Crime Analysis and Prediction in New York City

prepared for

STAT4201 Advanced Data Analysis

by

Group No.: 4
Bharath Sharma (bcs2150)
Mayank Mahajan (mm4399)
Parth Pareek (pp2547)
Shruti Sinha (ss4940)

Abstract

This project involved use of various geographical, temporal, economic and socio-demographic information, in combination with spatial techniques, to analyze crime patterns and trends. The goal was to understand the underlying factors affecting crime rates like income, education, female-male ratio, age distribution and facilities in a locality at a tract level in New York City. A Poisson regression was explored as a method for predicting the number of felonies at any tract and also determining the most significant predictors. Other regression techniques such as random forests were also tested and compared. Geometrically Weighted Regression (GWR) was employed to test for spatial relationships in felony counts and to further investigate anomalies obtained in the OLS. The results suggest that several structural measures that vary with location have a relationship with crime rate and the global model for crime prediction could be erroneous.

Introduction

The goal of the project was 3 fold - to understand the crime data for NYC, be able to predict crime and to propose safer neighborhoods in NYC for new residents.

The data sources used for the analysis are listed below:

- NYPD incident level crime data (with approximately 75,000 felonies for first 9 months of 2015)
- US Census data for various demographic covariates
- NYC facilities data

The data was rolled up at a US census tract level. A tract is approximately the size of 2-3 blocks (much smaller than a zip code) and can be considered homogeneous in terms of demographic variables.

Crime Statistics in New York City

The incident level crime data was analyzed by time, borough and types of crime.

- No particular seasonal trend in crime was seen for 2015 data except that the number of felonies increase in the summer months
- Highest number of felonies have been reported in Brooklyn followed by Manhattan
- There are 7 major types of felonies out of which grand larceny, assault, robbery and burglary are particularly higher in frequency

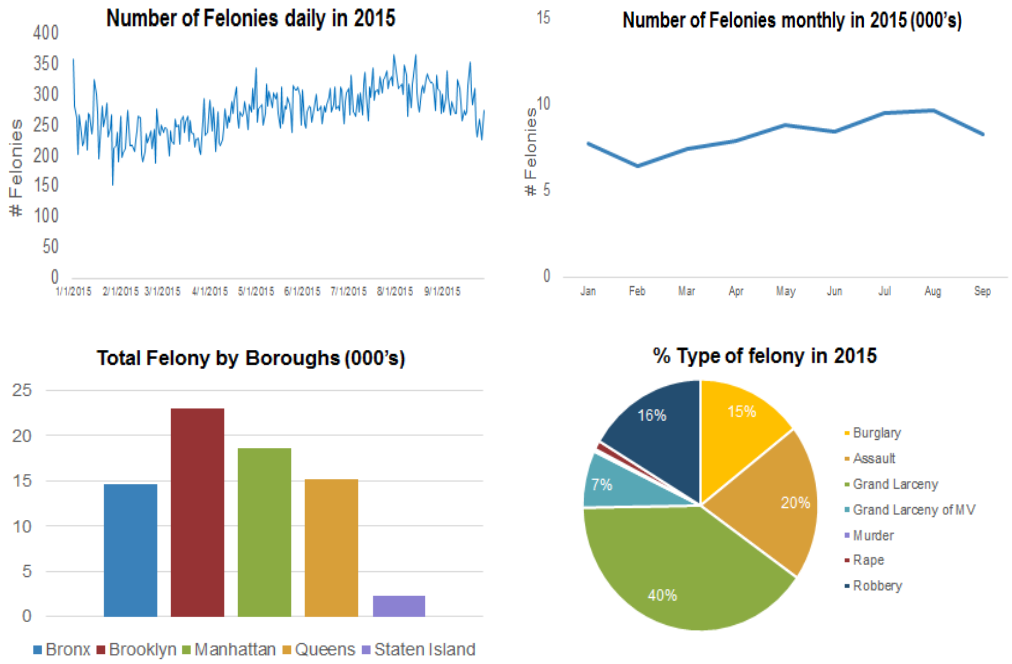


Figure 1: Crime statistics in NYC

Exploratory Data Analysis

Univariate analyses of individual variables were performed to understand the data distributions followed by them. The 2 major variables, total felony and total population, are right skewed. Transformations like log, log-log, square-root were tried on these 2 variables to get to normality as shown below in Figure 2.

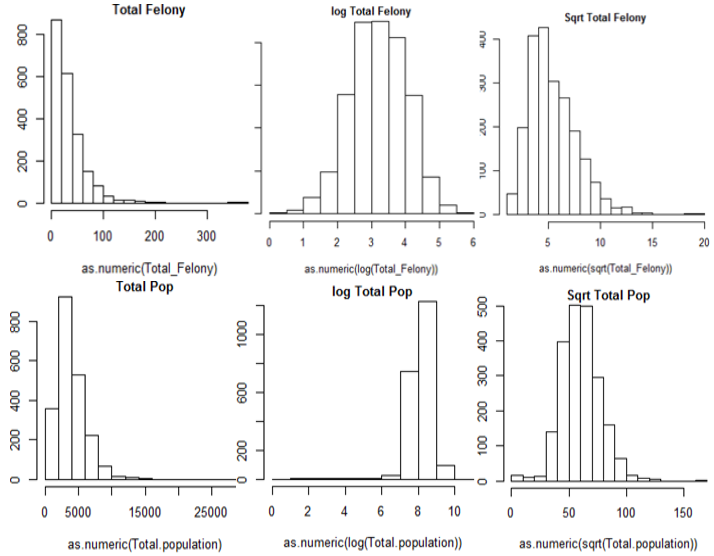


Figure 2: Distributions of Total felony and population

Similar to total felony and population, most of the variables were not normally distributed. On the other hand, % population with college degree had a normal distribution (Figure 3). It was also interesting to note that median household income followed a uniform distribution as seen below.

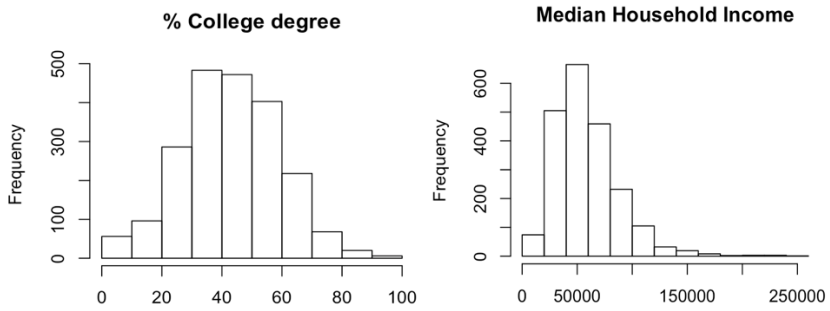
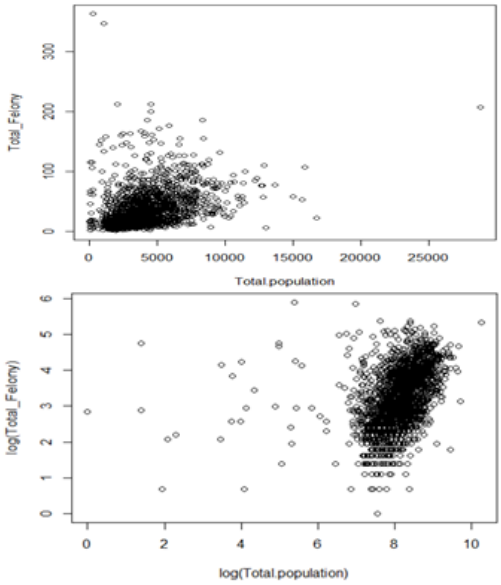


Figure 3: Distributions of Household income and Education

Bivariate analysis between different pairs of untransformed and transformed variables was performed. The only pair which had a slightly linear relation were total felony and population after log transforming both variables. But as none of the other covariates showed a linear relation with total felony and the model did not show much improvement using log population, population was kept untransformed in the final model.

The highest correlated variables with total felony were seen to be population, household income and number of homeless shelters in the tract with population and shelters having a positive correlation and income having a negative correlation (Figure 4).



	Total_Felony
Total_Felony	1.000000000
Total.population	0.382393510
Median_Household_Income	-0.333527312
LogTotal.population	0.289837669
Homeless_shelters	0.241181913
Pharmacy	0.230913006
Schools	0.204564580
Health	0.204310888
Percentage_college_degree	-0.191222894
Mental_Heath	0.181027458
Crimi0l_Justice	0.171322417
Daycare	0.170212283
Senior_homes	0.153610382
Race_White	-0.110496599
Recreatio0l	0.103999205
Disabilities	0.100917488
Race_Other	0.094878614
Race_Black	0.093927320
Race_Asian	-0.080811834
Over_18_year	0.077419484
Unemployment_rate	-0.069480099

Figure 4: Distributions of Felony vs. Population; Correlation of Felony vs. other covariates

Subset Selection and Model Building

Three poisson GLM models were built for the analysis:

- **Population model** - with only population as the covariate
- **Best model** - using "bestglm" function to perform subset selection, the best set of variables were chosen
    - Covariates in best model: Population, Median income, %male, age, %college degree, race(race_black, race_asian) and facilities like criminal justice, pharmacy and homeless shelters
- **Full model** - with all the covariates included

Next, the 3 models were compared with each other to understand the amount of variance in number of felony explained by each of them. There was seen a huge drop in deviance from the population model to the full model.

This implies that population alone is not enough to explain felony frequencies. But the Δ drop in deviance for best vs. full model was only 733. This suggested that the best model explains most of the variance in the number of felony and that the additional 15 covariates in the full model should not be included in the model as they do not seem to add a lot of value (Figure 5).

	Residual Deviance	# Covariates
Full Model	20,311	25
Best Model	21,043	10
Population Model	33,960	1

	Full vs Population Model	Best vs Population Model
Drop in Deviance	13,649	12,916
Additional Covariates	(25-1)=24	(10-1) = 9

Figure 5: Comparison of full, best and population models

The final poisson GLM model after subset selection had an AIC of 29,604 and residual deviance of 21,043. The model coefficients and standard errors (in parentheses) were as below:

Total Felony = 2.76 + 8.66e-05*population + 0.017 * age + 0.015 * sex + 0.0044 * race_black
(0.0324) (1.4e-06) (0.00058) (0.00102) (0.00015)

+ 0.092 * criminal_justice + 0.126 * pharmacy + 0.052 * homeless_shelter
(5.012e-03) (0.0054) (0.00397)

- 6.77e-03 * education - 4.39e-04 * median_income - 0.0021 * race_asian
(0.000258) (7.88e-06) (1.82e-04)

All the covariates above are significant with p-values < 0.00001.

Model Validation and Comparison

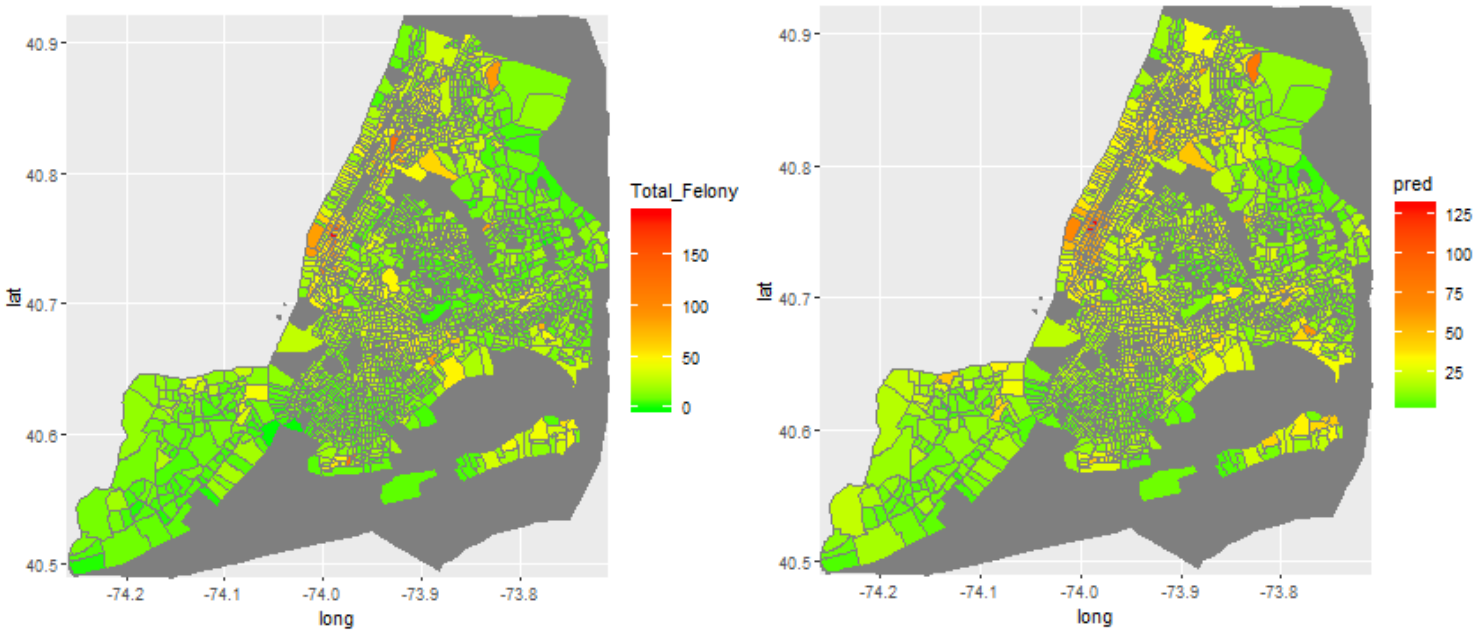
The tract level dataset with number of felony and other demographic and facilities data was split into 80-20 to form a train and test datasets (by random selection of tracts). The 80% of the data was used for model building and the rest was used for model validation. The mean square error (MSE) between the train and test set was calculated to see how each of the model performs.

Different types of model employed were linear model, GLM poisson, Trees and Random forest for the purpose of finding the best model for predicting crime rates. As seen in the table below the random forest has the lowest MSE between test and train datasets and hence it was the model of choice for crime prediction (Table 1).

Using random forest on the 2015 felony data, felony counts were predicted as seen in the heatmap in Figure 6.

Model	Difference in MSE for train and test data	
	With all covariates	Covariates after subset selection
Linear Model	638.46	671.66
GLM - Poisson	679.26	717.30
Trees	644.57	636.41
Random Forest	**546.09**	**559.50**

Table 1: Comparison of different models of crime prediction



Actual Felony Heatmap **Predicted Felony Heatmap**

Figure 6: Prediction using Random Forest for 2015 data

Geometrically Weighted Regression

The issue with a regular OLS solution to predict crime is that OLS estimates global coefficients for the regression, and hence, variations over space are suppressed. Therefore, in the previous Poisson regression, it is assumed that the relationship between demographic parameters and crime level is constant throughout NYC, which can be grossly inaccurate, especially for a large city like NYC. For this purpose, we introduce the geographically weighted regression (GWR), to obtain global and spatially local estimates of the coefficients in order to capture spatial relationships. The GWR is expressed as:

$$y_i = \sum_{k=0}^{n} \beta_{ki} x_{ki} + \varepsilon_i$$

Where β_{ki} is the value of β_k at point i. Each point has a corresponding nxn weight matrix W_i with off diagonal elements equal to 0 and diagonal represents geographical weighting of observed data for point i.

$$W_i = \begin{pmatrix} w_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{in} \end{pmatrix}$$

$$W_{ij} = \exp\left(\frac{d_{ij}^2}{h^2}\right)$$

Where d_{ij} represents the distance between centroids of tracts i and j. The bandwidth parameter, h, can be pre-specified or determined by the model selection.

$$\widehat{\beta_\iota} = (X^T W_i X)^{-1} X^T W_i X$$

The summary statistics of the estimates of the GWR are presented below along with the analysis of deviance table

Variable	**Mean**	**SD**	**Min**	**Max**
Intercept	3.35	0.49	-1.10	5.05
Total Population	0.31	0.21	-0.86	1.60
Median Income	-0.13	0.22	-1.40	2.44
Percentage Male	0.02	0.32	-2.34	6.59
Over 16 years	0.06	0.29	-2.64	1.24
Over 60 years	-0.07	0.35	-1.46	7.07
Percentage with College Degree	-0.04	0.16	-1.45	1.19
Unemployment Rate	0.00	0.33	-3.41	4.92

Table 2: Geometrically Weighted Regression Summary Statistics

Source	**Deviance**	**DOF**	**Deviance/DOF**	**Source**
Global model	28350.98	2094.00	13.54	Global model
GWR	7160.18	1004.62	7.13	GWR
Difference	21190.80	1089.38	19.45	Difference

Table 3: Deviance values for different models

The t-values for two of the parameters, Percentage of population over 16 and Median Household Income are plotted below. Blue regions are regions where the parameter is statistically significant negative value. Red regions are where the parameter is a statistically significant positive value.

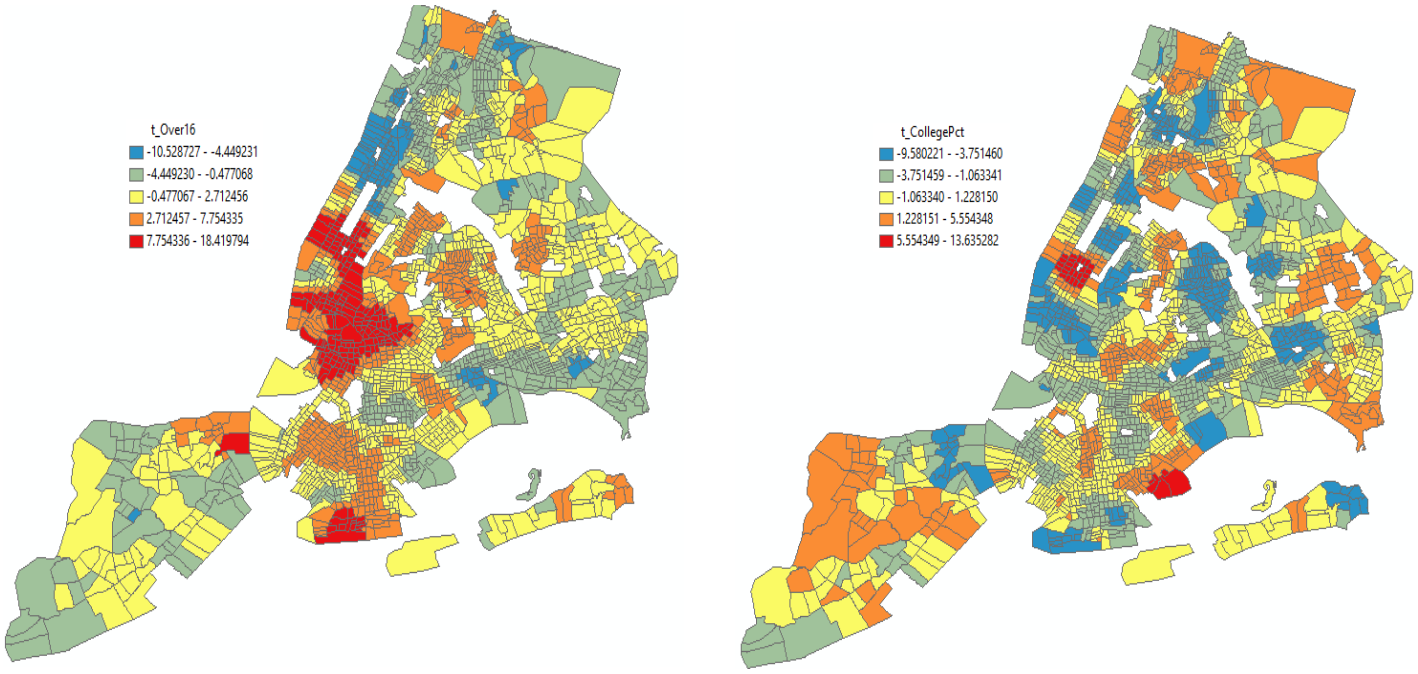


Figure 7: Results of geographically weighted regression

As we can see from the above diagrams, global estimates for the coefficients lead to erroneous results as they do not account for spatial relations. For example, the percentage of population above 16 years has a positive influence on crime rates in lower Manhattan and upper Brooklyn but has a negative influence on crime in upper parts of Manhattan and other places, reasons for which would require further investigation. Similarly, every coefficient takes a range of values depending on the spatial location.

GWR also helps investigate certain anomalous results obtained in the regular OLS model. The OLS showed a negative relationship between Unemployment and number of crimes. However, using the GWR, we see that this negative relationship is obtained due to few outliers which skewed the global estimate to a negative relationship. However, most of the tracts showed a positive relationship between unemployment and crime.

Housing zone predictive analysis

We used the collated dataset to predict the best housing zones on a set of parameters. Crime rate, age, median household income, race, and facilities available were used to cluster all the tracts into 10 zones. Expectation Maximization algorithm was used to perform the clustering. After the clustering was performed, housing requirement information is fed – household income, age of buyer, race, and facilities needed, and probability of buying house in each zone is predicted. The data was scaled for the calculations to avoid stronger effect of any variable.

See housing zone prediction in Figure 7 below for income = 100,000, age > 18 and < 60, college degree holder, Asian, facilities needed – schools, healthcare, pharmacy, and transport.

Figure 8: Prediction of housing zones

Conclusion

This study provides us a few insights on crime patterns in NYC. While the Global Poisson regression did capture most of the key parameters and presented logical relationships between crime and demographic and geographical predictors, it did not sufficiently capture the information in the training data. It was evident that a global model would not be able to predict crime accurately in a large city such as New York and spatial relationships were being suppressed. The Geometrically Weighted Regression (GWR) fit a better model and answered some anomalies obtained in the OLS. The models presented key predictors, such as percentage of people above 16 years of age, number of homeless shelters, unemployment rate, male to female ratio, and median income which were aligned with our intuition. However, further investigation is required to obtain a concrete crime prediction model.