Predicting San Francisco Crime Rate by Census Block

Kenneth Noddings

2023-04-23

Abstract

I attempted to build a model that would predict the property-crime rate by census block in the San Francisco county area using a small number of explanatory data points (property value, proximity to police stations and bus stops, elevation, and approximate street grade) as well as data on crime incidents in San Francisco over the past 20 years. My personal area of interest was in finding how powerful of a predictor street grade is of property crime rates; the hypothesis being that steep streets may causally impact crime by disincentivizing travel up and down them (this would follow general literature on urban crime that suggests there is a causal impact from neighborhood "permeability" on crime rates). I used a gradient boosting machine (gbm) to build this model and found an out-of-sample RMSE of 0.0026/20, which compares to true average property-crime rate of 0.0055/20 crimes per meter squared of area per year (this translates to an average error of 47%), or to the RMSE of a linear model, initialized using the same explanatory variables, of 0.0047/20 (this translates to an average error of 85%). In the GBM model, I found that property valuations had the highest predictive power on property crime rate, followed by police station proximity, absolute elevation, and bus stop proximity. The predictive power of my more direct attempts to measure street grade were all rather low.

Introduction

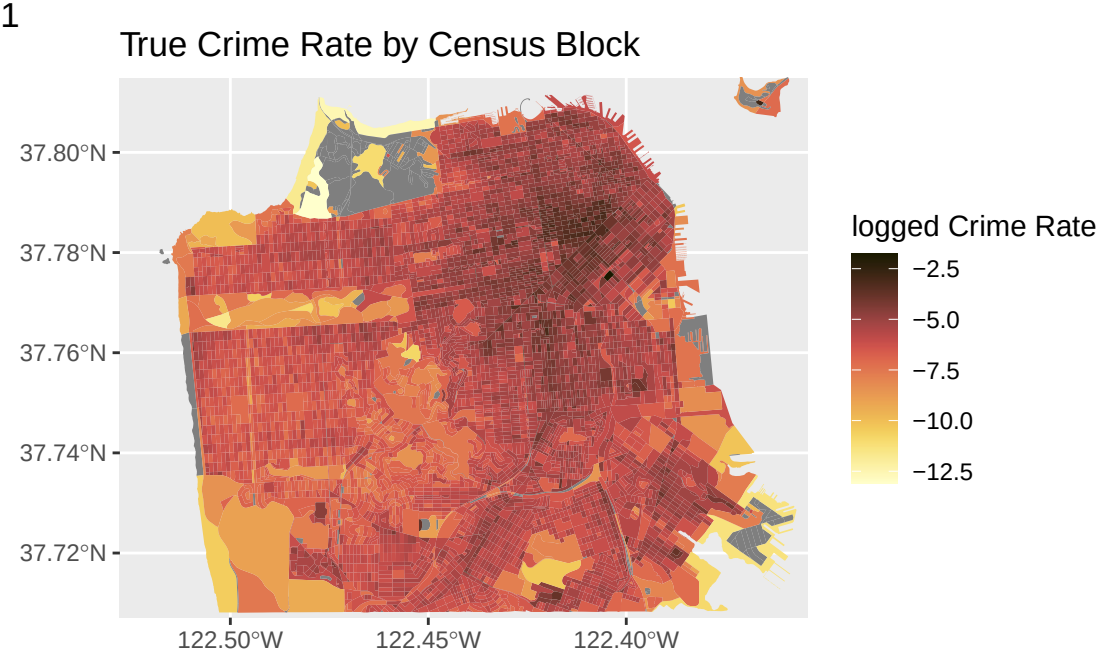
A recent trip to visit my sister in San Francisco, who happens to live on one of the cities ubiquitous steep streets, spawned a passing remark that caught my interest: "The steeper streets have less crime." Certainly the claim seems sensible: criminals, just like anyone else, don't enjoy walking up hill, and so, all else being equal, they should prefer to choose a street that is flat compared to one that is not. My decision to pursue this hypothesis ultimately lead my to the more general project of modeling crime rates in San Francisco outlined in this report. Here I see what the best model of crime rates I can build is using only a few key data points: property value, proximity to the nearest police station, proximity to the nearest bus stop, absolute elevation, and approximate street grade. These metrics are extremely easy to measure or estimate, so an accurate model built off of them could be very valuable at approximating unknown property crime rates–a variable of interest who's importance need no further explanation.

Methods

As mentioned previously, I have used several sets of data in the making of this report, the majority of which were sourced from the "DataSF" website, which holds a large collection of interesting data tables relating to the county of San Francisco. I originally sourced data on police station and bus stop locations, elevation (from amazon web services), median housing values (from the American Community Survey/Census), and property crime incident reports from 2003 to present (see figure 1 below for the map of true incidence of crime over this period). However, I was unsatisfied with the median housing data that I retrieved from the census, as it was summarized by block group rather than block ("block group" is one step larger than "block" in the census designation scheme), and a fair number of areas in San Francisco had no reported

values (likely because of a lack of residential property in those block groups) (see figure a.1 in the appendix). In order to come up with a more accurate measure of property value by block, I additionally sourced a data set from DataSF comprised of the property values of individual properties (a single apartment complex, for instance) as assessed for the purposes of property taxation. These valuations were further broken down into categories, including "land" valuation. Using this data, I was able to assign a separate average measure of both individual land value, and total (including improvements etc) property value for almost every census block in the county (see figures a.2, and a.3 in the appendix). I further engineered an approximation for street grade using elevation data. To do this, I compared the elevation of a given block to each of its neighbors. I then recorded the minimum, average, and maximum of these measures for each block. Using all of this data, I fit a gradient boosting machine (gbm) with a maximum of 500 trees, interaction depth of 12, and shrinkage factor of 0.1. Using four-fold cross validation, I measured the out-of-sample performance of each potential number of trees, and found the best model at 474 trees. I also initialized five additional models for comparison, where each has left out one of the sets of explanatory variables.

Figure 1



Results

For the full model, I found an out-of-sample RMSE of 0.0026/20. This RMSE translates to a an average error of approximately 0.05% compared to the average property crime rate per square meter per year of 0.0055/20, and an approximate increase in out-of-sample accuracy of 1,500% over a linear model initialized with the same variables (RMSE of 0.0047/20). See figure 1 below for the map of the models predictions and figure 2 below for the map of the residual error between the predictions and true values.

Figure 2

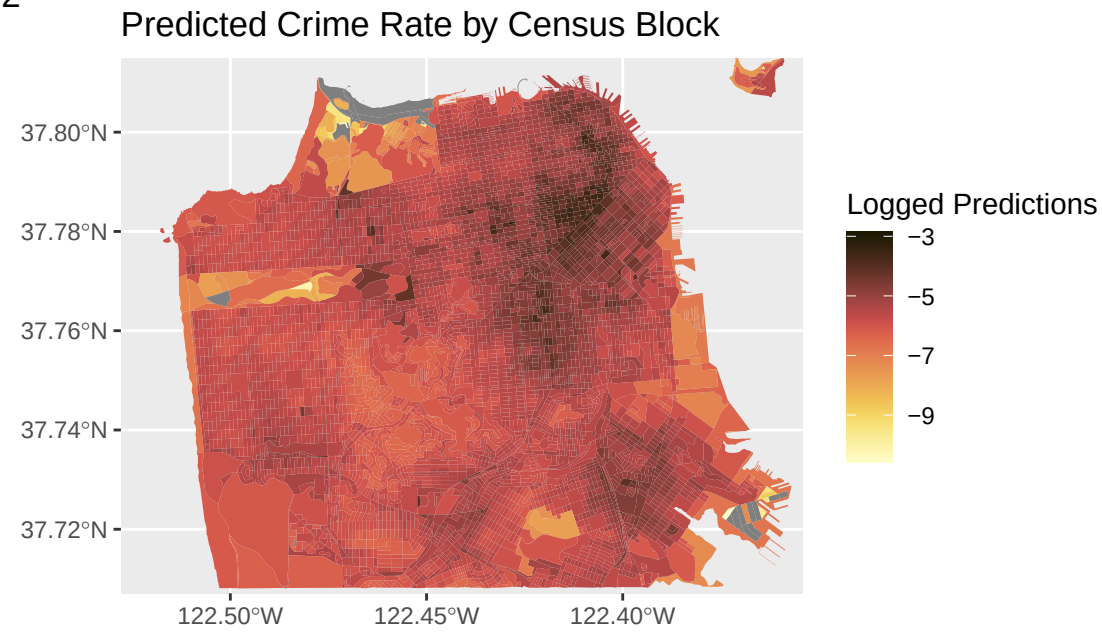
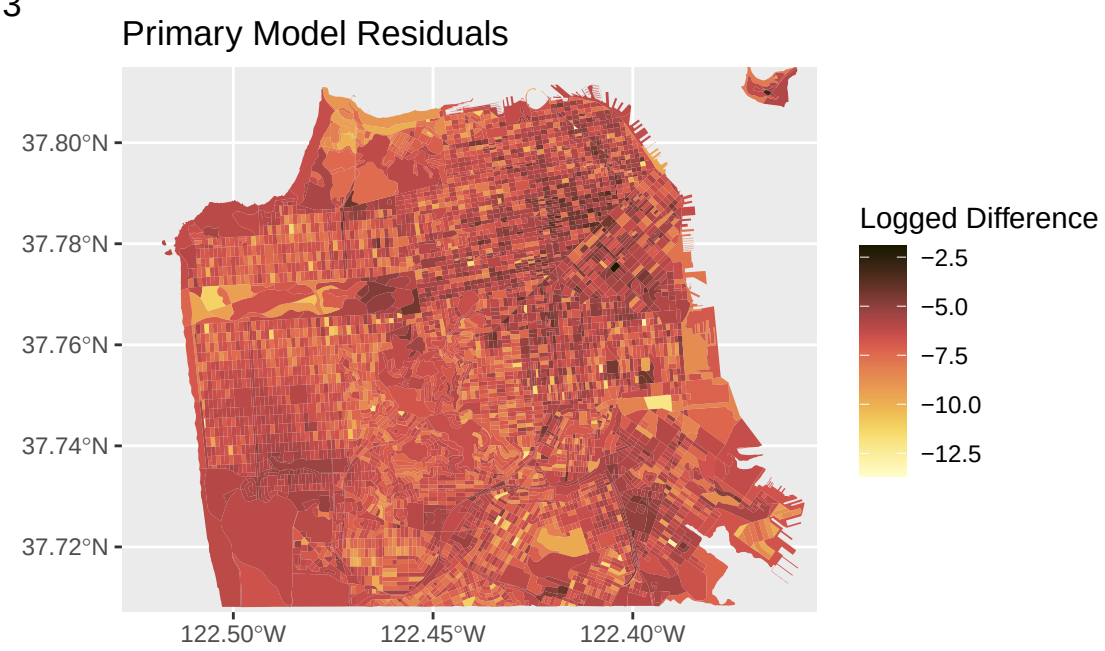

Predicted Crime Rate by Census Block
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Predictions
−3
−5
−7
−9

Figure 3



I was also interested in measuring the relative importance of each variable in the model, which you can see in table 1 below. Note that the measures of steepness and property value are separated out into multiple sub variables in this chart, so to understand the actual influence of those data sets, we have to consider the related variables as if they were one. The advantage of not having these variables combined is that we are now able to see how the different methods of measuring these categories compare. It is interesting to notice that the land values from property tax assessments ("meanLandValue") were much better than the census estimate of housing values ("estimate"), but that the total measure of property tax assessments ("meanTotalValue") were rather lackluster. The steepness measures are rather lackluster in general, but we can see that the mean appears to underperform here relatively speaking.

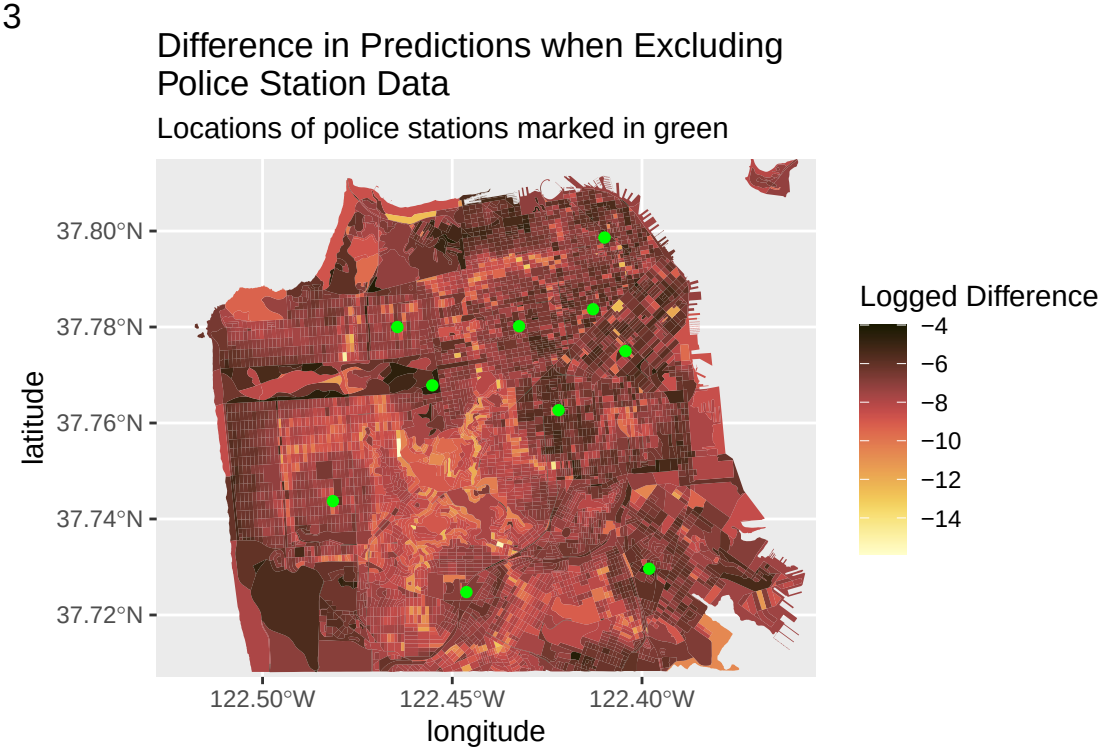
Table 1: Table 1: The relative influence of each variable in the main model

	Relative Influence
policeDist	27.791637
elevation	17.093228
meanTotalValue	12.192372
busDist	11.209223
meanLandValue	10.478382
estimate	9.907374
maxSteepness	4.559611
minSteepness	3.719318
meanSteepness	3.048855

My final visualization step was to create maps that showcase the difference in predictions between the main model and a model that leaves out one of the variable groups. Figure 3 below, the prediction differences

between the full model and a model without information on proximity to police stations, is the most striking of these. We can see very clear rings of higher differences around several of the marked police stations (the farthest west police station, for instance). I at first wondered if the strength of police station proximity in the predictions might have sprung from an abundance of filed reports, rather than a lack of crime, but the coefficient in the linear model on police station proximity is negative, so it seems that the expected mechanism is likely in evidence. The other partial maps can be seen in figures a.4, a.5, a.6 and a.7 in the appendix.

Figure 3



Conclusion

Even with such a small set of variables, I was able to create a very accurate model for predicting property crime rates in San Francisco. It appears that property value is the largest predictor (when considering all the variables for property value that were included), which is not surprising. I was disappointed to see that my measures for steepness had such low predictive power, however I believe that this is because absolute elevation ended up accounting for the effect that I was expecting to measure through approximate street grade: people don't like to walk up hill, so higher elevations see less property crime.

Appendix

Figure a.1

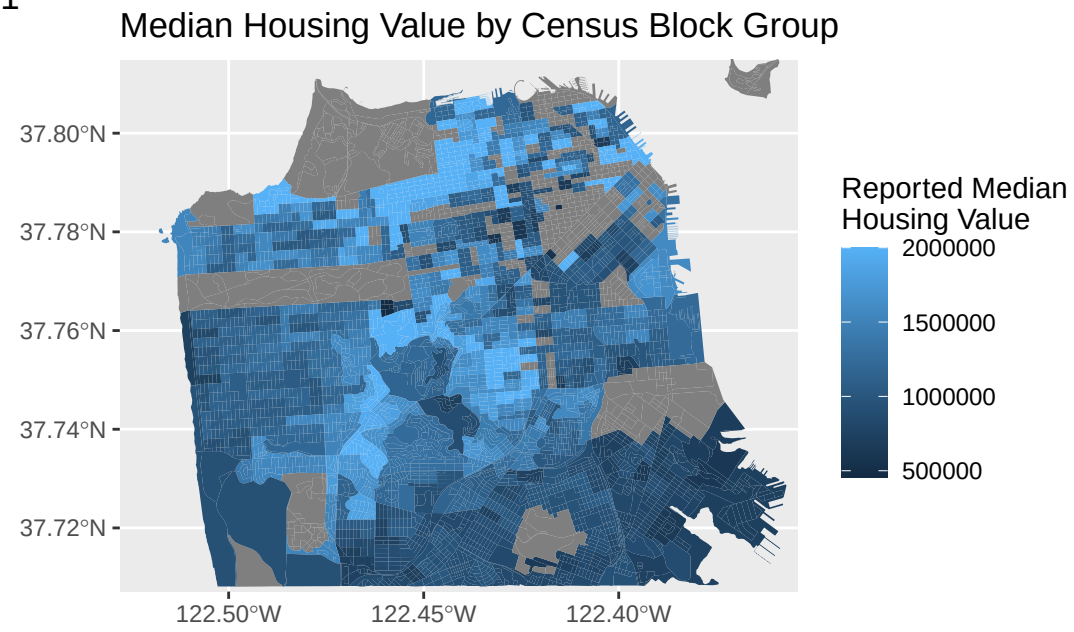


Figure a.2

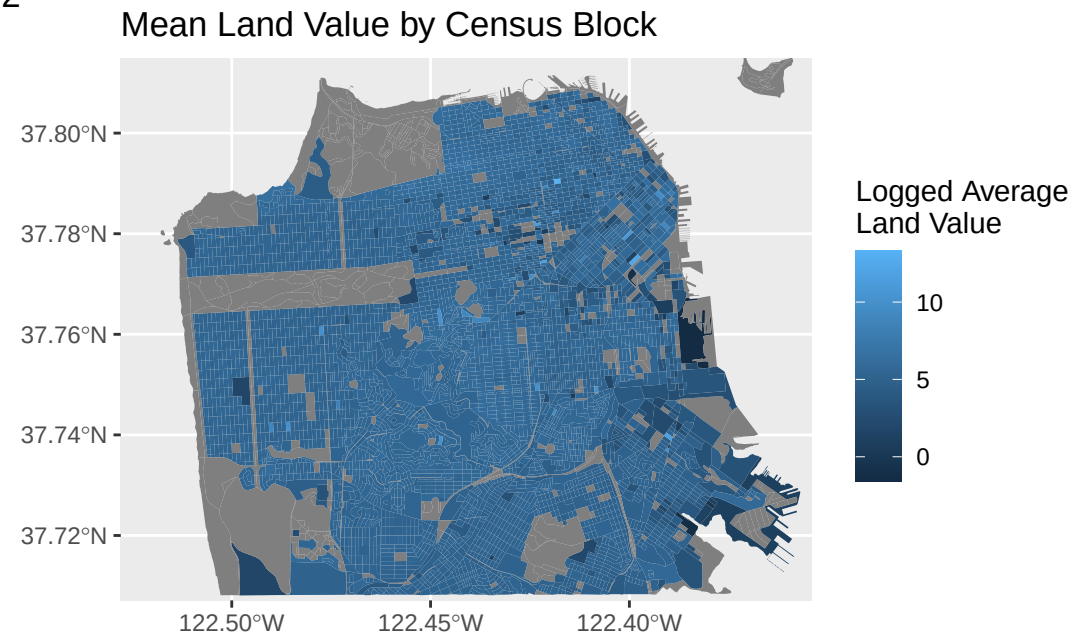


Figure a.3

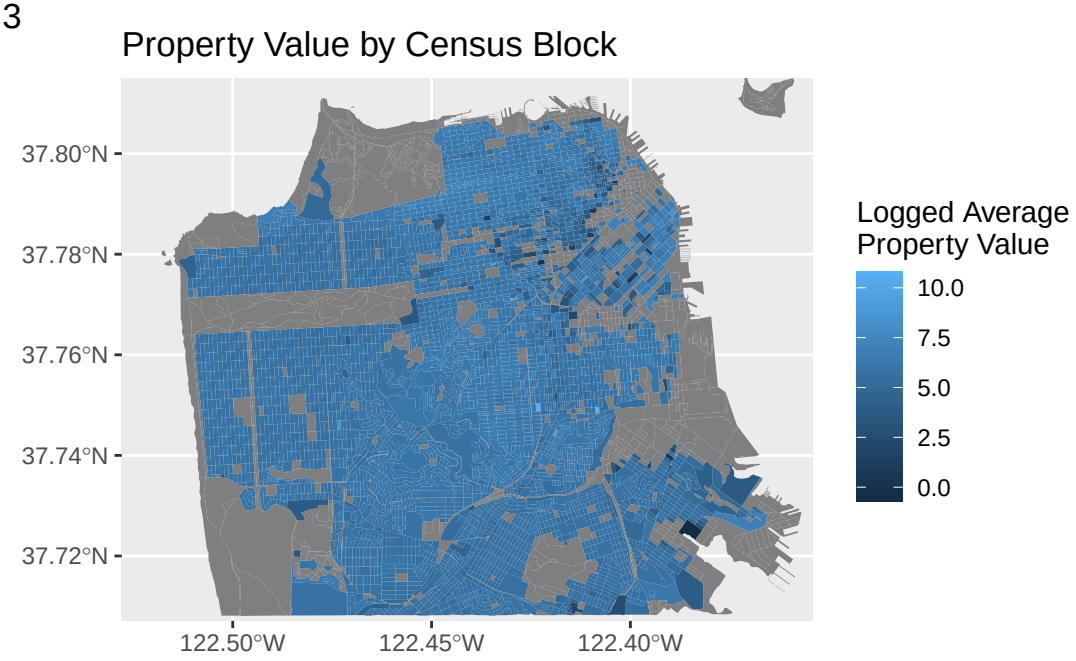

Property Value by Census Block
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Average Property Value
10.0
7.5
5.0
2.5
0.0

Figure a.4

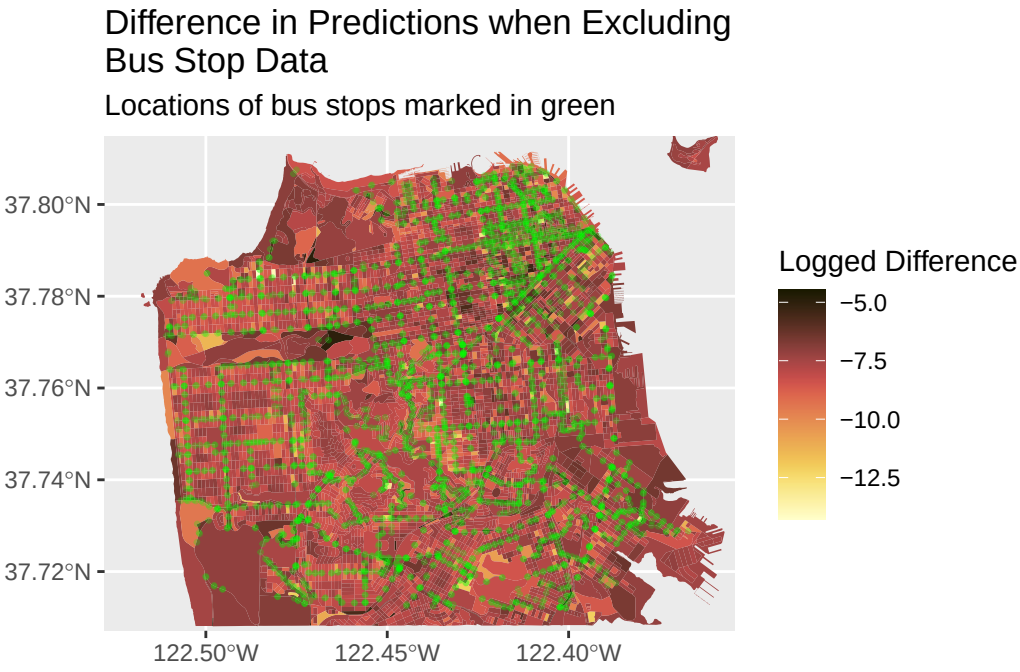

Difference in Predictions when Excluding Bus Stop Data
Locations of bus stops marked in green
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Difference
−5.0
−7.5
−10.0
−12.5

Figure a.5

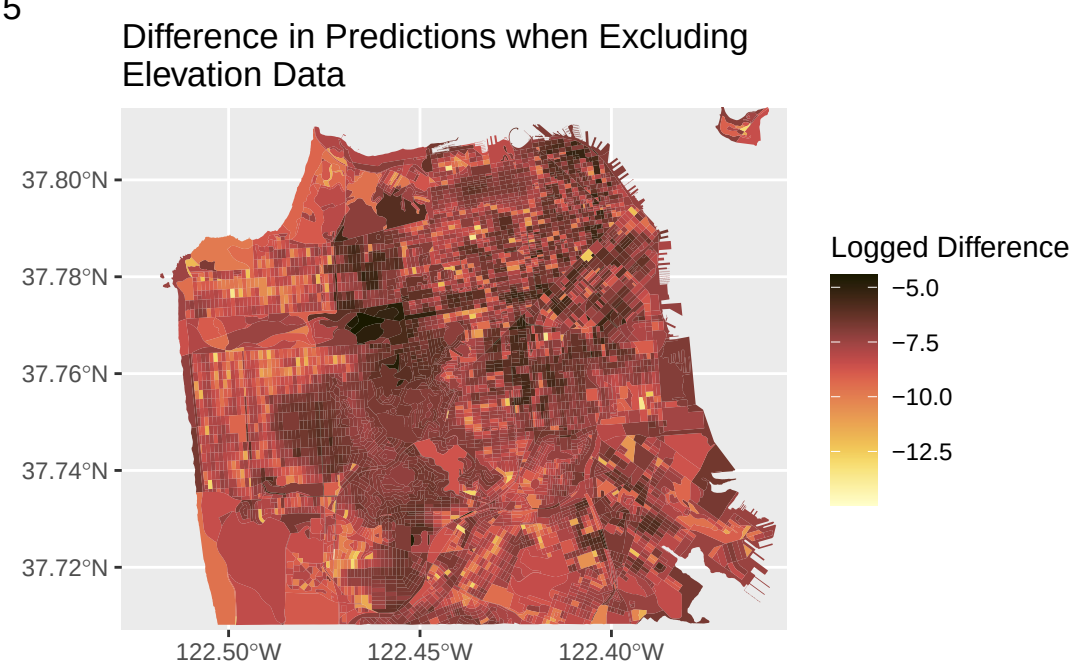

Difference in Predictions when Excluding Elevation Data
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Difference
−5.0
−7.5
−10.0
−12.5

Figure a.6

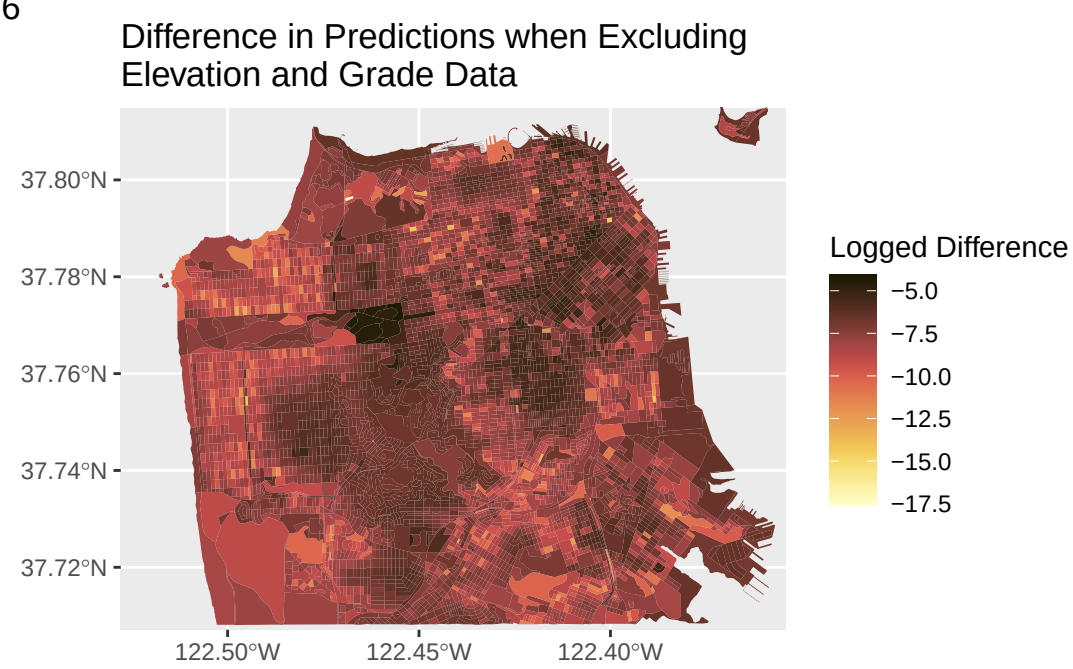
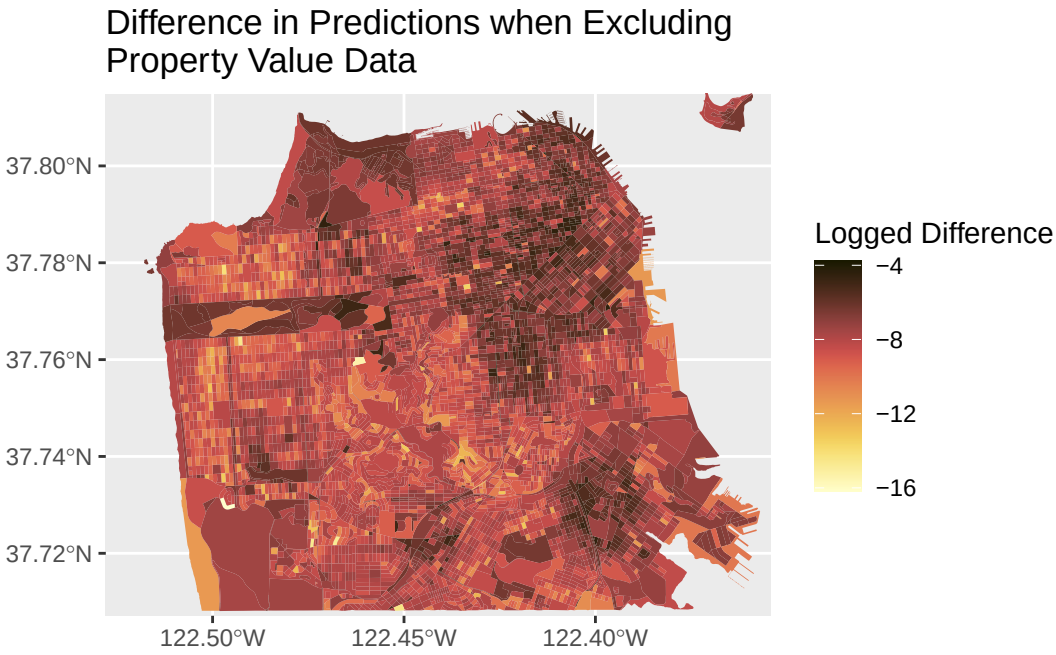

Difference in Predictions when Excluding Elevation and Grade Data
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Difference
−5.0
−7.5
−10.0
−12.5
−15.0
−17.5

Figure a.7


Difference in Predictions when Excluding Property Value Data
37.80°N
37.78°N
37.76°N
37.74°N
37.72°N
122.50°W
122.45°W
122.40°W
Logged Difference
−4
−8
−12
−16