Technical Memo

Date: 4 February 2019

To: Spring 2019 CE395A Data Science for Environmental Site Characterization

From: Andrea J. Elhajj, Graduate Student

Subject: 2018 Staten Island Felonies Analysis

Executive Summary

In this analysis, we examined a selection of felonies in the New York City borough of Staten Island and occurring in 2018. These felonies included: murder, kidnapping, grand larceny, dangerous weapons, and criminal mischief. The first set of objectives were to determine whether the race, age, or gender of the suspect was correlated to the type of crime committed. Chi-Squared Tests of Independence were conducted to address this, and a correlation at the 95% confidence level were found for all three variables. The second set of objectives were to create a semivariogram and determine if the data exhibited spatial structure. A crime severity index on a scale of 1 through 5 was assigned to each of the crime categories in the respective order as listed above. The crimes and their associated indices were mapped to a shapefile of Staten Island, and a semivariogram was produced. It was determined that spatial structure did indeed exist within the data, and a Kriged interpolation of the surface was produced and mapped. However, the spatial scale relative to the range of the semivariogram meant that this model was best applied to smaller areas of land, such as blocks, or neighborhoods.

Introduction

The subject of this investigation is a dataset from the New York Police Department (NYPD) containing all valid felony, misdemeanor, and violation crimes reported to the NYPD in 2018 [1]. The anonymized crime data are released annually on the NYC Open Data site as a spreadsheet. The complex "system" included in this data set is the city of New York, comprised of five boroughs. These the are: Brooklyn, Bronx, Manhattan, Queens, and Staten Island, sitting where the Hudson River meets the Atlantic Ocean. The "media" sampled are all CCRB jurisdiction complaints. For the purposes of this data analysis, the "system" will be reduced to focus solely on the borough of Staten Island. This is due to the sheer quantity of data. The collection of this "media" occurs in accordance with the New York State Penal Law. The NYPD records reported crimes and offenses in their database and makes them available to the public. A description of the field names and descriptions can be found in the appendix (Table 1).

The objectives of this analysis are to answer the following questions about selected felony crimes occurring in the borough of Staten Island:

Is there a correlation between the suspect's age, race, or gender to the type of felony that they commit?

If a crime severity index is assigned to selected felonies, does the data exhibit spatial structure?

Methods

The data was first trimmed to contain only the following felony crimes for 2018 in the Borough of Staten Island: murder, kidnapping, dangerous weapons, grand larceny, criminal mischief. The data was then cleaned to remove any crimes with coordinates referencing non-land locations (i.e. in the Hudson River and New York Bay). These felonies were then respectively assigned the following severity indices: 1, 2, 3, 4, and 5.

Next, a shapefile containing all five boroughs of New York City was brought into ArcGIS, and the polygon containing Staten Island was extracted and exported as a new shapefile. This was then brought into R, and the selected felony crimes were plotted on the polygon, color-coded by severity (Figure 1).

To determine whether there is an association between categorical variables, a nonparametric test known as the Chi-Square Test of Independence is commonly used [2]. This test was performed at the 95% confidence level to determine whether the categorical variables of crime type and suspect age range are independent or related, and the contingency table utilized in this test was visualized in the form of a mosaic plot (Figure 2). This test and visualization method were then repeated for the variables of race and gender (Figure 3 and Figure 4, respectively).

A semivariogram was created using the geoR package, which provides a wide variety of functions for geostatistical data analysis [3]. The empirical variogram was binned using the equal-points binning method, and a model was fit to the resulting semivariogram. This model was used to create a Kriged map over the entire grid of Staten Island (Figure 6). The origin (bottom left corner) of the grid represents a longitude of 913,845 ft and a latitude of 121,174ft in New York State Plane Coordinate System, Long Island Zone, NAD 83. A visual was also created to show the variance in this interpolated surface (Figure 7). This process was then repeated on a smaller subset of plot of land to produce an interpolated surface and illustrate the variance (Figure 8 and Figure 9, respectively).

Results/Conclusion

The result of the Chi-Square Test of Independence on suspect age and committed crime type indicates that suspect age group is not independent of type of crime committed ($\chi 2 = 40.79$, $p = 0.004$). For instance, the mosaic plot of suspect age group and felony type shows that murders on Staten Island in 2018 were committed predominantly by the 25-44 year-old age group and the 45-64 year-old age group (Figure 2). As expected, the 65+ year-old age group produced the least number of suspects across all crime types (Figure 2). The Chi-Square Test of Independence also showed evidence that the suspect's race and type of committed are not independent ($\chi 2 = 36.482$, $p = 0.013$). Lastly, we can see that suspect gender is not independent of the type of crime committed ($\chi 2 = 31.7$, $p = 6.81e\text{-}06$). Males are more likely to commit all 5 types of the crimes analyzed (Figure 4).

The fit of the model to the semivariogram shows that there is indeed spatial structure in the data (Figure 5). A cubic model was fit to the resulting semivariogram, with a nugget of 0.23 km, a range of 0.15 km, and a sill of 0.59. This shows that the spatial structure exists for a distance of up to 0.15km. This makes it difficult to visualize over the entire Earth system of Staten Island (Figure 6). However, the effects of the spatial structure show up well in the extracted 1km x 1km subset plot of land (Figure 8). The variance of the interpolated surface shows regions of high variance that indicate the areas where crimes are least likely to occur (Figure 9).

The crime indices assigned to these felonies carry an implicit level of ambiguity and uncertainty. Additionally, only select felonies from the entire batch of felonies, misdemeanors, and violations were

included in this analysis; therefore, the analysis may not reflect the true areas in which severe crimes are most likely to occur.

This analysis leaves many interesting questions to explore. The non-independent relationships between age group, race, and gender with the type pf crime committed should be investigated further. Normalizing the data to the population may result in intriguing societal justice insights. Furthermore, with more research into the penal code, all crimes could be Kriged according to a more standard index rating system. This would create a better picture of high and low (severe) crime areas. Additional census block data containing economic or sociological demographic data could be mapped to this surface to discover what attributes of specific crimes are possibly correlated to lower crime severities.

References

[1] "NYPD Complaint Data Current (Year To Date)." (2019). NYC Open Data, NYC Open Data,

<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data> (May 3, 2019).

[2] "Chi-Square Test of Independence." (n.d.). LibGuides, Kent State University,

<https://libguides.library.kent.edu/spss/chisquare> (May 3, 2019).

[3] Ribeiro, P. J., and Diggle, P. J. (2018). "Package 'geoR' Reference Manual." CRAN, The

Comprehensive R Archive Network,

<https://cran.r-project.org/web/packages/geoR/geoR.pdf> (May 3, 2019).

Attachments

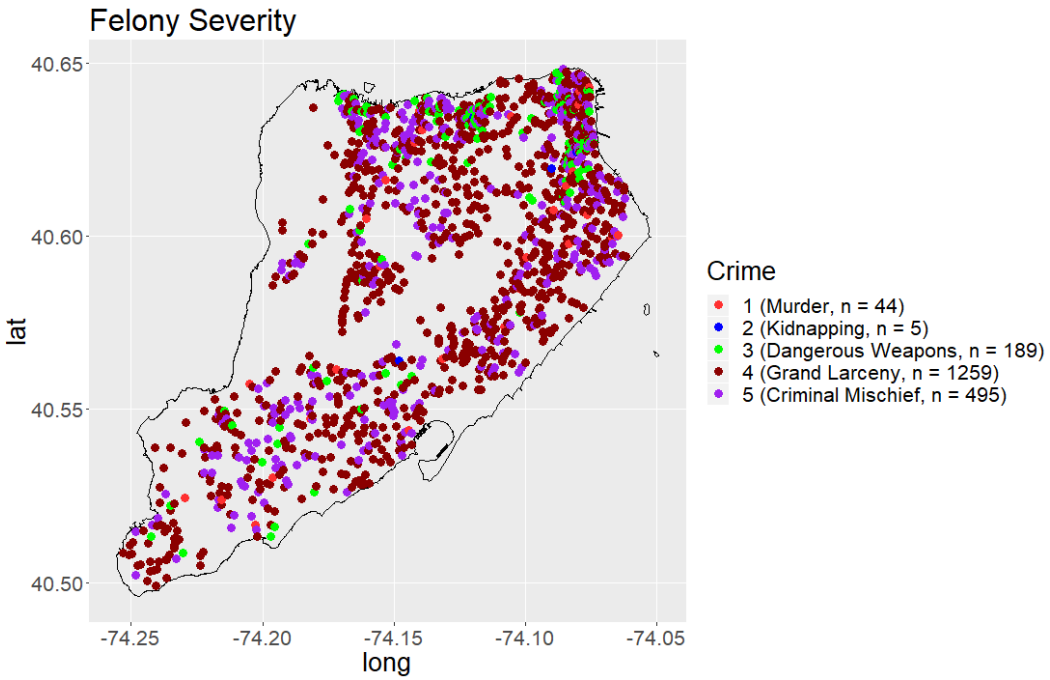

Figure 1: Locations of 2018 Felony Crimes Categorized by Severity Index

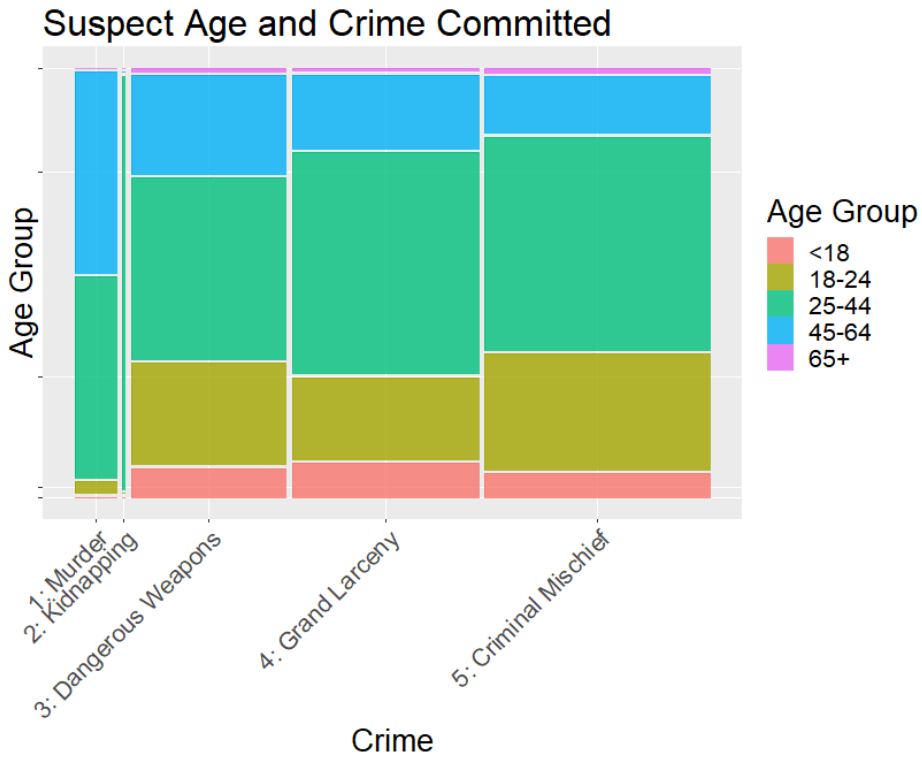

Figure 2: Mosaic Plot of Suspect Age Group and Felony Type

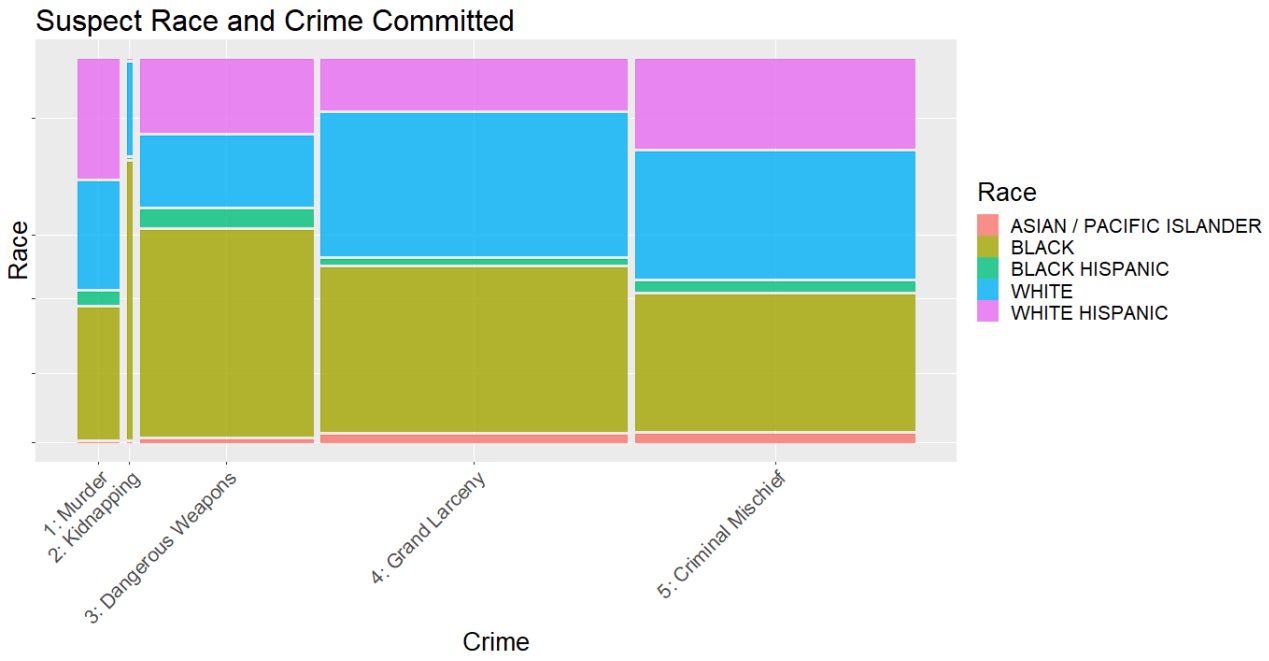

Figure 3: Mosaic Plot of Suspect Race and Felony Type

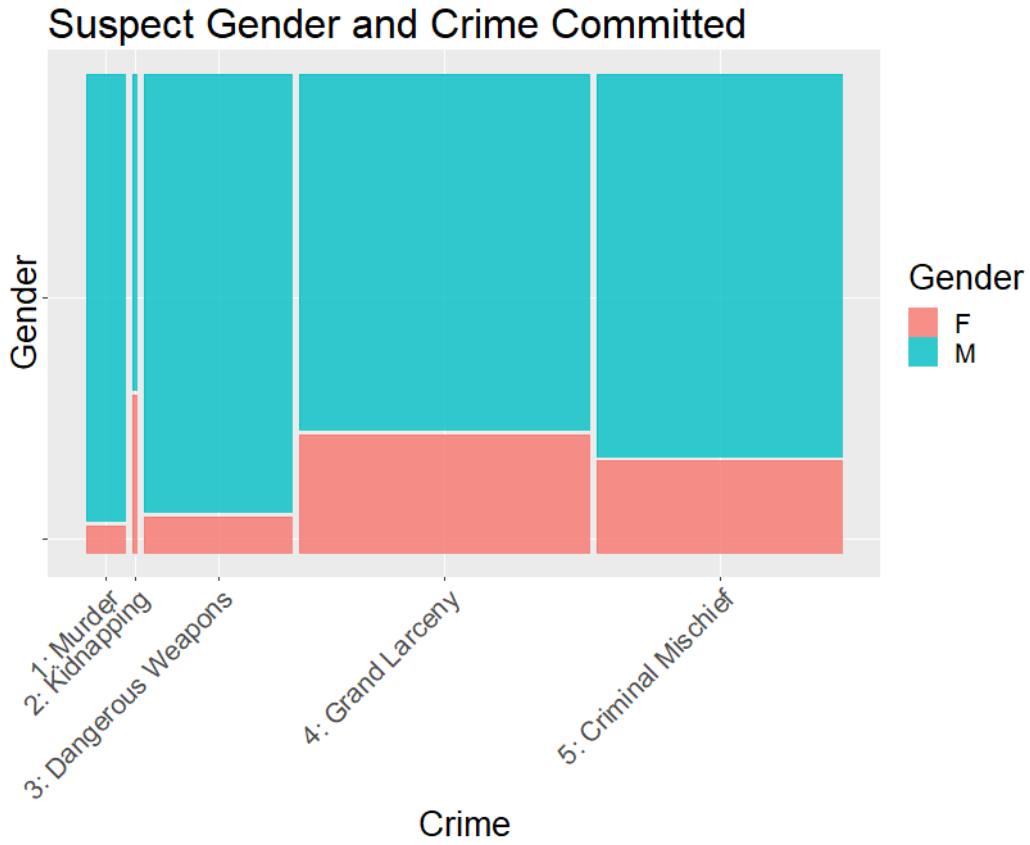

Figure 4: Mosaic Plot of Suspect Gender and Felony Type

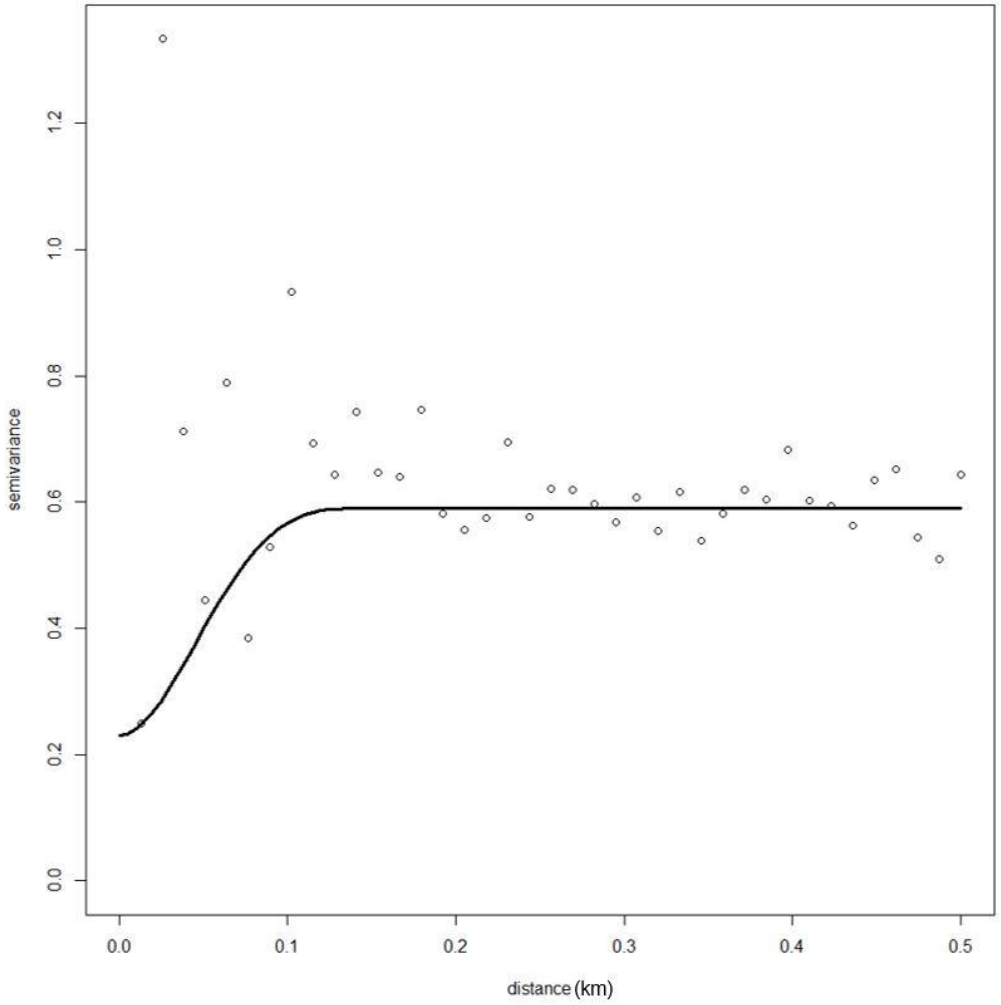

Figure 5: Semivariogram of Crime Severity Index

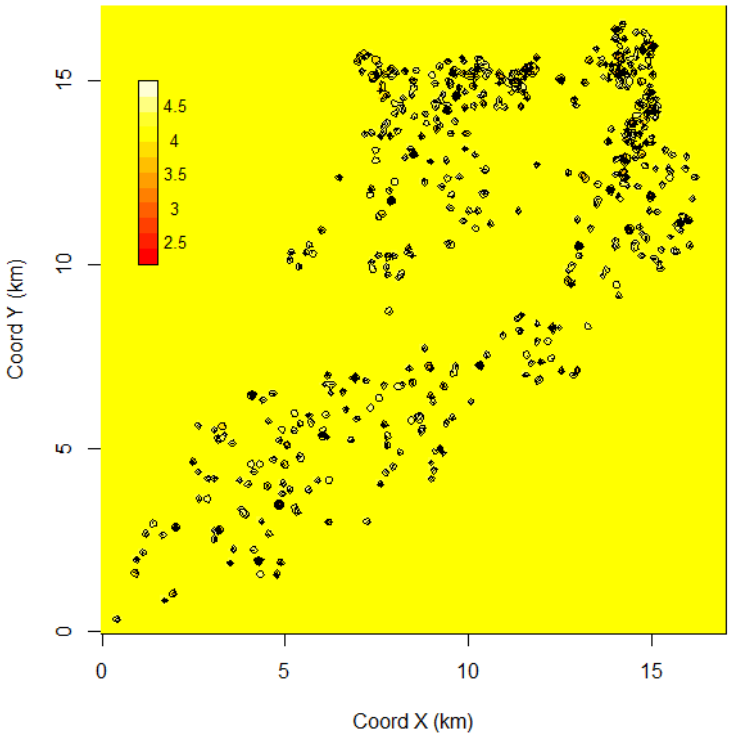

Figure 6: Kriged Map of Crime Severity for Staten Island

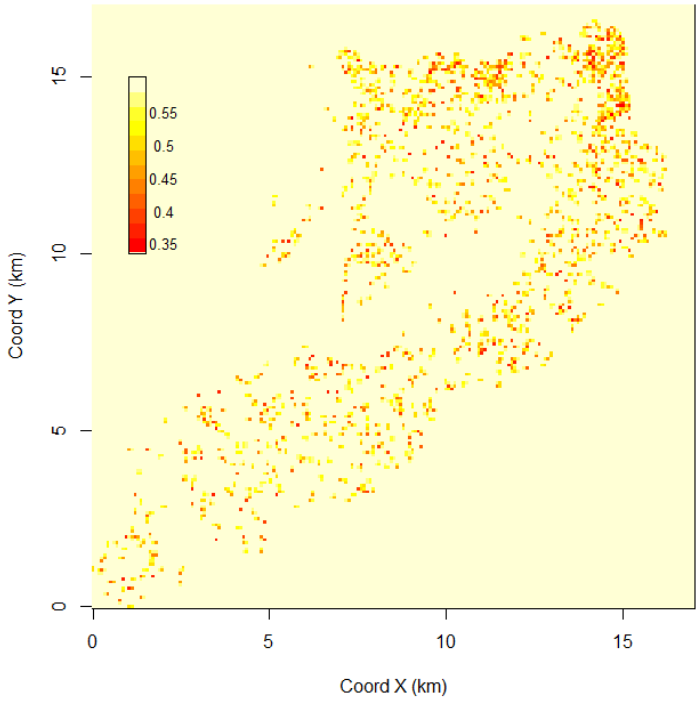

Figure 7: Variance of Crime Severity for Staten Island

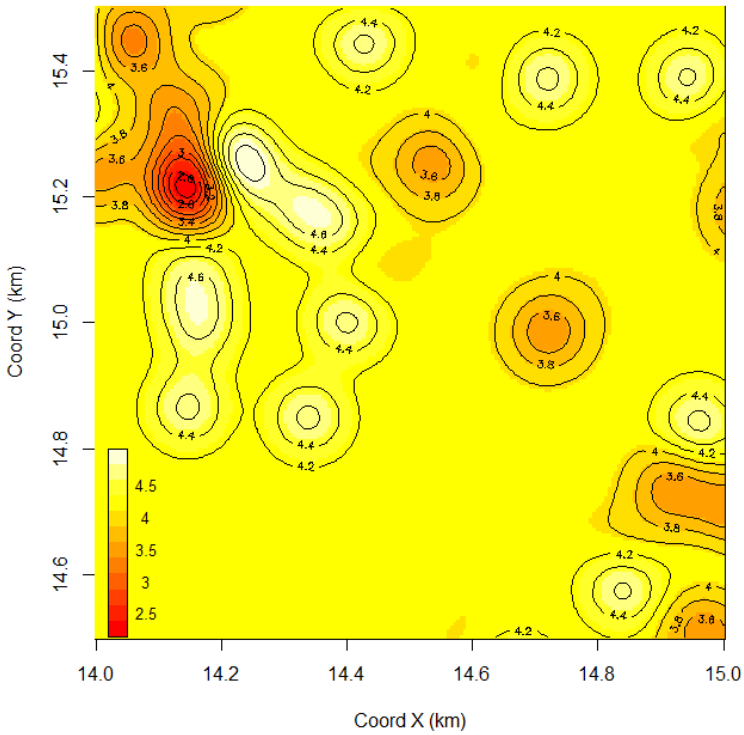

Figure 8: Kriged Map of Crime Severity for 1km x 1km Subset Land Plot

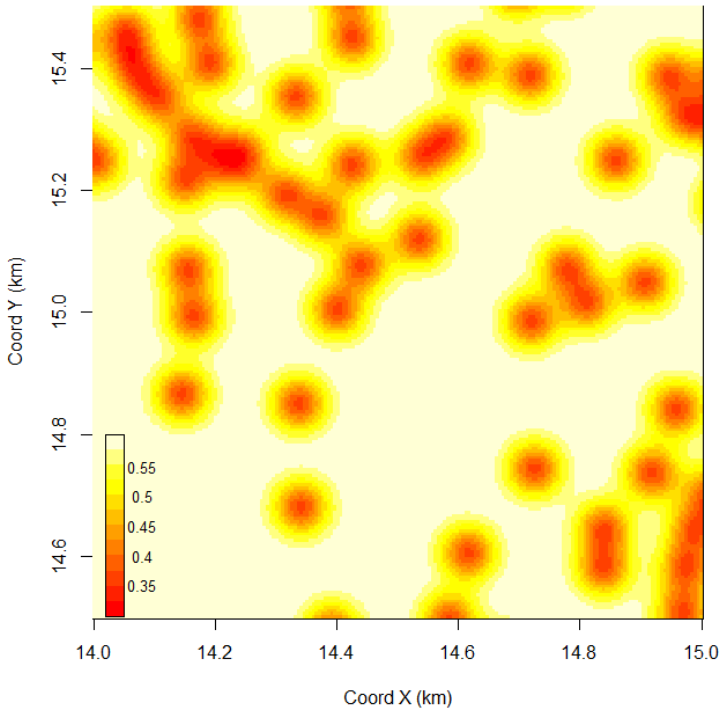

Figure 9: Variance of Crime Severity for Subset Land Plot

Appendix

Table 1: Field Names and Descriptions of Data

Field Name	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
ADDR_PCT_CD	The precinct in which the incident occurred
BORO	The name of the borough in which the incident occurred
CMPLNT_FR_DT	Exact date of occurrence for the reported event
CMPLNT_FR_TM	Exact time of occurrence for the reported event
CMPLNT_TO_DT	Ending date of occurrence for the reported event
CMPLNT_TO_TM	Ending time of occurrence for the reported event
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
HOUSING_PSA	Development Level Code
JURISDICTION_CODE	Jurisdiction responsible for incident. Either internal, like Police (0), Transit (1), and Housing (2); or external (3), like Correction, Port Authority, etc.
JURIS_DESC	Description of the jurisdiction code
KY_CD	Three-digit offense classification code
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
OFNS_DESC	Description of offense corresponding with key code
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable
PATROL_BORO	The name of the patrol borough in which the incident occurred
PD_CD	Three-digit internal classification code
PD_DESC	Description of internal classification corresponding with PD code
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
RPT_DT	Date event was reported to police
STATION_NAME	Transit station name
SUSP_AGE_GROUP	Suspect's Age Group
SUSP_RACE	Suspect's Race Description
SUSP_SEX	Suspect's Sex Description
TRANSIT_DISTRICT	Transit district in which the offense occurred.
VIC_AGE_GROUP	Victim's Age Group
VIC_RACE	Victim's Race Description
VIC_SEX	Victim's Sex Description
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)