Reporte de Validación de Modelos de Criminología

Juan Felipe Cerón Uribe

21 de julio del 2017

El objetivo de este reporte es establecer una metodología de referencia de validación y comparación de modelos de predicción de crimen. En este se discuten el proceso y los resultados de la validación por detección de hotspots de los modelos de predicción de crimen en Bogotá implementados en Quantil hasta el momento. Presentaremos brevemente el modelo, los dos métodos para su estimación desarrollados hasta el momento, explicaremos la metodología de validación y finalmente discutiremos los resultados.

Índice

1. Los modelos predictivos

Ambas metodologías, entrenadas con datos reales de criminalidad en Bogotá, pretenden acomodar un modelo ETAS (Epidemic Type Aftershock Sequence) a los datos disponibles de criminalidad en Bogotá. La propuesta principal del modelo es que, además de los factores de transfondo que incrementan los índices de criminalidad en ciertas zonas, la criminalidad es un proceso auto-exitante. Esto quiere decir que la ocurrencia de un evento incrementa la probabilidad de observar otros en una vecindad del primero. De modo que la intensidad condicional de eventos en un punto espacio-temporal λ se puede dividir en dos componentes:

$$\lambda(t,x,y) = \nu(t)\mu(x,y) + \sum_{\{k:t_k<t\}} g(t-t_k, x-x_k, y-y_k) \tag{1}$$

Llamamos al factor $\nu(t)\mu(x,y)$ la componente de transfondo, pues esta explica el impacto del contexto espacio temporal en la intensidad puntual. Llamamos a $g(\Delta t, \Delta x, \Delta y)$ la función de desencadenamiento, pues esta explica como un crimen desencadena la ocurrencia de otros en una vecindad del mismo.

En ambos métodos implementados la estimación de cada una de estas funciones es un proceso iterativo en el cual:

1. Se ejecuta una tarea de declustering para separar los datos en eventos de transfondo y réplicas (eventos desencadenados por otros eventos).
2. Se estiman las funciones μ y ν utilizando los datos de transfondo y la función g con las réplicas mediante estimación de densidad por kernels (kernel density estimation).

Para una discusión completa del algoritmo vea el artículo de Mohler [1]. La diferencia entre los metodos de estimación radica en la estimación por kernels de las funciones μ, ν y g (son iguales en cuanto a la metodología de declustering). A continuación explicaremos estas diferencias.

1.1. Ancho de banda fija

El primer método de estimación que consideramos utiliza la forma clásica de un estimador de densidad por kernels. Este busca estimar una función de densidad de probabilidad $p(x)$ a partir de observaciones $\{x_i\}_{i\leq n}$:

$$\hat{p}(x) = \frac{1}{nh}\sum_{i\leq n} K\left(\frac{x-x_i}{h}\right) \tag{2}$$

Donde K es un **kernel**, esto es, una función real no negativa tal que:

- $\int_{\mathbb{R}} K(x)dx = 1$
- $K(0) = 0$

En todas las estimaciones utilizamos un **kernel normal**:

$$K(x) = \frac{1}{2\pi}e^{-\frac{x^2}{2}}$$

El parámetro h, conocido como el ancho de banda, define el grado de *suavización* (smoothing) del estimador, y juega un papel crucial en la estimación. Los anchos de banda usados para entrenar los modelos fueron determinados en un ejercicio de validación previo al presente:

$$h_g = h_\mu = h_\nu = 0{,}5$$

1.2. Ancho de banda variable

Nuestro segundo método de estimación implementa el algoritmo descrito por Mohler [1]. En este paper, Mohler observa que dado que el número de eventos de transfondo y réplicas cambia en cada iteración (y por ende en cada estimación de g, μ y ν), un ancho de banda fijo suavizará las funciones demasiado en algunos casos y no lo suficiente en algunos otros. De modo que Mohler utiliza un ancho de banda variable como el que describiremos a continuación:

Para una muestra $\{x_i\}_{i\leq n}$ sea σ la desviación estándar de la muestra y sea D_i la distancia del punto i a su m-simo vecino más cercano. Definimos entonces la densidad estimada como:

$$\hat{p}(x) = \frac{1}{n\sigma} \sum_{i \leq n} \frac{1}{D_i} K\left(\frac{x - x_i}{\sigma D_i}\right) \tag{3}$$

Observe que el ancho de banda corresponde al factor σD_i con lo cual la suavización de un kernel es proporcional al grado de agrupamiento de puntos cerca al punto i.

En este artículo [3], citado por Mohler, los autores recomiendan el uso del 10-mo al 100-mo vecino más cercano en la distancia D_i. Para la elaboración de este reporte utilizamos el 10-mo en todos los casos (g, μ y ν) dado que utilizamos menos datos de entrenamiento que la bibliografía en general.

Es importante resaltar que, tanto en esta metodología como la anterior, una parametrización adecuada es fundamental para descubrir los aspectos relevantes de los datos, con el fin de obtener un mayor poder predictivo.

1.3. Dimensión temporal cíclica

Además de lo referente al ancho de banda en la estimación de kernels, los modelos estimados difieren en cuanto a su manejo de la dimensión temporal de los eventos:

En el método de ancho de banda fija se definió una dimensión temporal T cíclica como función del tiempo t en el cual ocurrió un evento. $T \in [0, 42)$ indica el momento de la semana al que corresponde t, donde $T = 0$ representa las 12am de un lunes y $0 \equiv 42$. La componente temporal de transfondo (nu) se entrenó con la dimensión temporal $T(t)$, sin embargo se usó t para calcular la componente de desencadenamiento. De modo que este modelo en realidad tiene la forma:

$$\lambda(t, x, y) = \nu\left(T(t)\right) \mu(x, y) + \sum_{\{k : t_k < t\}} g(t - t_k, x - x_k, y - y_k) \tag{4}$$

El modelo estimado por el método de ancho de banda variable es exactamente el descrito por la ecuación 1.

2. Metodología de validación por hotspots

Por medio de esta metodología, desarrollada por Cheng [2], evaluamos la capacidad de un modelo de predecir **hotspots**; subregiones espaciales que presentan la mayor criminalidad en una ventana temporal w_t.

Para una ventana de tiempo w_t dividimos la región considerada, de área A, en subregiones cuyo tamaño depende de la granularidad del análisis deseada. Siguiendo a los autores definimos un **grado de cobertura** $\frac{a}{A}$ donde a es el área cubierta por hotspots. Posteriormente elegimos las subregiones de más alta intensidad estimada como hotspots hasta que la suma de sus áreas sea $\geq a$. La cobertura puede definirse siguiendo límites prácticos como la máxima cobertura que puede lograr la policía de una ciudad en la duración de w_t.

Una vez identificados los hotspots en una ventana temporal w_t calculamos el **hit-rate** como la proporción de eventos capturados por hotspots en la ventana temporal w_t. Sea N_t el número total de eventos en w_t y n_t el número de eventos capturados por algún hotspot:

$$hit\ rate = \frac{n_t}{N_t}$$

Note que esta métrica es comparable únicamente bajo un grado de cobertura constante. Evaluamos esta métrica a lo largo de varias ventanas temporales consecutivas con el fin de reducir la varianza de nuestra estimación.

Luego de tomar estas métricas podemos usar el hit-rate promedio de dos modelos para decidir cuál tiene un mayor poder predictivo. Alternativamente, si asumimos que la diferencia en el poder predictivo entre dos métodos es independiente de las tasas de criminalidad que analizan, la secuencia de diferencias es una muestra de variables aleatorias *iid*. Esto último nos permite usar pruebas estadísticas estándar para comparar los modelos. Para permanecer alineados con la metodología de Cheng utilizamos la prueba WSR (Wilcoxon signed-rank test), una prueba no paramétrica que en este caso nos dirá si uno de los métodos resulta en hit-rates significativamente más altos que el otro.

Para su artículo, Cheng [2] realizó un ejercicio de validación por hotspots con las siguientes características:

- Entrenó cuatro modelos diferentes, entre los cuales está el de Mohler.
- Utilizó datos del 1 de marzo del 2011 hasta el 6 de enero del 2012 proporcionados por la policía de Chicago, dejando los últimos 100 días del conjunto de datos (a partir del 28 de septiembre) para validar los modelos comparados, con ventanas de validacion de un día.
- Utilizó como subregiones una grilla regular de cuadrados de lado $250m$.
- Definió una cobertura del 20 %.

3. Implementación

En esta sección describiremos los detalles de la implementación de los dos métodos de estimación y de la metodología de estimación utilizados en la elaboración de este reporte. El código utilizado, al igual que un documento que contiene una descripción extensa de los datos de criminalidad que Quantil tiene a su disposición, se encuentran en un repositorio de Bitbucket llamado "Validación crimen".

3.1. Datos

En Quantil contamos (entre muchos otros) con datos de criminalidad en Bogotá reportados por la Policía Nacional de Colombia desde el 1 de enero del 2004 hasta el 9 de marzo del 2014. Sin embargo, la ejecución del entrenamiento y validación del modelo presente son tareas muy complejas (tanto en memoria como en procesamiento) y fueron ejecutadas en computadores de uso personal, por lo cual fue necesario limitar el número de datos a utilizar.

Estimamos el modelo ETAS bajo ambas metodologías vistas usando como datos de entrenamiento los periodos de abril a junio (incluyendo ambos meses completos, 8923 registros) del 2010, de mayo a junio (5899 registros) y los datos de junio únicamente (2837 registros), para un total del seis estimaciones diferentes del modelo. Utilizamos las primeras dos semanas del mes de julio como periodo de validación (1356 registros), con ventanas temporales de validación de un día.

A lo largo del ejercicio de entrenamiento y validación de los modelos utilizamos tres computadores de uso personal (dos de 4GB y uno de 8GB de RAM).

3.2. Ancho de banda fija

Como mencionamos anteriormente, enternamos tres modelos con el método de ancho de banda fija con uno, dos y tres meses de datos de entrenamiento. El algoritmo de entrenamiento en este caso estaba implementado previo a este ejercicio (fue desarrollado por Mónica Ribero). Sin embargo, en su estado original, la implementación tiene dos problemas que hacen inviable su ejecución en computadores de uso personal:

1. Dependende del cálculo previo de una estructura que contiene las restas entre cada evento y todos sus eventos anteriores (facilita el cálculo de la función g). Con n datos de entrenamiento, esta estructura tiene $\frac{(n-1)(n-2)}{2}$ filas y tres columnas: latitud, longitud y tiempo. Por esta razón ninguna de las máquinas disponibles pudo sostener esta estructura con dos meses de datos de entrenamiento.

2. Desaprovechamiento de la memoria RAM en el cálculo de la función λ, la cual se calcula en cada dato en el algoritmo de entrenamiento.

Con respecto a lo primero la solución se redujo a evitar el uso de esta estructura. Esto no representó un gran sacrificio en tiempo de cómputo dado que el algoritmo implica operaciones mucho más complejas como el cálculo de kernels y la agrupación de datos.

Con respecto a lo segundo, desarrollamos para este ejercicio de validación una versión vectorizada de la implementación original. A modo de prueba calculamos la intensidad estimada λ en 54730 puntos espacio temporales (la validación requiere operaciones de este orden) con 678 datos de entrenamiento. La versión vectorizada demoró 65 segundos mientras que la original demoró 188; la primera fue mucho más ágil aun usando pocos datos de entenamiento.

Cuadro 1: Resultados del entrenamiento con ancho de banda fija

Meses	Transfondo	Réplicas	Ejecución (horas)
Junio	340	2484	0.17
Mayo-Junio	630	5255	1.24
Abril-Junio	973	7947	1.75

3.3. Ancho de banda variable

La implementación de este algoritmo se desarrolló en su totalidad para este ejercicio de validación. Una ventaja del anterior sobre este es que existen paquetes (en Python y R, al menos) que implementan la estimación de densidades por kernels. Sin embargo no existen paquetes que implementen esta estimación usando anchos de banda variables. Los paquetes implementados comunmente son escritos por profesionales de la computación numérica, por lo cual podemos contar con que están altamente optimizados. La estimación implementada en este caso está altamente vectorizada, sin embargo, como veremos, tanto el entrenamiento de los modelos como su validación fue mucho más eficiente con ancho de banda fija.

Cuadro 2: Resultados del entrenamiento con ancho de banda variable

Meses	Transfondo	Réplicas	Ejecución (horas)
Junio	13	2811	23.4
Mayo-Junio	14	5871	53.4
Abril-Junio	3	8917	130.12

Note como este algoritmo tiende a identificar un efecto mínimo de transfondo. Más adelante discutiremos las posibles consecuencias de este fenómeno.

3.4. Identificación de hotspots

Para identificar los hotspots en una ventana temporal w_t es imperativo contar con una división del área de estudio en subregiones. Utilizamos una grilla regular de cuadrados de lado $\approx 380m$, con lo cual obtenemos un total de 10946 subregiones. De modo que los hotspots en la ventana w_t son las 2190 regiones de mayor intensidad a lo largo de w_t, donde la intensidad de una (sub)región S se define como

$$I(S) = \int_S \lambda(s) ds \tag{5}$$

Dada la limitada capacidad de cómputo del ejercicio estimamos esta integral mediante un método de Monte Carlo tomando 10 puntos aleatorios por subregión. De modo que en la validación de la ventana w_t debemos calcular λ en

$$10946 \text{ subregiones} \times 10 \text{ puntos}$$

Este es el número de veces que calculamos ν y μ. Pero por cada uno de estos debemos calcular un componente de desencadenamiento, lo cual implica E cálculos de g (donde E es el número de datos de enternamiento), de modo que ejecutamos un total de

$$10946 \text{ subregiones} \times 10 \text{ puntos} \times E \text{ datos}$$

cálculos de g. Finalmente, recordemos que el cálculo de una densidad estimada por kernels implica la suma de un kernel por cada punto usado en la estimación (g se estima a partir de E_r réplicas identificadas). De modo que la validación completa tiene una complejidad de:

$$10946 \text{ subregiones} \times 10 \text{ puntos} \times E \text{ datos} \times E_r \text{ réplicas}$$

Por supuesto, es posible vectorizar y paralelizar muchos de estos cálculos para agilizar su cómputo, pero solo hasta cierto punto. Dado el tamaño del problema, la estructuras de datos utilizadas por el algoritmo superan facilmente la memoria RAM de los computadores de uso personal, como los que utilizamos para la elaboración de este reporte. La solución a esta problemática fue el desarrollo de un programa de validación que recibe como parámetro el número máximo de subregiones de las cuales la máquina que lo ejecuta puede calcular la intensidad simultaneamente.

Esta parametrización habilitó la ejecución de la validación en las máquinas disponibles, sin embargo tuvo un costo predecible en el tiempo de ejecución, ya que nos vimos obligados a limitar la concurrencia de las operaciones realizadas. Un computador de 4GB de RAM consiguió calcular la intensidad en tan solo 100 subregiones concurrentemente, donde el máximo es 10946.

En los modelos de ancho de banda fija la validación del modelo con tres meses de datos de entrenamiento tardó cuatro días. El modelo de ancho de banda variable con datos de entrenamiento de un mes tardó catorce días en completar su ejecución. Los dos restantes no conluyeron su ejecución; en catorce días conseguimos validar 2 de las catorce ventanas temporales con el modelo de ancho de banda variable y dos meses de datos de entrenamiento, y ninguna ventana con el de tres meses de datos. De modo que incluímos únicamente los resultados parciales de la validación de estos últimos dos modelos en este reporte.

Concluímos que, si bien limitar la paralelización de los cálculos involucrados en la validación habilita su ejecución en un mayor espectro de máquinas, estas deben ser capaces de cierto nivel de concurrencia para completar la validación en un tiempo razonable. En la última sección discutiremos por qué el algoritmo de ancho de banda variable presenta una complejidad tan alta y cómo podríamos reducirla.

4. Resultados

Ahora presentamos los resulados del proceso de validación de ambos métodos de estimación y la prueba WSR. Tenga en cuenta que, dado el grado de cobertura considerado de 20 %, si elegimos los hotspots aleatoriamente el valor esperado del hit-rate es de 20 %.

4.1. Ancho de banda fija

Cuadro 3: Hit-rates con uno, dos y tres meses de entrenamiento

Ventana	Observados	1 Mes	2 Meses	3 Meses
1	95	0.147	0.189	0.221
2	99	0.202	0.242	0.152
3	91	0.209	0.208	0.187
4	99	0.202	0.212	0.202
5	75	0.213	0.213	0.173
6	99	0.192	0.202	0.162
7	95	0.274	0.189	0.168
8	101	0.218	0.188	0.188
9	107	0.196	0.215	0.243
10	102	0.245	0.255	0.176
11	85	0.2	0.224	0.224
12	86	0.267	0.256	0.198
13	112	0.17	0.17	0.25
14	110	0.209	0.173	0.218
Promedio	96.9	0.21	0.217	0.197

4.2. Ancho de banda variable

Cuadro 4: Hit-rates con uno, dos y tres meses de entrenamiento

Ventana	Observados	1 Mes	2 Meses
1	95	0.568	0.653
2	99	0.596	0.616
3	91	0.484	
4	99	0.515	
5	75	0.453	
6	99	0.475	
7	95	0.484	
8	101	0.564	
9	107	0.467	
10	102	0.48	
11	85	0.424	
12	86	0.477	
13	112	0.491	
14	110	0.464	
Promedio	96.9	0.496	0.634

4.3. Wilcoxon Signed Rank Test

Esta es la prueba de hipótesis no paramétrica que utiliza Cheng [2] para determinar si uno de dos modelos predictivos es significativamente mejor (en términos de hit-rate) que otro. Para cuatro de los modelos entrenados contamos con la validación de todas las ventanas temporales: todos los entrenados con ancho de banda fija y el entrenado con un mes de datos con ancho de banda variable. De modo que necesitamos seis pruebas para probar todas las combinaciones:

Cuadro 5: Pruebas WSR

Modelo 1	Modelo 2	p-valor
Fija 1 mes	Fija 2 meses	0.48
Fija 1 mes	Fija 3 meses	0.42
Fija 2 meses	Fija 3 meses	0.24
Fija 1 mes	Variable	0.00098
Fija 2 meses	Variable	0.00098
Fija 3 meses	Variable	0.00097

Concluímos que el único modelo cuyo poder predictivo difiere de los demás es el entrenado con ancho de banda variable. Es razonable, dados los resultados parciales de la validación obtenidos, pensar que los otros modelos de ancho de banda variable también tienen un mayor poder predictivo que todos los de ancho de banda fija, cuyas capacidades de predicción no parecen superar la elección aleatoria de hotspots.

5. Conclusiones

El principal resultado de este ejercicio de validación es el establecimiento de una metodología de comparación y validación de los modelos de predicción de crimen desarrollados y por desarrollar en Quantil. Ahora que esta existe tiene sentido explorar diferentes modelos y diferentes parametrizaciones de los que ya conocemos.

En cuanto a la validación realizada puntualmente, los modelos entrenados con ancho de banda fija mostraron tener un poder predictivo nulo (hit-rates muy cercanos al esperado con predicciones aleatorias). Concluimos en este aspecto que para utilizar modelos de este tipo es necesario estudiar mucho mejor su parametrización, por ejemplo a través de validación cruzada. Además concluímos que la estimación de los modelos de ancho de banda variable es más robusta dado que aprovecha las características de los datos para definir el ancho de banda automáticamente. Esto es importante pues, como lo dice Mohler [1], los datos de la estimación de densidades por kernels cambian en cada iteración de este algoritmo.

Otro factor que pudo afectar el poder predictivo de los modelos de ancho de banda fija es un problema que presenta el uso de la dimensión temporal cíclica definida, el cual describimos mediante un ejemplo: Si un crimen sucede un domingo a las 23:59, con una estimación estándar de kernels esto no propaga ninguna probabilidad adicional de crímines el lunes en la madrugada. Esto se debe a que en este caso $T(t) \to 42$, mientras que el lunes en la madrugada es representado por números cercanos al 0. Para corregir este error es necesario implementar un kernel cíclico.

Aunque los modelos de ancho de banda variable mostraron cierto grado de poder predictivo, los tres mostraron un potencial problema en común; identificaron muy pocos eventos de transfondo en comparación con las réplicas. Esto implicaría que el efecto de desencadenamiento mutuo de crímenes es mucho más alto que el de las características de algunos sectores de Bogotá. Esto puede ser cierto, o tal vez lo observado en la validación revela un sesgo del algoritmo de entrenamiento hacia el efecto de desencadenamiento. Es posible estudiar este problema como lo hace Mohler [1]. Él simula un proceso puntual bajo el modelo ETAS usando funciones μ, ν y g predefinidas. Esto le permite validar que su algoritmo es capaz de identificar exitosamente cierta estructura de réplicas-transfondo. Esto no dice nada acerca del poder predictivo de un algoritmo, pero puede ayudarnos a explorar los posibles problemas que presenten los nuestros.

Finalmente concluímos que en el proceso de desarrollo y validación de estos modelos es imperativo contar con máquinas de alto rendimiento (o alquilarlas en la nube). Esto le permitirá a los investigadores de Quantil experimentar diferentes parametrizaciones y modelos con una mayor flexibilidad, además de hacer más eficiente cualquier proceso de desarrollo al disminuir el tiempo durante el cual cierto experimento se encuentra en ejecución y es imposible avanzar.

Otras conclusiones menos relevantes:

- La precisión de la estimación de la integral de λ en una región mediante el método de Monte Carlo puede sesgar la elección de los hotspots en una ventana temporal. Es necesario revisar resultados que acoten el error de esta estimación para elegir el número de puntos necesarios en la misma.
- Cheng [2] sugiere un límite superior a la distancia de desencadenamiento: Un crímen no puede desencadenar otro crímen a una distancia mayor que 300m o 60 días. Esto es altamente consistente con el modelo y puede aliviar el sesgo observado por el efecto de desencadenamiento en los modelos de ancho de banda variable.
- El planteamiento de una visualización de la intensidad estimada $\hat{\lambda}$ y sus componentes podría ser útil para entender los problemas que presentan.

Referencias

[1] P. J. Brantingham F. P. Schoenberg y G. E. Tita G. O. Mohler M. B. Short. "Self-exciting point process modeling of crime". En: *Journal of the American Statistical Association* (2011).

[2] G. Rosser y T. Cheng M. Adepeju. "Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study". En: *International Journal of Geographical Information Science* 30.11 (2016), págs. 2133-2154.

[3] Ogata Y. y Vere-Jones D. Zhuang J. "Stochastic declustering of space-time earthquake occurences". En: *Journal of the American Statistical Association* (2002).