# Residential Segregation and Educational Segregation

## 1. Research Questions, Hypothesis and Variables

In this research, I am trying to find out the relationship between residential segregation and educational segregation in New York City.

I presume that residential segregation (on the basis of racial segregation in geographic level) can indicates the educational segregation (high-ranking schools clustering in certain places), to be more specific, the higher the segregation level one place has, the lower the education quality this place will have.

In this research, I choose zip code area as my analytical units. This choice is based on these reasons: 1) compared to census tracts, residents in every neighborhood know their zip code better, and this advantage in subjective perception can give this research a concrete sense. 2) compared to school districts, the data for zip code areas are more feasible and the units are smaller, which can help us avoid modifiable areal unit problem.

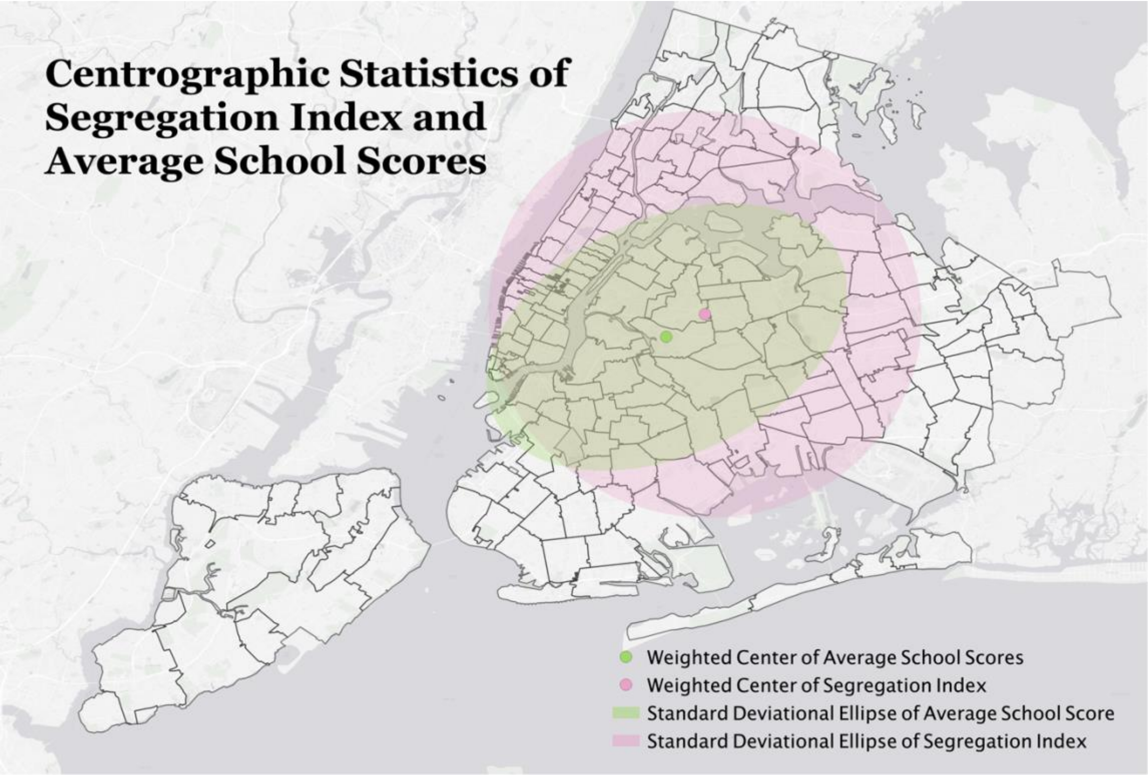
To testify the validity of my hypothesis, I, first of all, use the racial distribution in every zip code area to calculate the segregation index for each zone. Therefore, the residential segregation is represented by this index as my major independent variable. In terms of the dependent variable, I use the data from a website ranking all the high schools in New York City in 2019. The scores for these high schools indicates the education quality of them.

In terms of my control variable, I introduce income, age, sex ratio and the percentage of school-age students in each zip code area. I do not include race in this set mainly because I have used it to calculate the segregation index, so to avoid multicollinearity, I decide not to include them again.

The median income for each zip code area is the major control variable here, which means that income might be the reason for educational segregation instead of residential segregation. The mechanism behind this competing hypothesis is that the richer a place is, people in this place might devote more into local education development, and therefore has a better quality of education. So if income is the reason for the education inequality in space level, then my hypothesis will be falsified.

## 2. Descriptive Bivariate Analysis

As mentioned above, my dependent variable of interest is the average school scores in New York City zip code areas, and my independent variable of interest is the segregation index for each zip code area. In this section, I will show some graphs to depict the basic geographic information about these two variables, as well as their spatial relationship.


Centrographic Statistics of Segregation Index and Average School Scores
● Weighted Center of Average School Scores
● Weighted Center of Segregation Index
🟩 Standard Deviational Ellipse of Average School Score
🟪 Standard Deviational Ellipse of Segregation Index

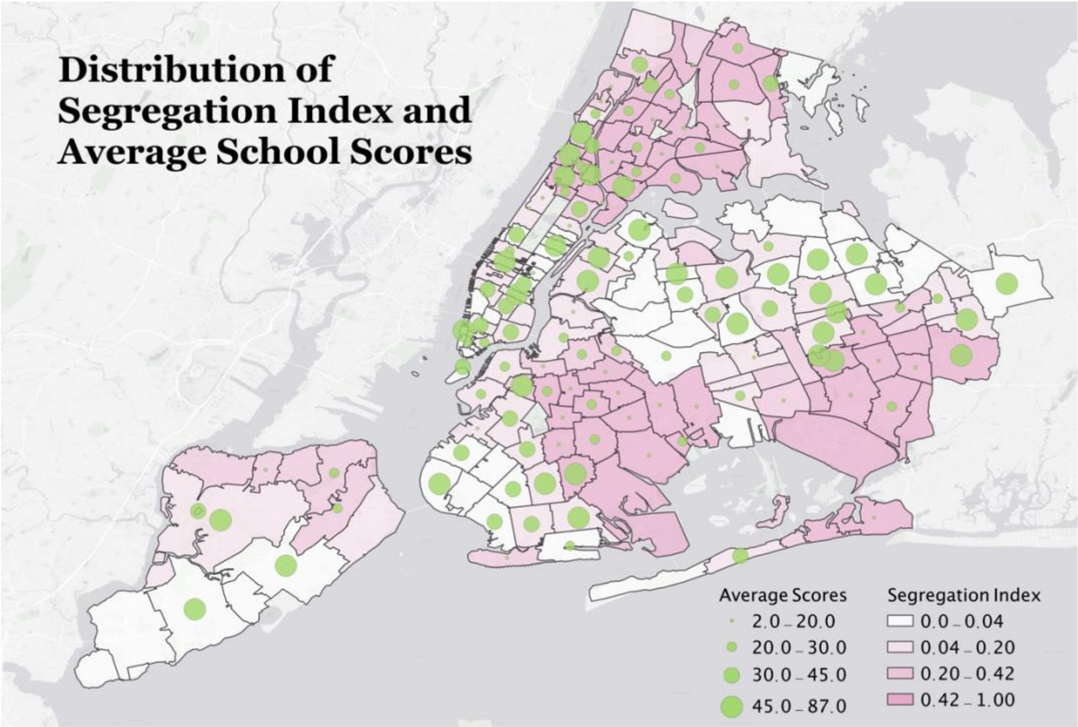
Source: U.S. Census Bureau; Schooldigger

*Figure 1 Centrographic Statistics of Segregation Index and Average School Scores*

In Fig.1, the distributions of these two variables weighted by themselves are shown. The blue dot and ellipse are the weighted distribution of average school score in each zip code zone, while the red dot and ellipse are the weighted distribution of segregation index.

The comparison between these two mean centers shows that the greatly segregated areas are relatively located in the Northeastern area of New York City, while the schools with higher scores are more likely sit in the Southwest of the city.

In terms of the standard deviational ellipse, they are totally different in terms of shapes. While the shape of the standard deviational ellipse for segregation index is more like a circle, the one for the scores is a narrow ellipse, which is more consistent with the shape of New York City. Therefore, the school scores are more widely separated in the direction from northeast to southwest; by contrast, the spread of the segregation index on the map is equal from northeast to southwest as from northwest to southeast, but if we take the shape of New York City into consideration, then we can conclude that the data for segregation index in the direction of northwest to southeast are more disperse.

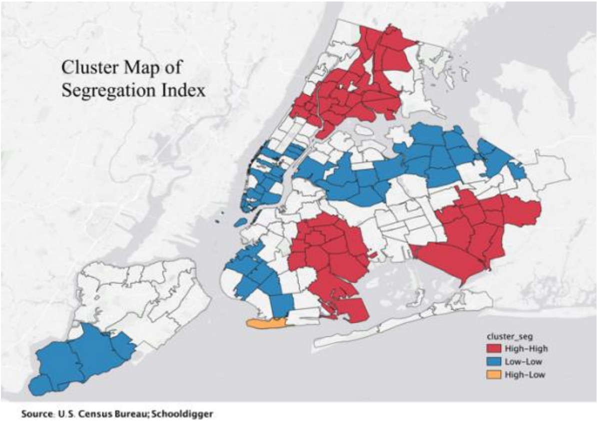


Source: U.S. Census Bureau; Schooldigger

*Figure 2 Distribution of Segregation Index and Average School Scores*

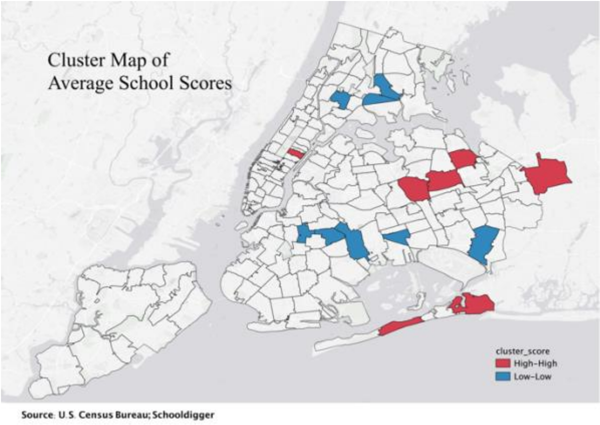
In Fig2, we can have a rough image of the relationship between residential segregation and areal educational quality (average school score for each zip code area). Generally speaking, the place with higher segregation level is less likely to have better schools, as we can witness in this graph that areas with deep orange (meaning highly segregated) are usually with small pink dots (lower ranked schools). For example, in the east side of Brooklyn and the south side of Queens, despite few areas with the inverse pattern. Furth more, areas having light orange filling color or even white are normally of larger pink dots, like the downtown area of Manhattan, upper side of Queens and the west corner of Brooklyn.

Of course, there are some areas showing the opposite pattern, which can falsify my hypothesis, like the uptown Manhattan and most areas in Bronx.



Source: U.S. Census Bureau; Schooldigger

*Figure 3 Cluster Map of Segregation Index*



Source: U.S. Census Bureau; Schooldigger

*Figure 4 Cluster Map of Average School Scores*

Fig3 and Fig4 are the clustering maps for my major variables. As there is some zip code areas missing school score data, the clustering map for it is not perfect. But by comparing these two maps together, we can roughly see that places with the segregation index having high-high value clusters have no pattern of high school scores clusters, and several of them even have a low-low value of score clusters (for example, in upper Brooklyn, west Queens and west Bronx). even just, and vice versa, the high-high value clusters of school scores are always appearing in places where the low-low value of segregation index clustered or no significant clusters. This is another approximation of the negative relationship between residential segregation and educational quality.

## 3. Regression Analysis

### 3.1 Classic OLS Model

*Table 1 Classic OLS Model Result*

| | Average School Score |
|---|---|
| Segregation index | -15.9* |
| | (6.4) |
| Log(income) | 3.5 |
| | (4.5) |
| Sex Ratio | 0.3 |
| | (0.2) |
| Median Age | 0.6 |
| | (0.4) |
| Percentage Under 18 | 0.2 |
| | (0.4) |
| Constant | -41.7 |
| | (61.4) |
| Observations | 124 |
| $R^2$ | 0.19 |
| Adjusted $R^2$ | 0.15 |
| Log Likelihood | 515.3 |
| Residual Std. Error | 15.8 (df = 118) |
| F Statistic | 5.4*** (df = 5; 118) |
| *Notes:* | *P < .05; **P < .01; ***P < .001 |

In the result generated from the naive OLS model (Table.1), we can find that the coefficient of segregation index is about -16, and the figure is highly statistically significant, meaning that when there is 1 score increase in one place's segregation index (indicating it become more segregated), there will be around 16 less score in the average score in school(s) in this area, when net of the effect of areal income, median age, percentage of school children and sex ratio. Therefore, we can say there is a negative relationship between segregation and school test score.

In terms of the coefficient for log(income), first of all, I use log(income) instead of the absolute income because this variable is not normally distributed. To avoid the coefficient being distracted by outliners, I use log(income) as control variable. The coefficient of this control variable is approximately 3.5, representing a positive relationship between the median income of one place and the test score of the school(s) in the same area. However, this coefficient is not statistically significant, so the coefficient has no difference from 0 — this positive relationship between income and score may not exist.

There is also no relationship between the other three control variables and the dependent variable -- school scores, as all the other coefficients are not significant.

### 3.2 Relative Diagnostics

*Table 2 Diagnostics for Spatial Dependence*

| Test | MI/DF |
|---|---|
| Moran’s I (error) | -0.15** |
| Lagrange Multiplier (lag) | 1* |
| Robust LM (lag) | 1 |
| Lagrange Multiplier (error) | 1* |
| Robust LM (error) | 1 |
| Lagrange Multiplier (SARMA) | 2 |
| *Notes:* | *P < .05; **P < .01; ***P < .001 |

The Moran’s I score in my classic OLS model is 0.15 (Table.2), and it is highly significant, thus rejecting the null hypothesis that the residuals are independently distributed. Therefore, I have to adopt either spatial error model or spatial lag model to deal with the violation of the BLUE assumption in this classic OLS model.

As for the diagnostics for spatial error and spatial lag (Table.2), both the simple LM tests for lag and for error are statistically significant, meaning there exists spatial dependence. However, the dilemma here is that both robust LM test for error and for lag are not significant, therefore, I felt lost when choosing one model over the other. I will just try both of them and see if there is any huge difference.

Since the probability of robust LM error test is less than the that of robust LM lag test. I will, therefore, try the spatial error model first.

### 3.3 Spatial Models

First of all, I chose rook contiguity weights to create the weight matrix.

Because, first of all, the contiguity weights can help us see the spatial dependence or relationship between areas sharing the same boundary. Additionally, I chose rook contiguity weights, instead of queen contiguity weights, to take how many proportion of the boundary two areas share into consideration. So that it can distinguish the relationships of areas share little boundary and areas share more boundary.

In my case, if people live in two places sharing more boundary, they are more likely to communicate with each other, share information, and affect each other's idea over children education, it will also affect their attitude towards people from other ethnic groups. Therefore, using rook contiguity weights makes sense in this case.

As the result of the spatial error model (Table.3) shows, the negative relationship between segregation and score still exists and it is still significant, with the coefficient for the segregation index are almost the same (around -16), so we can still state that the higher one area is segregated, the lower the education quality schools in this area will have, even when we control the effect of income, age, proportion of school children, and sex ratio.

This is similar for coefficient for log(income) in this model. The positive coefficient for income does not change, and it is still not statistically significant, despite a trivial increase in the absolute figure from coefficient in the OLS model to the one in the spatial error model.

*Table 3 Spatial Error Model Result*

| | Average School Score |
|---|---|
| Lambda | 0.3* |
| | (0.1) |
| Segregation index | -15.9* |
| | (7.1) |
| Log(income) | 3.7 |
| | (4.8) |
| Sex Ratio | 0.1 |
| | (0.2) |
| Median Age | 0.4 |
| | (0.4) |
| Percentage Under 18 | 0.2 |
| | (0.4) |
| Constant | -31.8 |
| | (65.0) |
| Observations | 124 |
| $R_2$ | 0.23 |
| Log Likelihood | -512.9 |
| Residual Std. Error | 15.0 (df = 118) |
| *Notes:* | *P < .05; **P < .01; ***P < .001 |

Lambda in this model means the spatial correlated error. In this model, the coefficient of Lambda is positive and is statistically significant, therefore the effect of the spatial autocorrelation is controlled, and the fit of the model is improved.

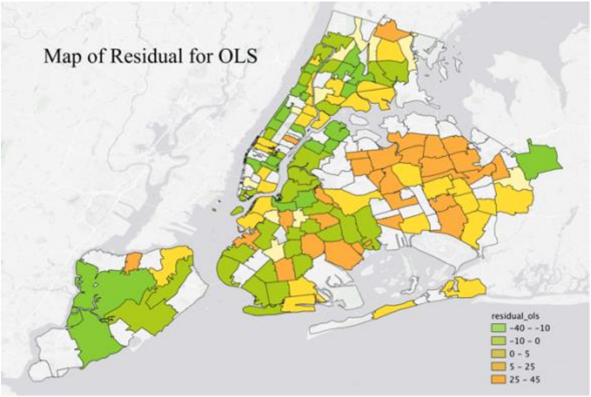
I also tried the spatial lag model (Table.4), the results of my major variable and control variable are similar with the ones from spatial error model, even though there is a drop between the size of the coefficient for segregation index (from -16 to -12). The negative relationship between segregation index and scores, and the null relationship between all the control variables and my dependent variables are all stay the same. So our conclusion stay robust among different models.

As for the correction of auto-correlation from OLS model to the Spatial Error Model, the comparison between the maps of residuals in these two models is shown in Fig.5 and Fig.6.

By comparing them, we can see that the auto-correlation error term is largely modified when we adopt spatial error model.

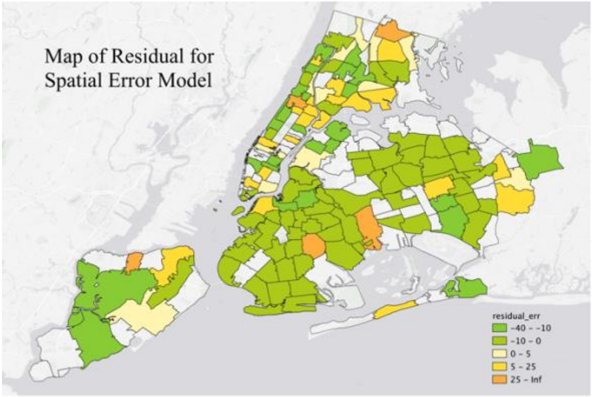
*Table 4 Spatial Lag Model Result*

| | Average School Score |
|---|---|
| Weighted average score | 0.2* |
| | (0.1) |
| Segregation index | -12.9* |
| | (6.2) |
| Log(income) | 3.4 |
| | (4.3) |
| Sex Ratio | 0.1 |
| | (0.2) |
| Median Age | 0.5 |
| | (0.4) |
| Percentage Under 18 | 0.2 |
| | (0.3) |
| Constant | -42.0 |
| | (58.2) |
| Observations | 124 |
| $R^2$ | 0.23 |
| Log Likelihood | -512.7 |
| Residual Std. Error | 15.0 (df = 117) |
| *Notes:* | *P < .05; **P < .01; ***P < .001 |



Source: U.S. Census Bureau; Schooldigger

*Figure 5 Map of Residual for OLS*



Source: U.S. Census Bureau; Schooldigger

*Figure 6 Map of Residual for Spatial Error Model*