Clustering Analysis of Chicago ZIP Codes by their COVID-19 Curves

Sashank Rayapudi

I. Introduction

In the wake of an unprecedented, novel global pandemic, many cities and states have been experiencing hotspots where cases are increasing at a relatively high rate. Some areas go through multiple hotspots over time, but at different times and in different waves. Each of these areas has a more nuanced story than just being classified as a 'hotspot' and can give us a more detailed picture of an area's current situation.

In this paper, I look to analyze Chicago ZIP codes based on their COVID-19 case curves. For example, curves could be labeled as 'worse second wave' or 'late first wave'. Clustering these curves can give us a story on why the curve takes on a certain shape based on certain socio-economic determinants or location. However, my motivation for clustering these curves goes beyond labeling these curves - I want to use these labels to predict when or why a curve for a ZIP code could turn into a 'hotspot' based on their socioeconomic determinants and/or location. Utilizing B-Splines, hierarchical clustering , and other data processing methods, I allow for scalability and eliminate bias when clustering the ZIP codes[1].

II. Research Statement

In an effort to analyze the similarities and differences between the Chicago ZIP codes, I proposed the following preliminary questions:

> How are Chicago ZIP codes clustered or grouped by their COVID-19 curves?
> What are the socio-economic determinants of curve shape by ZIP code, or location?

Using these questions as a template, I hope to find out why certain determinants and location result in a particular COVID curve and how these ideas can be used to implement non-pharmaceutical interventions in ZIP codes that could potentially experience hotspots.

III. Dataset and Methods

Data for this analysis was taken from 'COVID-19 Cases, Tests, and Deaths by ZIP Code' from cityofchicago.org and includes data collected over 53 weeks from March 1st, 2020 to March 6th, 2021[2]. The set includes data on weekly and cumulative cases, tests, positive tests and deaths.

Throughout the data cleaning process, I took many steps to ensure that high quality data was being produced for a proper analysis. First, I dropped the 'Unknown' ZIP code due to a lack of information and the '60666' ZIP code as it is the code for the O'Hare International Airport and case data wasn't available. Second, I dropped all of the columns except the ones I needed for my analysis - *ZIP code*, *Week Number*, and *Cases - Weekly*. Since each ZIP code is a feature in this dataset, I pivoted the table and gave each ZIP code its own column, set the Week Number as the index, and Weekly Cases as my values. I also had to reset the Week Number index since Week 10 was 03/01/2020 and I wanted to set that as my Week 1. The 'NaN' values in the dataset were defined by cityofchicago.org in this manner - "Values are removed for privacy until the cumulative total for the ZIP Code has reached 5 cases. A blank indicates a suppressed number from 0 to 4." For simplicity, I replaced each null value with a 0 as it would not mess with the shape of the curve significantly. Lastly, since I am interested in the shape of each curve and not the absolute count of cases, I had to normalize the data so that ZIP codes with different populations were treated similarly. I did this by converting weekly case counts into a percentage of the total cases in their respective ZIP codes. At the end of the data cleaning process, I was left with 58 ZIP codes over 53 weeks. For the purposes of visualizing my process, I will use Hyde Park (ZIP code 60637) case data (Figures 1 & 2).

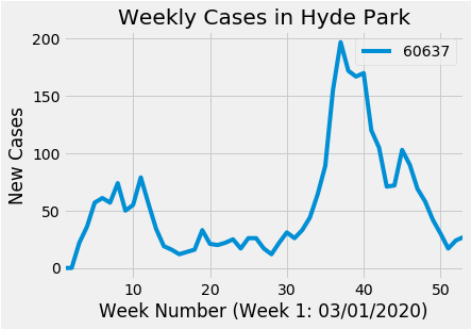

Figure 1

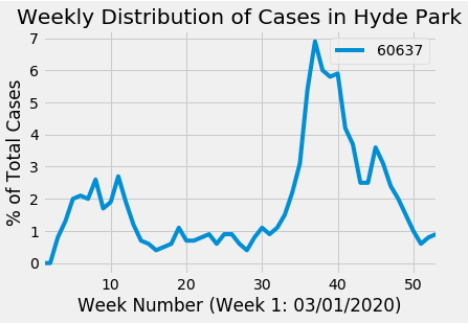

Figure 2

IV. Exploratory Analysis

In order to cluster each ZIP code by its curve shape, I needed to transform the data in such a way to describe the trend over 53 weeks. The actual data is too volatile, and a single regression line would oversimplify the COVID pattern. I opted to use B-Splines since this method could approximate non-linear functions by using a piece-wise combination of polynomials and describe the shape of each ZIP code's curve[1]. Using 12 knots, I was able to produce a B-Spline approximation and 12 coefficients for each of the 58 ZIP codes in my dataset. A few B-Splines are shown below, including Hyde Park's:

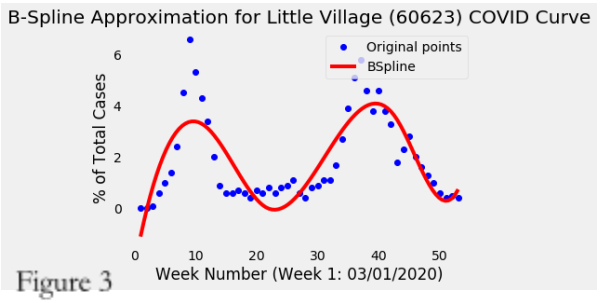

Figure 3

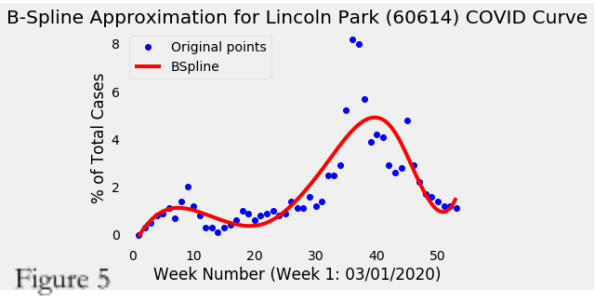

Figure 5

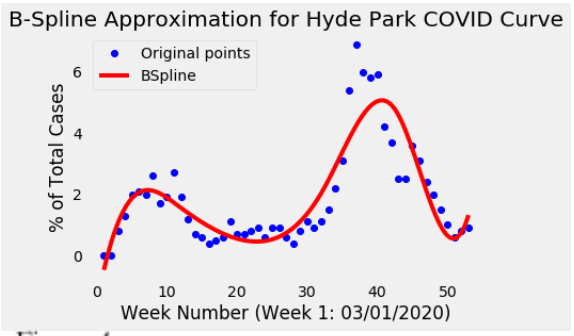

Figure 4

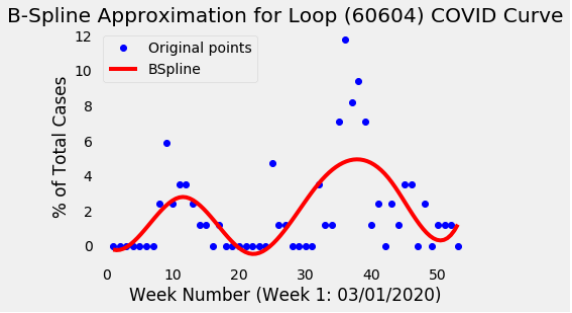

Figure 6

These curves illustrate the general trend of each ZIP code without encapsulating too much of the noise and volatility. I decided to use hierarchical clustering over K-means clustering since I had to feed 12 features, or coefficients into the model. After creating a new dataframe with 12 columns for each coefficient and 58 rows for each ZIP code, I fed the

dataframe into a hierarchical clustering model and produced the following dendrogram (Figure 7):

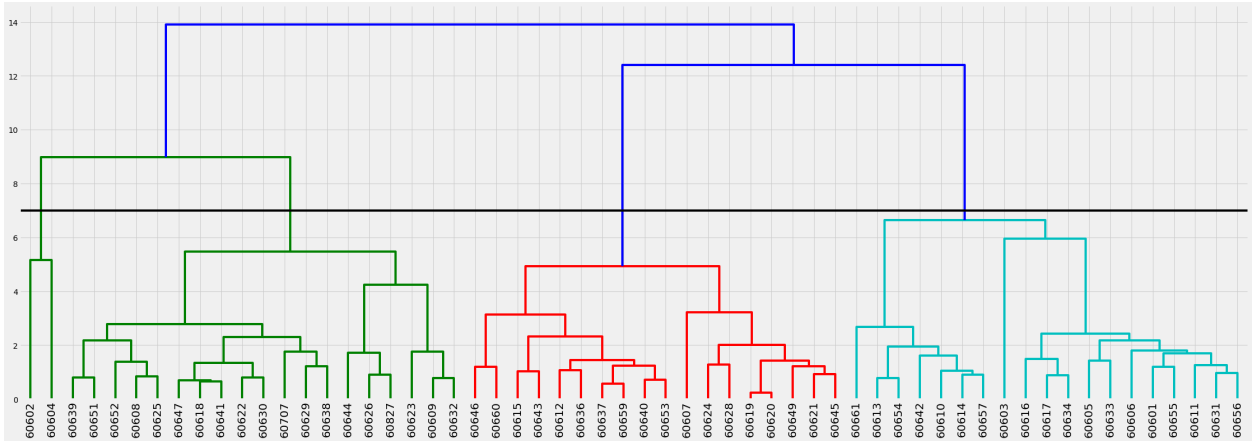

Figure 7

After visualizing the different ZIP codes based on the dendrogram, it was clear that all of the ZIP codes experienced a wave near week 10 and week 40. I came up with the following four cluster labels:

Cluster 1: 'Strong 1st and 2nd waves' (Figure 3)
Cluster 2: 'Moderate 1st wave, Strong 2nd wave' (Figure 4)
Cluster 3: 'Weak 1st wave, Strong 2nd wave' (Figure 5)
Cluster 4: 'Volatile waves' (Figure 6, 60602 and 60604 are the only ZIP codes in this cluster)

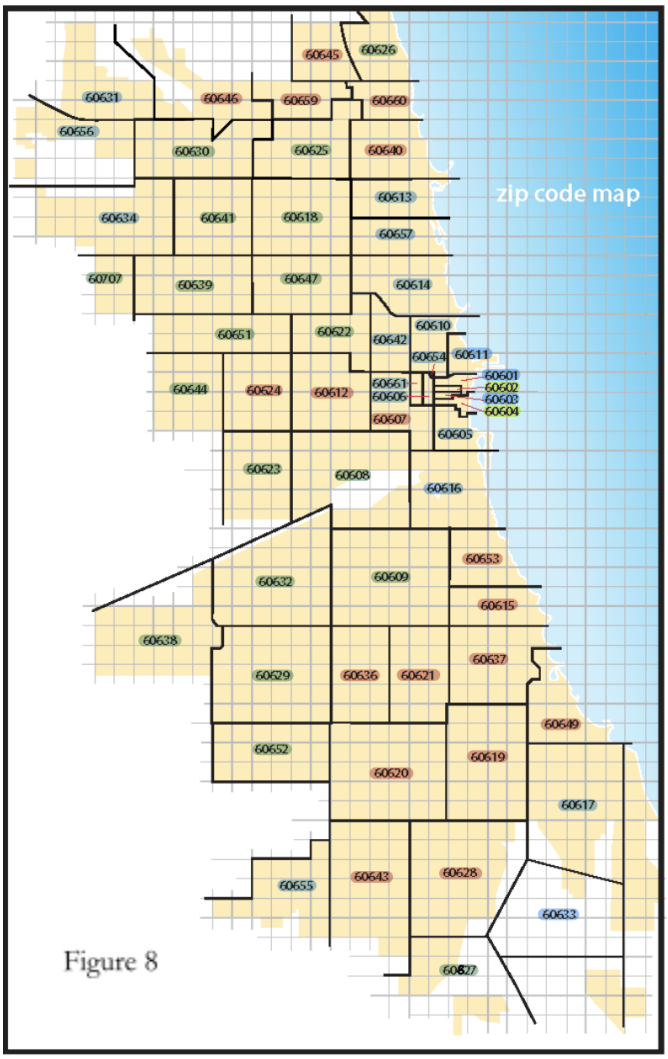

Figure 8

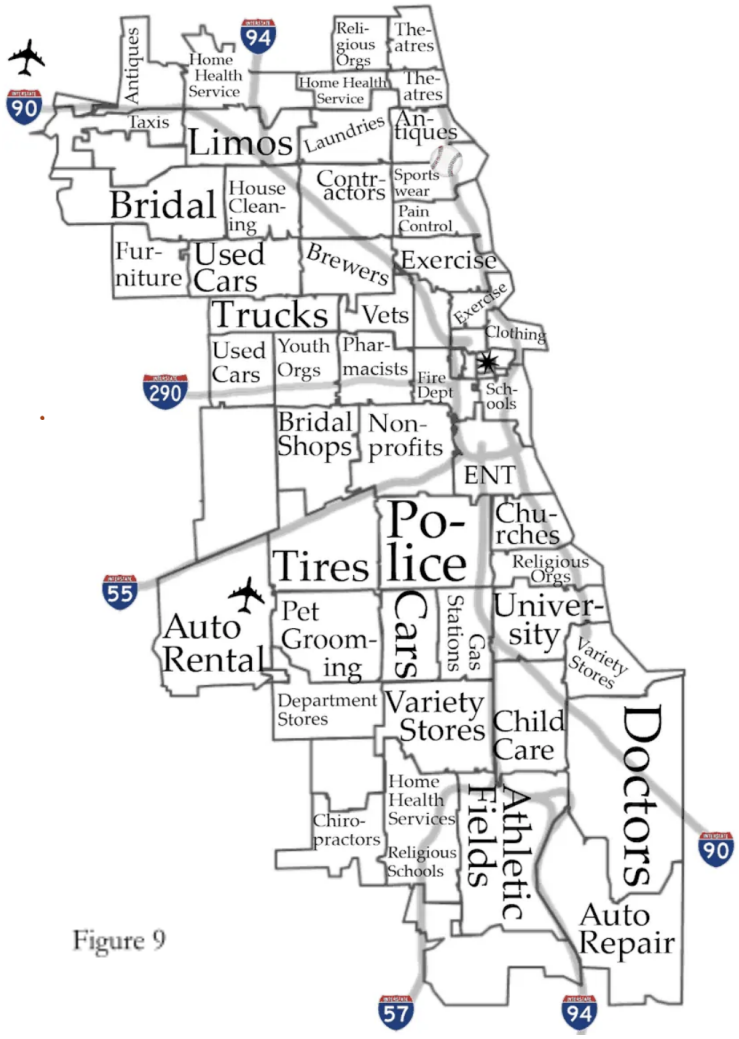

Figure 9

Surprisingly, the ZIP codes in cluster 1 (blue) with the highest population densities experienced a weak 1st wave when compared to the other clusters (Figures 8 & 9). This is likely due to the fact that metropolitan areas and hospitals enforced strict social distancing and lockdown procedures in the early stages of the pandemic. The ZIP codes in cluster 2 (red) consist of universities, religious organizations, and home health services. There was some level of social distancing enforced during week 10, but still resulted in a moderate first wave due to a lack of a total lockdown. Cluster 3 (green) consists of small businesses, car rentals, and airports. All of these businesses and services required constant contact and turnover to keep them going, and is likely the reason for a strong 1st wave. Lastly, cluster 4 (neon green) consists of COVID curves with high volatility and a number of waves due to its location in a populous environment.

V. Conclusion

Using B-Splines and hierarchical clustering, I was able to cluster Chicago ZIP codes with similar trends in COVID-19 cases. In addition to this, the unsupervised clustering model gave us a more nuanced story about certain hotspots - when the wave occurred and how intense it was. This information can be used to predict waves and implement non-pharmaceutical interventions in areas where deemed necessary. For example, in ZIP codes with small businesses, there should be more enforcement during hotspot times to prevent the spread of the virus. Individuals in these areas must be active for their businesses and services to flourish. If the conditions were to change and new patterns were to emerge from COVID or even another virus, we can utilize this information and properly prevent hotspot cases with unsupervised learning methods.

VI. References

[1]Michelen, Rory. "Using B-Splines and K-Means to Cluster Time Series." *Medium*, Towards Data Science, 22 June 2020, towardsdatascience.com/using-b-splines-and-k-means-to-cluster-time-series-16468f588ea6.

[2]Chicago, City of. "COVID-19 Cases, Tests, and Deaths by ZIP Code: City of Chicago: Data Portal." *Chicago Data Portal*, 12 Mar. 2021, data.cityofchicago.org/Health-Human-Services/COVID-19-Cases-Tests-and-Deaths-by-ZIP-Code/yhhz-zm2v.

[3]Dunlevy, John. "Classifieds." *Chicago Reader*, Chicago Reader, 11 Mar. 2021, www.chicagoreader.com/chicago/classifieds/Content?category=61309067&redirString=.

[4]"This Is How Chicago Looks According to Most Popular Businesses." *Apartment Living Tips - Apartment Tips from ApartmentGuide.com*, 18 Dec. 2018, www.apartmentguide.com/blog/chicago-most-popular-businesses-maps/.