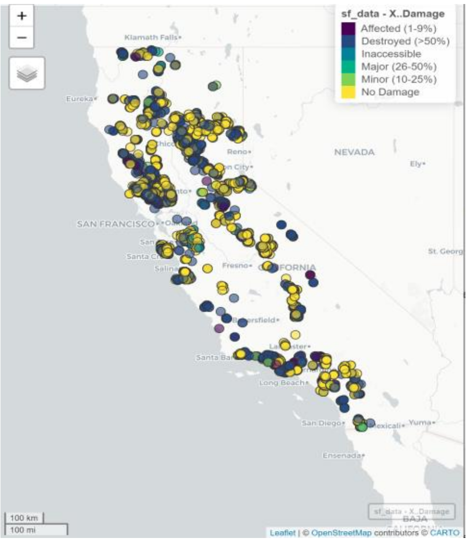
An Exploration Into Applying Spatial Methods for Wildfire Data

Benson, Daniel, Steven

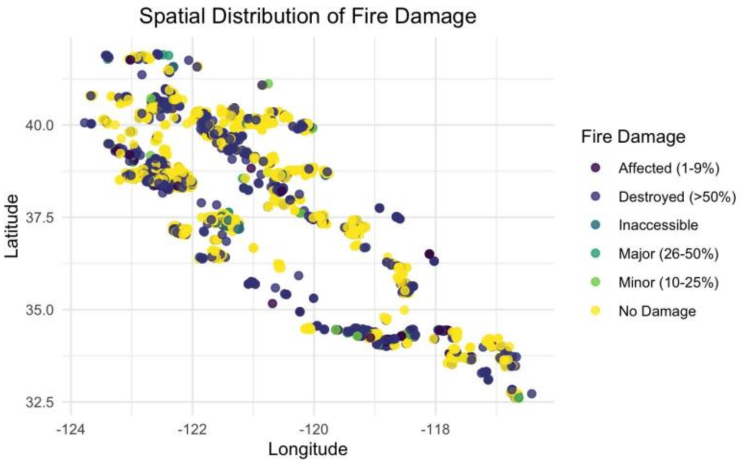
Introduction:

Among the natural disasters humans deal with on a regular basis, wildfires tend to rank on the lower end in terms of calamity and catastrophe. That being said, they are still the cause of many deaths, damage to the environment, and destruction of property. Though wildfires can be beneficial to ecosystems by clearing overgrown vegetation and dead organic material, the amount of occurrences has substantially increased in recent years as a result of climate change and other factors. Recent headlines include record breaking wildfires in Hawaii and Greece (2023). We believe that spatial methods can help us learn more about the distribution of wildfires in a particular area. The raw dataset includes data on over 100000 California wildfires including latitude, longitude, county, and building damage (No Damage, Affected (1-9%), Minor (10-25%), Major (26-50%), and Destroyed (>50%)). For computational efficiency, we sampled 10000 wildfires from this dataset stratified by county and removed ones with low counts (1 or 2 fires).

Initial mapview of the wildfires:



Distribution by Latitude and Longitude:



Methods:

We expect wildfires to exhibit clustering and spatial autocorrelation due to their nature. To quantify this, we ran Geary's C and Moran's I tests. Geary's C measures local spatial autocorrelation, which tells us how similar or different nearby locations are in terms of wildfire damage. (Do nearby locations have similar or notably different wildfire damage?) Moran's I measures global spatial autocorrelation, which tells us if similar wildfire damage levels cluster across California. (Are areas with similar levels of wildfire damage in California clustered together?)

We also wanted to use explore fitting SAR (spatial autogressive) and CAR (conditional autoregressive) models for our data. The main issue is that the response (building damage) is an ordinal variable. Though it represented the % of building damage, it only is recorded as a range and not a precise value. To convert it into a continuous response, we mapped the categories such that "No Damage" -> 1, "Affected (1-9%)" -> 2, "Minor (10-25%)" -> 3, "Major (26-50%)" -> 4, "Destroyed (>50%) -> 5". There were 58 wildfires with an "Inaccessible" response which were removed from the dataset. For computational practicality, we also subsetted the (already subsetted data) to around 2000 wildfires.

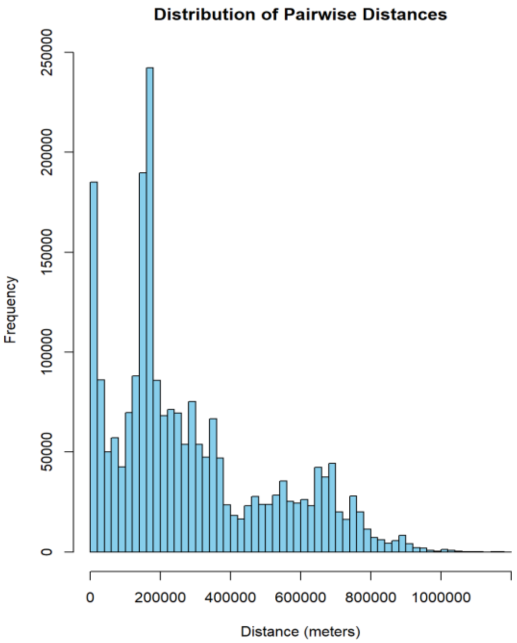
Finally, we went back and fitted a variogram to analyze the spatial variability of building damage in in terms of semivariance (half of the average squared distances in the response between wildfire locations). We wish to see if these results conflict or are consistent with each other.

Results:

	Test Statistic	P-value	Expectation	Variance
Geary's C	0.395494	$2.2*10^{-16}$	1	$4.12*10^{-5}$
Moran's I	0.603004	$2.2*10^{-16}$	$-9.46*10^{-5}$	$4.06*10^{-5}$

In creating the spatial weights matrix, we used k = 4 nearest neighbors. Under the null hypothesis (no spatial autocorrelation), we would expect the test statistic for Geary's C to be 1 and Moran's I to be 0. We see from our p-values that there is strong evidence of positive global and local spatial autocorrelation. This is consistent with what we would expect from our initial visualization and beliefs about wildfires.

For the SAR/CAR models, to pick a reasonable cutoff (maximum distance that defines spatial data), we plotted the distribution of the pairwise distances between wildfires and took the 25th percentile. This resulted in a cutoff of 136178.8.



quantile(distances_subset, 0.25) = 136178.8

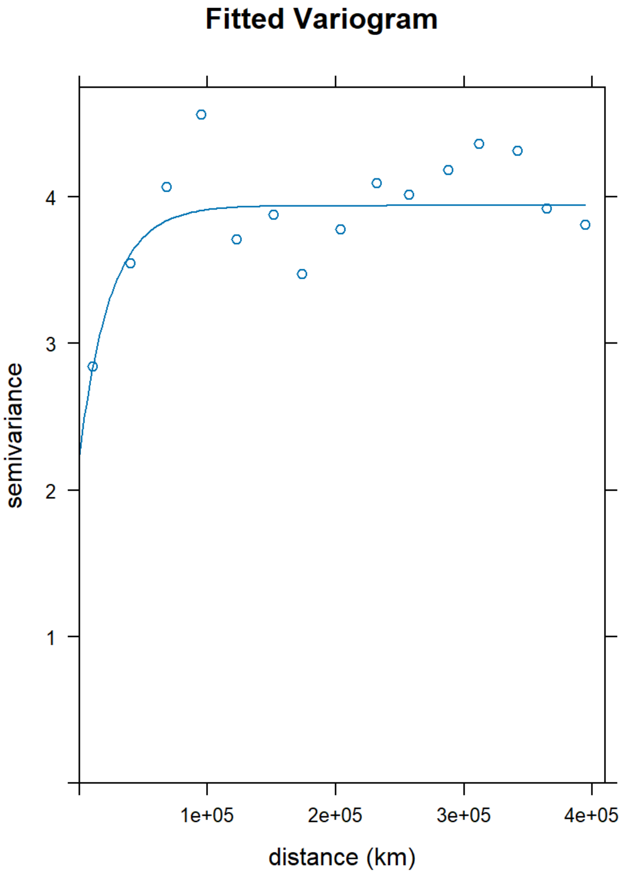
Cutoff = 136178.8	Intercept	Standard Error	de	range
SAR Model	2.58959	0.08197	3.662600	0.001165
CAR Model	2.64590	0.07088	3.662600	0.001178

The results are very similar for both models. Using no other predictors, the mean wildfire level is around 2.5 (which would correspond to between minor and major damage or around 37.5%). A high de (spatial variance) with low range (spatial decay) suggests a strong local spatial dependence. It implies that damage levels are only highly correlated within close distances.

Changing the cutoff to 5000 (restricting many neighbors), we obtain these results:

Cutoff = 5000	Intercept	Standard Error	de	range	extra
SAR Model	2.81456	0.04569	3.142689	0.003149	3.597203
CAR Model	2.84336	0.04552	3.305546	0.003154	3.560854

The “extra” term is extra variance added by the model to compensate for missing long-range effects. We see that the mean wildfire level has increased slightly, but for the most part, drastically changing the cutoff doesn’t seem to affect the results very much. This implies that fire damage is highly localized (i.e. the strongest influence comes from the nearest neighbors).



We chose a spherical model for the fitted variogram due to the “plateauing” of points. It shows relatively high semivariance at almost all distances. The parameters are represented here.

model	psill	range
Nug	1.652569	0.00
Sph	2.073052	22157.62

We can calculate the total sill from the nugget and partial sill (sph)

$$Total\ sill = 1.652569 + 2.073052 = 3.725621$$

This represents the maximum semivariance the variogram reaches once spatial correlation disappears. The relatively short range (22km) implies that spatial correlation drops off very quickly. This is sort of at odds with the results from our earlier models. Perhaps the discreteness of the response isn't great for this type of model. The distribution of our data (the initial stratified sampled one) is very binomial as seen in this table.

	1	2	3	4	5
# of responses	4412	369	107	54	5571

Since the data is dominated by 1's and 5's, many pairs likely yield a squared distance of $(5-1)^2$ = 16, which push the semivariance up dramatically. The SAR and CAR models showed us the predicted average wildfire damage level, which would be around the middle given the distribution. If anything, it shows that one method or model might not tell the whole story.

Note: The supplementary variogram cloud was omitted due to plotting issues

This project was intended more as a way to utilize several of the methods discussed in class on a real dataset than to draw any ground-breaking inference. There were some notable issues that may affect the validity of our results. Namely, our response was an ordinal one that was converted and approximated to a continuous response. This simplifies many of the calculations but also means we should take the values with a grain of salt. This dataset also doesn't provide information that could've been very useful in modelling such as human response time and various environmental factors that would make some areas more prone to wildfires than others. Overall, this was a modest albeit clumsy dive into applying spatial methods for a limited dataset.