Project Report

Exploring Spatial Data on Crime Analysis

Matheus Paes de Souza

mpaes.souza292@gmail.com

Supervision: Jorge Poco

Escola de Matemática Aplicada

December 22, 2021

Contents

Abstract

This research project aims to analyse the spatial relation between the distribution of crime and the presence of amenities in the city of São Paulo. To that aim, we employ a spatial-aware regression model, Geographically Weighted Regression (GWR). This model takes into account the spatial distribution of the input data, and describes the manner in which the importance of features for the prediction of a variable varies in space.

1 Datasets

We used two datasets for this task, describing crime occurrences and amenities throughout São Paulo. Both datasets and a processed one relation crime with amenities are available at google drive folder as SPdataEstabelecimentos&Crime.zip.

1.1 Crime occurrences dataset

This dataset is a list of crime occurrences that were reported in São Paulo from 2006 through 2017. The occurrences were sourced from official Police Reports. Each entry lists the date and time of the occurrence, whether it was against passersby, vehicles or stores, and the geographical coordinates of the occurrence. For this work, only the occurrences reported in 2017 were considered.

1.2 Amenities dataset

This dataset provides information for amenities located throughout São Paulo. Each entry indicates the amenity's name, category, and geographical coordinates. The data distinguishes between 108 categories (explained in the codebook available at google drive folder as SPdataEstabelecimentos&Crime.zip). This dataset was sourced from information from Google Maps.

1.3 Discretization and aggregation of the datasets

Both datasets indicate individual geographical coordinates for each point of data. In order to identify spatial patterns in the crime distribution and amenities in São Paulo, we need to discretize the city in small, preferably nearly identical regions. In order to achieve this discretization, we used Uber's H3 hexagonal spatial discretization system[1]. H3 provides us with small, nearly identical hexagonal regions of a controllable size covering the entire city's area. The previous datasets were then aggregated in these regions. H3 provides a parameter to control the resolution of the discretization, i.e. the size of the regions. We

[1] `https://h3geo.org/`

chose to work with resolution levels 8 and 9, as lower resolutions were not fine enough and higher resolutions were not computationally efficient.

The resulting dataset for each resolution was a list of hexagonal regions covering São Paulo, in which each entry indicates the number of crimes reported inside the region in 2017 for each type of crime, and the number of amenities present inside the region, for each type of amenity. We also created a datasets aggregating only the downtown area.

2 Methodology

The spatial analysis was performed through the training of a prediction model for the number of crime occurrences in each region. The code of this methodology is available at google drive folder[2] at SpatialCrime.zip.

2.1 The model

The model used for predicting the number of crimes was Geographically Weighted Regression (GWR) [1]. GWR takes into account the spatial structure of the data, when divided into discrete regions. In contrast to simpler prediction models such as Linear Regression, this class of models encloses a prediction model for each region. In GWR, each local model takes into account not only the features of it's local region, but also the features of the surrounding regions. The contribution of each region to a local model depends on it's distance to the local region, and is weighted by a kernel function. The kernel function has a bandwidth parameter for controlling the radius of influence and weight decay for surrounding regions. In a related work, Silva et al. [2] used Geographically Weighted Regression to model homicide rates in the state of Pernambuco, Brazil.

In this work, we used the Gamma kernel for GWR, with bandwidth parameters varying from 900 to 6000, depending on the resolution level of the discretization.

We used the implementation of Geographically Weighted Regression provided by the Python package mgwr [3].

2.2 Data transformations

Some further transformations were applied to the dataset indicated in Section 1.3.

[2] `https://drive.google.com/drive/u/0/folders/1bjx4DdHJw2OzilAp25yvv-78kcaL3_yR`

2.2.1 Treatment of outlier values in crime levels

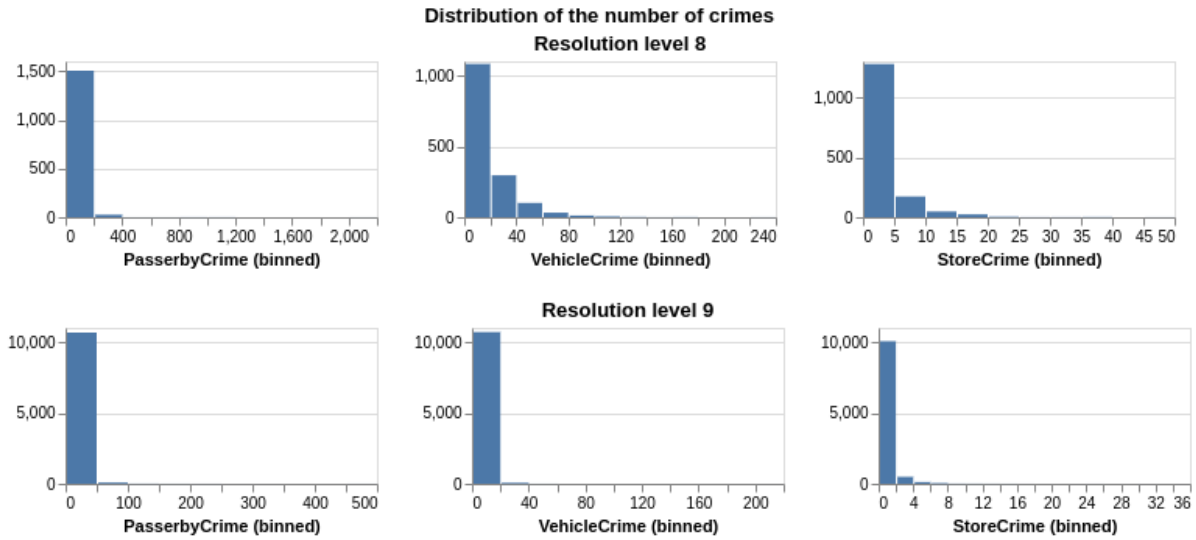

Figure 1: Distribution of the number of crimes

Figure 1 shows the distribution of the number of crimes across the regions. It can be noted that small values are very frequent, with this frequency decreasing as the values rise. Still, there are a few regions which have extremely high values. The presence of these outliers cannot be ignored, since they are hostpots. However, their presence in the data have a degrading effect on the performance of the prediction model. We explore two solutions to this problem, both involving the application of a monotonic transformation to the data.

The first solution is simply to sum 1 and apply the natural logarithm to the data values. This transformation is continuous and it's power increases exponentially as the values increase, bringing about the desired effect.

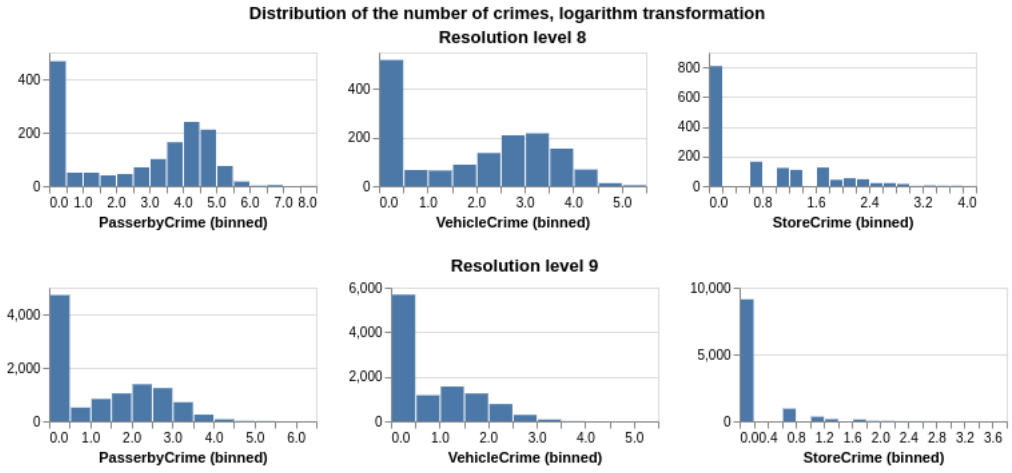

Figure 2: Distribution of the number of crimes, logarithm transformation

The second method is to apply the inverse quantile function of the data distribution. This will replace each data point by it's quantile, producing values between 0 and 1. As this method utilizes a transformation that depends on the data, we calculated the inverse quantile function using only the training data in order to avoid data leakage. Then, this same function was used to transform the test data.

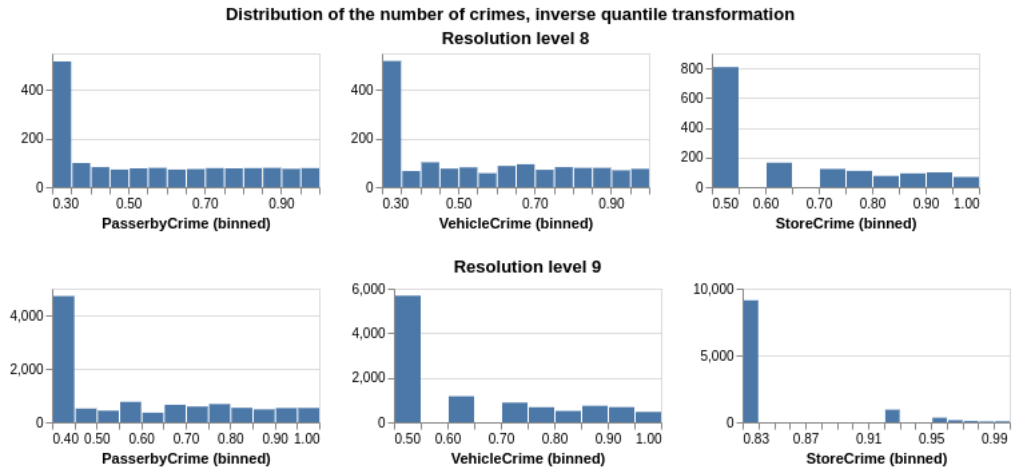

Figure 3: Distribution of the number of crimes, inverse quantile transformation

2.2.2 Treatment of multicollinearity on the input data

As the input data has many variables, a possible problem is the presence of multicollinearity in the data. Furthermore, the high number of variables also makes the possibility of overfitting more likely. To mitigate this, we treated the input data by removing variables according to correlation measures. We performed hierarchical clustering of the variables using the Spearman correlation coefficients and Ward's linkage criterion [4]. This method requires a parameter (threshold) for the generation of the clusters.

3 Evaluation

We now describe the evaluation process, with the choosing of the transformations, multicollinearity treatment threshold parameter and kernel bandwidth.

3.1 Resolution level 8

We trained and evaluated predictors for passerby crimes, for both logarithm and inverse quantile transformation, and for several threshold and bandwidth parameters. The full list of parameters can be found in in the codebook available at google drive folder in SP-dataEstabelecimentos&Crime.zip). We then calculated the equivalent of the R2 measure for the test data. The result of the experiments is shown in the figures below. Figure 4

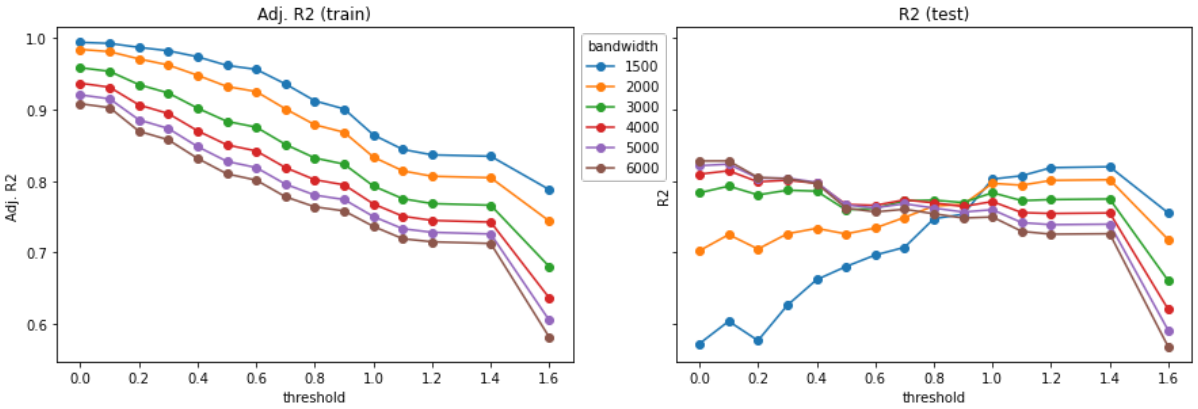

Figure 4: Results for inverse quantile transformation.

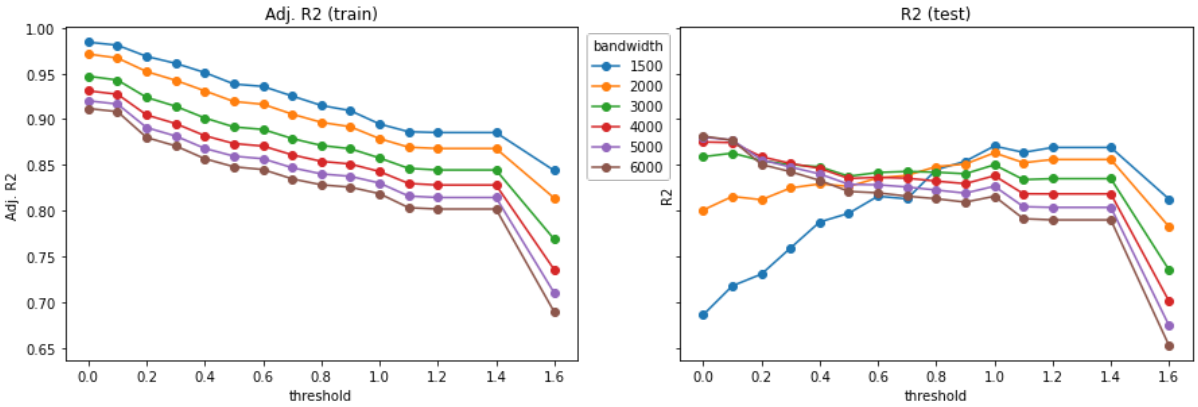

Figure 5: Results for logarithm transformation.

shows the results for the inverse quantile transformation, and Figure 5 the results for the logarithm transformation.

The best result for the inverse quantile transformation was 0.83, while the logarithm transformation 0.88. The best results for both transformations were achieved using threshold 0 (equivalent to no multicollinearity treatment) and bandwidth 6000.

The remainder of the experiments were performed using the logarithm transformation.

Figure 6 shows the results obtained for predicting the number of crimes against vehicles. The best score was 0.83 with a threshold of 1.2 and bandwidth 1500.

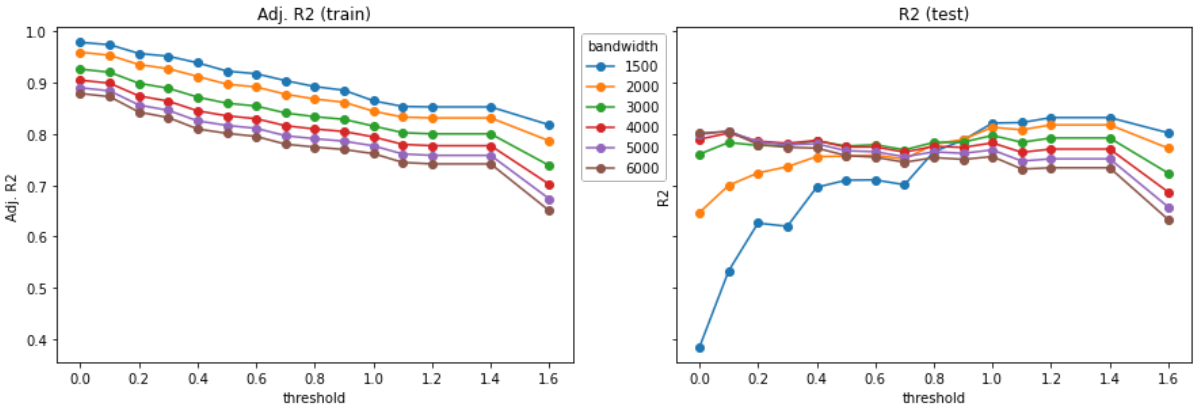

Figure 6: Results using R2 metrics for crimes against vehicles.

Figure 7 shows the results obtained for predicting the number of crimes against stores. We achieved a best score of 0.64 with threshold of 0.1 and bandwidth 6000.

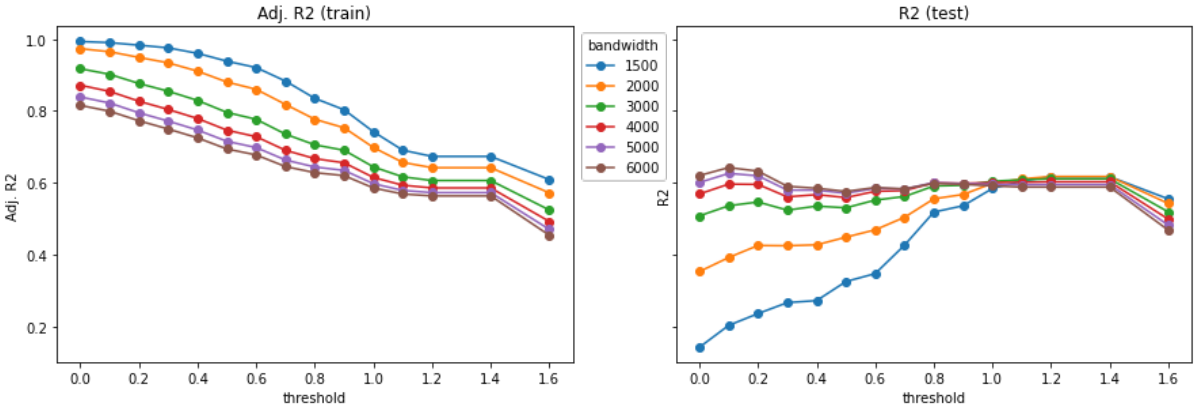

Figure 7: Results using R2 metrics for crimes against stores.

3.2 Resolution level 9

We performed similar experiments with the resolution level 9. The full list of parameters can be found in in the codebook available at google drive folder in SPdataEstabelecimentos&Crime.zip).

Figure 8 shows the results obtained for predicting the number of crimes against passersby. The best result was a score of 0.76, for a threshold of 0 and bandwidth 2700.

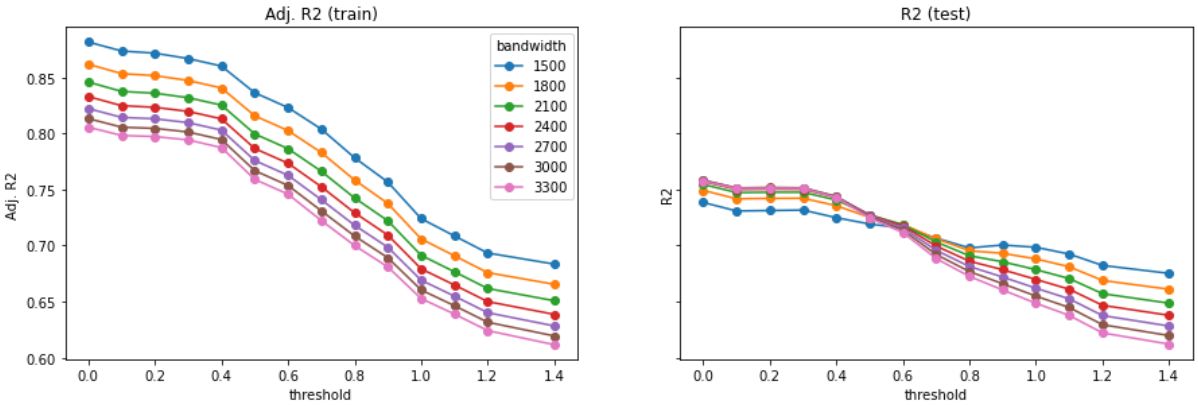

Figure 8: Results using R2 metrics for crimes against passersby.

Figure 9 shows the results obtained for predicting the number of crimes against vehicles. The best result was a score of 0.57, for a threshold of 0 and bandwidth 2700.

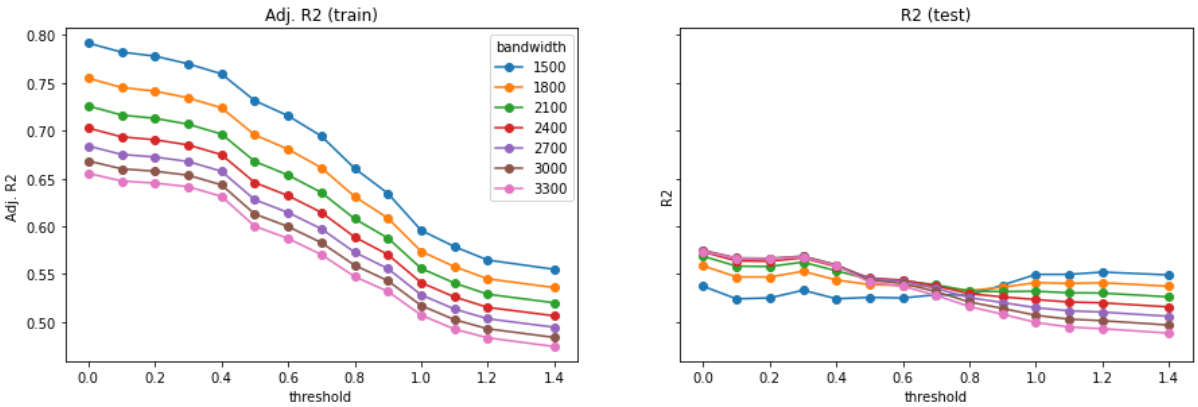

Figure 9: Results using R2 metrics for crimes against vehicles.

Finally, figure 10 shows the results obtained for predicting the number of crimes against stores. The best result was a score of 0.29, for a threshold of 0.5 and bandwidth 3300.

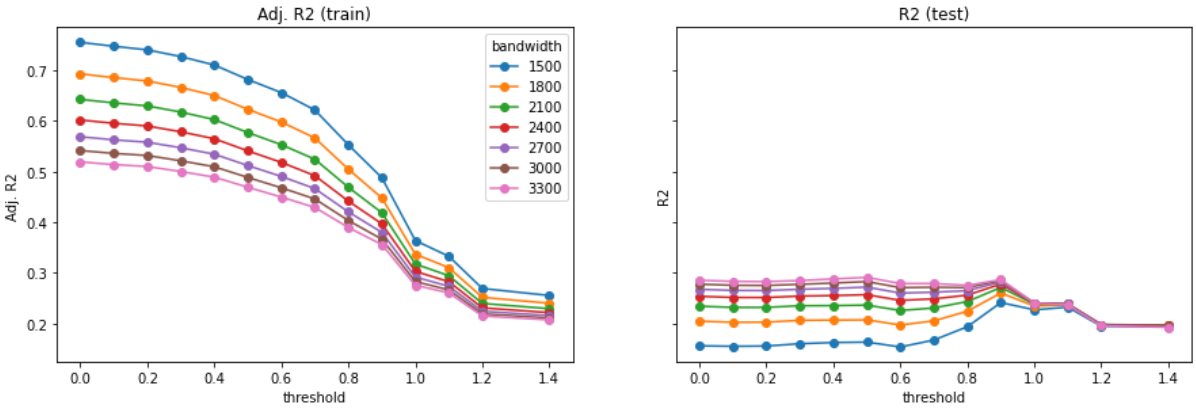

Figure 10: Results using R2 metrics for crimes against stores.

4 Discussion

The experiments showed that the logarithm transformation for the number of crimes leads to the best results in the regression. The best results were also generally observed with little to no multicollinearity treatment. Finally, we observed that the discretization with resolution level 8 lead to better results than with resolution level 9. We summarize the best result for each regression variable in Table 1:

	Resolution level 8			Resolution level 9		
Crimes against	R2	threshold	bw	R2	threshold	bw
Passersby	0.88	0.0	6000	0.76	0.0	2700
Vehicles	0.83	1.2	1500	0.57	0.0	2700
Stores	0.64	0.1	6000	0.29	0.5	3300

Table 1: Summary of results using R2 metrics

One of the simplest models available to predict the number of crimes in a given region is the Linear Regression model. While the goal of the experiment is to identify spatial patterns in the crime distribution using Geographically Weighted Regression, the predictive power of the models are also important. Thus, we provide a comparison of the performance of both models for predicting the number of crimes against passersby in resolution level 8.

Both models achieved good results, though the Geographically Weighted Regression model had slightly better performance. The latter achieved a score of 0.88 for the measure of the equivalent of the R2 coefficient for the test dataset, while the Linear Regression achieved around 0.83.

4.1 Case analysis

We now present an analysis of the regression for the number of crimes against passersby in resolution level 8, in two cases:

4.1.1 Case 1 - Hotspot region analysis

We analyse the results obtained by utilizing a Geographically Weighted Model to predict the number of passerby crimes in the four adjacent regions with the highest number of recorded incidents (5047 incidents), located downtown. The figure below shows the 10 features identified to have the biggest importance for the prediction on each of the regions:

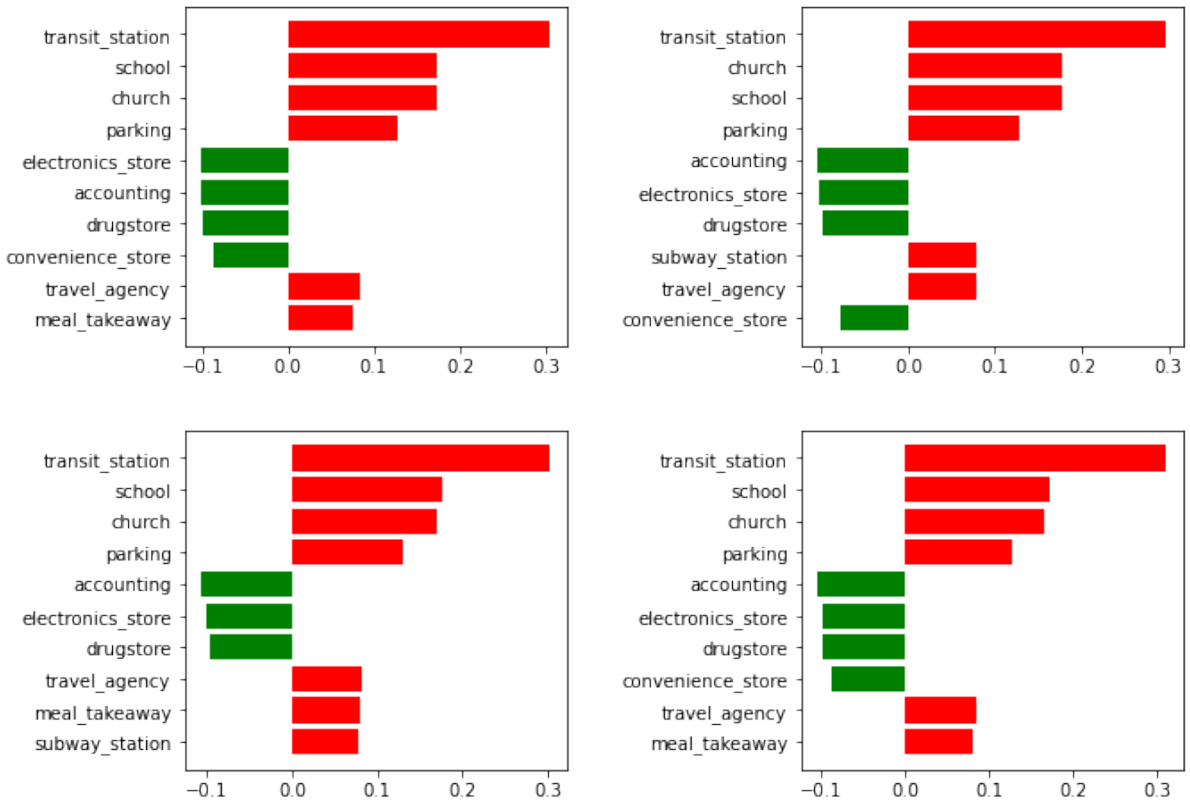

Figure 11: Positive (in red) and negative (in green) importance of the variables to occurrence of crimes.

It can be noted that the feature `transit_station` is the most important predictor for all four regions, with an increasing effect in the crime level predictions. In fact, this feature was found to be frequently the most important predictor. Furthermore, the feature `subway_station` also has high importance and increasing effect in the predictions for two of these four regions. This could be interpreted as bus stops and subway stations being possible hotspots for crimes against passersby (i.e., muggings).

The importance for the other features are similar across the regions. Schools, churches, parking structures, travel agencies and takeaway restaurants seem to have a positive cor-

relation with the number of crimes reported in the area, while the presence of accounting offices, electronics stores, drugstores and convenience stores were found to have the opposite effect. These are perhaps not immediately interpretable, and can serve as a starting point for investigation or the refining of the model.

4.1.2 Case 2 - Low crime region analysis

In contrast, we now discuss the results of the same regression for four additional adjacent regions with much lower crime rates, recording only 39 cases. The figure below shows the calculated importance for the main features:

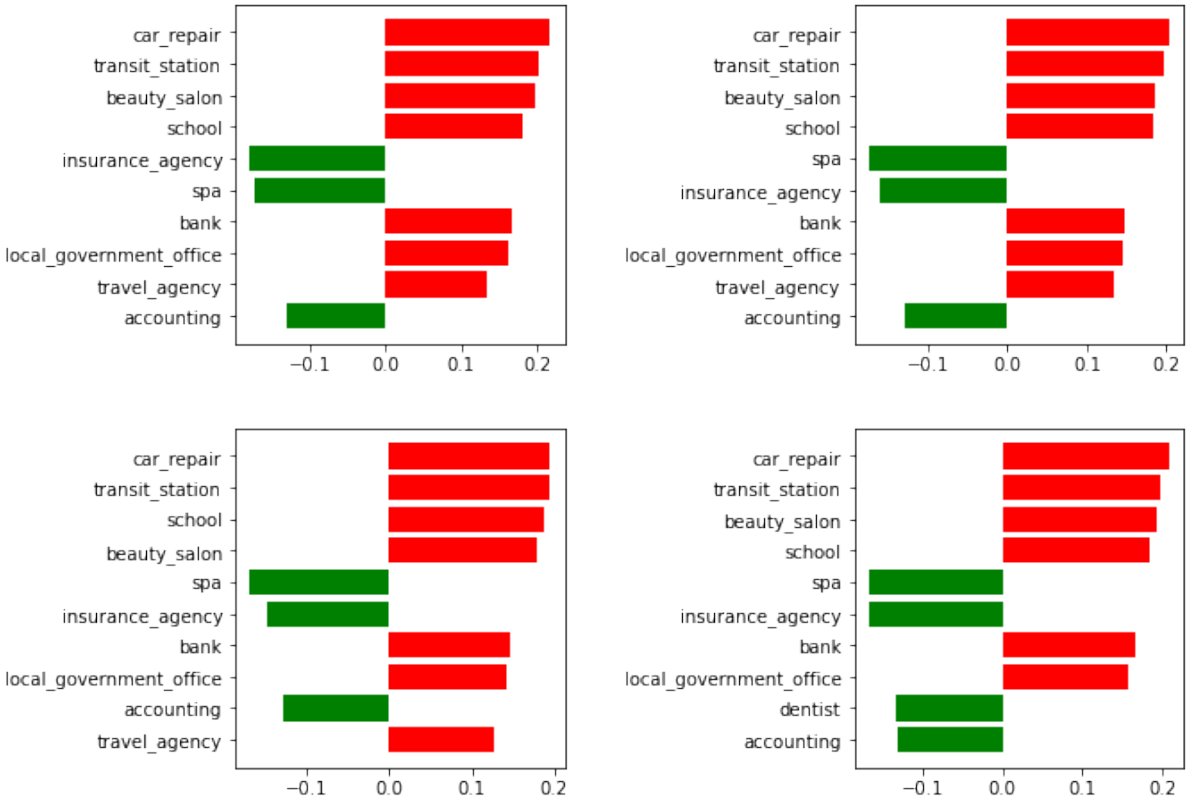

Figure 12: Positive (in red) and negative (in green) importance of the variables to occurrence of crimes.

In this part of the city, the feature with the most influence in the increase of the prediction is the presence of car repair shops, though closely followed by the already known `transit_station` feature. Again, we have a certain commonality in all of the predictors for these regions. Beauty salons, schools, banks, offices for local government and travel agencies have a positive correlation with the increase in crime there. In this case, the increase in muggings near banks is very easily explained. Meanwhile, the presence of spas, insurance agencies, accounting and dentists offices were found to have negative effect on the prediction, though this behaviour also lacks a simple explanation.

4.2 Visualization

In Figures 13 and 14, we show a visualization of the predicted values for passerby crimes in the whole city and limited to downtown. The values have been scaled for the training. We can observe that the predicted values agree with the actual data.

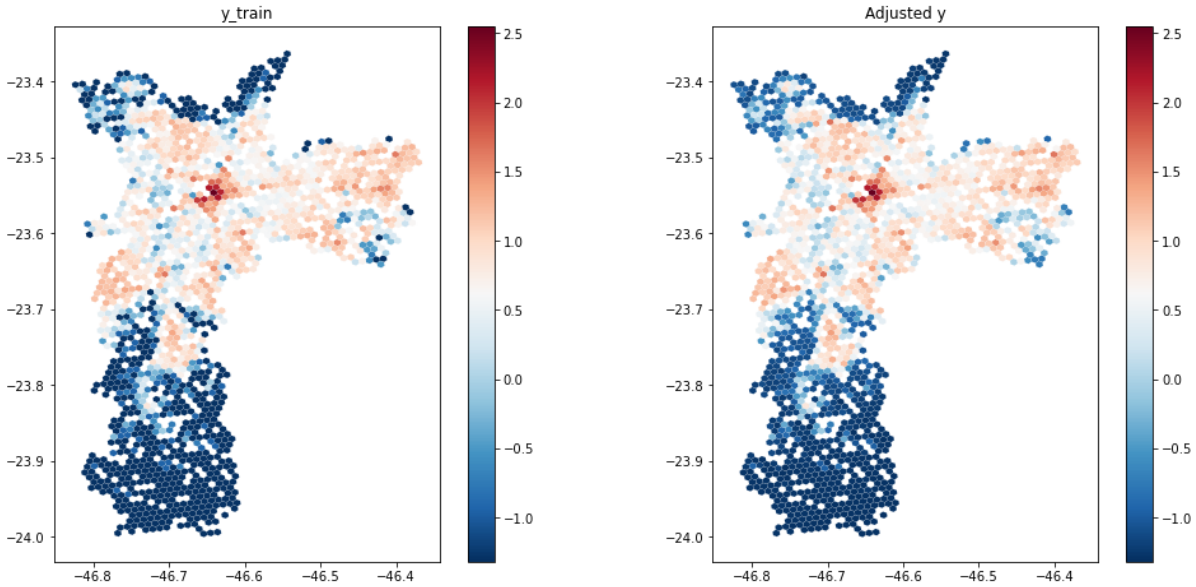

Figure 13: Heatmap of number of crimes in the whole city of São Paulo with threshold=0.0, bw=6000.

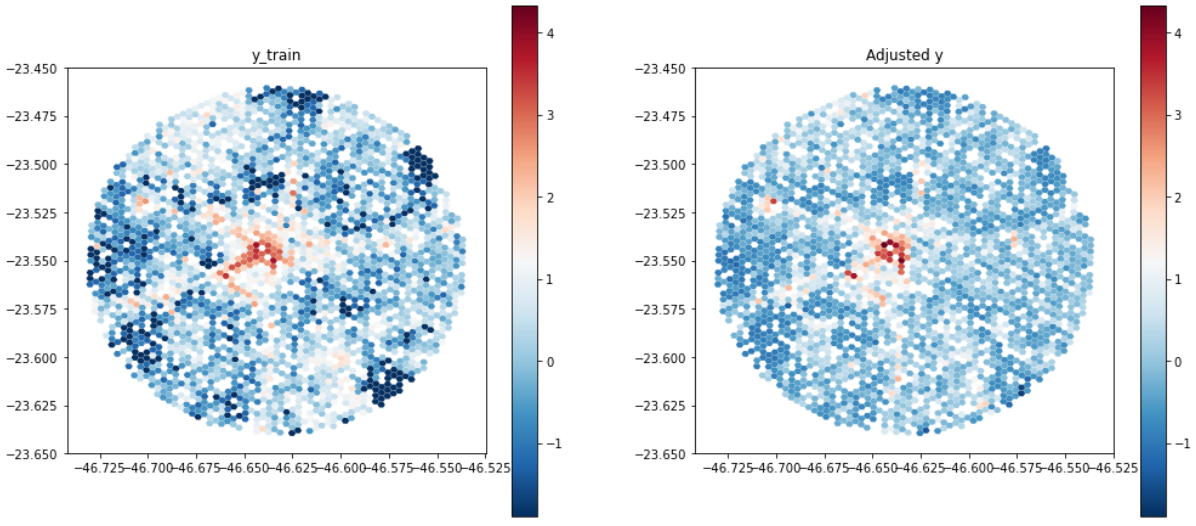

Figure 14: Heatmap of number of crimes in the whole city of São Paulo with threshold=0.0, bw=3300.

We can observe that the predicted values (in the right) have a good resemblance with the original values (in the left) for the whole city (Figure 13) preserving the scale, thus making a good prediction. Now, for the data focused in the downtown (Figure 14) despite the scale not being preserved, thus not making a really good prediction the patterns where the data highlight criminal activities is preserved.

5 Conclusion

In this report we studied the impact of models that only takes account of regions near to the the observed one. Our dataset investigates the impact of certain amenities on crime. Since we create a model for each region, we can observe what are the most important variables for each region and then observe which amenity has a deeper impact on each part of the city. Our experiments show a small increase in performance using Geographically Weighted Regression, as another gain using that model, we could observe that the presence of amenities have different impacts on each region.

References

[1] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 3, pp. 431–443, 1998.

[2] C. Silva, S. Melo, A. Santos, P. A. Junior, S. Sato, K. Santiago, and L. Sá, "Spatial modeling for homicide rates estimation in pernambuco state-brazil," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 740, 2020.

[3] T. M. Oshan, Z. Li, W. Kang, L. J. Wolf, and A. S. Fotheringham, "mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 269, 2019.

[4] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.