Misceláneo

doi.org/10.17141/urvio.34.2022.5395

URVIO, Revista Latinoamericana de Estudios de Seguridad, No. 34, septiembre-diciembre 2022
pp.83-98 • RELASEDOR y FLACSO Sede Ecuador • ISSN 1390-4299 (en línea) y 1390-3691

Prediciendo el crimen en ciudades intermedias: un modelo de "machine learning" en Bucaramanga, Colombia[1]

Predicting Crime in Middle-Size Cities. A Machine Learning Model in Bucaramanga, Colombia

Juan-David Gelvez-Ferreira[2], María-Paula Nieto-Rodríguez[3] y Carlos-Andrés Rocha-Ruiz[4]

Recibido: 28 de marzo de 2022
Aceptado: 1 de junio de 2022
Publicado: 30 de septiembre de 2022

Resumen

El uso de tecnología para prevenir el crimen es una práctica cada vez más frecuente. Sin embargo, la evidencia se ha concentrado en ciudades principales, que cuentan con gran cantidad de datos y mejores capacidades locales. El objetivo de esta investigación es presentar los resultados de un modelo de "machine learning" para predecir el delito en Bucaramanga, una ciudad intermedia de Colombia. Se utilizó el procesamiento de señales para grafos y una adaptación al caso del modelo de vectorización de texto TF-IDF. Se identificó que los mejores resultados en la predicción del crimen se dieron con modelos espaciales de grafos por semanas. Además, encontramos evidencia de que existen diversas dificultades de predicción, en dependencia de la periodicidad del modelo. La mejor opción posible (con los datos disponibles) es una periodicidad semanal. El mejor modelo encontrado es un KNN de clasificación, que alcanza un 59 % de exhaustividad(*recall*) y más de 60 % de exactitud (*accuracy*.). *Concluimosque* los modelos de predicción del delito constituyen una herramienta útil para construir estrategias de prevención en ciudades principales; sin embargo, existen limitaciones para su aplicación en ciudades intermedias, que cuentan con poca información.

Palabras clave: análisis de datos; Colombia; crimen; prevención del crimen; Policía

Abstract

The use of technology to prevent and respond to citizen security challenges is increasingly frequent. However, empirical evidence has been concentrated in major cities with large amounts of data and local authorities' strong capacities. Therefore, this investigation aims to capture a series of policy recommendations based on a machine learning crime prediction model in an intermediate city in Colombia, Bucaramanga (department

1 Agradecemos los comentarios y aportes en la investigación de Eduardo Escobar Gutiérrez, Juan Sebastián Franco, Lina María González, Pablo Montenegro Helfer, Michael Weintraub, dos evaluadores anónimos y a los editores de la revista *URVIO*. Los errores que persistan son responsabilidad de los autores.

2 University of Maryland, Estados Unidos, jgelvez@umd.edu, orcid.org/0000-0001-7526-8384

3 Departamento Nacional de Planeación, Colombia, mp.nieto@uniandes.edu.co, orcid.org/0000-0002-6272-5627

4 Departamento Nacional de Planeación, Colombia, candres.rocharuiz@gmail.com, orcid.org/0000-0001-8635-8537

of Santander). The model used signal processing for graphs and an adaptation of the TF-IDF text vectorization model to the space-time case, for each of the cities' neighborhoods. The results show that the best crime prediction outcomes were obtained when using the models with spatial relationships of graphs by weeks. Evidence of the difficulty in predictions based on the periodicity of the model is found. The best possible prediction (with available data) is weekly prediction. In addition, the best model found was a KNN classification model, reaching 59 % of recall and more than 60 % of accuracy. We concluded that crime prediction models are a helpful tool for constructing prevention strategies in major cities; however, there are limitations to its application in intermediate cities and rural areas in Colombia, which have little statistical information and few technical capabilities.

Keywords: crime; crime prevention; Colombia; data analysis; police

Introducción

¿Qué tan eficiente es predecir el crimen en ciudades intermedias?[5] ¿Debería la política pública concentrase en herramientas de predicción del delito? Esta investigación responde ambas preguntas usando un modelo de predicción del crimen en Bucaramanga, Colombia. El modelo de "machine learning" que presentamos usa señales para grafos y una adaptación al caso del modelo de vectorización de texto TF-IDF, el cual busca predecir el crimen. Además, presentamos recomendaciones de política pública orientadas a la implementación de modelos de predicción del delito para ciudades intermedias, considerando los sesgos de los datos, así como otras alternativas para prevenir el crimen.

La literatura académica ha mostrado que los asuntos de seguridad ciudadana deben ser atendidos a través de estrategias costo-efectivas, pertinentes y basadas en la evidencia empírica (DNP y CESED 2020). Durante los últimos años, el Gobierno colombiano avanzó en la construcción de métodos de análisis delictivo, la optimización de la investigación criminal, los análisis y la simplificación de audiencias judiciales, entre otras políticas (ver, por ejemplo, Collazos et al. 2021; Mejía et al. 2021). El siguiente paso, de acuerdo con lo que afirman diferentes autoridades, es crear herramientas para predecir el delito (El Tiempo 2021; Alcaldía Mayor de Bogotá 2019).

Sin embargo, poco se ha avanzado en la construcción de modelos de predicción delictiva en América Latina, especialmente en ciudades pequeñas o intermedias. El modelo que presentamos en este artículo constituye una de las primeras aproximaciones en la región, basada en estimar si en una semana determinada ocurrirán delitos en cada sección territorial de la ciudad. Se implementó en una ciudad intermedia como Bucaramanga dado que en Colombia no existen modelos predictivos de crimen en ciudades intermedias, y se desconoce su eficiencia.

Los resultados de la investigación, como se verá más adelante, no son concluyentes. Ciudades con información estadística limitada, como Bucaramanga, pueden obtener modelos de predicción limitados; así que otro tipo de herramientas para la atención del crimen pueden ser más eficientes. Por tanto, existen al menos dos alternativas para realizar modelos de predicción en ciudades intermedias. Primero, aumentar la capacidad analítica y de datos de una ciudad; y/o segundo, implementar

5 Entiéndase como ciudades intermedias "aquellas que tienen entre 50 000 y un millón de habitantes" (Organización Mundial de Ciudades y Gobiernos Locales Unidos s.f.). Las ciudades intermedias, más allá de su importancia demográfica, se identifican por las funciones que desempeñan en la mediación de flujos entre las zonas rurales y los territorios urbanos (Instituto de Estudios Urbanos-Universidad Nacional de Colombia 2016).

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

otro modelo de predicción. La investigación revela que pueden existir sesgos y discriminación de los modelos predictivos, por lo tanto, se invita a considerar estos antes de su implementación.

La estructura del artículo es la siguiente. Tras esta introducción, se presenta la revisión de literatura sobre los modelos de predicción del delito. Luego, la sección de metodología, que describe las variables descriptivas de los datos usados, así como la explicación del modelo predictivo. Por último, se exponen los resultados y las limitaciones del modelo, y recomendaciones de política pública.

Soporte teórico

A lo largo de los años, las estrategias enfocadas en la prevención del delito han demostrado tener óptimos resultados en reducir la criminalidad (Gélvez 2018; Wright y Beaver 2012). A su vez, existen diferentes casos que muestran cómo la efectividad de la prevención del delito se maximiza cuando las instituciones policiales concentran sus recursos en pequeñas unidades geográficas (Weisburd y Telep 2014; UNODC 2010). Por ejemplo, en Bogotá, Colombia, se encontró que al duplicar los tiempos de patrullaje en zonas determinadas se reducen los niveles de delincuencia, y al combinar tiempos extra de patrullaje con una mayor provisión de otros servicios públicos esa reducción es aún mayor (Blattman et al. 2017).

El éxito de las estrategias enfocadas en lugares se basa en la premisa de que existen puntos calientes de crimen (Brantingham, Brantingham y Taylor 2005), y es allí donde deben aumentar los riesgos de delinquir (Cornish y Clarke 2003). Un ejemplo es el uso de cámaras de vigilancia en Medellín, Colombia, donde se redujeron los delitos contra la vida y la propiedad respecto al año anterior a la instalación, al igual que una disminución en el número de arrestos (Gómez et al. 2019).

Esta política ha sido utilizada con frecuencia por su costo-efectividad y por ofrecer la oportunidad de usar el pie de fuerza disponible de manera que se obtengan mayores impactos en los índices de criminalidad (Braga et al. 2014; Abt et al. 2019). Sin embargo, persiste el debate sobre los efectos secundarios de estas iniciativas, como el desplazamiento del crimen a zonas aledañas (Braga et al. 2014; Johnson et al. 2014). La experiencia latinoamericana se puede observar en las ciudades como Medellín (Collazos et al. 2020), Cali (Blair y Weintraub 2020) y Bogotá (Blattman et al. 2017) con resultados diferenciados. En Cali y Bogotá se encontraron resultados positivos reflejados en la reducción en la ocurrencia de delitos, no obstante, en Medellín no se halló ningún impacto.

El uso de fuentes de datos y registros estadísticos para la predicción del delito es un método que cobra fuerza en la academia y en los tomadores de decisión. Por ejemplo, en los Centros de Apoyo a Decisiones Estratégicas, implementados en Chicago, Estados Unidos, se utiliza la inteligencia de datos junto con las capacidades e inteligencia humana para identificar los problemas prioritarios de criminalidad (Abt et al. 2019). Con base en esto, se desarrollan y evalúan con regularidad las estrategias para focalizar la atención de los recursos policiales, usando la tecnología para mejorar la eficiencia y rapidez en la atención de eventos de delincuencia y criminalidad. En la actualidad, el Departamento de Policía de Chicago realiza diversas pruebas piloto que buscan evaluar los resultados de esos centros (Hollywood et al. 2019).

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

El Departamento de Policía de Nueva York, Estados Unidos, también ha desarrollado varias pruebas con base en modelos de predicción del delito. Con un gran conjunto de bases de datos, *softwares* e infraestructura de apoyo, encontraron gran precisión en las predicciones de diferentes eventos criminales, en especial, de los eventos protagonizados por el uso de armas de fuego (Levine et al. 2017). Igualmente, en Los Ángeles y Kent se implementó otro modelo de predicción de delito llamado el Modelo de Secuencia de Réplicas de Tipo Epidémico (ETAS, por sus siglas en inglés), que calcula el riesgo de comportamiento delictivo en puntos críticos a largo plazo y los riesgos que se presentan con alguna regularidad en el corto plazo (Mohler et al. 2015). El crimen se redujo en la prueba del modelo gracias a los algoritmos predictivos (Mohler et al. 2015; Ridgeway 2018).

Sin embargo, otros estudios han presentado resultados no concluyentes de la implementación del tipo de sistemas que se analizan (Santos 2014; Hunt et al. 2014; Saunders et al. 2016). Parece que, entonces, los modelos de predicción del delito producen diversos resultados, de acuerdo con las particularidades de las modalidades de crimen en la zona y a las características contextuales. Así como lo mencionan Meijer y Wessels (2019), las estrategias con base en la predicción del delito pueden reducir de manera eficiente varios crímenes, pero no todos y, en consecuencia, cada autoridad de policía debe adecuarlos según sus propias necesidades.

En particular el modelo de grafos utilizado en esta investigación ha sido implementado para analizar datos de crimen, epidemiológicos, inventarios de bienes comerciales y redes de transporte (Shuman et al. 2013). Por ejemplo, el modelo espaciotemporal de grafos ponderados y una red neuronal generalizable de grafos estructurados se utilizaron conjuntamente para la predicción del delito asociando cada nodo de grafo con una serie temporal delictiva, con buenos resultados de predicción de delito en una escala temporal de horas (Wang et al. 2018). El uso de esos modelos se potenció gracias al uso del proceso Hawkes para la predicción del delito, en donde la precisión de sus resultados lo convirtieron en un *software* comercial utilizado hoy en día en diferentes lugares del mundo (Mohler et al. 2015). Otros estudios también han aplicado modelos espaciotemporales como el ST-ResNet, con los cuales se demuestra que puede predecirse el delito con reducciones pequeñas en la precisión (Wang et al. 2017).

Para el caso de Colombia se conoce de una sola aproximación a los modelos de predicción del delito (Riascos et al. 2020). Dicho estudio replica el modelo de Mohler et al. (2011) para Bogotá, el cual es piloteado desde noviembre de 2017, con apoyo de la Secretaría de Seguridad, Convivencia y Justicia de Bogotá, la Policía de Bogotá y Quantil S.A. En este modelo se destaca que en la predicción se pueden presentar limitaciones como el refuerzo del sesgo de retroalimentación y de discriminación en los resultados obtenidos. Por lo tanto, es esencial para la implementación de este tipo de modelos contar con información que abarque una gran cobertura y, en lo posible, especificidad geográfica. Allí se practicaron diferentes líneas de estudio como el espaciotemporal, la distribución de la infraestructura física, más el entorno ambiental y visual para complementar la información (Riascos et al. 2020). Sin embargo, los modelos de predicción no han sido diseñados en contextos con poca información disponible, por lo que la presente investigación se centra en una ciudad intermedia, la cual no cuenta con una estrategia de predicción diseñada.

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Metodología

La principal fuente de datos usada en esta investigación es el Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional (SIEDCO).[6] Esta base de datos cuenta con información de hechos delictivos -homicidios, hurtos a personas y lesiones personales- y con una descripción en cada una de las siguientes variables: modalidad del delito, departamento, municipio, fecha (mes, día, año), edad de la víctima, género de la víctima, zona urbana o rural, hora del día en el que ocurrió el hecho, código de identificación (DIVIPOLA), país de la víctima, latitud y longitud del hecho y número de casos ocurridos en el hecho. La información reporta datos desde el primero de enero de 2016 hasta el 31 de diciembre de 2019, distribuidos en 17 columnas y con 1 865 869 observaciones.

Las variables geográficas disponibles en esta fuente de datos fueron las de latitud y longitud, por lo que para obtener información georreferenciada desagregada por barrio o manzana se utilizarán los mapas (*shapefiles)* publicados en las páginas web de entidades públicas colombianas. De esta manera se desarrollaron los modelos de predicción del delito con resultados desagregados por barrio o manzana. Al revisar las categorías y la frecuencia de cada una de las variables (mes, delito, zona y género), no se encuentra que contengan valores atípicos y cada una contiene las categorías esperadas. Lo anterior, evidencia que los datos cuentan con las características necesarias para efectuar estudios detallados. Además, el desarrollo de análisis descriptivos de la frecuencia de las variables utilizadas en el modelo de predicción del delito a escala nacional permite identificar la existencia de diferencias estructurales en el comportamiento y las modalidades del delito según el mes, la hora, la zona y el género. En ese modelo se cuenta con la información completa de cuatro delitos: "hurto a personas", "lesiones personales", "violencia intrafamiliar" y "homicidio" en diferentes horas, días y meses del año, teniendo en cuenta la zona en la que sucedió el delito -sea urbana o rural- y el género de la víctima. Las particularidades de los datos permiten interpretar los resultados obtenidos del modelo de predicción en un nivel geográfico más limitado.

Selección del municipio y estadísticas descriptivas

Con el objetivo de desarrollar una prueba piloto de predicción del delito, se decidió seleccionar una entidad territorial que cumpliera con las siguientes condiciones: 1) que fuera uno de los municipios con mayor cantidad de hechos delictivos -homicidios, hurtos a personas y lesiones personales-, y por ello aportara significativamente a las tasas nacionales de criminalidad; 2) la información georreferenciada correcta, tanto de crimen como de malla vial; 3) que no use un modelo de predicción del delito o esté en proceso de construcción, como

6 SIEDCO se encuentra implementado desde el año 2003, teniendo a la fecha más de 26 millones de registros de delitos y actividad operativa. De ello, tres millones han sido ingresados de la fiscalía general de la Nación, con un promedio diario de 3500 registros correspondientes a delitos (Policía Nacional 2021; SEN 2021). De esta manera, este Sistema que consolida la información de manera anual, es reconocido, articulado y coordinado por el Sistema Estadístico Nacional (SEN), bajo normas, principios, políticas y procesos técnicos del Departamento Administrativo Nacional de Estadística (DANE). En el año 2021 recibió la más reciente certificación por parte del DANE a cinco años, lo que "evidencia la pertinencia, exactitud, puntualidad, oportunidad, accesibilidad, interoperabilidad, coherencia, transparencia, integridad y consistencia de los registros administrativos conocidos por la institución" (Buitrago, Rodríguez y Bernal 2015).

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

en los casos de Bogotá, Medellín o Villavicencio, ciudades donde se desarrollan iniciativas similares a la planteada en este documento; y 4) que idealmente fuera una ciudad intermedia, dado que enfrenta desafíos de seguridad ciudadana y de acceso a la información diferentes a los de ciudades principales.

Para la determinación de la ciudad de análisis, se tuvo en cuenta la magnitud de la ciudad. Como este es un estudio sobre ciudades medianas, se eliminan ciudades como Bogotá o Medellín o municipios pequeños como Barichara en el mismo Santander. De tal modo que para determinar si es ciudad mediana o no, se tuvo en cuenta la cantidad de habitantes entre 500 000 y 1 000 000, de tal modo que alguna ciudad que entre en esos parámetros puede ser elegida. Una vez determinadas las ciudades candidatas, se eligió de manera aleatoria con distribución uniforme la ciudad y se determinó Bucaramanga como ciudad para el estudio.

Entre las alternativas de selección se exploró adelantar el estudio en la ciudad de Bucaramanga, que cuenta con 35 838 delitos registrados entre 2016-2019, que la ubican como la quinta ciudad del país con el mayor número de casos, después de Bogotá, Medellín, Cali y Barranquilla. La desagregación espacial de los *shapefiles* se encontró por sector, sección y manzana, provenientes del DANE. Así mismo, la mayor desagregación está por manzana, pero no cubre la totalidad de territorio urbano de Bucaramanga, por lo que se escogieron los polígonos por sección DANE para mostrar los mapas de calor y hacer el análisis geoespacial. Se eligió la zona urbana del municipio, ya que la inmensa mayoría de los delitos sucedieron allí y la información geográfica más desagregada se encuentra en esta zona.

Metodología para la predicción del delito

A partir del uso de la metodología desarrollada para procesamiento de señales para grafos (Stanković y Sejdić 2019) y en la metodología de vectorización del análisis de texto -con la matriz TF-IDF- aplicada a este problema, con una vectorización de delitos a través del tiempo se elaboró un modelo preliminar de clasificación con el propósito de determinar si en una sección definida por el DANE ocurre un delito con diferentes desagregaciones temporales. La relación plana entre los grafos de cada una de las secciones -de ahora en adelante denominados polígonos-, muestra que cada uno tiene relación con sus vecinos más no tiene relación consigo mismo, es decir, no existe conexión de cada polígono con sí mismo.

Con base en lo anterior, se establecieron los pesos de la relación entre polígonos a partir de la vectorización de una matriz TF-IDF, ampliamente utilizada para el análisis de texto y que se compone de documentos y términos, pero aplicada a este contexto de las categorías de delitos y sin el análisis textual (Stanković y Sejdić 2019). Para el desarrollo del ejercicio se tomaron los polígonos como equivalencia y el tiempo, es decir, cada polígono se le asignó una etiqueta uno (1) si hubo algún delito en un lapso específico.

La matriz 1 que se encuentra en la tabla 1 muestra la transformación de los datos iniciales donde *d* corresponde a *1* si hubo delito y *0* cuando no hubo; *p* es el polígono al cual se refiere, *n* el número identificador del polígono, *t* el periodo actual de tiempo y *T* el periodo total. Lo anterior significa que cada columna corresponde a un periodo y cada fila corresponde a cada polígono -dentro de la equivalencia de la matriz TF-IDF cada polígono sería cada *documento* y cada periodo serían los *términos*-.

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Tabla 1. Matrices, ecuaciones y formulas utilizadas

Matriz 1	Matriz de delitos-tiempo	$\begin{pmatrix} d_{p_1,t-T} & \cdots & d_{p_1,t} \\ \vdots & \ddots & \vdots \\ d_{p_n,t-T} & \cdots & d_{p_n,t} \end{pmatrix}$
Ecuación 1	Cálculo de matriz de frecuencia delito-tiempo	$f(t,p) = \frac{x_{p,t}}{\sum_{t' \in p} x_{t',p}}$
Ecuación 2	Cálculo de matriz inversa de frecuencia delito-tiempo	$if(t,P) = \left(\frac{\lvert P \rvert}{\{\lvert p \in P; t \in p \rvert\}}\right)$
Ecuación 3	Cálculo de matriz F	$F = f * if$
Ecuación 4	Cálculo del modelo de clasificación	$d_{i,t} = g\left(\sum d_{i,t-j} * F\right)$

Fuente: elaboración propia.

Con base en la matriz 1 de la tabla 1, se realiza la transformación de datos a partir de la ecuación 1, que es la proporción de la categoría *delito* en un periodo de tiempo específico relativo a todos los polígonos. Así mismo, la ecuación 2 de la tabla 1 muestra la matriz inversa que representa la proporción de la categoría delito en cada polígono relativo a todos los periodos de tiempo. De tal modo, la ecuación 3 de la tabla 1 muestra la multiplicación de las matrices ya planteadas. Esto, lleva al cálculo de la matriz *F* (matriz TF-IDF) que tiene dimensiones *PxP* y representa la relación de cada polígono con respecto a los demás. Hecho el cálculo de la matriz, esta se multiplica por cada periodo de tiempo dentro de la matriz original teniendo una matriz de dimensiones *PxT* que servirá como insumo para el desarrollo del modelo de clasificación. La ecuación 4 de la tabla 1 muestra la operación descrita.

El *algoritmo de pronóstico* de nuevos periodos de tiempo se basa en estimaciones sobre pronósticos, el cual toma los primeros datos de entrada -la matriz resultante de la multiplicación de cada vector de tiempo con respecto a la transformación de la matriz *F*-, y los datos (*1* y *0*) de salida del periodo inmediatamente posteriores. Después, se entrena un modelo de clasificación y se ingresan los datos de salida de entrenamiento a la matriz de datos de entrada -se hace la misma transformación con la matriz *F*-, y se pronostica el siguiente periodo al de la salida en el modelo originalmente estimado. Los datos pronosticados sirven como nuevos datos de entrenamiento del siguiente modelo para estimar el siguiente periodo.

Para asegurar que las dimensiones en los datos de entrada sean siempre los mismos -y así poder agregar un nuevo periodo cada vez -, se realizó una transformación con análisis de componentes principales (PCA, por sus siglas en inglés) que corresponde a una transformación lineal a los datos de entrada, en el cual se calcula lo que se denomina *componentes principales.* De esta manera, cada componente tiene un porcentaje de varianza explicada, es decir, cada componente explicará un porcentaje de variabilidad del total de los datos. Esta técnica permite reducir las dimensiones de una matriz con K variables. Pues, al calcular sus *componentes principales*, una matriz K-S mantiene una alta capacidad de explicar los datos.

Modelos de aprendizaje supervisado

Existen dos categorías de algoritmos de aprendizaje de máquina: los *modelos de aprendizaje supervisado* y los *modelos de aprendizaje no supervisado*. La principal diferencia entre ambos está en el planteamiento del problema; mientras que los modelos de aprendizaje supervisado se centran en predecir una variable objetivo, los modelos de aprendizaje no supervisado tratan con las relaciones entre los datos y buscan encontrar patrones intrínsecos en ellos. Dada esa divergencia, los modelos utilizados para la predicción de crimen empleados aquí son modelos supervisados, ya que se busca predecir una variable objetivo, en este caso la cantidad de delitos por cada sección en la ciudad de Bucaramanga. Por tal motivo, el modelo resultante es uno de clasificación binaria donde *1* corresponde a que sí hubo un delito en una sección específica, en un periodo de tiempo determinado, y *0* corresponde a que no hubo delito.

Para ello, se probaron los modelos de *K-Nearest Neighbors* y *Support (KNN)* y *Support Vector Machine* como clasificadores. La idea principal dentro de los modelos de KNN es encontrar los vecinos más cercanos de cada vector dentro de un espacio vectorial, esto permite que los pronósticos de cada modelo correspondan a la proporción de clases de los N vectores más cercanos. En este sentido, el hiperparámetro principal en tales modelos es la cantidad de vecinos por tener en cuenta para realizar el pronóstico de la clase.

Por su parte, los *Support Vector Machine para clasificación* son modelos que dividen el espacio vectorial con base en los vectores borde en donde termina una clase y empieza otra. Estos vectores se denominan *vectores de soporte* y generarán la separación del espacio de dichos vectores de soporte en un espacio de $K+p$ dimensiones, donde K son las dimensiones originales y p es la diferencia entre el espacio original y la transformación del espacio a uno nuevo. Esta ampliación en las dimensiones del espacio vectorial es necesaria, ya que la separación de datos en el espacio original no suele ser lineal, mientras que, al aumentar las dimensiones del espacio, la separación en el hiperplano planteado tiende a linealizarse. El modelo elegido en el desarrollo del presente estudio es el modelo KNN, pues presentó mejores resultados de clasificación y tiempos de ejecución, los cuales serán explicados en apartados posteriores.

Discusión y resultados

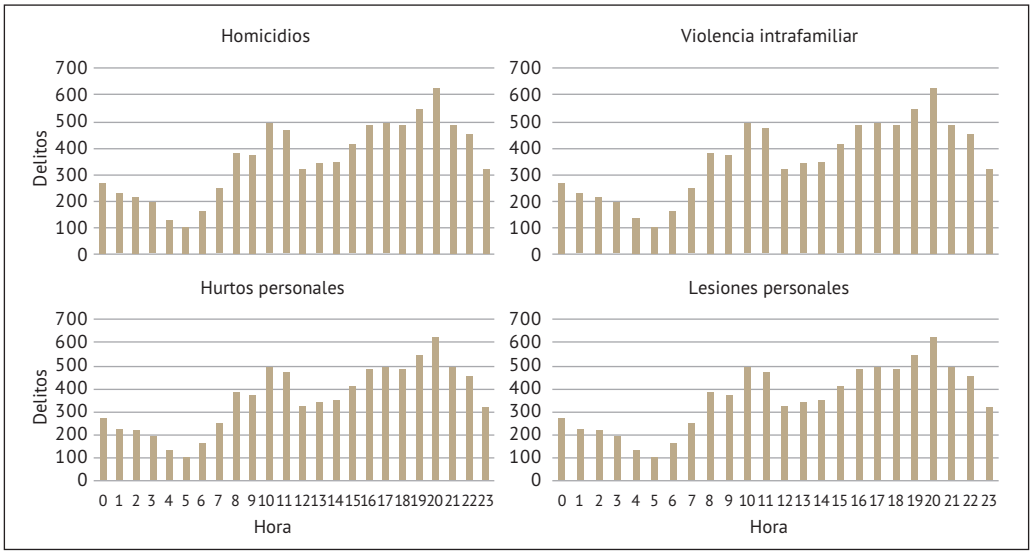
Estadísticas descriptivas

Las estadísticas descriptivas de delitos en Bucaramanga se construyeron teniendo en cuenta la totalidad de los delitos en el tiempo y el espacio. El tipo de delito se separa en *homicidio*, *violencia intrafamiliar*, *hurtos* y *lesiones personales*, y cada hecho puede clasificarse en una de estas cuatro categorías. La diferenciación por el tipo del delito es fundamental para los efectos de este estudio, ya que cada uno tiene un comportamiento distinto tanto en el tiempo como el espacio y su frecuencia varía bastante. Los datos muestran que el 55,5 % de los delitos se compone son hurtos a personas, seguido por lesiones personales 23,6 %, violencia intrafamiliar 19,7 % y, con el 1,1 %, los homicidios.

Igualmente, los datos muestran un incremento en la cantidad de delitos año tras año desde 2016 a 2019. Lo anterior se pudo pre-

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Figura 1. Distribución de los delitos de interés por hora del día



Fuente: elaboración propia con base a datos del SIEDCO.

sentar por el cambio metodológico que tuvo SIEDCO y la puesta en marcha del aplicativo "¡A denunciar!", presentado por Rodríguez et al. (2018). Por su parte, el hurto a personas es el delito que ha experimentado un mayor incremento a través del tiempo. En contraparte, las lesiones personales y la violencia intrafamiliar han presentado una ligera disminución. Ello supone que el aumento de los delitos totales en la ciudad de Bucaramanga se debe al aumento de hurtos a personas, o al cambio de recolección de las denuncias.

Al analizar el número total de delitos por día para cada año, no se encuentra una tendencia clara o un día que resalte con mayor cantidad de hechos. Sin embargo, al graficar el número de delitos por tipo y día de la semana se observa que los homicidios y lesiones personales son más comunes los fines de semana, la violencia intrafamiliar los primeros días de la semana, y los hurtos son mucho menores el domingo que el resto de los días.

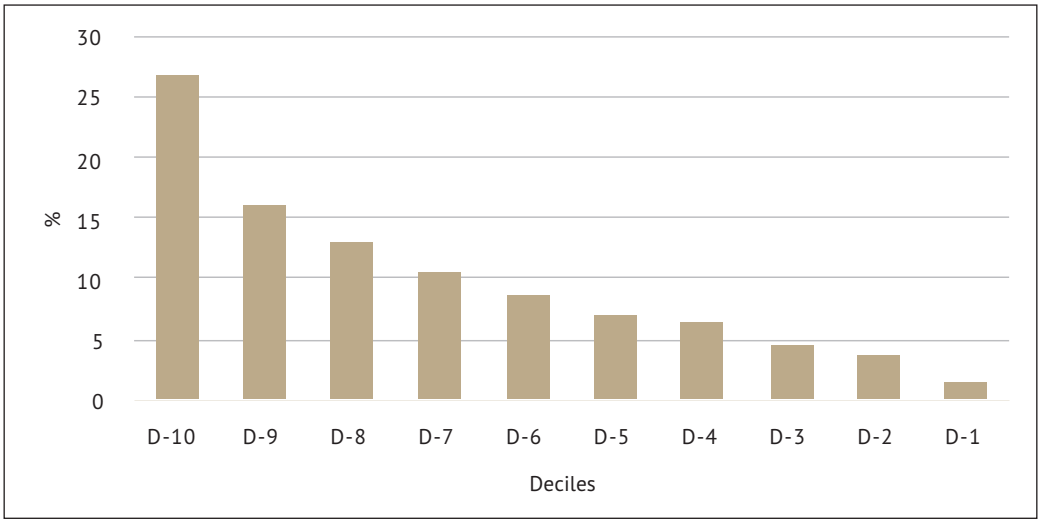
La figura 1 muestra la distribución de los cuatro delitos en Bucaramanga para cada hora del día. Los delitos ocurren con menos frecuencia entre la 1:00 a. m. y las 5:00 a. m. y suben a partir de las 6:00 a. m. El punto máximo sucede a las 10:00 a. m., con más de 2500 registrados a esa hora, dado que es el pico de hurto a personas. A partir de ahí se registra una disminución hasta menos de 1500 delitos a la 1:00 p. m., un crecimiento hasta las 7:00 p. m. y luego una gran caída hasta el final del día, al parecer por lesiones personales, homicidios y hurtos a personas.

Estadísticas espaciales

Como se presentó en la sección 3, se analizó la concentración de los delitos por secciones territoriales se agruparon las 311 secciones DANE en 10 grupos, ordenados según la cantidad de delitos perpetrados en cada grupo durante el período de 2016 a 2019. Lo

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Gráfico 1. Cantidad de delitos por decil, 2016-2019



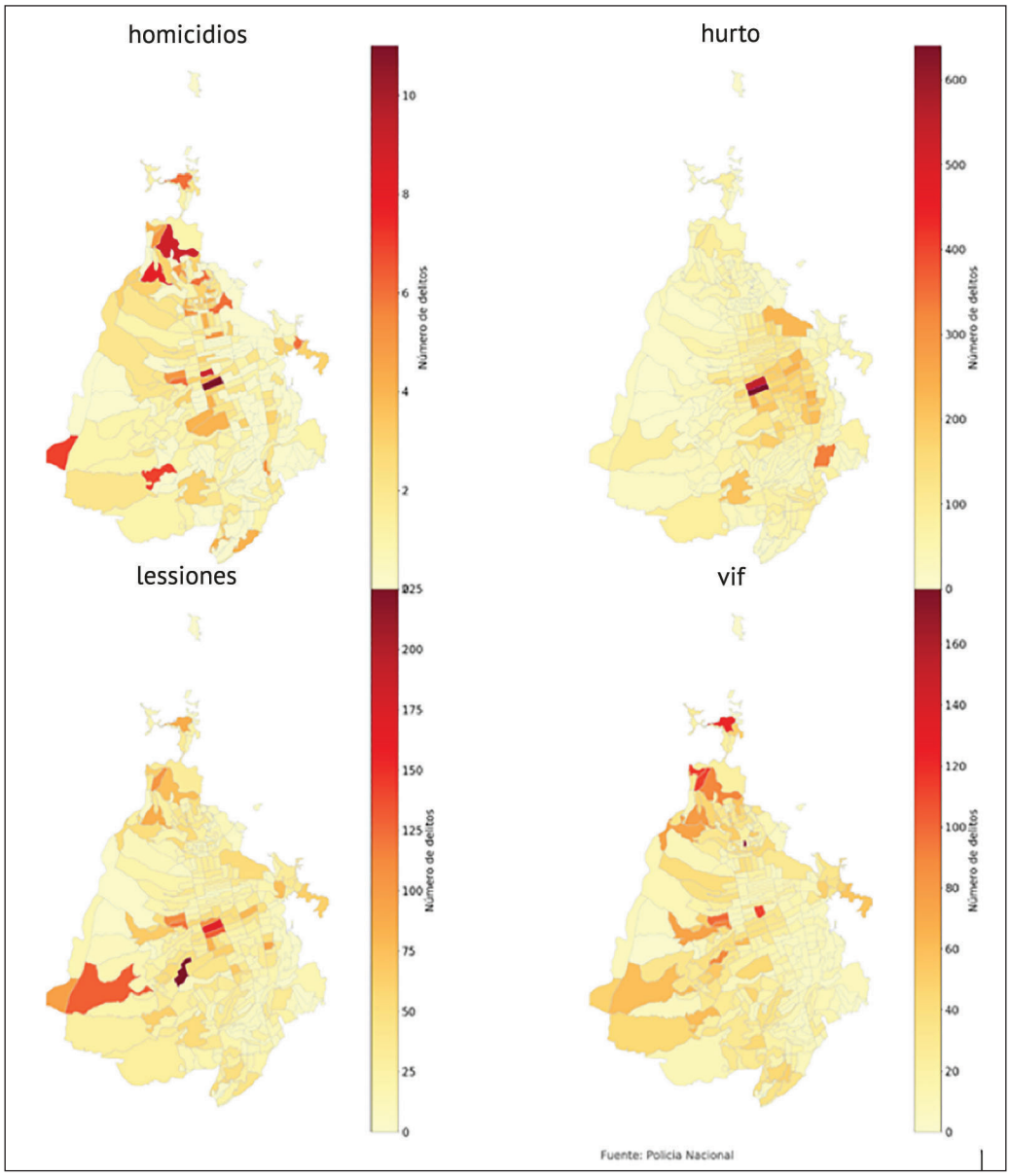
Fuente: elaboración propia con base a datos del SIEDCO.

anterior significa que se dividió la distribución de delitos de las secciones DANE por deciles y se calculó el porcentaje de ellos de cada decil sobre el total, como se observa en el gráfico 1. En el eje horizontal se ubican los deciles de las secciones, mientras que en el eje vertical está el porcentaje de delitos de cada decil con respecto al total.

Como se observa en el gráfico 1, más del 25 % de los delitos sucedieron en el decil con mayor cantidad de delitos; es decir, más del 25 % de los delitos se agrupa en el 10 % de las secciones territoriales -o en un poco más de 30 secciones-. El segundo decil con más delitos concentra un 16 % de los delitos totales, un poco más de la mitad del primer decil. La concentración disminuye significativamente para los demás deciles y el último reúne el 2 % de los delitos totales.

También se presenta una concentración significativa en los diferentes tipos de delitos -homicidio, violencia intrafamiliar, hurtos a personas y lesiones personales-; cerca del 30 % de cada uno de los delitos sucedieron en el 10 % de las secciones. El mapa 1 muestra la cantidad de delitos totales cometidos en Bucaramanga, tanto por sección geográfica DANE, como por tipo de delito. La ciudad de Bucaramanga cuenta con datos georreferenciados de delitos en un nivel de desagregación máximo de manzana, sin embargo, esta información no cubre la totalidad de territorio urbano de Bucaramanga. Por lo tanto, el nivel de desagregación geográfico que se utilizó para realizar los análisis geoespaciales y el modelo de predicción del delito fueron los polígonos por sección DANE. De acuerdo con lo anterior, el sector donde se concentra la mayor cantidad de hechos se encuentra en el centro de la ciudad. De igual manera, hay ciertos sectores en las periferias que presentan altos niveles de delitos en el periodo total de estudio.

Mapa 1. Cantidad de delitos por tipo de delitos en Bucaramanga, 2016-2019



Fuente: elaboración propia con base a datos de la Policía Nacional.

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Resultados del modelo y limitaciones

Los mejores resultados de la predicción del crimen en Bucaramanga se dieron al utilizar los modelos con relaciones espaciales de grafos por semanas. Más específicamente, el modelo que contó con una mayor precisión fue el modelo *KNN* con frecuencia semanal, con el que se obtuvo un nivel de precisión en la predicción entre el 50 % y el 60 %; es decir, cerca de la mitad de los pronósticos que efectuó el modelo eran verdaderos delitos. Este porcentaje de precisión es coherente con experiencias previas de predicción del crimen.[7]

Este tipo de hallazgos permiten llevar un análisis inicial acerca de la dinámica de delito que se presenta en Bucaramanga. Al obtener indicadores de precisión por debajo de la mitad, se intuye que los delitos en esta ciudad no tienen un carácter aleatorio, sino que refieren un componente estructural que, de tenerse suficiente información, permite el desarrollo de modelos de predicción del delito con buenos resultados. Con base en lo anterior, es pertinente continuar con el desarrollo de estudios de modelos de predicción en diferentes ciudades del país, a fin de encontrar qué modelo ofrecería mejores resultados en la predicción del delito, así se lograría mejor gestión de los recursos de las autoridades de policía.

La tabla 2 contiene las matrices de confusión total de los modelos de frecuencia diaria y semanal, respectivamente. Los modelos de frecuencia diaria tienen mayor dificultad, por lo que a pesar de ser el mejor modelo (SVM), su matriz de confusión registra resultados inferiores, frente al de frecuencia semanal.

Tabla 2. Resultados predicción con SVM con frecuencia diaria y KNN con frecuencia semanal

		Predicción	
SVM		No delito	Delito
Original	No delito	8693	755
	Delito	123	70
KNN		No delito	Delito
Original	No delito	6527	3322
	Delito	2042	3037

Fuente: elaboración propia.

Por otro lado, la tabla 3 muestra los pronósticos del mejor modelo (KNN con frecuencia semanal) con diferentes umbrales para considerar el pronóstico de la clase delito. Al igual que en la sección anterior, al aumentar el umbral, se busca aumentar la precisión. Para validar los resultados de la predicción se utilizaron dos métricas *Precision* y *Recall*. La métrica de *Precision* mide el porcentaje de delitos predichos que verdaderamente fueron delitos, mientras que la métrica de *Recall* calcula el porcentaje de delitos que se lograron predecir de manera correcta entre el total de los delitos. De tal modo, para llegar a un pronóstico, depende del error que se esté dispuesto a tolerar entre *Precision* y *Recall*.

En general, se encontraron buenos resultados al utilizar la frecuencia semanal, pero hubo dificultades al realizar el proceso con frecuencia diaria. Sin embargo, depende del usuario y su disposición para priorizar una métrica frente a otra, pues se puede lograr un *recall* mucho más alto, pero a costo de un *precision* relativamente más bajo. Además, estos

7 De acuerdo con Wang (2017) y Kang y Kang (2017), para las ciudades de Los Ángeles y Chicago en Estados Unidos, la validación de los modelos se encuentra entre el 65 % y 80 % de precisión. Mientras que, para el caso de Bogotá, el modelo de predicción desarrollado por Riascos et al. (2020), la validación se encuentra entre un 46 % y 59 %.

URVIO, Revista Latinoamericana de Estudios de Seguridad No. 34 • septiembre-diciembre 2022 • pp.83-98
ISSN 1390-4299 (en línea) y 1390-3691

Tabla 3. Métricas y umbral mínimo para estimación por KNN semanal

Umbral mínimo	0,5	0,6	0,7	0,8
Accuracy	64,07 %	64 %	67,59 %	67,59 %
Recall	59,80 %	59,80 %	19,10 %	19,10 %
Precision	47,76 %	47,76 %	57 %	57,09 %

Fuente: elaboración propia.

modelos muestran que los delitos en Bucaramanga responden a una estructura, más no son dados a un componente estrictamente aleatorio. En consecuencia, con una mayor diversidad de datos podría llegarse a mejores resultados y se facilitaría el uso de variaciones o diferentes técnicas.

Estos hallazgos permiten concluir que, en zonas rurales y centros urbanos pequeños o con pocos delitos, la predicción del delito puede ser ineficiente, dado que la cantidad de hechos registrados es insuficiente para elaborar pronósticos precisos. Asimismo, es importante resaltar que los resultados obtenidos para la ciudad de Bucaramanga no son generalizables para otros municipios del país. En consecuencia, se requiere estimar modelos de cada municipio o área metropolitana, donde se aspire implementar modelos de predicción del delito. Los municipios pequeños o intermedios en el país pueden beneficiarse de otras estrategias de atención del delito basadas en evidencia, como la identificación de puntos calientes de crimen.

Así mismo, dada la crisis de salud pública causada por la COVID-19 en 2020 y las restricciones a la movilidad impuestas por los Gobiernos para reducir el nivel de contagios, los delitos han tenido un comportamiento atípico (Alvarado et al. 2020). Por lo tanto, es probable que los crímenes en Bucaramanga se hayan reducido o cambiado de ubicación. En consecuencia, existe la probabilidad de que las predicciones de delitos como los homicidios, lesiones personales y hurtos a personas, basadas en datos históricos no sean efectivas para las predicciones del año 2020. Será necesario validar con los nuevos datos -conocida como *backtesting*- para confirmar si hubo cambios drásticos en los delitos de ese año que pudieran afectar el desempeño de los modelos. Si ese es el caso, se necesitará calibrar o entrenar de nuevo el modelo. Este proceso de *backtesting* y recalibración puede realizarse de manera periódica para mantener el modelo actualizado y con buena capacidad de predicción.

Conclusiones

El modelo de predicción del delito planteado en este artículo permite formular cuatro conclusiones. Primero, la predicción del delito en entidades territoriales con poca información estadística puede ser una herramienta útil, pero compleja y costosa en su desarrollo e implementación. Las ciudades pequeñas e intermedias cuentan con un limitado acervo de información delictiva y georreferenciada, dados los obstáculos procedimentales y logísticos para un ejercicio riguroso de recolección de información. Por tanto, se hace necesario generar puentes con organismos del tercer sector o la academia, de manera que estas herramientas puedan ser más efectivas y eficientes para prevenir el delito. Otras herramientas de visualización y análisis de datos podrían generar información suficiente para la toma de decisiones de política pública (por ejemplo, mapas de calor, econometría espacial, etc.).

Segundo, es indispensable contar con un mayor acervo de datos e información geoespacial. Con los nuevos modelos de predicción del delito, se abre una ventana para generar estrategias y mecanismos de recolección de información periódica y sostenible sobre las dinámicas del delito y los comportamientos contrarios a la convivencia. Con la coordinación entre las diferentes entidades, se pueden obtener mejores bases de datos. Por ejemplo, con la articulación entre los datos de criminalidad que tiene la Policía Nacional, los datos ofrecidos por las líneas telefónicas de atención a emergencias (123), y la información sobre medidas correctivas, se obtendrán datos más completos y una mejor coordinación entre las entidades involucradas en la atención del delito.

Tercero, no existe un único modelo de predicción del delito, pues estos deben ajustarse a las características de cada entidad territorial. El modelo desarrollado para la ciudad de Bucaramanga no pretende ser el definitivo o replicarse en otras ciudades intermedias. Al contrario, sus resultados invitan a la adaptación, la implementación y el estudio de modelos de predicción del delito de acuerdo con el contexto y la información disponible. Entre otras herramientas, se encuentra el modelo de elipses espaciales y el de estimación de densidad por Kernel (KDE) y KDE con temporalidad. Este último es una buena alternativa a los modelos trabajados antes, pues no requiere desagregación geográfica (Mohler et al. 2011), con lo cual se obtienen resultados óptimos en comparación con otros modelos para la ciudad de Bogotá (Barrera et al. 2016).

Finalmente, al utilizar los modelos de predicción del delito como insumos para formular e implementar políticas públicas, se deben tener en cuenta las implicaciones éticas y los sesgos de su uso. Por un lado, los resultados de estos modelos pueden tener efectos discriminatorios sobre un determinado grupo poblacional o una zona de la ciudad (Karppi 2018). En consecuencia, la verificación y validación respecto a la presencia de sesgos reviste igual importancia en el uso de los resultados de la formulación de estrategias de prevención del crimen. Por otro lado, al utilizar solamente los datos disponibles, los modelos de predicción no tienen en cuenta el posible subreporte de casos y denuncias en determinados delitos, lo cual se traduce en una pérdida de eficiencia. Por consiguiente, las estrategias de prevención basadas en los modelos de predicción deben desarrollarse en paralelo con el fortalecimiento de la justicia y con el acompañamiento de la ciudadanía.

Bibliografía

Abt, Thomas, Chris Blattman, Beatriz Magaloni y Santiago Tobón. 2019. "¿Qué funciona para prevenir y reducir la violencia juvenil? Revisión sistemática de la evidencia sobre prevención y reducción de la violencia juvenil, con un análisis aplicado al contexto mexicano*". USAID*.

Alcaldía Mayor de Bogotá. 2019. "Bogotá desarrollará un método de predicción de delitos", https://bit.ly/3JrFqnr

Alvarado, Nathalie, Ervyn Norza, Santiago Pérez-Vincent, Santiago Tobón y Martín Vanegas-Arias. 2020. "Evolución de la seguridad ciudadana en Colombia en tiempos del COVID-19". *Nota de Política CIEF* 1. doi.org/10.18235/0002780

Barrera, Francisco, Carlos Díaz, Álvaro Riascos y Mónica Ribero. 2016. "A comparison of different crime prediction models for Bogotá". *Documentos CEDE* 34.

Blair, Rob, y Michael Weintraub. 2020. "El Ejército y la seguridad ciudadana: un experimento de campo en Cali, Colombia", https://bit.ly/3SnxCHz

Blattman, Christopher, Donald Green, Daniel Ortega y Santiago Tobon. 2017. "Pushing Crime Around the Corner? Estimating Experimental Impacts of Large-Scale Security Interventions", doi.org/10.2139/ssrn.3050823

Braga, Anthony A., Andrew V. Papachristos y David M. Hureau. 2014. "The effects of hot spots policing on crime: An updated systematic review and meta-analysis". *Justice quarterly* 31 (4): 633-663. doi.org/10.1080/07418825.2012.673632

Brantingham, Patricia, Paul Brantingham y Wendy Taylor. 2005. "Situational crime prevention as a key component in embedded crime prevention*". Canadian Journal of Criminology and Criminal Justice*: 271-292. doi.org/10.3138/cjccj.47.2.271

Buitrago, Julián Ricardo, Jair David Rodríguez y Pedro Aleksander Bernal. 2015. "Registros administrativos de policía para la consolidación de cifras de criminalidad en Colombia". *Revista Criminalidad* 57 (2): 11-22.

Cornish, Dereck, y Ronald Clarke. 2003. "Opportunities, precipitators and criminal decisions: a reply to Wortley's critique of situational crime prevention". *Crime Prevention Studies* 16: 41-96.

Gélvez, Juan David. 2018. "¿Cuáles determinantes se relacionan con la percepción de inseguridad? Un análisis estadístico y espacial para la ciudad de Bogotá, D. C.". *Revista Criminalidad* 61 (1): 69-84.

Gómez, Santiago, Daniel Mejía y Santiago Tobón. 2019. "The Deterrent Effect of Surveillance Cameras on Crime". *Documentos CEDE*.

Hollywood, Jhon, Kenneth McKay, Dulani Woods y Denis Agniel. 2019. "Real-Time Crime Centers in Chicago: Evaluation of the Chicago Police Department's Strategic Decision Support Centers", doi.org/10.7249/rr3242

Instituto de Estudios Urbanos-Universidad Nacional de Colombia. 2016. "Las ciudades intermedias como resultado del proceso de urbanización", https://bit.ly/3BCXx83

Johnson, Shane D., Rob T. Guerette y Kate Bowers. 2014. "Crime displacement: What we know, what we don't know, and what it means for crime reduction". *Journal of Experimental Criminology* 10 (4): 549-571. doi.org/10.1007/s11292-014-9209-4

Kang, Hyeon-Woo, y Hang-Bong Kang. 2017. "Prediction of crime occurrence from multi-modal data using deep learning". *PLoS ONE* 12 (4): e0176244. doi.org/10.1371/journal.pone.0176244

Karppi, Tero. 2018. "The Computer Said So: On the Ethics, Effectiveness, and Cultural Techniques of Predictive Policing". *Social Media + Society* 4 (2): 205630511876829. doi.org/10.1177/2056305118768296

Levine, E. S., Jessica Tisch, Anthony Tasso, y Michael Joy. 2017. "The New York City police department's domain awareness system". *Interfaces* 47 (1): 70-84. doi.org/10.1287/inte.2016.0860

Meijer, Albert, y Martijn Wessels. 2019. "Predictive Policing: Review of Benefits and Drawbacks". *International Journal of Public Administration* 42 (12): 1031-1039. doi.org/10.1080/01900692.2019.1575664

Mejía, Daniel, Ervyn Norza, Santiago Tobón y Martín Vanegas-Arias. 2021. "Broken Windows Policing and Crime: Evidence

from 80 Colombian Cities", doi.org/10.2139/ssrn.3917187

Mohler, George-, Martin Short, Jeffrey Brantingham, Frederic Paik Schoenberg y George Tita. 2011. "Self-Exciting Point Process Modeling of Crime". *Journal of the American Statistical Association* 106 (493): 100-108, doi.org/10.1198/jasa.2011.ap09546

Mohler, George, M.B. Short, Sean Malinowski, Mark Johnson, George Tita, Andrea Bertozzi y P. Jeffrey Brantingham. 2015. "Randomized Controlled Field Trials of Predictive Policing". *Journal Of the American Statistical Association* 110 (512): 1399-1411. doi.org/10.1080/01621459.2015.1077710

Organización Mundial de Ciudades y Gobiernos Locales Unidos. s.f. "Ciudades intermedias- Nexo vital entre lo local y lo global", https://bit.ly/3d2CEJi

Ridgeway, Greg. 2018. "Policing in the Era of Big Data". *Annual Review of Criminology* 1 (1): 401-419. doi.org/10.1146/annurev-criminol-062217-114209

Riascos, Álvaro, Mateo Dulce, Juan Sebastián Moreno y Francisco Gómez. 2020. "Prediciendo el crimen en Bogotá". *Nota de Política CEDE 38*. Universidad de los Andes.

Santos, Rachel. 2014. "The effectiveness of crime analysis for crime reduction: Cure or diagnosis?". *Journal of Contemporary Criminal Justice* 30 (2): 147-168. doi.org/10.1177/1043986214525080

Saunders, Jessica, Priscillia Hunt, y John S. Hollywood. 2016. "Predictions put into practice: A quasi-experimental evaluation of Chicago's predictive policing pilot". *Journal of Experimental Criminology* 12 (3): 347-371. doi.org/10.1007/s11292-016-9272-0

Shuman, David, Sunil K. Narang, Pascal Frossard, Antonio Ortega, y PierreVandergheynst. 2013. "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains". *IEEE Signal Processing Magazine* 30 (3): 83-98. doi.org/10.1109/msp.2012.2235192

SEN (Sistema Estadístico Nacional). 2021. "Resultados de evaluación de la calidad estadística y trabajo conjunto para la encuesta de convivencia y seguridad ciudadana", https://bit.ly/3JtYKAq

Stanković, Ljubisa, y Ervin Sejdić. 2019. *Vertex-Frequency Analysis of Graph Signals*. Signals And Communication Technology.

Wang, Bao, Penghang Yin, Andrea Bertozzi, Jeffrey Brantingham, Stanley Osher, y Jack Xin. 2017. "Deep Learning for Real-Time Crime Forecasting and Its Ternarization". *Chinese Annals of Mathematics*. doi.org/10.1007/s11401-019-0168-y

Wang, Bao, Xiyang Luo, Fangbo Zhang, Baichuan Yuan, Andrea Bertozzi, y Jeffrey Brantingham. 2018. "Graph-Based Deep Modeling and Real Time Forecasting of Sparse Spatio-Temporal Data". *Cornell University*. doi.org/10.48550/arXiv.1804.00684

Weisburd, David, y Cody Telep. 2014. "Hot Spots Policing". *Journal Of Contemporary Criminal Justice* 30 (2): 200-220. doi.org/10.1177/1043986214525083

Wright, Jhon, y Kevin Beaver. 2012. Parenting and Crime. En *The Oxford Handbook of Criminological Theory*, editado por Francis T. Cullen y Pamela Wilcox, 40-68. Oxford University Press.