Crime and the city:
Spatial planning for crime reduction

Ruben Sharpe

March 21, 2021

Abstract

It is hypothesised that crime in a neighbourhood is correlated, albeit in a complex manner with the availability of public venues. Rather than hypothesising about causal relations, a machine learning approach is proposed in which high-crime, medium-crime and low-crime neighbourhoods are characterised by their public venues. These characterisations may serve as templates for good and bad practices in spatial planning. As a test case, the city of Tornonto (Ontario, Canada) was characterised based on information about public venues, obtained from Foursquare, and the City government of Toronto Open Dataset on crime.

Neighbourhoods were clustered by similarity of their composition in terms of venues using K-means clustering. Spatial information on crime was visualised in choropleth maps of Toronto. Visual comparison of neighbourhood clusters with crime maps shows the feasibility of the approach.

Decision trees and support vector machines were trained to classify neighbourhoods as high-, medium- or low-crime. With such models, city planners may try to predict how modifications would impact the crime classification or,in other words, which modifications correlate with what crime level.

Contents

1 Introduction

It is intuitively clear that public venues, such as bars and restaurants, but also public service locations such as transport hubs or banks impact the wellbeing of local communities. However, it is not straightforward to predict whether this impact will be positive or negative.Bars and restaurants can enliven a neighbourhood and may contribute to greater safety because of increased presence of people in the streets, because potential witnesses deter crime. On the other hand, the presence of bars and restaurants is also correlated with alcohol abuse related crime. Similarly, there is a duality in the social impact for the presence of shops and banks. These may have a positive impact because of associated prosperity, but may also attract criminal activity at night because of reduced traffic compared to residential areas.

The complexity and the interrelatedness of these issues carries over to the complexity of city planning. The difficulty for city planners, moreover, is that they rarely get to design a neighbourhood from scratch. Typically, they have to deal with neighbourhoods with for the most part a fixed set of public facilities, and which may be suboptimal from a design point of view. The question for a city planner, therefore, is often : "What would be, *for a certain neighbourhood and given all its other attributes*, the impact of building (or removing) a shopping mall? Or a park? Or adding, or removing a bus stop?"

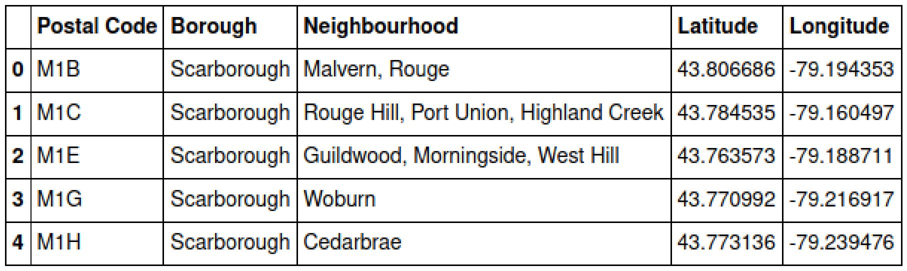
There is, of course, a lot of scientific literature on this topic but that is out of the scope of this project. This report is aimed at city planners and is intended to show the feasibility of machine learning as a tool for spatial planning.
Machine learning allowes an automated approach to compare neighbourhoods both in terms of their public venues and in terms of social markers, such as crime rates. This provides city planners, apart from their professional expertise, with examples of good and bad practices that may help them in their decision making. For reasons of convenience, the analysis in this report will look at the city of Toronto (Ontario, Canada). Crime rates will be used as the proxy for the wellbeing of communities, although other markers such as employment rate or health statistics could also be used.

2 Data

Location data, information about public venues and crime statistics will be used to classify neighbourhoods and map them with respect to the prevalance of crime. From this, generalised conclusions may be drawn with respect to the makeup of ‘high-crime’ and ‘low-crime’ neighbourhoods.

2.1 Location data

Location data for the city of Toronto can be acquired by scraping postal codes, boroughs and neighbourhoods from Wikipedia.[1] This data needs to be enriched with the GPS coordinates of the respective neighbourhoods. These coordinates can be obtained from the internet in the form of a ‘.csv’ file from the Coursera website for the Data Science Capstone.[2] The enriched dataset consists of the postal codes, borough, and neighbourhood names and the neighbourhoods’ latitude and longitude (Figure 1).



	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 1: Example of data enriched with coordinates.

2.2 Public venues

Foursquare is used to obtain information about the public venues, specifically their category, in each neighbourhood (Figure 2). For this purpose, the neighbourhood is arbitrarily defined as the area within a 1.500 m radius around a postal code’s geographical centre. The top 10 most prevalent venues will be used to characterise the neighbourhood (Figure 3).

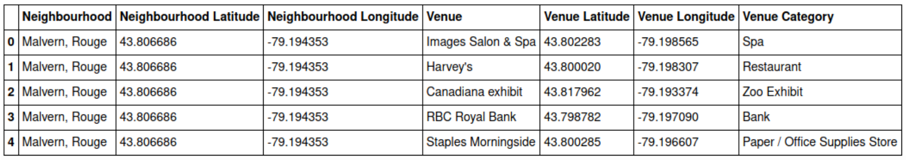
2.3 Crime data

The city of Toronto provides a lot of open data related to the city.[3] This data ranges from polls conducted by the government, to inventory lists of street furniture, to ‘bicycle count and locations’ and is ever increasing. From this

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

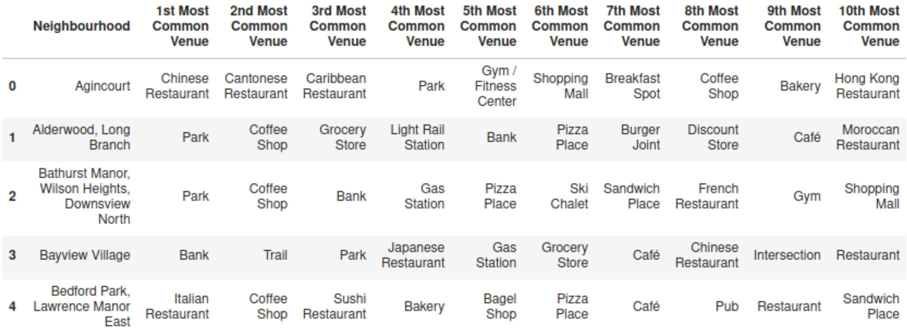
[2] https://cocl.us/Geospatial_data

[3] https://www.toronto.ca/city-government/data-research-maps/open-data/



	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Malvern, Rouge	43.806686	-79.194353	Harvey's	43.800020	-79.198307	Restaurant
2	Malvern, Rouge	43.806686	-79.194353	Canadiana exhibit	43.817962	-79.193374	Zoo Exhibit
3	Malvern, Rouge	43.806686	-79.194353	RBC Royal Bank	43.798782	-79.197090	Bank
4	Malvern, Rouge	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store

Figure 2: Example of which venue information can be obtained.



	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Chinese Restaurant	Cantonese Restaurant	Caribbean Restaurant	Park	Gym / Fitness Center	Shopping Mall	Breakfast Spot	Coffee Shop	Bakery	Hong Kong Restaurant
1	Alderwood, Long Branch	Park	Coffee Shop	Grocery Store	Light Rail Station	Bank	Pizza Place	Burger Joint	Discount Store	Café	Moroccan Restaurant
2	Bathurst Manor, Wilson Heights, Downsview North	Park	Coffee Shop	Bank	Gas Station	Pizza Place	Ski Chalet	Sandwich Place	French Restaurant	Gym	Shopping Mall
3	Bayview Village	Bank	Trail	Park	Japanese Restaurant	Gas Station	Grocery Store	Café	Chinese Restaurant	Intersection	Restaurant
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Coffee Shop	Sushi Restaurant	Bakery	Bagel Shop	Pizza Place	Café	Pub	Restaurant	Sandwich Place

Figure 3: Example of neighbourhoods with top 10 most prevalent venues.

repository, a list of crime rates can be obtained.[4] This list contains, for each neighbourhood the number of incidences of certain types of crime (Table 1). Population information from the 2016 census, included in the file, can be used to normalise crime rates relative to population size. Conveniently, the file also contains geographical information so that this information can be easily mapped. The crime categories that are accessible from the data files are given in Table 1.

Assault	from 2014 – 2019
Auto Theft	from 2014 – 2019
Breaking and Entering	from 2014 – 2019
Homicide	from 2014 – 2019
Robbery	from 2014 – 2019
Theft	from 2014 – 2019

Table 1: Crime data categories.

[4] https://open.toronto.ca/dataset/neighbourhood-crime-rates/

3 Methodology

3.1 Exploration of crime information

Crime information is available for six categories. This information may be visualised in choropleth maps (Figure 4). For direct comparison with the neighbourhood information, the neighbourhood names in both datasets must be correlated. However, upon joining the dataset for the geographical data with that for the crime data, it is found that the names are so dissimilar that only 20 neighbourhoods are found to match up (Figure 5).[5] Fortunately, the Toronto city government provides a 'geojson' file for the crime data so that neighbourhood information and crime information may be mapped, independently of each other. This allows their visual inspection and subsequent manual labeling of neighbourhoods in the venue dataset with respect to 'crime proneness' (Figure 6).

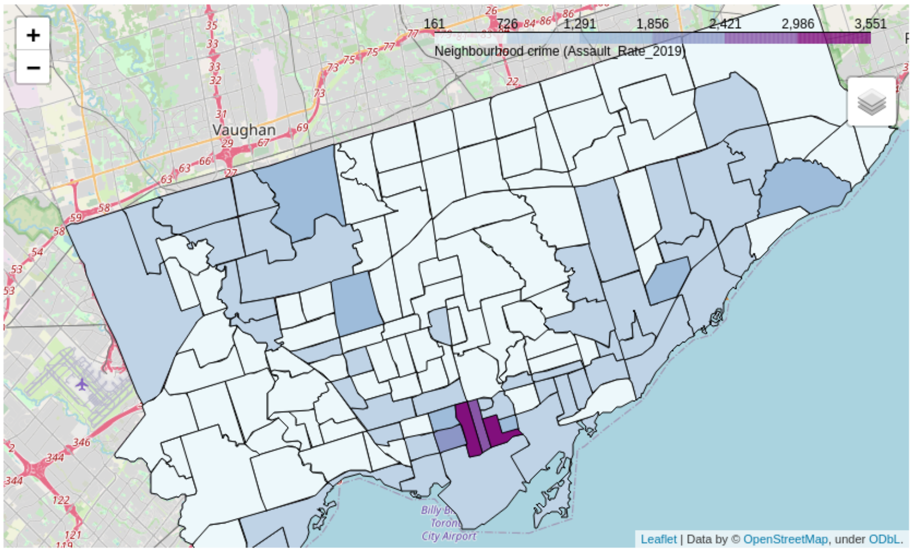
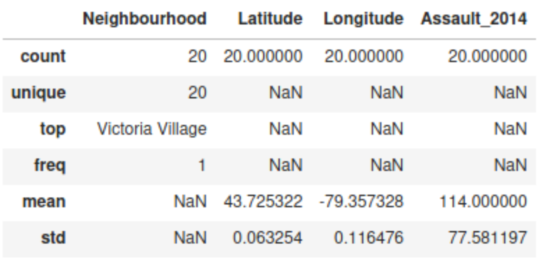

161
726
1,291
1,856
2,421
2,986
3,551
Neighbourhood crime (Assault_Rate_2019)
Vaughan
Billy B
Toron
City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 4: Choropleth map of assaults in 2019.

Since there are six crime categories, a neighbourhood classification in 'high-', 'medium-' or 'low-crime' must either relate to one of the six categories or a new crime index must be constructed. For this, each crime category is first normalised with respect to the maximum number of occurences across all neighbourhoods. Then these indices are summed over all the six crime categories and normalised again with respect to the maximum of this new 'overall' category. This can be interpreted as follows: an index of '1' means that the neigbourhood has scored, on average, the worst in all crime categories, whereas an index of '0'

[5] This is true, even after the individual neighbourhood names were reconstituted from those that were originally combined because they share the same postal code.



	Neighbourhood	Latitude	Longitude	Assault_2014
count	20	20.000000	20.000000	20.000000
unique	20	NaN	NaN	NaN
top	Victoria Village	NaN	NaN	NaN
freq	1	NaN	NaN	NaN
mean	NaN	43.725322	-79.357328	114.000000
std	NaN	0.063254	0.116476	77.581197

Figure 5: Statistical information of the dataset after an 'inner' merge of geographical and crime datasets. Note the count of only 20 neigbourhoods in stead of the expected 140.

means that there have been no incidences of crime in any of the categories in that neighbourhood. When using the crime *rate* for each crime category, these are already represented as incidents per 100.000 population and therefore the population size of each neigbourhood does not need to be taken into account.

3.2 Machine learning

3.2.1 Exploration of neighbourhood clusters

Data exploration is done mostly by visual inspection. Similarity of neighbourhoods with respect to their public venues, as well as the spatial distribution of similar neighbourhoods is explored using K-means clustering. The number of clusters is arbitrary so it is good to get an idea of the inner-cluster cohesion by inspecting the output of the clustering algorithm for increasing numbers of clusters. From this, it can be observed which clusters break up first and which stay in tact, thereby indicating a greater dissimilariy with respect to the other clusters. Comparison of clusters with the spatial distribution of crimes may give an indication of how these are correlated.

3.2.2 Modelling

K-means clustering provides some qualitative insight in the correlation between neighbourhood composition and the occurence of crime. Clusters of neighbourhoods may therefore be classified according to some crime rating (e.g. high-, medium- and low-crime). This classification, however, is not inherent in K-means clustering. It only takes into account similarity and dissimilarty of neigbourhoods, not how well this correlates with crime. It is, moreover, not necessarily obvious how changes in the neighbourhood composition may cause this classification to change. In other words, it is from K-means clustering not obvious which changes would cause a certain neighbourhood to become more or less similar neighbourhoods with a more desirable crime rating. For this reason, a decision tree model will be built. Such a model will allow parsing of changes to a neighbourhood's composition and a prediction of the new crime rating. De-

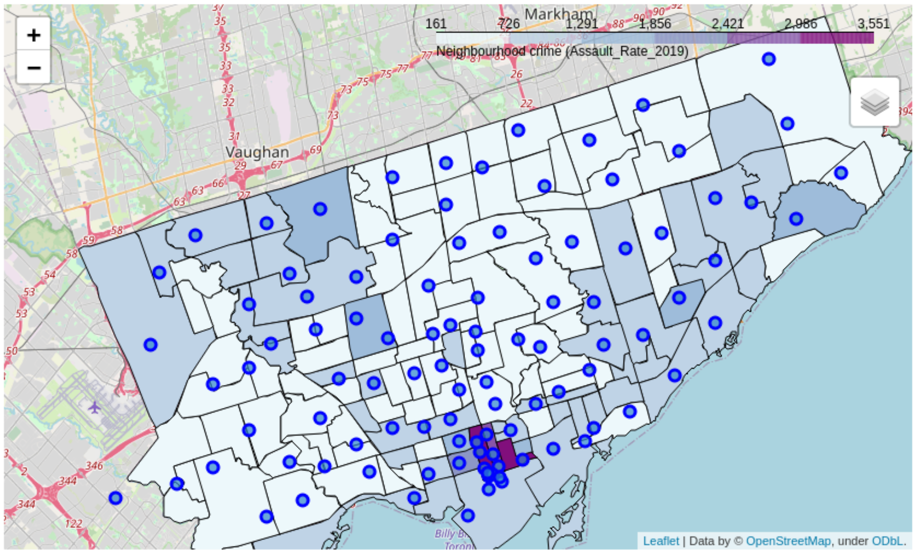

161
726
1,291
1,856
2,421
2,986
3,551
Neighbourhood crime (Assault_Rate_2019)
Markham
Vaughan
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 6: Crime (assault) choropleth, plotted together with neighbourhood information. Neighbourhoods (dots) that lie on purple areas may be manually labeled 'high-crime', neighbourhoods on blue areas 'medium-crime' and neighbourhoods on light blue areas 'low-crime'.

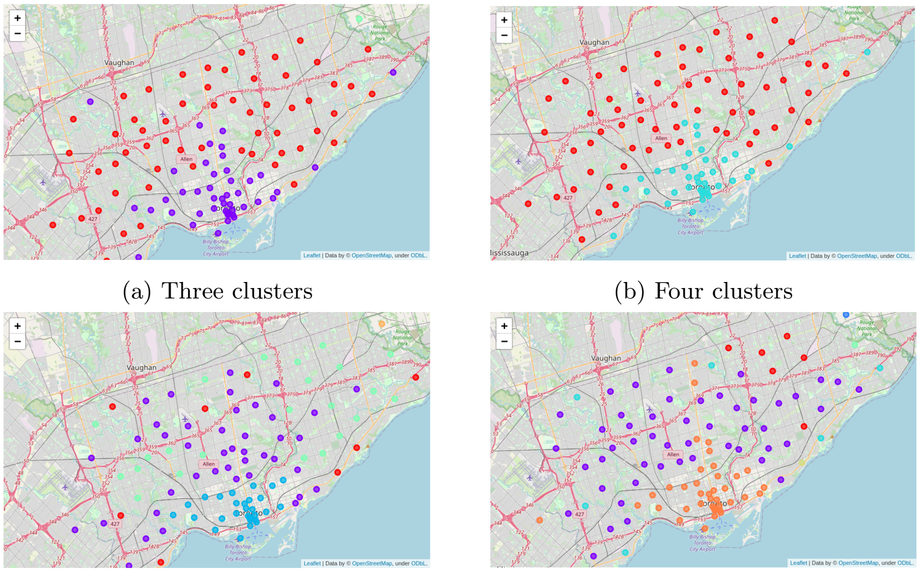
cision trees are very intuitive and from its explicit structure it is easy extimate up front to which changes the outcome will be most sensitive to. For reasons of comparison, however, a support vector machine classifier will also be built.

4 Results

4.1 Visual inspection

Neighbourhoods in Toronto can, based on their public venues, broadly be grouped into two categories: one centred around the harbour and one broad band from east to west (Figure 7). From inspection of the most prevalent venues in each cluster, it is not intuitively clear how these clusters differ from one another (Figure 8).

Spatially, the clusters correlate somewhat with the crime map, insofar that a relatively high crime index is found in the neighbourhoods around the harbour and in the North-West corner of Toronto. This agrees reasonably well with the spatial distribution of cluster '0', which is marked with red dots in Figure 9. A significant amount of neighbourhoods belonging to cluster '0', moreover are found in low-crime areas. The cluster distribution, furthermore, does not clearly discriminate between medium and low crime indices which both are equally well covered by 'cluster '2', which is marked with blue dots in Figure 9.


Vaughan
Allen
427
Toronto
Billy Bishop Toronto City Airport
Rouge National Park
lississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

(a) Three clusters

(b) Four clusters

(c) Five clusters

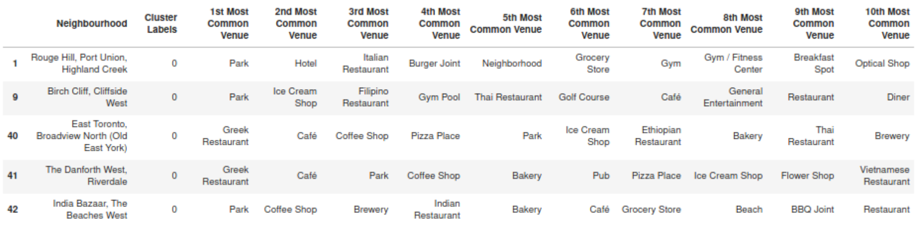
(d) Seven clusters

Figure 7: Different distributions of three, four, five and seven clusters. Note that there are basically two stable clusters, concentrically oriented around the harbour. Increasing the number of clusters only causes small split offs.

4.2 Decision tree and SVM

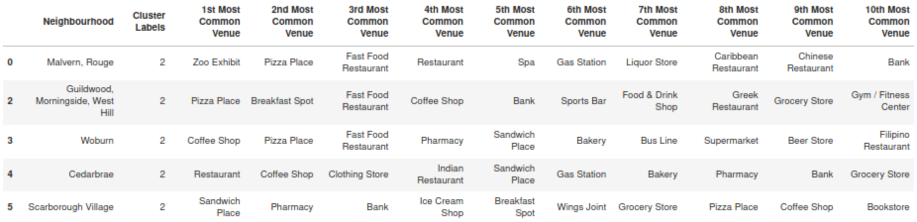
5 Discussion

6 Conclusion



	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Rouge Hill, Port Union, Highland Creek	0	Park	Hotel	Italian Restaurant	Burger Joint	Neighborhood	Grocery Store	Gym	Gym / Fitness Center	Breakfast Spot	Optical Shop
9	Birch Cliff, Cliffside West	0	Park	Ice Cream Shop	Filipino Restaurant	Gym Pool	Thai Restaurant	Golf Course	Café	General Entertainment	Restaurant	Diner
40	East Toronto, Broadview North (Old East York)	0	Greek Restaurant	Café	Coffee Shop	Pizza Place	Park	Ice Cream Shop	Ethiopian Restaurant	Bakery	Thai Restaurant	Brewery
41	The Danforth West, Riverdale	0	Greek Restaurant	Café	Park	Coffee Shop	Bakery	Pub	Pizza Place	Ice Cream Shop	Flower Shop	Vietnamese Restaurant
42	India Bazaar, The Beaches West	0	Park	Coffee Shop	Brewery	Indian Restaurant	Bakery	Café	Grocery Store	Beach	BBQ Joint	Restaurant

(a) Venue composition of first five neighbourhoods in the 'harbour' cluster.



	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Malvern, Rouge	2	Zoo Exhibit	Pizza Place	Fast Food Restaurant	Restaurant	Spa	Gas Station	Liquor Store	Caribbean Restaurant	Chinese Restaurant	Bank
2	Guildwood, Morningside, West Hill	2	Pizza Place	Breakfast Spot	Fast Food Restaurant	Coffee Shop	Bank	Sports Bar	Food & Drink Shop	Greek Restaurant	Grocery Store	Gym / Fitness Center
3	Woburn	2	Coffee Shop	Pizza Place	Fast Food Restaurant	Pharmacy	Sandwich Place	Bakery	Bus Line	Supermarket	Beer Store	Filipino Restaurant
4	Cedarbrae	2	Restaurant	Coffee Shop	Clothing Store	Indian Restaurant	Sandwich Place	Gas Station	Bakery	Pharmacy	Bank	Grocery Store
5	Scarborough Village	2	Sandwich Place	Pharmacy	Bank	Ice Cream Shop	Breakfast Spot	Wings Joint	Grocery Store	Pizza Place	Coffee Shop	Bookstore

(b) Venue compostition of the first five neighbourhoods in the 'broad band' cluster.

Figure 8: Compositions of the most stable clusters. Note that although these are stable and therefore very dissimilar, the dissimilarity is not intuitively obvious.

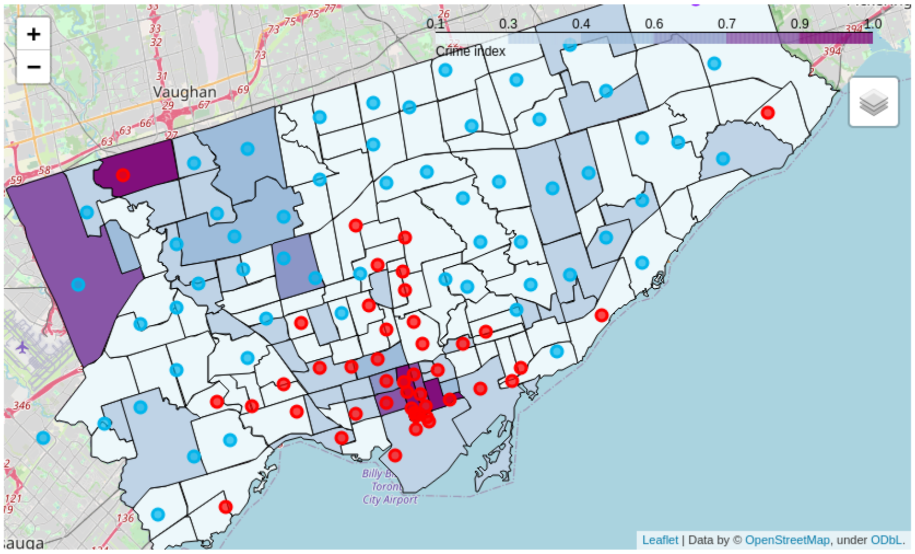
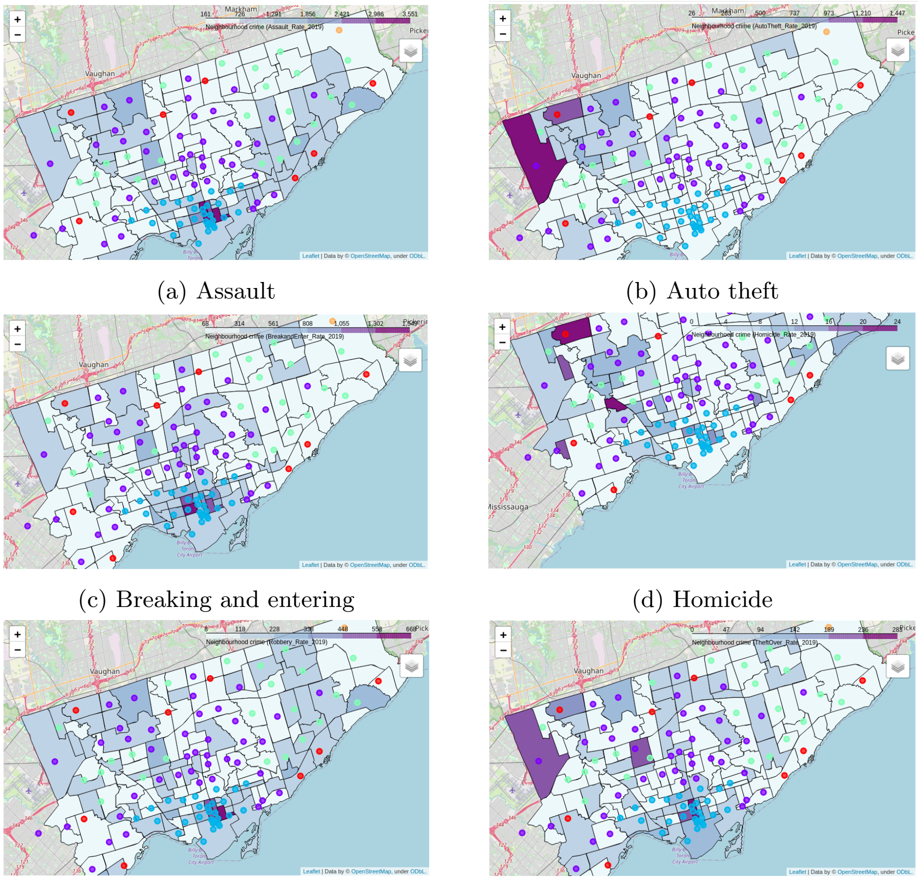


Figure 9: Spatial distribution of crime and of neighbourhood clusters.


161 726 1,291 1,856 2,421 2,986 3,551
Neighbourhood crime (Assault_Rate_2019)
Markham
Vaughan
Picke
Leaflet | Data by © OpenStreetMap, under ODbL.
26 263 500 737 973 1,210 1,447
Neighbourhood crime (AutoTheft_Rate_2019)
Markham
Vaughan
Picke
Leaflet | Data by © OpenStreetMap, under ODbL.
68 314 561 808 1,055 1,302 1,549
Neighbourhood crime (BreakandEnter_Rate_2019)
Vaughan
Pickerin
Billy Bi Toront City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.
0 4 8 12 16 20 24
Neighbourhood crime (Homicide_Rate_2019)
Mississauga
Billy Bi Toront City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.
8 118 228 338 448 558 668
Neighbourhood crime (Robbery_Rate_2019)
Vaughan
Pick
Billy Bi Toront City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.
0 47 94 142 189 236 283
Neighbourhood crime (TheftOver_Rate_2019)
Vaughan
Picke
Billy Bi Toront City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

(a) Assault

(b) Auto theft

(c) Breaking and entering

(d) Homicide

(e) Robbery

(f) Theft other

Figure 10: The six crime categories, plotted with the five neighbourhood clusters.