

Exploring Residential Crime Prediction with GIS - Demographic Profiles vs Top Offender Location: A Rochester, Minnesota USA Case Study

Ryan A. Robert

Department of Resource Analysis, Saint Mary's University of Minnesota, Winona, MN, 55987

Keywords: Geographic Information Science, Residential Crime, Crime Reduction, Intelligence-led Policing, Cluster Analysis, Hot-spot Detection, Kriging Interpolation, Statistical Analysis, Spatial Analysis

Abstract

This study focused on whether or not and to what extent certain demographic, spatial, and temporal characteristics influence crime within Rochester, Minnesota, USA. The research question proposed is: "To what extent do certain socio-economic and geospatial variables influence crime rates in Rochester, Minnesota? If significant correlations do exist, can the relationships be used to accurately predict future crimes? Finally, is offender location and offender density a better predictor of crime than the demographic makeup of the city?" Statistical software including SPSS, Excel, and ArcGIS Geostatistical Analyst were used to analyze relationships between crime rate and several independent variables. A geographic information system (GIS) was also used to spatially analyze data and help visualize patterns of crime locations. Correlation and regression results suggest that while socio-economic make up of block groups is a better predictor of crime than is top offender location, both methods produce relatively weak R^2 values of .22 and .12 respectively. The model for socio-economic variables predicts 82 out of 90 block groups within a standard error of -1 and 1. The model for offender location variable predicts 85 out of 90 block groups within a standard error of -1 and 1.

Introduction

The city of Rochester, Minnesota USA is home to a diverse population, largely due to the presence of the healthcare giant Mayo Clinic and technology giant, IBM. Rochester serves as the study area for this research (Figure 1). Over the last decade Rochester has experienced rapid population growth, increasing in population 20% since 2000. This represented the largest rate of population growth of any Minnesota city and

launched Rochester into the rank of third on a list of largest cities in Minnesota (MCDC, 2012). The period of rapid growth experienced in Rochester occurred simultaneously with a period of budget cuts to local and state government entities, including areas pertaining to law enforcement agencies. Police departments need to be able to show efficiency in terms of most effectively allocating limited, financial and human resources.

GIS and crime mapping have been used to produce prediction models that

have helped to alleviate hours spent patrolling areas which are not likely to have high crime rates. GIS technology has also helped evolve crime mapping and analysis from the use of simple, hard copy pin maps to real-time data collection in an easier analyzed digital format (Ratcliffe, 2002). Major American cities like Baltimore, Maryland (Weisburd and McEwen, 1997), Cincinnati, Ohio (Davin and Lin, 2009), and Jersey City, New Jersey (Weisburd and Green, 1995) have all seen successful use of GIS in crime analysis efforts.

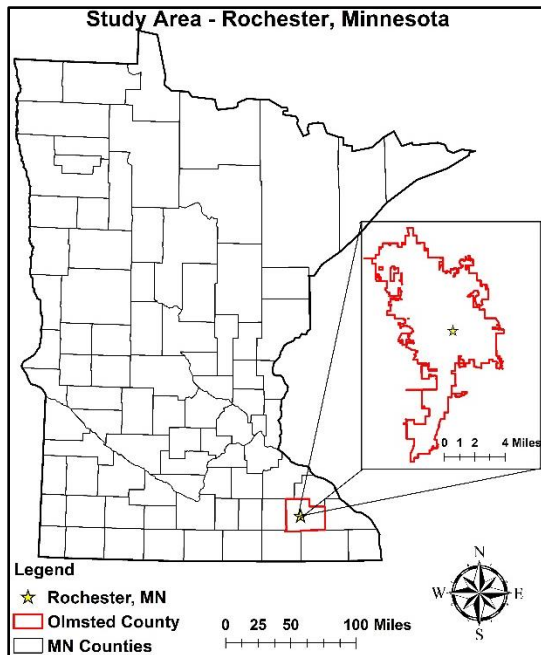


Figure 1. Study area location map. Located in southeastern Minnesota's Olmsted County, the city limits of Rochester (population 110,742) serve as the boundary for this research.

This research aims to develop an accurate model for forecasting crimes within the city of Rochester, Minnesota. A prediction model would help to allocate limited police department resources more efficiently and effectively. In order to create such a model, preliminary analysis of variable relationships was conducted. Preliminary research produced several maps visualizing the spatial distribution of crimes as well as the demographic make-

up of the city which can help in other aspects of city governance, including planning and zoning departments.

Methodology

Data Collection Process

Multiple data sets were acquired to complete this study. Geographic data, demographic data, zoning data, top offender data, and crime data were all necessities. Each data set was acquired as raw data, manipulated for GIS use, and then integrated into the GIS.

Geographic data needed included shapefiles for the city limits of Rochester and boundaries for census block groups. These were obtained from ESRI 2010 boundary files. The polygons served as a general study area for analysis.

Demographic data incorporated in this study included measures of education, income, population, employment, age, race, and home ownership. Previous studies have shown these variables may influence crime within an area (Sampson and Groves, 1989).

Crime data were obtained from the City of Rochester Police Department (RPD). Crime data obtained were for the calendar years of 2011-2013. Top offender data was provided by RPD. These data were provided electronically, via email and in the form of an Excel spreadsheet.

Demographic data were obtained from the Census Bureau website. Census 2010 data as well as census estimate data collected were Summary 3 tables aggregated to the block group level. The block group level was used as it was the smallest common area for which all data attributes were collected. For example, several demographic variables are only collected down to the block group level whereas others are collected at the smaller, block level.

Land use and zoning information was from 2012 and provided by the City of Rochester Planning and Zoning Department.

Data Preparation

Most of the data used in this project was in tabular form. Several steps were taken to ensure the data was consistent and spatially accurate. In order to conduct thorough analysis, each respective data set also needed appropriate data attributes. Perhaps the most important attribute of the entire project was crime location. Crime location and geographic information associated with a criminal event can provide clues about the identity of suspects, assist in the design of prevention or apprehension strategies, aid in the evaluation of programs, and help gain a better understanding of environmental factors that may be associated with crime (Brantingham and Brantingham, 1991). Ultimately, this attribute was used to derive the dependent variable, crime density. Thus, it was essential that the locations were spatially accurate. The methodologies used required point data (crime location) to fall within relatively small polygon features (block groups). Therefore, it was essential to be consistent with geographic projections. UTM 15N NAD83 was the projection used.

Crime and Geographic Boundary Data

Geographic coordinate (X,Y) data were collected and recorded as part of a police report and updated into the city's database as an attribute for the crime. Geocoding was not necessary. Instead, coordinates of each crime were integrated by adding the Excel table to the GIS and performing the "Add XY Data" function. Table 1 shows crime data in raw, spreadsheet form.

Crimes included in this study included damage to property, vehicle theft,

theft, burglary, and robbery. Table 2 shows residential crime occurrences.

Table 1. Crime data attribute table. Date, location, and type of crime are collected as part of crime report. XY data was used to determine dependent variable, crime density.

Case Number	Date of Crime	Latitude	Longitude	Crime Type
2011-00001028	1/1/2001	44.017	92.458	03 Robbery
2011-00002421	1/16/2011	44.049	92.454	03 Robbery
2011-00002629	1/17/2011	44.023	92.479	03 Robbery
2011-00003608	1/23/2011	44.017	92.456	03 Robbery
2011-00005634	2/5/2011	44.035	92.485	03 Robbery
2011-00016404	4/9/2011	44.026	92.472	03 Robbery

Table 2. Analysis of residential crime by type. Theft and vandalism made up 80% of residential crimes researched.

Crime type	# of Crime Type	PCT of Total Res. Crime
Robbery	168	1.70%
Burglary	1420	14.50%
Theft	5910	60%
Motor Theft	292	3%
Pos Stol Prop	74	0.80%
Vandalism	1949	20%

These are common forms of property or residential crimes. Figure 2 shows all reported residential crimes for all of Olmsted County. To provide for spatial integrity and spatially accurate analysis, the newly created feature class was also projected into UTM 15N NAD 1983.

Table 1 also displays key attribute fields for analysis. Attributes for crime data included:

1. *Location of crime* so that crime information could be spatially analyzed and density could be calculated; and
2. *Crime type* so that crimes could be classified and analyzed according to the type of crime committed; and
3. *Date of crime* so that temporal analysis could be included as needed.

Crime data incorporated in this study included residential crimes committed within residential areas. Figure

3 shows the zoning data classified into areas zoned “residential” and areas zoned as anything else, or “nonresidential.”

Crime features were selected by location to select only features which occurred within areas zoned as residential. This narrowed field was then overlaid with the 2010 census block group boundary files. Figure 4 shows an overlay of crime and zoning data. Block group data served as the refined study area for exploration. Olmsted County consists of 111 block groups. The study area for this analysis contained 90, or 80% of the total block groups within the county.

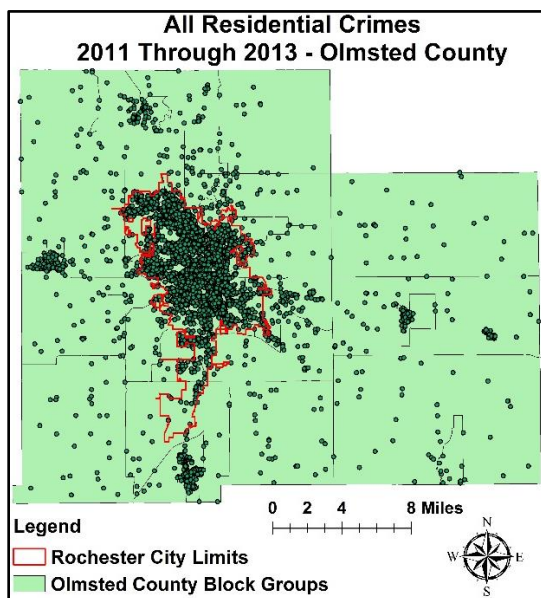


Figure 2. Residential crimes within Olmsted County. XY data for crime locations were imported into a GIS for analysis. The figure shows most residential crimes from 2011-2013 were concentrated within the study area boundary.

According to data provided by the Rochester Police there were 11,457 residential crimes reported in Olmsted County between 2011 and 2013. Of these crimes, 10,351 were committed within the study area census block groups. Eighty-eight percent, or 9,810, of crimes committed were reported in residential areas within the Rochester area study area.

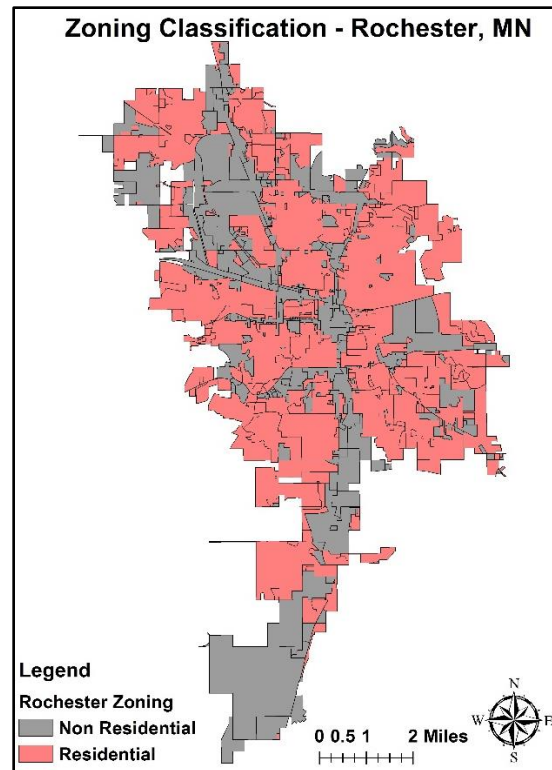


Figure 3. Zoning data as classified for analysis. Residential areas are depicted in red. All other classifications or “non-residential” areas are depicted in grey.

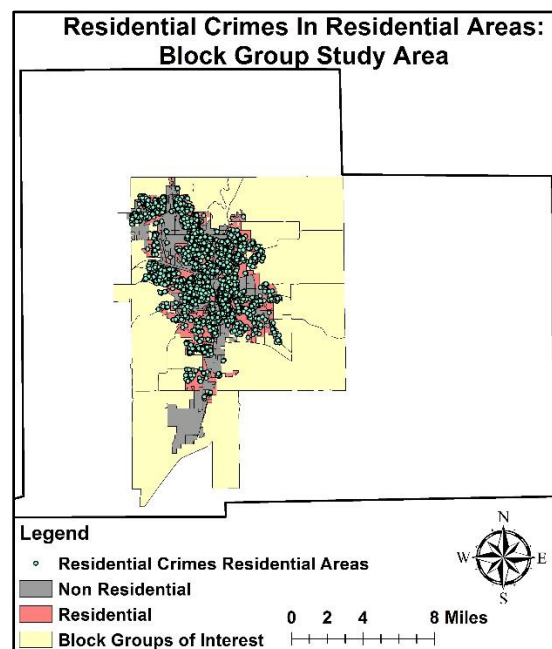


Figure 4. Rochester area block groups including residential crimes in residentially zoned areas. 90 block groups were included for analysis.

Top Offender Data

Addresses of offenders were included in police reports and government documents. Average distance from a crime location to the offender's place of residence can vary, but some studies have shown average distances of 3 miles (Ratcliffe, 2001). Addresses of offenders are public record and served as an *offender location* table attribute. These addresses were an attribute field in the Rochester Police Department's (RPD) database. RPD compiles this data into a daily, "Top Offenders" report. The report is a running, weighted ranking of offenders based of quantity of offenses and severity of crime. An offender who commits several crimes and/or commits a severe offense will rank higher on the list than someone committing a single crime or a crime of less severity. This project uses a list of the top 100 offenders as of January 2013.

Demographic Data

Attributes of demographic data were joined with geographic data creating a larger geographic attribute table including demographic measures. Attributes include:

1. *GEOID* to serve as the primary key in table joins; and
2. *PCT_ASSDEG_UP* to serve as the measure of educational attainment field; and
3. *MED_HH_INC* to serve as the measure of income field; and
4. *PCT_NON_WHT* to serve as the measure of race field; and
5. *PCT_MALE* to serve as the measure of sex field; and
6. *TOTAL_POP* to serve as the measure of population field; and
7. *MED_AGE* to serve as the measure of age within the block group; and
8. *PCT_Zown_OCC* to serve as a measure of home ownership; and

9. *PCT_NILF* to serve as a measure of employment; and
10. *PCT_FEM_FAM_NoHP* to serve as a measure of household type; and
11. *NUM_CRIMES_RES* to serve as a measure of crime density; and
12. *CrimePP_Den_3yrAverage* to serve as a measure of crime density.

New data were normalized or standardized by total population, creating a percentage for each variable. For instance, the raw data might provide the number of persons living below the poverty level per block group. The raw poverty data were then calculated as a percentage of the total population of the block group. This process provides for easier and more accurate comparison among blocks.

This project did require the creation of new data. The different demographic variables examined in this study (i.e. education, income, race, etc.) are a compilation of various census table data attributes. Table 3 shows raw education variables combined for use in this study.

Table 3. Raw education data from census files. Education data was compiled from U.S. Census tables. A summation of all degrees – associates to doctoral – were computed and standardized by population.

B15003 Total Pop.	B15003 Associate's	B15003 Bachelor's	B15003 Master's	B15003 Professional	B15003 Doctorate
667	24	106	56	142	46
648	19	74	87	11	12
515	43	88	34	0	0
1476	179	141	62	0	19
321	41	70	46	18	0

The eight demographic variables analyzed for this study were defined as follows and visualized in spreadsheet form in Table 4:

1. *TOTAL_POP* represents the total population for each block group.

Table 4. Spreadsheet of variables used in analysis. Eight socio-economic variables were compared against crime rates.

GEOID_2	Population	Crime Density	Education	Income	Home Own	Sex	AGE	Unemployment	Single Mom	Race
271090001001	667	0.24	56	32604	13	45	79	80	2.5	12.9
271090001002	648	0.11	31	22946	9	38	50	49	0.0	43.9
271090002001	515	0.08	32	41771	50	55	25	29	11.3	21.9
271090002002	1476	0.02	27	54844	70	77	41	61	0.0	37.5
271090002003	321	0.12	55	47132	74	37	42	40	15.1	30.5
271090002004	629	0.13	39	56250	72	52	29	24	8.1	3.8
271090002005	672	0.12	30	32951	24	41	24	33	24.4	48.2
271090003001	608	0.24	39	38988	40	63	35	44	14.0	19.4
271090003002	876	0.06	39	42115	76	53	31	18	12.5	6.3
271090003003	513	0.08	34	43553	81	45	27	18	17.9	7.4
271090004001	781	0.03	71	48029	50	48	33	33	7.4	16.1

It is a standalone figure in the B15003 table and also serves as a standardizing value for determining several variables included in this study.

2. *PCT_ASSDEG_UP* represents the percentage of the population with higher education. It uses the sum of all degrees, associate through doctorate, and is based on B15003 census tables.
3. *MED_HH_INC* represents median household income. It comes directly from the B19013 census tables.
4. *PCT_NON_WHT* represents the percentage of the total population that is NOT white. It is derived from the standalone B2001 table column “White Alone” over total population.
5. *PCT_MALE* represents the percentage of the total population that is male. It is derived by dividing the B01002 table data by total population within each block group.
6. *MED_AGE* represents the median age within each block group. It comes directly from the standalone B01001 table.
7. *PCT_Zown_OCC* represents the percentage of total population who live in owned homes rather than rented apartments or homes. It is a summation of the “Owned free and clear” and “Owned with a mortgage or loan” columns in census 2010 Summary File 1 H11 tables.

8. *PCT_NILF* represents the percentage of the population that is not in the labor force. It is derived from employment data in the B02325 tables and normalized by total population.
9. *PCT_FEM_FAM_NoHP* represents the percentage of female head of households with no husband. It is obtained from the B11001 tables as a standalone and standardized by total population.

Statistical Analysis

The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and regression. Correlation quantifies the strength of the relationship between pairs of variables whereas regression expresses the relationship in the form of an equation.

To quantify the strength of the relationship, the correlation coefficient, or *r* value, was calculated. The value of *r* always lies between -1 and +1. A value of the correlation coefficient close to +1 indicates a strong positive linear relationship. A value close to -1 indicates a strong negative linear relationship. A value close to 0 indicates no linear relationship.

After performing correlation analysis, the regression statistics were used to form an equation for the “line of best fit.” The regression line is a plot of

the expected value of the dependent variable for all values of the independent variable. An optimum regression line is applied to the data when the prediction of the independent variable based on the given values of dependent variables is as close to the real y values as possible. Successful regression equations should be able to predict areas of future crime based on the variables observed in this study.

Both standard multiple linear regression and stepwise multiple linear regression were used as methods for analysis. Standard multiple regression includes every independent variable in its prediction output whereas stepwise multiple regression creates subsets of significant and not significant independent variables. Stepwise regression eliminates variables deemed not significant and excludes them from the prediction model.

The coefficient of determination (R^2) is the square of the correlation coefficient. Its value may vary from zero to one. The coefficient of determination indicates the amount of variation in the dependent variable that is 'explained' by variation in the independent variables. For example, an r-squared value of .49 means that 49% of the variability in the dependent is explained by the regression equation.

The standard error of the estimate for regression measures the amount of variability in the points around the regression line. It is the standard deviation of the data points as they are distributed around the regression line. The standard error of the estimate can be used to develop confidence intervals around a prediction.

Results

Results included maps of demographic variables, scatterplots of independent versus dependent variables, figures displaying correlation and regression

findings, and maps of standardized residuals from regression.

Variable Maps

One of the advantages of using a GIS for analysis is that tabular attribute data can be mapped and visualized when a spatial component is added. The following maps (Figure 5 – Figure 15) are visual representations of the variables explored. They provide an easy way to obtain a general assessment of the demographic makeup of the study area and create a medium in which general patterns can be determined for potential analysis in the future. Relationships between the dependent variable and the independent variables seem to exist.

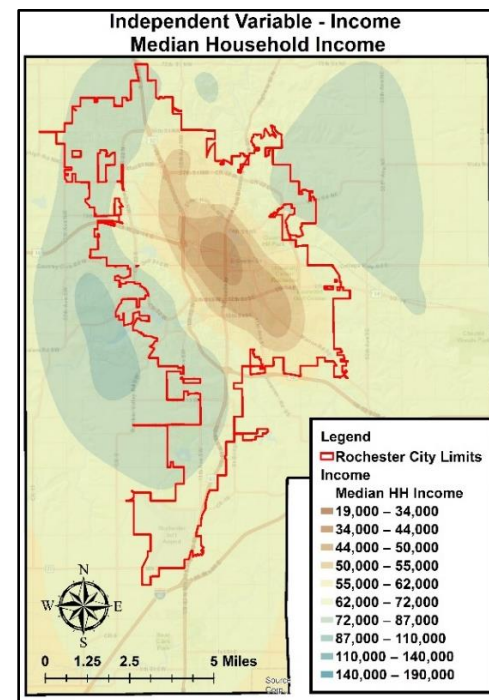


Figure 5. Kriging interpolation of income variable within study area block groups. Clusters of higher median income appear around the edges and outside the city limits.

Variable Charts

Each independent variable was plotted against the dependent variable, crime density.

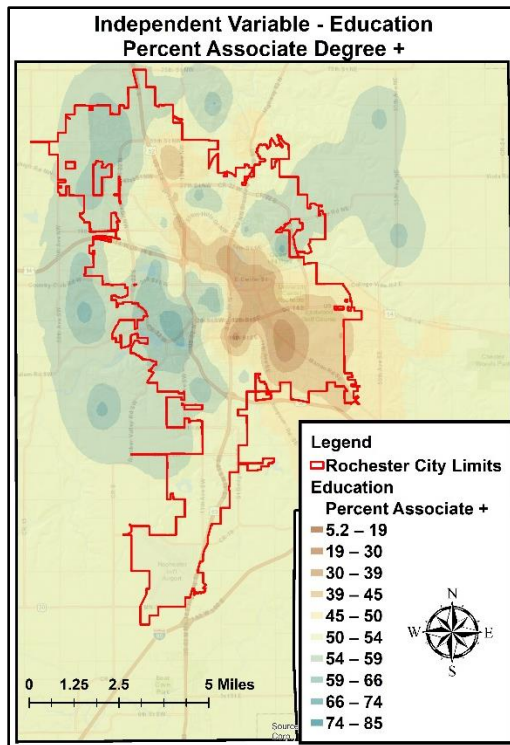


Figure 6. Kriging interpolation of independent education variable within study area block groups. Clusters of higher median income appear around the edges and outside the city limits.

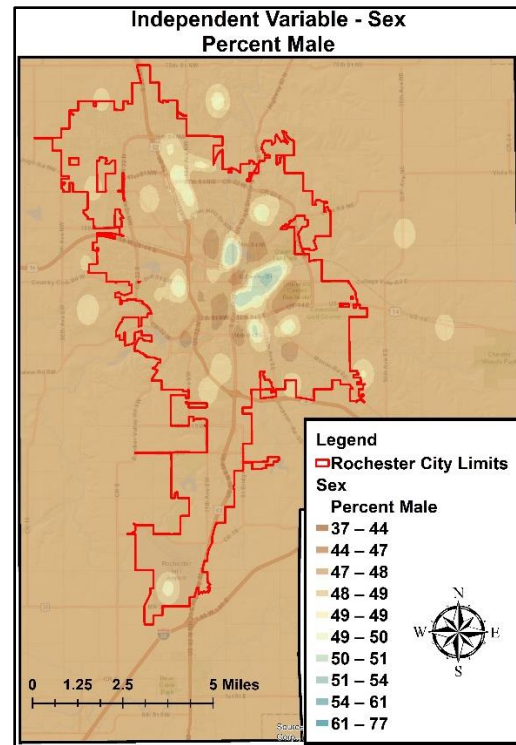


Figure 8. Kriging interpolation of sex variable within study area block groups. No pattern seems to exist.

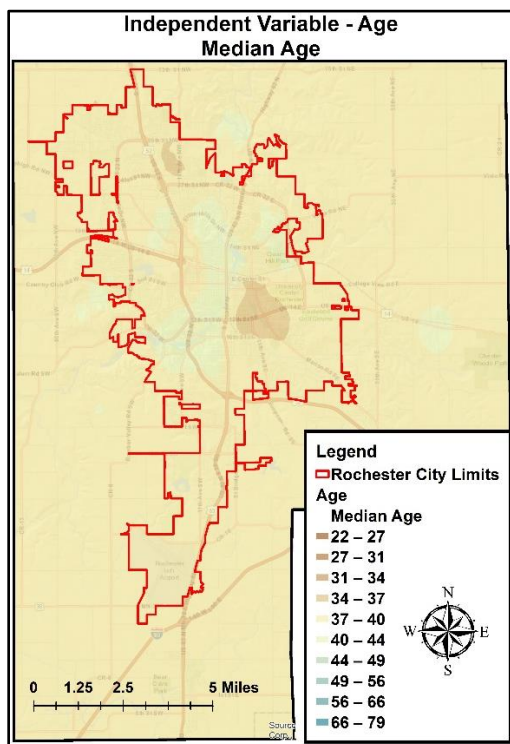


Figure 7. Kriging interpolation of age variable within study area block groups. There does not seem to be a pattern present.

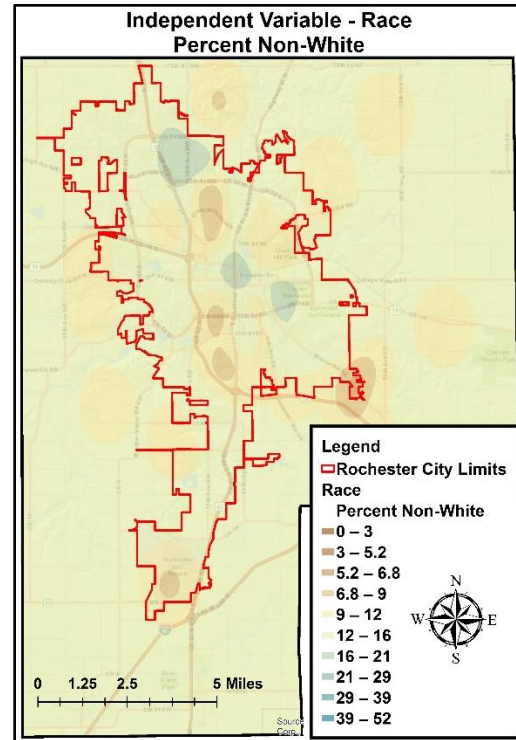


Figure 9. Kriging interpolation of race variable within study area block groups. Clusters of high and low areas of non-white population seem to exist at random.

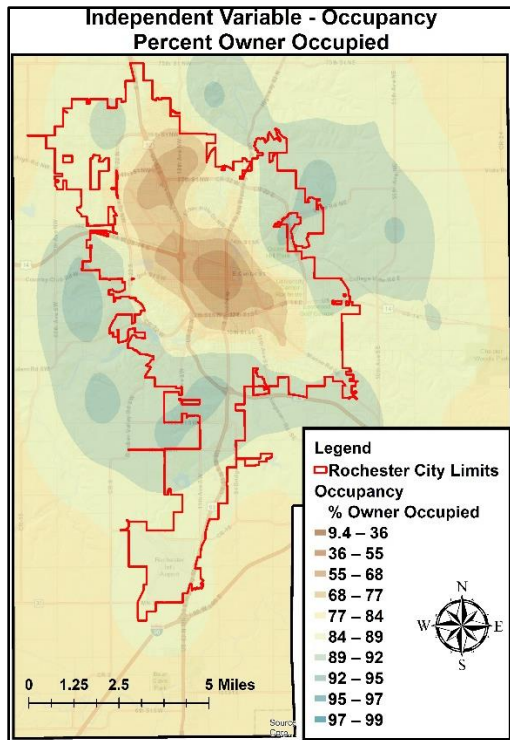


Figure 10. Kriging interpolation of occupancy variable within study area block groups. Clusters of higher home ownership appear to the southwest and to the northeast.

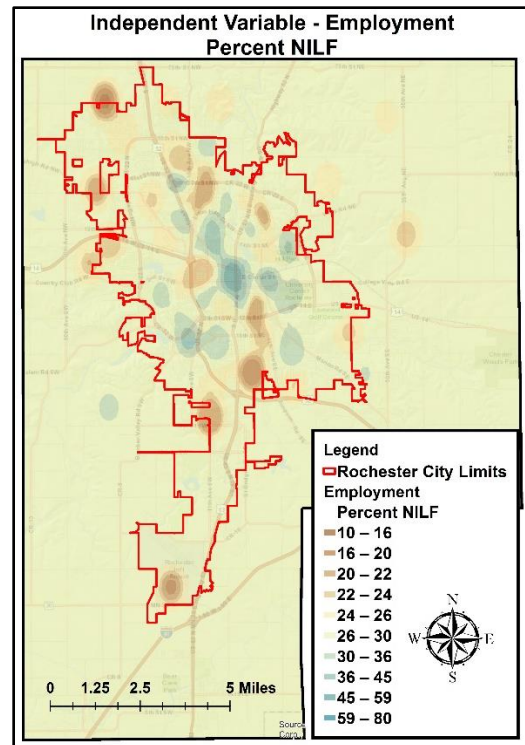


Figure 12. Kriging interpolation of independent employment variable within study area block groups. Clusters of high and low concentrations appear near the city center.

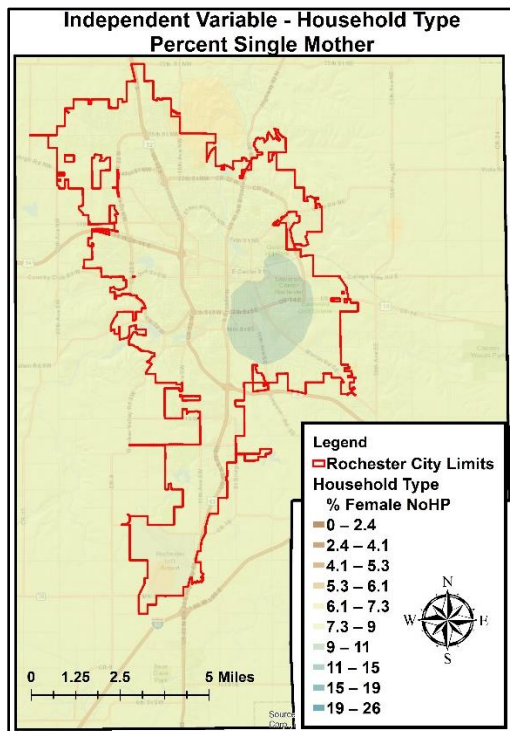


Figure 11. Kriging interpolation of independent household type variable within study area block groups. There is a concentration of single mother homes in the southeastern area of the city.

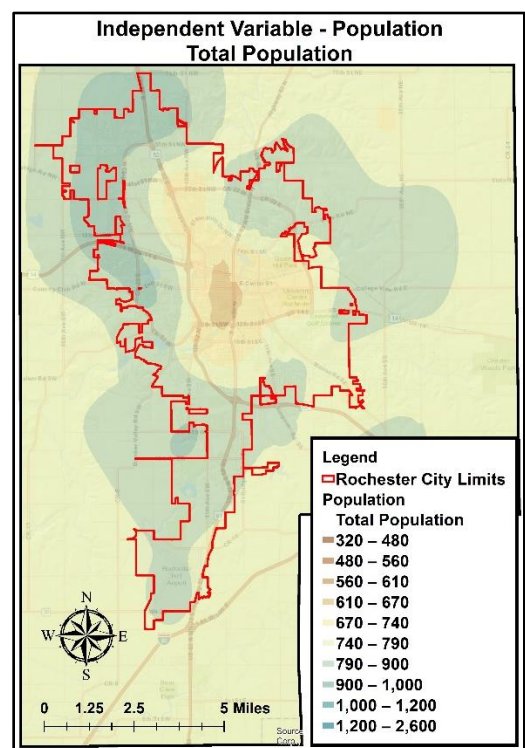


Figure 13. Kriging interpolation of independent population variable within study area block groups. A cluster of low concentrations appear near the city center.

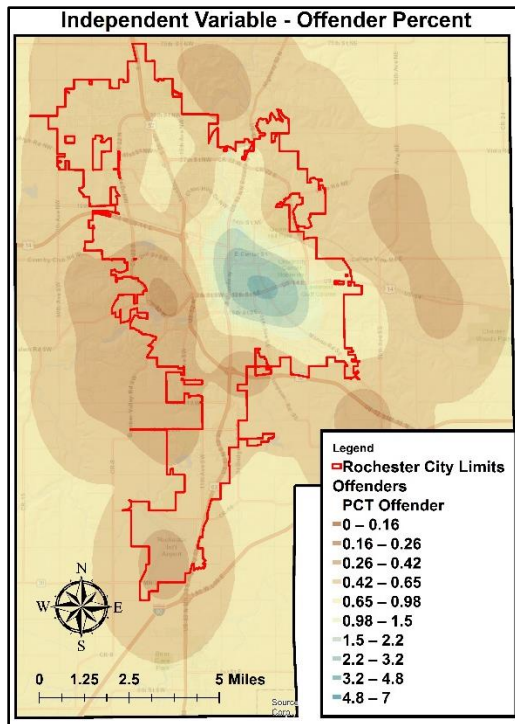


Figure 14. Kriging interpolation of independent offender percentage variable within study area block groups. A cluster of high concentrations appears near the city center.

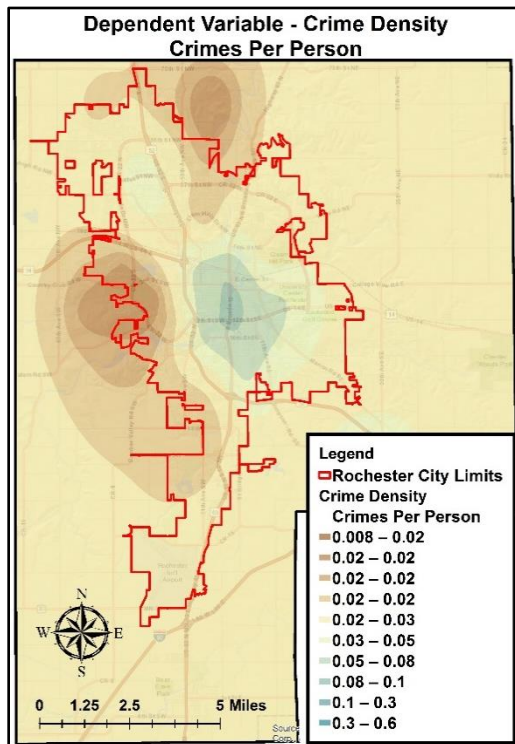


Figure 15. Kriging interpolation of dependent variable crime within study area block groups. Clusters of low crime rates appear to the north and to the west of the city. A concentration of high crime rates appears near the city center.

Scatterplots (Figures 16 – 23) were the result. When the line of best fit is placed, its slope and direction show the strength of these relationships.

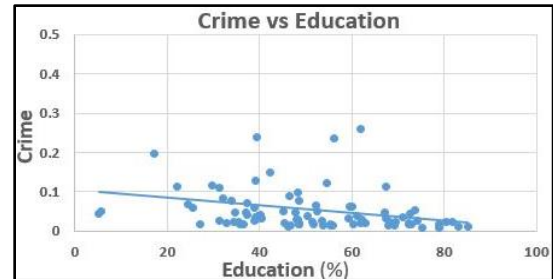


Figure 16. Crime density vs Percent Education. With a p value of -.24, the correlation is significant at the .05 level.

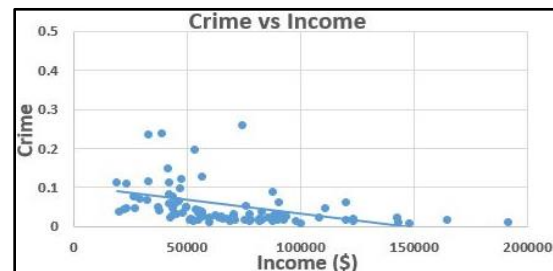


Figure 17. Crime density vs Median Household Income. With a p value of -.33, the correlation is significant at the .01 level.

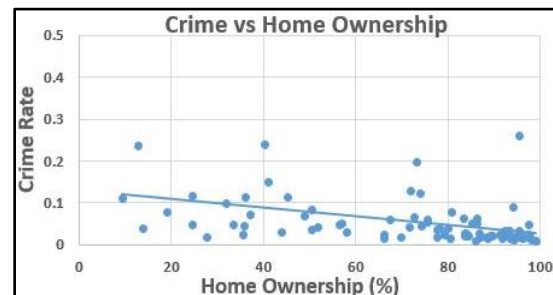


Figure 18. Crime density vs Home Ownership. With a p value of -.35, the correlation is significant at the .01 level.

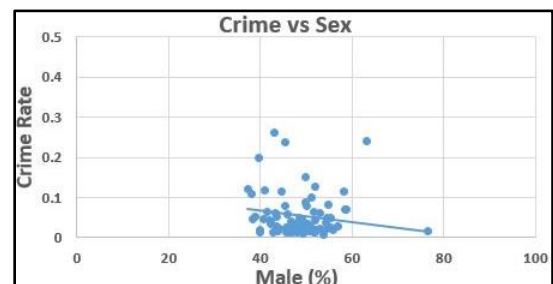


Figure 19. Crime density vs Percentage Male. There is a weak, indirect relationship between the variables. Correlation is not significant.

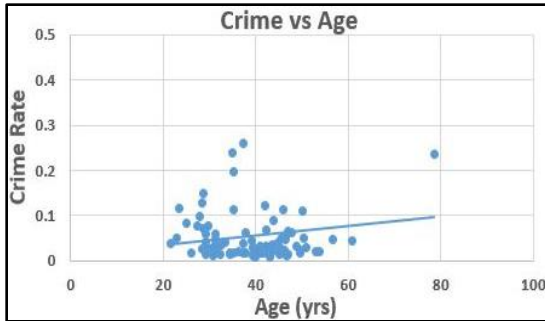


Figure 20. Crime density vs Age. There is a weak, direct relationship between the variables. Correlation is not significant.

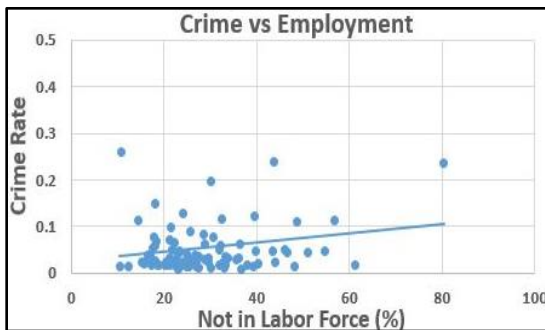


Figure 21. Crime density vs Employment. There is a weak, direct relationship between the variables. Correlation is not significant.

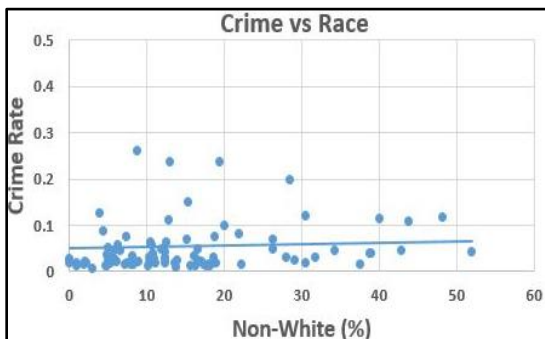


Figure 22. Crime density vs Percent Non-White. There is a weak, direct relationship between the variables. Correlation is not significant.

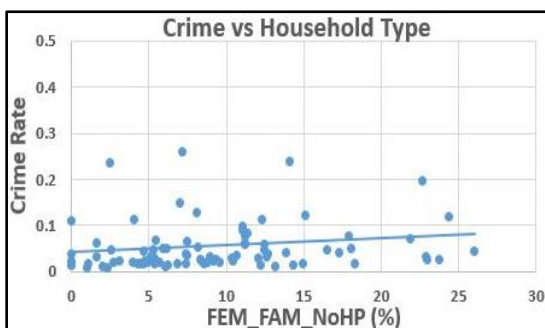


Figure 23. Crime density vs Household Type. There is a weak, direct relationship between the variables. Correlation is not significant.

Correlation and Regression

Table 5 shows results of correlation statistics. Of the eight independent variables, only three were found to be significant at the $p < .05$ level. Measures of income, education, and home ownership all had a significant correlation to crime density. Significantly correlated variables are highlighted in green. Yellow and red variables were not significant and were excluded from the prediction model.

When comparing crime density with offender location, an R value of .35 and an R^2 value of .12. When analyzing demographic variables against crime regression produces an R value of .47 and an R^2 value of .22. Therefore, demographic variables were a better predictor of crime within the city than were the locations of top criminal offenders.

Residual Maps

Residual maps are an easy way to visualize how well a regression equation predicts crime. Areas of over and under prediction can be spotted. Figures 24 and 25 show residual maps for both offender location and for demographic variables. The model for socio-economic variables predicts 82 out of 90 block groups within a standard error of -1 and 1. The model for offender location variable predicts 85 out of 90 block groups within 1 a standard error of -1 and 1.

Discussion

Correlation and regression results were much lower than expected. A number of decisions might have led to these results and making minor changes could possibly produce a more accurate depiction of the relationship between crime and other variables.

For instance, defining crime was a

Table 5. Correlation Matrix for Socio-Economic Variables vs Crime Density. While most independent variables are not strongly correlated with crime density, race, single mother families, percentage in labor force, age, and sex have the weakest correlations. All of these variables have a p value of less than .2.

	Education	Sex	Age	Employ	Income	Ownership	Occupancy	Race	# Offenders	# Crimes	Crime Density
Education	1.00										
Sex	-0.16	1.00									
Age	0.30	-0.21	1.00								
Employment	-0.13	0.03	0.58	1.00							
Income	0.73	0.04	0.17	-0.22	1.00						
Ownership	0.44	-0.03	0.12	-0.31	0.70	1.00					
Occupancy	-0.46	-0.17	-0.39	-0.07	-0.43	-0.21	1.00				
Race	-0.41	0.09	-0.29	0.23	-0.41	-0.61	0.28	1.00			
# Offenders	-0.52	0.28	-0.38	-0.01	-0.47	-0.44	0.45	0.41	1.00		
# Crimes	-0.17	-0.04	0.09	0.13	-0.27	-0.36	0.11	0.08	0.32	1.00	
Crime Density	-0.24	-0.11	0.13	0.16	-0.32	-0.35	0.13	0.05	0.22	0.86	1.00

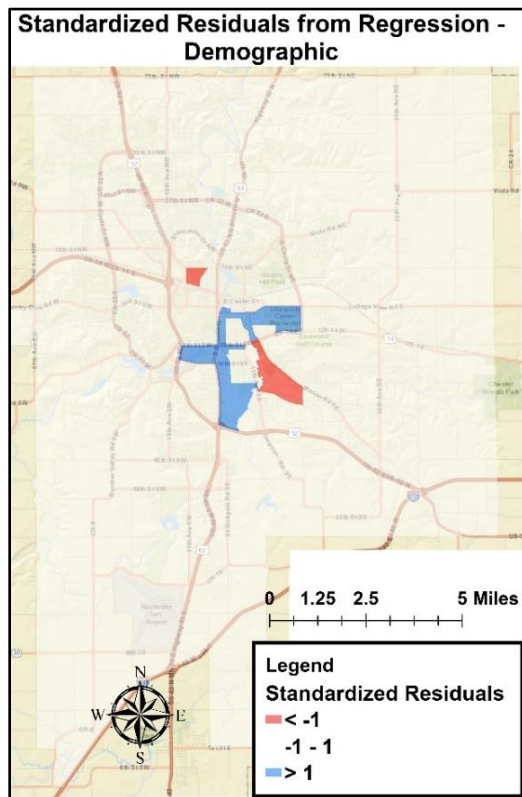


Figure 24. Standardized Residuals from Demographic Variable Regression Model. A majority of block groups are predicted with a standard error between 1 and -1.

major decision in this work. This research defined crime as “residential crimes in residentially zoned areas.” Broadening the definition of crime to include *all crime* or *all residential crimes* could reveal more patterns and produce a more reliable prediction model.

Second, independent variables

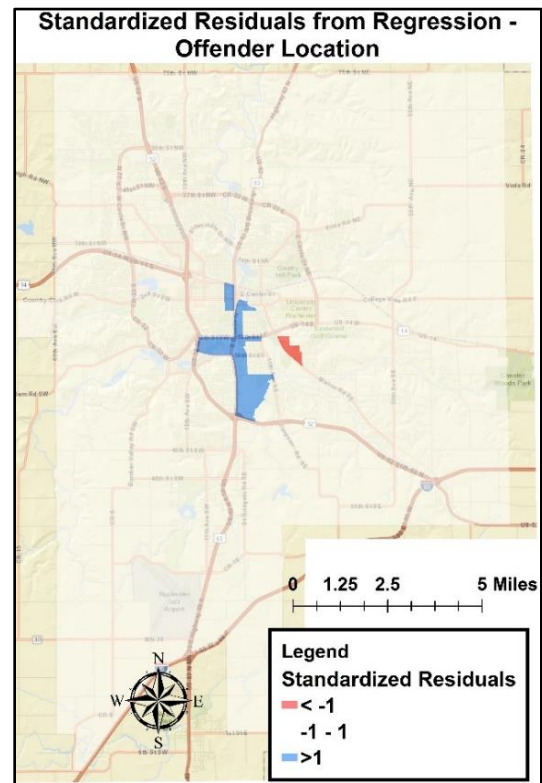


Figure 25. Standardized Residuals from Offender Location Regression Model. A majority of block groups are predicted with a standard error between 1 and -1.

chosen also had direct impact on the output results. This research created variables based on U.S. Census Bureau summary files. Most variables were a sum of multiple, more specific indicators. This research used the example from Sampson and Groves (1989) as a base for selecting demographic predictor variables.

It is important to note variables also become limited by aggregation level. MAUP, or modifiable areal unit problem should always be considered. MAUP may serve as a source of statistical bias. Different results could be found based on what geographic boundary is used (Openshaw, 1984). This research used block groups as a geographic boundary. Using a larger census tract file could produce more desired results but would result in a significantly smaller sample size.

All of the data used for this analysis was considered to be static. Introducing a temporal element for analysis might not only be interesting, but may also produce more dynamic results that could be degraded into smaller units of time and thus more desirable results.

An important element to consider regarding offender location is that the formula was created by the Rochester Police Department. Presently, offender locations do not exclude persons who have been imprisoned as it relates to this study. Offender scores are calculated and decrease over time with inactivity of the offender. The static nature of this research does not directly account for a criminal being removed from the streets. Such a situation might influence any prediction model not taking this fact into consideration.

Conclusion

GIS created maps can be useful visuals when searching for crime and demographic patterns. However, statistical analysis is required to explore such relationships. Correlation and regression results for the data used in this research suggest that while socio-economic make up of block groups is a better predictor of crime than is top offender location, both methods produce relatively weak R^2 values of .22 and .12 respectively. These R^2

values suggest the relationship of these variables to crime is weak and therefore not a sufficient predictor of crime on their own. Model predictability might become more accurate as examination and inclusion of additional significantly correlated independent variables are included.

Acknowledgements

The completion of this project could not have been possible without the assistance of John Ebert, Professor and Chair of the Department of Resource Analysis at Saint Mary's University of Minnesota and Joe Koncur of the Rochester Police Department. Their guidance and contributions were vital in the project's success and were sincerely appreciated. In addition I would like to thank my friends and family for their unconditional love and support.

References

- Brantingham, P. J., and Brantingham, P. L. 1991. *Environmental Criminology*. Waveland Press. Prospect Heights, IL.
- Davin, H., and Lin, L. 2009. Cops and robbers in Cincinnati: a spatial modeling approach for examining the effects of aggressive policing. *Annals of GIS*, 15(1), 61-71.
- Minnesota State Demographic Center. 2012. *Minnesota Population Projections 2015 to 2040*.
- Openshaw, S. 1984. The modifiable areal unit problem. CATMOG 38. GeoBooks, Norwich, England.
- Ratcliffe, J. H. 2001. Policing urban burglary. *Trends and Issues in Crime and Criminal Justice*, 213.
- Ratcliffe, J. H. 2002. Damned if you don't, damned if you do: Crime mapping and its implications in the real world. *Policing and Society*, 12(3), 211-225.

- Rochester Area Economic Development, Inc. 2012. Rochester, Minnesota MSA Data Book. Historical and Projected Growth. Retrieved August 1, 2015 from <http://www.pineislandeda.org/>. About_Region_Rochester_MSA_Data_Book.pdf.
- Sampson, R., and Groves, W. B. 1989. Community Structure and Crime: Testing Social Disorganization Theory. *American Journal of Sociology*, 94, 774-802.
- Weisburd, D., and Green, L., 1995. Policing drug hot spots: the Jersey City drug market analysis experiment. *Justice Quarterly*, 12, 711–735.
- Weisburd, D., and McEwen, T. 1997. Geographic Information Systems and Crime Analysis in Baltimore County, Maryland. *Crime Mapping and Crime Prevention*, 157-190.