

Predicting Crime rates based on spatial environment <Tania Vara, TaniaVM, tvn237>

Abstract

The objective of this project is to find common factors in the population towards a better identification in the incidence of crime across the precincts in New York and through this factors create a model to try to identify precincts that are more susceptible to this problem. The factors measured in this study are Unemployment Rate, School Dropout Rate and different Complaints by precinct trying to identify the relationship between the variables mentioned and predict the phenomenon of crime. The reached conclusions are that Unemployment and Drop out rate are highly correlated in addition with complaints about animal abuse. The model using most of the factors reached a 94% of R-square.

Introduction

Criminological research has shown that crime can spread through local environments via a contagion-like process (Johnson 2008) [1]. The research here discussed focuses on finding this common local environments of Crime through New York in order to have a possible predictor using characteristics of the environment.

Crime does not need to be capital (nor d School, Drop , Complaints... is, can we , or is: "can we... The question to answer is, Can we measure and predict Crime rates based on unemployment rate, School Dropout Rate, and certain kind of Complaints?

The measure of this problem is important across urban studies to help to reduce the incidence of Crime. In previous studies from the College of Arts and Sciences Nova Southeastern University, they explored a social interaction model for the evolution of the attractiveness of the crime environment for criminal activity. The environment in which the crime occurs may play a role in the generation and accessibility of crime opportunities and may even provoke criminal activity. The attractiveness for one crime type as the description of social interaction given is representative of a 'communication of risk' about areas where crime has occurred. [2] move the reference up to the first mention of the study

The first step used in this study was making a Pearson test to find the correlation with the different factors mentioned before. The factors that showed more correlation were plotted inside a map and compared with a map showing the "hot spots" across New York in Crime.

The second step was creating a multiple regression to quantify the relationship between the dependent variable in this case is the number of events of the crime and independent or predictor variables.

The method used was a 'backward stepwise regression' (using OLS) to get the combination of weights of different factors across the precincts that give us the higher correlation in this case measured by the "R-Square".

Data.

The data identified as available to answer the question are:

- *Crime by precincts in 2014:* From <http://data.beta.nyc/dataset>. New York City is divided into 77 geographical areas called precincts. This data contains information about the number of crimes occurred in each precinct in a period of time. The period measured is from 2014 as a complete period of time.
- *Dropout and Unemployment rate:* From <http://www.socialexplorer.com/tables/>, from this web page was obtained the Dropout Rate and Unemployment rate. Information of Census Survey: *ACS 2013 (5-Year Estimates)*.
- *Complaints of 311 data.* From: <https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9>. This data was used as a parameter for the kind of environment that the crime zone has. The Complaints was filtered in the period of 2014 and just the Complaints related to the NYPD agency. This data contains the exact address where the call was made.
- *Shape files of precincts and Census tracks in New York:* http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml

Weakness of the data.

very good!

The number of Complaints could be biased by income, there is a known tendency to over report in wealthy neighborhood: the wealthy population is "more engaged" and tends to call 311 more easily, whereas in poorer neighborhoods people, possibly because of less trust in the government, they feel less entitled to service, are less likely to call. **excellent: most people using 311 data missed this**

Dropout Rate and Unemployment rate was measured by census track and was joined through a spatial join with the shape of the precinct. The sizes of the shape are not equal, then the data is an approximation of the real rates per precinct.

Data Wrangling:

1. *Mapping the number of crimes per precinct.* Using the number of crimes per precinct and the shape file of precincts in New York. This data was joined by the number of precinct.
2. *Mapping the number of Drop Outs and unemployment per Census track.* Data was obtained by Census survey; this data is organized by Census track. To join the data, the steps were the following:

Shape file per census track contains the following data:

Shape	CTLabel	Boro Code	BoroName	CT2010	BoroCT2010	CDEligibil	NTACode	NTAName	Shape_Length	Shape_Area
Polygon	1	2	Bronx	100	2000100		BX98	Rikers Island	18903.35	18154596
Polygon	1	1	Manhattan	100	1000100		MN99	park-cemetery-etc-Manhattan	11023.05	1844421

Census file contains the following data:

Geo_QName	Geo_TRA CT	Total Population	Civilian Population 16 to 19 Years	Not high school graduate, not enrolled (dropped out)	High school graduate, or enrolled (in school)	Civilian Population in Labor Force 16 Years And Over	Employed	Unemployed
Census Tract 1, Bronx County, New York	100	9191	1246	1057	189	0	0	0
Census Tract 34, Kings County, New York	3400	3331	137	1	136	1842	1746	96
Census Tract 39, New York County, New York	3900	5253	132	0	132	2960	2732	228
Census Tract 91, Queens County, New York	9100	2717	97	0	97	1774	1679	95
Census Tract 248, Richmond County, New York	24800	5065	321	0	321	2388	2240	148

3. *Join both tables:* Identify the names of each county in both tables and create a key that contains the same information in both tables in this case is the necessary information in both tables is column BoroCT2010:

Geo_QName	Geo_TRACK	BoroName	BoroCode	BoroCT2010
Census Tract 1, Bronx County, New York	100	Bronx		2 2000100
Census Tract 39, New York County, New York	3900	Manhattan		1 1003900
Census Tract 91, Queens County, New York	9100	Queens		4 4009100
Census Tract 34, Kings County, New York	3400	Brooklyn		3 3003400
Census Tract 248, Richmond County, New York	24800	Staten Island		5 50024800

4. *Create the column BoroCT2010 in the Census file:* Subtract from Geo_QName, the name of the county, then add the BoroCode in this table and create the column Geo track with 6 digits.

Geo_QName	NEW Geo_QName	Geo_TRACK	NEW Geo_TRACK
Census Tract 1, Bronx County, New York	Bronx	100	000100

- 4.1 Identify each Geo_QName with its correspondent number:

NEW Geo_QName	NEWBoroCode
Bronx	2
Manhattan	1
Queens	4
Brooklyn	3
Staten Island	5

this level of detail about the data and joining steps is not needed in the report, especially if you provide the code that supported your research

- 4.2 Join "NEWGeo_TRACK" with "NEWBoroCode" and I obtained the new BoroCT2010 to join both tables.

NEW Geo_TRACK	NEWBoroCode	BoroCT2010
000100	2	2000100
003900	1	1003900
009100	4	4009100
003400	3	3003400
024800	5	5024800

5. *Spatial join.* With the data per precinct, the spatial join was performed using GIS tools. The intersection is made by joining both shape files, I this case precincts and census tracks. The final join groups the data in the census by sum of cases per precinct. This was the last step to have the final table of the data. All the factors were normalized by the population in each precinct.

Methodology. 1.Hotspots and its environment.

Identifying hotspots is the first step a policing or crime reduction agency needs to take when discerning where best to prioritize their resources. [3] In addition, identifying the factors that are more correlated in the incidence of crime could give a framework of the important characteristics of each precinct. In order to know which factors were more correlated with the incidence of Crime a Pearson

Test were performed. This analytical tool measures the Correlation between sets of data and shows how well they are related. Pearson test gave the following results:

Dependent variables			Pearson Test	
SOURCE	NAME OF PREDICTOR	Variable	R square	P Value
Census Data	DropOut Rate Normalized by Population per precinct	X1	(0.71369827402173003,	3.2181586290079372e-13)
	Unemployed Rate normalized by population in labor force	X2	(0.82474972339379748,	2.9946901442061772e-20)
311 Complaints filtered by agency: NYPD and Normalized by population per precinct	Animal Abuse	X3	(0.74497243160624382,	7.95237296942606e-15)
	Blocked Driveway, Traffic Illegal and Parking	X4	(0.63326207440169335,	6.4011355336111015e-10)
	Derelict Vehicle	X5	(0.55654797659306232,	1.4740584556986782e-07)
	Disorderly Youth, Graffiti, BikeRoller Skate Chronic	X6	(0.055942414316575942,	0.62892378252634262)
	Drinking	X7	(0.65060756970764244,	1.5071729467908762e-10)
	Illegal Fireworks	X8	(0.54192325552656295,	3.5801298545643582e-07)
	Noise	X9	(0.3843464767089001,	0.00055844841216215913)
	Panhandling Homeless Encampment	X10	(0.1211377993776749,	0.29396457878553955)
	Vending	X11	(0.115938180774041,	0.31532837168888994)
	311 total of Complaints	X12	(0.65467926473889781,	1.0588263036873779e-10)
NYPD Open Data	Events of Crime in the precinct	Y		

As it can be observed in the table above, the factors of Drop out Rate, Unemployment rate and the complaint about animal abuse are the factors that are highly correlated with the incidence of crime.

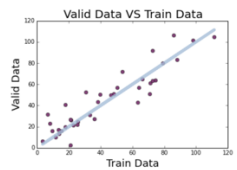
To see this high correlation easily the factors were mapped per precinct ordered by its standard deviation. The red ones show the precinct with the higher rate. (Additional figure 1)

1.Forecasting number of crimes in the precinct based on its environment. To perform this point a regression using ordinary least squares (OLS) was used, a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data. This method was computed using all the regressors and a 'backward stepwise regression' was made. This regression calculates the R Square using all the parameters of the model. Then exclude one regressor and compute again the model. This was made with all the regressors in order to maximize the R- square and exclude the factors that are minimizing the R-square.

With this process the best combination for predicting give us the following results:

excellent!!

Taking out X8, and X10, the regression is better with a R square of 0.949

<table><tr><td>Dep. Variable:</td><td>CrimeNorm</td><td>R-squared:</td><td>0.949</td></tr><tr><td>Model:</td><td>OLS</td><td>Adj. R-squared:</td><td>0.930</td></tr><tr><td>Method:</td><td>Least Squares</td><td>F-statistic:</td><td>51.82</td></tr><tr><td>Date:</td><td>Tue, 15 Dec 2015</td><td>Prob (F-statistic):</td><td>2.18e-15</td></tr><tr><td>Time:</td><td>01:45:35</td><td>Log-Likelihood:</td><td>-151.24</td></tr><tr><td>No. Observations:</td><td>38</td><td>AIC:</td><td>322.5</td></tr><tr><td>Df Residuals:</td><td>28</td><td>BIC:</td><td>338.8</td></tr><tr><td>Df Model:</td><td>10</td><td></td><td></td></tr><tr><td>Covariance Type:</td><td>nonrobust</td><td></td><td></td></tr></table>	Dep. Variable:	CrimeNorm	R-squared:	0.949	Model:	OLS	Adj. R-squared:	0.930	Method:	Least Squares	F-statistic:	51.82	Date:	Tue, 15 Dec 2015	Prob (F-statistic):	2.18e-15	Time:	01:45:35	Log-Likelihood:	-151.24	No. Observations:	38	AIC:	322.5	Df Residuals:	28	BIC:	338.8	Df Model:	10			Covariance Type:	nonrobust			<table><tr><th></th><th>coef</th><th>std err</th><th>t</th><th>P> t </th><th>[95.0% Conf. Int.]</th></tr><tr><td>X1</td><td>1.8617</td><td>0.608</td><td>3.063</td><td>0.005</td><td>0.617 3.107</td></tr><tr><td>X2</td><td>0.1388</td><td>0.064</td><td>2.163</td><td>0.039</td><td>0.007 0.270</td></tr><tr><td>X3</td><td>3.4381</td><td>8.753</td><td>0.393</td><td>0.697</td><td>-14.492 21.368</td></tr><tr><td>X4</td><td>0.2027</td><td>7.978</td><td>0.025</td><td>0.980</td><td>-16.139 16.544</td></tr><tr><td>X5</td><td>0.2357</td><td>8.086</td><td>0.029</td><td>0.977</td><td>-16.327 16.799</td></tr><tr><td>X6</td><td>-10.1370</td><td>21.142</td><td>-0.479</td><td>0.635</td><td>-53.444 33.170</td></tr><tr><td>X7</td><td>-6.9781</td><td>15.798</td><td>-0.442</td><td>0.662</td><td>-39.339 25.383</td></tr><tr><td>X9</td><td>0.1255</td><td>8.096</td><td>0.015</td><td>0.988</td><td>-16.459 16.710</td></tr><tr><td>X11</td><td>1.5881</td><td>13.054</td><td>0.122</td><td>0.904</td><td>-25.152 28.328</td></tr><tr><td>X12</td><td>-0.2066</td><td>7.972</td><td>-0.026</td><td>0.980</td><td>-16.537 16.124</td></tr></table>		coef	std err	t	P> t	[95.0% Conf. Int.]	X1	1.8617	0.608	3.063	0.005	0.617 3.107	X2	0.1388	0.064	2.163	0.039	0.007 0.270	X3	3.4381	8.753	0.393	0.697	-14.492 21.368	X4	0.2027	7.978	0.025	0.980	-16.139 16.544	X5	0.2357	8.086	0.029	0.977	-16.327 16.799	X6	-10.1370	21.142	-0.479	0.635	-53.444 33.170	X7	-6.9781	15.798	-0.442	0.662	-39.339 25.383	X9	0.1255	8.096	0.015	0.988	-16.459 16.710	X11	1.5881	13.054	0.122	0.904	-25.152 28.328	X12	-0.2066	7.972	-0.026	0.980	-16.537 16.124	
Dep. Variable:	CrimeNorm	R-squared:	0.949																																																																																																					
Model:	OLS	Adj. R-squared:	0.930																																																																																																					
Method:	Least Squares	F-statistic:	51.82																																																																																																					
Date:	Tue, 15 Dec 2015	Prob (F-statistic):	2.18e-15																																																																																																					
Time:	01:45:35	Log-Likelihood:	-151.24																																																																																																					
No. Observations:	38	AIC:	322.5																																																																																																					
Df Residuals:	28	BIC:	338.8																																																																																																					
Df Model:	10																																																																																																							
Covariance Type:	nonrobust																																																																																																							
	coef	std err	t	P> t	[95.0% Conf. Int.]																																																																																																			
X1	1.8617	0.608	3.063	0.005	0.617 3.107																																																																																																			
X2	0.1388	0.064	2.163	0.039	0.007 0.270																																																																																																			
X3	3.4381	8.753	0.393	0.697	-14.492 21.368																																																																																																			
X4	0.2027	7.978	0.025	0.980	-16.139 16.544																																																																																																			
X5	0.2357	8.086	0.029	0.977	-16.327 16.799																																																																																																			
X6	-10.1370	21.142	-0.479	0.635	-53.444 33.170																																																																																																			
X7	-6.9781	15.798	-0.442	0.662	-39.339 25.383																																																																																																			
X9	0.1255	8.096	0.015	0.988	-16.459 16.710																																																																																																			
X11	1.5881	13.054	0.122	0.904	-25.152 28.328																																																																																																			
X12	-0.2066	7.972	-0.026	0.980	-16.537 16.124																																																																																																			
The model takes out the X8(Illegal Fireworks) and X9(noise) regressors for optimizing the model		This plot compares how well the model fits to the real data.																																																																																																						

Conclusions: The Unemployment rate, drop out rate and Complaints about Animal abuse are highly correlated with the number of crime through the precincts. Compared with the expectations, the Complaints about Animal abuse was a finding having a high correlation with the incidence of crime per precinct. The model for predicting was tried in the half of the sample and shows an accuracy of 94%.

Future work: The model could be improved by taking into consideration the income per precinct to exclude the fact that the number of complaints could be biases by income.

very very good work!

Bibliography

- [1] [1] Johnson, Shane D. "Repeat Burglary Victimization: A Tale of Two Theories." *J Exp Criminol Journal of Experimental Criminology* 4, no. 3 (08, 2008): 215-40. doi:10.1007/s11292-008-9055-3.
- [2] Haskell, Evan C. "A Social Interaction Model For Crime Hot Spots." *ECMS 2014 Proceedings Edited By: Flaminio Squazzoni, Fabio Baronio, Claudia Archetti, Marco Castellani*, 05, 2014. doi:10.7148/2014-0745.
- [3] Chainey, Spencer, Lisa Tompson, and Sebastian Uhlig. "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime." *Secur J Security Journal* 21, no. 1-2 (02 2008): 4-28. doi:10.1057/palgrave.sj.8350066.

Additional figure 1

