Comparison of different crime prediction models in Bogotá

Francisco Barreras
Carlos Díaz
Alvaro Riascos
Mónica Ribero [*]

May 25, 2016

[*] We wish to thank Polkan García for useful comments and conversations and Sergio Camelo for excellent research assistance. We are also thankful with Daniel Mejía for his support and direction during the course of this research and we gratefully acknowledge financial support from the Center of Study For Security and Drugs (CESED) of the University of los Andes in Bogotá. The usual disclaimer applies.

Contents

Abstract

Hotspot mapping is a useful technique used to identify where to direct crime preventing resources. It consists of the geographical identification of areas with high crime frequency in the past in hopes of preventing future crime, relying on the existence of fixed determinants of crime as well as the existence of temporal phenomena that nonetheless have certain persistence. There are a number of techniques for identifying crime hotspots based on different models like point estimate models, kernel density estimation, spatial ellipses and spatio-temporal models. The model to be used is often selected in base of the ease of use or interpretability. The purpose of this article is to define a method to evaluate different models and find the most appropriate using historical crime data from Bogotá, Colombia.

1 Introduction

Hotspot mapping is a tool that has been used for a long time by academics and police departments to identify areas of high crime probability and for efficiently assigning limited crime preventing resources. Several mapping techniques exist and have evolved to become more visually succinct and easy to interpret. The choice of model, however, has implications over the resulting hotspots and therefore over the efficacy of crime prevention.

In this research we explore four different modelling techniques for crime prevention. Namely a point model, spatial ellipses model, kernel density estimation and a spatio-temporal model. We explore a measure of efficacy that allows a comparison between models in a train set - test set methodology, the "Precision Accuracy Index" (PAI) [6] . The models are fitted on crime data of the city of Bogotá, Colombia for the period of 2011-2012.

Our results indicate that Kernel Density Estimation models are superior than the other two in prevention of crime. This results are robust, as the expected measure of the PAI rather than the PAI was calculated by averaging over 20 different test weeks of data.

For the three models models that do not include explicitly the temporal dimension optimum results were obtained when estimating different models for each day of the week and ranges of 6 hours, accounting for the shifting dynamics of crime in the city of Bogotá.

2 Different Methods in Crime Prediction

Hotspot mapping has been a useful tool for predicting where crime tends to be higher. It's application has been used to aid police patrols (Hough and Tilley, 1998 ; LaVigne and Wartel, 1998, 1999). It's based on the premise that crime has a higher probability of occurring in geographical areas where it has occurred in the past. This provides an opportunity to exploit the large body of geocoded crime data that has become available in the last decades thanks to the sofistication of GPS and geocoding systems.

There are several techniques for grouping and representing the spatio-temporal data which have been developed to facilitate visualization and practical use in police departments. However there are few studies that compare different techniques in their ability to predict future crime [6].The aim of this article is to replicate results from [6] which compare a normalized "hit rate" for the different models and find which is more suitable for the task of predicting crime in Bogotá.

The difference in performance of different models can be explained by the trade-off that is present whenever making hotspot mapping. If the grouping of data is too granular, there is a failure in identifying clusters of crime [5] as the visualization becomes too "speckly". On the other hand, techniques which group data according to different geographical criteria encounter the MAUP (Modifiable Aereal Unit Problem) [14], in which the resulting statistics are largely sensible to the arbitrarily chosen boundaries for grouping data.

Some measures have been proposed in the past few decades for comparing different hotspot mapping techniques in their ability to predict the occurrence of crime. A very intuitive one is the "hit rate" that counts how many future crimes occurred in the areas marked as hotspots by the model. However, his measure is of little use since it is not normalized by area.Thus, one can have a model with 100% hit rate by marking the whole study area as a hotspot. Coping with this problem *Bowers et. al.* [4] proposed a measure called the "Search Efficiency Rate" defined as the number of events per square Kilometer in areas marked as hotspots.

As noted in [6], the "Search Efficiency Rate" does not consider the relation with the size of the study area. Two settings with a Search Efficiency Rate of 20 crimes per Km^2 can largely differ in the percentage of crimes contained in the hotspots and the size of them. We would like to capture the fact that more specific hotspots define a more useful model when thinking of the deployment of crime-preventing resources.

In [6] the authors introduce a measurement that normalizes by the number of total crime in the evaluation of the study area as well as the total area of study. Namely, the **PAI** or “Precision Accuracy Index”. This measure was found as the most convenient for comparing between models, not only by the normalizations that allows comparability but also because it captures neatly the relation between type I and type II errors, both of which should be small in a prediction model.

2.1 Points Model

The points model is the simplest aggregation strategy to detect hotspots. It is founded on the idea of hotspots being micro areas defined on the order of one-block long street segments [15, 16], so aggregation over a fixed size grid of this order of magnitude allows for the detection of such hotspots when crimes concentrated in a single cell exceed certain threshold.

This model is very simple to implement; the only parameters to estimate are grid size and the hotspot threshold and it is been a practical tool in the prediction of small temporary hotspots [9]. However it has obvious limitations as an actual point cluster might not be detected if distributed over contiguous cells. Moreover, the fixed nature of the grid makes it impossible to detect concentrations of crime that conform a hotspot larger than the grid size failing consequently to capture the spill-over effect of hotspots in neighbouring cells.

2.2 Spatial Ellipses Model

The first ellipses model started of as a program to identify “hot circles” using crime spatio-temporal data [2]. However, these circles often overlapped and degenerated into ellipses. The main idea remains the same, to identify highly concentrated points in the map and fit a “standard deviation ellipse” in each one. The ellipses’ size and orientation indicate the underlying distribution of crime.

Spatial Ellipses model have been widely used thanks to their implementation in the ”Spatial and Temporal Analysis of Crime” software (STAC). Its applications range widely; see for example [12] and [3]. They are, undoubtedly, a popular and useful method for outlining clearly regions were actions should be taken.

Its benefits include their ability to outline regions without using predefined boundaries like police quadrants or political-administrative boundaries [12].Also, it is a method compatible with multiple GIS applications and it needs few parameters for estimation. However, it has been criticized because typically crime distributions do not follow ellipses and because it doesnot allow significant analysis inside the outlined regions or negate any comparison with the events outside such regions [6].

The fitting of a standard deviation ellipse for a set of points is straightforward. The magnitude of the two axes of each ellipse is given by the following equations:

$$SDE_x = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n}}$$

$$SDE_y = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2}{n}}$$

Each ellipse is oriented in the direction of maximum dispersion for each set of points. The other parameter to be estimated is the search radius which is the underlying parameter in a clustering algorithm that assigns every point in the data set to a cluster. The search radius is the maximum distance that can separate two points belonging to the same ellipse.

2.3 Kernel Density Estimation

KDE is a smoothing technique that fits gaussian distributions centered at every point in the sample and adds up to produce a smooth map over the study area which captures the density or volume of point/crime across the study area without conforming to predetermined areas like ellipses or fixed grids [6]. The search radius for every gaussian distribution to be fit (bandwidth) captures the level of aggregation desired to build the hotspots, and the result is represented as intensity values on a grid whose dimensions also ought to be calculated and determine the granularity in the visualization of such hotspots. However there is no universal doctrine on how to set these parameters [6].

Hotspots can be derived from a KDE estimation by choosing a threshold for the intensity of crime at every cell. E.G. the top $x\%$ of cells with higher intensity. This procedure clearly defines geographical zones which are not bounded by arbitrary divisions like quadrants or census that could bias the analysis. Other advantages are the visual simplicity and allure and the increasingly high availability of its implementation among GIS software [6]. It has widespread applications documented in literature like [8] and [7].

For the purpose of this research the most relevant comparison of KDE and other models is in [6] where KDE is proven to outperform other models like Ellipses and Thematic Grid Mapping. This probably follows the shape flexibility that KDE allows and the capturing of spill-over effects.

Kernel density estimation, differing with the other two models, does not require clustering of points. It simply consists of fitting a gaussian density of probability about each point in the set, and then the corresponding intensity map consists of the sum of all those, sometimes overlapping, densities. For each point we set the following density.

$$K(x, x_i, \Sigma) = \frac{1}{\sqrt{(2\pi)^3 det(\Sigma)}} exp\left[-\frac{1}{2}(x - x_i)^t \Sigma^{-1}(x - x_i)\right] \quad (1)$$

The resulting intensity map u is therefore:

$$u(x, \Sigma) = \frac{1}{n+1}\sum_{i=0}^{n} K(x, x_i, \Sigma) \quad (2)$$

2.4 Temporal Dimension

A crucial role in crime data analysis is the time dimension. It is clear that crime events do not have the same intensity through time. Criminal activity depends on factors like potential victims, motivated offenders and capable guardianship [9] that are dependent on the time of day and the day of the week. Furthermore, many theories have shown a contagious behaviour of crime of events [10]. Hence temporal models have been enriched by drawing a parallel with seismology where seismic movements are structurally different than their aftershocks [13]. Some suppositions on the estimated models in this paper attempt to capture these relations to produce more accurate models.

Of interest for this paper was the model proposed by [13] that identifies crime clusters in time and space using KDE and a self-exciting point process with intensity

$$\lambda(t, x, y) = \nu(t)\mu(x, y) + \sum_{k:t_k<t} g(t - t_k, x - x_k, y - y_k; M) \quad (3)$$

.

Here, as in seismology, training data is declustered into background and afterschock events. The first group is modeled with a Poisson process with intensity $\mu(x, y)$. Aftershock events occur in time and space accordingly to kernel g that is supposed to capture the decreasing probability in time and space, this risk being higher when closer to a background event.

Due to the size of data needed to fit kernel g, a nonparametric montecarlo simulation that avoids using all data is used to fit the model. Let $\{(t_k, x_k, y_k)\}_{k=1}^{N}$ be a sample from the preceding model, p_{ii} the probability that event i is a background event and p_{ij} the propability that event j is an aftershock of event i. Then

$$p_{ii} = \frac{\mu(t_i, x_i, y_i)}{\lambda(t_i, x_i, y_i)} \tag{4}$$

$$p_{ji} = \frac{g(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(t_i, x_i, y_i)} \tag{5}$$

The next iterative algorithm is used to estimate ν_n, μ_n, g_n.

1. Use $P_{n-1} = [p_{ij}]$ to sample $\{(t_k, x_k, y_k)\}_{k=1}^{N}$background events.
2. Estimate ν_n, μ_n, g_n accordingly with the previous samples.
3. Update P_n using 4 and 5 with the new values of ν_n, μ_n, g_n

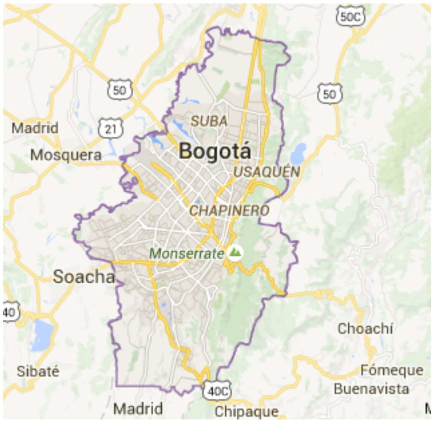
Gaussian kernels were used for all intensities.

3 Data and Methodology

The data employed corresponds to the historical crime data of the urban area of Bogotá, Colombia. Bogotá is the capital, the largest and crowdest city in Colombia. It has an area of 1,587 square kilometres divided in 20 localities or districts. The city has an urban population of 7,878,783 habitants to 2015, giving a density of 4,310 habitants per squared kilometer. By 2014 the proportion of people in poverty was 10,1 % and the average income was $999.195, equivalent to 1.62 Minimum salaries.

During the 90's Bogota was considered one of the most dangerous cities in the world, thus crime has being a popular policy topic.

Figure 1: Bogotá, Colombia. Map Data 2016 Google



The data used was collected and geocoded by the *Policía Metropolitana de Bogotá* from 2004 to 2014. This database comprises 329.793 individual observations, each containing its coordinates, time of occurrence (to the minute) and crime type. The study area is the whole urban area of Bogotá and the models were trained with data from year 2011 and evaluated in a train test- test set methodology in 20 randomly chosen weeks of 2012.

The data are coded by crime type, with category 2 being a redundant subset of category 3. The different categories are listed in table 1 and their distribution in **??**. Since it was not the purpose of this study to compare dynamics of different types of crimes, we made predictive models over a pooling of crime types 1,3,5,6,7 and 9, types of crime which have a victim involving burglary or aggression. The choice to pool the data follows the need to have enough data locally since data would be segmented by time.

Table 1: Crime Type Coding

Crime	Code
Homicide	1
Injury (Old)	2
Injury	3
House Theft	5
Motorcycle Theft	6
Auto Theft	7
Personal Theft	9
Drug Traficking	14

The models that did not incorpore temporal variables to map hotspots were fitted on disjoint aggregations of data based on the day of the week and

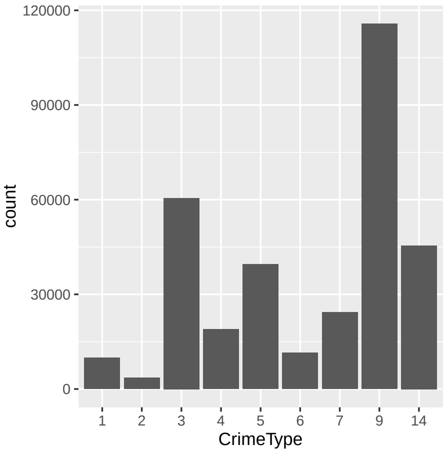


Figure 2: Frequency by crime type

certain hour ranges, in order to compare their performance across time. The optimal choice of hour ranges was decided using the PAI on a test set, balancing the trade off between a specific model for every hour of the day and the lack of enough data to make accurate predictions on each model if the division was to granular. Models with a 6 hour range were chosen.

In order to validate the models we performed a standard training and test set methodology (common in most machine learning applications) where models are tested on data that were not used during the training. Testing models on the data they were trained in leads to erroneous conclusions because models tend to ”overfit” the data, meaning that they can identify apparent or noisy relations in the training data that do not hold in general.

In our framework, models were trained over different time windows before 2012 and tested in 20 randomly chosen weeks of 2012. Testing in different weeks is important to avoid season specific crime phenomena that could lead to conclude that models are not well suited to predict. E.g. the first week of the year, or weeks with very particular weather phenomena. The model used for each time period (day and hour) is used in the corresponding testing.

The measure used to compare models is the aforementioned Precision Accuracy Index or PAI which can be calculated once a model marks certain areas as hotspots. The PAI consists of the quotient of two measures, the hit rate and the percentage of area. In order to make this measure more robust, the PAI was averaged over the 20 weeks used in the test set.

$$\text{PAI} = \frac{\text{Hit Rate}}{\text{Percentage of Area}}$$
$$\text{Hit Rate} = \frac{\text{Crimes included in Hotspots}}{\text{Total Crimes}}$$
$$\text{Percentage Area} = \frac{\text{Area of Hotspots}}{\text{Total Area}}$$

All of the estimated models were fitted using R statistical software and the libraries **sp** and **rgeos**. As it was mentioned before, the point model is the simplest in estimation. It requires only to specify the size of each cell in the grid and then counting how many events in the considered time frame fall within each cell. For this estimation, six hour intervals were chosen for every day of the week and the counts were for such period in the year of 2011. Meaning, for example, that a model was created for Monday in the 00:00-06:00 time period, by aggregating all events fow the same period of every Monday of 2012. The cells in the grid were set to 150 Mt. Hotspots were defined as the top 2% cells with higher counts. Notice that hotspots are different depending on the moment of the day and of the week, capturing crime dynamics.

Due to the low level of aggregation that comes from the point model (there is no spill over from cell to cell, so hotspots are very sparse and narrowly defined) the PAI is not the best measure for model precision in prediction. This is because the area defined as hotspots is forcefully very small (only about 10 % cells have crimes at all). Thus the PAI is often 0 when the model misses and has a very large value when the model hits something. In order to solve this issue we compare the point model only to KDE using very similar areas (about 10 %) and calculating the hit rate (which is essentially equivalent to the hit rate). In general the authors don't recommend the PAI as a measure with very narrowly defined hotspots as it works better with medium sized hotspots.

The ellipse model was fitted in the same time periods using the **aspace** R library. The search radius was set to 250 Mt. as it is documented to be an efficient radius. [11]

The only parameters to be set in the KDE model are the bandwidth (captured by the Σ) which determine how disperse each density is and therefore allowing for the overlapping of such densities and the forming of hotspots, the optimal bandwidth was used based on the best resulting PAI and was set to 150 Mt. The other parameter concerns visualization and application and it is the cell size of the grid of outputted intensities, a cell size of 150

Mt. was chosen as well.

Figures 4, 5, 6 below show visualizations for three of the models, restricted to Usaquén, a locality in Bogotá.

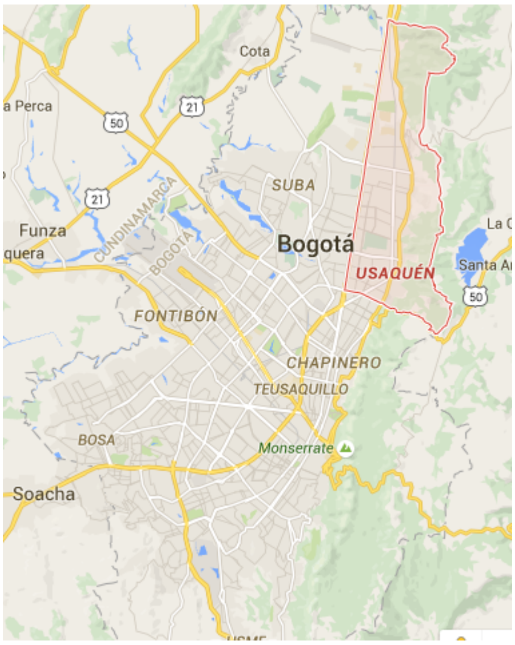


Figure 3: Locality of Usaquen

4 Results and Conclusions

Next we present the PAI results for the ellipses and KDE models. The results are presented for each of the training weeks and the average (which is the expected value of the PAI for that model).

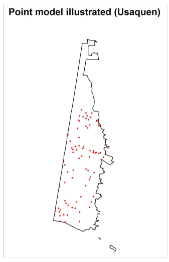


Figure 4: Visualization of point model in the locality of Usaquen.

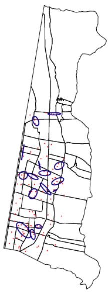


Figure 5: Example of spatial ellipses in the locality of Usaquen.

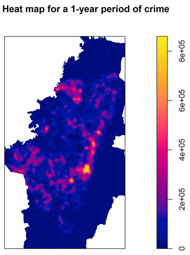


Figure 6: Heatmap generated with KDE

Week	PAI		
	Ellipses	KDE	Temporal
1	3.76	8.10	5.083324
2	6.65	6.64	3.913320
3	3.70	6.95	6.023466
4	3.24	5.91	4.573324
5	3.98	6.02	3.243143
6	3.67	6.30	4.023410
7	2.93	7.10	3.015682
8	2.46	5.77	4.126600
9	3.02	6.07	4.775211
10	3.56	7.57	4.601124
11	5.24	7.82	3.124406
12	3.45	7.38	4.105612
13	4.55	6.03	4.553678
14	3.23	6.86	3.875501
15	3.68	5.85	3.305000
16	4.80	5.68	3.301474
17	3.24	5.28	7.065500
18	2.55	5.08	4.025316
19	3.99	5.94	4.562433
20	5.92	6.09	3.679918
Avg	3.88	6.42	4.248872

Table 2: PAI of models

Now we compare the hit rates of the point model and KDE model on a comparable area (about 10 %) to evaluate which model is more accurate.

Week	Hit Rate	
	Points	KDE
1	0.09	0.41
2	0.11	0.74
3	0.12	0.77
4	0.10	0.66
5	0.10	0.67
6	0.11	0.70
7	0.08	0.36
8	0.10	0.64
9	0.10	0.67
10	0.08	0.38
11	0.09	0.39
12	0.08	0.37
13	0.10	0.67
14	0.11	0.76
15	0.10	0.65
16	0.09	0.63
17	0.09	0.59
18	0.08	0.56
19	0.10	0.66
20	0.10	0.68
Avg.	0.09	0.59

As it can be appreciated, KDE models outperform the ellipses and point model. This is consistent with results in the literature [6]. This is probably due to both the ability of KDE to aggregate or cluster events in a flexible way, a feature that the points model lacks; But also because it can define hotspots in a very fitted way, and not necessarily in the restricted geometries of the ellipses model.

5 Developed Software

As an application of the conducted research, the authors designed a mobile-compatible software for predictive policing in the urban area of Bogotá. It is intended to direct crime preventing-resources to where they are most needed. The underlying model is a Kernel Density Estimation (KDE) trained over a year long sample (intended to be the year before the application is used).

5.1 General Interface

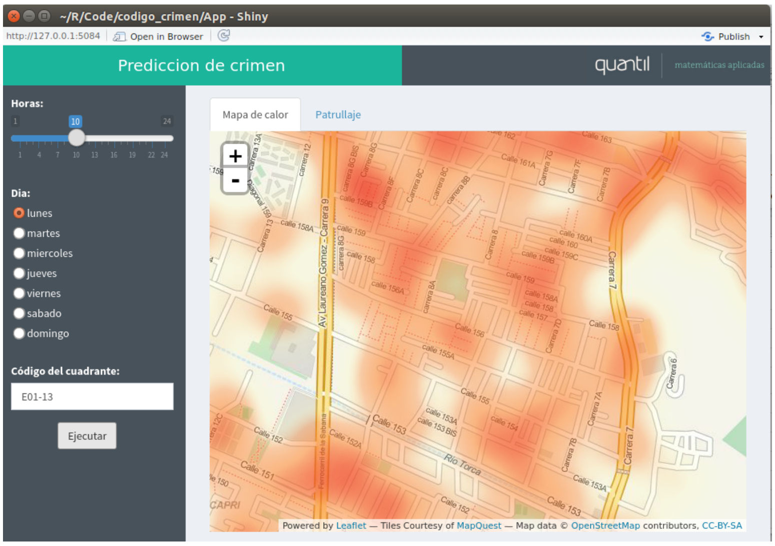
The application was developed using an RStudio package called Shiny [1] which provides an HTML compatible environment for programming within R. It uses interactive lightweight maps to show the user the distribution of crime and the recommended segments to patrol or deploy crime preventing resources.

The application requires the end user to specify a day of the week and the time of the day and loads the appropriate crime intensity matrix which is displayed in the interface. Additionally, the user is required to input in a text box the name of the police quadrant where the map should be centered.

It's compatible with multiple operating systems and web browsers, including mobile phones capable of processing HTML (most smart phones).

5.2 Heatmap Functionality

The first functionality is an interface which displays a heatmap of crime intensities over a map of Bogotá centered at the desired police quadrant. It shows the user where areas of high intensity crime are located by colouring them in a scale of red proportional to the intensity in the crime intensity matrix calculated by the KDE.



Horas: 10

Dia:
● lunes
○ martes
○ miercoles
○ jueves
○ viernes
○ sabado
○ domingo

Código del cuadrante: E01-13

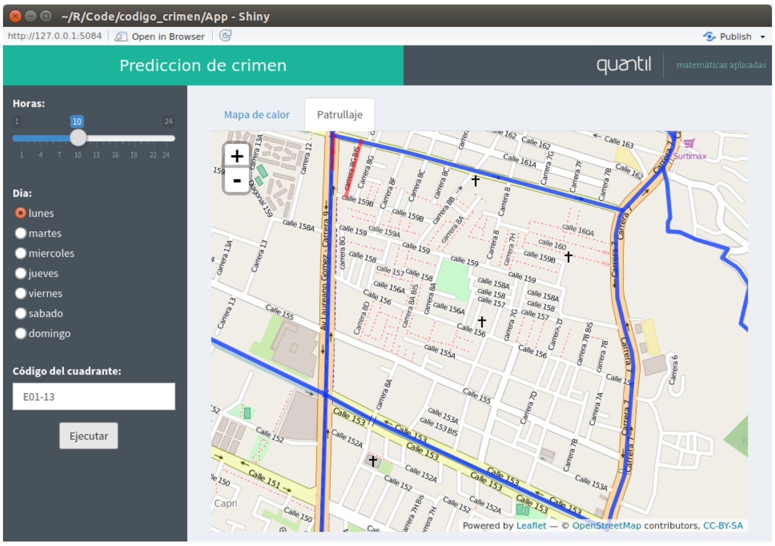
Figure 7: Heatmap Functionality

The user may scroll and zoom about the map to compare intensities with other areas. The visualization of this heatmaps is a quick way of spotting crime patterns and hotspots. The application displays different intensities of crime depending on the time and day of the week chosen, capturing temporal crime dynamics.

5.3 Segment Patrol Functionality

The second functionality is intended to be a more specific display intended to point out the exact street segments on which resources should be deployed. According to the current restriction that police officers must restrict their patrols to their quadrants, this functionality shows the segments which intersect with the highest-intensity areas of the quadrant.

This high-intensity areas are defined as the 1% of cells with highest crime intensity in the underlying crime intensity matrix generated by the KDE. As before, the recommended patrol may change depending on the day and time of the day that the user inputs. When the user chooses a given quadrant, the display shows blue line segments delimiting the specified quadrant and red line segments signalling the streets to be patrolled as shown in the following examples:



~/R/Code/codigo_crimen/App - Shiny

http://127.0.0.1:5084

Prediccion de crimen

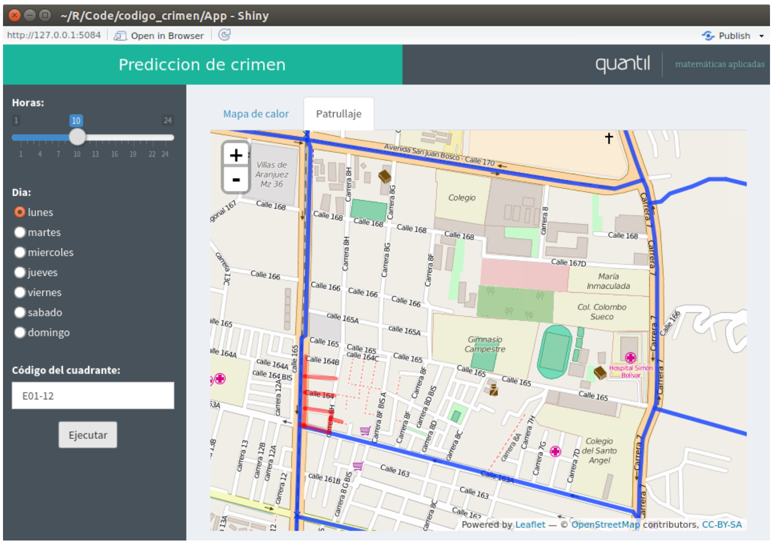
Horas: 10

Dia: ● lunes ○ martes ○ miercoles ○ jueves ○ viernes ○ sabado ○ domingo

Código del cuadrante: E01-13

Powered by Leaflet — © OpenStreetMap contributors, CC-BY-SA

Figure 8: Segment Patrol Functionality



~/R/Code/codigo_crimen/App - Shiny

http://127.0.0.1:5084

Prediccion de crimen

Horas: 10

Dia:

● lunes
○ martes
○ miercoles
○ jueves
○ viernes
○ sabado
○ domingo

Código del cuadrante: E01-12

Figure 9: Segment Patrol Functionality

References

[1] *Easy web applications in R*, 2013. Available online at www.rstudio.com/shiny/.

[2] Carolyn Rebecca Block. Stac hot-spot areas: A statistical tool for law enforcement decisions. In *Crime analysis through computer mapping. Washington, DC: Police Executive Research Forum*, pages 15–32. Citeseer, 1995.

[3] Kate Bowers. Exploring links between crime and disadvantage in north-west england: An analysis using geographical information systems. *International Journal of Geographical Information Science*, 13(2):159–184, 1999.

[4] Kate J Bowers, Shane D Johnson, and Ken Pease. Prospective hotspotting the future of crime mapping? *British Journal of Criminology*, 44(5):641–658, 2004.

[5] Spencer Chainey and Jerry Ratcliffe. *GIS and crime mapping*. John Wiley & Sons, 2013.

[6] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[7] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding hotspots. 2005.

[8] Victor Goldsmith, Philip G McGuire, John B Mollenkopf, and Timothy A Ross. *Analyzing Crime Patterns: Frontiers of Practice: Frontiers of Practice*. Sage Publications, 1999.

[9] Wilpen L Gorr and YongJei Lee. Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, 31(1):25–47, 2015.

[10] Shane D. Johnson. Repeat burglary victimisation: a tale of two theories. *Journal of Experimental Criminology*, 4(3):215–240, 2008.

[11] Ned Levine. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. National Institute of Justice, 2013.

[12] Barnes E. Martin D. and Britt D. *he Multiple Impacts of Mapping it Out; Police, Geographic Information Systems (GIS) and Community Mobilization During Devil?s Night in Detroit, Michigan.* Crime Mapping Case Studies: Successes in the Field . USA: Police Executive Research Forum., 1998.

[13] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2012.

[14] S. Openshaw. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, 38, 1984.

[15] David Weisburd, Shawn Bushway, Cynthia Lum, and Sue-Ming Yang. Trajectories of crime at places: A longitudinal study of street segments in the city of seattle*. *Criminology*, 42(2):283–322, 2004.

[16] David L Weisburd, Elizabeth R Groff, and Sue-Ming Yang. *The criminology of place: Street segments and our understanding of the crime problem*. Oxford University Press, 2012.