Reshaping Gotham City: Identifying Key Factors behind High Crime Rate in Chicago and Providing Public Policy Solutions

Final Project Report for CAPP 30234

Ruize Liu (liu10)
Yimeng Qiu (qym)
Wenjun Shi (wshi1)
Jiaqi Yang (stephyang)

Abstract

This research aims at reevaluating public safety problem in Chicago, more specifically, identifying key features behind blocks with high crime rate and devising public policy solutions accordingly. In most social science literature, demographics are known to exhibit strong correlations with crime; yet, we believe demographics alone are not sufficient to explain this complex issue. Therefore alongside traditional demographics, we introduce two new groups – infrastructure and traffic flow – to better capture neighborhood characteristics[6].

Output of our model suggests that instead of a simple sum of human behaviors, high crime rate is an aggregated result of people interacting with environment. Impacts of societal environment on individuals play a more important role than their demographic features in terms of race, income and academic background.

Based on these, we propose to 1) initiate large-scale urban renewal projects with focus on renovating outdated houses, 2) reduce law enforcement actions that target individuals or groups based on demographic features 3) collaborate with neighbor cities on gentrification projects. Efficiency of proposed policies and projects have been validated using data from other metropolitan cities quantitatively and qualitatively.

Three types of machine learning techniques are used in this research: unsupervised clustering, supervised regression, and synthetic control method. In unsupervised learning, we apply hierarchical clustering and K-means to select states shared similar demographic features with Illinois. Then, scope is narrowed down to Chicago in the supervised regression part in order to key factors behind its high crime rate. Lastly, we constructed a simulated Chicago using synthetic control method with data of real cities selected from unsupervised learning. This method helps us better understand results of a policy or project.

1 Background and Overview

January 24th, 2017, President Trump posted on Twitter, "if Chicago doesn't fix the horrible 'carnage' going on, 228 shootings in 2017 with 42 killings (up 24% from 2016), I will send in the Feds". Though controversial, the fact of high crime rate in Chicago is nonnegligible. Since 2011, overall crime rate per one hundred thousand population of the U.S. has been decreasing steadily from 1500 to 1200, while the statistics of Chicago remain strikingly high as figure 1 shows. Dr. Florida, Sociology Professor of NYU, also demonstrated in his op-ed on CityLab that despite the nationwide great crime decline, crime rate of Chicago, particularly violent crime rate, rebounded recently [7][8].

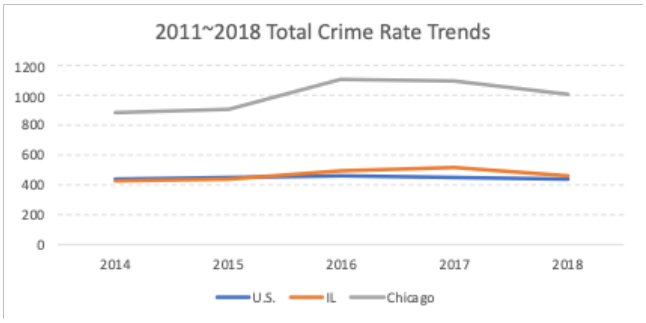


Figure 1. 2011 2018 Total Crime Rate Trends

Main target of this research is to identify key factors behind high crime rate and devise public policy solutions accordingly. For this purpose, we apply machine learning techniques in three steps: unsupervised clustering for selecting comparable states, supervised regression for identifying important features, and synthetic control method for evaluating proposed policies and projects. Details of data and statistical methods will be discussed in detail in following sections. [2][9][12]

, ,

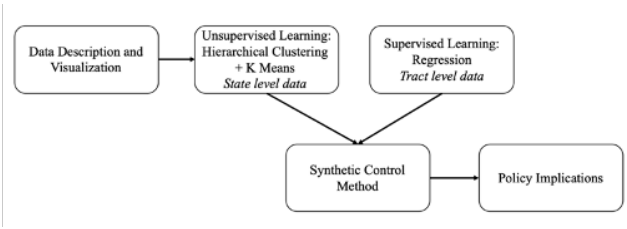


Figure 2. Model Construction Workflow

After policy research and statistical analysis, we propose large-scale urban renewal projects with focus on renovating outdated houses as the central solution to address this problem.

This report is prepared for three groups of audiences: policy makers in Chicago Housing Authority (CHA) and Chicago Police Department, Nonprofits for affordable housing like A Safe Haven and Deborah's Place, and potential real estate investors. CHA policy makers are the main target audience because they have direct governmental authority to launch urban renewal projects. For officers from Police Department, we hope they could understand the complexity behind crime incidents and reevaluate their law enforcement actions. Nonprofits have more connections with individuals and could facilitate communication with neighbors. Lastly, we attempt to convince real estate investors of potentials of outdated blocks and attract capital support from them.

2 Data

2.1 Features: demography, infrastructure, and dynamic traffic

Features in this research were collected in State, Tract, and City level for unsupervised clustering, supervised regression, and synthetic control prediction respectively.

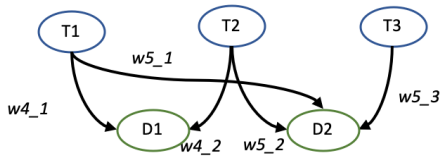
In state level, we selected demographic variables including percentage of African American, percentage of White, percentage of Asian American, median income, and median age as features for unsupervised clustering. Each row represents a state. All data were retrieved via Census API; no missing or outlier issues exist. Median income ranges from $43,567 (MS) to $82,604 (DC) with median of $59,116; median age ranges from 39.6 (NJ) to 28.5 (UT).

In tract level, we narrowed down scope to Chicago and incorporated infrastructure data and dynamic traffic flow information alongside with demographic data for supervised regression. Infrastructure data gathered from Chicago Data Portal include number of affordable units, number of grocery stores, count of library, count of public arts institutes, count of police stations, count of abandoned houses, and sum of graffiti complaints through 311.

The traffic aspect contains two features: static traffic_sum that sums up traffic flow of representative street corners within a tract in one day, and dynamic traffic_flow that reflects strength of connection among tracts. Specifically, we take percents of inflow taxi from each tract as weights and multiply by crime rates of corresponding tracts as [15]:

$$traffic\_flow = \sum_{i=1}^{n}\left(\frac{W_i}{\sum(W)} * crime\_rate_i\right)$$

where n = total number of tracts in Chicago. [15]



$$Index_4 = \frac{w_{4-1}}{w_{4-1}+w_{4-2}} * CrimeRate_1 + \frac{w_{4-2}}{w_{4-1}+w_{4-2}} * CrimeRate_1$$

$$Index_5 = \sum \frac{w_{5-i}}{\sum w_5} * CrimeRate_i$$

Figure 3. Illustration for Traffic Flow Index

For the last part, we synthesized city-level data of New York City, Houston, Las Vegas and seven other cities to create a simulated Chicago and verified effects of urban renewal projects. Data for this part are the same as above, but in city level.

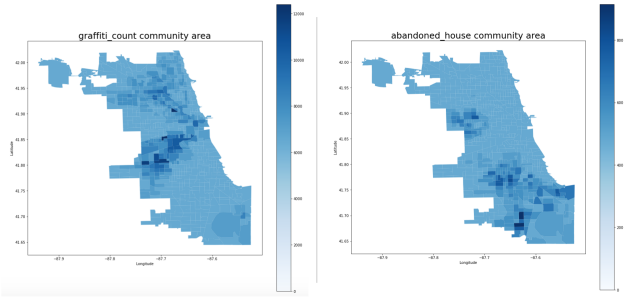


Figure 4. Distribution of Abandoned Houses and Graffiti in Chicago

2.2 Target: crime rate

Target variable throughout this research is crime rate, which is computed as total number of criminal cases divided by 100 thousand population within the area. Units of target are consistent with features – in state, tract, and city level respectively in clustering, regression, and synth method sections.

State-level data were downloaded from *FBI Offenses Known to Law Enforcement - Crime in the U.S. Archive*; tract- and city-level data were from City Data Portal. The two offices provide data collected in different statistical methods and thus have distinct features. FBI provides count of each type

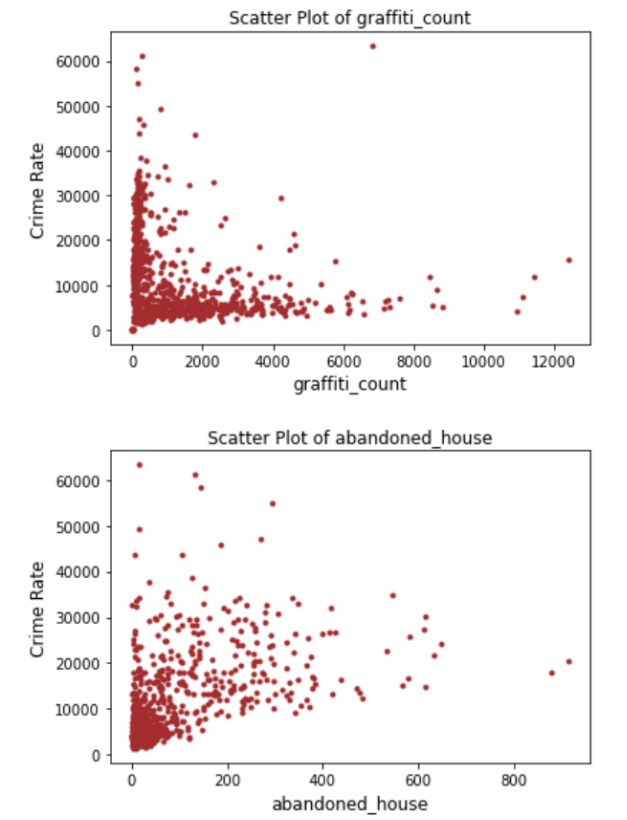


Figure 5. Scatter Plot of Two Features against Crime Rate

of crimes in city level over years holding consistent standards and categories, but lacks geographical details for each case. City Data Portal contains richer information on individual cases, while are less consistent across cities. Because of these features, FBI data are more suitable for general macro-level analysis; we took them for visualization, preliminary research, and clustering. Detailed crime records from City Data Portal were used in supervised regression and synth method sections.

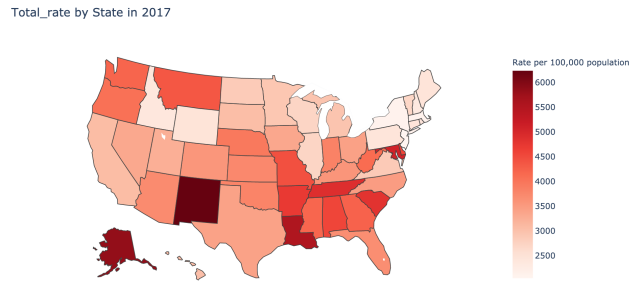


Figure 6. Total Crime Rate by State in 2017 (FBI data)

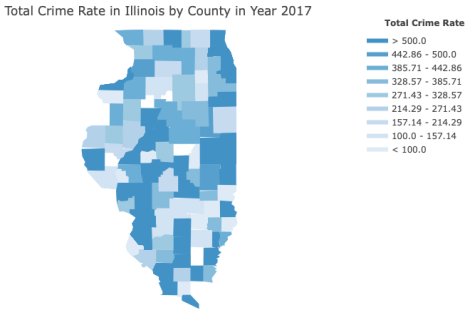


Figure 7. Total Crime Rate by County in Illinois in 2017 (FBI data)

3 Quantitative Machine Learning Methods and Discussion

3.1 Unsupervised Learning – Hierarchical Clustering and K-means

Unsupervised learning is one type of machine learning working for pattern detection, dimensionality reduction and classification based on input features alone. Common algorithms under unsupervised learning include hierarchical clustering, K-means and PCA. In this research, unsupervised approach was adopted for state clustering. By inputting demographic features of 50 states, we expected to split states into several groups and find states sharing similar demographic features with Illinois. Results of clustering will be used later in synth control analysis.

We applied both divisive hierarchical clustering and K-means methods, and selected states falling in the same group with Illinois in two methods for robust grouping results. Divisive hierarchical clustering goes through a top-down process that starts from all points and split them into branches until only one left in each branch. This algorithm is applied firstly because prior knowledge of number of groups is not required. Within hierarchical clustering, several criteria can be used for classification. Ward's minimum variance calculates Euclidean distance between points; single and complete method concern minimum and maximum of distance. We chose Ward's over others because our input features (median income, black percentage, white percentage, and median age) are unevenly distributed; taking minimum or maximum distance might lead to serious biases.

Dendrogram plotted based on hierarchical clustering indicates three is a reasonable choice for number of groups. Illinois falls into the second group with North Dakota, Rhode Island, and 12 other states.

Unlike hierarchical clustering, K-means starts from random centroids and requires prior knowledge about number of groups, yet it has advantage in less time complexity. Inputting number of group as 3, we got a K-means grouping result as below. Notably, though only three variables are displayed in the figure for clearer visualization, all four were

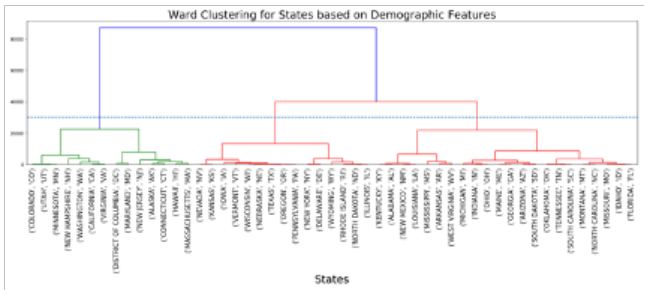


Figure 8. Dendrogram from Hierarchical Clustering

taken in the algorithm. Illinois belongs to the purple group and lies close to PA, RI and NY.

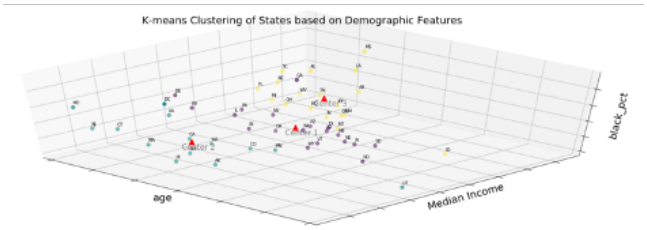


Figure 9. K-means Clustering

States in the same group with Illinois under two algorithms are: Texas, New York, Pennsylvania, Wisconsin, Rhode Island, Nevada, Wisconsin, Nebraska, Pennsylvania, Iowa, North Dakota, Kansas, Oregon, Rhode Island, Wyoming, Delaware, and Vermont. Representative cities from these states will be used in later policy analysis and synth section.

3.2 Supervised Learning – Multivariate Regression

In this part, we focused on tracts in Chicago only, aiming at identifying key factors having significant relationships with crime rate by tract. Traditionally, criminologists and policy researchers use demographics to analyze regional crime rate. In this research, alongside with basic demographic features, we hypothesized that infrastructure conditions and dynamic traffic flow are also correlated with neighborhood crime rate and have significant influence on it.

Multivariate linear regression technique fits our research purpose and dataset perfectly; all data for this question are continuous and we would like to check in detail how one-unit change of a specific feature will influence the overall crime rate.

Correlation plot suggests a strong positive relationship between percentage of white and percentage of black; in order to avoid multicollinearity, we dropped percentage of white. Sanity check and features normalization were also conducted before running regression.

Regressing crime rate against demographic features only, we got a linear model with 0.439 R-square. Only one out of six variables shows significant relationship with the target.

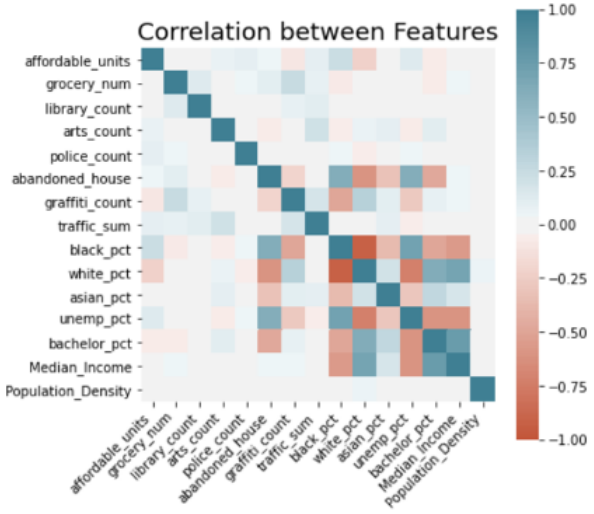


Figure 10. Correlation Matrix of Features

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.439
Model:	OLS	Adj. R-squared:	0.433
Method:	Least Squares	F-statistic:	83.46
Date:	Thu, 11 Jun 2020	Prob (F-statistic):	4.44e-77
Time:	12:12:52	Log-Likelihood:	-6667.8
No. Observations:	648	AIC:	1.335e+04
Df Residuals:	641	BIC:	1.338e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.076e+04	281.303	38.266	0.000	1.02e+04	1.13e+04
black_pct	6089.8370	437.043	13.934	0.000	5231.628	6948.046
asian_pct	310.0334	309.894	1.000	0.317	-298.496	918.563
unemp_pct	911.4325	445.504	2.046	0.041	36.610	1786.255
bachelor_pct	540.8667	467.107	1.158	0.247	-376.378	1458.112
Median_Income	237.3329	477.826	0.497	0.620	-700.961	1175.627
Population_Density	-285.3849	283.454	-1.007	0.314	-841.995	271.226

Figure 11. Coefficient Significance Summary of Demographic Features

Coefficient summary table of full regression is displayed below. Overall R-square of simple multivariate regression increases to 0.577 with adjusted R-square 0.567, much higher than partial regression. Using 20% of data as testing set, mean squared error of the model is 34240474.16 and bias is 38191793.8. Most infrastructure features and both two traffic features have significant linear relationship with crime rate. This result confirmed our hypothesis that demographics factors alone are not sufficient to explain crime rate.

Regularization

To achieve higher model precision, we repeated the training-and-testing process on polynomial regression and apply multiple regularization methods. Metrics of testing set shows little change, indicating multivariate linear regression is a rather robust model (Figure 13).

Bootstrapping

Since observations in the train set are only 648, in order to improve the performance, reduce the bias and variance, we used bootstrapping to verify the outcome of our OLS regression as follows:

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.577
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	57.54
Date:	Wed, 10 Jun 2020	Prob (F-statistic):	1.97e-107
Time:	02:54:58	Log-Likelihood:	-6575.9
No. Observations:	648	AIC:	1.318e+04
Df Residuals:	632	BIC:	1.326e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.076e+04	245.825	43.789	0.000	1.03e+04	1.12e+04
affordable_units	719.2527	260.622	2.760	0.006	207.464	1231.042
grocery_num	-16.9622	261.233	-0.065	0.948	-529.952	496.028
library_count	-68.3377	250.647	-0.273	0.785	-560.539	423.864
arts_count	496.9719	253.361	1.962	0.050	-0.560	994.504
police_count	1014.1293	248.223	4.086	0.000	526.688	1501.570
abandoned_house	883.5830	346.710	2.548	0.011	202.741	1564.425
graffiti_count	754.2945	299.978	2.514	0.012	165.220	1343.369
traffic_sum	1163.2933	291.600	3.989	0.000	590.670	1735.916
black_pct	5606.9411	462.801	12.115	0.000	4698.127	6515.755
asian_pct	-260.3059	278.732	-0.934	0.351	-807.659	287.047
unemp_pct	510.9499	404.599	1.263	0.207	-283.571	1305.471
bachelor_pct	301.2299	425.535	0.708	0.479	-534.405	1136.864
Median_Income	-427.3290	437.279	-0.977	0.329	-1286.025	431.367
Population_Density	-107.8280	248.399	-0.434	0.664	-595.616	379.960
traffic_flow	2423.2247	303.878	7.974	0.000	1826.493	3019.957

Omnibus:	291.824	Durbin-Watson:	1.962
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2219.796
Skew:	1.840	Prob(JB):	0.00
Kurtosis:	11.287	Cond. No.	4.76

Figure 12. Coefficient Significance Summary of Full Regression

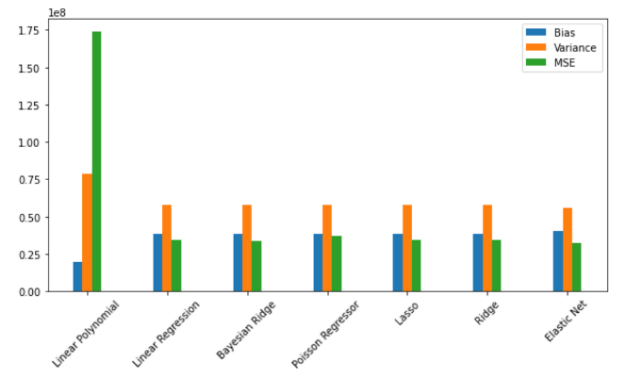


Figure 13. Testing Metrics under Different Methods

Step 1) We run the OLS regression based on bootstrapped sample to get our $\beta 1$

Step 2) Repeat previous step 10000 times, then we will get $\beta 1 \ldots \beta 10000$

Step 3) Use following formulate calculate the bootstrapped S.E., which B = 10000.

$$\hat{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}}_b)^2$$

In order to test whether our bootstrapped samples are fairly large enough to get an accurate estimate of standard error. We visualized results for each of number of samples (5,10,15,20,30,50,80,100) in box-plot by estimating the bootstrapped S.E. for 100 times.

As we can see, the mean and standard deviation for the estimated standard errors become stable with increasing number of bootstrapped samples. It looks like the outliers are suppressed when having sufficiently large number of samples. 10000 bootstrapped samples give us a really stable result.

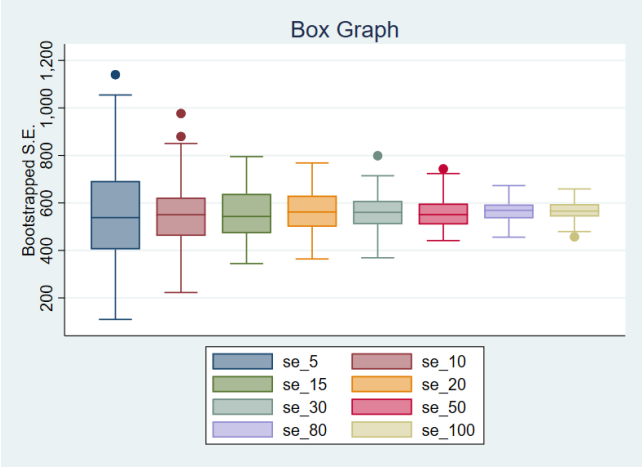


Figure 14. Example for black$_pctbootstrappedS.E$

In the conclusion, the infrastructure features such as abandoned house, graffiti count and dynamic features like traffic flow are crucial in the crime rate model. Traditional model with only demographics likely leads to biases and inconsistent results.

3.3 Comparative Case Study – Synthetic Control Method

Synthetic control methods were originally proposed by Abadie, Diamond Hainmueller [1] with purpose of estimating the effects of aggregated interventions, that is, how will a particular intervention implemented at an aggregate level affects a small number of large units (such as a cities, regions, or countries) on some aggregated outcomes of interest. It involves construction of a weighted combination of comparable cases used as controls. This comparison is used for estimating what would have happened to the treatment group if it had not received the treatment.

In our research we use this approach in a different way. We set Chicago as our treatment group, and then use major cities in states from unsupervised learning and pre-determined variables (not affect by our treatment) to construct a synthetic control group. Treatment in this case is a hypothetical urban renewal project and we set a cutoff point to test whether any real change occurs.

Advantages of this approach include:

- It suggests a method for determining the control group before researchers compare data. This could help guard against human bias of selectively constructing a control group in order to produce more favorable results.
- The synth method allows for a better match between the treatment and control groups based on observable covariates and pre-period values of the dependent variable.
- Researchers can run a "placebo" treatment on each potential control city using the same synth specification. This allows them to check whether the observed outcome in the "treatment" state is due to chance (e.g. differential city-time trends) or really due to the treatment in question. This improves on traditional diff-in-diff type studies that expose to this critique.

Drawbacks:

- This method is still somewhat susceptible to manipulation because the cutoff time point and length of observation period to construct the control is basically ad hoc.
- It is a process that is not always entirely transparent, which can be a downside in presenting to nonprofessional audiences.
- While the placebo tests provide a method for hypothesis testing, the placebo distribution is limited by the number of plausible geographic counterfactuals and ones ability to fit synthetic controls to each placebo.

4 Qualitative Policy Research and Synthetic Control Method Discussion

4.1 Insights from Linear Regression

The regression output reveals several interesting points. First, though still have some explanation power, most features in traditional demographic aspect do not show significant relationship with crime rate. P-values of unemployment rate, college degree percentage, median income, and population density are all above 0.05; using 95% as the threshold, statistically speaking, there is not enough evidence to suggest significant relationships between these variables and crime rate. This confirms our hypothesis that when including more infrastructure related factors, significance of demographics decreases, implying that crime is a complex outcome of people and environment.

Second, the infrastructure aspect as a whole has strong relationship with regional crime rate and further breaking down the issue, we found that renovating existing buildings will contribute more effectively to lower crime rate than constructing public institutes. The two features related to renovation, number of abandoned houses and count of street graffiti, have clearly significant positive relationship with crime rate with 99% confidence interval. Holding all other input unchanged, one less abandoned house in a tract will lead to 883 less crime per 100 thousand population. Yet on the other hand, significance level of public cultural institutes like library and public arts is much lower. Though one more library will lead to 68 less crime, their relationship is not robust.

Third, coefficient of index traffic flow index is large with high significance level, indicating that interconnection among tracts and dynamic flow are important in predicting crime rate.

4.2 Policy Research on In-Cluster Cities

From selected states in unsupervised learning, we further chose New York City (NY) Houston (TX), Philadelphia (PA), Las Vegas (Nevada), Milwaukee (WI), Wilmington (DL), Fargo (ND), and Ohama (NB) as control cities. These cities have comparatively complete data and hold important socioeconomic positions in their states, which make them comparable to Chicago.

Looking at crime rate, differences between these cities' crime rate and national average have been decreasing since 2015. Therefore we set 2015 as the cutoff time point and researched on policy change around this time.

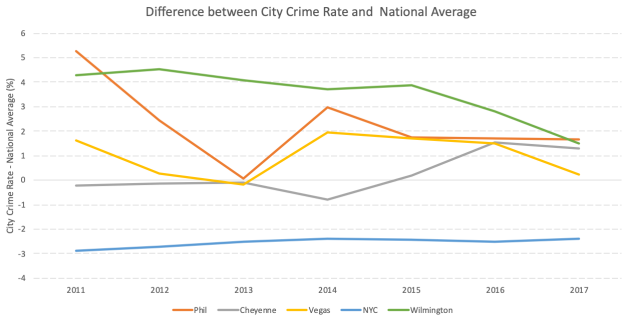


Figure 15. Difference between City Crime Rate and National Average

- **Cheyenne: $1 for historical houses**: In 2017, Cheyenne Mayor Marian Orr formed the Fight the Blight Committee to deal with abandoned house problem by offering some 'free' historical houses with only $1 [10][5].
- **Philadelphia: $1 for abandoned houses**: Over years, Philadelphia has sold thousands of abandoned properties for $1. New owners would earn their bargains by fixing up abandoned houses and returning once-derelict parcels to the tax rolls.[14][3]
- **Wilmington: Vacant Property Registration Program**: Wilmington has a Vacant Property Registration Program to encourage owners of vacant properties to immediately rehabilitate the property or to sell the property to an individual or an agency that will make the property attractive for sale or rental. Meanwhile, The Wilmington Neighborhood Conservancy Land Bank was established in 2017 to return vacant, dilapidated, abandoned, and delinquent properties to productive use, while strengthening and revitalizing the neighborhoods and inspiring economic development. [4][11]
- **New York City: Combat "Zombie Homes" and Graffiti Removal Project**:
  1. In order to "return zombie homes to productive use", New York legislature passed a package of laws in 2016 requiring lenders to inspect, maintain, and report zombie homes to the state. Later in the fall of 2017, HPD launched a "Zombie Homes Initiative" to track and identify these properties, conduct exterior surveys, provide resources to homeowners at risk of foreclosure [13].
  2. New York City takes strict actions on graffiti, including providing full-time, street-by-street graffiti removal service. The program began in Brooklyn, and early success prompted its expansion to the Bronx, Queens, Staten Island, and Manhattan.

Chicago also has Homebuyer Direct Program that "beak[s] down barriers to homeownership and help transform communities by offering fixer-upper homes at below-market prices in neighborhoods across Cook County". This program starts from 2017, while its effect and evaluations are hard to find online.

4.3 Synthetic Control Output

Given information about the treatment, control, dependent variable, and time periods, the synth function creates matrices averaged over the pre-treatment predictors for Chicago and the control unit. And Combine with the outcome variable in the 2010, 2011, 2012, 2013,2014 as pre-trend. We can calculate the optimal weights over the control group by solving:

$$W = \underset{W}{\operatorname{argmin}}(Z_1 - Z_0 * W)'(Z_1 - Z_0 * W)$$

where Z1 (treatment)and Z0(control) are vectors containing observed pre-period covariates, and V is variance matrix.

Inputting our data, we got synthetic weights as followings:

Unit Weights:

Co_No	Unit_Weight
Cheyenne	.224
Fargo	0
Houston	0
Las Vegas	.007
Los Angeles	0
Milwaukee	0
New York	.032
Omaha	0
Philadelphia	.666
Wilmington	.072

Figure 16. Synthetic Weights

Clearly, the synthetic Chicago and real Chicago start split from 2015. Imply that highly likely there is some policy or

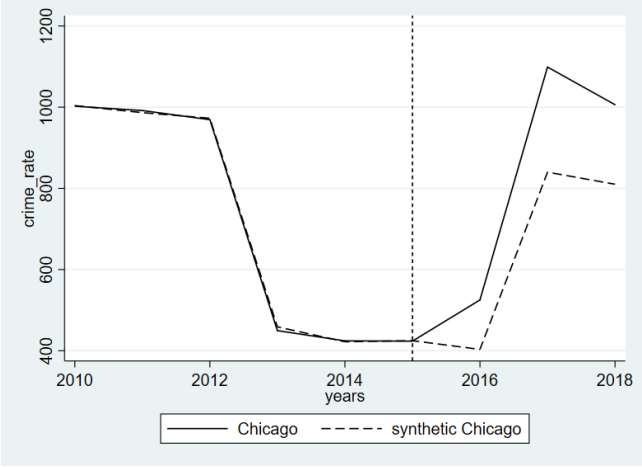


Figure 17. Trends of Real and Synthetic Chicago

action related to non-demographic features is responsible for part of the Chicago violent spike. However, from the synthetic Chicago we can see that there still exists a large spike, meaning that despite the potential treatment policy, the large proportion of the increase may due to the Chicago's crime rate Natural change or Periodic change.

In order to test the treatment significance, we will run a nonparametric permutation test:

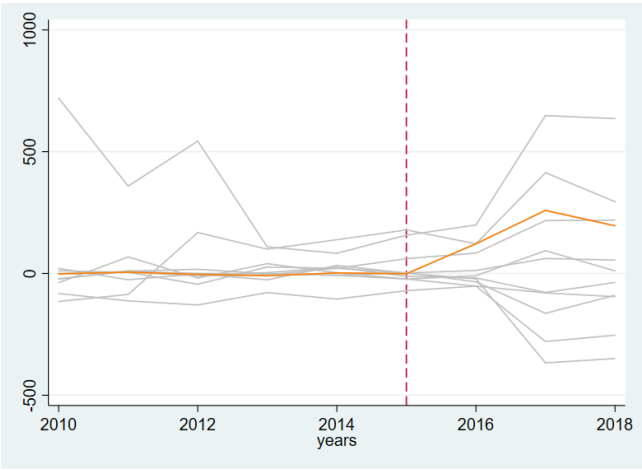


Figure 18. Nonparametric Permutation Test

From the test plot we know that, If the treatment effect were significant, we would expect to see that the orange line was visually below or above the cluster of gray lines after the beginning of the treatment year, indicating that the gap between the treatment unit and synthetic control is not consistent with the null hypothesis(treatment effect is zero). However, we see that the orange line is certainly not in the lower 1%, 5% of treatment effects; thus, we cannot graphically reject the null hypothesis. On the other hand, some of the cities have very poor pre-trend fits.

4.4 Conclusion of Policies and Synth Method

In this part, we qualitatively researched on urban renewal policies and projects of similar cities and conducted synth test to evaluate effect of such policies.

To conclude, we will say that high crime rate is a complex outcome of various social, economic, political, and human conditions. In Linear Regression section, we proved that demographic factors alone are not sufficient and infrastructure plays a crucial role. Here using synth method, we found policies related to non-demographic features indeed responsible for part of the Chicago violent spike, even though it's still not the full picture. Other plausible explanations include public emergent issues like recent demonstration and global crisis like COVID-19, or simply natural change and periodic fluctuation.

5 Policy Recommendations

Based on machine learning results, we provide policy recommendations from three perspectives: launching large-scale urban renewal projects, reducing direct and discriminatory action on individuals and groups, and lastly cooperating closely for unified management.

First, facilitating infrastructural development and launching abandoned houses renewal programs will accelerate urban revitalization and thus help control crime rate. More importantly, based on our regression result, urban renewal projects should prioritize renovating abandoned houses and removing graffiti. Presence of obsolete buildings and messy graffiti sends signal of disorder and ungoverned, which causes negative impressions and directly triggers criminal conducts. Size of negative effect of these factors is profound. Yet, on the other hand, positive effect brought by public cultural institutes, like library and art institutes, is quite limited. The program $1 for abandoned houses that achieved success in Cheyenne and Philadelphia can be used as a good reference.

Second, since virtually no demographic features have significant relationship with tract crime rate, prediction on potential crimes based on individual's demographic features and police's direct action taken on certain groups should be reduced. This point is supported by plenty of other research and practices. One instance is the restriction on stop-and-frisk, which once believed would result in violent crime spike. Yet in fact, U.S. cities that saw sharp declines in street stops didn't experience rapid increases in homicides and other violent crimes afterwards. In 2011, the New York Police Department began to abandon its stop-and-frisk policy amid widespread public criticism of the practice and an ongoing legal battle over its constitutionality. Chicago has taken active action on this as well; "around Dec, 2015 Chicago agreed to complete an additional report after every street stop on January 1, the same day a state law came into effect that added further restrictions". We believe this policy should

be carried on and be expanded to other procedures in law enforcement actions.

Lastly, cooperating with neighbor cities closely and developing unified strategies are vital in combat crimes. Strong mutual effect indicating by traffic flow index between tracts demonstrates that no city can detach itself from the interconnected world; only through collaboration can cities restore peace and safety together.

6 Ethics

Our research suffers potential ethical issues and bias from three aspects.

Firstly, many people and traditional researchers may think that there is some positive correlation between percent of black and crime rate. However we do not agree with them. The traditional model does not take features such as the abandoned housing, the living environment of black people, and authority's attitudes (like police may be more likely to convict black people) into consideration. These features result in bias in many traditional models. And In order to address these potential biases, we included those variables such as abandoned housing, graffiti count, and included more than two races as well. We put our best efforts to eliminated bias, yet there are still possible some features are ignored.

2. For the unsupervised learning, we use different states' demographic features for clustering. However, the major city of the state may not be able to fully represent the state and even though in the same group, they may not represent other cities as well. Therefore, directly comparing them with each other will lead to bias. In order to address this problem, we adopt the synthetic control method that allows us to simulate a synthetic Chicago sharing similar demographic pattern as the real Chicago.

3. Another potential bias is about law and definition of crime. States may use slightly different definitions in data collection. Since it's almost impossible to unify our cross-cities crime data under one same principle, in the third part when introducing other state cities as control, we start to focus on violent crime only. Definition of violent crime (murder, rape, and so on) is rather consistent in all states.

7 Limitations and Suggestions for Future Work

The major limitations of this research are insufficiency in data and inconsistency in statistical methods. First, though most large cities provide open data on a uniform City Open Data System, items included and statistical methods vary city by city. For instance, count of abandoned houses and number of graffiti complaints are the two central features in this research; Chicago provide yearly data on them, while New York City only has graffiti after 2016 and Los Angeles has neither. Meanwhile, statistical scale varies year by year even for the same variable within the same city, which increases difficulties for our comparative studies.

Second, Census Bureau, FBI, and City Data Portal different geospatial identification system and adopt different statistical scales. Census Bureau provides data using tract as the minimal unit, but FBI considers city as a whole. We found a state-county-city relational dataset to reconcile, while since there're duplication of city names and nonstandard name input, some have to be dropped because of no match even after using Jelly Fish fuzzy matching. Hence, we hope in future, a uniform system of data collection and geographical identification could be used across bureaus.

For future work, if more data available, we plan to push unsupervised learning to county level and select counties similar to Cook County; this will yield more accurate predicting results. Similarly for synth method, inputting more variables will help researchers clearly identify the most important factor. Also, with more data in different types, more other machine learning methods, especially random forest and neural networks, can be used for method comparison and for more reliable prediction.

References

[1] Diamond A. Hainmueller J. Abadie, A. 2007. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program. (2007), 635–644. https://doi.org/10.3386/t0335

[2] Ribeiro H. V. Rodrigues F. A. Alves, L. G. 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications* (2018), 505, 435–443.

[3] M. Chambers. (2019, September 09). *Northwest Philadelphia: Jumpstart Germantown Looks to Revitalize the Community One Property at a Time.* Retrieved June 2020 from https://philadelphianeighborhoods.com/2017/05/30/northwest-philadelphia-jumpstart-germantown-looks-to-revitalize-the-community-one-property-at-a-time/

[4] W. Delaware. (2015). *Wilmington, DE.* Retrieved June 2020 from https://www.wilmingtonde.gov/government/city-departments/licenses-and-inspections/vacant-property-registration-program

[5] R. Doyle. (2015, March 17). *No One Seems to Want Cheyenne's 'Free' Frontier Homes.* Retrieved June 2020 from https://www.curbed.com/2015/3/17/9980056/free-1890s-house-in-cheyenne-wyoming

[6] Lederman D. Loayza N. Fajnzylber, P. 2002. What causes violent crime? *European economic review* 46, 7 (2002), 1323–1357.

[7] R. Florida. (18 Oct. 2019). *The Great Crime Decline Is Over in Some Chicago Neighborhoods.* Retrieved June 2020 from www.citylab.com/equity/2019/10/chicago-crime-rate-statistics-data-violence-segregation/599317/

[8] M. Ford. (2017, January 25). *What's Actually Causing Chicago's Homicide Spike?* Retrieved June 2020 from https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/

[9] Javideh M. Ebrahimi M. R. Keyvanpour, M. R. 2011. Detecting and investigating crime by means of data mining: a general crime matching framework. *Procedia Computer Science* 3 (2011), 872–880.

[10] M. Mullen. (2017, August). *Fight The Blight Campaign Reveals Affordable Housing Problem In Cheyenne.* Retrieved June 2020 from https://www.wyomingpublicmedia.org/post/fight-blight-campaign-reveals-affordable-housing-problem-cheyenne

[11] L. Nagengast. (2016). *Land bank seeks to address Wilmington's abandoned and blighted properties.* Retrieved June 2020

from https://www.delawarepublic.org/post/land-bank-seeks-address-wilmingtons-abandoned-and-blighted-properties

[12] S. V. Nath. 2006, December. Crime pattern detection using data mining. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* (2006, December), 41–44.

[13] E. NYC. (2015). *Graffiti-Free NYC*. Retrieved June 2020 from https://edc.nyc/program/graffiti-free-nyc

[14] Williams C. Purcell D. Vargas, C. (2020, March 03). *Philly gave away 2,300 properties for $1 each. Now it is owed nearly $900,000 in back taxes.* Retrieved June 2020 from https://www.inquirer.com/news/philadelphia/philadelphia-1-dollar-properties-back-taxes-land-bank-councilmanic-prerogative-20200303.html

[15] Kifer D. Graif C. Li Z. Wang, H. 2016, August. Crime rate inference with big data. *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016, August), 635–644.