# A Manifold Learning Data Enrichment Methodology for Homicide Prediction

Juan S. Moreno Pabón* Mateo Dulce Rubio* Yor Castaño* Alvaro J. Riascos*† Paula Rodríguez Díaz*

*Quantil
†Universidad de los Andes

October 2020

***Abstract*—Not all types of crime have the same priority in the agendas of policymakers since society tends to be more reluctant to more violent and costly crimes such as homicide. However, relative to other types of crime, homicides are statistically more challenging due to its sparsity and low frequency. For instance, over the last five years the average number of homicides across the city of Bogota has been roughly a thousand events per year, compared to the more than one hundred thousand robberies reported in the same period. Nevertheless, more than 80% of the homicides in the city occur during street fights suggesting a strong spatial and temporal correlation between these two types of crime. With this in mind, we used a manifold learning approach that capitalizes on a rich dataset of street fights to discover a criminal manifold that we use to penalize a KDE model of homicides where sparsity and low frequency is an issue. To implement this we follow a Kernel Warping methodology (Zhou & Matteson, 2015). The methodology reduces the relevant space for homicide prediction to regions of the city where homicides or street fights have occurred, giving more weight to the homicide episodes. We also introduce a temporal decay component to place a larger importance to recent events. The proposed model outperforms a standard KDE trained with homicide data, a KDE trained in both homicide and street fights data for homicide prediction, and a standard self-exciting point process on homicide data: flagging just the 5% of the area of the city with the highest estimated density, the Kernel Warping model correctly identifies between 30% and 35% of the homicides in the test set.**[1]

## I. Introduction

Predicting crime is extremely desired by police departments to allocate law enforcement resources more efficiently. Given that society tends to be more reluctant to more violent and costly crimes, such as homicide, not all types of crime have the same priority in agendas of policymakers. For instance, [1] estimated that the average cost per murder exceeds $17.25 million in the United States and discussed how a prediction tool could be used to assist homicide prevention. In Bogota, the capital city of Colombia, the homicide rate has been trending downwards with roughly 14 murders per 100.000 inhabitants in 2019. However, this rate is considerably higher than the international average, and it is still relevant to develop predictive models that assist law enforcement agents in their work.

[1]Results of the project "Diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en Bogotá" funded by Colciencias with resources from the Sistema General de Regalías, BPIN 2016000100036. The opinions expressed are solely those of the authors.

Previous models like [2], with a further extension to explicitly predict homicides [3], successfully model crime through self-exciting point processes and are considered the state-of-the-art of crime prediction models. However, homicide occurrences in Bogota present two main challenges for statistical analysis and predictive modeling that make self-exciting point processes not suitable for them: sparse spatial distribution and low frequency. For instance, over the last five years, the average number of homicides has been roughly a thousand events per year, in contrast to the more than one hundred thousand robberies reported in the same period. Nonetheless, more than 80% of the homicides occur during street fights suggesting a strong spatial and temporal correlation between these two types of crime. The proposed methodology addresses these challenges through a Manifold Learning approach that capitalizes on a rich dataset of street fights to discover a criminal manifold that we use to penalize a KDE model for homicide prediction.

The Manifold Learning approach assumes that the homicide data lives on a low-dimensional subspace of the sample space, to reduce the region of interest to a subset of Bogota were homicides are likely to occur. The methodology can be seen as a semi-supervised learning technique in which the labeled and the unlabeled data (homicides and street fights) are used to estimate the underlying lower-dimensional space where the data lives, and the labeled data (homicides) is then used to estimate the decision boundary over this subspace [4].

Specifically, we follow [5] to capitalize on a rich dataset of street fights, highly correlated to homicides, to discover a criminal manifold that is later used to penalize a KDE model of homicides occurrences. The methodology deforms the estimation made by the KDE to take into account the regions where street fights occurred while highlighting the homicide occurrences. In addition, we introduce a temporal decay component to place higher importance on recent events. The proposed model outperforms a standard KDE model trained with homicide data, a KDE trained in both homicide and street fights data for homicide prediction, and a standard self-exciting point process on homicide data.

The article is organized as follows. Section II presents the Kernel Warping methodology for homicide prediction, Section III explains the data used to train and test the models, Section IV states the main results, and Section V discusses the findings.

## II. Methodology

We use the Kernel Warping methodology used in [5] to predict ambulance demand in Melbourne, to predict homicides in Bogota that are not only scattered but also rare. The Kernel Warping methodology is used in this work to enrich the homicide data using street fights, a criminal event highly correlated to homicide occurrence.

In the first place, the Kernel Density Estimation (KDE) model which places a non-negative kernel function at each one of the points in a dataset $\mathcal{D}$ and adds them to produce an estimation of its empirical distribution. Typically, the kernel function is a Gaussian distribution with smoothing parameter $\sigma$ known as the bandwidth of the kernel. Then, for a point $x$ in the relevant domain, the model predicts the following intensity:

$$f(x) = \frac{1}{|\mathcal{D}|} \sum_i k_\sigma(x, x_i). \tag{1}$$

The Kernel Warping methodology assumes that the homicide data lives in a manifold embedded in $\mathbb{R}^2$. Given than 80% of the homicide events occur during street fights, we assume that the aforementioned manifold is defined by the location of the historical occurrences of street fights in the city. Thus, to estimate the criminal manifold, we use a point cloud $\mathcal{Z} = \{z_i\}_i$ of historic homicides and street fight events in the training set. These events give an approximation of the manifold we want to estimate to reduce the relevant area for homicide prediction to the spatial region where homicides are likely to occur.

We use the adjacency graph defined on the point cloud data as an empirical discrete approximation of the manifold. Specifically, we define the adjacency graph matrix $A$ with entries $a_{ij} = 1$ if the event $z_i$ from the point cloud data $\mathcal{Z}$ is amongst the $n-$nearest neighbors of $z_j$, or vice versa, and $a_{ij} = 0$ otherwise. Moreover, we define the set $\mathcal{S}$ as the historic homicides in the train set. In a semi-supervised fashion, the set $\mathcal{S}$ represents the label data we want to predict while the point cloud $\mathcal{Z}$ includes both the labeled and the unlabeled data from which we aim to learn the region where labeled data lives.

For instance, Figure 1 plots the adjacency graph constructed for the events between March and August of 2019. Point cloud data (homicides and streets fights) corresponds to the blue points, while labeled data (homicides) are the red points.

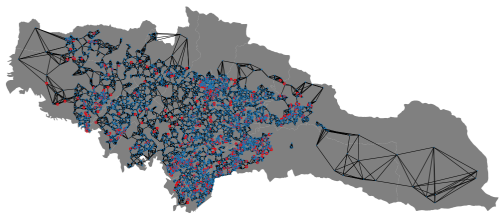


Fig. 1. Adjacency graph of street fights (blue) and homicides (red).

We then construct the graph Laplacian matrix $L = D - A$ from the adjacency graph matrix $A$ and the diagonal degree matrix $D$, with its diagonal entries equal to the row sum of $A$. The graph Laplacian matrix $L$ gives an empirical discrete approximation of the Laplace-Beltrami operator on the manifold and penalizes differences between adjacent nodes [4], [5].

Finally, from a standard kernel function $k_\sigma(\cdot, \cdot)$, we construct a warped kernel $\tilde{k}$ following [5], [6] towards the point cloud data:

$$\tilde{k}_\sigma(x, s) = k_\sigma(x, s) - k_{xz}^T (I + \lambda K_{zz})^{-1} \lambda L k_{sz}, \tag{2}$$

where $k_{xz} = [k_\sigma(x, z_1), \ldots, k_\sigma(x, z_Z)]$ and $k_{sz} = [k_\sigma(s, z_1), \ldots, k_\sigma(s, z_Z)]$ are vectors of kernels evaluated at $x$ and $s$, respectively, with respect to the point cloud data $\mathcal{Z}$. The matrix $K_{zz} = [k_\sigma(z_i, z_j)]_{i,j}$ is a symmetric matrix of kernels evaluated at all pairs of points of the cloud data, and $I$ is a $Z \times Z$ identity matrix. Finally, $\lambda$ accounts for the degree of deformation: if $\lambda = 0$ then $\tilde{k} = k$, while $\lambda \to \infty$ implies $\tilde{k}$ approaches a positive constant on the point cloud.

The warped kernel in equation (2) is computed for every $x$ in Bogota and each $s \in S$. Finally, the warped kernel (2) is replaced in equation (1) to predict the expected crime intensity for a given point $x$ in the city:

$$\tilde{f}(x) = \frac{1}{|\mathcal{S}|} \sum_i \tilde{k}_\sigma(x, s_i). \tag{3}$$

Extending [5], [6], we introduce a temporal decay component to place a larger weight to more recent events. This time decay seeks to account for the hotspots' dynamics and the displacement of crime. Specifically, we use an exponential decay with parameter $\omega$ that controls the reduced weight imposed on older events. We then multiply this spatial decay to the kernel warped to the spatial region of interest defined by the point cloud. Then, we obtain a Kernel Warping estimation with temporal decay for a given point $x$ in Bogota at a time period after the end of the training set:

$$\hat{f}(x) = \frac{1}{|\mathcal{S}|} \sum_i \exp(-\omega(t_x - t_{s_i})) * \tilde{k}_\sigma(x, s_i). \tag{4}$$

## III. Data

The data was provided by the Security Office of Bogota and contains criminal records from 2010 to 2020, with information about the type of the crime, georreferenced location of occurrence, time stamp, among others. The model was estimated using 6-fold cross validation with each training set consisting of the crime events within a six months period. The model was then evaluated using the homicides occurring in the month following the training set. The first training set consisted of the events between January and June of 2019 and the first test set was the homicides reported in July of 2019. The following training and respective test sets were defined as the 4 weeks lag ahead from the previous training-test sets.

For each training set, the point cloud $\mathcal{Z}$ was defined as both the homicide and the street fights and the labeled data $\mathcal{S}$ as the homicide episodes only. On average, each point cloud consists of 9,312 (std = 364.13) homicide and street fight records, each labeled data set of 416 (std = 19.33) homicides, and each test set of 68 (std = 9.98) homicide observations.

## IV. Results

The cross validation procedure was used to select the parameters needed to train the model, as well as to asses the variability of the predictive capacity of the methodology. Specifically, the number of neighbors $n$ used to construct the adjacency graph (and therefore the Laplacian matrix), the bandwidth of the kernels $\sigma$ and the parameter of deformation $\lambda$ were selected to maximize the average predictive power of the model on the test sets using grid search and the cross validation procedure described above.[2]

To evaluate the predictive capacity of the models we use a discrete grid of Bogota with ∼33,000 cells of 54 by 54 meters. As validation metric we used the Hit Rate measure, which captures the portion of homicides of the test set that occurred in regions flagged by the model as hotspots:

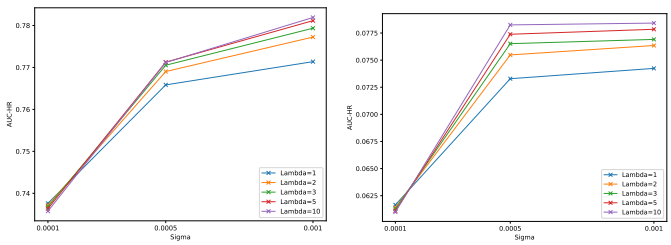
$$\text{Hit Rate} = \frac{\text{\# of homicides in hotspots}}{\text{\# of homicides}}. \tag{5}$$

In detail, for each fold and training set, we trained the model and evaluate it using equation 3 at the coordinates of the centroids of each one of the grid cells to construct a homicide intensity over the city. The top $\mathrm{x}\%$ of the grid cells are flagged as hotspots and the Hit Rate of the model is computed as the portion of homicides in the test set that occurs in any of these hot cells. Furthermore, we vary the percentage of the area of Bogota flagged by the model as hotspots and produce Hit Rate (HR) vs. Percentage of Area Covered (PAC) curves. The area under the HR-PAC curve gives a metric of the global predictive performance of the model.

Fixing the number of neighbors $n = 7$ (which was the number of neighbors that maximized the average area under the HR-PAC curve[3]), Figure 2 presents the average area under the HR-PAC curve for the cross validation trained models for different values of $\sigma$ (bandwidth kernel) and $\lambda$ (warping deformation parameter). Furthermore, we compute the HR-PAC AUC for $20\%$ of the grid cells flagged as hotspots. The motivation behind this metric is the practical use of the model: in a city as large as Bogota with limited resources, police cannot cover large areas of the city permanently, which is why a model with a more concave curve in the initial part might be more valuable relative to one with a larger total area.

We found that increasing the deformation parameter $\lambda$ improved the model performance under all of the validation metrics used. This validates our approach as it suggest that the more the original KDE is deformed and enriched with the manifold defined by the street fighting data the better the model perform. Further, the bandwidth parameter $\sigma = 0.001$ consistently gives the better predictive power. These results remain the same using the HR-PAC AUC for 5 or 10 percent of the area.

Additionally, we fix the number of neighbors $n = 7$, the kernel bandwidth $\sigma = 0.001$ and the deformation parameter $\lambda = 10$, and fine tuned the temporal decay parameter $\omega$. The cross validation procedure is analogous to the one used for the rest of parameters using equation (4) to train and test the model. The temporal distance of two events was measured as the number of months between them. We tested a range of $\omega$ values from 0 to 0.2 and found that a temporal decay $\omega = 0.1$ slightly improves (around 0.01) the predictive capacity of the kernel warping approach relative to a case of $\omega = 0$ (equivalent to having no temporal decay), suggesting some temporal dynamics that should be included into the model.




(a) Total HR-PAC AUC. (b) HR-PAC AUC, PAC $\leq 20\%$

Fig. 2. Average area under the Hit Rate - Percentage of Area Covered as hotspots curve for homicide prediction. 6-fold cross validation varying $\sigma$ and $\lambda$ parameters.

Lastly, we compare the predictive accuracy of the kernel warping approach with and without the temporal decay component, and against standard KDE models using the homicide events (labeled data) and using both homicide and street fights data (point cloud).[4] Figure 4 plots the predicted intensity maps for Bogota for the training set of crime events between March and August of 2019. The actual homicides reported during September in Bogota (60) are presented as white points. Furthermore, we compare our approach against a state-of-the-art self-exciting point process on homicide data [2]. Figure 3 presents the Average HR - PAC curves for the 6-fold cross validation for the different train models: KDE with homicide data, KDE with both homicide and street fights data, Kernel Warping, Kernel Warping with temporal decay, and Self-Exciting Point Process with homicide data.

The kernel warping proposed model outperforms the competing models for homicide prediction. This is particularly relevant for the first portion of the HR-PAC curve as it is likely the actual covering capacity of Bogota police department. For instance, flagging the 5% of the area of the city with the highest estimated density, the kernel warping model correctly identifies between 30% and 35% of the homicides in the test set. Furthermore, including a temporal decay slightly improves the predictive capacity of the model.

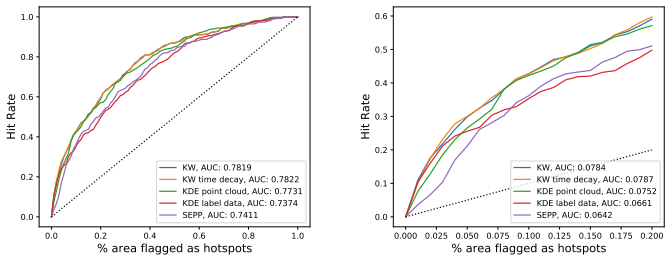
## V. Discussion

We implement a manifold learning approach that capitalizes on a rich dataset of street fights to discover a criminal manifold that we use to penalize a KDE model of homicides where sparsity and low frequency are challenging issues. We follow a Kernel Warping methodology that reduces the relevant space

[2]We used binary weights and nearest neighbors to construct the adjacency graph matrix in all the experiments.

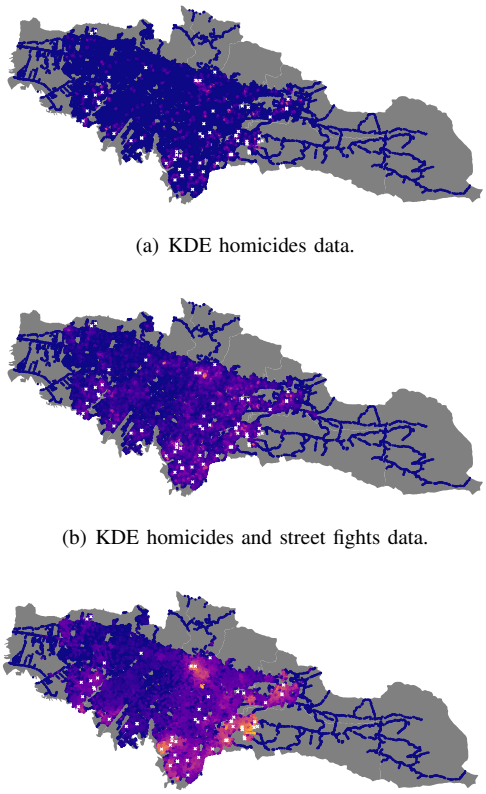
[3]Results not shown.

[4]For these KDEs we used the same optimal bandwidth $\sigma = 0.001$.




(a) Total HR-PAC curve. (b) HR-PAC curve, PAC $\leq 20\%$.

Fig. 3. Average Hit Rate - Percentage of Area Covered as hotspots curves for homicide prediction.



(a) KDE homicides data.

(b) KDE homicides and street fights data.

(c) KW with time decay homicides and street fights data.

Fig. 4. Predicted intensity maps by different models. Actual homicides reported in September of 2019 in white.

for homicide prediction to regions of the city where homicides or street fights have occurred, giving more weight to the homicide episodes. The proposed model outperforms a standard KDE trained with homicide data, a KDE trained in both homicide and street fights data and a standard self-exciting point process on homicide data. We further introduce a temporal decay component to place a larger importance to recent events slightly improving the predictive capacity of the methodology. It should be noted that a standard KDE model that uses both homicide and street fighting events for homicide prediction outperforms a KDE trained with just homicide events. This shows the relevance of enriching the training set with other highly correlated events in order to overcome the challenging sparsity and low frequent nature of homicide events.

The proposed model seeks to assist policymakers and law enforcement agencies in the allocation of scarce resources intended to prevent homicide events. Since homicide is one of the most violent and costly crimes, developing robust predictive models will likely impact in a positive way society as a whole. In particular and to a greater extent, it will benefit those that are more vulnerable to suffer directly from this fatal type of crime. However, crime prediction models suffer from several sources of bias. For instance, they operate in partial feedback or bandit settings where the decisions made by the algorithms affect the data collected which is used to retrain the models [7]. Moreover, the data used to train these models frequently present a selection bias product of historic patrolling patterns of law enforcement agents [8]. When these two issues are present together, predictive patrolling models can reproduce and accentuate biases against some populations leading to over and/or under patrolling these communities. This is called in the literature as runaway feedback loops [9].

Homicide data in Bogota is checked weekly between different public entities to generate high-quality homicide data with low reporting or selection biases. However, our methodology capitalizes on enriching this high-quality dataset with other types of crime that likely suffer from these biases. Further research must be done to quantify and address these potential biases of the proposed model for homicide prediction.

## References


[1] M. DeLisi, A. Kosloski, M. Sween, E. Hachmeister, M. Moore, and A. Drury, "Murder by numbers: Monetary costs imposed by a sample of homicide offenders," *The Journal of Forensic Psychiatry Psychology*, pp. 501–513, 2010.

[2] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011.

[3] G. Mohler, "Marked point process hotspot maps for homicide and gun crime prediction in chicago," *International Journal of Forecasting*, vol. 30, no. 3, pp. 491–497, 2014.

[4] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[5] Z. Zhou and D. S. Matteson, "Predicting Melbourne Ambulance Demand using Kernel Warping," *arXiv:1507.00363 [stat]*, July 2015. arXiv: 1507.00363.

[6] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 824–831, 2005.

[7] S. Kannan, J. H. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu, "A smoothed analysis of the greedy algorithm for the linear contextual bandit problem," in *Advances in Neural Information Processing Systems*, pp. 2227–2236, 2018.

[8] K. Lum and W. Isaac, "To predict and serve?," *Significance*, vol. 13, no. 5, pp. 14–19, 2016.

[9] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Runaway feedback loops in predictive policing," *arXiv preprint arXiv:1706.09847*, 2017.