Predicting crime in Bogota using Kernel Warping

Sergio Garrido

Universidad de los Andes

September 2017

Outline

Motivation

- Big cities municipalities spend a certain amount of their budget on security.
- Since the resources allocated are scarce, an economic problem arises and it is the work of economists to do this in an efficient fashion.
- Hence being able to accurately and efficiently predict crime improves cities livability.

Literature review

- (Kianmehr & Alhajj, 2008) evaluate different types of SVM to predict crime in Columbus, Ohio and St. Louis, Missouri. Compared to other machine learnings models, neural networks and k-nearest neighbors, the SVM performs considerably better.
- (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011) Use a self-exciting point process model, from the seismology literature, to predict crime in Los Angeles.
- (Kang & Kang, 2017) Propose a deep learning neural network to predict crime using both spatio-temporal, socio-demographic and urban data to predict crime in Chicago, Illinois. The architecture of the Neural Netowrk, a Convolutional Neural Network, offers better predictions than SVM and Kernel Density Estimation models

Literature review

- (Barreras, Diaz, Riascos, & Ribero, 2016) tested several machine learning methods previously developed on the crime literature. They find that the model proposed by (Mohler et al., 2011) is the best performer for the colombian case.
- (Zhou & Matteson, 2016) developed the kernel warping methodology to predict ambulance demand in Melbourne, Australia. Their model performed better than industry standard models.

Data

The data we use to train the model comes from the Bogotan Metropolitan Police. They collected and geocoded all the data. Data was collected between 2004 and 2014 accounting. However, to make it comparable with previous works, the model will be trained only for 2011 and forecasted out of sample using 20 weeks in 2012. The available variables on the data set are a discretized temporal domain i.e. $T = \{1, 2, ...\}$ and a continuous spatial domain $S \subset R^2$

Table No. 1

Crime type	Code	Frequency
Homicide	1	1132
Injuries (before)	2	423
Injuries	3	6615
House theft	5	4077
Motorbike theft	6	1467
Car theft	7	2317
Robery	9	13398
Drug Trafficking	14	3695

Graph No. 1

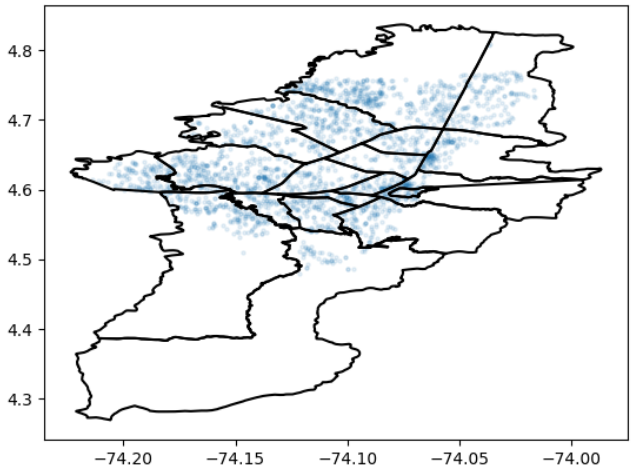


Figure 1: Saturday night

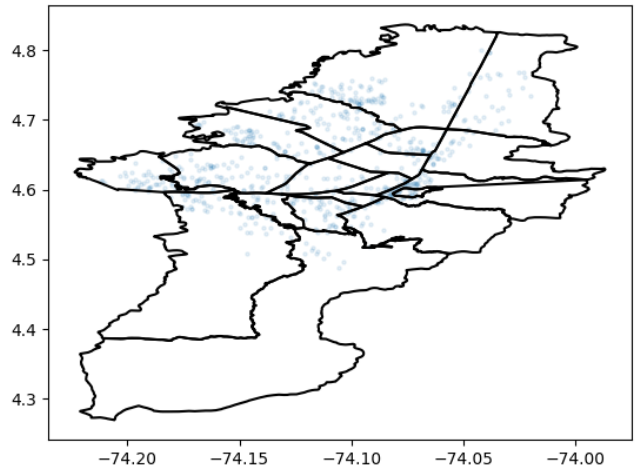


Figure 2: Tuesdays morning

Figure 3: Crime distribution for two different time frames

Graph No. 2

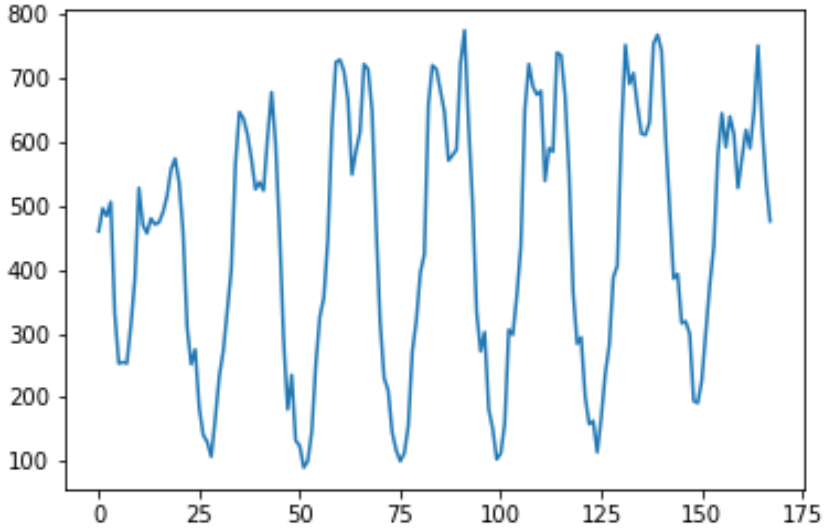


Figure 4: Distribution across the weekly aggregation

Methodology

- This paper uses the Kernel Warping methodology inspired in (Zhou & Matteson, 2016). This model is a transformation of the industry standard Kernel Density Estimation (KDE).
- This methodology is superior to previously used methodologies in the crime literature because it takes into account the geometry of the places where the crime occurs, thus some local relations can be exploited improving the accuracy of the out of sample forecast.
- However, this model can be both computationally and mathematically expensive

Methodology

- KDE, let K be a distribution function, n the number of observations and h the bandwidth:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{(x - x_i)}{h})$$

- The kernel warping methodology modifies the function K_h such that it approximates to the geometry of the city.
- A bandwidth too Large wipes out local features where there is sufficient data, a bandwidth too small leads to spurious peaks where the data is sparse.

Definitions

- Let s_t be the location of the i-th crime arising from the t-th time period, for $i \in \{1, ..., n_t\}$, where n_t is the total number of crimes in the t-th period.
- We assume $\{s_t, i : i = 1, ...n_t\}$ for each period follow an NHPP (Non Homogeneous Poisson Process) over S, with positive intensity function λ_t.
- $\lambda_t(s) = \delta_t f_t(s)$, for $s \in S$. Here $\delta_t = \int_S \lambda_t(s)$ is the aggregate demand intensity over the spatial domain and $f_t(.)$ is the continuous spatial density of the demand at time t such that $f_t(s) > 0$ and $\int_S f_t(s)ds = 1$.

Methodology

The proposed methodology does this by:

- Spatio-temporal KDE
- Choosing a point cloud from the training data
- Creating an adjacency matrix based on some decision rule. Creating a Laplacian matrix based on the adjacency matrix
- Modifying the function K_h using the previous Laplacian matrix
- Estimating a spatio-temporal version of equation (1)
- Choosing the best performing parameters (KNN, h, etc.) to make predictions

Spatio-Temporal KDE

- For a certain one hour period (of the week) we can predict crime using KDE. The training data for this learning task are the observations from the same hour (and same day of the week) of the last M weeks.
- M is a number of weeks chosen a priori and is a sliding window which takes into account seasonality (or periodicity). Large vs. Small.
- Let $T_u = \{u - 168m : m \in \{1, ...M\}\}$, be the training data, then:

$$f_u(x) = \frac{1}{\sum_{t \in T_u} n_t} \sum_{t \in T_u} \sum_{i=1}^{n_t} k(x, s_t, i|H) \tag{1}$$

- is the Spatial-temporal KDE for any $x \in S$.

Choosing a point cloud

- Which points? Any point from M until the predicted period.
- How many points? There is a trade off, more points is computationally more costly.
- Points or mesh? Using a mesh would reduce the resolution but could achieve higher computation speed.
- Global or local? Instead of choosing from the whole spatial domain, we could choose from certain regions, this can achieve computational and accuracy advantages.

Choosing a point cloud

- For the purpose of this paper, we will sample 1000 points as the point cloud for each component. We will denote the set of points in the point cloud: $\{z_i\}$ for $i \in \{1, ..., Z\}$
- A component a region discretized from the spatial domain for each 1-hour period. (More on this later)

Components (digression)

- Components are defined as a discretization of the spatial domain.
- The discretization is made through a clustering algorithm. (Zhou & Matteson, 2016) Find that the K-means algorithm using euclidian distance is the more appropriate one for this task.
- The clustering is done based on the labeled data and it is important to make sure that the clusters are not too small because otherwise cross-validation is not possible.

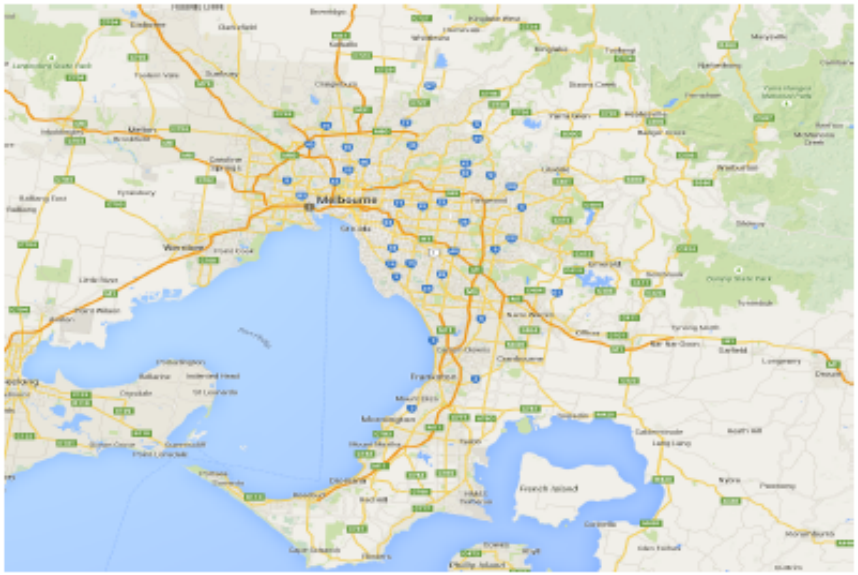
Constructing the adjacency graph

- We construct a graph with node at each point cloud and edges connecting points that are close.
- It is possible to represent this graph using a symmetric positive semidefinite adjacency matrix A.
- Which point to connect? If we had knowledge about the spatial domain and how crimes interact we could connect it with priors, however this information is costly and nodes can be connected using the Nearest Neighbor (NN) algorithm (Not symmetric).
- To perform this algorithm, we must choose an n. n should be big enough to ensure that the point cloud is sufficiently connected but small enough to emphasize local relationships.

Constructing the adjacency graph

- Weighted edges? in the simplest case it is possible to set $A_{ij} = 1$ if nodes z_i and z_j are connected and 0 otherwise. Another possible idea suggested in (Belkin and Niyogi) is to define weighted edges depending on the distance between points ($A_{ij} = exp(-||z_i - z_j||^2/r)$).
- However the choice of r is not clear, (Zhou & Matteson, 2016) propose choosing r by fitting an exponential distribution on all distances between connected nodes.
- For the purpose of this paper, we choose 5 NN and binary weights to construct A.

Caso Melbourne



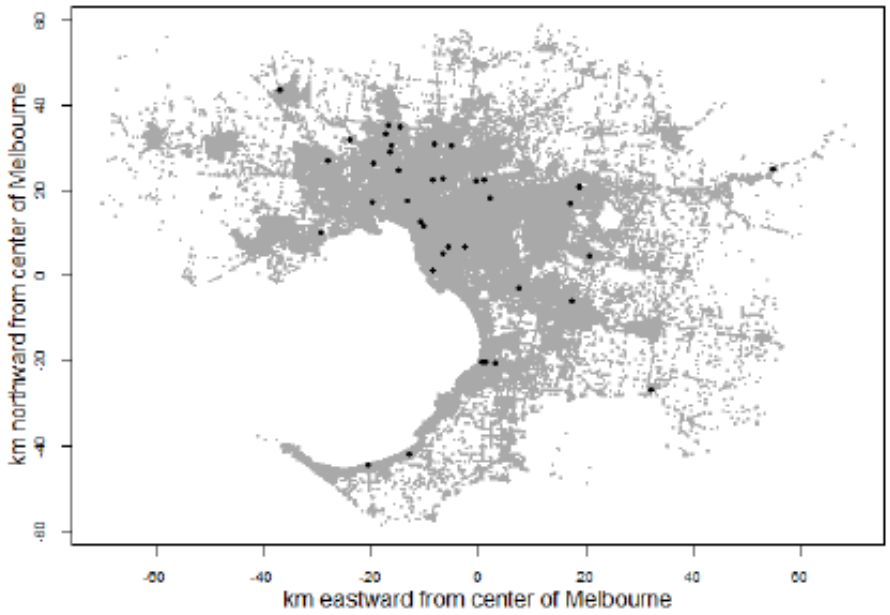


Fig 1: Left: map of Melbourne (Google Maps, 2015); right: spatial locations of all 696,975 Melbourne ambulance demand incidents from years 2011 - 2012 (in gray), and 38 demand incidents for a typical 1-hour period (in black). We observe complex boundary and geographical features (e.g., highways, roads, satellite suburbs).

Caso Melbourne

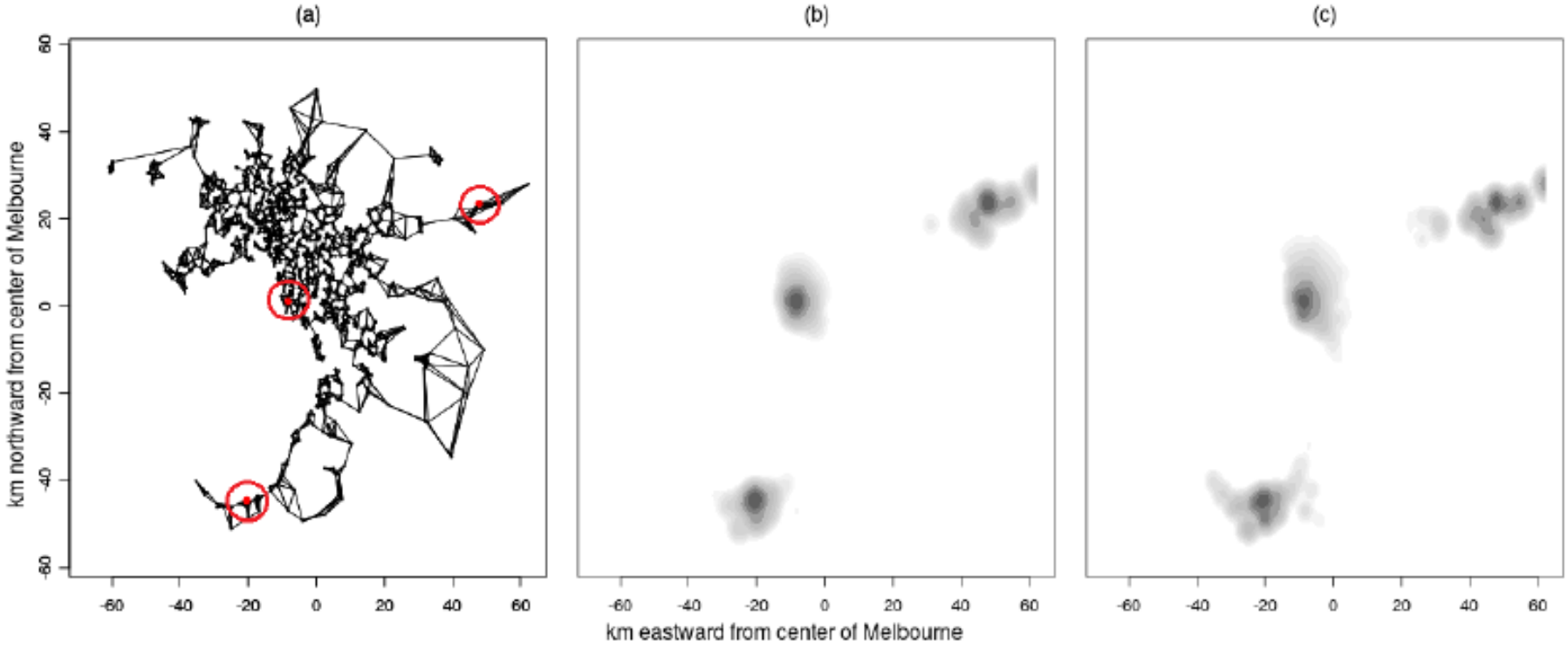


Fig 2: Examples of kernel warping: (a) the adjacency graph of a sample point cloud of size 1000; three observations are highlighted in red; (b) and (c), warped kernels centered at the these three observations with degrees of deformation $\lambda = 0.5$ and 2, respectively.

Constructing the Laplacian Matrix

- The graph Laplacian matrix is defined to be $L = D - A$ where D is the diagonal degree matrix with its diagonal being the column (or equivalently, the row) sum of A:

$$D_{ii} = \sum_j A_{ij}$$

- L is a symmetric positive semidefinite matrix. if the graph has multiple connected components, L can be rearranged into a block diagonal matrix, where each block is the respective Laplacian matrix for the component.

Warping the kernels

- We warp the kernel in equation (1) to the point cloud to generate a new warped kernel $\tilde{k}$.
- Let $k_x = [k(x, z_1|H), ..., k(x, z_Z|H]$ and $k_s = [k(s, z_1|H), ..., k(s, z_Z|H)]$ be vectors of kernels evaluated at x or s and the point cloud data $\{z_i\}$.
- Matrix $K = [k(z_i, z_j|H)]_{i,j\in\{1,...,Z\}}$ is a symmetric matrix of kernels evaluated at all pairs of point cloud data and I is a Z by Z identity matrix.
- Then for any $x \in S$ and any s in the labeled data:

$$\tilde{k}(x, s|H) = k(x, s|H) - k_x^T (I + \lambda K)^{-1} \lambda L k_s \tag{2}$$

Warping the kernels

- The parameter $\lambda > 0$ represents the degree of deformation. When $\lambda = 0$, we have $\tilde{k} = k$ and when $\lambda \to \infty$, $\tilde{k}$ approaches a positive constant on the point cloud.
- We replace the regular Gaussian kernel k in equation (1) with the new warped kernel $\tilde{k}$ defined in (2) to predict the density of crime demand f_u at a future time period u.
- We estimate the Gaussian kernel bandwidth H and the degree of deformation λ through cross-validation.

Caso Melbourne

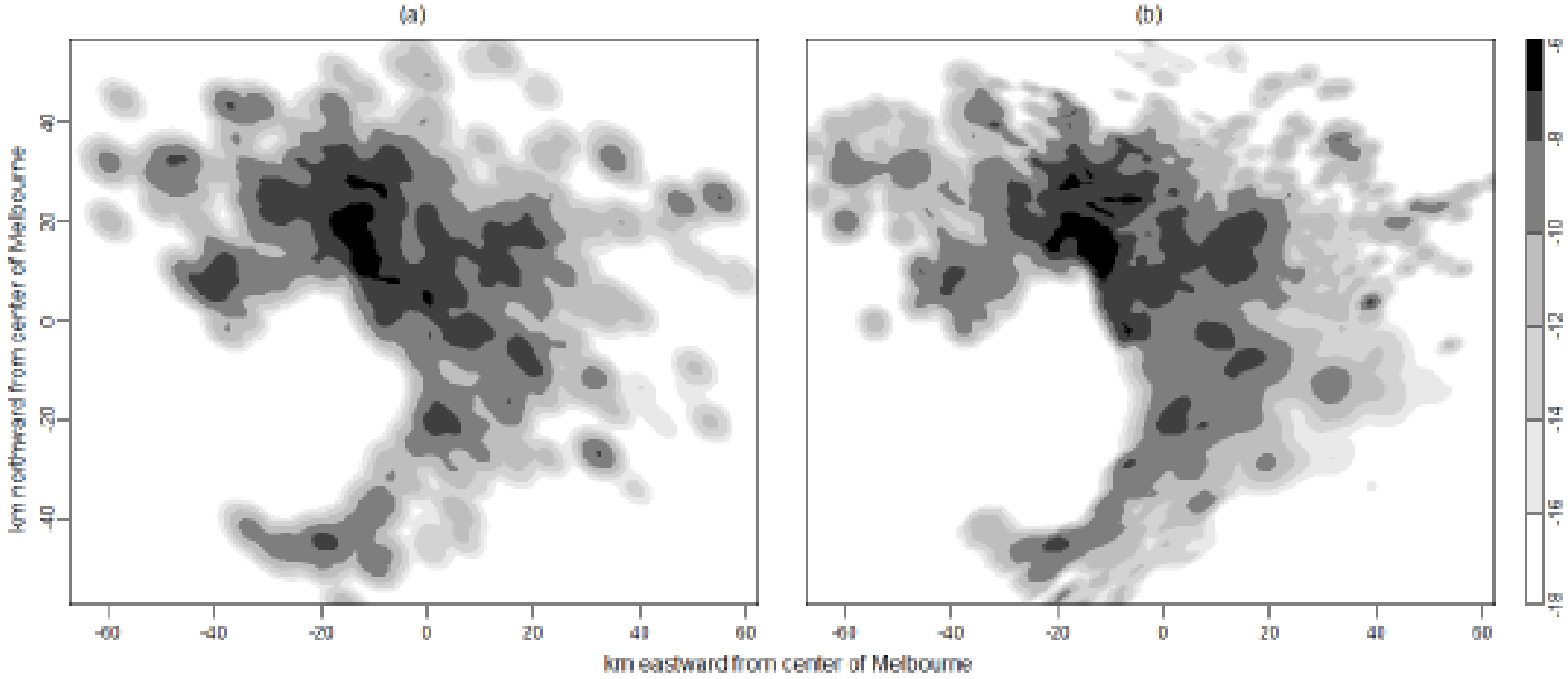


Fig 3: Log predictive densities using spatio-temporal kernel warping for March 2, 2011 (Wednesday) at (a) 2 - 3 am (night), and (b) 2 - 3 pm (day). For time period (a), we have sparse data and cross-validate to choose 1 spatial component. For time period (b), we have more data and choose 5 spatial components

Current challanges

- Dealing with considerably less data, this might make the cross-validation difficult.
- Dealing with optimal programming methods (paradigms).
- Choosing the right comparison metrics to evaluate the extent of this methodology.

References I

References

Barreras, F., Diaz, C., Riascos, A., & Ribero, M. (2016, November). *Una comparacion de diferentes modelos para la prediccion del crimen en Bogota.* Universidad de los Andes.

Kang, H.-W., & Kang, H.-B. (2017, April). Prediction of crime occurrence from multi-modal data using deep learning. *PLOS ONE*, *12*(4), e0176244. Retrieved 2017-09-04, from http://dx.plos.org/10.1371/journal.pone.0176244 doi: 10.1371/journal.pone.0176244

Kianmehr, K., & Alhajj, R. (2008, May). EFFECTIVENESS OF SUPPORT VECTOR MACHINE FOR CRIME HOT-SPOTS PREDICTION. *Applied Artificial Intelligence*, *22*(5), 433–458. Retrieved 2017-09-04, from http://www.tandfonline.com/doi/abs/10.1080/08839510802028405 doi: 10.1080/08839510802028405

References II

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011, March). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, *106*(493), 100–108. Retrieved 2017-09-04, from `http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.ap09546` doi: 10.1198/jasa.2011.ap09546

Zhou, Z., & Matteson, D. S. (2016). Predicting Melbourne ambulance demand using kernel warping. *The Annals of Applied Statistics*, *10*(4), 1977–1996. Retrieved 2017-09-08, from `http://projecteuclid.org/euclid.aoas/1483606848`