Color-Coded Deviance: Spatial Point Pattern Analysis of NYC 911 Complaint Log Case Study: St. Patrick's Day 2015

STAT 787: Spatial Statistics

INTRODUCTION

This project seeks to explore a subset of crime data for the borough of Manhattan during St. Patrick's day in 2015. The use of statistics in criminological research started back in the 1980's as a supplement to law enforcement agencies to better understand the behavior of macro trends in crimes committed. But back then technology in data collection and computing were still in their infancy and formal statistical methodologies were not as rigorously adhere to. Block (1995) was one of the applied researchers who created a visualization tool for the Illinois State Police Department. Later on, other academics like Hermann (2013) recognized that formal spatial statistical methodologies were needed to analyze the vast publicly available crime data released by city agencies, which was done for the Bronx. We propose this project as an extension of Hermann's work but for Manhattan isolated with in a narrow holiday week known for it's particularly debauchery.

DATA

The data used for this project is entitled "NYPD Complaint Data Historic", which is free to download for public use from the NYC Open Data Initiative. It contains observations detailing the 911 calls and their associated investigation by the NYPD. Of most interest here are the longitude and latitude of where these incidences took place, the time and date of when these phone calls were made, and the crime classification of each offense, of which possible are Felony, Misdemeanor, and Violation respectively in terms of severity.

We are primarily interested in the exploration of this data in the form of a spatial point pattern. Preliminary data cleaning and restructuring require the conversion of the raw lat/long into a grid coordinate in the X-Y plane. The GIS standard used for this conversion is "epsg: 2831", specifically designed for the Eastern New York State geography including NYC and Long Island. By law, it is initially standardized in feet and we follow this convention for the analysis. The bulk of the analysis is done with the R package "spatstat".

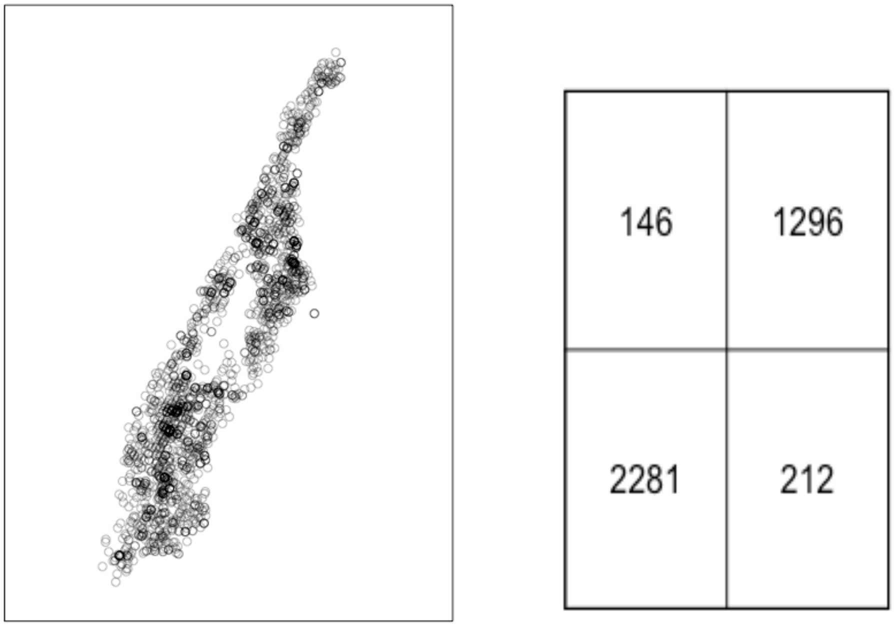
The data consists of 1984 observations with the associated variables described above. The time frame was taken over the week of St. Patrick's Day from 3/16/15 to 3/22/15 starting from midnight. The data provided by Open Data NYC is for all boroughs but for this project only those from Manhattan were used.

METHODOLOGY

The main philosophy behind the analysis is that of a non-parametric approach. As this project is mostly exploratory, this is a natural way of starting with minimum assumption of data's probability distribution. One method to start off with is to detect for the presence of Complete Spatial Randomness (CSR), a central concept in spatial analysis that is based on uniform distribution of points in the spatial structure. This is accomplished by a chi-square test based on counting the number of points in a divided window, a technique pioneered by Ripley (1981). The analysis then continues into calculating the average intensity function for each location of the point pattern. See Bivand (2013) for an overview. Relatedly, the main tools of non-parametric kernel smoothing are employed to determine the best bandwidth. Three bandwidth models are proposed. Once the presence of CSR is rejected we continue in examining the second order property of the spatial point process. This is accomplished by the use of the inhomogeneous K-function and the inhomogeneous G-function. Baddeley (2016) provide an overview of these functions. The analysis also includes an examination of multi-type spatial point patterns, which incorporates factors of crime offense categories.

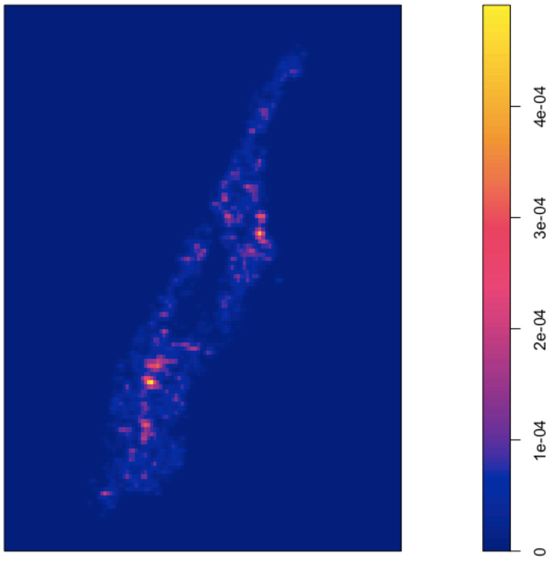
RESULTS

The first step in our exploration is to test for the presence of CSR. Its interpretation suggests that the arrangement of the points follow that of an uniform distribution. Such a result is rarely interesting and it is the goal of the exploration to provide evidence against this phenomenon. Hence, a chi-square test is performed based on count data. The converted X-Y coordinates are plotted on a window, which is then divided into four equal square subpanels. Care is taken to make sure that in each quadrat that there is at least 5 points. Here the null hypothesis is H_0: CSR is present. In practice, however, the almost two thousand points created an almost perfect silhouette of Manhattan; CSR is obviously not possible. But for the sake of formality the quadrat count and chi-square test are provided. The p-value is less than 2.2e-16 and therefore, null hypothesis of CSR is rejected.



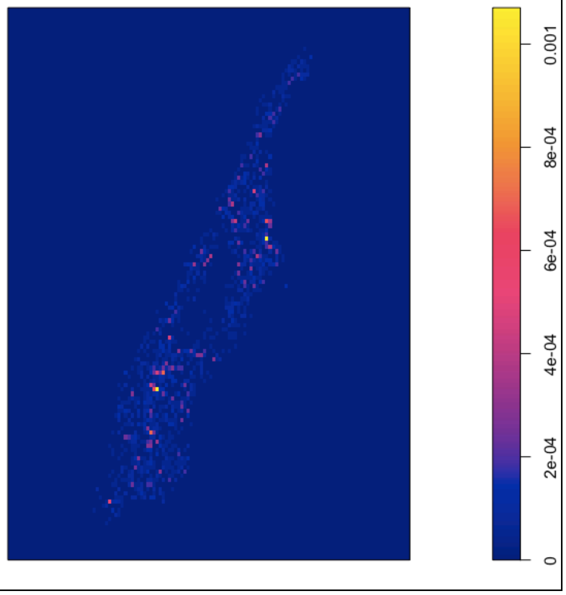
Left: Plot of point pattern without contours. Almost perfect shape of Manhattan
Right: Quadrant count of points in each panel.

The second step is to use kernel smoothing of bandwidth to aide in our examination of intensities of occurrences in a particular area. The three models used are 1) the R spatstat package's default method based on Likelihood Cross Validation, see Silverman (1986) 2) Diggle's bandwidth based on the minimization of the Mean Square Error (2014) and 3) Scott's bandwidth, based on multivariate kernel estimation; see Scott (2015). The best results came from using the default method and that of Diggle's bandwidths. Scott's bandwidth resulted in an inconclusive interpretation. Of particular note is that standard errors of the mean are available for the two aforementioned bandwidths and are small enough for us to be confident.

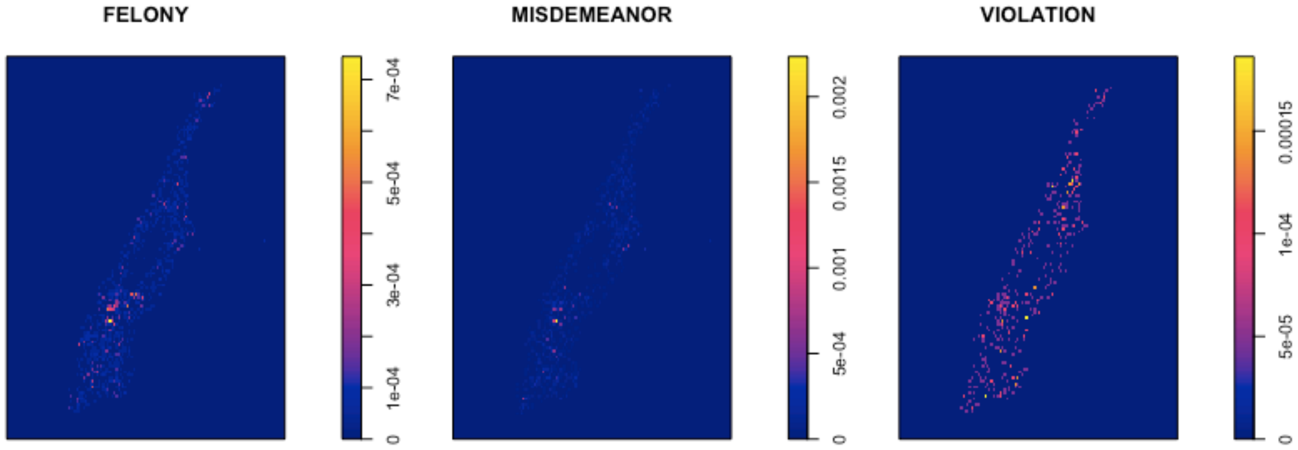


Above: Kernel smoothing based on Default Method provided spatstat

Below: Kernel smoothing based on the Diggle bandwidth

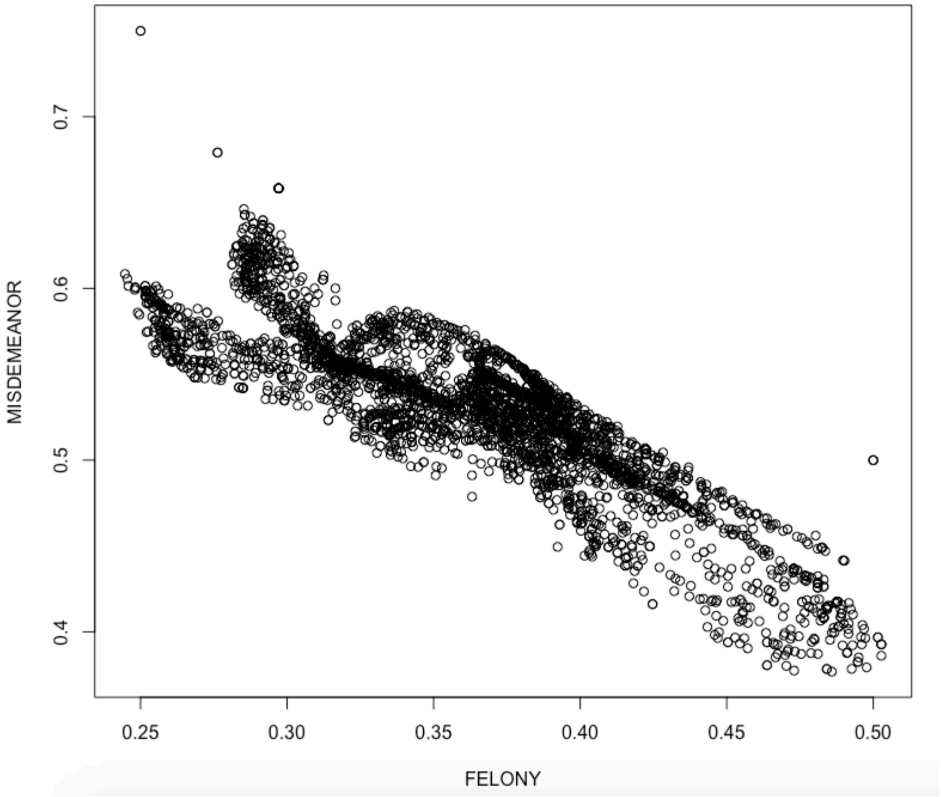


A finer breakdown of the point pattern can be rendered by separating the observations into factors based on crime classification. We discover that the Diggle bandwidth provided the best visual interpretation of hotspots on the map. Intuitively, we see that felonies are noticeable in Hell's Kitchen neighborhood of Manhattan. We also see that violations are present all throughout the borough but note that they only represent about 12% of the observations. The reader is advised to examine the plots with the scales in mind.



Point pattern for incidences of Felonies (**Left**), Misdemeanors (**Middle**), and Violations (**Right**). All based on the Diggle bandwidth.

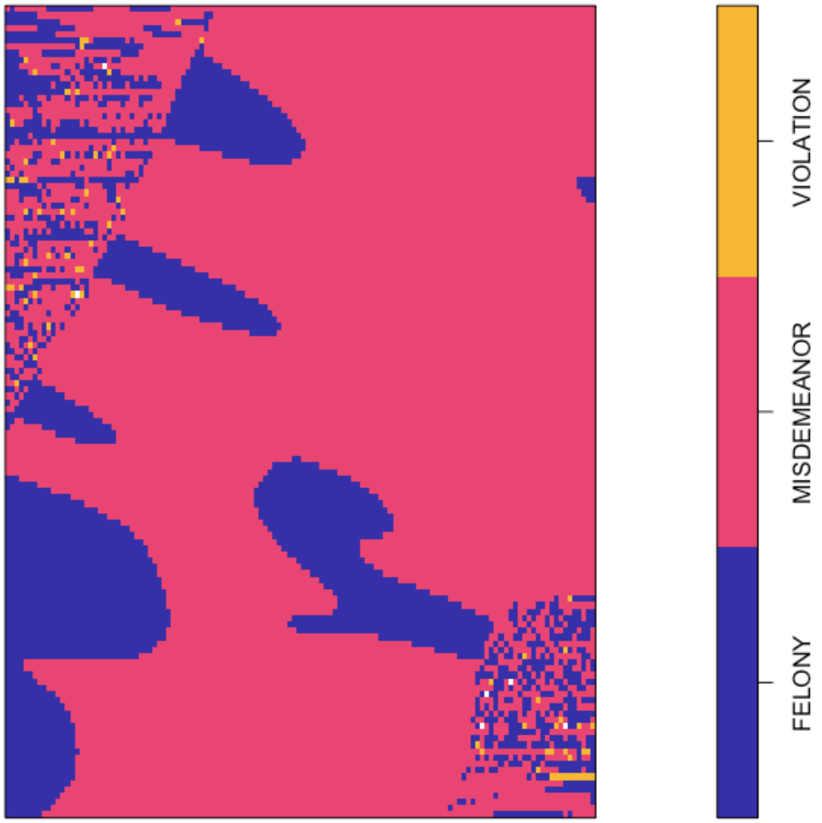
Following up with the multi-type point pattern, we are then interested in the relationships these types of crimes have among each other and how dependent they are. The main tool for this task is in the use of the relative risk function. In survival analysis, it is the ratio of the probabilities of an interested category to that of an excluded group. The package spatstat invokes leave-one out cross validation, a technique related to the Jackknife method of sampling. Here, Violations were left out but note that it only represents 12% of the total observations. We discover that there is an inverse relationship between the probabilities of Misdemeanors to that of Felonies.



Above: P-P plot of the two major crime categories showing inverse relationship.

With this result we can create a texture plot showing the highest probability of which crime type might occur in a geographic location. The below figure shows that Misdemeanors are more likely to occur in Manhattan but there are pockets of Felonies, especially closer to areas near the water surrounding the borough. Most importantly is that the crime categories separate from each other in a discrete manner and that there are no obvious homogeneous mixing of probabilities.

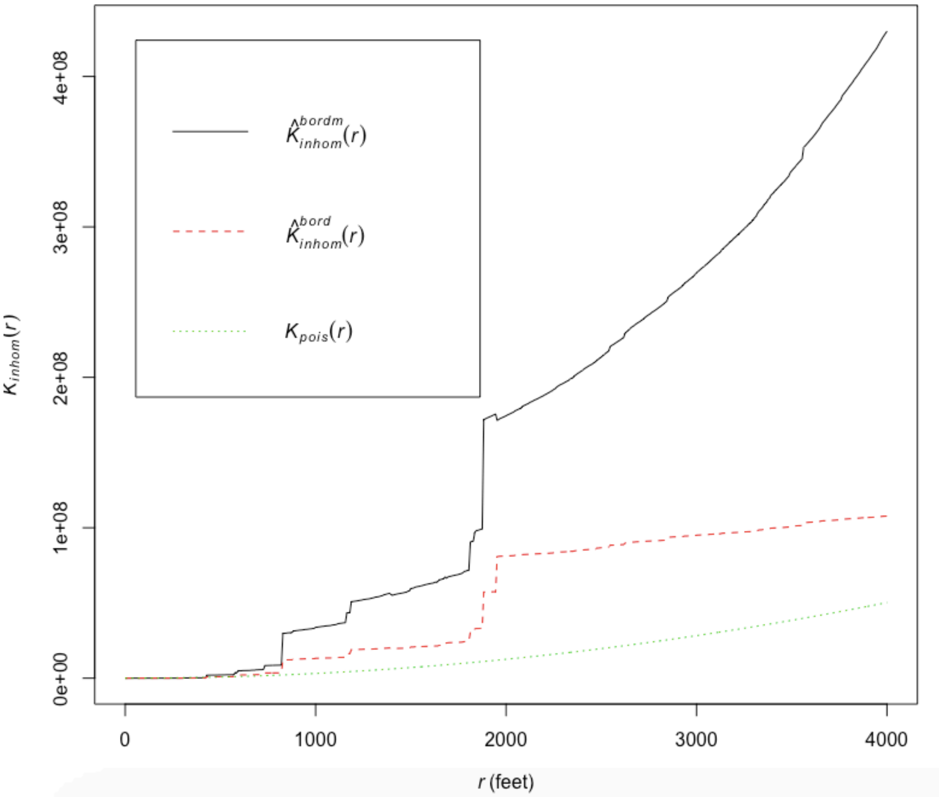
A more rigorous and formal statistical procedure is available in the form of a separation of type test, proposed by Baddeley (2016), where the null hypothesis is H_0: Offense types are not separated. This test, based on Monte Carlo simulation of sampling schemes, resulted in a p-value of exactly 0.05. In a case like this, it is a reminder that alpha level for type I threshold is not set in stone. With the help of the P-P and texture plots of crime types, we reject the null hypothesis and see that there is evidence that the factors of crime categories are indeed independent from each other. A result like this is an important step in the exploratory data analysis. See the discussion later in the paper for possible uses of this information.



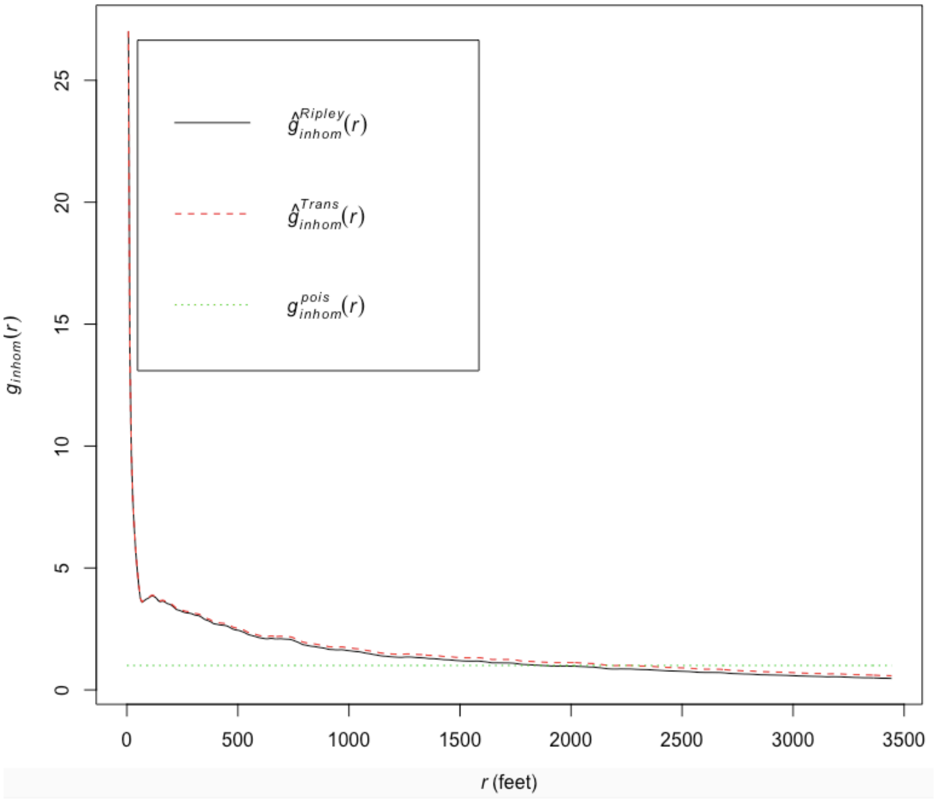
Above: Texture plot indicating the highest probability of crime type in the window frame. Note that the relative discrete separation of the types.

The last task is to examine the second order property of the point pattern. The first order property, represented in the form of the intensity function, is based on the average number of occurrences in a unit area. It might be not be helpful enough to lead to a formal model of regression.

By examining the second order property, we gain insights into the distribution of the underlying process. In particular interest is the justification for using a Poisson point process as the foundation for further modeling. The R package calculates both the inhomogeneous K-function by Baddeley (2016) and the inhomogeneous pair correlation function. These two functions are used for when non-stationary processes are observed and since we have rejected CSR, these seem appropriate to use. The below example is based on the univariate point pattern.



Above: Inhomogeneous K-function. Indicates the presence of clustering.



Above: Inhomogeneous Pair Correlation function based on G-function. Represents nearest neighbor approach.

The inhomogeneous K-function is the non-stationary version of Ripley's K-function and it indicates whether there are spatial dependencies among the points. The plot shows the curve of the inhomogeneous K-function dominating the curve for the homogeneous K-function, which is based on CSR. With evidence of spatial inhomogeneity we then ask whether there is evidence of clustering. For this, we use calculate the inhomogeneous pair correlation function which uses the G-function as a basis. It shows that points between at most around 1200 feet between each other are detected within the nearest neighborhood.

DISCUSSION

The Poisson model used as a basis for many of the current spatial techniques available to us can finally be justified after the exploratory data analysis of the first and second order processes. Future work can include a spatial regression of multi-type factors as covariates. We also concluded that the crime classification types are independent from each other and so one natural application can be a logistic binary regression to predict whether number of incidences correlate with each type of crime. The dataset also includes location data of where these incidences took place. These can further be integrated into the analysis as covariates of a regression model. Kriging the predicted number of incidences based on distance lag is also possible now that the second order process reveals an appropriate bandwidth distance.

References

Baddeley, A., Gregori, P., Mateu, J., Stoica, R., Stoyan, D., 2006. Case Studies in Spatial Point Process Modeling. *Lecture Notes in Statistics*. Springer.
Baddeley, A., Rubak, E., Turner, R., 2016. Spatial Point Patterns: Methodology and Applications with R. CRC Press.
Bivand, R., Pebesma, E., Gomez-Rubio, V., 2013. Applied Spatial Data Analysis with R. Springer.
Block, C., 1995. STAC Hot Spot Areas: A Statistical Tool for Law Enforcement Decisions. *Crime Analysis Through Computer Mapping*. Police Executive Research Forum.
Diggle, P., 2014. Statistical Analysis of Spatial and Spatial-Temporal Point Patterns 3rd Ed. *Monographs on Statistics and Applied Probability*. CRC Press.
Hermann, C.,2013. Street-Level Spatiotemporal Crime Analysis: Examples from Bronx County, NY (2006-2010), in *Crime Modeling and Mapping Using Geospatial Technologies*, Leitner, M. (Editor). Springer.
Ripley, B., 1981. Spatial Statistics. Wiley.
Scott, D., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization 2nd Ed. Wiley.
Silver, B.W., 1986. Density Estimation for Statistics and Data Analysis. CRC Press.