MACS30200 – Perspectives in Computational Research

Data, Methods, & Initial Results (Draft)

Kevin Sun

Tuesday, May 8, 2018

Theoretical Constructs & Defining the Model

As previously mentioned in the Literature Review section, the purpose of this research paper is to connect two important nodes in the field of education. Scholarship on college retention and persistence rates has extensively studied student demographic, student academic, and tertiary institution characteristics as factors that impact retention and persistence rates. Relatedly, but never explicitly connected with this scholarship, is the broad field of work documenting the differentiating treatment and impact white versus non-white teachers have on students. Thus, the model this paper defines hinges on a variable of interest we will call *critical mass.*

The *critical mass* variable takes on a value between 0 and 1 and represents a percentage of a given Chicago high school's teaching staff that is predicted to be non-white. Recall that the reason we are interested in *critical mass* as a variable is rooted in the theory of *representative bureaucracy*. This theory originating from Kingsley argues that institutional bureaucracies should be demographically representative of the constituents they serve in order to best meet their needs. Thus, when a particular "critical mass" threshold of non-white teachers is surpassed at a school, we want to examine whether a significant increase in college retention and persistence rates of graduates of that high school can be observed.

$$Persistence = \beta_0 + \beta_1\, CriticalMass + \beta_2 ELL + \beta_3 SPED + \beta_4 FreeLunch + \beta_5 White + \beta_6 Black + \beta_7 NativeAmericanAlaskan + \beta_8 Hispanic + \beta_9 MultiRace + \beta_{10} Asian + \beta_{11} HIPI + \varepsilon$$

The model above illustrates the *critical mass* variable as the variable of interest as well as including a host of demographic control variables. The model will control for: 1) *ELL* as the proportion of a school's student population that has English Language Learners (or as CPS classifies them as "bilingual"); 2) *SPED* as the proportion of a school's student population that requires special education services or Individualized Education Plans (IEP); 3) *FreeLunch* as the percentage of students of a school's student population that qualifies for free or reduced lunch – a socioeconomic signifier for a school; 4) *White, Black, NativeAmericanAlaska, Hispanic, MultiRace, Asian,* and *HIPI* as the proportion of a school's student population that are white, African-American, Native American/Alaskan, Hispanic, multiracial, Asian, or Hawaiian/Pacific Islander, respectively.

The dependent variable of interest – *Persistence* – is the percentage of graduates of a specific Chicago high school who returned the college or university after their first year enrolled at a college or university.

Data Source & Collection

The data utilized in this research is publicly available data published by Chicago Public Schools. The data on school-level demographics of students is published in this data portal under the "Demographics" tab. The data on college enrollment and persistence is also housed on this data portal. The data on individual teachers at each individual school is taken from the Employee Position Files, also published and maintained by Chicago Public Schools. These are the raw data

files that were used in this research. Data collection process consisted of downloading the excel files and/or csv files from the Chicago Public Schools Data Portal. Further description in the following section details the process of calculating and obtaining the *critical mass* variable from the Employee Position Files.

Data Wrangling & Imputation & Dealing with Missingness

Most of publicly available data published by Chicago Public Schools is granular only down to the school-level. Furthermore, individual school profiles reveal only demographics of the student body, not the teaching staff. Demographics for the teaching staff is available publicly only at the district level, which is not helpful for the purposes of our model. Thus, we use the Employee Position Files which contains individual-level data of every employee on payroll in Chicago Public Schools.

We first filter the dataset by the keyword "teacher" to include only employees who are classified as "teachers" in some capacity by Chicago Public Schools. This yields roughly 21,000 individuals which is slightly larger than the 20,626 teachers the CPS website identifies is on their teaching staff. This discrepancy may be due to the inclusion of individuals who have "teacher" in their job title but are not assigned to a specific school as well as teacher's assistants. The former inclusion does not impact our variable of interest, *critical mass*, as individuals not directly affiliated with any specific school will not count towards that school's *critical mass* variable. We choose to include "teacher's assistants" as individuals in these positions on occasion will execute instruction, discipline, and other actions similar to classroom teachers.

To obtain our variable of interest, *critical mass*, we use the *ethnicolr* python package to impute the race of a teacher based on either their last name or their full name from the Employee Position Files dataset. The *ethnicolr* package contains multiple methods for imputing race based

on either a given last name or full name – we utilize an ensemble of three of these methods to predict race based on a name. Each of these three methods is based on a specific dataset: 1) the 2010 United States Census 2) Wikipedia 3) Florida voter registration files. The first method takes an individual's last name and predicts their race based on the proportion of individuals in the 2010 U.S. Census that identified as a particular race. The second method utilizes Wikipedia data to predict a specific region of the world (i.e. 'Greater African, Muslim' or 'Greater European, East European') from which the last name appears and from there, we generalize those regions into broader race categories like "white" or "Asian". The third method takes an individual's full name and predicts their race based on the proportion of individuals in Florida's voter registration files that identify as a particular race. Finally, if two or more of the three methods predicted the same race, then we classified that teacher as that race. Following that process, we then calculated for each school the proportion of teachers at each school that identified as "non-white" versus "white" – that proportion is our *critical mass* variable for each school.

Next, we had to solve issues with missingness in our dependent variable – *persistence.* There are 5 years of enrollment and persistence rates for each high school in the dataset. If a school has 2 or fewer years of missing data, the enrollment and persistence rates are imputed from the average of the 3 years for which the school does have data. Otherwise, a school with 3 or more years of missing enrollment and persistence rates is dropped from the dataset.

To connect these two data sets, we initially attempted using record linkage to link on school name between the college retention & persistence data set and the critical mass data set. Given two different sets of school identification numbers provided, linking on school ID was not possible. Further, there was relatively large gap in naming conventions native to both datasets. A

single school may go by multiple names (e.g. “Hyde Park HS” and “Hyde Park Career Academy” are the same school; “Payton HS” and “Walter Payton College Prep High School” are also the same school). This proved challenging for efficiently measuring the distance between two strings using the *recordlinkage* package in python using Levenshtein, qgram, and Jaro-Winkler methods. Thus, the school names were manually recoded and then merged.

Our final dataset yields a total of 71 high schools in Chicago from which we run our model. (*Note: All code for data cleaning, wrangling, imputation can be found in the import_data.py file on Github)*.

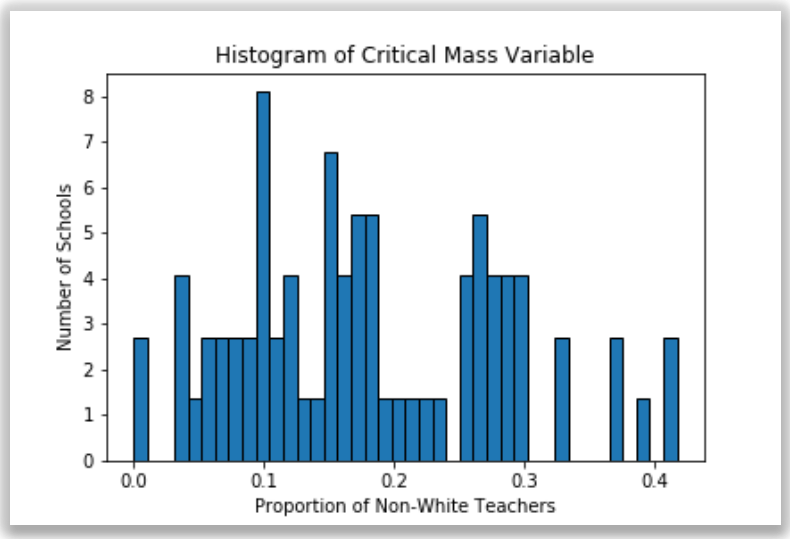
Summary Statistics

Critical Mass Variable	
# Obs.	71
Mean	0.1823
Std. Dev.	0.1019
Min.	0.0000
25%	0.1033
50%	0.1728
75%	0.2670
Max.	0.4167

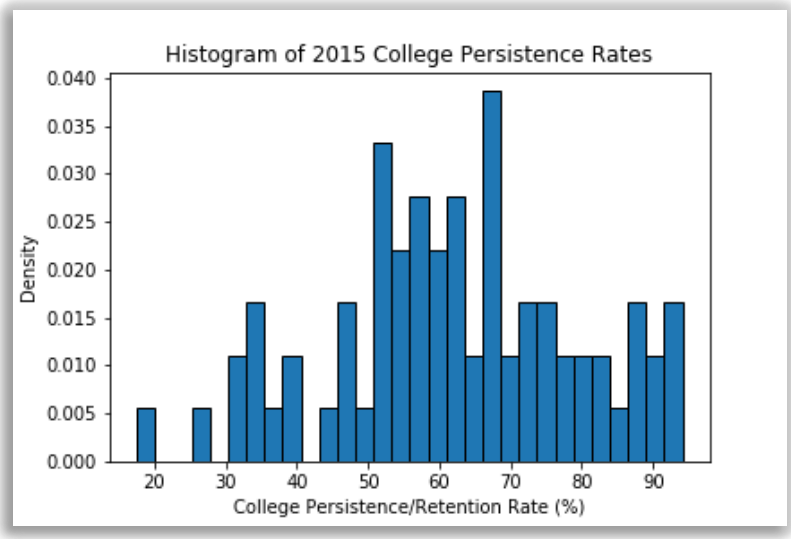
Exploratory Data Analysis

The distribution of our variable of interest *critical mass* is plotted in the histogram below. We observe a multi-modal distribution. Both an interesting and problematic observation of this data for our model is that this distribution of proportion of non-white teachers is more

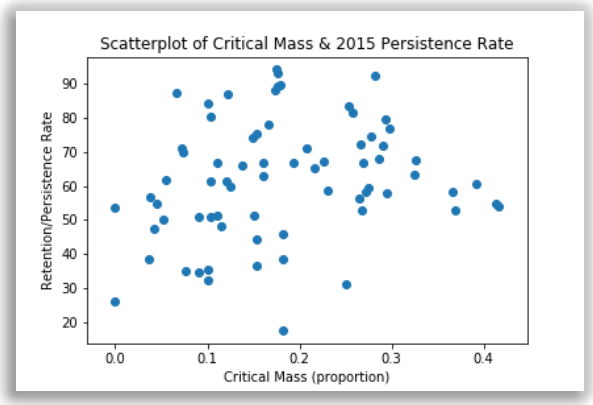
representative of the national demographic of teachers in public schools than of the demographic of teachers in Chicago Public Schools. CPS reports that ~50% of its teachers are not white.

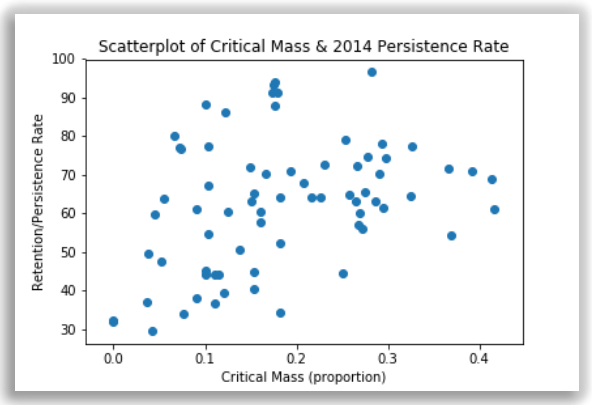


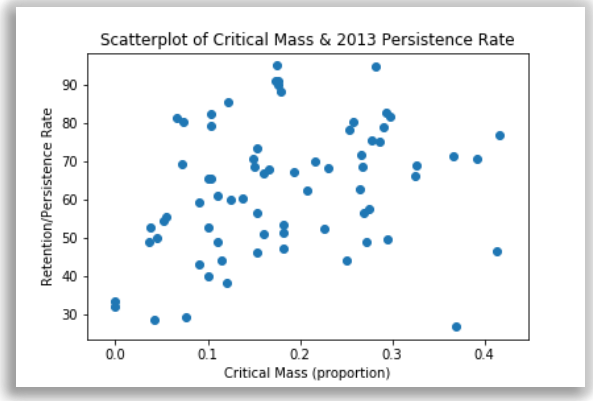
The distribution of our dependent variable *persistence* for the year 2015 is plotted in the histogram below. We see a trend that is skewed slightly to the left with the median of high school's having graduates persisting at 61.4%.

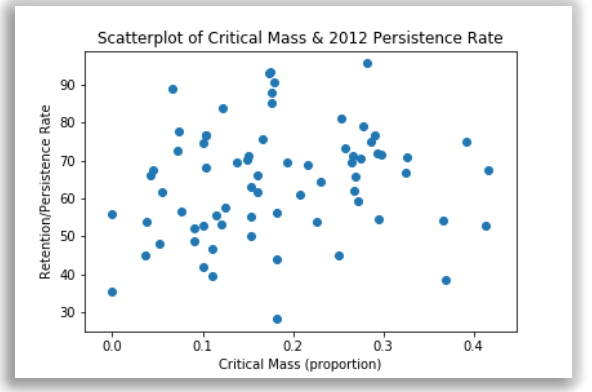


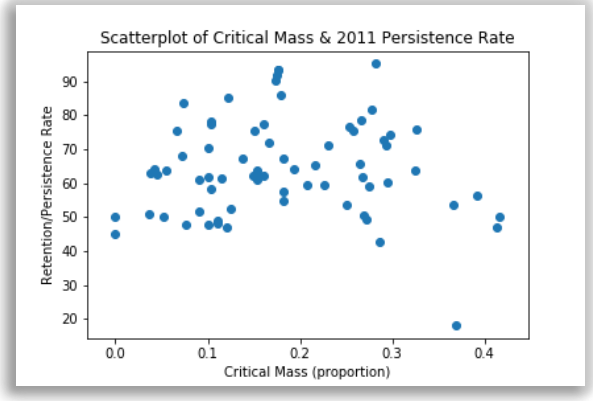
Additionally, we plot the relationship between *critical mass* and *persistence* for years in which we have data 2010-2015 below.

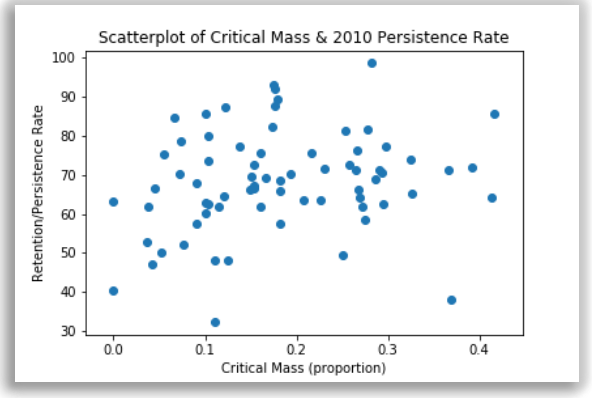












From the plots above, there appears to be no conclusive or consistent relationship between *critical mass* and college *persistence* rates. While the literature signaled that we might reasonably hypothesize that the relationship between *critical mass* and *persistence* would be non-linear, we do not observe any clear or evidence "threshold" of *critical mass* at which *persistence*

noticeably shifts, increases, decreases, etc. At this point, it may be necessary to revisit the race imputation process given the underestimation of the number of non-white teachers in Chicago Public Schools by our process. This will be discussed further in following sections.

Initial Results

After running OLS regression on college persistence data from the year 2014, the reported estimates are in the table below. While the sign on our variable of interest, *critical mass*, is what we might expect, it is not statistically significant.

Dependent Variable: College Persistence Rate (Year: 2014)

	Estimate	**Standard Error**
Intercept	1149.4920	2381.417
Critical Mass	28.8329	18.092
White	-10.4704	23.800
Black	-10.0601	23.813
Native American/Alaskan	-1.2104	24.246
Hispanic	-10.0127	23.812
Multiracial	-12.7024	24.263
Asian	-9.4830	23.773
Hawaiian/Pacific Islander	-21.1136	25.071
Unknown	-10.5216	24.204
ELL	-0.2883	0.484
SPED	-50.8926	16.014
Free Lunch	-88.5535	22.019

Weaknesses of Methods & Next Steps

The next steps include addressing some of the weaknesses of our *critical mass* variable. As observed in the summary statistics table for *critical mass*, the mean proportion of non-white teachers at each high school that we obtain is roughly 18% of the teaching staff. Thus, there is heavy underestimation from our imputation method of non-white teachers in sample given that it is reported that 50% of teachers in Chicago Public Schools are non-white. Although, interestingly enough, our race imputation methods align almost exactly with national trends where roughly 18% of teachers are not-white on a national level (as covered in the literature review). This could be the datasets on which the *ethnicolr* package draws to make predictions of race based on name; those datasets – the U.S. Census, Wikipedia, and Florida voter registration rolls – are not representative of the population of Chicago. Although, the predictions from the Florida dataset produced the closest aggregate predictions to total number of non-white teachers of our dataset.

Secondly, in alignment with critical mass theory literature, we will attempt a linear spline. The literature suggests both a non-linear relationship and that around 30-50% of non-white people in an organization (teachers at a school) and noticeable difference in outcome (retention/persistence rate) may be observed.