A Socio-economic Analysis of Chicago

Suraj Krishnamurthy

1. Introduction

1.1. Background

The socio-economic status of a region, be it a city, country, or anything like it, can be a powerful gauge of the efficiency of governments, business, and economic policy. Researchers have recently begun leveraging the power of machine learning to analyze the effects of economic and financial changes on the well-being of a population, and vice versa. Understanding this multi-faceted relationship is of utmost importance in creating and maintaining a fair social structure, along with a happy citizenry, which will only result in the further growth of the region.

1.2. Scope of the project

In this project, the city of Chicago will be studied between 2008-2012, and the effects of various socio-economic factors on the health and so-called "hardship index" of its citizens will be explored. The project aims to predict the hardship index of a neighbourhood based on the interplay of a number of factors, including per capita income, poverty level, and housing situation.

The universally beneficial effect of public parks on the economic, environmental and mental well-being of a community's residents is well-documented; see 1 and 2, for example The second aim of the project is thus to guide city authorities by suggesting locations to open new parks.

1.3. Scope of impact

An analysis such as this would be of use to a number of groups: governments can use it to guide policy formulation, businesses can use it to enter markets that would benefit from their presence, and the citizens themselves can use it to understand their community better, perhaps forming their own groups to improve their situation, or that of their neighbours.

2. Data sources

- Census data for Chicago between 2008-2012, obtained from the city’s data portal.
- Crime data for the city between 2008-2012, obtained from Kaggle.
- Chicago public school data for the academic year 2011-2012, obtained from the city’s data portal.
- The list of neighbourhoods by community area in Chicago, scraped from Wikipedia.
- The list of public parks in a 500 m radius of each neighbourhood, obtained using the Foursquare API.

3. The relationship between the hardship index of a community and socio-economic factors

3.1. Data acquisition and cleaning

The census dataset contained almost wholly relevant information. It contained the community area name and number, which were needed for cross-referencing data in the crime and school datasets. It also contained the target variable, the hardship index, which is given for each of Chicago’s 77 communities, along with some key indicators, such as the percentage of households below the poverty line, the average per capita income, and the percentage of people living in crowded households. There was only minor manipulation to be done: renaming and typecasting of a couple of columns, and dropping a row of unnecessary information.

The data regarding crimes in Chicago was split in two files at the source: crimes reported between 2008-2011, and crimes reported in 2012. They were concatenated after some cleaning. The crime datasets contained many columns which did not add value to the analysis, such as the crime ID, FBI code, etc., so those columns were picked out which were most useful. In this case, this was the community area number, the latitude & longitude of the crime location, and whether the reported crime eventually led to an arrest. However, the information in this dataset is of all *reported* criminal activity. Since an arrest is a key indicator of a crime actually taking place, those crimes which were reported but did not lead to an arrest were removed from the dataset. The total number of crimes for each community were calculated and added to the dataset. After cleaning the dataset of NaN values, dropping redundant information, and some typecasting, this dataset was ready.

Moving on to the public school data, the important columns had to be picked out, just as with the crime data. Here, this was be the name of the school, the community area number in which the school is located, its latitude and longitude, the average student attendance, and the number of students who ended up in college. Since there are multiple schools in a community, the data was grouped by community.

These three datasets were then merged into one table for the final analysis.

3.2. Exploratory data analysis and visualization

As a first step, the distribution of the target variable, the hardship index, was visualized, by plotting a histogram of its values. This variable ranges from 0-100; 10 bins were made.

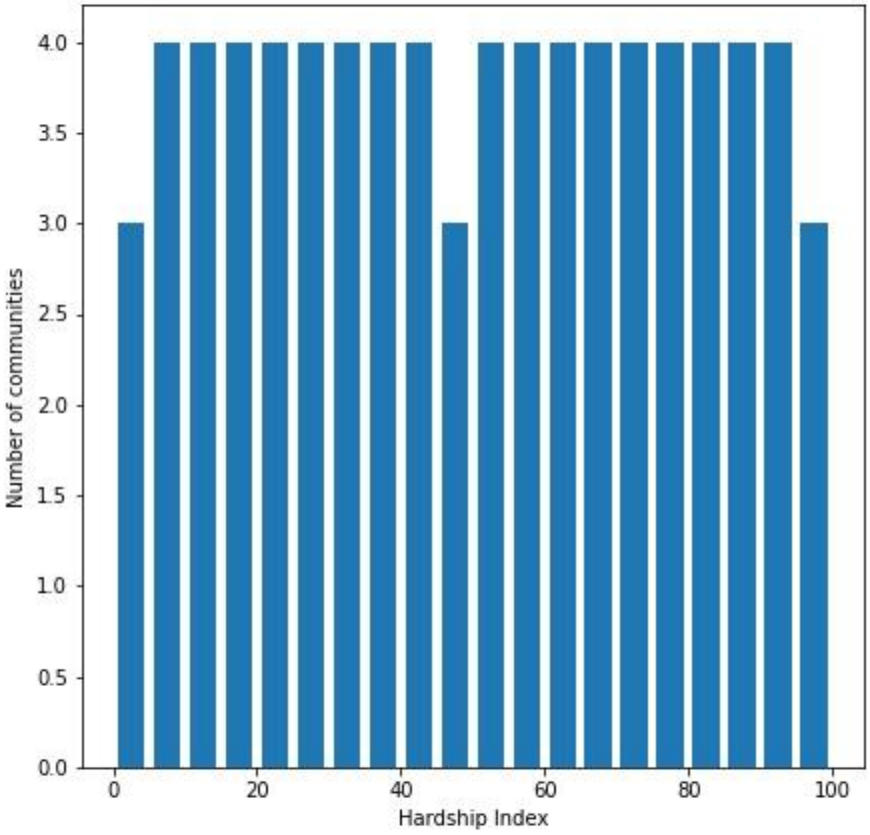

Figure 1: The distribution of hardship index over Chicago's communities.

This was not very informative. The city's communities show an even spread of the metric.

The hardship index was then plotted as a function of the various socio-economic factors that were selected for in the three datasets, in order to see which of them should be chosen for training the machine learning model. For example, the figure below shows the hardship index as a function of student attendance, and the number of schools & eventual college goers.

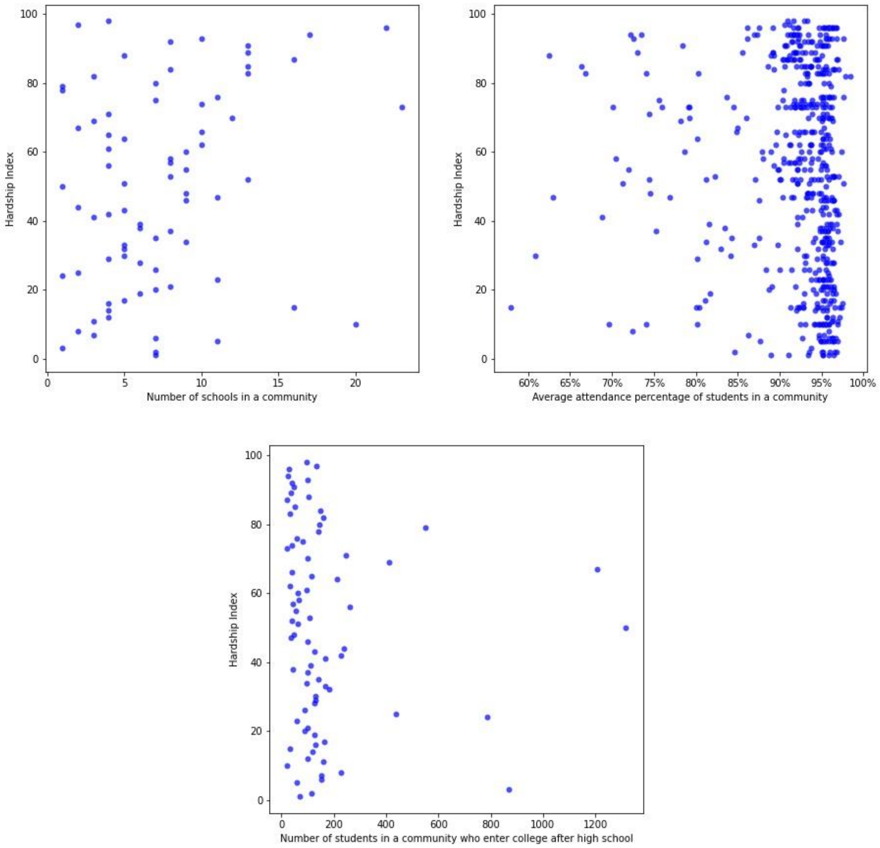

Figure 2: The hardship index as a function of the number of schools in a community (top left), the attendance percentage of students (top right), and the number of students who would go on to college (bottom).

We can see that none of these factors have any impact on the hardship index. Perhaps some pattern would emerge if we could factor in population numbers for each community as well. Unfortunately, this information is not available on the Chicago Data Portal.

Next, influence of the percentage of people living in crowded households, as well as those living in poverty was examined. Simply from a reading of the description of the data, we can guess that as the percentages increase, people must feel more hardship, especially if they live under the poverty line. The figure on the next page shows what the data says.

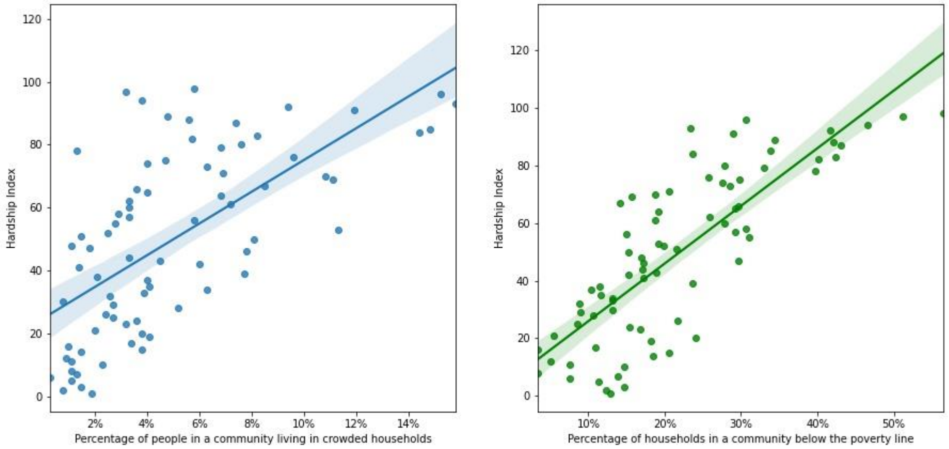

Figure 3: The hardship index as a function of the percentage of people living in crowded households (left), and households below the poverty line (right).

The data corroborates our suspicions, although there are a few observations to be made.

- There is a surprising amount of variation in the hardship index of people living in crowded households. We can see that even if a community has a low percentage of people living in cramped houses (say about 1%), the hardship index varies from close to 0 all the way to 50, with one point even at an index of 80. It is likely that there are other factors coming into play. For example, people living in close quarters but not in poverty probably feel less hardship than their neighbours who do suffer from poverty.
- The correlation between the hardship index and poverty is tighter, although there is some variation, which is to be expected. Curiously, a community having a relatively large number of households below the poverty line (look at the data around 11-12%) can still have a hardship index close to zero.

What's the relationship between the age-related statistics and a community's hardship index? Let's have a look.

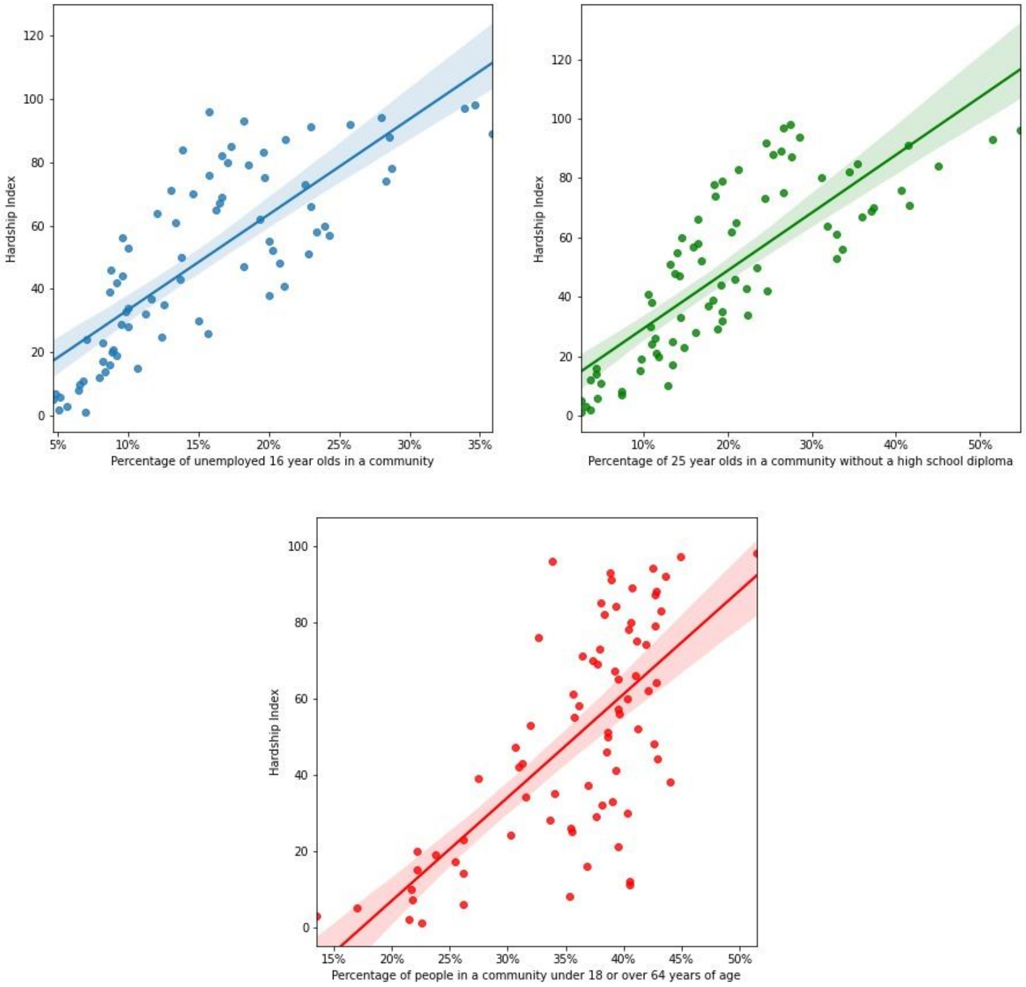

Figure 4: The hardship index as a function of the percentage of unemployed 16 year olds in a community (top left), percentage of 25 year olds without a high school diploma (top right), and percentage of people below 18 or above 64 years of age(bottom).

There is a positive correlation between each of the variables and the hardship index.

- Unemployment has an obvious detrimental impact on a family or community's hardship. We see that the absence of a job for young adults in high school (16 year olds) can place a strain on the community. The importance of a high school-level education, at minimum, is also seen in the top-right graph.

- An interesting pattern can be seen in the bottom plot. There appears to be some boundary of the percent of people in a community either aged below 18 or above 64 (around 27-30%) below which the hardship index is contrained to be less than 20. Once this percentage goes above 30%, however, the hardship increases dramatically for a community. More accurately, the potential for hardship increases dramatically, since the index spans almost the entire range. An increase in this percentage means a larger number of people who are not in the workforce (18 being just about the age at which one would enter, and 64 being around retirement age). That may be at least part of the explanation for this correlation.

The effect of per capita income on a community can be easily guessed: the more the income per capita, the easier the lives of its members should be.

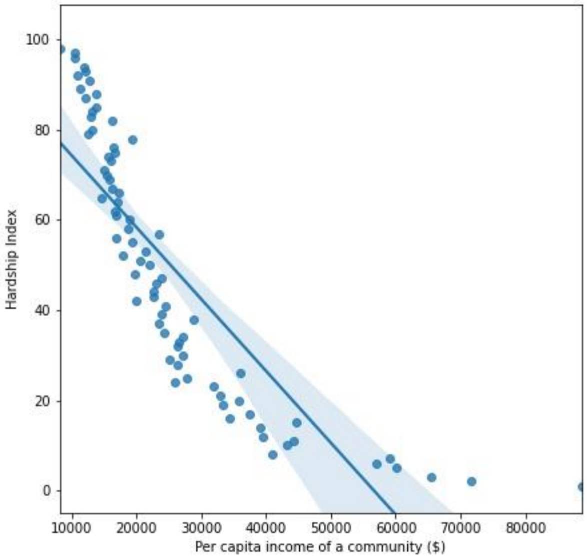

Figure 5: The hardship index as a function of a community's per capita income.

A steep drop of the hardship index is observed with increasing per capita income. When it becomes more than about 60,000$, the index becomes effectively zero.

Finally, the effect of the cumulative number of crimes over the period of 2008-2012 on the metric was examined.

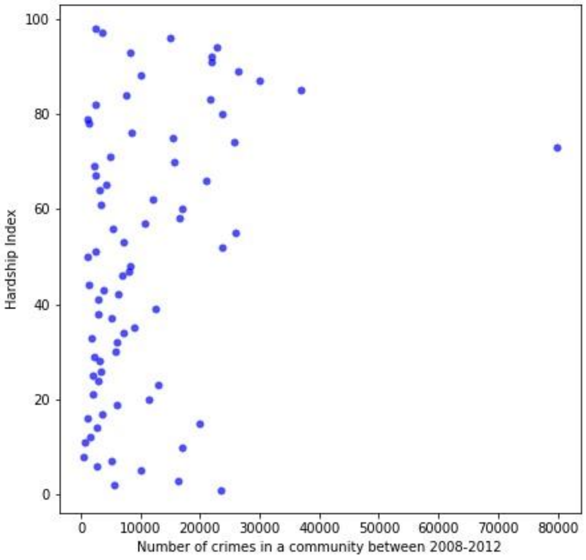

Figure 6: The hardship index as a function of the number of crimes.

Surprisingly, this does not seem to have an impact on the hardship index.

The variables to use for the machine learning model could now be picked: percent of people living in crowded households, percent of households living in poverty, percent of unemployed 16 year olds, percent of 25 year olds without a high school diploma, percent of community residents under 18 or over 64 years of age, and per capita income.

3.3. Predictive modeling and analysis

Although the independent variables appear to have approximately linear correlation with the target variable, 3 models were tested: a multiple linear regression model (with 6 independent variables), as well as polynomial regression models of degrees 2 & 3. Especially since we do not have a large dataset (only 77 data points), the different models were then evaluated using cross-validation. The R^2 metric was used for this purpose.

The data was split into training and testing sets, using an 80/20 ratio, and scaled using the `StandardScaler` method of the `preprocessing` module of the scikit-learn library. The training data was then split into five folds in the k-fold cross-validation model; here are the scores for each machine learning model:

	Multiple Linear Regression	**Polynomial Regression; n=2**	**Polynomial Regression; n=3**
Fold 1	0.9751	0.9566	0.8675
Fold 2	0.9788	0.9171	0.8019
Fold 3	0.9653	0.9280	0.9577
Fold 4	0.9775	0.9434	0.9625
Fold 5	0.9497	0.7903	0.9768
Average	**0.9693**	**0.9071**	**0.9133**

Table 1: Cross-validation scores for the three models that were tested.

A very good score was obtained with the multiple linear regression model, whereas R^2 shows quite sizeable fluctuations depending on the slice of data used for training for a second degree polynomial model: the last fold has a score of 0.79, compared to the other scores, which are all > 0.9. The same problem is observed in the third degree polynomial model.

So, the multiple linear regression model was chosen, and the full training set was fed to the model, which returned the following intercept and coefficients:

Intercept	**Percent of crowded housing**	**Percent of households below poverty**	**Percent of 16 year olds w/o job**	**Percent of 25 year olds w/o high school diploma**	**Percent aged below 18 or above 64**	**Per capita income**
48.5081	2.8012	8.9105	6.4735	10.4676	5.18867	-1.9559

Table 2: Intercept and coefficients of the multiple linear regression model.

Using the model with these values on the test data gave an R^2 score of 0.95322.

The performance of the model was visualized by plotting the predicted values versus the actual values.

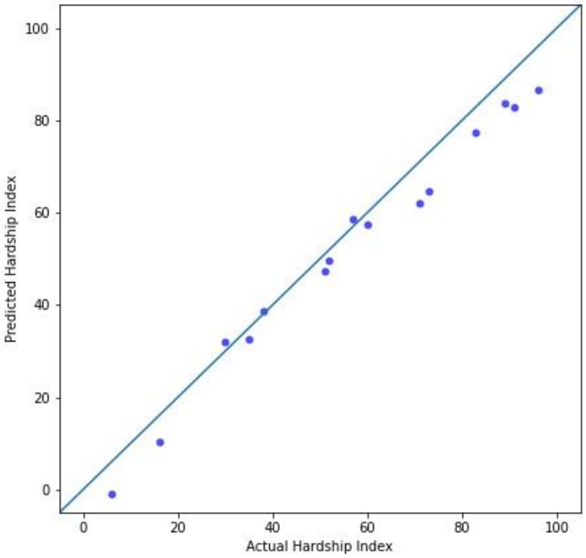

Figure 7: Predicted hardship index using the multiple linear regression model versus the actual values.

If the predicted values were exactly equal to the test values, the result would have been a straight line, since x=y in that case; in other words, the straight light blue diagonal line represents a perfect fit. It seems like the model consistently underpredicts the data, evidenced by most of the points lying below the line. This is likely a consequence of the small number of data points we are working with.

4. Predicting useful locations of public parks in the city

4.1. Data acquisition and cleaning

The list of neighbourhoods in Chicago by community area was scraped from Wikipedia. An immediate problem with the scraped data was that there were a few community names which were not present in any of the previous datasets (census, crime, school), so only those neighbourhoods were be retained whose community names matched with the names in the census dataset, thus maintaining consistency across all datasets.

The latitude and longitude of each neighbourhood was then retrieved using the geocoder module. Unfortunately, some of the neighbourhoods had redundant coordinates; these were clearly neighbourhoods in close proximity to each other which the geocoder module could not distinguish finely enough. Therefore, the dataset was checked for duplicates, and whichever neighbourood came first in such redundant sets was retained.

The next step was using the Foursquare API to retrieve some information about parks located around the neighbourhoods of Chicago. First, a general list of the top 50 venues in a 3 km radius of each neighbourhood was retrieved. The public parks were then picked out this list. Given the radius of the search, it was likely that duplicate results for multiple neighbourhoods would have been returned, so the unique parks that the API found were retained. This was done by sorting the data by the distance between a park and the neighbourhood coordinates. Then, whichever occurrence of a park came first was picked, since the farther away a park is, the more likely it is that it got picked up by the API when searching for venues in a different neighbourhood.

4.2. Exploratory data analysis and visualization

Using the folium library, a map of Chicago, its community areas and its neighbourhoods was created.

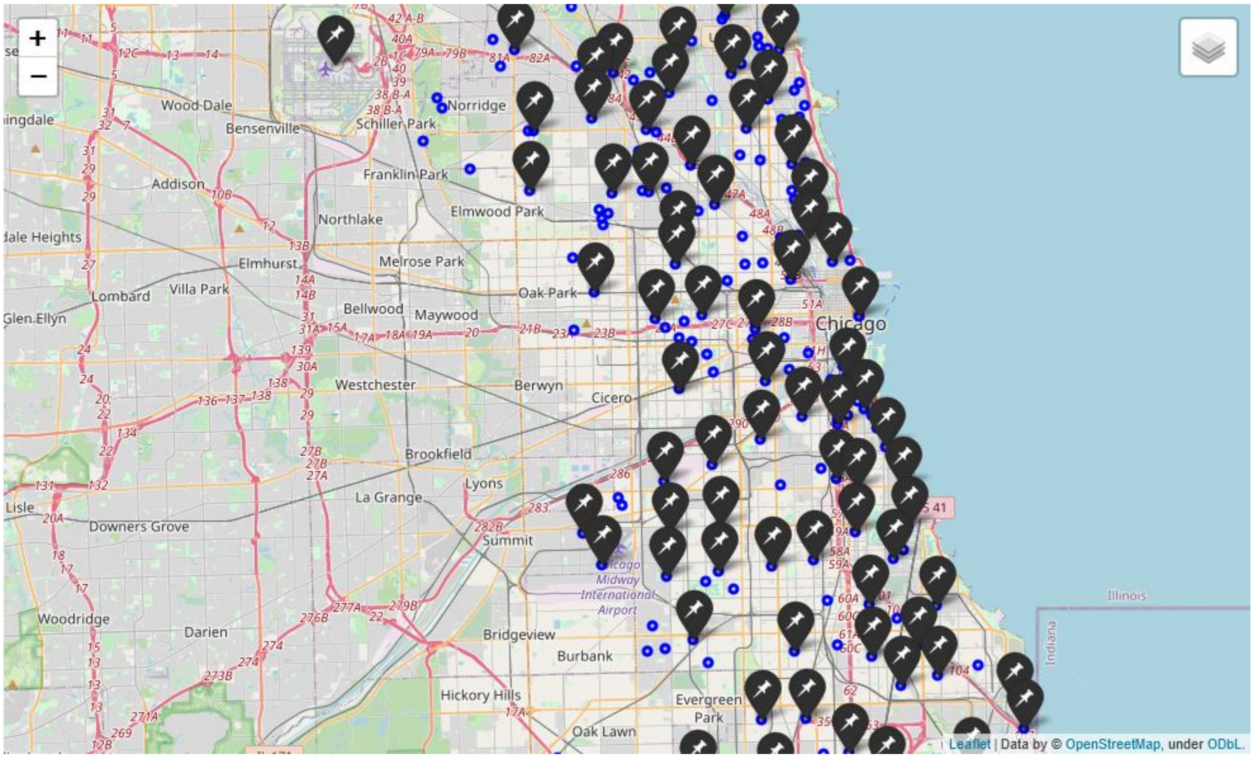

Figure 8: A map of Chicago with its community areas marked with the black pins, and its neighbourhoods denoted by the blue circles.

The distribution of the various parks that were retrieved in the data acquisition stage is shown below.

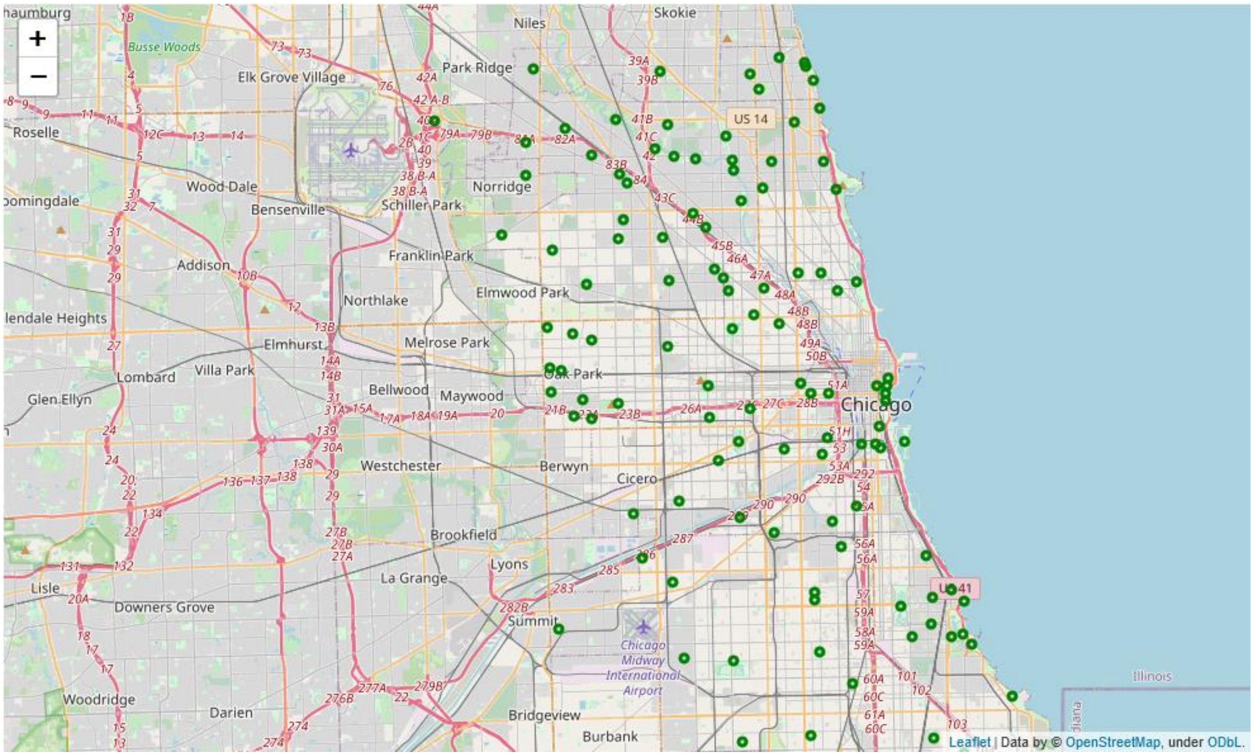

Figure 9: The public parks of Chicago, retrieved by the Foursquare API.

A heatmap of the parks was then used to obtain more insight into their distribution.

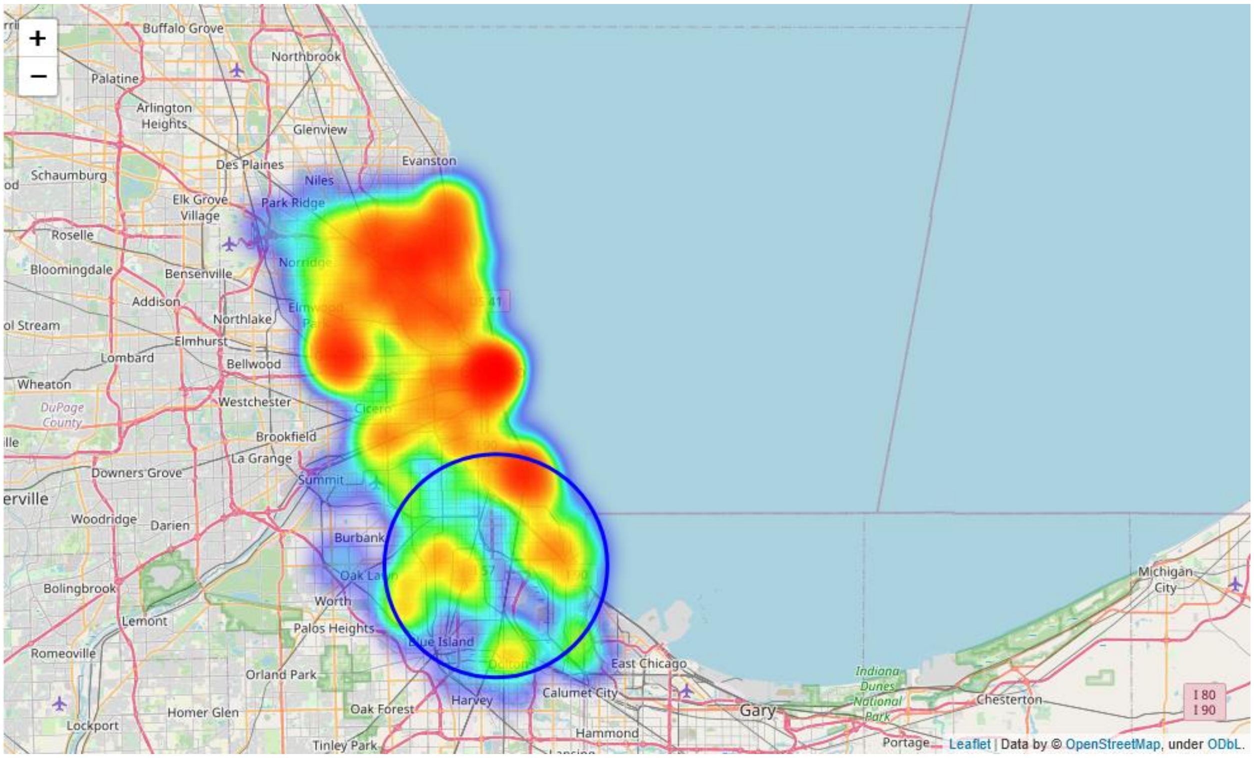

Figure 10: Heatmap of the public parks in Chicago. Areas with relatively few parks are enclosed within the blue circle.

There appear to be pockets of relatively low park density, especially in the south of Chicago.

4.3. Machine learning modeling and analysis

The search was thus restricted to the south of Chicago, centering near the I-57/I-94 intersection; the Harlan Community Academy High School was picked as the central point, and a 7.5 km radius around it was examined.

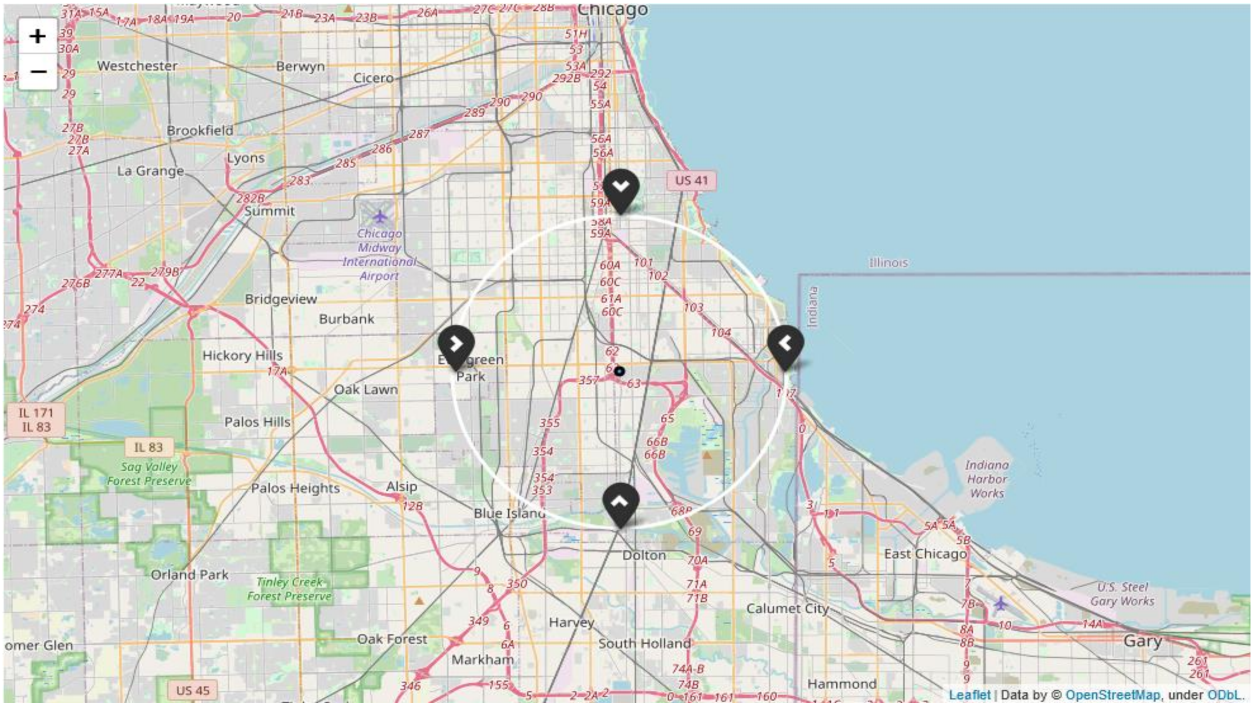

Figure 11: The region of interest of the study's search.

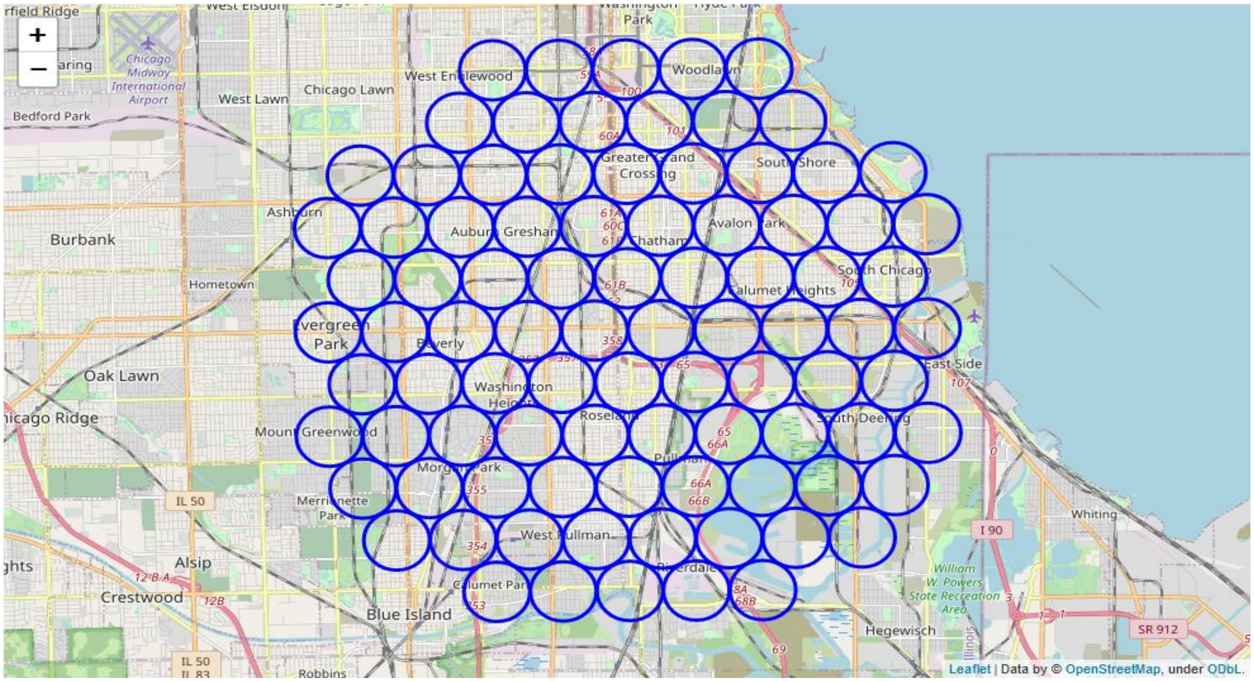

Figure 12: The hexagonal grid within the region of interest.

The approach taken was to count the number of parks in the vicinity of each grid cell, as well the distance to the closest park. It would then be possible to check if there were any cells in which there were few parks nearby, and if the closest park was, well, not so close by. All right. In order to narrow down the filtering criteria, the average number of parks (2.42) & average distance to a park (1.18 km) in a community was calculated. The filtering criteria were thus set: areas with 2 parks or less in their vicinity, and where the distance to the closest park is more than 1 km. With these restrictions, 88 areas were found with 2 parks or less in their vicinity, and 63 were found where the nearest park is more than 1 km away. The number of areas satisfying both conditions was also 63; these were the same areas where the parks were more than 1 km away. The centers of these cells are shown in green in the map below.

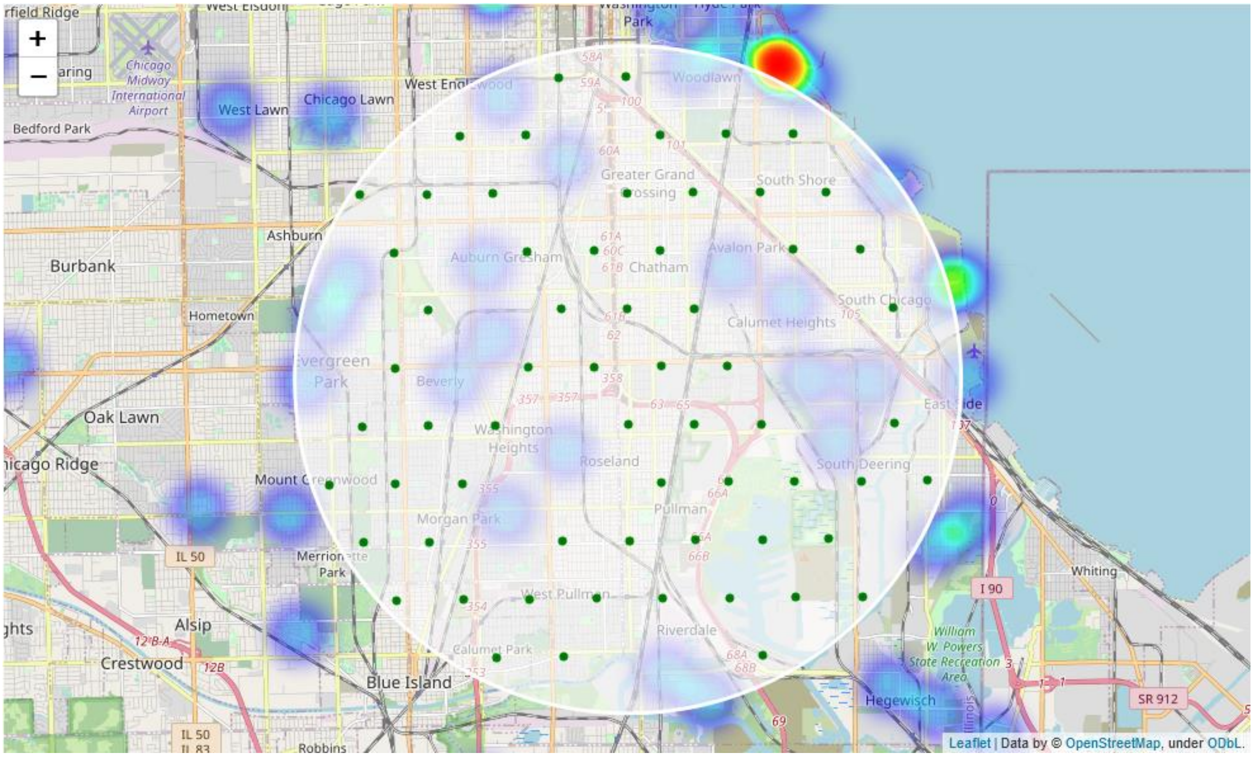

Figure 13: Candidate locations for public parks.

k-means clustering was then performed on these locations as a way to focus the search even more, and 6 clusters were found.

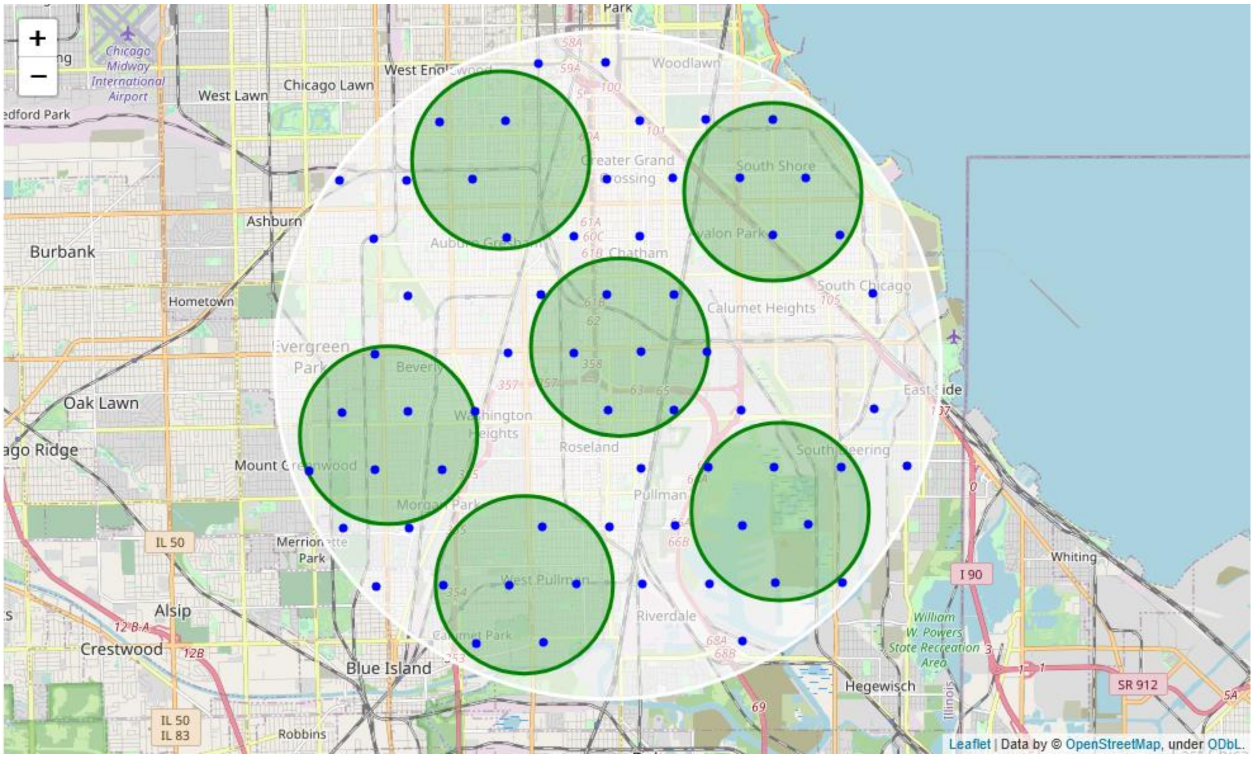

Figure 14: The 6 candidate zones returned as a result of performing k-means clustering on the candidate locations.

Finally, the addresses of the cluster centers were reverse geocoded, giving approximate locations to begin further investigation:

10411 South Claremont Avenue, Beverly, Chicago, Illinois 60643

1949-1951 East 78th Street, South Shore, Chicago, Illinois 60649

South Green Street, West Pullman, Chicago, Illinois 60827

234 East 95th Street, Roseland, Chicago, Illinois 60617

7359 South Carpenter Street, Englewood, Chicago, Illinois 60621

11555 South Stony Island Avenue, South Deering, Chicago, Illinois 60617

5. Results and discussion

The analysis has yielded insight into the factors that have the greatest effect on the economic hardship of a community, represented by the hardship index metric. Aside from what may perhaps be an obvious factor, living below the poverty line, it is observed that household density, percentage of community members not in the workforce, education level, and per capita income impact this metric strongly. By training a multiple linear regression model on data between 2008-2012 from the city of Chicago, a fairly robust predictive pipeline has been demonstrated, in spite of the small size of the dataset. As mentioned, access to details of population by community area could help in pinpointing other factors on which the hardship index in dependent. A single value of the metric was provided over the four year period for each community, resulting in the non-ideal amount of data. Access to multiple years' worth of the metric would allow for a more precise and generalized evaluation of the socio-economic conditions of a community.

Armed with the well-established fact of the beneficial effects of having parks in a community, the study then tackled the problem of finding new areas to open parks for public access. By performing a geospatial analysis of the city, it is evident that parts of the south have space for this. Dividing this area into search grids and performing k-means clustering on the results provided six candidate zones which can be investigated further. The final decision will have to take other factors into consideration: geographical suitability, real estate availability, and prices, for example.

6. Conclusion

This study had two aims: analyzing the socio-economic factors that affect the well-being of a community's members, developing a predictive model in the process, and to find areas lacking in public parks, enabling municipal powers to be proactive in providing this proven resource to their constituencies.

The result of pursuing the first aim, a linear regression model that can predict the hardship index of a community based on socio-economic statistics, can be very useful to government, for example, in order to improve their policies so that communities provide a healthy, nurturing environment for its residents. Pursuing the second aim led to an algorithmic way to find areas lacking a particular resource, providing a great starting point for action. Although parks in Chicago were used in this study, the methodology can be utilized for any public good.