The Impact of Different Venues on COVID-19 Transmission in the city of Toronto, Canada

Yicong Li

June 25th, 2021

1. Introduction

1.1 Background

Coronavirus disease 2019 (COVID-19) pandemic is an ongoing global public health issue caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It caused great damage around the world. According to the data records provided by Centers for Disease Control and Prevention (CDC), the number of infection cases in the United States until June 25th, 2021, is about 34 million [1], but there may still be undetected potential infected cases. In some countries, the virus has even mutated to spread more widely. Thus, we need to know the possible methods which could be used to control the virus transmission in order to prevent further spread of pandemic. Currently, the main known route of transmission of SARS-CoV-2 is through close direct or indirect person-to-person contact [2]. Based on the rapid spread of the virus, complete closure of the area is an effective and feasible solution. At the same time, we also need to know which specific facilities have enhanced the spread of the virus.

1.2 Problem & Interest

In objective terms, the occurrence of cases is highly related to both local and global geospatial properties of places. Toronto is a typical global city with different culture groups. Its diversity and population mobility might provide valuable indication on transmission patterns of virus, moreover, the information about and appraisal of local epidemic might be able to provide reference to other places in the world.

In this project, we will explore the influence of different venues categories on the occurrence of COVID-19 cases in the city of Toronto, Canada, with some basic data science techniques. The data analysis of possible transmission points could not only provide some reference for controlling the spread of the virus, but also offer valuable cases for preventing the spread of other infectious diseases in ways of analyzing the characteristics of these transmission points and executing targeted public health management on similar locations. The result might be concerned by the local government for decision-making on related issues.

2. Data acquisition and pre-process

2.1 Data Sources

Data about the confirmed and probable cases of patients infected with SARS-CoV-2 virus is provided by Open Data Portal of the city of Toronto [3]. The dataset is refreshed weekly and contains all cases since the beginning of pandemic, January 2020. One of the most important features of this dataset is that is the cases are assigned with geographical information, which is the neighborhoods name.

Neighborhood Data of Toronto city is also provided by Open Data Portal [4]. It includes the boundary information of Toronto's 140 geographically distinct neighborhoods. These data will give us basic geospatial reference for both graph plots and association analysis.

Foursquare Place API [5] offers access to its global database which contains location-based data of venues, including venue trending, venue categories, venue latitude and longitudes and so on. We would like to apply categories defined and given by the Foursquare database and assume venues of the same category have the same influence on case occurrence. It returns a list of venues mainly based on specific location with radius to search within. Since there is no direct data source about neighborhood centroids in Neighborhood data, we apply k-means clustering on the boundaries coordinates to find the latitudes and longitudes of centroids. One of the disadvantages of this API is that we could only get data on a limit of 50 venues with each call of any endpoint. Therefore, here we mainly call for category and geospatial information of 50 trending venues on an assumption that population mobility and human contacts are prior influences on virus transmission as mentioned in the introduction. Although database renews frequently, the API versioning parameter provides control over the time period of the location data. This feature contributes a significant influence on the accuracy of analysis.

2.2 *COVID-19 Cases in Toronto* Dataset Pre-Processing (Table 1)

After obtaining a preliminary understanding of the data, we confirm that all rows of the dataset are unique cases. Not all data could be used in the later analysis: cases of 'Probable' Confirmation Classification are removed based on consideration of cases occurrence; since we try to research the virus transmission in a localized region, cases which have the Source of Infection on 'Travel' or 'No information' should not be considered.

Besides, as Toronto city declared its first call of local state of emergency [6] on March 23, 2020, which prevented essential businesses from operating and gave direct warning to local people, thus caused interference on general businesses of venues and population mobility, cases infected after this date might bring bias to the dataset and should not be considered.

We need to know the maximum latency period of disease to estimate the latest date that the disease was acquired. One of the many studies about the development of COVID-19 symptoms at the beginning of 2020 presents that the longest time for symptoms to be shown is 24 days [7], which is about a month. Given episode dates, which refers to the earliest date recorded for symptoms shown, laboratory specimen collected, or infection reported, of cases in the dataset, all cases of episode date before April 23, 2020, which is 30 days after the first

call of the local state of emergency, are chosen from the dataset.

Dataset Name	Kept features	Dropped features
COVID-19 Cases in Toronto	Neighborhood Name, Source of Infection, Classification, Episode Date	_id, Assigned_ID, Outbreak Associated, Age Group, FSA, Reported Date, Client Gender, Outcome, Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU and Ever Intubated

Table 1. Feature selection of *COVID-19 Cases in Toronto* Dataset

2.3 *Boundaries of City of Toronto Neighborhoods* Dataset Pre-Processing (Table 2)

K-means is Applied to find the centroid of each Neighborhood (Figure 1).

Values in 'AREA_NAME' are strings composed of both neighborhood names and area codes in parenthesis, in this case, we add a new column 'Neighborhood' which contains 'AREA_NAME' with area code removed for further matching process. However, when we try to merge the processed dataset of *COVID-19 Cases in Toronto* and the processed dataset of *Boundaries of City of Toronto Neighborhoods* by neighborhood names, there are still some neighborhoods names that have slight differences which make them never match with each other, for example, 'Danforth East York' and 'Danforth-East York'. Thus, we process a manual substitution handling to make neighborhood names of two datasets consistent.

Dataset Name	Kept features	Dropped features
Neighbourhoods	(Neighbourhood), AREA_NAME, geometry	_id, AREA_ID, AREA_ATTR_ID, PARENT_AREA_ID, AREA_SHORT_CODE, AREA_LONG_CODE, AREA_DESC, X, Y, LONGITUDE, LATITUDE, OBJECTID, Shape__Area, Shape__Length, CLASSIFICATION, CLASSIFICATION_CODE

Table 2. Feature selection of *Neighbourhoods* Dataset

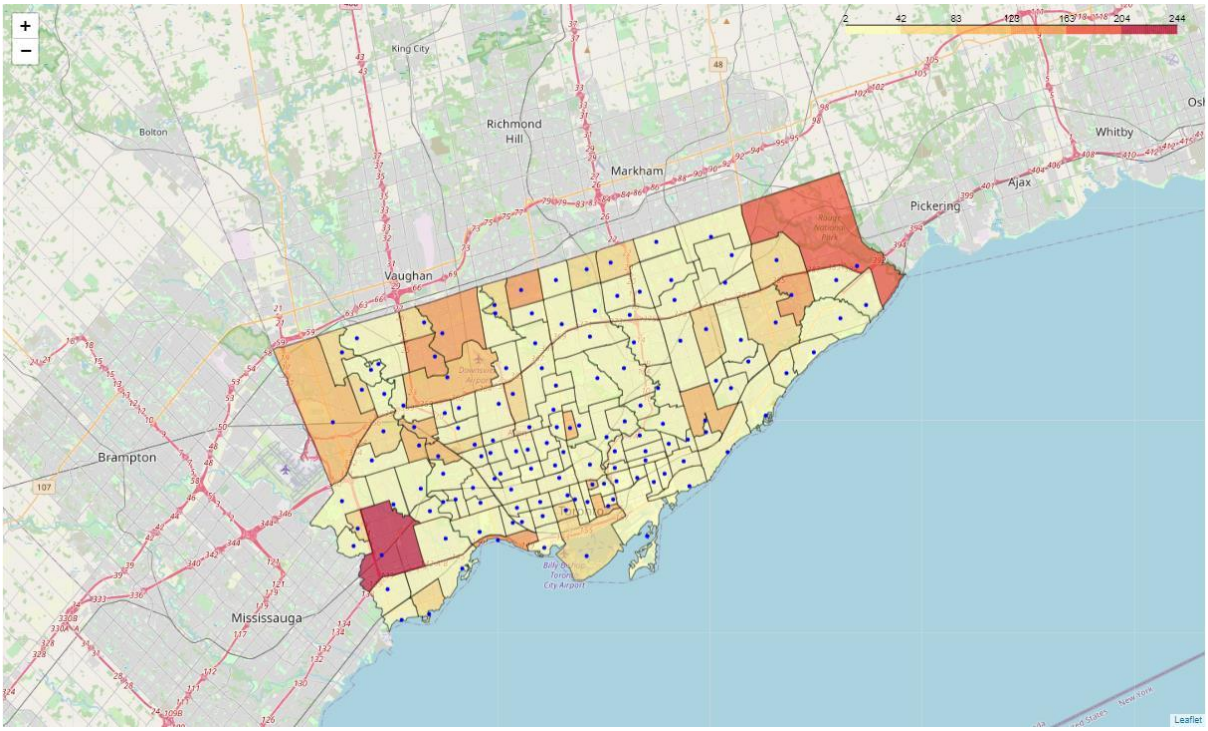

Figure 1. Choropleth map of Toronto Cases with neighborhood centroids

2.4 Venue Data: *Foursquare* Data Retrieving (Table 3)

Modern people might travel a long distance through transportation from places to places. It has been shown in a research about COVID-19 pandemic effect on Sweden Population Mobility that people might still travel as far as 10000 meters or more under mild policies [8]. On the other hand, there is also research shown that most regular basis travel of individuals ranged in a finite neighborhood [9]. As we assume that the disease was mostly transmitted in public through crowd population, we could retrieve trending venues by utilizing the 'trending' section of 'explore' end point, which returns a list of venues near the current location with the most people checked in. Given the limit of project time and cost, we apply trending with a radius of 10000 meters to make mobility relatively well-concerned.

We define the version parameter of API as '20200101' to make sure information about venues is reasonable in terms of time. After going over all neighborhoods to get all nearby trending venue information, the unique venues from dataset based on their venue id is shown (Figure 2).

Dataset Name	Kept features	Dropped features
Venue Data: Foursquare	Neighbourhood, Venue_id, Venue, Category, Latitude, Longitude	

Table 3. Feature selection of Venue Dataset

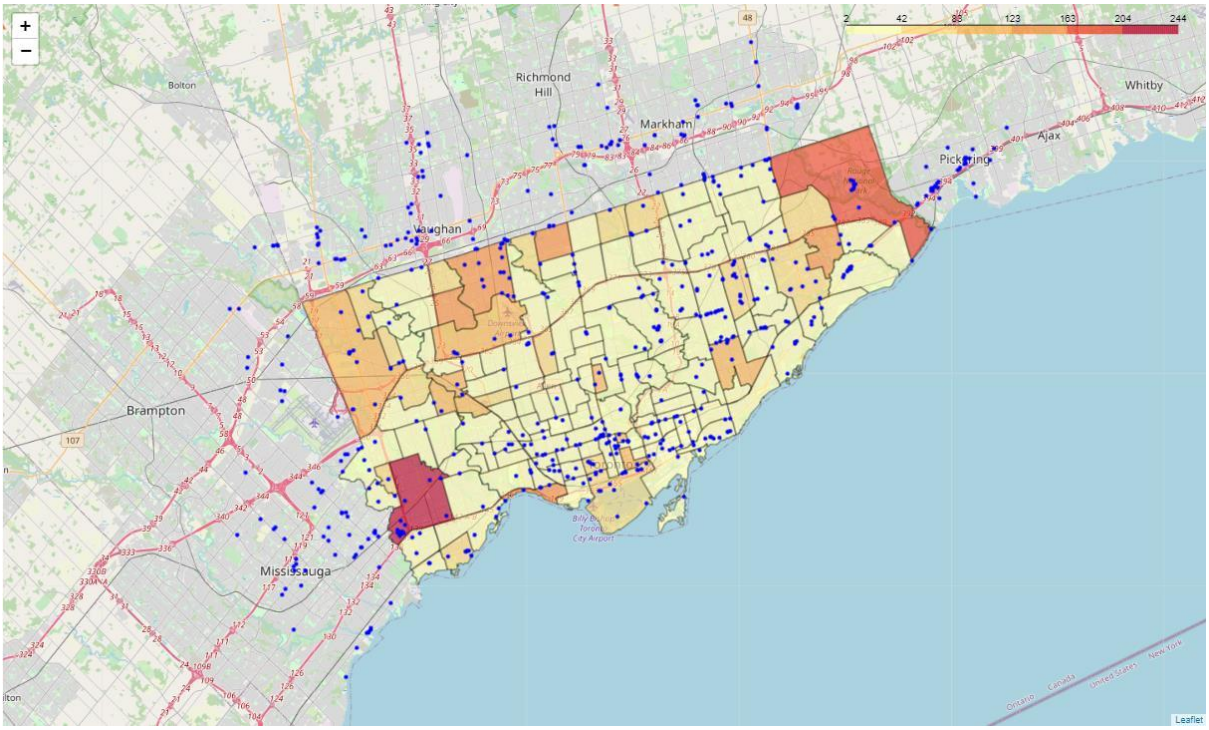

Figure 2. Choropleth map of Toronto Cases with trending venues

2 Modeling

Given the limit of Foursquare API, only 50 venues available for each centroid in defined radius. We have tested that there are 6959 venues for 140 neighborhoods in radius of 3,000 meters on version 20200101, and there are certainly some venues of neighborhoods not returned. Even for 1604 returned venues in radius of 500 meters, there exist neighborhoods which reach 50 venues limit. It makes histogram visualization and Density-based clustering invalid here. In this case, we focus on finding the difference of influence of venue categories on case occurrence with these 140 neighborhood samples, therefore, multi-linear regression is applied to find the weight of each category.

We make a further process on Venue dataset, applying One-hot Encoding to turn Venue category into dummy variable. It maps the category values into columns with values of 0 and 1. Then we group venues by neighborhood. Venue dataset and Case dataset are merged as final input for Multi-linear Regression modeling. The training result shows both positive and negative weights. The positively weighted categories, marked by red signs on the map (Figure 3), may have potentially high influence on the occurrence of COVID-19 cases.

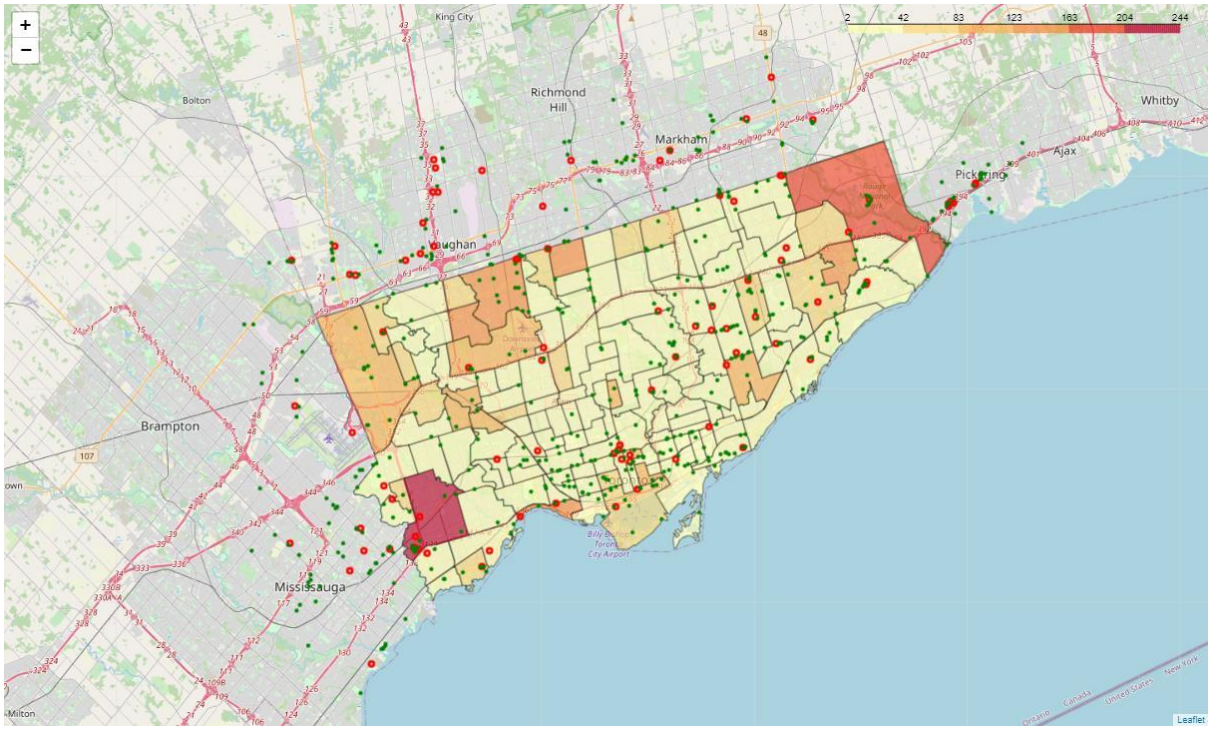

Figure 3. Choropleth map of Toronto Cases with identified trending venue candidates

3 Conclusions & Discussion

Purpose of this project was to find possible venues categories which may contribute to the occurrence of COVID-19 cases based on trending venues in the city of Toronto, Canada, during the beginning of the pandemic in order to provide local government and other possible research of interest to this topic with ideas which might narrowing down the scope of possible geographical search on virus transmission. By research on the phenomenon, we have decided the time and space range of the cases, which could justify further analysis and then generated evaluation on trending venue categories. Venues with positive or negative weights were then visualized on the map to offer support on understanding the results.

However, the project may not end at this point. Further improvements could be made in the project process, for example, if we could retrieve information of all venues around the city of Toronto, instead of just taking the trending venues, we were able to implement density-based clustering and other machine-learning models and compare their results and performances with each other, instead of just using multivariate linear regression. It is also possible to take the radius parameter of Foursquare venue selection as a hyper-parameter to see the possible changes on category weights caused by selected venue data based on different radius or maximum distances from the centroids of neighborhoods. In the meantime, more variables could be considered as hyper-parameters, including age group, gender, cultural distribution and so on to build more sufficient models in order to get more convincing results.

4 References

1. *How COVID-19 Spreads*. (n.d.). Centers for Disease Control and Prevention. Retrieved June 24, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html.

2. *United States COVID-19 Cases, Deaths, and Laboratory Testing (NAATs) by State, Territory, and Jurisdiction*. (n.d.). Centers for Disease Control and Prevention. Retrieved June 24, 2021, from https://covid.cdc.gov/covid-data-tracker/#cases_totalcases.

3. *COVID-19 Cases in Toronto* (Jun 23, 2021). (2020). [Dataset]. Toronto Public Health. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

4. *Neighbourhoods* (Mar 15, 2021). (2021). [Dataset]. Social Development, Finance & Administration. https://open.toronto.ca/dataset/neighbourhoods/

5. *Toronto Venue Data* (Jan 1st, 2020). (2020). [API]. Foursquare. https://developer.foursquare.com/docs/places-api/

6. *COVID-19 pandemic in Toronto*. (n.d.). Wikipedia. Retrieved June 22, 2021, from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Toronto.

7. Yang, Z., Zeng, Z., Wang, K., Wong, S. S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., … He, J. (2020). *Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions*. Journal of thoracic disease, 12(3), 165–174. https://doi.org/10.21037/jtd.2020.02.64.

8. Dahlberg, Matz & Edin, Per-Anders & Grönqvist, Erik & Lyhagen, Johan & Östh, John & Siretskiy, Alexey & Toger, Marina. (2020). *Effects of the COVID-19 Pandemic on Population Mobility under Mild Policies: Causal Evidence from Sweden*.

9. Song, Chaoming & Qu, Zehui & Blumm, Nicholas & Barabasi, Albert-Laszlo. (2010). *Limits of Predictability in Human Mobility*. Science (New York, N.Y.). 327. 1018-21. 10.1126/science.1177170.