Neighborhood Disadvantage and High School Dropout
Methods and Initial Results

Lerong Wang

May 9, 2018

1 Research Question

My research question is: To what extent do neighborhood disadvantages affect high school dropout rates? I plan to do (1) a spatial analysis of the high school dropout rate in NYC neighborhoods and (2) model high school dropout rate using neighborhood characteristics.

2 Data

2.1 Data Source

- Education and socio-demographic characteristics for New York City Census Tracts Data come from American Community Survey 2008-2012, US Census Bureau. I downloaded a compiled version from https://geodacenter.github.io/data-and-lab/. Neighborhood characteristics are contained in this data and high school dropout rate is computed using the variables contained in this data.

- NYC Neighborhood Tabulation Area Shapefile Data. Downloaded from https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page. The shapefile provided a boundary file for neighborhood area so that we can create a base map for the NYC Neighborhood Tabulation Area.

2.2 Data Processing

A list raw variables from the data that will be used for this project is provided below. Each observation in the original data corresponds to one census tract in New York City. There are 2166 observations in total. I performed spatial aggregation in GeoDa and combined the individual observations within the same neighborhood tabulation area (have same ntacode) and aggregated the sum of the each variable (male16to19, maledrop, female16to19, popunemplp, popinlabou, poptot, households, withpubass, poor and so on). Median household income is aggregated by computing the average median household income of different census tracts within the same neighborhood tabulation area.

high school dropout rate is calculated using

$$\frac{\text{maledrop} + \text{femaledrop}}{\text{male16to19} + \text{female16to19}}$$

unemployment rate is calculated using

$$\frac{\text{popunemplo}}{\text{popinlabou}}$$

poverty rate is calculated using

$$\frac{\text{poor}}{\text{poptot}}$$

proportion of households with public assistance income is calculated using

$$\frac{\text{withpubass}}{\text{households}}$$

The proportions of female population, male population, and people of different races are calculated separately using

$$\frac{\text{variable}}{\text{poptot}}$$

.

Table 1: List of Raw Variables

Variable Name	Desciptions
ntaname	New York City NTA (Neighborhood Tabulation Area) name
ntacode	Code associatd with the NTA
male16to19	Male Civilian Population 16 To 19 Years
maledrop	Male Not high school graduate, not enrolled (dropped out) - 16-19 age band
femal16to19	Female Civilian Population 16 To 19 Years
femaledrop	Female Not high school graduate, not enrolled (dropped out) - 16-19 age band
popunemplo	Unemployed total population
popinlabou	Total population in labour force
poptot	Total population
households	Total Households
withpubass	Households With Public Assistance Income
medianinco	Median household income (In 2012 Inflation Adjusted Dollars)
poor	Doing poorly as regard Ratio Of Income In 2012 To Poverty Level (Under 1.00)
female	Total Population Female
male	Total Population Male
pacific	Total Population Pacific Islander
mixed	Total Population Mixed race
hispanic	Total Population Hispanic
european	Total Population White
asian	Total Population Asian American
american	Total Population American Indian
african	Total Population African American

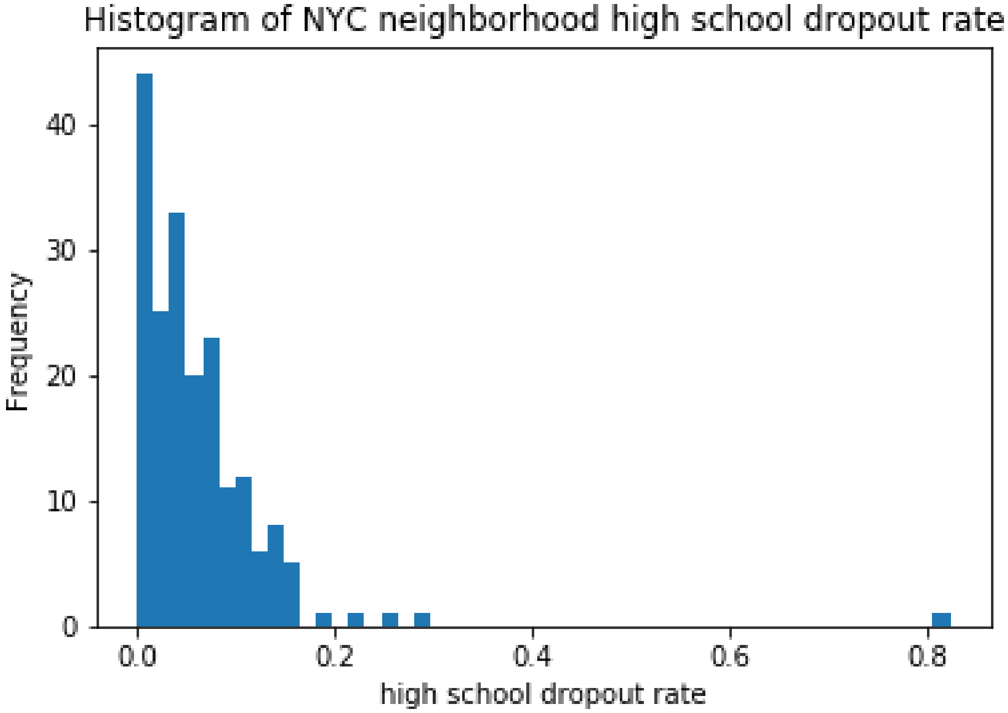


Figure 1: histogram of high school dropout rate

Here's an example of the exploratory data analysis I did. This is a histogram showing the distribution of high school dropout rate in NYC neighborhoods. It is not surprising that the data are strongly skewed, and most observations fall within the range of 0.0 and 0.2.

Table 2: Summary Statistics

Variable	N	Mean	St. Dev.	Min	Max
high school droupout rate	192	0.060419	0.074030	0	0.822222
unemployment rate	192	0.103509	0.037271	0.048299	0.224234
Households With Public Assistance Income	192	0.043396	0.032536	0	0.138621
poverty rate	192	0.188334	0.109335	0	0.555593
Median household income	192	58107	23671	18690	156604
male	192	0.476013	0.039998	0.346436	0.905554
female	192	0.523987	0.039998	0.094446	0.653564
pacific	192	0.000405	0.000862	0	0.004973
mixed	192	0.028857	0.018769	0.003587	0.162148
Hispanic	192	0.279489	0.212956	0.010923	0.858873
European	192	0.455828	0.28050	0.017508	0.976538
Asian	192	0.126666	0.142224	0	0.668978
American	192	0.003581	0.003224	0	0.017076
African	192	0.243516	0.268429	0.001474	0.938427

There are three undefined areas, which all have 0 high school dropout and 0 male16to19 + female16to19, so I choose to not consider these three observations. Hence, there are 195 neighborhood tabulation areas in total, but in this project, the number of observation will be 192.

3 Methods

First part of the project will be an analysis of the spatial distribution of high school dropout rate in NYC neighborhoods. Local spatial autocorrelation will be used. The Local Moran statistic was suggested in Anselin (1995) as a way to identify local clusters and spatial outliers. Rook contiguity spatial weight and queen contiguity spatial weight were generated before doing spatial autocorrelation.

Second part will model high school dropout using neighborhood characteristics. As I mentioned in the literature review section, a study conducted by Donnelly (2017) covered a relatively comprehensive range of neighborhood characteristics and comprised an index of neighborhood disadvantages. Donnelly did analysis on the effect of neighborhood disadvantage on high school dropout as well as a multilevel analysis on the extent to which school, neighborhood, and peer group mediate the effect of disadvantaged neighborhood. The neighborhood characteristics used in this study include unemployment rate, proportion of persons with income below poverty line, proportion persons age 25 and older with no high school diploma or equivalency and so on. Socio-demographic control variables were also added to the multivariate models that the study built. My theoretical model is largely based on Donnelly's theoretical constructs.

My basic model is:

$$\begin{aligned}\text{high school dropout rate} = {} & \beta_0 + \beta_1\text{unemployment rate} + \beta_2\text{poverty rate} \\ & + \beta_3\text{median household income} \\ & + \beta_4\text{demographic control}\end{aligned} \tag{1}$$

4 Initial Results

Figure 2 is a natural break map. It seems that high school dropout rate in NYC neighborhoods are much higher in the Bronx area, which is the upper right corner. In contrast, neighborhoods in the Staten Island have rather lower high school dropout rate. There's also an area in Queens with aggregated light yellow regions, which also means rather low high school dropout rate compared to other areas.
Moran's I scatter plot is shown in figure 3. A positive Moran's I statistics indicates positive spatial autocorrelation (high values clustered with high, low values clustered with low).
Local spatial autocorrelation was done using queen contiguity weight, and the LISA cluster map and LISA significance map were presented in figure 4 and figure 5. The high-high points and low-low points in the LISA cluster map basically corresponds to what we visualized from the natural break map with high-high points clustered in the northern neighborhoods and low-low points clustered in Staten Island and some eastern neighborhoods.

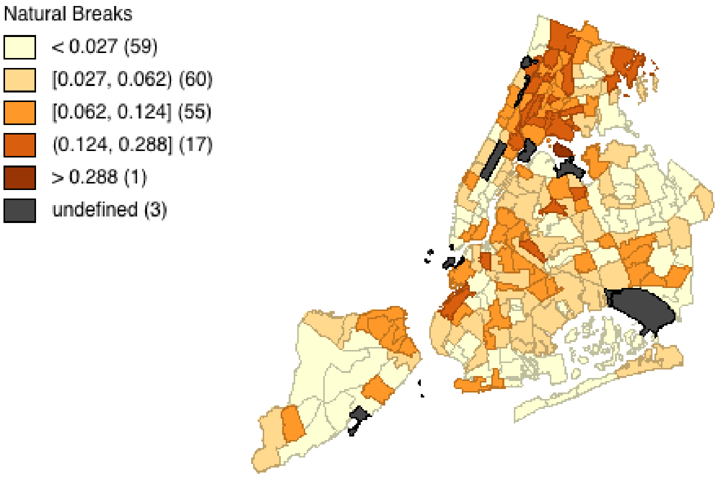


Figure 2: Natural Break Map of high school dropout rate

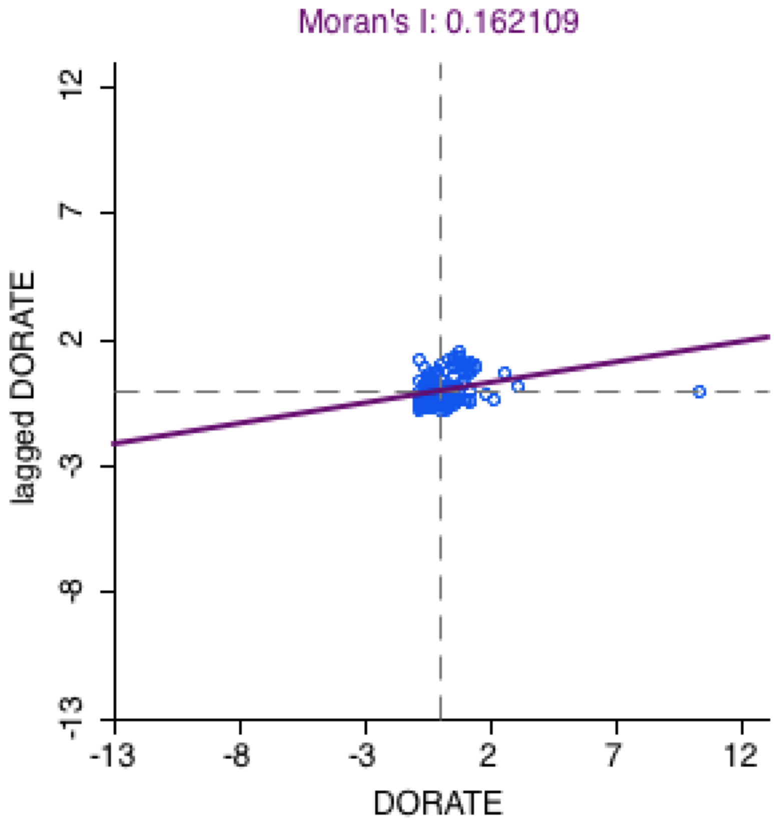
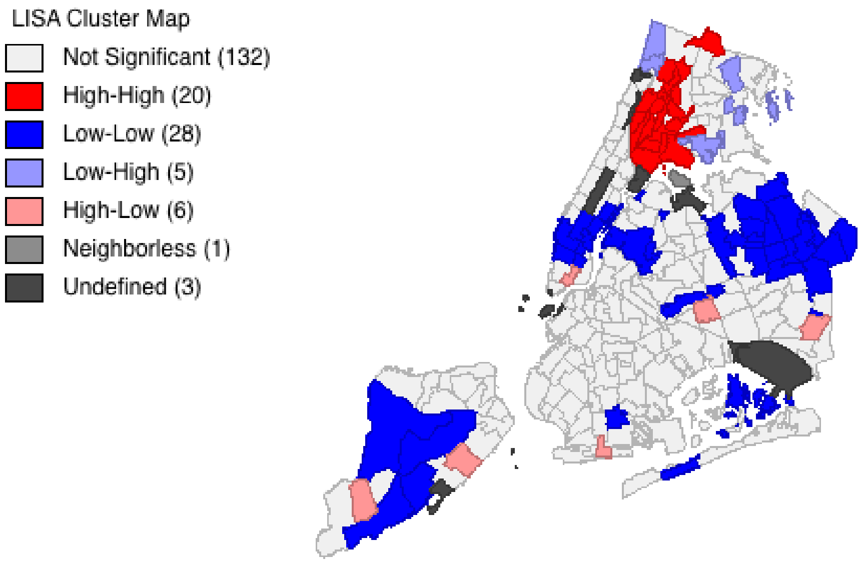


Figure 3: Moran's I Scatter plot



[H]

Figure 4: LISA Cluster Map

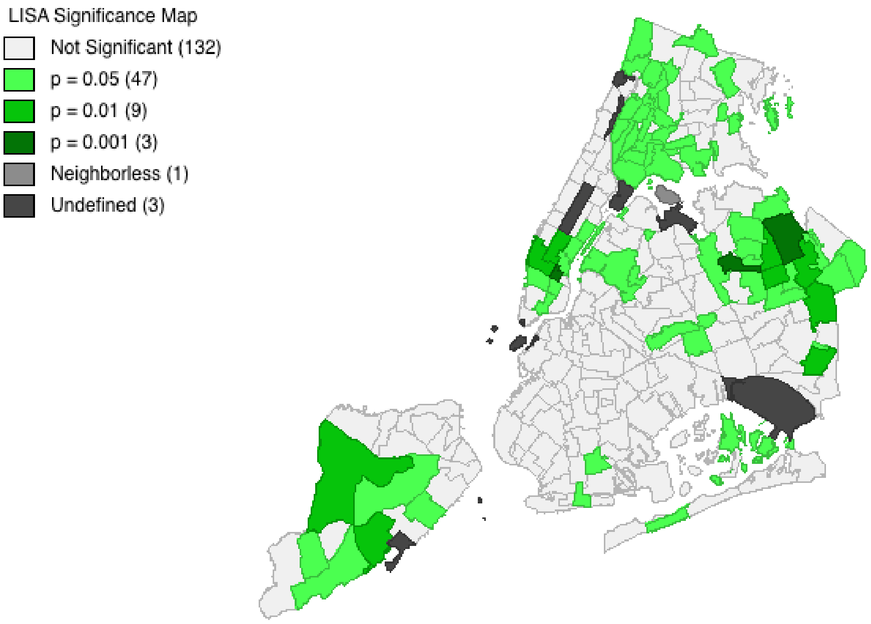


Figure 5: LISA Significance Map

The result of OLS is shown below. The response variable is dropout rate and the explanatory variables are unemployment rate, poverty rate, median household income, proportion of households with public assistance income and demographic control variables. The adjusted r-squared is 0.525, which means 0.525 of the variation can be explained by only the independent variables that actually affect the high school dropout rate. Only hispanic is significant at the level of 0.01.

Dep. Variable:	DORATE	**R-squared:**	0.555
Model:	OLS	**Adj. R-squared:**	0.525
Method:	Least Squares	**F-statistic:**	18.50
Date:	Wed, 09 May 2018	**Prob (F-statistic):**	1.51e-25
Time:	02:41:51	**Log-Likelihood:**	381.37
No. Observations:	191	**AIC:**	-736.7
Df Residuals:	178	**BIC:**	-694.5
Df Model:	12		

	coef	**std err**	**t**	**P> | t |**	**[0.025**	**0.975]**
constant	0.0298	0.040	0.753	0.453	-0.048	0.108
UNEMPRATE	0.0474	0.127	0.373	0.710	-0.203	0.298
medianinco	-1.928e-07	2.03e-07	-0.949	0.344	-5.94e-07	2.08e-07
POVERTRATE	0.0549	0.055	1.005	0.316	-0.053	0.163
ASSISTANCE	0.0830	0.182	0.457	0.648	-0.276	0.442
FEMALEPROP	-0.0986	0.064	-1.540	0.125	-0.225	0.028
MALEPROP	0.1285	0.067	1.929	0.055	-0.003	0.260
PACIFICPRO	0.8592	2.912	0.295	0.768	-4.888	6.607
MIXPROP	-0.0334	0.161	-0.207	0.836	-0.352	0.285
HISPANPROP	0.0877	0.035	2.505	0.013	0.019	0.157
EUROPROP	-0.0237	0.048	-0.494	0.622	-0.118	0.071
ASIANPROP	-0.0446	0.053	-0.838	0.403	-0.150	0.060
AMERIPROP	0.8769	0.896	0.978	0.329	-0.892	2.646
AFRICANPRO	-0.0023	0.049	-0.046	0.963	-0.100	0.095

Omnibus:	94.895	**Durbin-Watson:**	1.874
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	556.202
Skew:	1.812	**Prob(JB):**	1.67e-121
Kurtosis:	10.534	**Cond. No.**	1.97e+21

5 Next Step

A more in-depth spatial analysis will be performed based on the maps and methods that are already presented in this section. I also plan to use random forest to model high school dropout rate and do a model comparison.