New York University, Center for Urban Science and Progress, July 2020

Digital CEQR 2.0: Making Real-Time Predictions for City Planning Proposals

Akash Yadav ay1652@nyu.edu
Chenjie Su cs5998@nyu.edu
Eric Zhuang yz2936@nyu.edu
Guilherme Louzada gl1082@nyu.edu

Abstract

Planning in NYC is mostly based on neighborhood initiatives and private land-use projects. Among those projects, many of them require the changes of the initial land use classification, for example, an industrial warehouse being retrofitted as a residential building. This process requires a multidisciplinary approach, including environmental and socioeconomic analysis, which is performed by the Department of City Planning, to make sure there will not be a significant adverse environmental impact on the neighborhood. The analysis is typically conducted under the scope of CEQR, which stands for city environmental quality review. The objective of this project is to develop a solution that will enhance the process involved in the generation of the environmental and socioeconomic information needed for this process, making use of machine learning techniques aiming to make the process quicker and more accurate.

Keywords: CEQR, New York City, Department of Planning, Displacement, Gentrification, Eviction, Machine Learning, Time Series

1 Introduction

Currently environmental analysis processes carried out by the Department of City Planning in New York to ascertain the positive or negative impact of a new construction or rezoning proposal on a neighbourhood are all done manually and can take anywhere between 65 to 125 days to complete [1]. The analysis itself has received a lot of backlash from residents across different neighborhoods in New York for not being able to correctly predict the impact of proposals on transportation, school capacity and secondary displacement. The environmental review for Downtown Brooklyn's 2004 rezoning projected that 979 new apartments would be built by 2013; but according to an analysis done by Municipal Arts Society, the growth has far outpaced projections and some 3,000 apartments were created by 2013, and another 5,000 new housing units had been built by 2018 [15]. Another assessment for Long Island City saw 4400 new students enroll for Brooklyn Community School District 13 which was 10 times higher than the estimated projections. Being plagued by both time and accuracy based challenges, our research seeks to find ways to accurately estimate projections for several demography related projections in response to Chapter 5 of the CEQR Technical Manual in real-time.

2 Problem Definition

The CEQR Technical Manual, which is the document that encompasses the rules of what should be accomplished for a project to be accepted, has 24 chapters. Chapter 5 of the CEQR Technical Manual lists down several parameters to estimate the risk of residential and commercial displacement in a neighborhood should a CEQR rezoning be approved. Based on several incorrect predictions and assumptions, the Pratt Center for Community Development came up with a list of recommendations to make the CEQR review process more robust and universal to get closer to correctly predicting demographic changes within a neighborhood [11]. CEQR reviews have traditionally been done without keeping in mind the macro trends within the city and have also been highly criticized for not being able to accommodate racial composition based changes within their studies. The rezoning proposal for Williamsburg is said to have caused a racialized displacement, displacing 15000 Latino and black families compared to 950 as initially estimated by the CEQR studies [16]. The way we propose to achieve this is by:

1. How do we accurately predict the risk of gentrification for a neighborhood by taking into account race, income, and built environment related variables for 1, 5 and 10 years into the future?
2. How do we predict economic and demographic trends for 1, 5 and 10 years into the future to better accommodate CEQR decision process?

Based on the literature review and the data available, we wish to propose solutions for our two research questions by building a tool that would ease the decision process for CEQR officials by providing a more macro, timely, and accurate

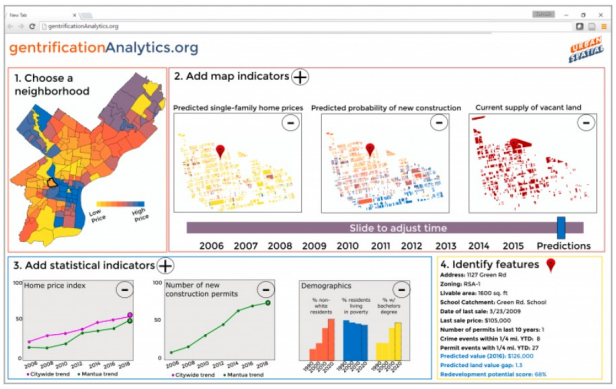

FIGURE 1: Example of an early warning tool for gentrification

prediction of neighborhood changes to study residential and commercial displacement better.

3 Literature Review

The design prototype and the analyses needed to get started with building predictive models were derived from technical documentation of several early warning systems and specific research papers dealing with gentrification. The CEQR App technical documentation [1,2,3,4] has clearly outlined the research, prototyping, testing, and reviewing phases needed for a typical urban planning tool that can be easily accessed by multiple stakeholders. In addition to the technical documentation, several Environment Assessment Statements issued publicly by the New York City Department of City Planning (DCP) for several rezoning proposals were reviewed. The statements contain information on the grounds on which a rezoning proposal was usually accepted and helped us in narrowing our search for the right data-sets.

Since gentrification is one of the common outcomes of neighborhood socio-economic change that also ultimately determines displacement, we decided to look closely at gentrification, especially from a prescient perspective as discussed in [9]. Documentation for tools like GentrificationAnalytics.org (Figure 1) created by the Urban Institute details topics of forecasting endogenous gentrification based on anticipating redevelopment of several neighborhoods [5]. Through research papers, technical blogs, and articles on gentrification and gentrification based tools, we were able to list down the data required, feature engineering techniques, models, and outputs that could be explored for building our tool for both residential and commercial displacement. To better understand motivations for displacement, we reviewed papers [6,7] discussing the migratory patterns of residents from different socio-economic backgrounds driven by compositional changes in neighborhoods, individual mobility patterns, and voluntariness of the individual to move to a different neighborhood. To estimate our target variables, we used research on gentrification from Governing that discusses a set of rules and preconditions to assess whether a census tract has been gentrified or not [10]. To study the incidence of commercial displacement in connection with gentrification, we reviewed literature suggesting the use of Yelp and Google data to estimate the correlation between restaurant types and affluence as a proxy for measuring gentrification [17]. Realizing the complexity of modeling urban phenomena using data science techniques, we also explore hybrid research techniques [8] on gentrification that make use of cellular automata models to determine action conditions based on which the probabilities of moving for individual households is modified within each time sequence.

4 Data & Methodology

American Community Survey (ACS) and the Social Vulnerability Index (SVI) was used to predict the risk of gentrification. ACS includes details on population spread, racial composition, income level composition, and several other demographic variables within a particular geography. We also used the Social Vulnerability Index (SVI) data-set, which, in contrast, has information about the extent of social risk within each tract of the United States.

We studied changes in income distribution, educational attainment distribution, minority population composition, crowding in buildings, median household income, median rent, etc. for five years to determine whether or not gentrification has occurred in a particular census tract. We used the research from the Gentrification Report Methodology (2015) [10] as a source to assign the gentrification labels. The report specifies two tests to determine if a tract has gentrified, and a census tract is deemed to have been gentrified if it passes these tests. The first test deals with determining if a neighborhood was eligible for gentrification. The second test lists down conditions based on which a neighborhood could be considered to have gentrified over a chosen period. The eligibility criteria deals with the following conditions:

1. The tract had a population of at least 500 residents at the beginning and end of the time period and was located within a central city
2. The tract's median household income was in the bottom 40^{th} percentile when compared to all tracts within its metro area at the beginning of the time period
3. The tract's median home value was in the bottom 40^{th} percentile when compared to all tracts within its metro area at the beginning of the time period

A tract would be considered to have been gentrified if it met the following conditions :

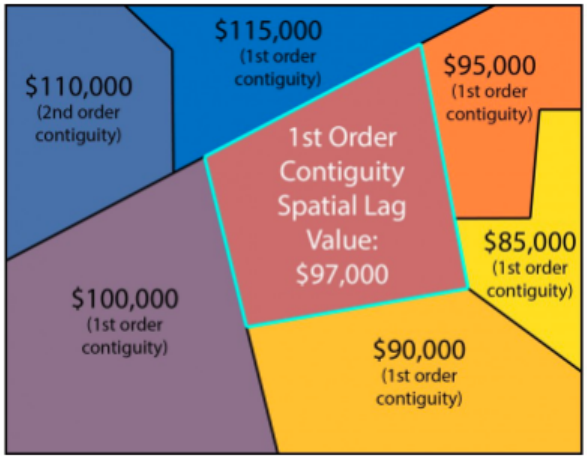

FIGURE 2: Using contiguity measure to assign spatial weights to neighboring tracts

1. An increase in a tract's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor's degrees, was in the top third percentile of all tracts within a metro area
2. A tract's median home value increased when adjusted for inflation
3. The percentage increase in a tract's inflation-adjusted median home value was in the top third percentile of all tracts within a metro area

The parameters that were ultimately used to build the models were all extracted based on literature review of prior gentrification studies and the critiques that have followed. Based on the recommendations of the Pratt Community Development Report, Flawed Findings, we shortlisted race and built environment related variables to be included along with income and education based demographic variables. A review of early warning systems also helped us engineer spatially relevant parameters that take into account the lagged weights of all the neighbors surrounding a particular tract. Models studying endogenous gentrification have often attributed spatial lags as important determinants for predicting risk of gentrification based on similar socioeconomic changes that have occurred temporally amongst the neighbor tracts. We construct these features by modelling New York City as an unweighted, undirected network with nodes being census tracts and edges occurring between tracts that are adjacent. The spatial lag was then calculated as a weighted average of all the surrounding neighborhoods for each demographic variable for the year 2011-2012 and then fed into our final models (Figure 2). Table 1 lists all the variables that were made use of for our modelling purpose.

Prior studies [13] describe the entire process of gentrification as a shift towards evenness in income profiles that indicate affluence and a predominant type of racial composition. To model this change as a response to our model, we attributed a measurement for inter-tract inter-year changes in the income distribution of the demography, the educational attainment of the demography and changes in median incomes, rent and home values. Feature engineering related to intercensal changes for demographic variables has all been replicated from an analysis to build a proximity-based early warning system for gentrification. To compute the intercensal change for demographic features such as income distribution, we used the Hellinger distance measurement as the distance between two distributions. For two discrete distributions P(X) and Q(X) with the same support, Hellinger distance is calculated as:

$$\triangle_{\text{Hellinger}} = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{P(X=x_i)} - \sqrt{Q(X=x_i)}\right)^2} \quad (1)$$

The second discrete distribution that we make use of to assess the evenness of income distribution is an ideal case scenario where we attribute the tracts with a baseline of perfect affluence. Tracts with low Hellinger distances calculated this way tend to be more high income whereas tracts with high Hellinger distances tend to be more low income. A tract can be considered to have become more affluent or gentrified from 2012 to 2017 if it has a negative difference and the inverse is true for a tract that becomes more low income. For other variables such as median income, median rent, minority composition, crowding within houses, and single-parent households, we computed the change from 2012 to 2017 within these parameters. For all our monetary related features from 2012 to 2017, we adjusted their values for inflation based on 2018 prices. The final set of variables that we chose for our analysis included the following:

To detail out all aspects of the CEQR Decisioning Process, we wish to complement the results of gentrification risk with predictions for home sale prices, rentals, and evictions. Displacement or migration usually happens when residents can no longer afford to remain in their houses due to rising prices or when they are forcefully evicted [12]. To better estimate the risk involving neighborhood changes, we, therefore, study the trends for rising housing prices and the rising risk of evictions around New York City. The data-set to build models for price prediction was extracted from the Department of Finance's Annualised Sales Data for Housing and Zillow's open data portal for data on Median Rentals across New York City neighborhoods for all types of housing. Both the data-sets were adjusted for inflation. For analysis of evictions and commercial displacement, Legally Operating Businesses and Evictions data-set from New York Open Data portal was used. To isolate the impact of CEQR proposals on rising prices and evictions, we used spatial

indexing to match each CEQR project to the smallest geography available in each of our other data-sets. This allowed us to understand the dynamics of changing prices, increased rate of evictions around particular BBLs, and within each zip with and without a CEQR project. Based on this, we were able to calculate the multiplier effects that a particular CEQR project has. Limitations around data availability for CEQR projects (8 projects with negative environmental impact) could not allow us to get estimates of neighborhood changes based on micro inputs of the CEQR process.

5 Models

We applied four machine learning methods for our gentrification based classification problem. We used unsupervised machine learning techniques such as K-Means Clustering to understand the different types of changes occurring within each neighborhood over 5 years and to assign labels of the perceived extent or risk of gentrification within each of these neighborhoods. We compared the results of our clustering techniques with supervised learning methods such as Random Forest Classifier, SVM Classifier and Logistic Regression where we were interested in the accuracy of these models to correctly predict between two labels (gentrification = 0 and gentrification = 1 where 0 means gentrification has not occurred in the particular tract and 1 means that it has).

We evaluated the performance of our supervised models based on using accuracy, precision, and recall scores. We were primarily interested in evaluating the recall scores, or the correctness of the model in predicting that a particular tract has gentrified. Since we had an imbalanced data-set with only 45 tracts having gentrification label as 1, we made use of stratified sampling techniques and SMOTE to balance the data-set. In the case of stratified sampling, we split the data-set into 5 samples having equal proportions of classes 0 and 1 within each sample. SMOTE on the other hand is an oversampling technique that balances out the data-set by generating randomly new examples or instances of the minority class from the nearest neighbors of a line joining the minority class sample. Based on these cross-validating techniques, we were able to then average out the accuracy, precision, and recall scores for our models. For models involving predictions of prices in a particular neighborhood, we made use of time-series modeling to get the mean estimates of prices 5 and 10 years into the future. We used seasonal decomposition to isolate the effects of seasonality from our data, ran stationarity tests, and then modeled the prediction problem using ARMA (Figure 3).

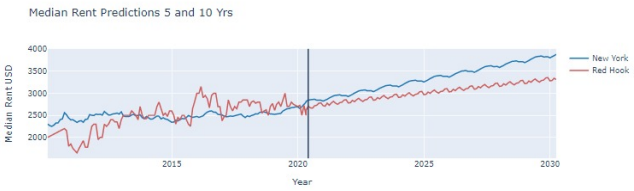

FIGURE 3: Predicting rental prices for all types of housing units across New York and Redhook

6 Results

For our clustering techniques, to evaluate the number of clusters needed, we made use of the silhouette score and the elbow method. Both Random Forest algorithms on each of our samples, after using the best parameters for maximum leaf nodes and tree depth. The accuracy of a Random Forest model was assessed by its ability to minimize the Gini impurity. The model performed poorly giving us an average accuracy score of only 69% and with precision and recall being 0. The features that showed up as important included spatial lags of white population composition, change (increase or decrease in the variable measured between 2012 and 2017) in median income, change in minority population composition, change in single-parent households, spatial lag of housing units available for sale around a particular neighborhood and changes in poverty, unemployment, crowding and people living in structures having 10+ units, mirroring some of the significant features leading to high separability from our clustering analysis. Using SVM Classifier yielded similar results with only marginal improvements in accuracy, precision, and recall scores. For logistic regression, we first balanced out the data-set using oversampling on our lower instances. After scaling the data-set and running recursive feature elimination, we were able to feed only the most important variables into our analysis. We further eliminated variables having p-values more than 0.05 to increase the performance of the model. The final results of this elimination resulted in an increase of accuracy scores to 82% with recall being close to 87%. Based on the results of the models, we were able to infer that a decrease in evenness of income distribution, decrease in evenness of educational attainment, decrease in the racial composition, decrease in crowding within structures with 10 or more units and spatial lags for home value, rent and income distribution all play a major role in suggesting whether or not a tract has gentrified (Figure 4).

To make predictions for 5 and 10 years into the future, we were able to use time-series modeling (ARMA) to simulate the changes in demographic variables based on ACS and SVI data available from 2010 to 2018. After simulating data for the years 2022 and 2027, we ran the regression models on this particular dataset and were able to predict the risk of gentrification for a future time.

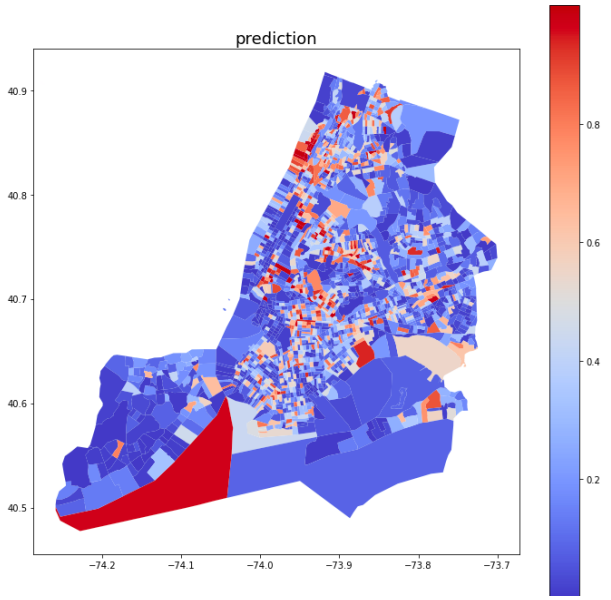

FIGURE 4: Prediction results for gentrification for each tract with probability ranging from 0 to 1

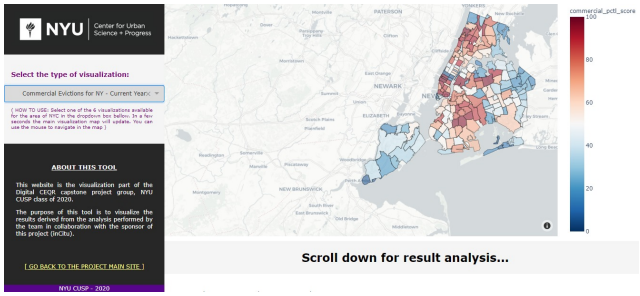

FIGURE 5: Early warning tools that can help accommodate multiple aspects of the CEQR process

7 Discussion

With the support of machine learning techniques, we can predict the risk of the gentrification of neighborhoods in New York City using the five-year ACS demographics and SVI index dataset. It is our understanding that other tools have previously produced similar results. Still, we could not find other tools that cover the many different aspects of CEQR decision process in one frame as we did. It is worth noting that the CEQR process itself does not use a macro approach (such as racial variables) and have rules that are considered flawed, some of which we found in the literature review. We believe our tool addressed these issues more consistently when compared to similar studies in this area. The findings of this research suggest that change in income and ethnicity related factors play important roles in predicting neighborhood gentrification risk. Since the city agencies currently only use project-specific CEQR analysis but excluding macro demographic and economic factors, our results can potentially provide them with a better understanding of city-wide displacement risk during the decision making process [11]. The unsupervised model results demonstrate that the declining population in structure with 10 units or more is also an important indicator of gentrification. This correlation conclusion also supplies a remedy option for the current CEQR analysis approach where only low-income tenants living in 1-4 unit buildings are considered vulnerable to displacement while excluding those living in larger buildings.

In addition to the prediction, we also built a web tool (Figure 5) that allows the audiences to toggle through different parameters of neighborhood change in NYC. The purpose of this tool is to provide auxiliary information for city officials to decide whether to approve or disapprove certain projects based on the current and future conditions of the neighborhood. To communicate the functionalities of this tool, we isolated Red Hook as an example. Based on the analysis and visualization, we summarize the socio-economic conditions of Red Hook including the safety of approving rezoning projects, the probability of gentrification in 5 years, the probability of residential eviction and commercial eviction along with the reasons, and the percentage increase of property price increase in the future.

8 Policy Implications & Future Work

The application of this framework may help CEQR process to be expedited and ease the decision process for approving proposals for city planning agencies. We believe that the implications of the tool can be extended for use by multiple stakeholders including architects, urban planners and policy makers. This framework brings visualisation to a bureaucratic process and this likely helps stakeholders get a clear understanding of the process.

The tool is currently limited in its scope for only one chapter of the CEQR process and geographically to the Red Hook area because this was an area of interest of the project team. One of the main limitations for this project was the time, since the machine learning process takes too long we could only generate results only for this area. We understand that this framework can be extended to other geographical regions and the tool can be updated in a way that it will be possible to compare different regions and also generate localised results for each specific region. This project covers the critique for traditional CEQR methods by extending its scope to include macro trends and demographic variables previously ignored in the manual processes.

For future implications, we want to develop a logic that calculates the safety index for approving rezoning proposals for a neighborhood-based on its eviction risk, property price, and gentrification risk indicators.

9 Planning & Team Roles

For planning we used an adaptation of the Agile methodology [18]. We decided to begin with a simple prototype and we evolved the idea with iterations involving project sponsors and mentors. Eventually, we achieved the goals that were set in the early stages of the project planning in four phases.

The team participants and their roles are as follows:

- Akash Yadav - Lead Data Scientist
- Chenjie Su - Lead Researcher
- Eric Zhuang - Project Coordinator/Liaison
- Guilherme Louzada - Planning/Front-End Developer

10 Acknowledgements

We would like to thank Chenglu Jin, our mentor for this project and Dana Chermesh, our sponsor for providing us with constant support and guidance in seeing this project through. Also we would like to extend our gratitude to NYU CUSP faculty and colleagues.

References

[1] New York City, Department of City Planning (2014). Procedures and Documentation, City Environmental Quality Review : Technical Manual

[2] New York City, Department of City Planning (2014). Establishing the Analysis Framework, City Environmental Quality Review : Technical Manual

[3] New York City, Department of City Planning (2014). Introduction to the Technical Guidance, City Environmental Quality Review : Technical Manual

[4] New York City, Department of City Planning (2014). Socioeconomic Conditions, City Environmental Quality Review : Technical Manual

[5] Steif, K. (n.d.). Predicting gentrification using longitudinal census data. Retrieved May 3, 2020, from http://urbanspatialanalysis.com/portfolio/predicting-gentrification-using-longitudinal-census-data

[6] Carlson, H. J. Measuring Displacement: Assessing Proxies for Involuntary Residential Mobility. City Community

[7] Marcuse, P. "Gentrification, abandonment, and displacement: Connections, causes, and policy responses in New York City." Wash. UJ Urb. Contemp. L. 28 (1985): 195

[8] Torrens, P. M., Nara, A. (2007). Modeling gentrification dynamics: A hybrid approach. Computers, Environment and Urban Systems, 31(3), 337-361.

[9] Pattabi, A. (2018). A Proximity-Based Early Warning System for Gentrification in California.

[10] Maciac M. (2015). Gentrification Report Methodology. Retrieved January 31, 2015, from https://www.governing.com/gov-data/gentrification-report-methodology.html

[11] Conte R. (2018) "Flawed findings, part 1, How NYC's approach to measuring residential displacement risk fails communities". PRATT CENTER FOR COMMUNITY DEVELOPMENT

[12] Way H., Mueller E., Wegmann J. (2018) Residential Displacement in Austin's Gentrifying Neighborhoods and What Can Be Done About It. The University of Texas Center for Sustainable Development in the School of Architecture the Entrepreneurship and Community Development Clinic in the School of Law.

[13] Chapple K., Zuk M. (2016) Forewarned: The Use of Neighborhood Early Warning Systems for Gentrification and Displacement. Volume 18, Number 3. U.S. Department of Housing and Urban Development

[14] Sims J., Iverson A. (2019) Multiple eviction: an investigation of chain displacement in Dane County, Wisconsin. University of Wisconsin–Madison, Madison, WI, USA

[15] Spivack, C. (2019, May 9) Rezonings transform NYC neighborhoods—but the city doesn't meaningfully study their impacts. Retrieved July 15, 2020, from https://ny.curbed.com/2019/5/8/18535693/nyc-neighborhood-rezonings-ceqr-environmental-review-city-council

[16] Iqbal, Z. (2019 December 9) Rezonings Underestimated development and displacement in both Williamsburg and Park Slope, study shows. Retrieved July 15, 2020, from https://bklyner.com/rezoning-displacement-study/

[17] Glaeser, E. L., Kim, H., Luca, M. (2018, May). Nowcasting gentrification: using yelp data to quantify neighborhood change. In AEA Papers and Proceedings (Vol. 108, pp. 77-82).

[18] Beck K. et. al. (2001) Manifesto for agile software development. Retrieved from http://agilemanifesto.org/

Appendix

Variable Name	**Description**
2017_2012_median_income	Difference in median income from 2012 to 2017
2017_2012_owner_occupied	Difference in value of housing units occupied by owners
2017_2012_median_rent	Difference in rent income
2017_2012_income_dist	Difference in income distributions measured using hellinger distances amongst a neighborhood
2017_2012_edu_dist	Difference in educational attainment measured using hellinger distances amongst a neighborhood
2017_2012_white_pop	Difference in white populations measured from 2012 to 2017
2017_2012_other_races	Difference in black, asisan, hispanic and native american populations
2017_2012_commuters_pt	Difference in percentage population using public transportation for commute
2017_2012_housing_units	Difference in number of housing units available for sale within each census tract
2017_2012_vacant_housing_units	Difference in number of vacant housing units within each census tract
2017_2012_bachelors_degree	Difference in number of people having a bachelor's degree
2018_2014_EP_POV	Difference between the poverty estimates (%) within a census tract
2018_2014_EP_UNEMP	Difference between the unemployment estimates (%) within a census tract
2018_2014_EP_SNGPNT	Difference between the proportion of single parent households within a census tract
2018_2014_EP_MINRTY	Difference between the proportion of minority populations within a census tract
2018_2014_EP_MUNIT	Difference between the proportion of population living in structures having 10 or more units
2018_2014_EP_MOBILE	Difference between the proportion of population living in mobile homes
2018_2014_EP_CROWD	Difference between proportions of people living in houses having more people than rooms
lagged_spatial_income	Weighted average of median income differences for all tracts in 2011-2012
lagged_home_value	Weighted average of differences in home values for all tracts in 2011-2012
lagged_spatial_rent	Weighted average of differences in median rent for all tracts in 2011-2012
lagged_income_dist	Weighted average of differences in income distribution for all tracts in 2011-2012
lagged_white_pop	Weighted average of differences in white population proportion for all tracts in 2011-2012
lagged_other_races	Weighted average of differences in racial composition of all tracts in 2011-2012
lagged_public_commute	Weighted average of differences in population commuting by public transportation in 2011-2012
lagged_housing_units	Weighted average of differences in availability of housing units for sale for all tracts in 2011-2012
lagged_vacant_housing_units	Weighted average of differences in availability of vacant housing units for all tracts in 2011-2012

TABLE 1: Variables Information.