Code ▾

Risk Modeling - Predicitve Policing

Palak Agarwal

October 23th, 2020

Introduction

One of the main aims of policing making is to make smart and efficient decisions in terms of resource and capital allocation which benefits large populations. It is always a comparison between initial investment to long trend usage and benefits. Policing has driven public-sector machine learning because law enforcement has significant planning and resource allocation questions. In this assignment, we look at the intersection of public sector planning and data science to ensure that the supply of a limited resource matches the demand for those resources. There are many traditional methods that exist but none have been proven effective.

In this assignment I looked at risk predictions for assaults. Given the current political and social scenario, we know that current policing and resource planning is a result of systematic racism, racial discrimination, and biased judgments. The most important result and outcome out of this prediction should be that it does not emphasize the same biases that are evident.

While we use a crime type and factors that affect it, one needs to be aware of the reporting bias that exists and not every crime location gets reported.

Set Up

The features in this model were engineered using the functions: **Nearest neighbor.** The nearest neighbor function finds the average distance from the measuring feature to the measured feature. The function requires 3 input variables - the dependent feature to measure from, the features to measure to, and the number of features. For example, a nearest neighbor feature could measure the average distance from each house to its three closest public parks.

Code

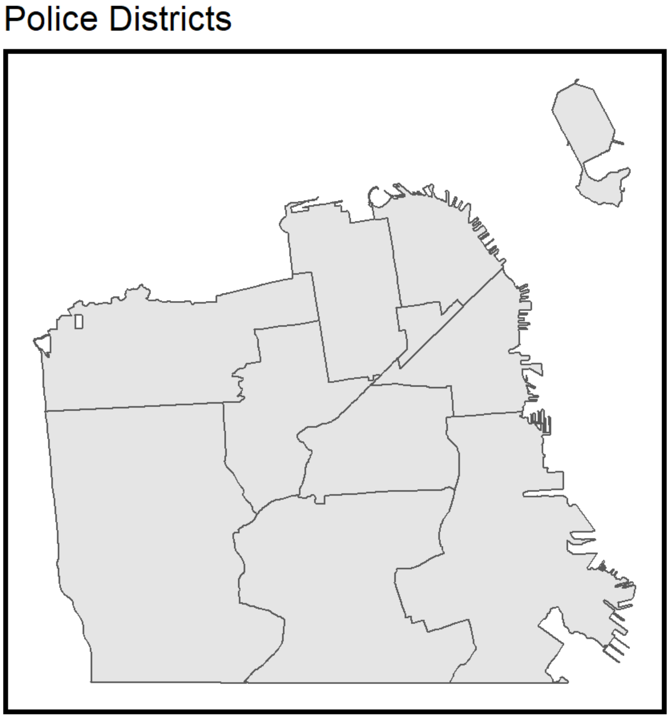
Data Wrangling

Loading data and creating fishnet

The crime that I have picked to look at is Assaults in San Francisco. By definition - it is the act to make a physical attack on. Here I have read in the assaults from 2017, San Francisco boundary and police districts. After that we create a fishnet over SF which will be used as the grid. I chose a grid 300 by 300 to account for finer resolution in the hotspots.

From Figures 1.1 and 1.2 you can see that the assault reports are clustered in the financial district which is the heart of SF. The area has a large foot fall and is very dense, hence the number of reports in the area is high and could be a result of reporting bias. The result is not surprising because downtown areas have higher population density and denser built environment which typically have higher rates of crime.

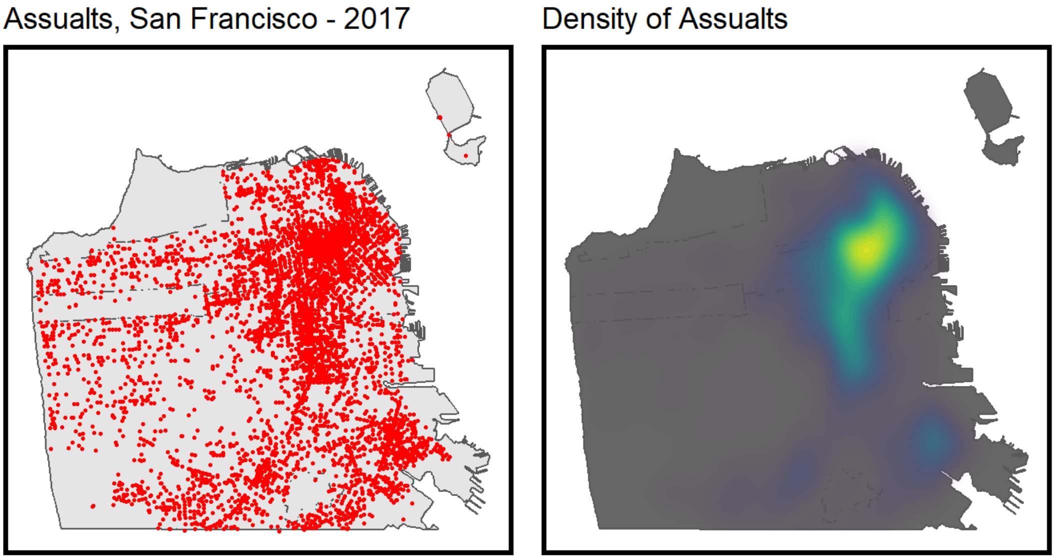
Code



Police Districts

Figure 1.1

Code



Assualts, San Francisco - 2017

Density of Assualts

Figure 1.2

Code

Fishnet for SF

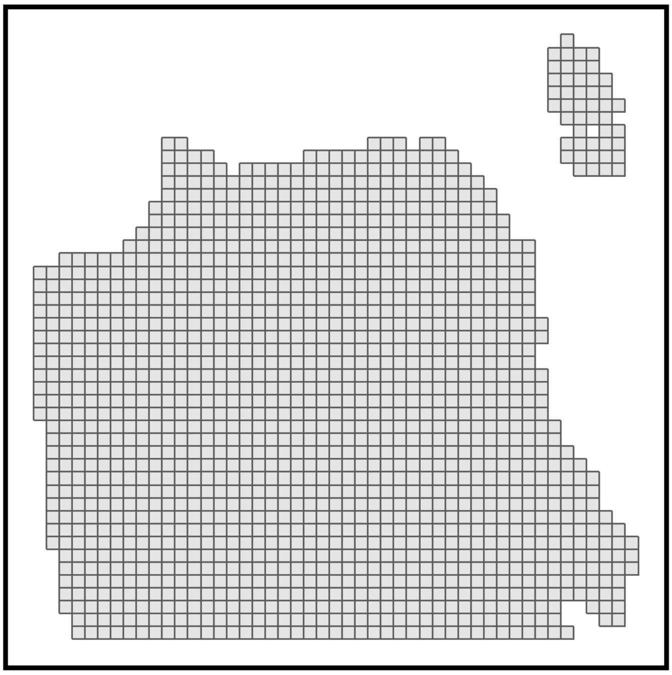


Figure 1.3

Joining assualts to the fishnet

We aggregate the total number of assaults to each grid cell by using the coordinates of the assault and finding which cell it falls within.

Code

Count of Assualts for the fishnet

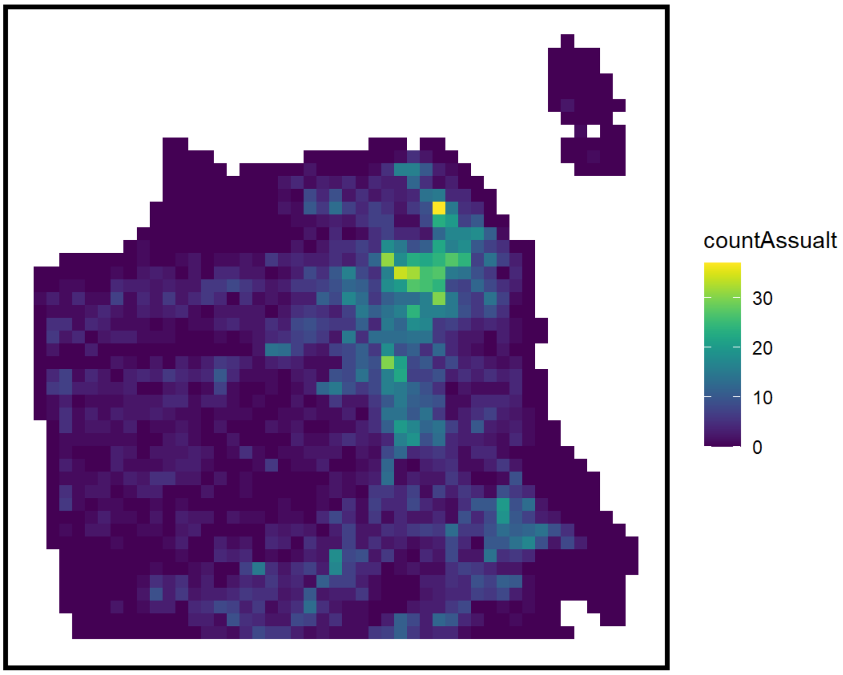


Figure 2.1

Modeling risk factors

Next, we move on to modeling the risk factors that are the independent variables. I used the following variables : **Encampments**, **Street lights out**, **Graffiti**, **Noise Reports**, **Abandoned Cars**, **Bars**, **Parks** and **Liquor stores.** The first five variables were taken from 311 calls and they likely have a reporting bias that one needs to aware of. Also, some variables marginalize and make assumptions connecting to one or more communities which add to the racial and systematic bias.

The other variables I looked at but didn't add to the final model are **Homeless calls**, **Drug crimes** and **Domestic violence crimes.** The reason for not using the homeless data was to avoid the unnecessary connection. As you can see from Figures 3.1 and 3.2 the location of drug and domestic violence crimes are clustered in the same location as that of assault crimes and that would have made the model overemphasize them over the other variables.

Code

Code

Assualt and Drug cases

Assualts in red and Drug in Blue

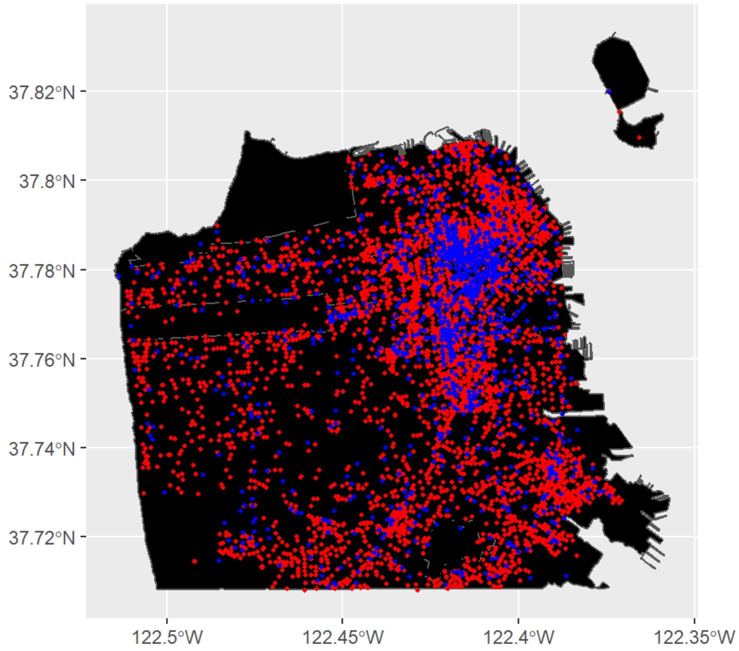


Figure 3.1

Code

Assualt and Domestic violence cases

Assualts in red and Domestic Violence in Blue

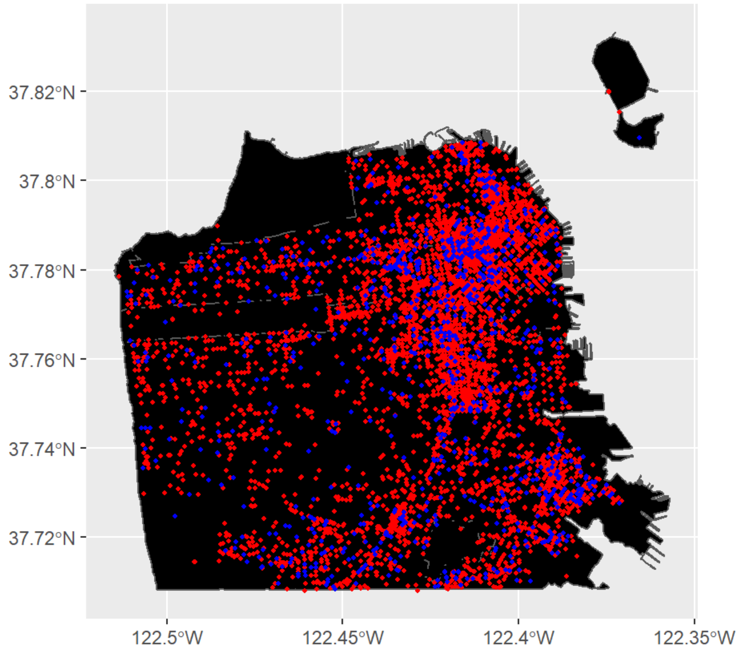
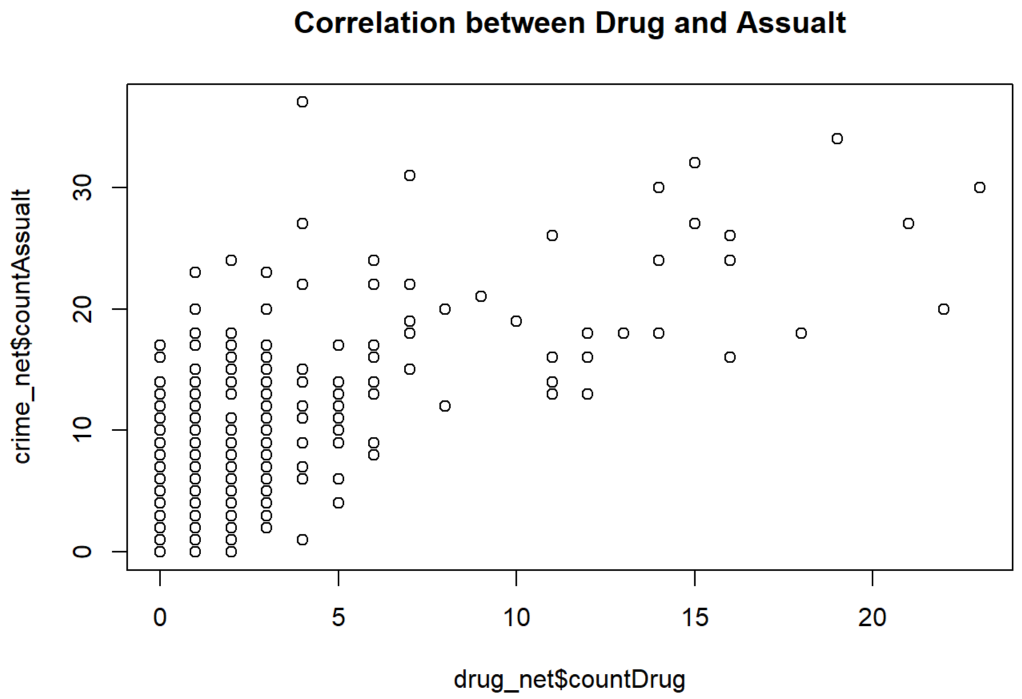
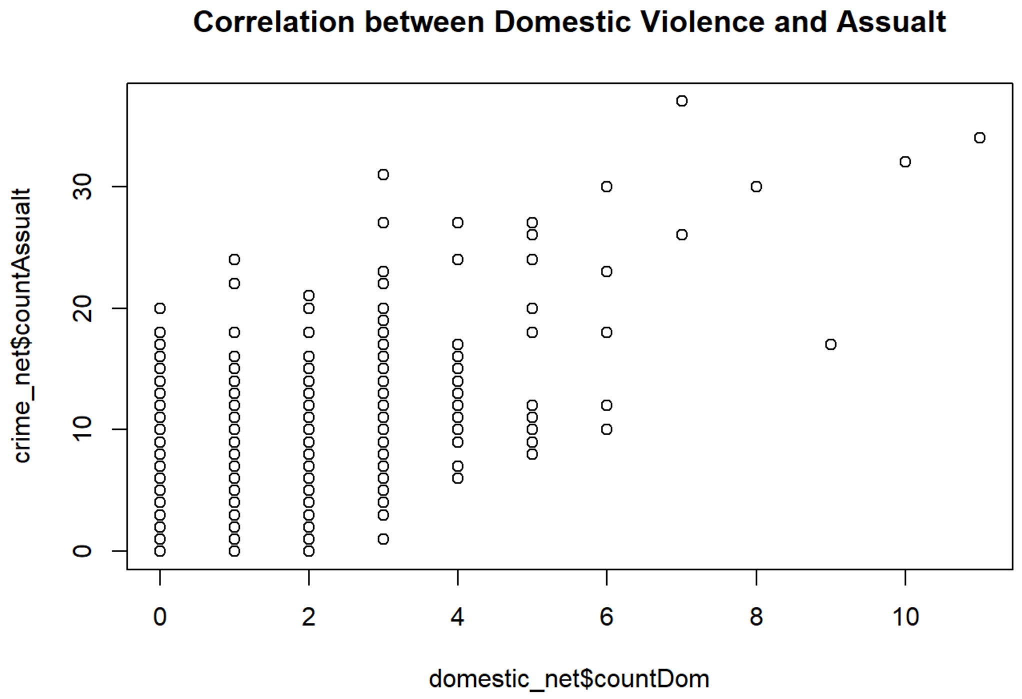


Figure 3.2

Code



Code



The above two scatter plots show that as the count of drug or domestic violence counts increase, so does the count of assaults. To avoid multicollinearity the variables were lefft out.

Feature engineering

Count of risk factors by grid cell

Now we bind all the variables and add it to the fishnet created above to calculate the count within each grid cell. In Figure 4.1 you can see the counts of each risk factor.

Code

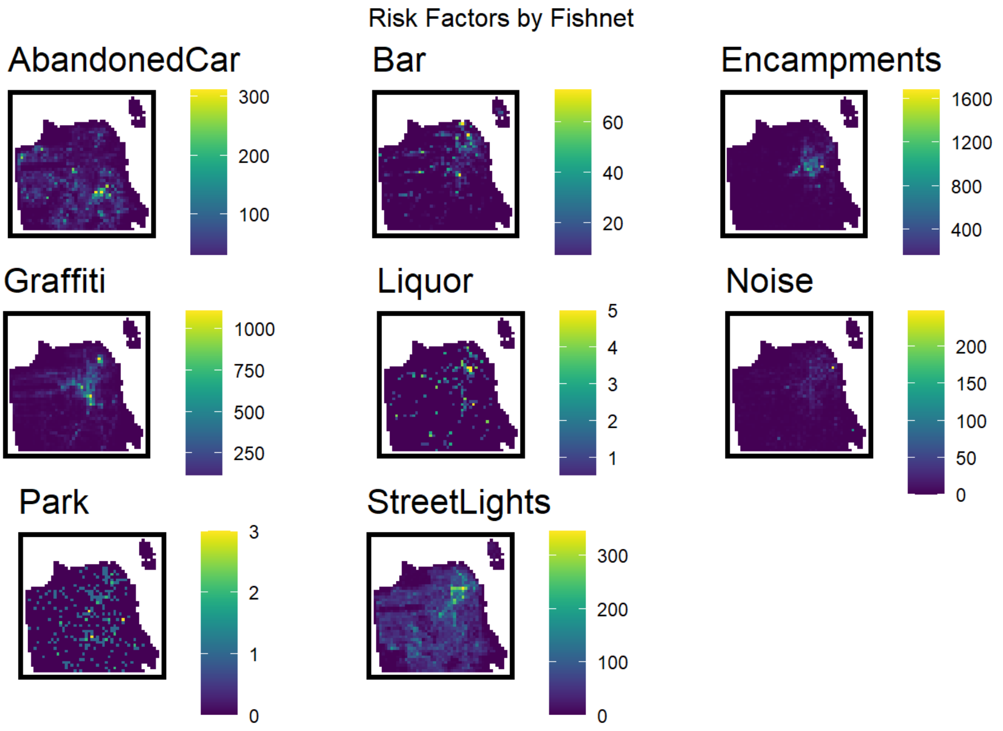


Figure 4.1

Nearest neighbor features

The grid cell imposes a very rigid spatial pattern over the variables. To loosen that up a little, we use the nearest neighbor function to create spatial relation between the risk factors. I ran the model on different k values for the risk factors, and decided which model worked better than the other based on the MAE values. Hence, below you see different k values. Figure 5.1 shows the updated counts after the adding the spatial feature engineering.

Code

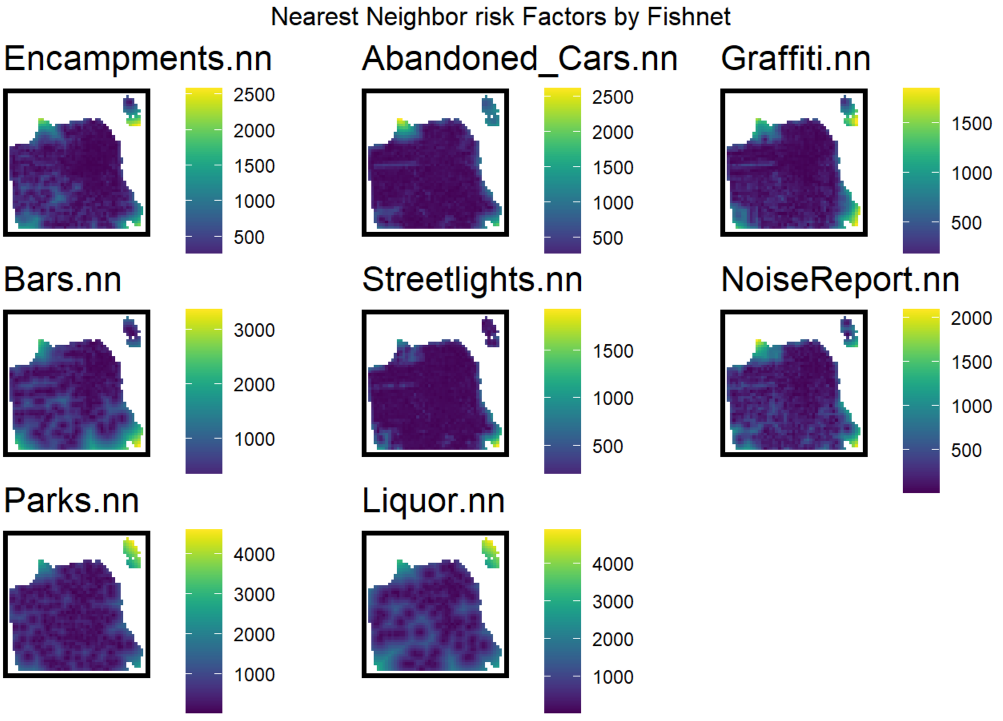


Figure 5.1

Measure distance to main spot in SF

Sf's hotspot is the financial district and the distance of each grid cell to that has been added to the model.

Code

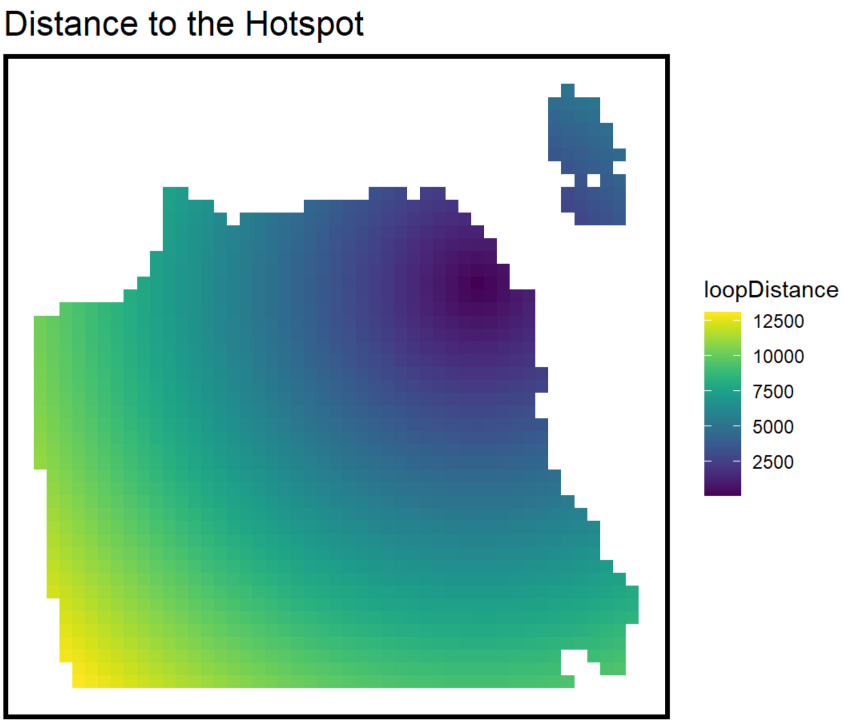


Figure 6.1

Create the final_net

Code

Local Moran’s I

‘Global’ Moran’s I is used to test for spatial autocorrelation at larger neighborhood scales. This information provided insight into the spatial process accounting for neighborhood scale clustering, but not clustering at more local scales. Hence, that local spatial process is explored using Local Moran’s I. Here, the null hypothesis is that the assault count at a given location is randomly distributed relative to its immediate neighbors. The tests results have been visualised in Figure 7.1.

Code

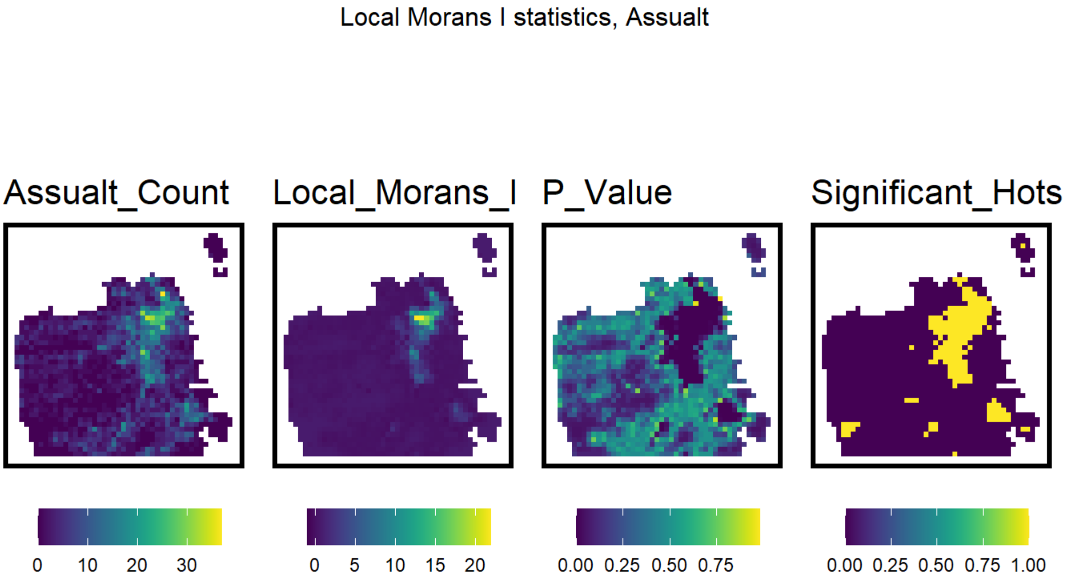


Figure 7.1

Highly significant hotspots

Using areas with p values less than 0.0000001 to identify significant spots. After that the distance of each grid cell is calculated to its nearest significant spot.

Code

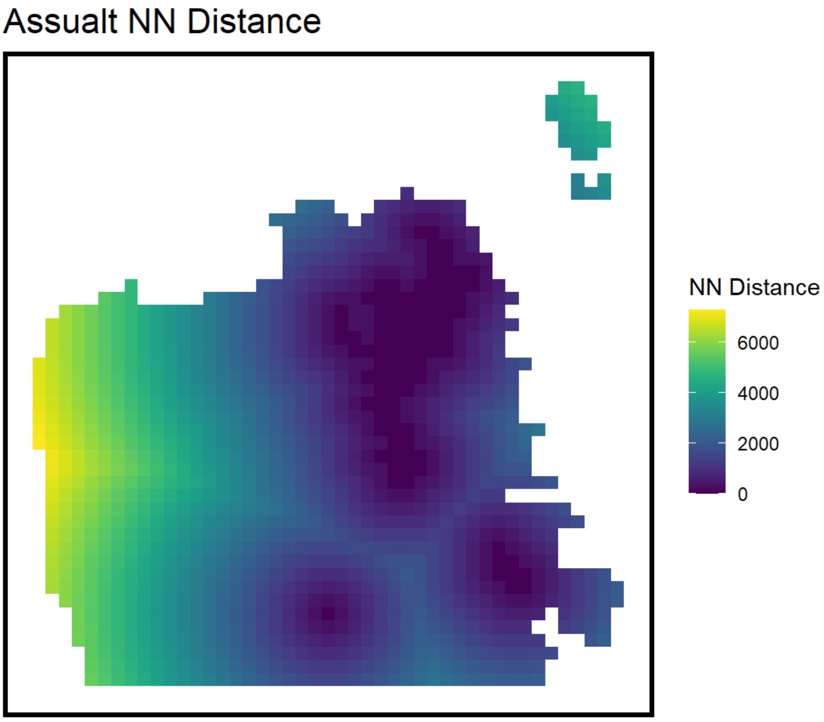


Figure 8.1

Correlation

Correlation gives important context while also providing intuition on features that may predict the count of assaults.

Code

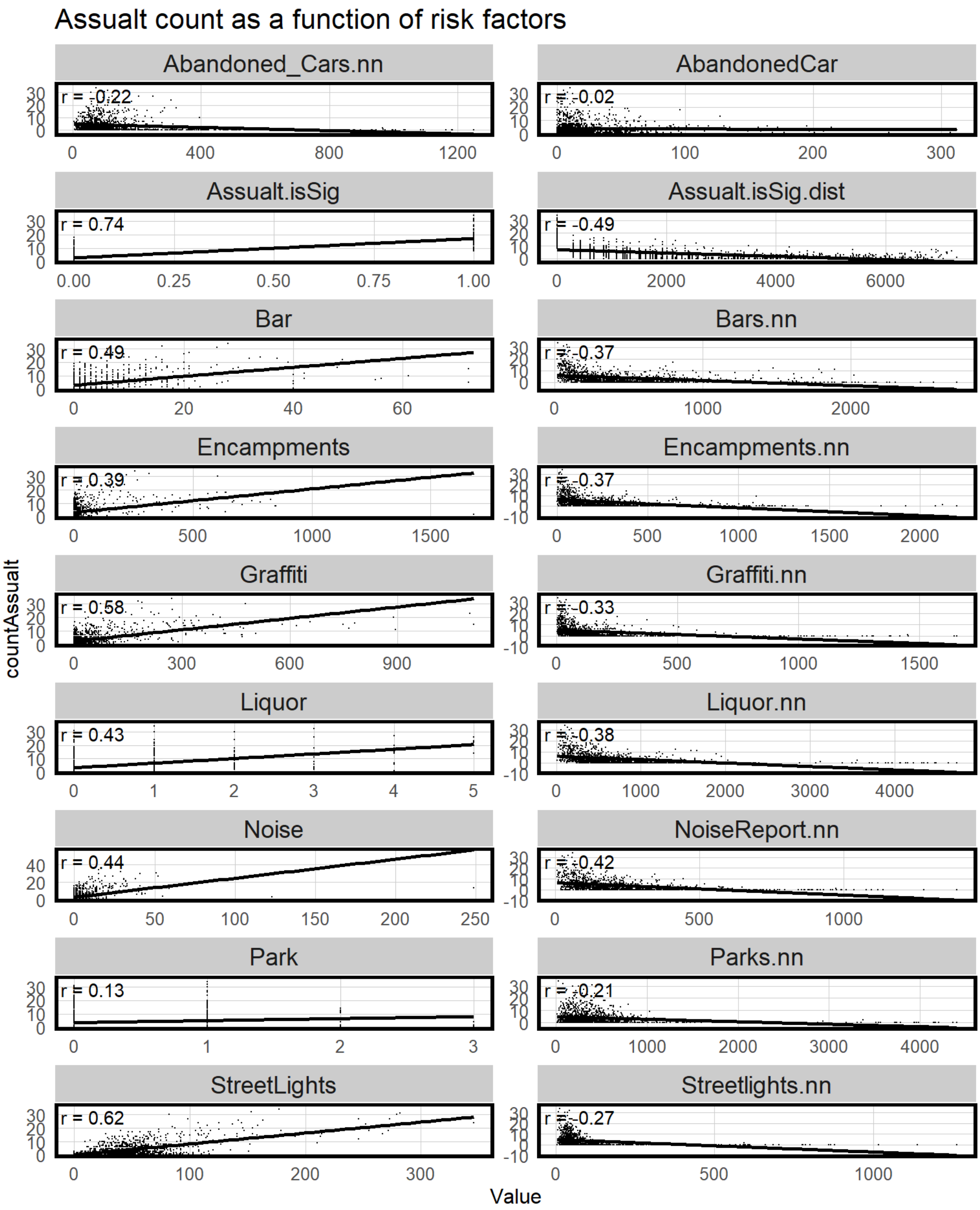


Figure 9.1

Poisson Regression

Cross validation

While predicting crime the model should be good at predicting the crime risk 'experience' at both citywide and local spatial scales. The best way to test for this is to hold out one local area, train the model on the remaining n - 1 areas, predict for the hold out, and record the goodness of fit. In this form of spatial cross-validation called 'Leave-one-group-out' cross-validation (LOGO-CV), each neighborhood takes a turn as a hold-out.

Also two different lists of variables were created, one has **Just risk factors** and the other is includes the risk facotrs with the Local Moran's I **Spatial Process**. Figure 10.1 and 10.2 show the predicted count of assaults and the observed counts. The model does a pretty good job with the predictions.

Code

Code

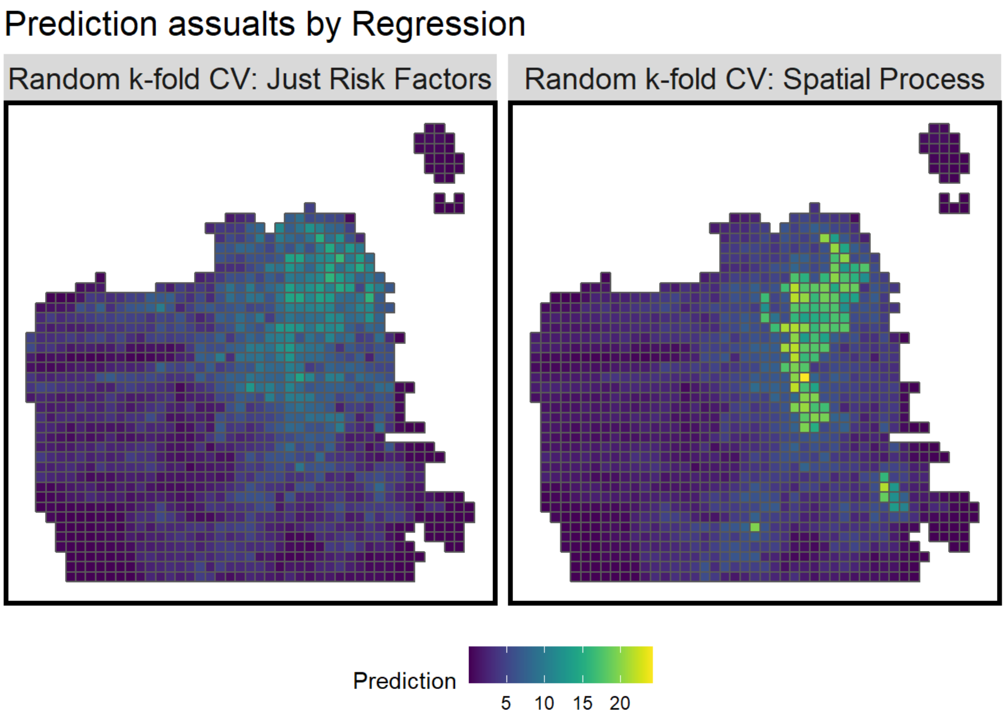

Prediction assualts by Regression
Random k-fold CV: Just Risk Factors
Random k-fold CV: Spatial Process
Prediction
5
10
15
20

Figure 10.1

Code

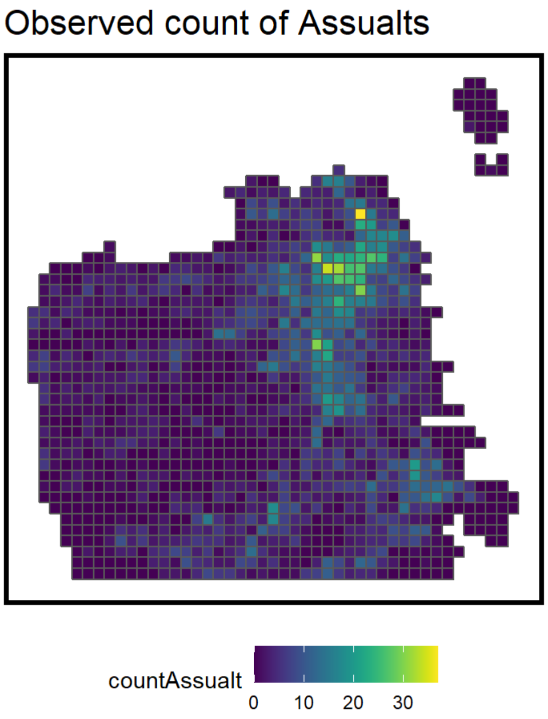


Figure 10.2

Accuracy & Generalzability

All the figures and tables below shows that adding spatial processes to the model, helps remove the large errors in the model and the mean MAE.

Code

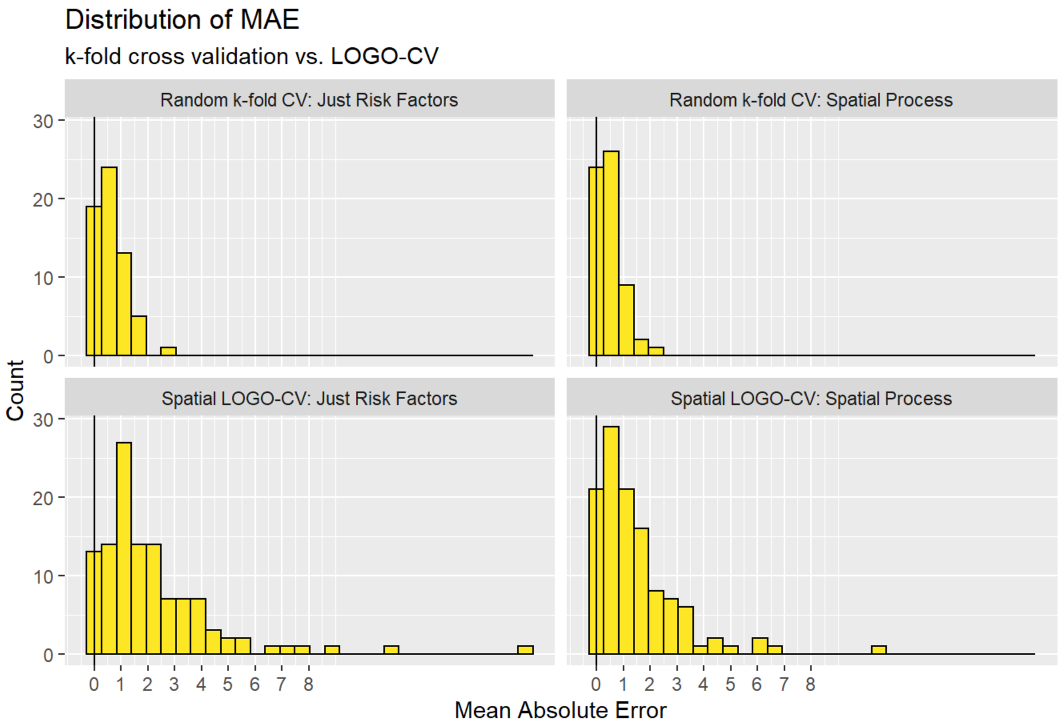


Figure 11.1

Code

MAE by regression

Regression	Mean_MAE	SD_MAE
Random k-fold CV: Just Risk Factors	0.65	0.53
Random k-fold CV: Spatial Process	0.52	0.44
Spatial LOGO-CV: Just Risk Factors	2.22	2.31
Spatial LOGO-CV: Spatial Process	1.48	1.58

Table 1.1

Code

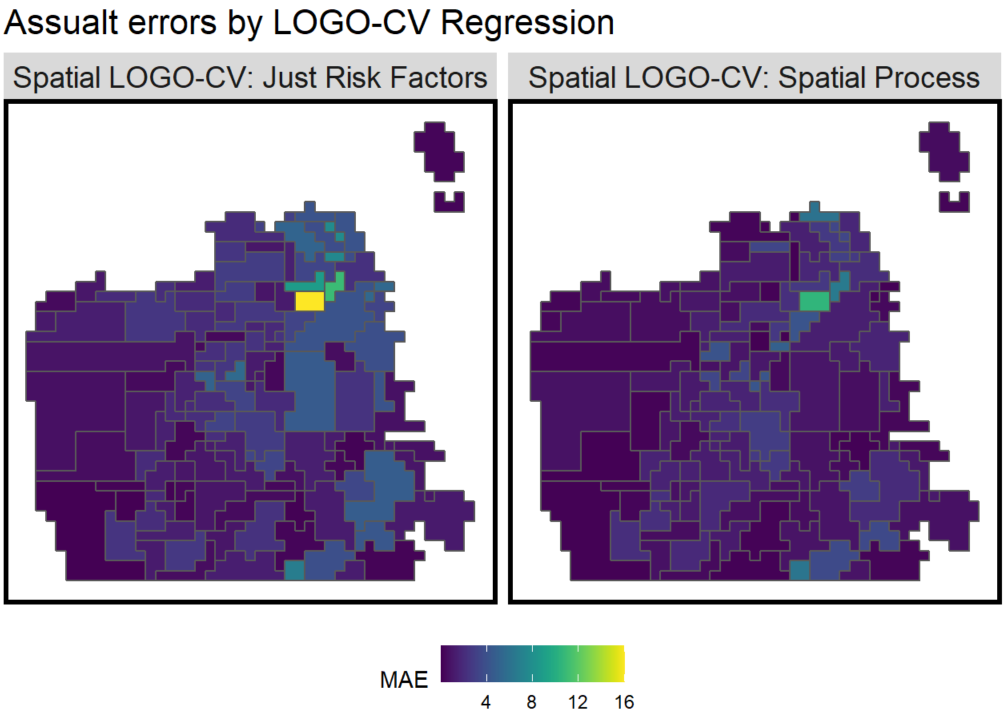

Assualt errors by LOGO-CV Regression
Spatial LOGO-CV: Just Risk Factors
Spatial LOGO-CV: Spatial Process
MAE
4
8
12
16

Figure 11.2

Local Spatial process

Initially the Local Moran’s I was calculated on queen neighbors, here we test it on neighborhood scale. We can see from Table 2.1 we can see that the model performs well at the neighborhood scale as well.

Code

Moran’s I on Errors by Regression

Regression	Morans_I	p_value
Spatial LOGO-CV: Just Risk Factors	0.3393603	0.001
Spatial LOGO-CV: Spatial Process	0.2087450	0.001

Table 2.1

Generalizability by neighborhood

From Figure 12.1 you can see that the model over predicts in low assault areas and under predicts in high assaults areas. This is not the most ideal scenario, and indicates more risk factors might need to be added to the model.

Code

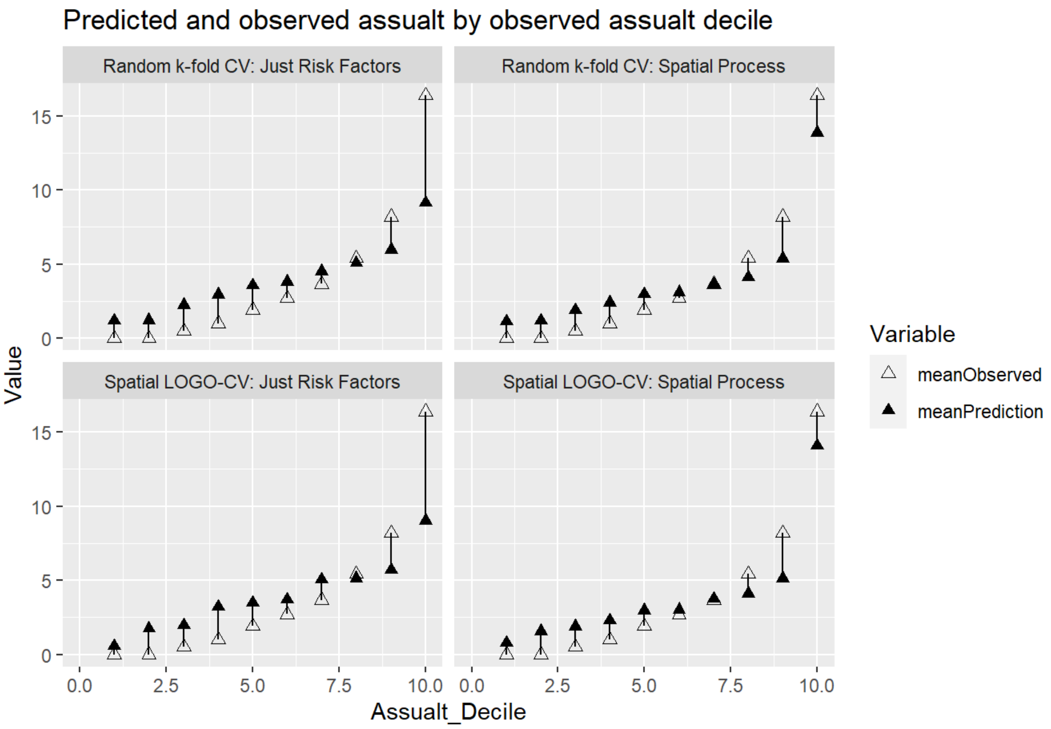


Figure 12.1

Generalizability by race

As mentioned before, current resource allocation is a result of systematic racism and hence the risk prediction model should be able to over come that bias. As its visible from table 3.1 the model under predicts in Major Non-white areas and over predicts in White areas which is the ideal result we hope to see.

Code

Code

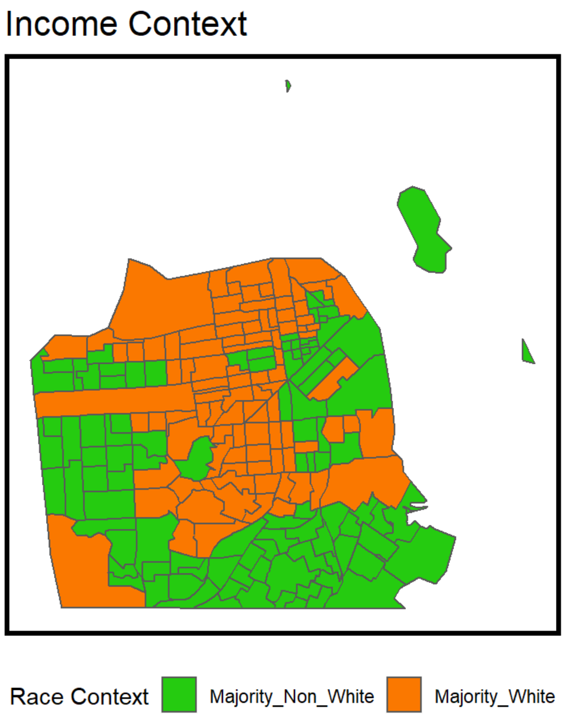


Figure 13.1

Code

Mean Error by neighborhood racial context

Regression	Majority_Non_White	Majority_White
Spatial LOGO-CV: Just Risk Factors	-1.1108024	1.2348325
Spatial LOGO-CV: Spatial Process	-0.6504104	0.7038001

Table 3.1

Kernal Density vs Risk Prediction

Traditionally, the police allocates its resources using kernel. density models which smooths out the hotspots into a continuous surface which is visualized in Figure 14.1. You can see the drawbacks of this model in Figure 14.2 in which the kernel density is overlaid with the sample from the assaults.

Code

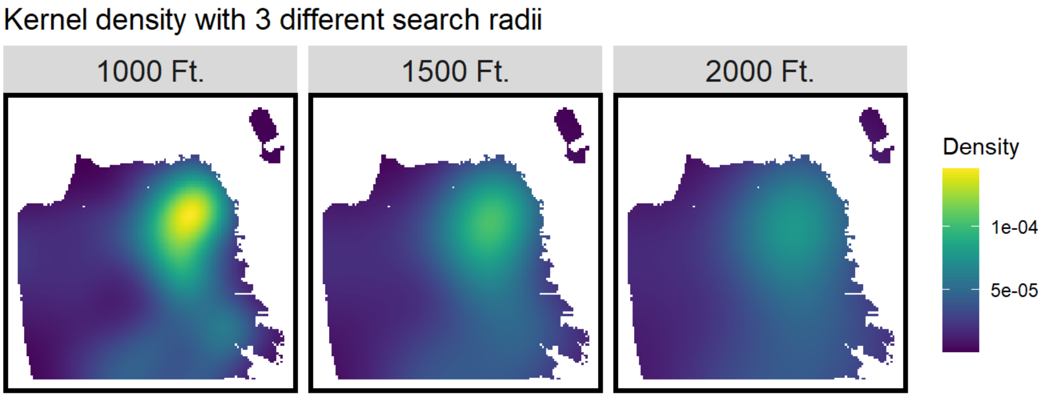


Figure 14.1

Code

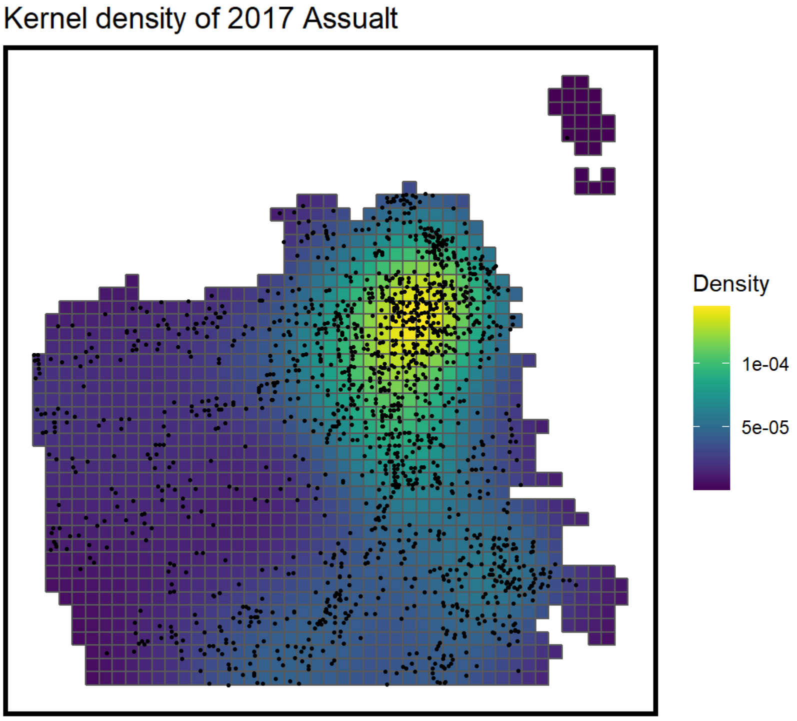


Figure 14.2

Goodness for fit

To test the model for goodness for fit, we run the model on 2018 assaults and see that the model does a better job predicting the locations.

Code

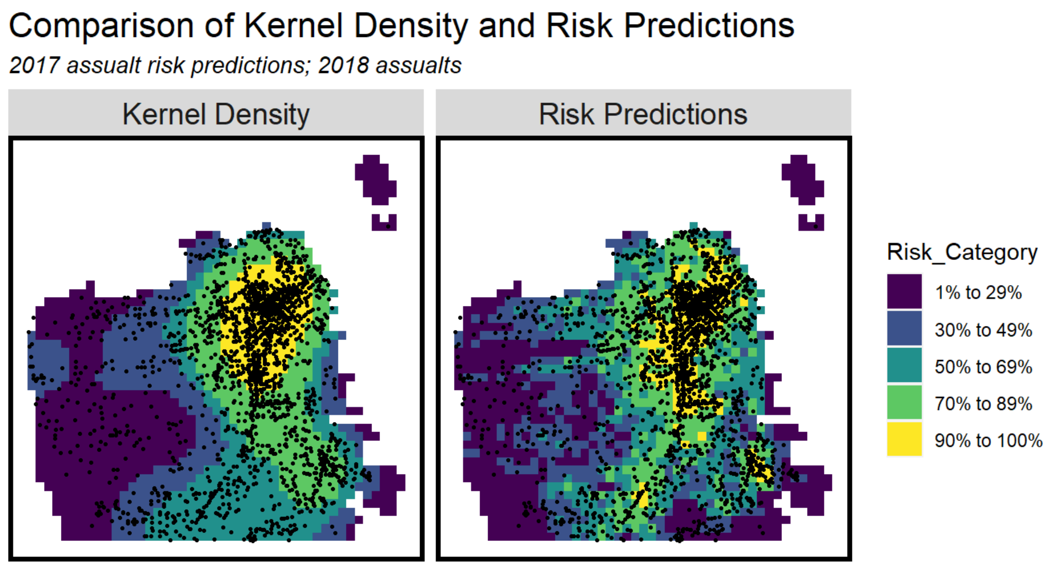

Comparison of Kernel Density and Risk Predictions
2017 assualt risk predictions; 2018 assualts
Kernel Density
Risk Predictions
Risk_Category
1% to 29%
30% to 49%
50% to 69%
70% to 89%
90% to 100%

Figure 15.1

Code

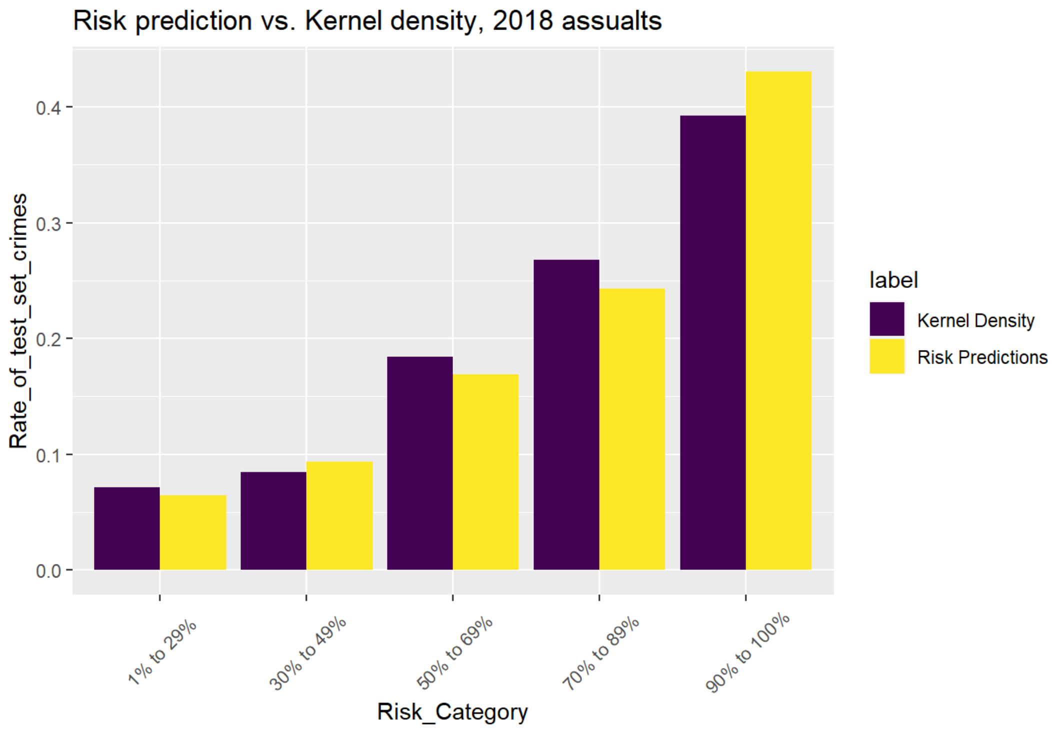


Figure 15.2

Conclusion

We have seen that the model does well generalizing and predicting across different scales but we cannot be sure the model doesn’t suffer from selection bias. As discussed in the introduction the data used has a reporting bias due to the self-reported nature of the data, which goes unaccounted for in the model.It is impossible to account for racial biases in existing resource allocation, creating a potential feedback loop for this kind of bias. While this models may not be completely appropriate for crime prediction, there are a host of other planning outcomes that could benefit greatly. The model does best with what data it is fed, so the issue is with the data itself over the model elements.

However, the predicted results match well with the reported data for 2018, hence it is fair to say that the benefits in cost reduction and resource allocation are high enough to make this a better modeling strategy over the traditional kernel density method. The model can be made smarter with the introduction of other variables and spatial processes, which can work towards reducing biases. If the model is implemented, future datasets become much smarter, leading to greater accuracy and generalizability in future predictions.