**Predictive Crime Hotspots**

Christian Carstens a , Mauricio Mendozab , Francisco Tobarc , Elwin van’t Woutd

a Major Computación e Ingeniería de Software, Escuela de Ingeniería, Pontificia Universidad Católica de Chile.

4º año, ctcarstens@uc.cl

b Major Investigación Operativa, Escuela de Ingeniería, Pontificia Universidad Católica de Chile.

4º año, msmendoza@uc.cl

c Major Computación e Ingeniería de Software, Escuela de Ingeniería, Pontificia Universidad Católica de Chile.

4º año, fvtobar@uc.cl

d Instituto de Ingeniería Matemática y Computacional, Escuela de Ingeniería, Pontificia Universidad Católica de Chile. Profesor asistente, [e.wout@uc.cl](mailto:e.wout@uc.cl)

Resumen

La **predicción** de **HOTSPOTS** y su posterior mapeo es una herramienta importante y fundamental en la actualidad, especialmente en el área de criminología. Esta permite enfocar los recursos y lograr una mayor tasa de efectividad en la prevención de crímenes.

Por ende, y dados los límites de recursos y alta tasa delictual en Chile, la policía se encuentra en un proceso de modernización, donde la predicción de **delitos** toma un papel fundamental. Por ende, lograr prevenirlos es una herramienta clave, tanto para la policía como para la población en general.

La hipótesis abordada durante el estudio es que, mediante métodos computacionales, se puede predecir los sectores con mayor probabilidad de ocurrencia de delitos, de manera de lograr una mayor efectividad y aprovechamiento de recursos.

Para lograrlo, se implementaron tres métodos en **PYTHON.** El primero, consiste en un *spatio-temporal* ***KERNEL DENSITY*** (STKDE), que basado en las coordenadas geográficas y temporales, entrega una estimación de delitos en base a estos datos. El segundo fue **PROMAP,** que solo se basa en las coordenadas geográficas. Finalmente, se utilizó el método ***RANDOM FOREST*** *Regressor*, que basado en las coordenadas geográficas y la aplicación de ***MACHINE LEARNING*** es capaz de realizar esta predicción.

Los resultados mostraron el gran desempeño promedio que presenta el STKDE, respecto a los otros dos analizados en la investigación. Además, se da en particular que este resulta aún más útil para abarcar áreas grandes, o un análisis en conjunto y no individual de los delitos. Sin embargo, se destaca la gran utilidad del RFR (Random Forest Regressor), para lo que son áreas promedio o tamaño intermedio, que si lo llevamos a la realidad, se podría plantear como el caso más aplicable o posiblemente requerido por un sistema policial. El caso de análisis de esta investigación fue de la ciudad de Dallas, USA.

***Palabras clave:***predicción, hotspot, delitos, STKDE, machine learning.

# 1. Introducción

Actualmente, el Big Data ha revolucionado muchas áreas de la ciencia, incluso en contextos cotidianos, donde los volúmenes de datos crecen de manera exponencial. En contraste con esta infinidad de datos, los recursos son mucho más limitados. Es por esto que entidades típicas, como la Policía de Investigaciones (PDI), están buscando el desarrollo de esta área para optimizar sus recursos de personal, de manera de predecir de mejor manera la ocurrencia de delitos del país. Actualmente, esta institución está implementando sistemas de predicción de comportamiento y cámaras de reconocimiento facial, los cuales podrían predecir el comportamiento delictivo de una persona o de un determinado sector. Estos requieren implementos complementarios y claramente se ven limitados nuevamente por los recursos.

Además, Hu, Wang, Guin y Zhu (2018) implementaron un **FRAMEWORK** que estima la predicción de crímenes a partir de un STKDE método, obteniendo el índice de precisión predictiva (PAI por sus siglas en inglés), el cual es la base para emularlo en el caso de estudio de esta investigación, comparando su efectividad con métodos también implementados en esta investigación, como lo son Promap y **RANDOM FOREST** regressor.

El objetivo de esta investigación es implementar un método capaz de predecir las áreas geográficas con una mayor probabilidad de ocurrencia de delitos. Esto permite, por ejemplo, señalar a la policía lugares estratégicos donde enfocar sus recursos, de tal manera de tener una mayor tasa de efectividad en la prevención de delitos.

La hipótesis del estudio, y directamente relacionada con el objetivo, es que existen áreas específicas que representan una baja porción del territorio total de un país, en las cuales ocurre la mayor cantidad de ilícitos (estas áreas serán denominadas de aquí en adelante como hotspot). Abarcar policialmente estos puntos implica aumentar significativamente la efectividad policial frente a estos. El caso de estudio de esta investigación es de la ciudad de Dallas, Estados Unidos. Para esto, se utilizó una base de datos proveniente de SOCRATA API, del año 2017, de la cual se consideraron delitos de robo en lugar habitado o entrada forzada. Esta contiene 150.000 datos aproximadamente, de los cuales se sacaron muestras aleatorias para entrenar los modelos.

# 2. Metodología

La metodología usada en esta investigación consistió, en términos generales, en la realización de simulaciones y generación de estadísticas de rendimiento y mapas de calor mediante **PYTHON**.

Primero, se entrenaron los tres distintos modelos mediante bases de datos de delitos ocurridos en la ciudad de dallas.

Estos es, generación de tres clases distintas (una para cada método) que contienen los atributos y métodos necesarios para generar distintos gráficos y estadísticas, con la predicción de los delitos mediante el entrenamiento anteriormente mencionado. Específicamente, las estadísticas (evidenciadas mediante gráficos) consisten en la tasa de aciertos (HR por sus siglas en inglés) y el índice de precisión predictiva (PAI). Por otro lado, se generan los mapas de calor, que muestran la probabilidad de ocurrencia de delitos en un área geográfica mediante distintos colores.

(1)

(2)

Donde *n* representa el número de delitos capturados o predecidos por los hotspots creados en el entrenamiento de cada método, de un total de *N* posibles.

*a* representa el área abarcada por los hotspots creados en el entrenamiento, de un área total *A.*

Particularmente los tres métodos utilizados fueron los explicados a continuación.

El método *Spatio-temporal kernel density estimation* (STKDE), en términos generales multiplica un kernel **BIVARIADO** espacial (plano XY) con uno univariado a través del tiempo. De esta forma, es posible estimar la función densidad de probabilidad de ocurrencia de delitos para coordenadas de espacio tiempo determinadas.

El segundo método utilizado fue promap. Este asigna a cada celda de la grilla inserta en el mapa, un porcentaje de riesgo acumulando los pesos de los delitos cercanos. Estos pesos () se definen como el inverso de la distancia del delito y el tiempo. La fórmula utilizada para calcularlos es la siguiente:

Imagen que contiene reloj

Descripción generada automáticamente

(3)

El último método utilizado, Random Forest Regressor, se basa en la aplicación de los principios de machine learning para realizar una clasificación y regresión capaz de predecir la ocurrencia de delitos en cierto plazo.

# 3. Resultados y discusión

Mediante la aplicación de la metodología anteriormente descrita, se obtuvieron las curvas que comparan los índices PAI y HR obtenidos según en nivel de área que se espera cubrir.

La primera comparación que surge a raíz de los resultados es respecto al estado del arte actual en este rubro, específicamente con la investigación de Hu, Wang, Guin y Zhu (2018)

El análisis de estos autores muestra que, en la comparación del método STKDE con Promap, el primero de estos obtiene un rendimiento superior en la gran mayoría de los rangos de área, excepto en rangos entre un 5% y 7.5% del área total.

Para el caso de esta investigación, se cumple algo muy similar. En efecto, para distintas combinaciones de **BANDWITH** especificados, el método STKDE tuvo un mejor rendimiento en comparación con el del promap, lo cual se ve reflejado en los gráficos de curvas PAI. Con esto, se analiza que ambos tienen un comportamiento muy similar, pero con un rendimiento mejor para el caso del STKDE.

En este mismo gráfico, se puede apreciar la comparación con el tercer método implementado. Este presenta un desempeño similar al resto, pero que difiere por tramos de áreas. Es decir, el comportamiento de la curva es el mismo, pero presenta mejores tendencias en determinados rangos.

Imagen que contiene monitor, pantalla, computadora, televisión

Descripción generada automáticamente

**Figura 1: Comparación de PAI**

En general, y dado a un análisis de sensibilidad realizado durante la investigación, se pudo apreciar que cuando se buscaba abarcar bajas proporciones de áreas, el método promap tendía a tener un desempeño similar o más cercano al método STKDE, pero al intentar abarcar áreas mayores, este último muestra claramente una tendencia a un mejor desempeño.

Además, se ve que el método basado en machine learning, para áreas pequeñas presenta un desempeño claramente inferior, pero a la hora de abarcar áreas medianas o no pequeñas, muestra un claro mejor funcionamiento. Sin embargo, cuando se considera una mayor o gran proporción de la ciudad, el método STKDE sigue presentando el mejor desempeño, seguido del RFR y finalmente el promap.

Lo anteriormente mencionado, también se puede reflejar en los distintos mapas de calor obtenidos para cada método. En efecto, en el caso de promap y Random Forest Regressor, los hotspots presentes son muy específicos y abarcan áreas más bien pequeñas, mientras que en el caso del STKDE, son hotspots más grandes y suaves, no tan sensibles a delitos individuales.

Lo anterior implica que, a la hora de tomar decisiones de dónde enfocar los recursos policiales en la ciudad de Dallas, el STKDE será una mejor opción para áreas grandes, mientras que en áreas pequeñas hay una mayor incógnita sobre el método más apropiado, e incluso mediante visualización, el método promap y RFR pueden resultar más útiles. Particularmente este último, entrega mayor confianza en la predicción para casos de lugares no tan específicos, pero que tampoco requieren abarcar gran parte de la ciudad

**Imagen que contiene animal, coral

Descripción generada automáticamente**

**Figura 2: Comparación de Heatmap**

Además, se incluye una visualización del método STKDE que muestra la propensidad de delitos para las dimensiones (x,y,t) Este permite visualizar y analizar cómo los *hotspots* se comportan al pasar cierto tiempo, representado por la dimensión Z en el gráfico. Esto muestra, por ejemplo, que el riesgo predicho por el STKDE, disminuye a medida que se encuentran en un tiempo más lejano al que ocurrieron los datos entrenados.

**Imagen que contiene interior, oscuro, monitor, luz

Descripción generada automáticamente**

**Figura 3: Visualización predicción espacio-temporal STKDE**

# 4. Conclusiones

Finalmente, mediante esta investigación se pudo profundizar en tres métodos distintos de predicción de delitos. Esto permitió obtener medidas de desempeño, análisis y visualización que permiten, en un contexto de la vida real, decidir qué método utilizar según los requerimientos del usuario.

Mediante la visualización y cálculo de medidas de desempeño como el PAI y HR, fue posible concluir que el método con mejor desempeño promedio fue el STKDE, el cual se presenta con más utilidad especialmente para abarcar áreas más grandes o contextos no tan específicos en cuanto a precisión individual. Por otro lado, para áreas de rango intermedio, como una comuna de una ciudad, el RFR pareciera ser una mejor opción para la predicción.

Gracias a los tres métodos, es posible lograr una mejor predicción y alcance de delitos que los actuales. Esto implica, por ejemplo, un considerable avance en la asignación de vigilancia policial según sectores, donde el método manual o arbitrario de asignación, es considerablemente menos efectivo que los métodos computacionales anteriormente analizados.

De esta manera, gracias a distintas tecnologías y avances computacionales, como se mostró en esta investigación, es posible visualizar y aportar a situaciones cotidianas, que hace unos años era impensado “automatizar”, y que actualmente existen distintas herramientas, tanto de visualización como de predicción, para mejorar la efectividad en contextos cotidianos. La predicción de delitos es solo un enfoque de lo que se puede lograr, pero el potencial aquí es enorme, para toda rama de investigación.

# Glosario

**BANDWITH:** capacidad de propagación del delito en el espacio o tiempo.

**BIVARIADO:** que involucra dos variables simultáneamente.

**FRAMEWORK:** conjunto estandarizado de conceptos, prácticas y criterios para enfocar un tipo de problemática particular, que sirve como referencia para enfrentar y resolver nuevos problemas de índole similar.

**HOTSPOT:** lugar o área con significativa actividad, peligro o violencia (es decir, mayor probabilidad de ocurrencia de delitos)

**KERNEL DENSITY:** herramienta que calcula la densidad de entidades en un vecindario alrededor de cierta entidad.

**MACHINE LEARNING:** disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente.

**PROMAP:** método para calcular la probabilidad de un delito a futuro en cada punto del mapa.

**PYTHON:** lenguaje de programación interpretado orientado a objetos.

**RANDOM FOREST:** algoritmo de aprendizaje supervisado que utiliza un método de aprendizaje conjunto para la clasificación y regresión.

# Agradecimientos

Agradecimientos al profesor Elwin Van't Wout, por su paciencia, apoyo y conocimientos entregados durante este proceso de investigación.

# Referencias

Hu, Y. Wang, F. Guin, C. Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation.

Kate J, B. Shane D, J. Ken, P. (2004). PROSPECTIVE HOT-SPOTTING. Oxford University.

Policía de Investigaciones (centro de prensa). (2018). Biometría para identificar autores de delitos. Recuperado de <https://www.pdichile.cl/centro-de-prensa/detalle-prensa/2018/05/02/biometr%C3%ADa-para-identificar-a-autores-de-delitos>.

Weisburd, D. Bernasco, W. Bruinsma, G. Putting Crime in its Place. Unidad de Análisis en Ciminología Geográfica. Editorial Springer.