IDENTIFICATION OF CRASH HOTSPOTS BY KERNEL DENSITY METHOD

PRESENTED BY NITASHA TULI

M.A STATISTICS, HUNTER COLLEGE

Introduction

With the urbanization process around the globe, traffic accidents have increased at a rapid pace in recent decades, causing significant loss to life and property. Road traffic incidents account for 1.25 million deaths across the world each year. For safety analysis it is important to understand spatial and temporal crash patterns to identify high risk areas which can help agencies allocate limited resources more efficiently. Dealing with geographical data, such as traffic crashes, requires special attention since sometimes it displays an arbitrary structure such as crashes being concentrated or clustered in space. This can be examined using point pattern analysis.

This study is intended in using kernel density method (a geostatistical approach) to identify crash hotspots in road network and trace high risk locations for potential intervention. Identification of hotspots is a systematic process of detecting road sections that suffer from an unacceptable high risk of crashes. It is a low-cost strategy in road safety management where a small group of road network locations is selected from a large population for further diagnosis of specific problems, selection of cost-effective countermeasures, and prioritization of treatment sites. These identified sites are often called, by various terms in literature, hazardous locations, hotspots, black spots, priority investment locations, collision-prone locations, or dangerous sites

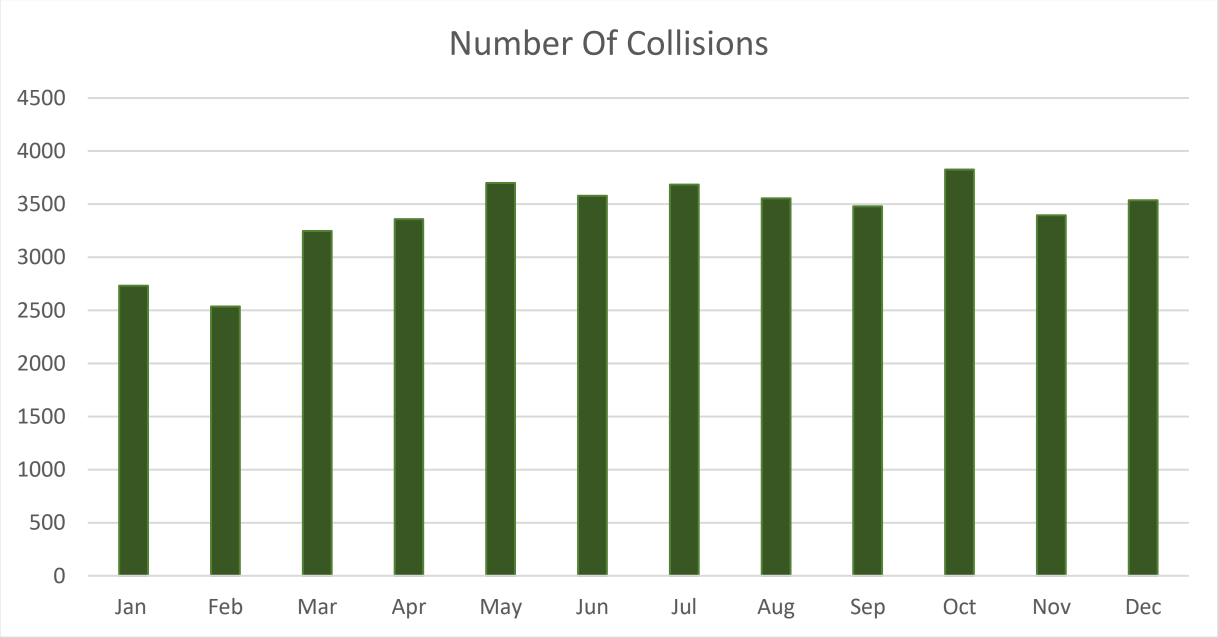
Study Area & Data Description

In the case study I will be looking at New York City Motor Vehicle Collisions Data for the year 2015. The dataset is hosted on the NYC Open Data website and includes all vehicle collisions recorded by the NYPD since July 2012. The records of collision have several key pieces of information attached to them, including what causes crashes, coordinates of majority of accidents, what type of vehicles are involved (motorcycles, buses, taxis, etc.), how many people are hurt or killed, whether they are motorists, passengers, pedestrians, or cyclists and where they happen. In this study I will try to find out intensity of collisions at different intersection in New York City. A total of 40,633 collisions were reported in New York City in year 2015. Below is the summary of collision and number of people injured/killed during that timeframe.

Year	2015	
Number of Motor vehicle collisions	40,633	
People injured or killed	Injured	Killed
MOTORISTS	2,945	3
CYCLISTS	1,150	1
PEDESTRIANS	2,207	15

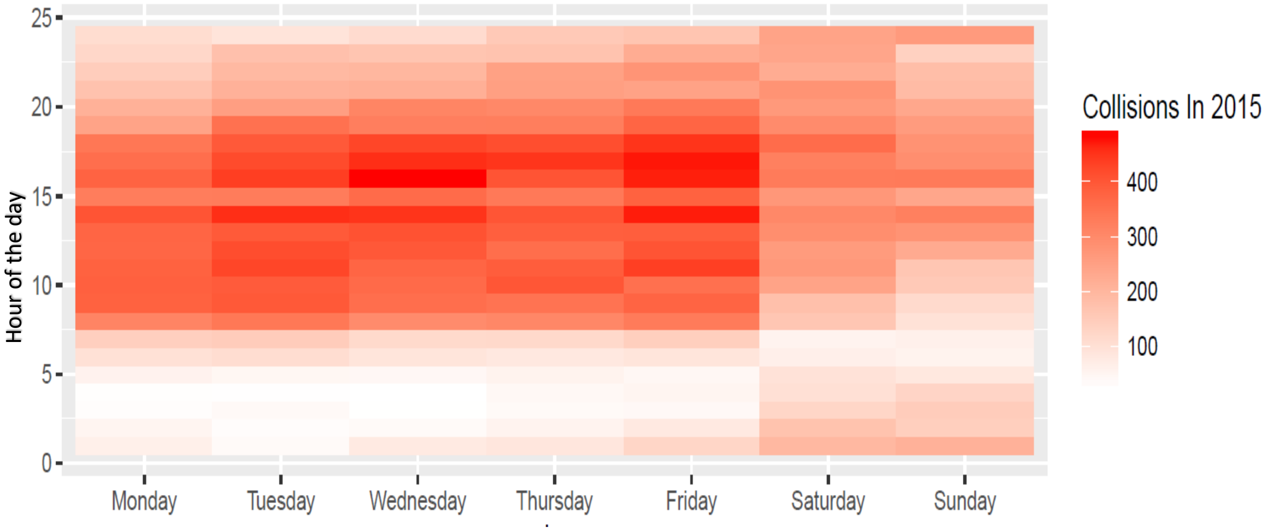
There are 40,633 collisions reported in New York City, almost 15% of which resulted in death or injury. Pedestrians are the most vulnerable having much higher casualty rate compared with other road users. 62% of people who got killed in road accidents were pedestrians. Usually the speculation is that the accidents were caused because of speeding but hypothesis is that the most accidents in NYC occur at difficult intersection and congested points. According to the data major contributing factor to collisions are driver inattention/distraction, turning improperly, failure to yield Right-of-Way, backing unsafely and many more.

The increase in population, employment and tourism has worsened the congestion in New York City. Below is the graph showing number of collisions in each month.

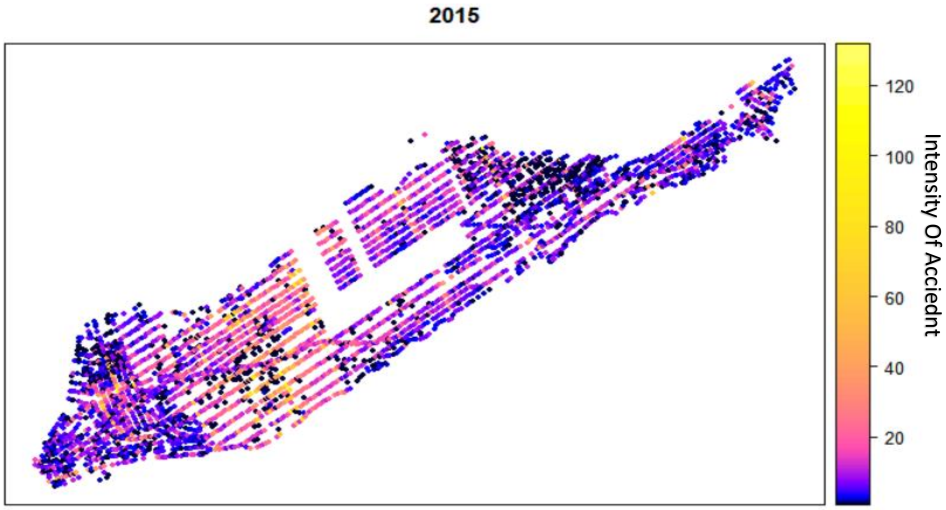


It shows less number of collisions in Quarter 1, possibly due to cold weather conditions reducing the traffic; but this gradually increases over rest of the months in the year.

Below is the Heat Maps to give better sense of accidents in New York City across time of the day vs. weekday -



The traffic collisions rise sharply after 8 AM, when hundreds of thousands of people are commuting into the city to get to work. While during the day, the frequency stays at similar levels, it begins to rise to its highest level after 4 pm. Statistics from the New York State Department of Motor Vehicles echo this finding. Traffic accidents are most likely to occur between 4 and 7 PM.



Above graph indicates number of collisions at any specific location in New York City, where yellow indicates most number of collisions.

Fatal vs Non–Fatal Accidents

A fatal-injury accident is defined as an accident in which at least one person (driver, passenger or pedestrian) was killed at crash. A non-fatal injury accident refers to an accident in which at least one person suffered injury but no fatalities occurred. For my study I have focused on collisions in which either someone is killed or number of injuries is greater than 4.

Complete Spatial Randomness

Spatial Randomness (CSR) describes a point process whereby point event occurs within a given study area in complete random fashion or not. It is also known as homogenous spatial point process. It implies that there are no regions where the events are more (or less) likely to occur and that the presence of a given event does not modify the probability of other events appearing nearby. Informally, this can be tested by plotting the point pattern and observing whether the points tend to appear in clusters or, on the contrary, they follow a regular pattern. Another way of doing this is quadrat test for CSR that uses chi-square statistics based on quadrat counts.

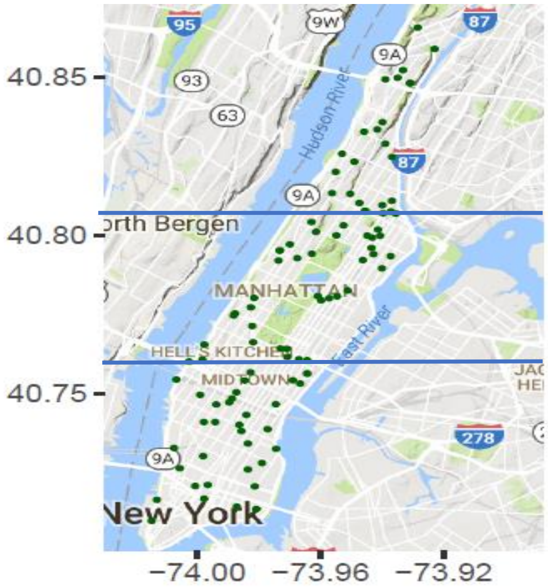
Quadrat Count Test for CSR

The quadrat test is a test of complete spatial randomness (CSR) that uses the χ^2 statistic based on quadrat counts. In the quadrat test, the study area window W is divided into sub regions called quadrats (W_1,W_2,....W_m) of equal area. The test counts the number of points that fall in each quadrat $n_j = n(X \cap W_j)$ for $j = 1, \ldots, m$. Under the null hypothesis of CSR, n_j are iid Poisson random variables. The following Pearson χ^2 test statistic assesses whether there is a departure from the homogeneous Poisson process:

$$\chi^2 = \frac{\sum_j (n_j - n/m)}{n/m}$$

A significant p-value indicates that the underlying point pattern is not a CSR

Testing CSR of 2015 Fatal Collision



Null Hypothesis :The given space have CSR

- Q1 = 23
- Q2 = 44
- Q3 = 31

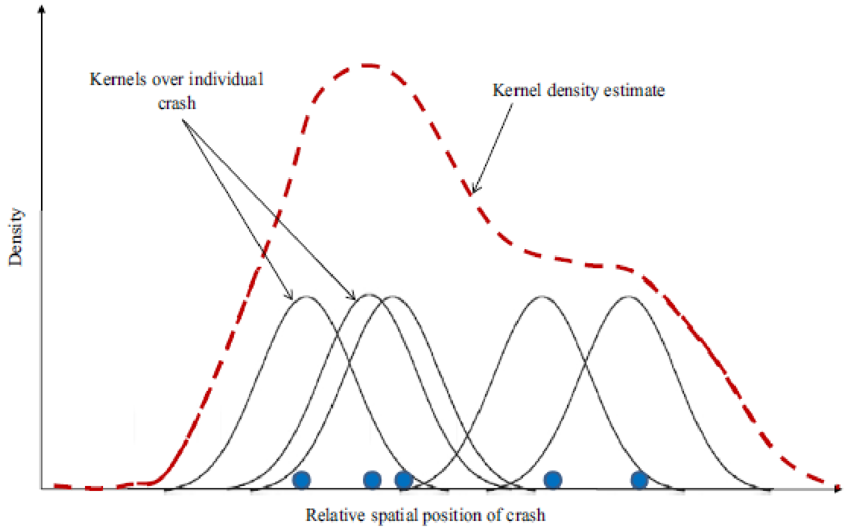
P value after running Chi square test = 0.0321

We will reject Null Hypothesis

No CSR

Kernel Density Method

The KDE, a non-parametric approach, is one of the most common used and well-established spatial techniques used to estimate the crash intensity for hotspot identification. In this method, a circular search area defined by a kernel function is placed over each crash (discrete points) resulting in individual smooth and continuous crash density surface as shown in the figure below.

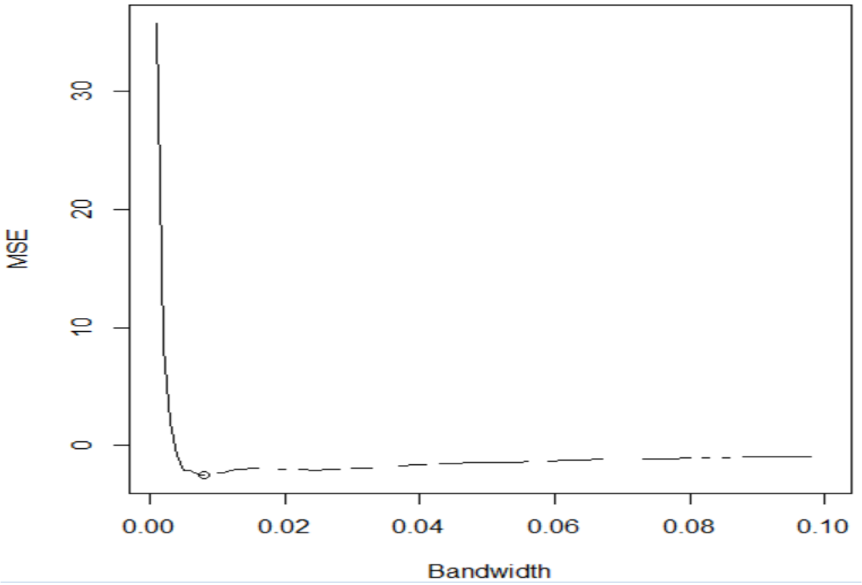

Kernels over individual crash
Kernel density estimate
Density
Relative spatial position of crash

Then, a grid of cells is overlaid over the study area. For a given cell, density is estimated by summing the overlapping density surface resulting from each crash point. This procedure is repeated for all reference grid cells. Note that kernel functions are symmetrical mathematical functions -

$$f(x,y) = \sum_{i=1}^{n} \frac{1}{n * 2 * \pi h^2} * \; W_i * K(\frac{d_i}{h})$$

Where $f(x,y)$ is the density estimate at the location (x,y), n is the number of observations, h is the bandwidth, K is the kernel function and d_i is the distance between location (x,y) and the i^{th} observation and W_i is the intensity of the observation. For the crash count, W_i is unit, which may vary when we consider different weights for different severities of crashes. The two main parameters which affect the KDE are bandwidth and cell size. Selection of bandwidth is more crucial than selection of kernel function. Bandwidth determines the extent of search area.

The optimal density of kernel density estimate is typically calculated by minimizing Mean Square Error. For my study, I have identified MSE using MSE 2d (splancs package) in R. Below graph shows different values of bandwidth and their associated MSE. The value that it minimizes it for Quartic Kernel is .008.



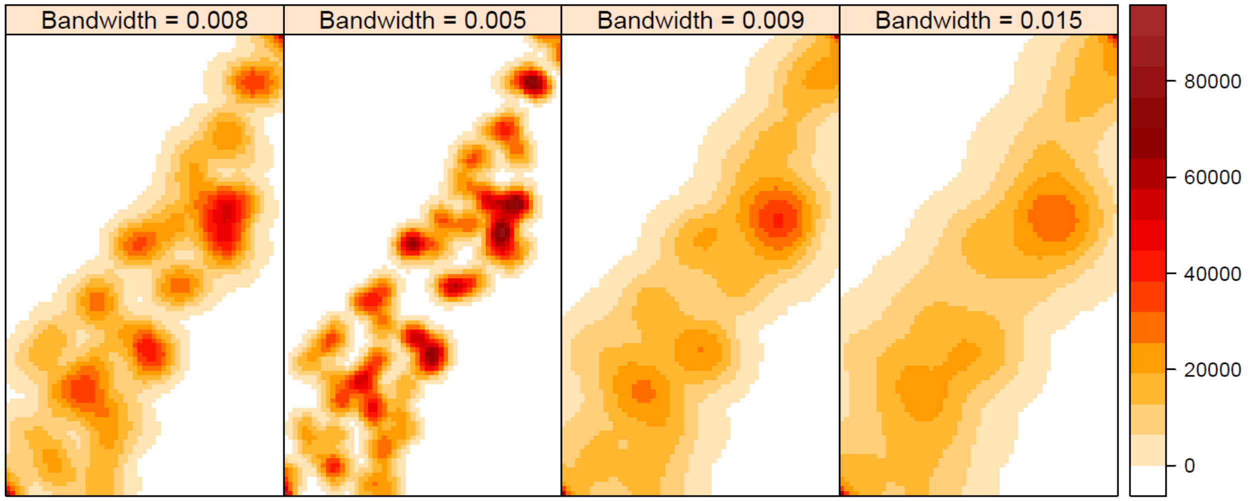
When estimating intensity by kernel smoothing, the key choice is not that of the kernel function but the bandwidth. Different kernels will produce very similar estimates for equivalent bandwidths, but the same kernel with different bandwidths will produce dramatically different results. The output of Kernel Density is presented in a raster format consisting of grid cells, where the size of cell has to be reasonable to represent crash cluster occurring in reality. Having a larger grid cell saves the processing time; however the information is likely to be averaged in larger area, thus resulting in loss of information. Meanwhile a smaller grid cell size increases the computational time and computes to a lower level of granularity which may not be necessary.

One of the attractive parts of the Kernel Density Method as compared to other variants of clustering method is that it takes into consideration of spatial autocorrelation of crashes. Moreover, this method is simple and easy to implement. This could be one of the reason that KDE method is widely used in road safety.

Results

After estimating the number of crashes over the grid, the output is presented in different color codes over a raster map. The next step is selecting a set of high risk zones i.e. the hotspots. There is no universal rule or threshold values to benchmark for what should be the hotspots. It is an arbitrary selection of cut-off values that screens relatively higher risk area over a given study area. Visually the proposed approach can detect spatial patterns of collision data and reasonably identify and rank unsafe pedestrian – vehicle crash locations.

Collisions in New York City 2015 (Kernel Density)



Bandwidth = 0.008 | Bandwidth = 0.005 | Bandwidth = 0.009 | Bandwidth = 0.015

The above graph is an output of Kernel Density Method purely basis of Collisions data characteristics. The finding seems obvious from above graph. The areas of Harlem and Midtown show much dense clusters of collisions than others. The study provides the framework for future research to incorporate crash predictions models to identify hotspots with an unreasonably high crash risk.

References

1) GIS based method for detecting high crash risk road segment using network kernel density estimation, by Afshin Shariat Mohaymany , Matin Shahri & Babak Mirbagheri.
   http://www.tandfonline.com/doi/abs/10.1080/10095020.2013.766396

2) Multiple Road traffic crashes and Injuries: A case study of Xi'an City, by Yong-gang Wang, Shen-sen Huang
   https://www.uni-obuda.hu/journal/Wang_Huang_Xiang_Pei_30.pdf

3) Accident Analysis and Prevention, by Jim Uttley, Steve Fotios
   www.elsevier.com/locate/aap

4) Identification of crash hotspots using kernel density estimation and kriging methods: a comparison, by Lalita Thakali • Tae J. Kwon • Liping Fu
   https://link.springer.com/article/10.1007/s40534-015-0068-0

5) NYC Collision Data on opendata
   https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

6) Factors affecting road crash modelling
   http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2238-10312015000200015

7) R Code References
   - https://stats.idre.ucla.edu/r/modules/subsetting-data/
   - https://rpubs.com/jimu_xw/crime_visualization
   - https://rstudio-pubs-static.s3.amazonaws.com/59317_9906f6e5167943a08f3e511849eb8831.html
   - http://www.people.fas.harvard.edu/~zhukov/spatial.R