Education Disparity in Chicago Public High Schools: A Statistical Analysis

Michael Kralis

December 2021

1 Introduction

Chicago has a history of being one of the most segregated cities in the U.S. which trickles down on the schooling system. This can be a problem when trying to provide education to all who need it. It also can help easily identify when one racial/ethnic group is being undervalued. The Chicago Public School system has a lot on their plate when trying to provide quality education to all walks of life. This begs the question, how well is this goal being achieved? This brings us to the goal of the paper. How can one use methods in statistics to visualize and test for the measure of educational disparity? The main focus will be to provide the methods to measure this disparity. In addition, we will visualize this disparity and provide meaningful interpretations of the statistical methods used.

In order to begin our descent into the Chicago public high school system, we must gather data. The data source used for this paper is from the Illinois State Board of Education. Their website contains a report card library dating back to 1996. Since the goal of this paper is to create analysis of Chicago public high schools in the present day, we will only go as far as 2015 for our data. These report cards contain data on every registered public school in Illinois. The pieces of data we are interested in are demographic data and academic data. We will then use these for our analysis going forward.

2 Clustering schools on Demographics

To start, we need to understand what data we are going to use. We will begin by selecting these factors from the report card data:

```
1. School Name
2. % Student Enrollment - White
3. % Student Enrollment - Black or African American
4. % Student Enrollment - Hispanic or Latino
5. % Student Enrollment - Asian
```

```
6. % Student Enrollment - Low Income
7. Student Attendance Rate
8. High School Dropout Rate - Total
9. High School 4-Year Graduation Rate - Total
10. % Graduates enrolled in a Postsecondary
     Institution within 12 months
11. # Student Enrollment
```

Then we will reduce our set of schools to high schools in Chicago. Now we create a K-means clustering algorithm to partition our high schools. However we will not be clustering these schools on the basis of anything but demographic indicators of the schools. In particular, we will be clustering schools based on factors 1) through 6). This means that each school in our clustering algorithm can be represented as a point in $\mathbb{R}^6$. Later we will see how these clusters have an affect on academic "success". In the context of this paper, we define academic success to be the factors 7) through 10) of our data.

In the preliminary step of the K-means algorithm we randomly select K points

$$\{p_1^{(0)}, p_2^{(0)}, ..., p_K^{(0)}\}$$

to be the starting means or centroids. Let $S \subset \mathbb{R}^6$ be the set of schools. With these random starting points, we will be able to construct a partition given by

$$\{S_1^{(0)}, S_2^{(0)}, ..., S_K^{(0)}\}$$

where

$$\bigcup_{i=1}^{K} S_i^{(0)} = S$$

and

$$S_i^{(0)} = \{s \in S : d(s, p_i^{(0)}) < d(s, p_j^{(0)}) \; \forall \; j \neq i\}$$

where the metric $d : \mathbb{R}^6 \times \mathbb{R}^6 \to \mathbb{R}$ is given by euclidean distance. The intuition behind the math here is that we drop K points in space that each have 6 randomly generated proportions for our school demographics. Then we partition schools into K subsets where the i^{th} subset, $S_i^{(0)}$, is the collection of schools that was "closest" to the i^{th} point, $p_i^{(0)}$. The superscript $^{(0)}$ denotes that this is the preliminary step. Moving forward, we recalculate our K points,

$$\{p_1^{(1)}, p_2^{(1)}, ..., p_K^{(1)}\}$$

where

$$p_i^{(1)} = \frac{1}{|S_i^{(0)}|} \sum_{s \in S_i^{(0)}} s$$

Which means that i^{th} point, $p_i^{(1)}$, is now the mean demographics of the schools from i^{th} subset, $S_i^{(0)}$, from the previous step. Then we create another partition given by

$$\{S_1^{(1)}, S_2^{(1)}, ..., S_K^{(1)}\}$$

using the same strategy as in the preliminary step. This algorithm will repeat in this fashion until the means converge to a point where the clusters don't change on the next iteration. Thus we will have our final clusters given by

$$\{S_1, S_2, ..., S_K\}$$

Once we have our clusters, we can now label each of the schools with the cluster it belongs in. For each cluster, S_i, we can give summary statistics. We present the implementation with $K = 3$ in figures 1, 2, and 3.

We notice from figures 1, 2, and 3 that cluster 0 was predominately black or African American, cluster 1 was predominately Hispanic, and cluster 2 was a mixed demographic cluster with a high amount of non-low-income students. These results, intuitively, give rise to some powerful concerns. The important aspect of the construction of these clusters is that the algorithm only used demographic factors to partition the schools yet was able to spot a group of schools with significantly higher academic "success" in cluster 2. Now it must be understood that how we are defining academic success is restricted to only 4 pieces of data. In reality, basing an individual student's academic success on these 4 factors would be preposterous. However, making a judgement about an entire school on these factors makes more sense.

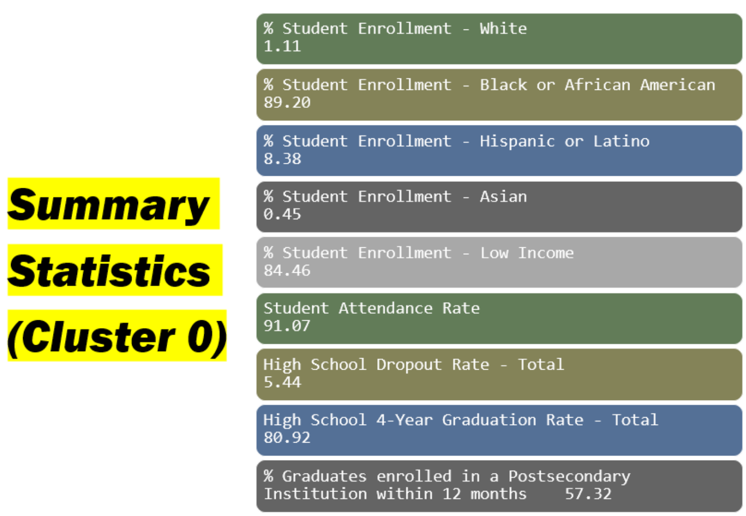

Figure 1: Summary Stats of Cluster 0

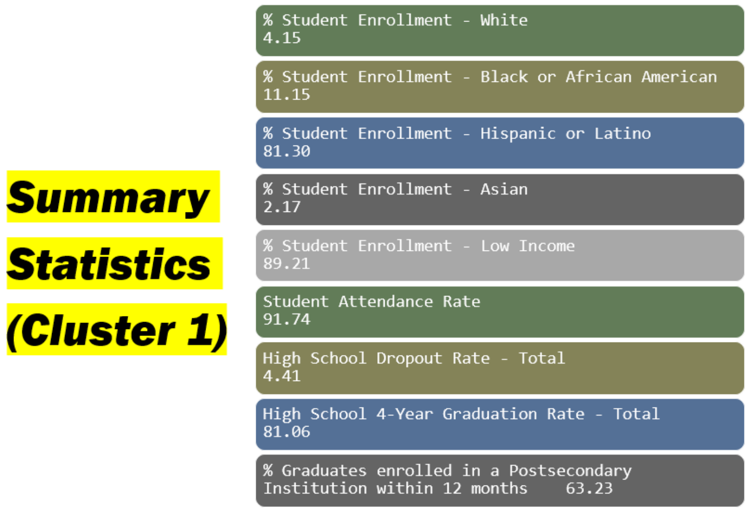

Figure 2: Summary Stats of Cluster 1

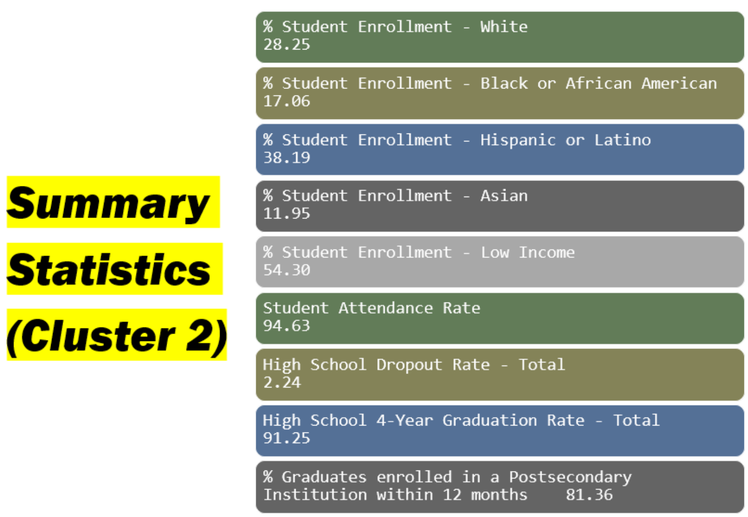

Figure 3: Summary Stats of Cluster 2

These clusters allow us to analyze even more about the school system in Chicago. They allow us to prove even more about how the segregation of Chicago has affected the school system. By plotting these clusters, we can create a visual of this effect.

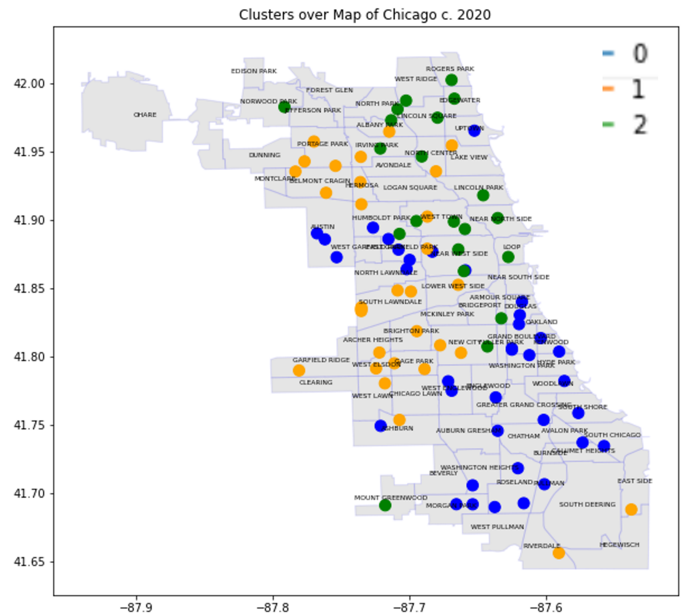

Figure 4: Mapping clusters over Chicago neighborhoods

It is clear to see that not only do these clusters separate schools efficiently on the basis of academic success (without even trying to) they also seem to be an efficient tool in splitting up the areas of Chicago. This can be problematic for students of color to receive a quality education.

3 Trends of clusters over time

The natural question to ask is, how do these disparities change over time? Or in the context of our previous work, how do these clusters change over time? The problem in answering such a question is that the K-means algorithm is a random algorithm. This means that if we cluster schools in the years 2019 and 2020, the clusters may look very different. However we can fix this problem by creating a stable matching algorithm.

The idea of a matching is a rigorously studied concept in discrete mathematics. However, the concept is rather simple. In the context of our problem, we have a bipartite graph as follows

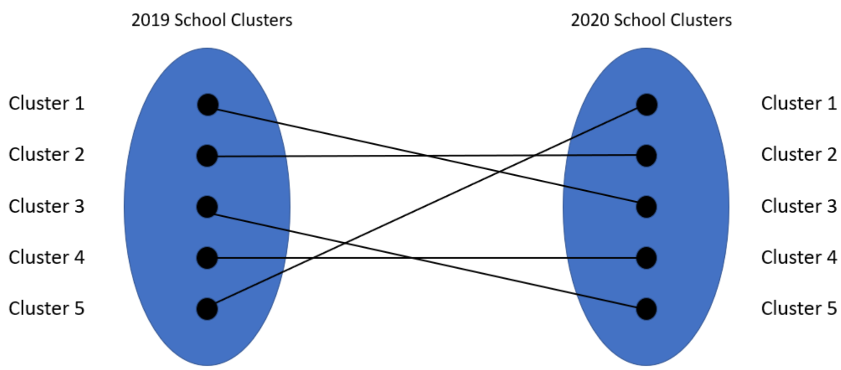

Figure 5: Implementation of stable matching

The lines across are the edges of the graph and they represent an element of the matching. For example, what these lines represent for this problem is that Cluster 1 is most like Cluster 3 and so on. The problem that arises is that we need to introduce cluster preferences. When introducing preferences into our graphical model, we step into the world of stable matchings.

In 1962, David Gale and Lloyd Shapley proved that for any set of preferences on a bipartite graph there exists a stable matching. It is also worth noting that in 1984 Alvin E. Roth observed that the same algorithm had been used for practical purposes in the early 1950s. The algorithm's history dates back to matching medical students to hospitals for their residency. Each hospital and student had a list of preferences so how do we make all parties happy? The algorithm is presented below

Algorithm
Inputs: List of clusters A, list of clusters B, set of preferences for each cluster
Output: Stable Matching

Initialize all $a \in A$ and $b \in B$ to free
while $\exists$ free cluster a who has a cluster b to match to, do
    b := first cluster on a's list to whom a has not yet matched to
    if $\exists$ some pair (a', b) then
        if b prefers a to a' then
            a' becomes free
            (a, b) become matched
        end if
    else
        (a, b) become matched
    end if
repeat

First of all, what is a stable matching? In the context of the medical students, say Hospital A wanted medical student B and vice versa. However, in the matching, medical student B was matched to Hospital B and Hospital A was matched with medical student A. This creates an unstable edge and thus the matching would be unstable. A stable matching is a matching where this situation does not occur. It is worth noting that even in a stable matching a hospital or medical student may not be matched with its preferred choice. However, that choice will not prefer it more than what it is already currently matched to.

Why do we need a stable matching in the first place? The purpose of this paper is to recreate results for any number of clusters. However, if the cluster size goes beyond 4, the matching algorithm can no longer be a greedy one. There will have to be clusters that don't get matched with their number one choice in order for a stable matching to be created. The key in choosing the cluster size of 4 is to avoid this problem. This way, we can explicitly show how these clusters change over time.

The next question to ask is, why don't we just save the schools clustered in the first year and track them over time? The problem with this approach is that a school's demographic and academic factors may change over time. They may change so much that the schools should be reclassified. By running a clustering algorithm each year and then mapping those clusters back, we can accurately assess the changes over time.

Now that we established the why, we will get down to the results. One factor we are interested in viewing is the proportion of low income students over time. From figure 6, we can see that the gap between clusters is immediate. Cluster 2 has significantly less low income students as compared to clusters 0 and 1. Not only this but the gap seems to be getting larger each year. Cluster 0 went from 91 to 85 percent low income students and cluster 1 went from 95 to 89 percent low income students while cluster 2 went from 68 to 55 percent low income students. This relative change is 3 times larger than in clusters 0 and 1. What this means is that the K-means algorithm is identifying that the schools with wealthy student bodies are getting wealthier over time.

Now we will turn our attention to the college enrollment rate of each cluster. From figure 7, like in figure 6, the differences are immediate. Cluster 2 has a much higher proportion of students enrolling in college compared to the other clusters. We can also see how the COVID-19 pandemic has affected schools. Over time we have seen more graduates move on to postsecondary institutions but in the 2019-2020 school year this growth comes to a screeching halt. More specifically, we can see the effect that the pandemic has on schools that are predominantly black or African-American students with cluster 0 having the largest drop in postsecondary institution enrollment.

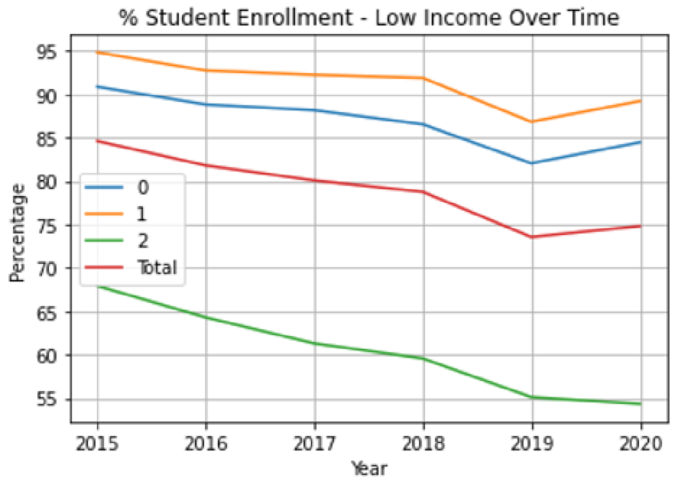

Figure 6: Proportion of Low Income students over time

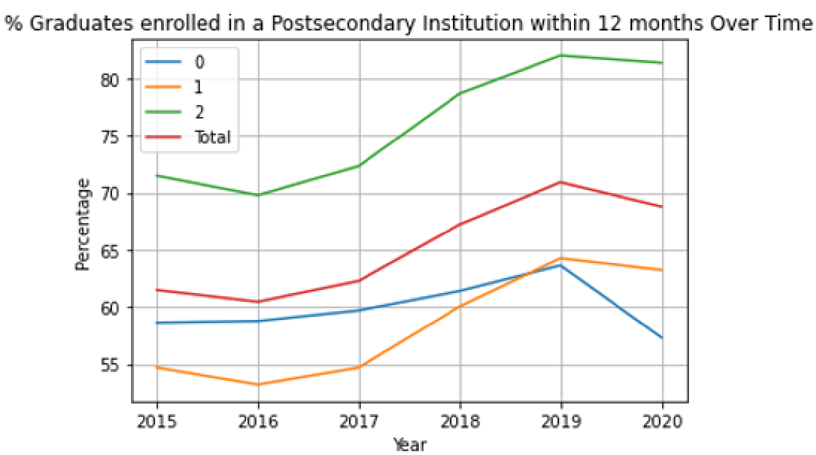

Figure 7: Proportion of students attending college over time

4 Traditional approach

In this section we will take a look at these factors from a more straightforward standpoint. What can we see about the correlations between all these factors?

Index	ıt Enrollmen	ent - Black c	ıllment - His	ıt Enrollmer	Enrollment - L	ıt Attendanc	ol Dropout I	'ear Graduat	stsecondary
% Student Enrollment - White	1	-0.459694	0.0413344	0.670453	-0.881595	0.485443	-0.31436	0.313289	0.507657
% Student Enrollment - Black or Afric...	-0.459694	1	-0.897257	-0.420106	0.341374	-0.430198	0.269335	-0.232523	-0.432968
% Student Enrollment - Hispanic or La...	0.0413344	-0.897257	1	0.0520113	0.0392081	0.236293	-0.145118	0.101725	0.22719
% Student Enrollment - Asian	0.670453	-0.420106	0.0520113	1	-0.569967	0.413457	-0.250182	0.25526	0.434239
% Student Enrollment - Low Income	-0.881595	0.341374	0.0392081	-0.569967	1	-0.642394	0.447039	-0.483383	-0.644065
Student Attendance Rate	0.485443	-0.430198	0.236293	0.413457	-0.642394	1	-0.722903	0.784383	0.726142
High School Dropout Rate - Total	-0.31436	0.269335	-0.145118	-0.250182	0.447039	-0.722903	1	-0.741412	-0.547078
High School 4-Year Graduation Rate -...	0.313289	-0.232523	0.101725	0.25526	-0.483383	0.784383	-0.741412	1	0.771031
% Graduates enrolled in a Postsecond...	0.507657	-0.432968	0.22719	0.434239	-0.644065	0.726142	-0.547078	0.771031	1

Figure 8: Correlation Matrix of Factors

Now the important part of this matrix to focus on are the correlations between are demographic factors and our academic factors. This data is presented below in figure 9. The first thing to notice is the disadvantage that schools with a large amount of low income students have when it comes to academic factors. Across the board, low income students is the most strongly correlated factor with respect to each academic factors. We can also see the disadvantage to schools that are predominantly black or African American.

Index	Student Enrollment - Whi	nrollment - Black or Africa	nt Enrollment - Hispanic c	Student Enrollment - As	dent Enrollment - Low In
Student Attendance Rate	0.485443	-0.430198	0.236293	0.413457	-0.642394
High School Dropout Rate - Total	-0.31436	0.269335	-0.145118	-0.250182	0.447039
High School 4-Year Graduation Rate -...	0.313289	-0.232523	0.101725	0.25526	-0.483383
% Graduates enrolled in a Postsecond...	0.507657	-0.432968	0.22719	0.434239	-0.644065

Figure 9: Correlations of academic factors against demographic factors

Referring back to figure 8 we notice a shocking discovery. The correlation between the proportion of low income students of a school and the proportion of graduates that enroll in postsecondary institution is about -0.644 while the correlation between the high school graduation rate and proportion of graduates that enroll in postsecondary institution is about 0.771. As an example as to why this is so problematic, let us consider someone trying to guess the academic profile of a school. Perhaps this person was limited to asking questions only about demographics. What these numbers tell us is that asking for the proportion of low income students is almost as good as asking the high school graduation rate when trying to predict the postsecondary institution enrollment! In short this person would most likely say, "tell us how wealthy the student body is and we will give a good ballpark estimate for its academics factors".

5 Potential for further analysis

The hidden goal of this project is to recreate new research projects for others in this scope. The proposed choice of analysis is not the only choice. The repository located at https://github.com/mkralis123/SoReMo contains scripts to recreate results. For example, what if the number of clusters chosen was 4 instead of 3? What if we chose to look at attendance rate over time instead? As we see below, figure 10 shows the main file called "clustering_ISBE.py". The file that contains the user-built functions is called "functions.py". Below we can see an implementation of the code. Note that there are some dependencies on unique packages in python such as geopandas.

```python
from mappingClusters import mapClusters
from functions import GaleShapleyAlgo, clusterDemographics, plotData, summaryStats

n_clusters = 3
folder = "ISBE_ReportCards/"
years = [2015,2016,2017,2018,2019,2020]
field_names = ['School Name',
               '% Student Enrollment - White',
               '% Student Enrollment - Black or African American',
               '% Student Enrollment - Hispanic or Latino',
               '% Student Enrollment - Asian',
               '% Student Enrollment - Low Income',
               'Student Attendance Rate',
               'High School Dropout Rate - Total',
               'High School 4-Year Graduation Rate - Total',
               '% Graduates enrolled in a Postsecondary Institution within 12 months',
               '# Student Enrollment']
feature_index = 9
cluster_indexnum = 2
year_index = -1


#############RUN CLUSTER ALGO#####################
demographic_dfs, clusters_over_time = clusterDemographics(folder, years, field_names, n_clusters)


##############GALE SHAPLEY ALGORITHM##############
demographic_dfs,clusters_over_time = GaleShapleyAlgo(clusters_over_time, demographic_dfs, years, n_clusters)


#################FOR PLOTS########################
data_for_plot = plotData(field_names[feature_index], demographic_dfs, clusters_over_time, years, n_clusters)


################PRINT SUMMARY STATS###############
summaryStats(cluster_indexnum, clusters_over_time, years)


###################DISPLAY GEOMAP#################
mapClusters(n_clusters)


###################GET CORRELATION MATRIX#########
correlation_matrix = demographic_dfs[year_index].filter(field_names[1:-1]).corr()
updated_matrix = correlation_matrix[field_names[1:6]].iloc[5:]
```

Figure 10: Implementation of Python code

6 Conclusions

To summarize our journey in uncovering the disparities present in the Chicago public high school system we will start from the beginning. We started this to analyze the disparity for a single school year, 2019-2020. We were able to immediately see differences in academic factors even though the clustering algorithm knew nothing about them. After plotting these clusters over the city of Chicago, we were able to understand the connection between the segregation of Chicago in the school system. Next we decided to observe how these clusters changed in time by making use of a graphical model and matching these clusters over time. Again, we were not only able to see clear differences year to year but we were able to see that these disparities have grown over time. Lastly, we observed the correlation matrix between all these factors and identified key values to illustrate this disparity.

In conclusion, there is much work to be done when it comes to Chicago public high schools. It is not sure whether or not the school system itself can make up for these inequalities but proof of that was not the goal of this paper. What we have shown is that it exists and it is strong. Low income students and students of color are at disadvantages when it comes to the academic factors selected.