MORE OF THE SAME

Diversity of Amenities in Neighborhoods as a Measure of Socioeconomic Wellbeing

Nasruddin Nazerali

More of the Same

Diversity of Amenities in Neighborhoods as a Measure of Socioeconomic Wellbeing

Nasruddin Nazerali

Introduction

Based on spatial data analysis of New York City and Toronto in which I segmented and clustered different neighborhoods based on the available popular amenities on Foursquare, I have made the following observations:

1. The types of amenities in Toronto appear to be more evenly spread around the city compared to New York City. Most postcodes were clustered to the same label/class based on the categories of the ten most popular Foursquare venues. This leads me to believe the city may not have a "good side" and "the other side", but offers a uniform experience to residents of any part of town.
2. Some of the clusters in neighborhoods in New York City, especially concentrated in certain burroughs geographically can be characterized by a lack of diversity in types of amenities. This leads me to hypothesize 'more of the same' types of amenities concentrated in certain neighborhoods. E.g., fast food, diner, cafe, ... and other variations on restaurants as a majority of the 10 popular/close Foursquare venues. Thus my title, 'more of the same', or equivalently, 'same difference'.

It is worth asking if the diversity of a city's amenities in different neighborhoods signify the socioeconomic well being of the residents in that area. We can guess that a neighborhood with a park, public swimming pool, gymnasium, museum, theater, cinema as well as cafes and eateries is a more affluent neighborhood, and the residents have more choice, and socioeconomic well being than residents who can choose only between near synonymous categories of eating/drinking establishments.

Another interesting and complicating factor in U.S cities is racial segregation in residential neighborhoods which persists in some cities as a legacy phenomenon despite legislation which fosters fairness. Chicago, IL, and Detroit, MI are cases that are cited as some of the more persistently segregated. Note this Washington Post article with good geospatial visualization.

As Chicago has extensive data on socioeconomic indicators, school performance, and crime, which we have examined in previous IBM classes, it is an ideal case study to examine if there is a correlation between diversity of amenity types concentrated in certain neighborhoods and socioeconomic well being.

Urban planners, city officials, developers and community organizers could benefit from a quantitative study of this type.

Data

This data analysis would benefit from some breadth as well as depth. In the course of the IBM Capstone class, I have performed the segmentation/clustering analysis of New York City (all boroughs) and Toronto using the Foursquare data. The following tasks would round out that analysis and add more insight in terms of breadth:

1. Reexamine the clustering parameters and analyze the performance of different models.
2. Quantify the diversity of categories of venues. This can be a heuristic parameter that makes a kind of entropy index from the clustering output, or the clustering can be re-parametrized as part of task 1.
3. Peform the same analysis on a representative sample of U.S. cities including Chicago. (San Fransisco - as we have crimes data as well - and Atlanta for a good geographic spread.)

For the case of Chicago an in depth analysis is possible due to extensive and well-curated data and can yield an answer to the hypothesis that socio economic factors will track in correlation to diversity of amenities. Data sets include

1. Chicago Public Schools - Progress Report Cards (2011-2012)
2. Crimes - One Year Prior to Present
3. Census Data - Selected Socioeconomic Indicators 2008-2012

The parametrization can be kept to a basic level to start in order to combine data sets, e.g., aggregate report card score, frequency of crimes and hardship index for all neighborhoods. This can be refined incrementally. Some work is needed to aggregate 'community areas', crime geolocations and school zipcodes into neighborhoods to query on Foursquare.

Methodology

1. Our analysis reproduces for the whole of New York City the clustering of Manhattan neighborhoods according to the category of the nearby venues queried from the Foursquare API. We can obtain (up to) the top 100 venues in a radius of 500 m.

Our data has 10295 venues in 306 neighborhoods in the 5 boroughs of NYC. There are 429 unique categories of which eateries of various sorts are clearly dominant as can be observed from the word cloud in figure 1. As we would expect, some more densely settled or busy work neighborhoods have 100 (maximum) venues returned, whereas some neighborhoods are in the single digits or just one venue. We encode categorical parameters with one-hot encoding that will indicate yes(1) or no(0) whether the type of venue is present within the designated radius in the given neighborhood. We can then create a numerical parameter by taking the mean of the frequency of occurrence of each category of venue. We then proceed to use the SciKitLearn K-Means Clustering package with k = 5 to cluster neighborhoods. Port Ivory in Staten Island did not return any venues in the 500 m radius and therefore returns an NaN in the dataframe and

array operations that are performed in the course of the clustering step. We can discard this row for the subsequent analysis.

(NB - in an earlier run, Howland Hook was the neighborhood that returned zero venues. The Foursquare request does not return identical outputs each time.)

Figure 1 It is clear from this simple word cloud that the majority of venues have something to do with food/drink. Parks, supermarkets, pharmacies, spas, gyms peak out with much smaller frequencies. Bodegas (small groceries) at least feature more prominently.

2. We repeated the above analysis with K-means clustering with only 2 clusters instead of 5. The motivation was to discover if two clusters show 'good' and 'bad' neighborhoods in terms of availability of amenities.
3. We will want to explore a different parameterization as indicated by the preponderance of eating establishments well as other synonymous types (supermarket - grocery, gym – fitness, etc.) which should be combined or simplified. The results from analysis step 1 above appear to reasonably differentiate neighborhoods by the types of common venues available. However, alternative models could possibly provide a higher level of insight that is masked by finer details.

We proceed to cluster by the category of venues higher in the hierarchy of categorization ('food' for all sorts of eateries, 'arts and entertainment' for a variety of venues such as museums and golf clubs, etc.) The Foursquare venue category hierarchy can be referred to. We obtain these categories by accessing the directory name above the venue category. As can be observed in the table below, ice cream shops, a Dunkin Donuts shop and an ethnic restaurant are all captured in the category of ‘food’. This reduced our categories from 429 down to only 9 ('food', 'shops', 'parks_outdoors', 'travel', 'arts_entertainment', 'nightlife', 'education', 'building', 'event'). We then

proceed to investigate clustering results for k = 5 and k = 2 for comparison with the previous parametrization.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	food
Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	shops
Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	food
Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	food
Wakefield	40.894705	-73.847201	Dunkin Donuts	40.890459	-73.849089	food

Table 1. A higher category designation for venue types

4. We will explore the city of Chicago by Community Area and simple measures of socioeconomic well-being obtained from three data sets, and perform a K-means clustering based on three scaled features obtained from the data.
    a. Crime rate: number of crimes reported at given addresses (geographic coordinates) which can be averaged and grouped by community area. The geographic coordinates are also averaged to obtain positions for the community area.
    b. Hardship index: obtained from census data and aggregated to community area. This is a summary index of various indicators which we have set aside to simplify the scope of this study
    c. Progress report card on schools, aggregated on the yes/no categorical answer to the question whether adequate progress was made from one year to the next. Schools are grouped by community area and a mean measure obtained from numerically encoded yes(0)/no(1) feature.
    (Note that we insure that the higher each of these parameters, the 'worse' the situation.)

5. Foursquare venues data is used to cluster the same community areas in Chicago – we limit our analysis to the higher category of venue types (8 unique types in this case), and to K = 2 in the K-means clustering algorithm.

Results

New York City

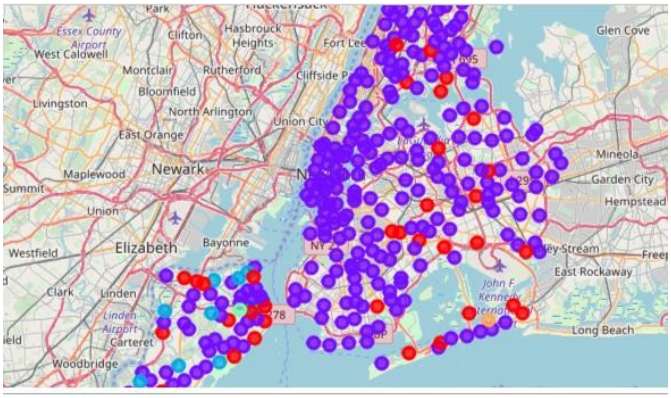

Figure 2 Venue categories: 429. K = 5 clustering. Shows some geographical concentration of clusters but otherwise quite homogenous.

Figure 3 Venue categories: 429. K = 2. The anomalous cluster clearly coastal or riverfront with amenities reflecting the geography: sites have parks, event spaces and other varieties non-food venues.

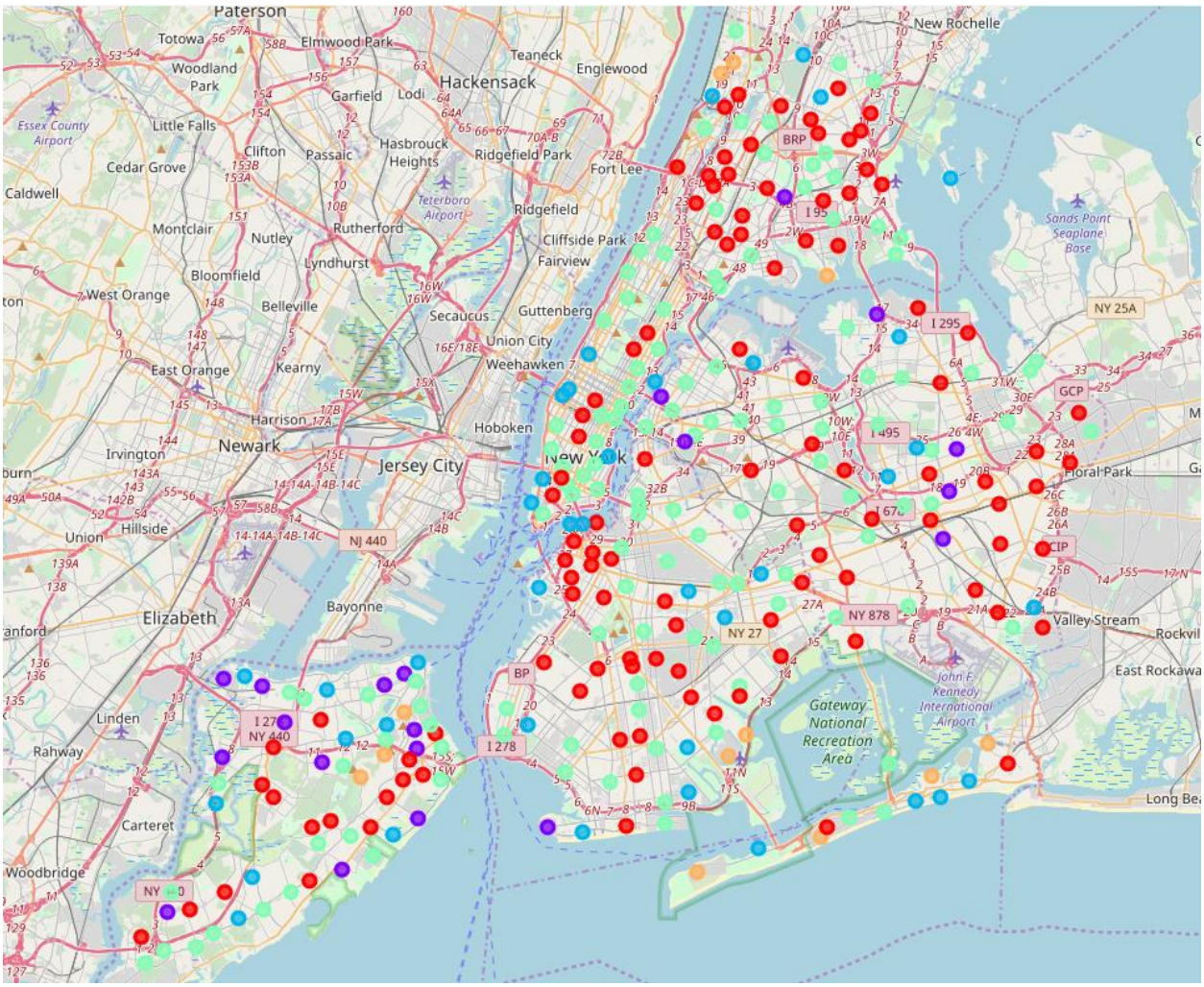

Figure 4 Venue categories: 9. K = 5.

Cluster 0. Red marker. Geographically not concentrated in one area. Ranking of venues -- (by quick inspection) food, shops, nightlife, parks.

Cluster 3. Cyan marker. As widespread and numerous as label 0, somewhat concentrated in Manhattan. Top venues include 'building' (e.g. gyms) and 'travel' (e.g. hotels) in addition to food and shops.

Cluster 4. Orange marker. This cluster appears clearly correlated with coastal/riverside/island locations geographically. It makes sense that travel and nightlife venues appear among the top ranking.

Cluster 1. Violet marker. This cluster appears mostly in Queens and Staten Island. It is prominently absent from Manhattan. Travel, parks and event spaces appear to rank high. It could be these venues are not prominent in the densely constructed and settled parts of the city.

Cluster label 2. Blue markers. These appear to be concentrated around coasts/riverfront neighborhoods. Parks stand out as a prominent venue.

Clustering based on the top level category of venues appears to be a promising direction for analysis. It is difficult to make general conclusions re. NYC based on this clustering. Anecdotal evidence of long-time residents of NYC appears to confirm the general availability of the basic amenities: food (including late night take-out), groceries/pharmacies, and transportation. It is also true that neighborhoods have ethnic majorities and similar businesses which tend to cluster in certain neighborhoods, and that the boundaries of these neighborhoods are fluid. Our analysis confirms some common sense observations about geographical features that tend to concentrate certain neighborhoods by the type of most common venue.

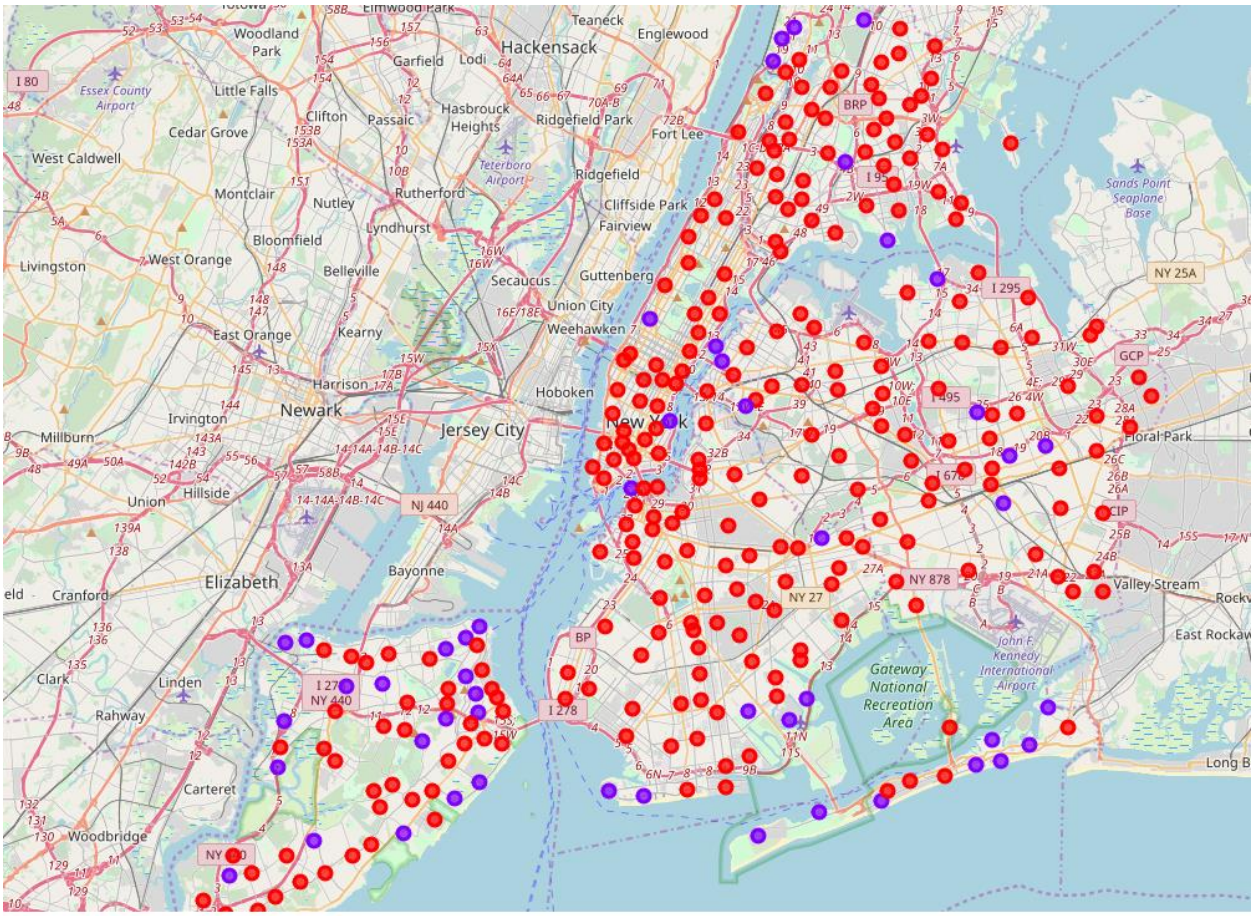

Figure 5 Venue categories: 9. K = 2. Cluster label 1 (violet marker) is certainly correlated with the geographic feature of being close to the shore or the river front (except a few neighborhoods in the center of Queens). Cluster 0 has the generic neighborhood venues whereas cluster 1 has more parks.

Chicago

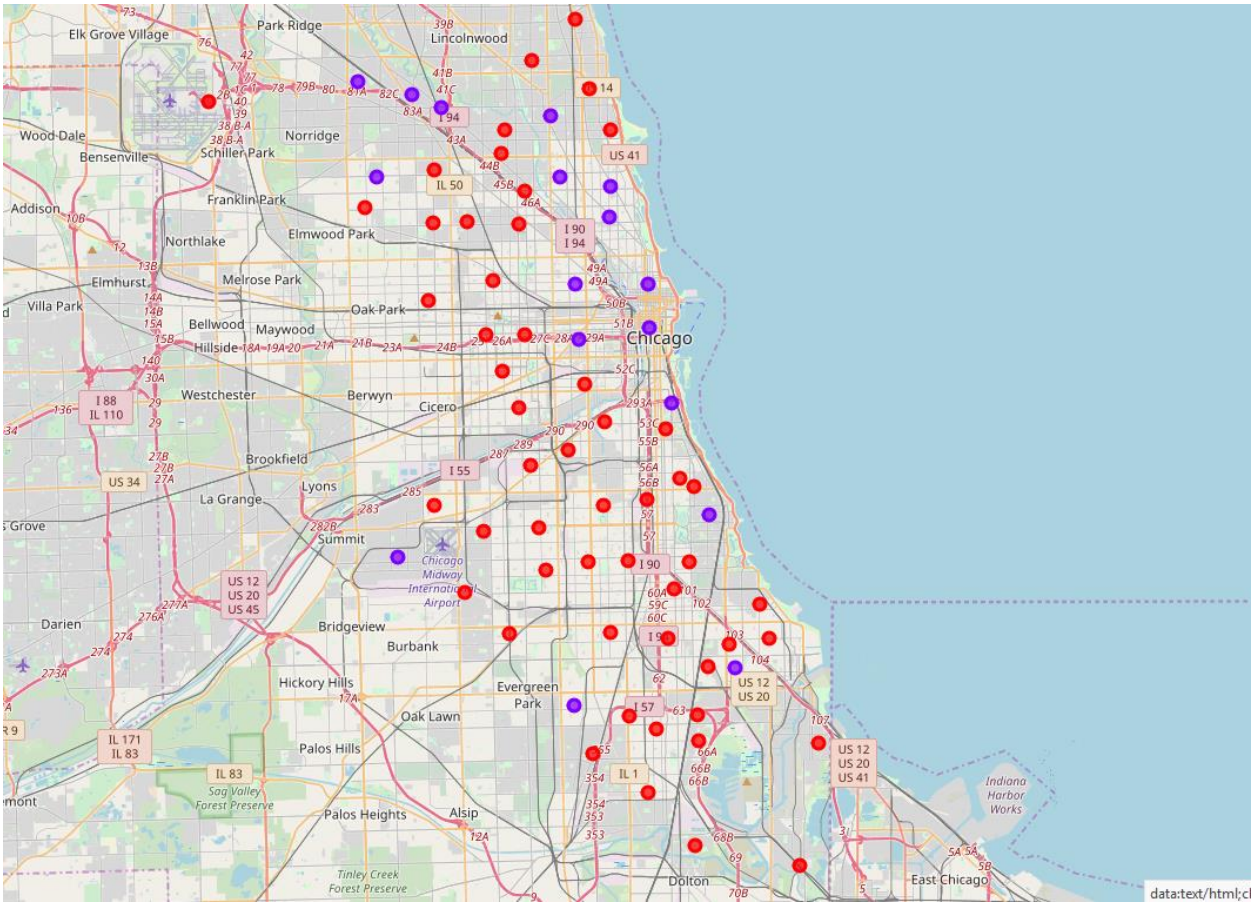

Figure 6 Chicago clustered by socioeconomic indicators.

Although it is not a pronounced effect, the cluster label 1 (blue marker) is concentrated in the northern part of Chicago and label 0 (red marker) is spread out more evenly throughout the city. The results clearly indicate that cluster 1 is doing better in terms of the three features we are investigating. See table 2. Recall, the higher the numerical values of these (scaled) features, the worse off the neighborhood.

	School	Crime	Hardship
Cluster 0 (Red)	0.9550	7.358	60.302
Cluster 1 (Blue)	0.485	5.706	14.353

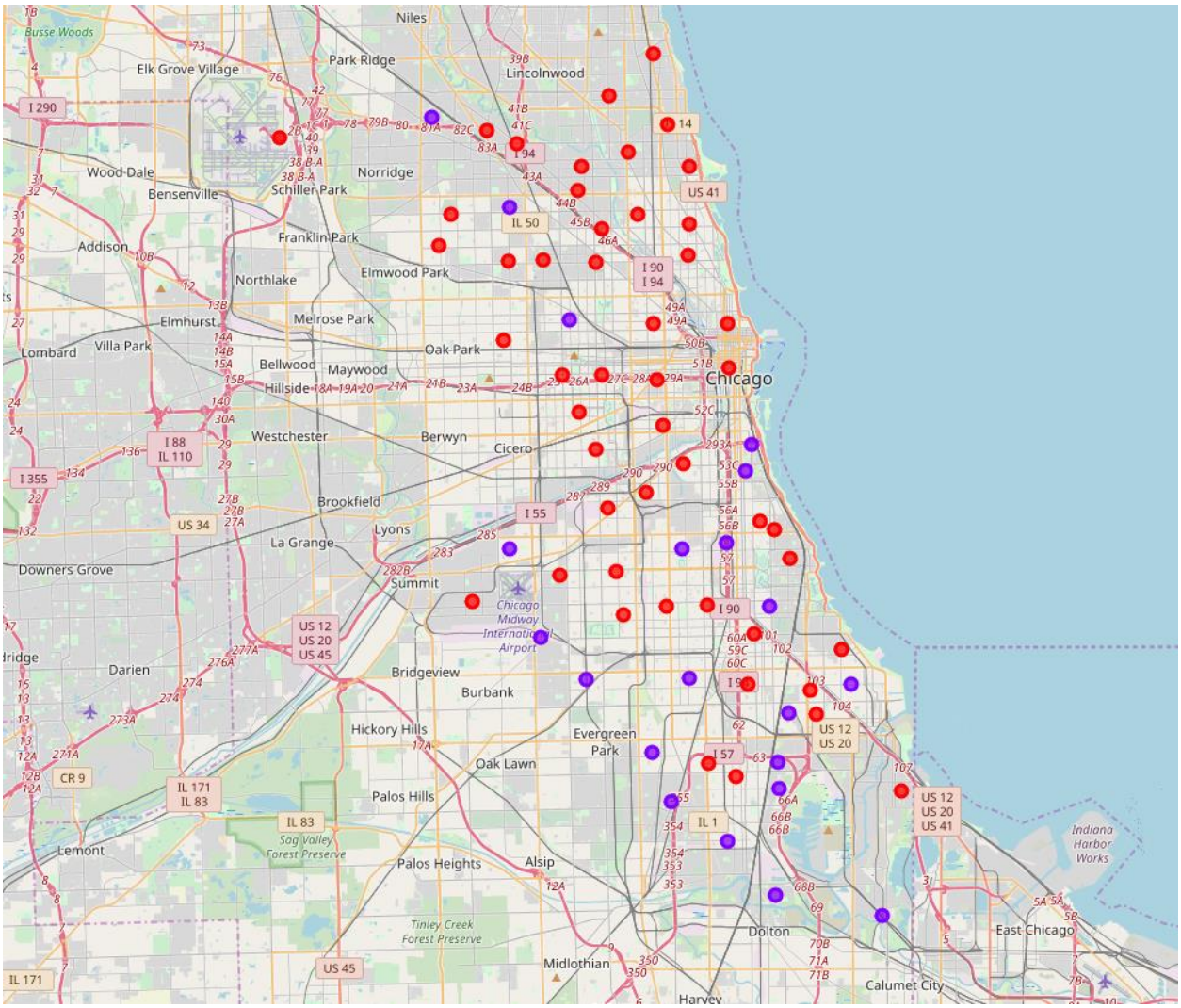

Figure 7 Chicago clustered by venue categories.

The clustering on Foursquare amenities shows a clear north-south geographical correlation, with cluster 0 (red markers, figure 7) to the north and cluster 1 (blue markers, figure 7) to the south. It is difficult to draw substantive conclusions based on the distribution of types of venues in the two clusters that can be observed from a ranking of the categories of venues. However, the analysis bears out the hypothesis that types of amenities available may indeed track with the socioeconomic indicators in a city.

If we focus on the most common venue, it appears clearly that the main cluster (0, red markers, figure 7) are dominated by food venues, whereas the other cluster (concentrated in the south geographically) has more shops, parks and other diverse amenities. This leads to the opposite conclusion from our hypothesis that the more socially and economically disadvantaged neighborhoods will have less diversity of amenities.

Conclusion

Our analysis started with the hypothesis that Foursquare venues data can be used to cluster neighborhoods in cities to reveal whether there is a lack or an abundance of diversity in amenities offered within a given walking distance. We performed a simplified analysis on a higher category of the venues instead of the more detailed venue types that were repeated for New York City as a base case. Chicago provided a good case study because of the easy availability of socioeconomic data which was also used to cluster neighborhoods. There was a confirmation of the hypothesis that socioeconomic factors and the distribution of types of amenities would be correlated. However, counter to our expectation, it is observed that the more affluent parts of Chicago are more homogeneous in terms of amenities provided.

Future Work

It is clear that this line of investigation can be rewarded from a more thorough and fine-tuned approach to the parameterization of the clustering problem. Instead of hundreds of types of venues, or 8-9 top-level categories of venues, it would be fruitful to engineer the venue category to be more theoretically apt to addressing the question whether some neighborhoods are underserved in terms of the diversity of venue types. It would also be illuminating to focus on one type of venue, such as food and examine the subcategories to try to find any correlation with economic or public health indicators if such data can be obtained at the neighborhood level.