Universidad de San Andrés
Departamento de economía
Licenciatura en economía

Detección de zonas calientes de homicidios en la Ciudad de Buenos Aires usando análisis de clusters

Autor: Agustín Nogara
Legajo: 25122
Mentor: Walter Sosa Escudero

Buenos Aires, Argentina
22 de junio del 2018

Abstract

En este trabajo se realiza un análisis de zonas calientes de homicidios en la Ciudad de Buenos Aires. Para esto, se utiliza el método de *hierarchical clustering* sobre una base de datos de crímenes recientemente creada por el gobierno local. Se eligen tres alturas que determinan el nivel de diferencia entre observaciones aceptado en un cluster y, en base a estas, se define la cantidad de clusters. Bajo el corte medio y alto se identifican ubicaciones en la zona sur y este como zonas calientes de homicidios en la vía pública. Esto podría tener un impacto tanto a nivel macro en la toma de decisiones políticas en temas de seguridad, como a un nivel microeconómico en las decisiones de asentamiento de las familias y las empresas.

Sección I: Introducción

La inseguridad es un tema de gran interés hoy en Argentina. Según encuestas, desde hace años la inseguridad encabeza las preocupaciones de los argentinos. En los últimos años se ha observado un considerable crecimiento de los barrios privados y del gasto en seguridad privada. De acuerdo al Instituto Nacional de Estadísticas y Censos (Indec) en el 2001 habían 281 barrios cerrados en el país, en 2010 ese valor escaló a los 450. Si bien no hay estadísticas oficiales recientes respecto a la cantidad de urbanizaciones privados, la Federación Argentina de Clubes de Campo estima que en 2017 la cifra superó los 900. Con respecto al gasto en seguridad privada, de acuerdo a la Cámara Nacional de la Industria de la Transformación, en 2008 el gasto promedio en seguridad privada por parte de empresas representaba el 3% del total del gasto, mientras que en 2017 dicho valor ascendió al 10%.

Claro está que, dentro de una ciudad, no todas las zonas tienen el mismo nivel de criminalidad; identificar las “zonas calientes” podría tener fuertes repercusiones en cuanto al diseño de políticas públicas y la asignación eficiente de recursos de seguridad. Además, desde una perspectiva microeconómica los resultados de este trabajo podrían resultar útiles en cuanto brinden más y mejor información a las familias y a las empresas privadas. Desde el punto de vista de las familias, una buena pregunta que incide en el largo plazo es en dónde vivir, conocer la ubicación de zonas calientes podría ayudar a realizar decisiones más informadas respecto a esta temática; esto aplica de igual manera a los inmigrantes que buscan asentarse en Buenos Aires. Desde el punto de vista de las empresas, el razonamiento es similar; para reducir potenciales costos por inseguridad una firma podría decidir evitar ubicarse en zonas calientes.

Johnson (2008) ha mostrado que la delincuencia puede propagarse en los entornos locales a través de un mecanismo similar al del contagio. Por ejemplo, ladrones podrían atacar repetidamente ciertas ubicaciones porque las vulnerabilidades de esos lugares son bien conocidas por los agresores. La propagación local y contagiosa del crimen conduce a la formación de clusters de delincuencia en el espacio y el tiempo.

La literatura criminalística llama a estos clusters "*zonas calientes*". Una zona caliente es un área geográfica delimitada donde se concentran incidentes delictivos repetidamente en el tiempo. Para que una ubicación sea denominada zona caliente debe poseer las siguientes dos características (Levine, 2013):

1. Característica espacial: debe existir cierta estructura de aglutinamiento en los incidentes.

2. Característica temporal: la característica espacial debe reiterarse en el tiempo.

Estas zonas calientes pueden definirse en base a la actividad o tipo de crimen cometido como por ejemplo, comercio de drogas (Weisburd and Green, 1995).

Con la aparición de nuevas bases de datos de crímenes, las herramientas de Data Mining resultan sustancialmente útiles para poder ubicar zonas calientes; lo que ayudaría mucho a la prevención y disminución del crimen.

El propósito de este trabajo es encontrar las zonas calientes de homicidios en la Ciudad Autónoma de Buenos Aires utilizando un análisis de clusters. Se encuentra evidencia a favor de la existencia de zonas calientes de homicidios a lo largo del sur y del este de la ciudad.

El trabajo se dividirá en cinco secciones adicionales. En la Sección II se realizará una breve revisión de la literatura previa. En la Sección III se describirá la base de datos a utilizar. En la Sección IV se explicará la metodología y la estrategia para validar resultados. En la Sección V se expondrán los resultados y finalmente, en la Sección VI, las conclusiones.

Sección II: Revisión de literatura

Existen diversos trabajos que emprenden el desafío de poder identificar zonas calientes de crímenes en ciudades urbanas. Muchos de ellos también abordan la temática mediante un análisis de clusters.

En el trabajo de Mohler, Short, Brantingham & Schoenberg (2011) se muestra que la metodología de space-time clustering utilizada en sismología para el análisis de terremotos es perfectamente aplicable al ámbito criminalístico. Dicho trabajo resulta relevante dado que prueba que una herramienta que era tradicionalmente empleada en el ámbito de la geografía y la estadística es compatible con el mundo de las ciencias sociales, en particular con los estudios del crimen. Al mismo tiempo, Johnson et al. (2008) muestran que la delincuencia se propaga de manera estructurada dando lugar a la formación de clusters.

Por otro lado, Grubesic & Murray (2002) destacan los potenciales beneficios y desventajas del *clustering* para la identificación de zonas calientes. Argumentan que, si bien estas herramientas son efectivas en exhibir concentraciones espaciales del crimen, es particularmente desafiante la tarea dado la dificultad en la elección de la cantidad de clusters.

En ese sentido, Agarwel et al. (2013) proponen una metodología para realizar análisis de clustering en bases de datos de crímenes mediante el uso de *K-means*. Además, Dwivedi, Pandey & Tiwari (2014) utilizan el mismo criterio para la prevención de actividades terroristas.

Este trabajo, sin embargo, se basa en el la metodología propuesta por Levine (2013). El autor describe los beneficios de utilizar *Hierarchical Clustering* sobre otros métodos de clustering. El principal beneficio que brinda es la posibilidad de no tener que decidir ex-ante la cantidad de clusters (k). Esto permite al analista hacer variar el valor de k para explorar las distintas posibles aglomeraciones, una vez aplicado el algoritmo.

Sección III: Datos

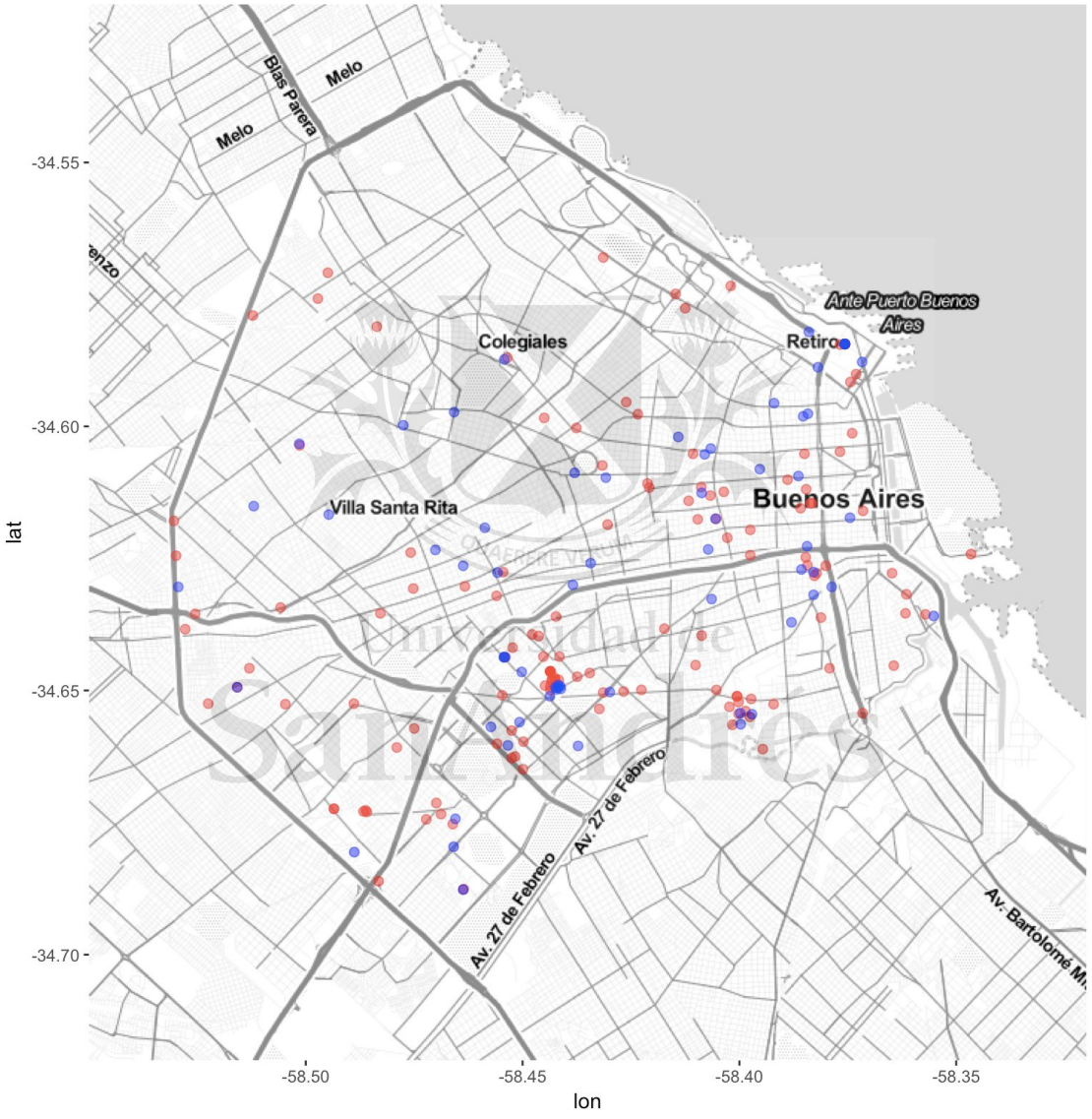
La base de datos a utilizar está compuesta por los distintos reportes de crímenes realizados en la Ciudad de Buenos Aires durante el año 2016 y el primer semestre de 2017. Es una base cuyo acceso al público general es reciente por lo que los estudios realizados con ella tienen un carácter innovador. Entre otras variables, cada observación contiene una localización geográfica exacta delimitada por la latitud y la longitud.

En lo que concierne a este trabajo se utilizará un subconjunto de está base comprendido solo por los reportes en donde el tipo de crimen es *homicidio doloso cometido en la vía pública*. Está elección se basa en el supuesto que, a diferencia de otro tipos de crímenes, el *homicidio doloso* se reporta la totalidad de las veces. Además se filtra por “cometidos en la

vía pública" dado que los homicidios cometidos en el ámbito privado no tendrían relación con zonas calientes.

En el Cuadro 1 se puede ver la geolocalización de los homicidios, en color rojo los de 2016 y en color azul los del primer semestre de 2017.

Cuadro 1: Homicidios en la Ciudad de Buenos Aires, 2016 (rojo) 2017[1] (azul).



Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

[1] Primer semestre 2017

A simple vista, es evidente cierta estructura en la distribución espacial de los puntos y cierta consistencia en el tiempo; hay aglomeramientos que coinciden en 2016 y 2017. El objetivo del análisis de clusters es poder identificar con herramientas estadísticas esta estructura.

Sección IV: Metodología y validación de resultados

Para el análisis de clusters se utilizará el método de *Hierarchical Clustering* que forma parte de los métodos de aprendizaje no supervisado; mayormente utilizado para la exploración de datos. Lo que se busca es encontrar cierta estructura en la distribución espacial de las observaciones que sea consistente con la definición de zona caliente. Con este objetivo en mente las únicas variables que se emplearán para clusterizar a las observaciones son la latitud y la longitud; el análisis se realiza sobre un espacio bidimensional.

El algoritmo que se utilizará para hacer Hierarchical Clustering es relativamente sencillo. Además, permite la visualización de los resultados en un dendrograma que resulta atractivo dado que le da libertad al analista para experimentar distintas cantidades de clusters (k).

Para empezar, se define la medida de disimilitud entre observaciones y clusters. En este caso, se utiliza la *distancia euclidiana* para comparar pares de observaciones:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,\, i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

En otras palabras, la medida disimilitud $W(C_k)$ para el k-ésimo cluster es la suma de todos los pares de distancias euclidianas al cuadrado entre las observaciones en el k-ésimo cluster, dividido por el número total de observaciones en el k-ésimo cluster.

Para medir la desemejanza entre grupos de observaciones (clusters), se utiliza el vínculo completo. Este tipo de vínculo tiene en cuenta la disimilaridad máxima dentro del cluster: calcula todas las diferencias por parejas entre las observaciones en el grupo A y las observaciones en el grupo B y registra la mayor; luego itera. El algoritmo empieza tratando a cada observación como su propio cluster $(k = n)$, esto se ve representado en la base del dendrograma. Luego, fusiona a los dos clusters que sean menos disímiles $(k = n - 1)$. Sigue fusionando los siguientes dos clusters menos disímiles $(k = n - 2)$. El algoritmo entonces continúa hasta que solo haya un único cluster $(k = 1)$, y esto se ve representado en la cima del dendrograma.

Un importante beneficio de *Hierarchical Clustering* sobre otros métodos de clustering, como *K-means*, es la libertad de no tener que decidir por adelantado el número de clusters. Esto

resulta crucial para este trabajo dado que, más que buscar una cantidad limitada de clusters, lo que se busca es una estructura en la distribución espacial de los datos.

La estrategia para la validación de los resultados es la siguiente: se separará la base de datos en dos partes; la primera contendrá las observaciones del 2016 y la segunda las del primer semestre de 2017. Se aplicará el método de *Hierarchical Clustering* sobre el subgrupo 2016 y se obtendrá el respectivo dendograma. Se harán tres cortes horizontales sobre el dendograma a distintas alturas: baja, media y alta. Luego se trazarán los clusters encontrados en el mapa de la ciudad. Este ejercicio se repetirá con con el subgrupo de 2017 y se compararán los resultados.

Al leer los dendrogramas hay que tener en cuenta que en el eje horizontal se presentan las distintas observaciones a clusterizar y en el vertical se define la altura del dendrograma medida en distancia euclidiana, que es la medida de disimilitud elegida. La altura del dendrograma determina el grado de clusterización: a mayor altura se permite mayor diferencia entre observaciones y grupos de observaciones, por ende menor cantidad de clusters (k). Es de esperar que los cortes altos sobre-agreguen las observaciones, mientras que los cortes bajos las sub-agreguen. En otras palabras, se espera que los cortes superiores en el dendrograma den como resultado clusters compuestos por más de una zona caliente y viceversa.

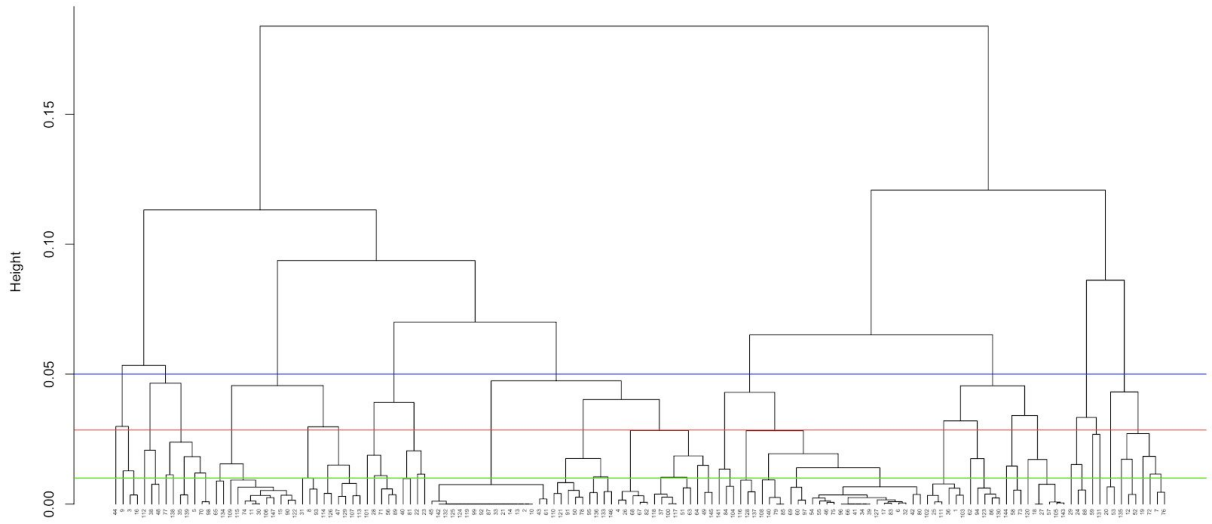
La estrategia de este trabajo contempla la elección de tres medidas de disimilitud para realizar tres cortes en los dos dendrogramas y comparar los clusters obtenidos en cada corte con sus respectivas parejas. De esta manera, se compararán los clusters producto del corte bajo del dendrograma 2016 con aquellos del corte bajo del de 2017 y así sucesivamente.

Para validar que los clusters encontrados sean zonas calientes debería ocurrir que la ubicación de los clusters en el grupo de 2016 sea igual, o similar, a la del de 2017. De esa forma, se validarían tanto la característica de espacio como la de tiempo.

Sección V: Resultados

Al aplicar el algoritmo propuesto sobre los dos grupos de datos se pueden obtener sus respectivos dendrogramas. En el Cuadro 2 se presenta el dendrograma correspondiente a los datos de 2016 y en el Cuadro 3 el dendrograma correspondiente a los datos de 2017.

Cuadro 2: Dendrograma datos 2016



Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

Cuadro 3: Dendrograma datos 2017

Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

Las líneas azules corresponden a los cortes altos (altura 0.05), las rojas a los cortes medios (altura 0.03) y las verdes a los cortes bajos (altura 0.01). Cada corte da como resultado distintas cantidades de clusters (k). De forma manual se puede obtener el valor k correspondiente contando las intersecciones entre cada corte horizontal y las líneas verticales del dendrograma. En la siguiente tabla se ilustra el valor de k correspondiente a cada corte por año.

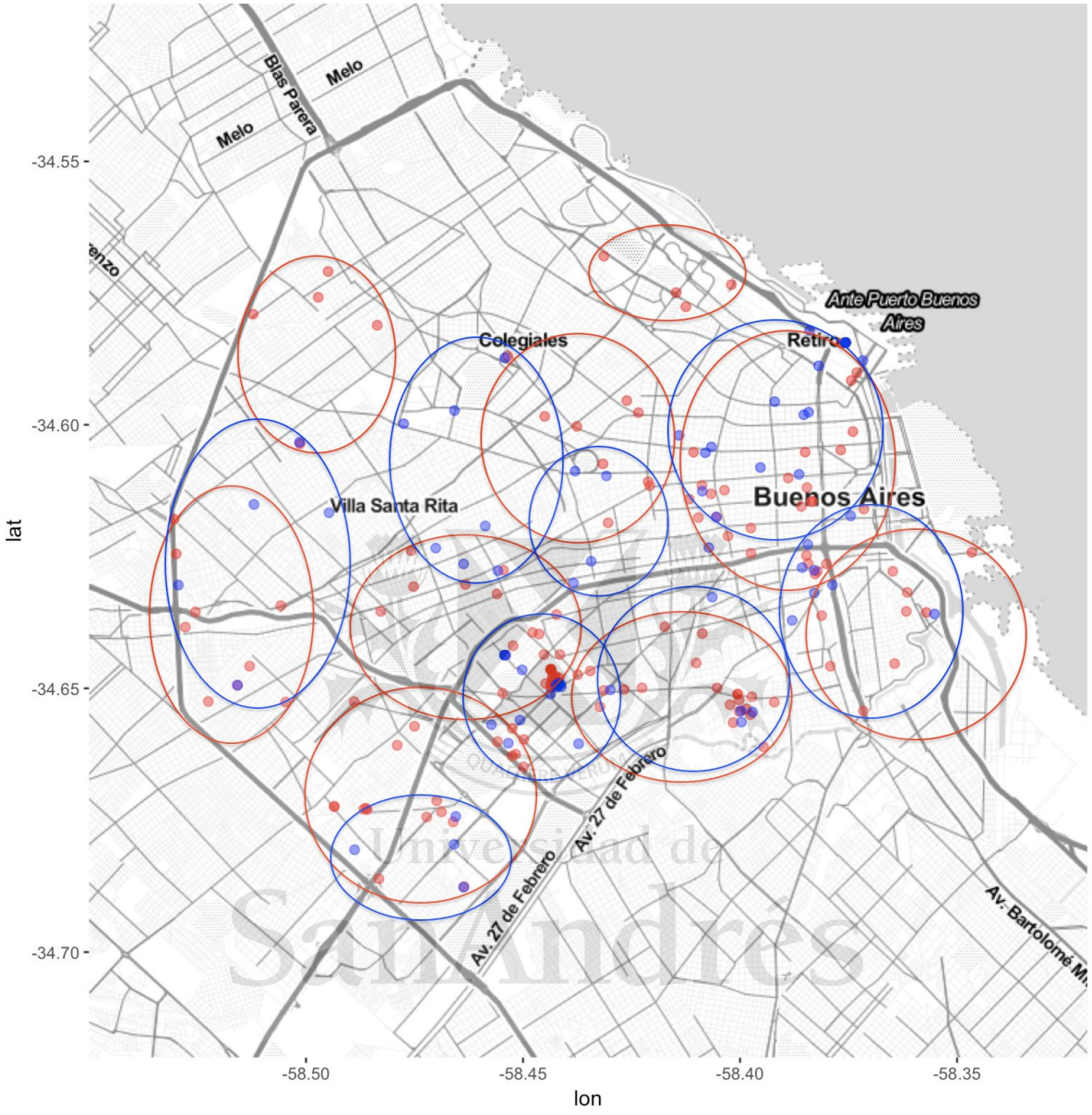
Tabla 1: Valores de k para cada corte transversal sobre los dendrogramas

	2016	2017
Corte Alto	$k = 9$	$k = 8$
Corte Medio	$k = 22$	$k = 15$
Corte Bajo	$k = 54$	$k = 32$

Se observa que para las mismas medidas de desemejanza para el año 2017 se obtiene menos cantidad de clusters en todos los cortes. Es posible que esta diferencia se deba a que solo estamos considerando datos del primer semestre del 2017 contra datos de todo el 2016.

En los siguientes 3 cuadros se muestran, superpuestos en el mapa de la Ciudad de Buenos Aires, los clusters obtenidos en cada año. Cada cuadro corresponde a la comparación de uno de los cortes sobre los dendrogramas.

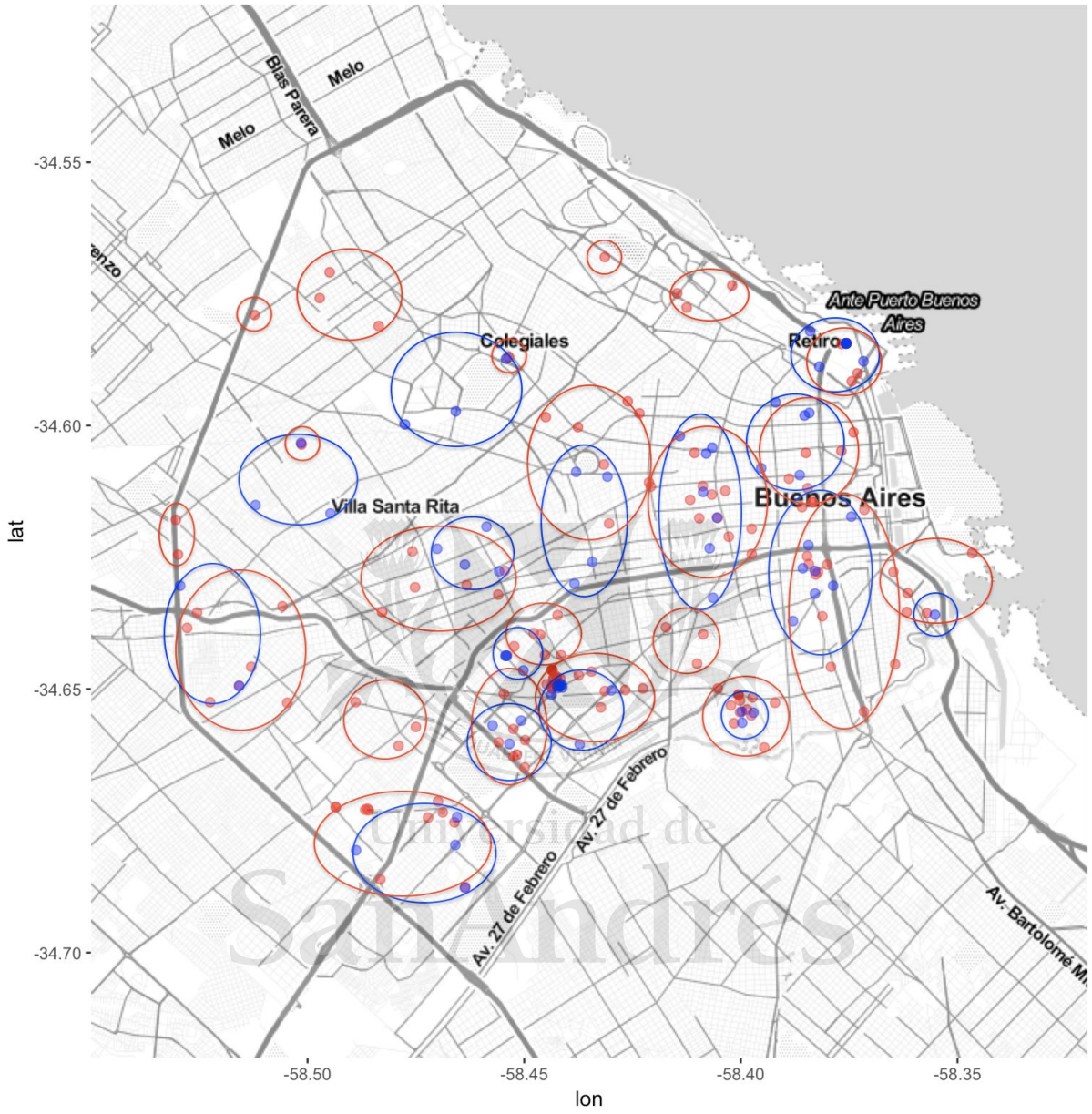
Cuadro 4: Clusters, Corte alto. Año 2016 (rojo) vs Primer Semestre 2017 (azul).



Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

Bajo el corte alto en los dendrogramas se obtienen relativamente pocos clusters compuestos por muchas observaciones. Se observa que hay cierta superposición de clusters en la zona este y a lo largo del sur de la ciudad.

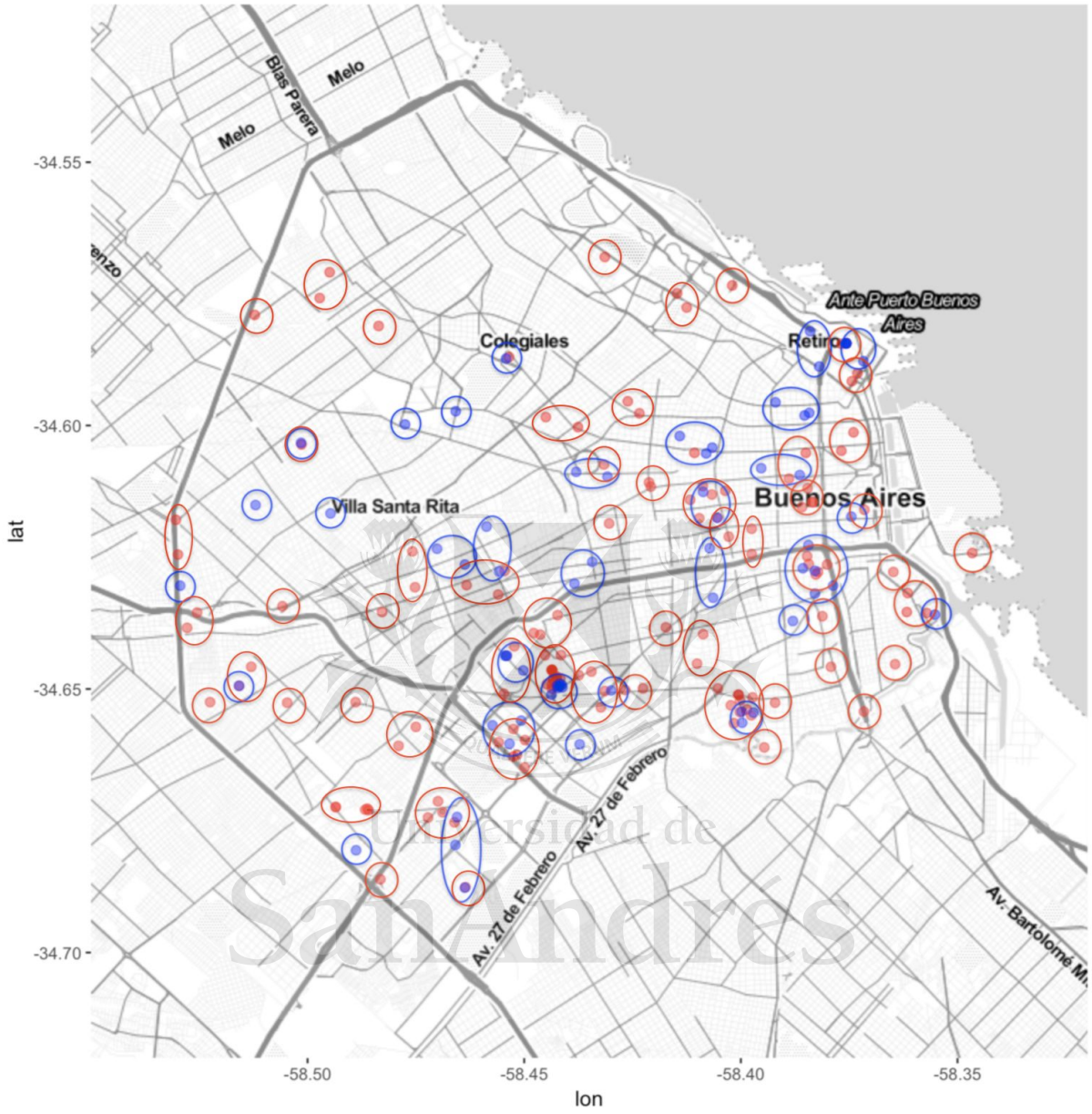
Cuadro 5: Clusters, Corte Medio. Año 2016 (rojo) vs Primer Semestre 2017 (azul).



Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

Bajo el corte medio, la cantidad de clusters obtenida es mayor comparado con el corte anterior. Estos clusters son menores en términos de cantidad de observaciones. En este caso, la superposición de clusters es más evidente. Se repite la ubicación de las superposiciones a lo largo del este, en el centro-este y a lo largo del sur de la ciudad. En particular se observan aglomeraciones en el Barrio Retiro, al noreste de la ciudad, a lo largo de la Avenida 9 de Julio y en barrios del sur de la ciudad como Flores, Barracas y Constitución.

Cuadro 6: Clusters, Corte Bajo. Año 2016 (rojo) vs Primer Semestre 2017 (azul).



Fuente: Elaboración propia en base a datos del Gobierno de la Ciudad de Buenos Aires.

Con el corte bajo, la cantidad de clusters es sustancialmente mayor en ambos años. Se observa que una gran cantidad de clusters está compuesta por una sola observación. En este caso la superposición de clusters entre años no es tan evidente como en los mapas anteriores.

Sección VI: Conclusiones

En este trabajo se aplicó el método de *Hierarchical Clustering* sobre una nueva base de datos de crímenes de la Ciudad de Buenos Aires para localizar zonas calientes de homicidios.

Se propusieron tres posibles clusterizaciones (distintos valores de k) para cada año y se compararon los resultados en búsqueda de superposición de clusters entre años. Se observó que bajo el corte más bajo sobre los dendrogramas, una gran cantidad de clusters estaba compuesta por solo una observación, lo que permitiría pensar que en ese caso ocurre una sobre-clusterizando de los datos. Sin embargo, bajo el corte medio y el corte alto se observó repetición en la ubicación de algunos de los clusters entre años.

Los resultados bajo el corte medio y el corte alto permiten inferir la ubicación de zonas calientes de homicidios en el sur y en el este de la ciudad: al haber superposición de clusters entre años se validan tanto la característica espacial como la temporal de una zona caliente. Estos resultados sugieren la existencia de barrios más peligrosos, en términos de homicidios en la vía pública, localizados a lo largo del sur y del este de la ciudad. En particular se observan observan zonas calientes en la zona sur del Barrio Flores, en Barracas, en Constitución y en el Barrio Retiro. Una posible extensión del trabajo sería explotar las herramientas econométricas espaciales para estudiar cómo los homicidios dependen de variables económicas como el desempleo y nivel socioeconómico de los barrios en las zonas calientes encontradas.

Si bien los resultados de este trabajo permiten inferir la ubicación zonas calientes en ciertos barrios de la ciudad, es necesario tener en cuenta que la cantidad de períodos de tiempo en el estudio es limitada y para probar robustez habría que repetir el ejercicio más períodos.

La detección de las zonas calientes no solo es importante para lograr una asignación más eficiente de los recursos del Estado, sino que también tiene una relevancia microeconómica dado que puede influir en las decisiones de asentamiento de las familias y las empresas.

Referencias

Baylé, F. (2016). Detección de villas y asentamientos informales en el partido de La Matanza mediante teledetección y sistemas de información geográfica.

Chalfin, A. & McCrary, J. (2017). Criminal Deterrence: A Review of Literature.

Dwivedi, S., Pandey, P. & Tiwari, M. S. (2014). Combating Terrorism Using Enhanced K-Means Clustering Algorithms.

Grubesic, T. H. & Murray A. T. (2002). Detecting Hot Spots Using Cluster Analysis and GIS.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). The Elements of Statistical Learning.

Hian, J. (2013). Spatial Clustering Methods in Data Mining: a Survey.

James, G., et al (2013). An Introduction to Statistical Learning.

Johnson, S. D., et al (2008). table and Fluid Hotspots of Crime: Differentiation and Identification.

Levine, N. (2013). A Spatial Statistics Program for the Analysis of Crime Incident Locations.

Malathi, A., et al (2011). An intelligent Analysis of a City Crime Data Using Data Mining.

McClendon, L. & Meghanatan, N. (2015). Using Machine Learning Algorithms to Analyse Crime Data.

Mohler, G. O., Short, M. B., Brantingham, P. J. & Schoenberg, F. P. (2011). Self-exciting Point Process Modeling of Crime.

VijayaKumar, M., et al (2013). The Day to Day Crime Forecasting Analysis Using Spatial-Temporal Clustering Simulation.

Weisburd, D., & Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment.