Neighborhood Disadvantage and High School Dropout

Lerong Wang
MACSS, Division of Social Science, University of Chicago
lwang11@uchicago.edu

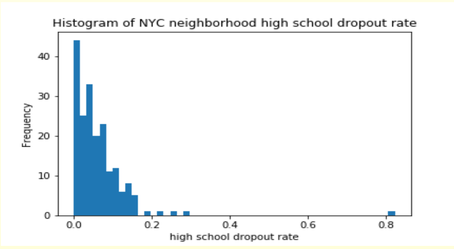
Background

Since the establishment of public school in the United States, obtaining a high school education has been viewed as the great equalizer to upward social mobility and opportunity. Educators and administrators in the US public school system have been dedicated to lowering high school dropout rates to ensure more equitable opportunities and outcomes for all students. Dropping out of high school has been associated with negative psychological adjustment and negative life outcomes. Thus, studying high school dropout has become an increasingly important social psychology, sociology, and public policy question.

Current research will focus on the environmental determinants that may affect high school dropout. My research question is: Can we use neighborhood disadvantages to predict high school dropout rates?

Specific Aims

In this research, high school dropout rate in NYC neighborhoods will be the outcome variable.



I plan to do a spatial analysis of high school dropout rate in NYC neighborhoods. Then, I will use different models to predict high school dropout rate in NYC neighborhoods and compare their predictive power.

Data

High school dropout data, socio-demographic data, and neighborhood characteristics for New York City Census Tracts come from American Community Survey 2008-2012, US Census Bureau. I downloaded a compiled version from Geoda Data and Lab website.

Each observation in the original data corresponds to one census tract in New York City. There are 2166 observations in total. I performed a spatial aggregation and combined the individual observations within the same neighborhood tabulation area.

NYC Neighborhood Tabulation Area Shapefile Data is used to create a base map for the NYC Neighborhood Tabulation Area.

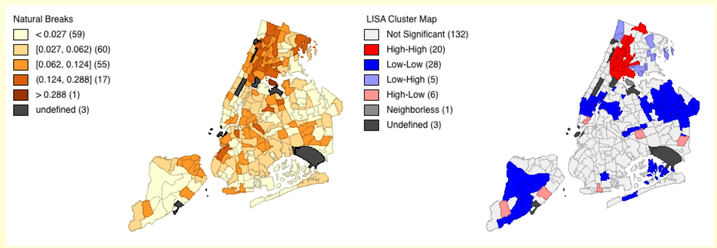
Methods

- Perform spatial analysis and local spatial autocorrelation
- Hypothesized model: High school dropout rate = $\beta_0 + \beta_1$unemployment rate + β_2 poverty rate + β_3 median household income + β_4 public assistance + β_5 demographic control
- Use OLS, random forest and decision tree to predict high school dropout rate in NYC neighborhoods.
- Use cross validation and compare the performance of different models

Spatial distribution

Natural break map

LISA cluster map

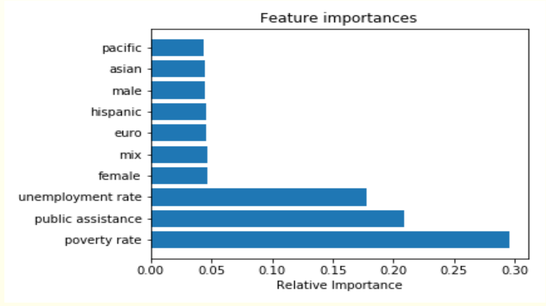


OLS

	coef	std err	t	P> | t |	[0.025	0.975]
constant	-0.0101	4.537	-0.002	0.998	-8.993	8.973
urate	-2.0131	15.157	-0.133	0.895	-32.023	27.997
pvrate	2.0633	6.929	0.298	0.766	-11.655	15.782
assistance	37.3056	22.102	1.688	0.094	-6.455	81.066
femalepr	-12.7820	7.076	-1.806	0.073	-26.792	1.228
malepr	12.7719	7.523	1.698	0.092	-2.123	27.666
pacificpr	182.9742	327.901	0.558	0.578	-466.248	832.196
mixpr	1.2515	18.379	0.068	0.946	-35.138	37.641
hispanicpr	9.9028	4.509	2.196	0.030	0.975	18.830
europr	2.4872	5.609	0.443	0.658	-8.618	13.592
asianpr	-1.1014	6.473	-0.170	0.865	-13.918	11.715
americanpr	166.1614	99.969	1.662	0.099	-31.770	364.093
africanpr	3.9238	5.790	0.678	0.499	-7.539	15.387
medianinco	-1.724e-05	2.26e-05	-0.761	0.448	-6.21e-05	2.76e-05

R-squared of OLS is 0.611. The only variable that is significant at the level of 0.05 is Hispanic population proportion. Poverty rate has a positive effect on dropout rate.

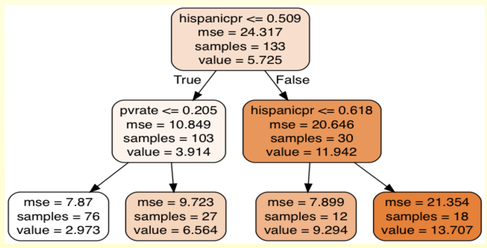
Random Forest



The feature importance ranking of the random forest model shows that top ten relative feature importance.

Decision Tree

After using cross-validation to determine the optimal level of tree complexity, I get the following decision tree.



Hispanic population and poverty rate are more useful for partitioning the tree.

Model Comparison

I use the validation set approach, break the data into training set and test set, and test the MSE of the predictor. I get the following results.

Model	Test MSE
OLS	15.025
Random Forest	14.326
Decision Tree	14.523

Findings

The high-high and low-low clusters in the LISA cluster map have positive local spatial autocorrelation. These clusters are more similar to their neighbors (based on the weighted average of the neighboring high school dropout rates, and the spatial lag).

Random forest has the best predictive performance for my feature set among the three models. The top three relevant features from the random forest model are poverty rate, public assistance, and unemployment rate.

Limitation

- The neighborhood characteristics considered in this research are limited. More neighborhood indicators such as crime rate are necessary.
- There are several ways to divide a city into small regions such as neighborhood tabulation area, zip code tabulation area and school district. More theoretical foundations are needed to determine which method makes most sense.
- Future research can use the spatial clusters as a starting point and analyze their common characteristics.