.

Log-Gaussian Cox Process Approach

Small Human Caused Wildfires

```
coco_wf_nat_sm.kppm <- kppm(unmark(coco_wf_nat_sm.ppp)~.,
                            data=predictors.im,
                            clusters = "LGCP",
                            model = "exponential")
plot(predict(coco_wf_nat_sm.kppm, eps = 15000))
```

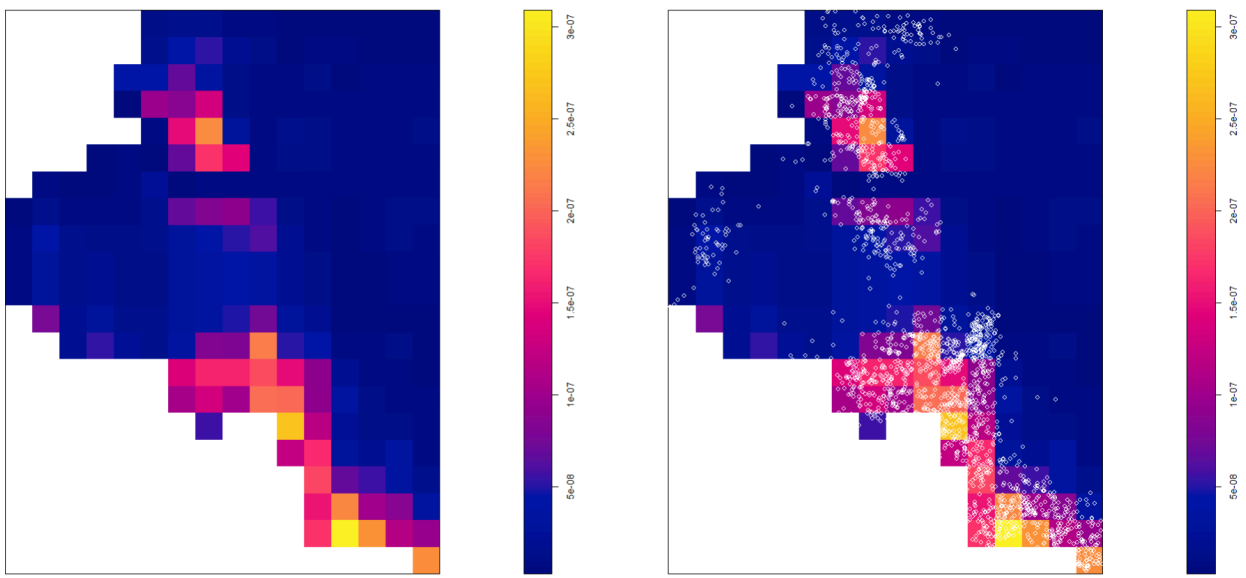


Figure 10: Small natural wildfire intensity surface.

Log-Gaussian Cox Process Approach

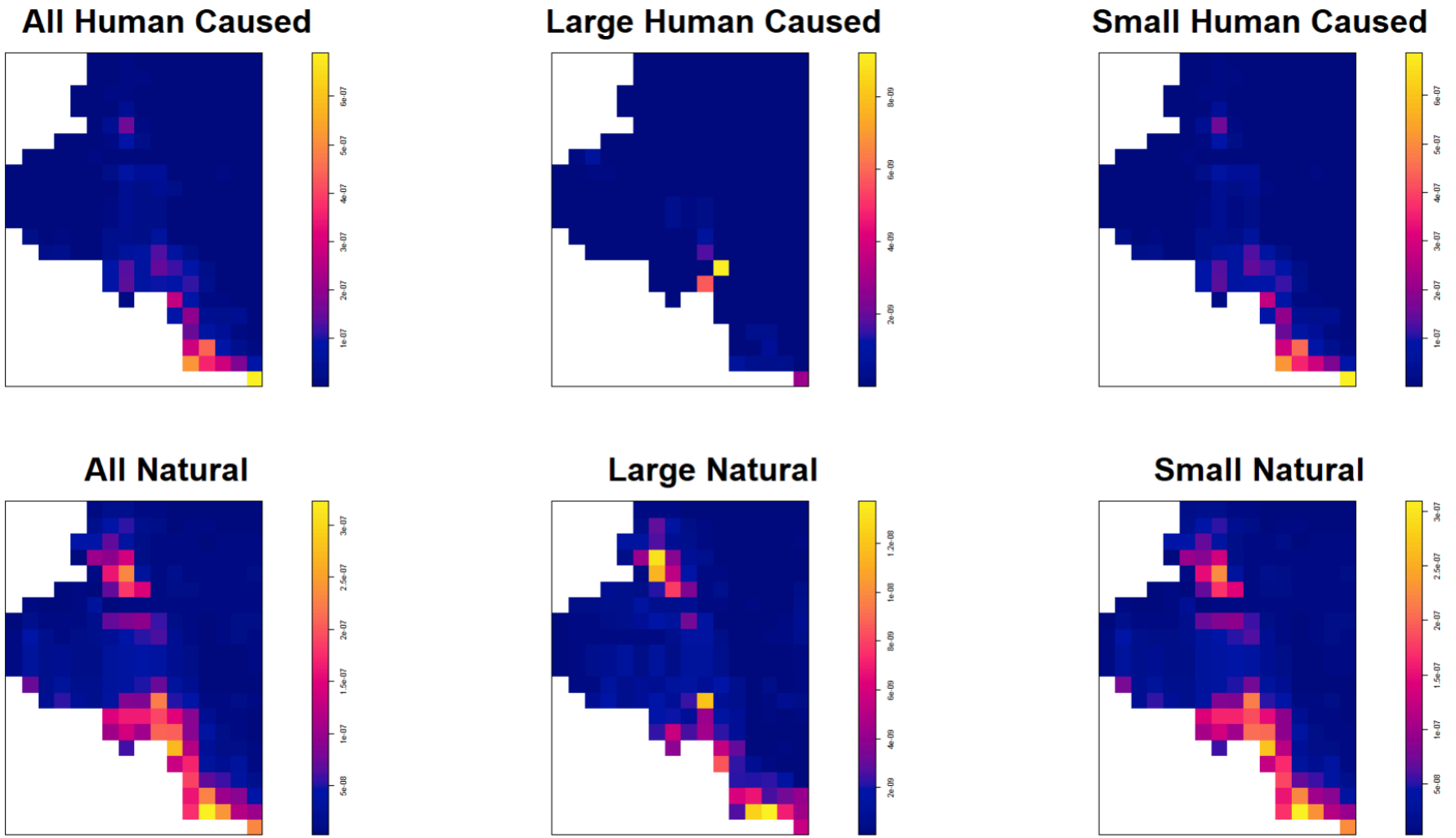


Figure 11

Log-Gaussian Cox Process Approach

$$C(u, u') = \sigma^2 e^{-||u-u'||/h}$$

	σ^2 **(partial sill)**	**h(range)**
All Human Caused	4.728370e+00	7791.704
Large Human Caused	3.220770e-09	29691.296
Small Human Caused	4.688047e+00	7806.519
All Natural	5.472859e-01	9910.462
Large Natural	1.400144e+00	5681.422
Small Natural	2.056924e+00	12054.552

Research Questions (3)

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

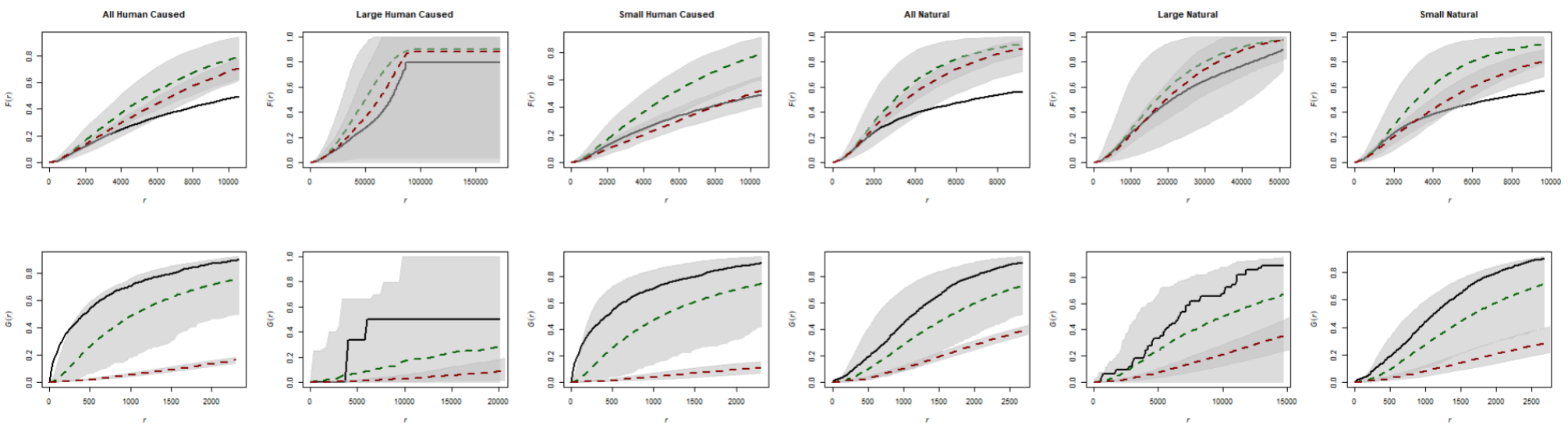


Figure 12: Green represents the intercept only model; red represents our model with all predictors.

Research Questions (3)

Arizona CSR Analysis

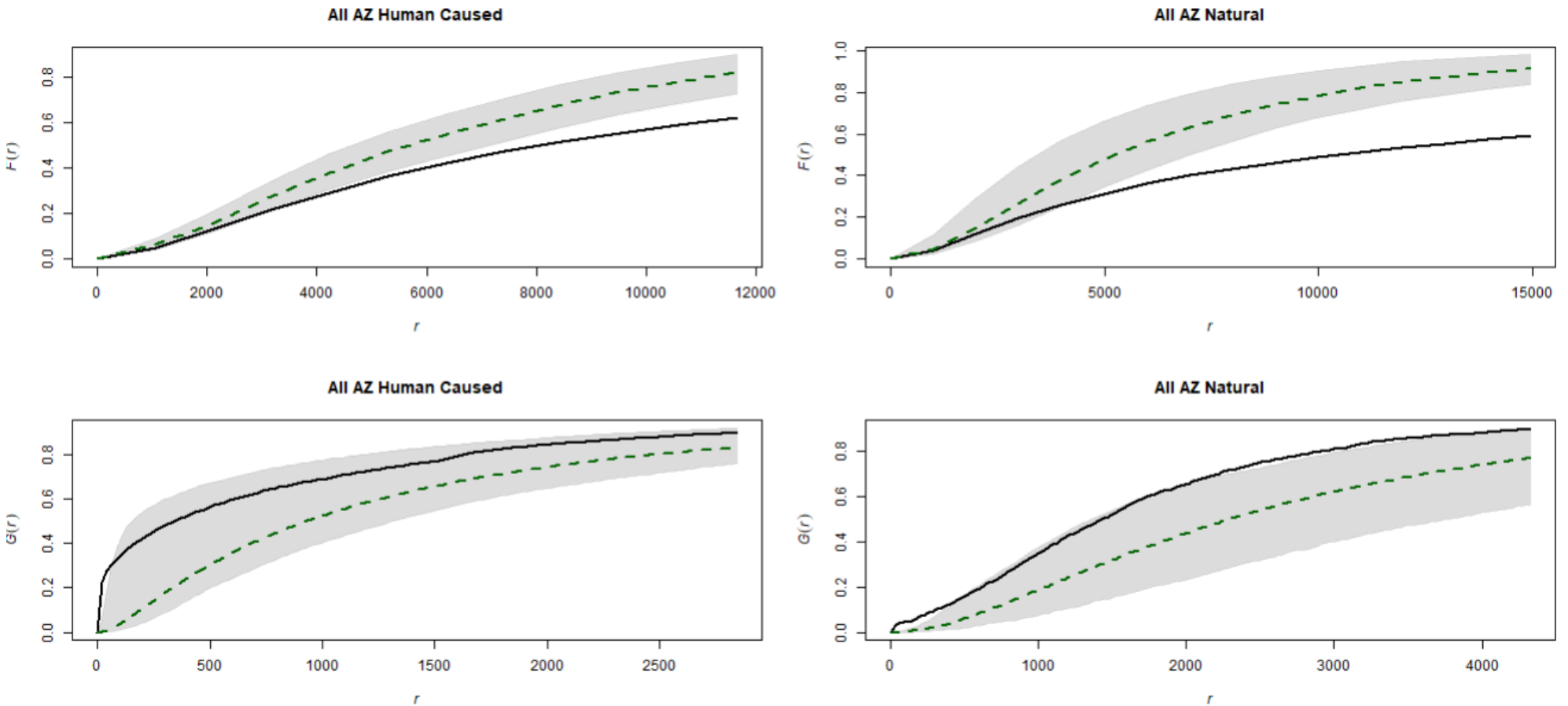


Figure 13

Binary GLM Spatial approach

- **Point process** - Using `x` and `y` as coordinates and `IncidentSize` as a *threshold*, we can study "large wildfires" (`IncidentSize` ≥ 1000 acres) as point process data.

Modeling approach

- **Binary Response Spatial Logistic Regression** - Module 5, similar to Gorillas Logistic Regression GLM example in *Non-Gaussian spatial data*.

$$\text{logit}(\lambda_1(\boldsymbol{s})) = \mathbf{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \mathrm{e}(\boldsymbol{s}) + \log(\lambda_0),$$

$$\mathrm{Y}(\boldsymbol{s}) \sim \text{Bern}(\mathrm{p}(\boldsymbol{s})), \quad \mathrm{E}[\mathrm{Y}(\boldsymbol{s})] = \mathrm{p}(\boldsymbol{s}) = \frac{\lambda_1(\boldsymbol{s})}{\lambda_0(\boldsymbol{s}) + \lambda_1(\boldsymbol{s})}$$

Binary GLM Spatial approach

We now want to study risk of occurence of large wildfire (`Indidentsize`≥ 1000 acres) spatially in terms of probabilities. We will also restrict occurrences to `FireCause=="Natural"`.

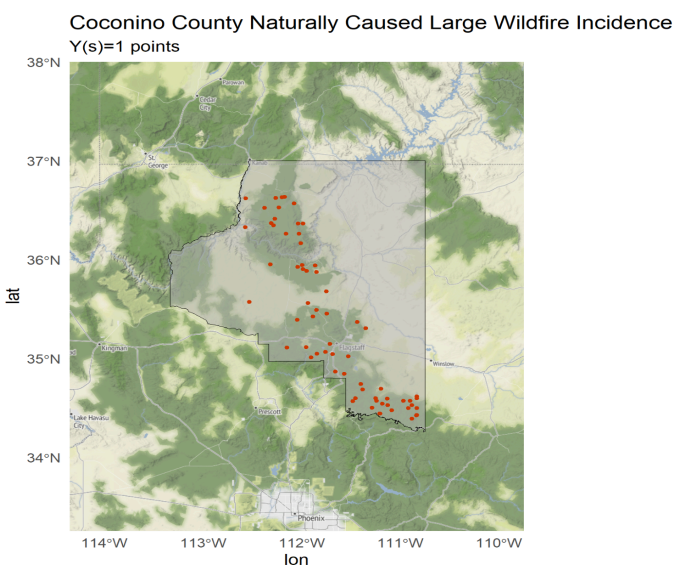
To build the model, we filter the data and treat large wildfire incident locations as a realization of a point process with intensity surface $\lambda_1(\boldsymbol{s})$.

For these points, we assign a response value of $\mathrm{Y}(\boldsymbol{s}) = 1$ (`Wildfires`).

$$\text{logit}(\lambda_1(\boldsymbol{s})) = \mathbf{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \underbrace{\mathrm{e}(\boldsymbol{s})}_{\text{spatial effect}} + \log(\lambda_0),$$

$$\mathrm{Y}(\boldsymbol{s}) \sim \text{Bern}(\mathrm{p}(\boldsymbol{s})),$$

$$\mathrm{E}[\mathrm{Y}(\boldsymbol{s})] = \mathrm{p}(\boldsymbol{s}) = \frac{\lambda_1(\boldsymbol{s})}{\lambda_0(\boldsymbol{s}) + \lambda_1(\boldsymbol{s})}$$



$\mathrm{n} = 66$ incidence points

Binary GLM Spatial approach

Next we generate a background realization of a known constant Poisson process λ_0, which represent locations that a large wildfire could have happened but did not (note incident date/time is randomly assigned 2014 to present).

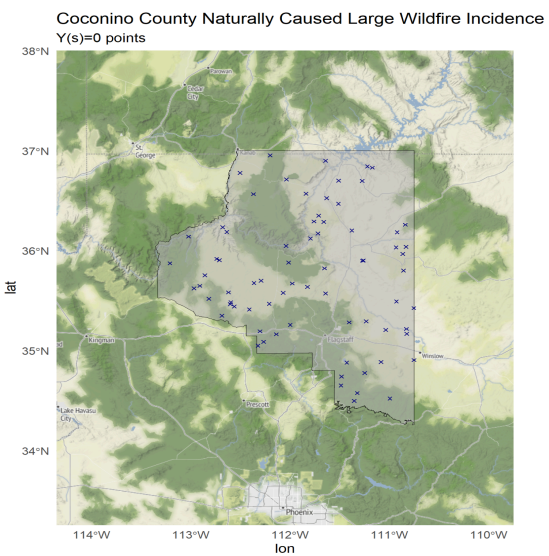
We also collect all the covariate data at these points (time consuming).

For these points, we assign a response value of $\mathrm{Y}(\boldsymbol{s}) = 0$ (`Wildfires`).

$$\text{logit}(\lambda_1(\boldsymbol{s})) = \mathbf{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \underbrace{\mathrm{e}(\boldsymbol{s})}_{\text{spatial effect}} + \log(\lambda_0),$$

$$\mathrm{Y}(\boldsymbol{s}) \sim \text{Bern}(\mathrm{p}(\boldsymbol{s})),$$

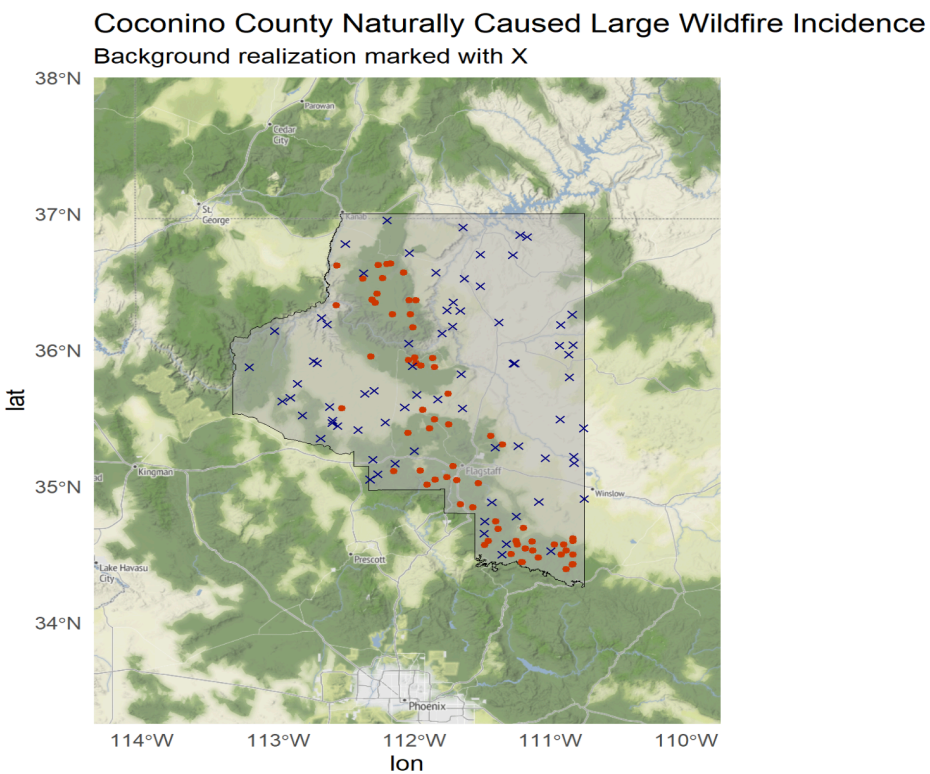
$$\mathrm{E}[\mathrm{Y}(\boldsymbol{s})] = \mathrm{p}(\boldsymbol{s}) = \frac{\lambda_1(\boldsymbol{s})}{\lambda_0(\boldsymbol{s}) + \lambda_1(\boldsymbol{s})}$$



$\mathrm{n} = 73$ background points

Binary GLM Spatial approach

Our model can then be built using selected covariates we desire for predictors and performing a binary logistic regression. The resulting prediction model will return the probability of the occurrence of a large wildfire for a given input location and corresponding values for the model's chosen covariates.



66 incidence points, 73 background points

Binary GLM Spatial approach

Model Selection

The selection of covariates of the model utilized a semi-automated process that minimized AIC. There were some issues with model convergence for some spatial covariance fits, and AIC did not necessarily favor significant predictors in the end.

```
spglm_formula <- Wildfires ~  I(sqrt(distance_rd_min_isprisec)) + I(log(pop.density)) +
  Precipitation_Buffered * Temp_Min_Buffered + I(get_season(FireDiscoveryDateTime)) +
  mean_grass * mean_forest

az_wf_spcov <- spcov_initial("wave")

az_wf_spglm <- spglm(spglm_formula, data = model_data_sf,
                     family = binomial, spcov_initial = az_wf_spcov)
```

The best AIC model tested utilized square root minimum distance from major roads, log population density data at the coordinates, interactions of environmental factors, and season.

Binary GLM Spatial approach

Model Summary Quick Look

```
Call:
spglm(formula = spglm_formula, family = binomial, data = model_data_sf,
    spcov_initial = az_wf_spcov)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.79364 -0.37217 -0.03644  0.38603  2.67112

Coefficients (fixed):
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.31282    7.91520  -0.292   0.7701
I(sqrt(distance_rd_min_isprisec))      -1.32153    0.81237  -1.627   0.1038
I(log(pop.density))                    -0.10851    0.49135  -0.221   0.8252
Precipitation_Buffered                  0.06916    2.08641   0.033   0.9736
Temp_Min_Buffered                      -1.12366    0.78317  -1.435   0.1514
I(get_season(FireDiscoveryDateTime))    0.54990    0.37428   1.469   0.1418
mean_grass                              1.07408    1.74733   0.615   0.5388
mean_forest                             1.30509    1.21915   1.070   0.2844
Precipitation_Buffered:Temp_Min_Buffered  0.39667    0.52455   0.756   0.4495
mean_grass:mean_forest                 15.29506    9.16978   1.668   0.0953 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Pseudo R-squared: 0.3684

Coefficients (wave spatial covariance):
```

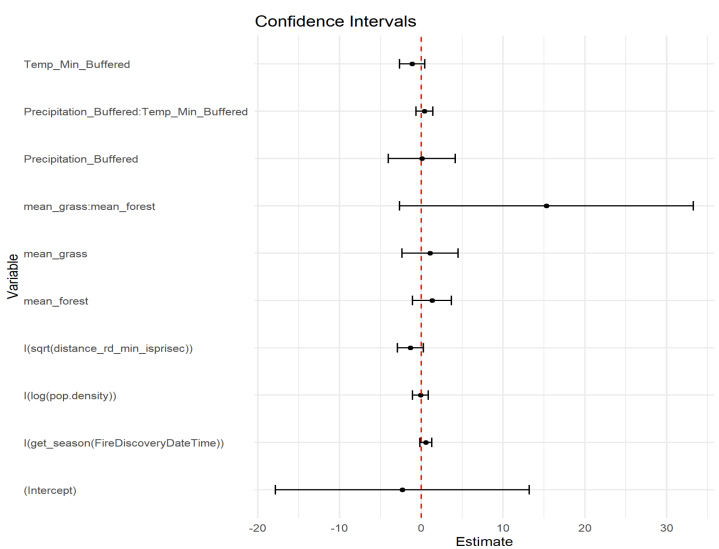
$$\sigma_{\text{sill}} = 2.832e+00$$
$$\sigma_{\text{nugget}} = 3.549e-0$$
$$\text{range} = 3.554e+0$$

Reasonable spatial fit diagnostics!

Binary GLM Spatial approach

CIs Quick Look

```
> confint(az_wf_spglm)
                                                2.5 %      97.5 %
(Intercept)                                -17.8263262 13.2006939
I(sqrt(distance_rd_min_isprisec))           -2.9137530  0.2706848
I(log(pop.density))                         -1.0715342  0.8545142
Precipitation_Buffered                      -4.0201412  4.1584522
Temp_Min_Buffered                           -2.6586537  0.4113346
I(get_season(FireDiscoveryDateTime))        -0.1836816  1.2834872
mean_grass                                  -2.3506191  4.4987742
mean_forest                                 -1.0844096  3.6945865
Precipitation_Buffered:Temp_Min_Buffered    -0.6314373  1.4247793
mean_grass:mean_forest                      -2.6773800 33.2675083
```

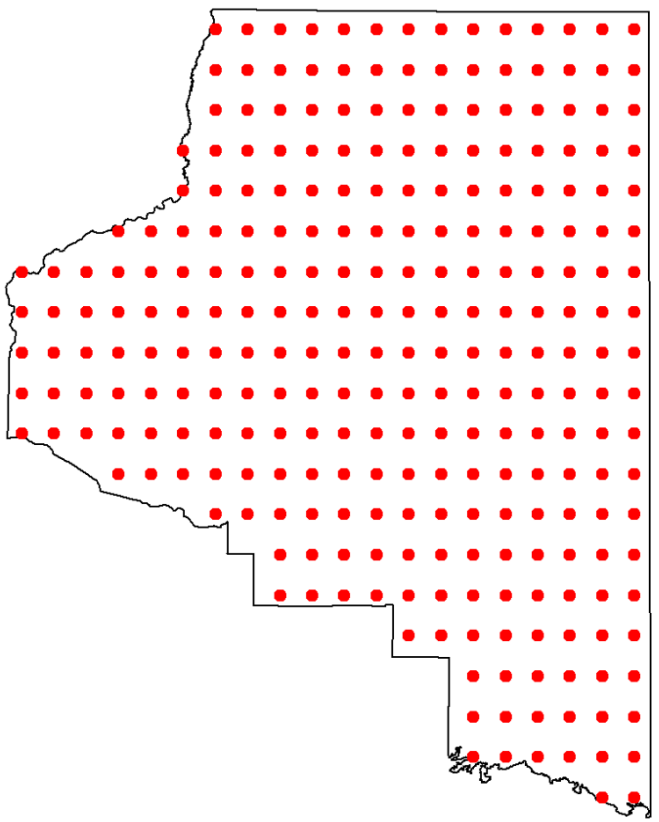


(a) CIs for fixed effects

Figure 14: Confidence Intervals

Binary GLM Spatial approach

Predictions

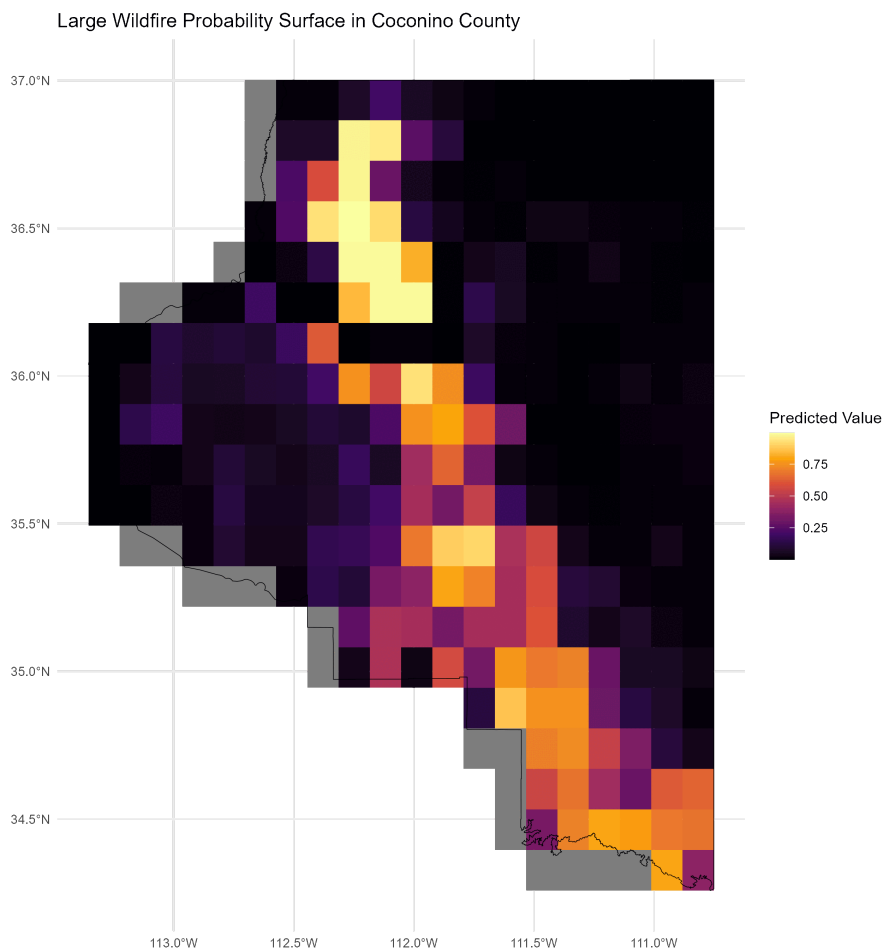


Due to some challenges we faced with the computational time/energy to recover realistic data for covariates, our ability to generate appropriate covariate data for predictions disallowed predictions with high resolution.

We generated a grid of equally spaced points within Coconino County, amounting to 293 total points to use as prediction locations. All prediction covariate data was captured for `2023-12-31 12:00:00`.

Binary GLM Spatial approach

Predictions

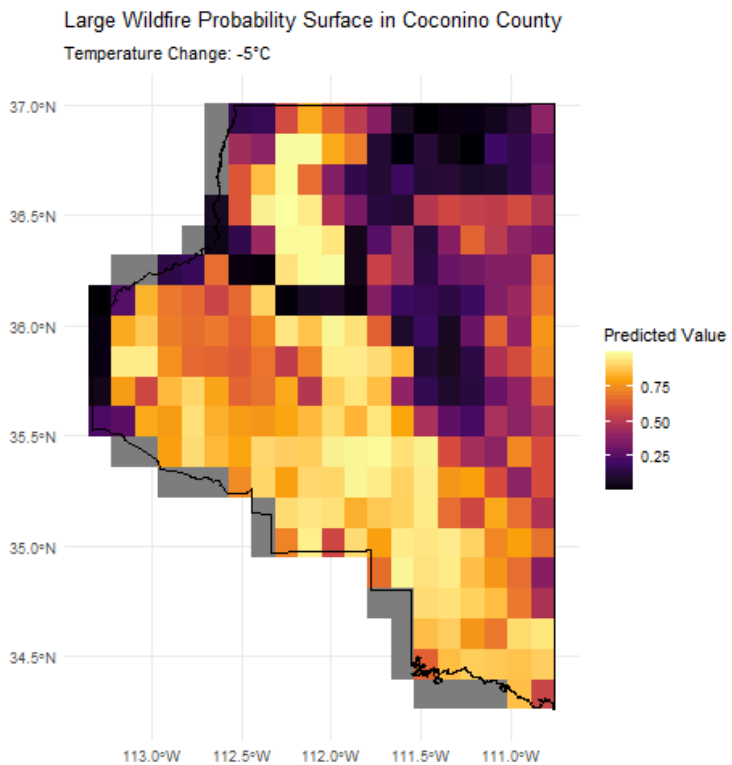


This is the result prediction surface (treating each grid point prediction as representative probability for the pixel's region), and the wildfire incidence points as well as the areas of national forests in the region overlayed.

Not totally outrageous!

Binary GLM Spatial approach

Well...maybe a little outrageous?



(a) +/- 5°C `Temp_Min_Buffered` sweep

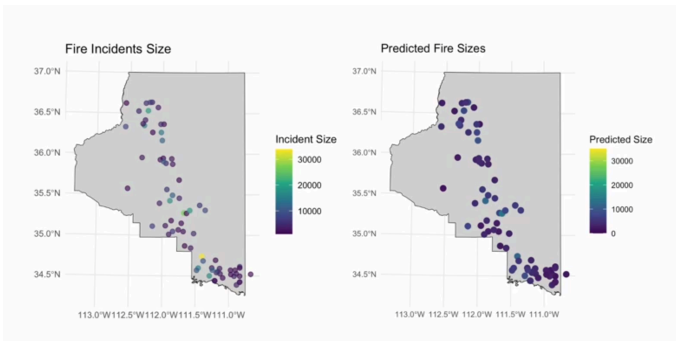
(b) +/- 10% `Precipitation_Buffered` sweep

Figure 15: Prediction surface for adjusting annual min temperature and precipitation

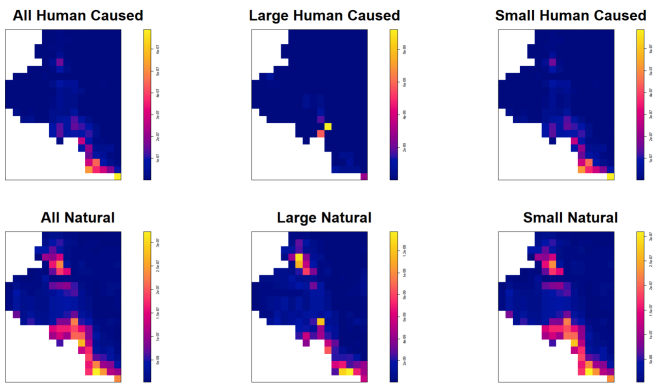
More background process points? Predictors correlated? Spatial confounding?

Conclusions and Future Research

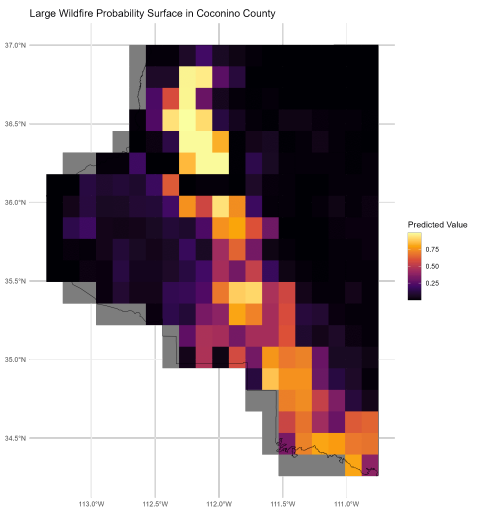
- Models at least demonstrate proof-of-concept (RQ1)
- We found covariate data that improved our models (RQ2)
- We found some interesting patterns of CSR/non-CSR for different types of fire incidence (RQ3)



(a) Spatial Linear Model



(b) Log-Gaussian Cox Process



(c) Binary Spatial GLM

Figure 16

Conclusions and Future Research

With some extra refinement, they could feasibly aid in better allocation of resources for fire prevention, possible forecasting, and information that can be coupled with other ecological models.

Some considerations for future model refinements.

- Better methods for recovering covariate data for better predictions
- Spatio-Temporal modeling?
- More careful treatment of highly correlated predictors
- More points for background realization for Binary model
- Other model scoring/tuning

There is also a lot more that can be done with this dataset, we encourage you to check it out!

Thank you!

Questions?