Universidad de Buenos Aires

Maestría de Explotación de Datos y Descubrimiento del Conocimiento

Análisis Inteligente de Datos

Trabajo Práctico

Matías Poullain

Primer Cuatrimestre 2021

1. Resumen

La Ciudad Autónoma de Buenos Aires (CABA) es la ciudad capital de la República Argentina. Esta subdividida en 15 unidades descentralizadas de gestión política y administrativa denominadas *comunas*. La necesidad de esta subdivisión es abordar las problemáticas territoriales particulares de forma diferencial, pero ¿es esta la forma más adecuada de subdividir a CABA para abordar la problemática de desigualdades al acceso de los servicios básicos y a la incidencia de delitos observados en la ciudad? Para contestar a esta pregunta se estudiaron, para cada manzana de CABA, su distancia a los denominados *centros de servicios básicos* (comisarías, cuarteles de bomberos, hospitales públicos, etc) más cercanos y la cantidad de delitos cometidos. Se buscó determinar si estas variables son suficientes para diferenciar a las comunas entre sí a partir del algoritmo clasificador SVM. Además se realizó una clasificación no supervisada con el algoritmo PAM a fin de proponer otra subdivisión de la ciudad más adecuada para diferenciar las necesidades particulares evidenciadas por las variables estudiadas. Si bien se obtuvo una buena performance de clasificación de las manzanas a sus respectivas comunas (exactitud: $IC_{95\,\%}$ = [0.66 ; 0.69]), la interpretabilidad de las reglas de clasificación encontradas por el algoritmo SVM es deficiente. La subdivisión más adecuada para las variables estudiadas presentó dos nuevas divisiones: una en el norte - noreste de CABA con relativa corta distancia a los centros de servicios básicos y con relativa alta incidencia de delitos y una zona sur-suroeste con características contrarias. La nueva subdivisión propuesta es más adecuada en cuanto a la clasificación de las manzanas de la ciudad y permite una mayor interpretabilidad de sus necesidades. Sin embargo entendiendo que las desigualdades socio económicas son de orígenes multifactoriales es necesario realizar un análisis más complejo a fin de optimizar la subdivisión y de abordar las problemáticas particulares de cada una zona de forma óptima.

2. Introducción

La Ciudad Autónoma de Buenos Aires (CABA) es la ciudad capital de la República Argentina. Con un estimado de poco más de 3 millones de habitantes [9], una extensión superior a los 200 km² y más de 12000 manzanas, es la mayor área urbana del país, la segunda de Sudamérica, Hispanoamérica y del hemisferio sur, y una de las 20 mayores ciudades del mundo [5]. Desde el punto de vista político-administrativo, CABA se subdivide en 15 comunas (Figura 1), unidades descentralizadas de gestión política y administrativa que, en algunos casos, abarcan a más de un barrio porteño [6]. Las Comunas tienen competencias exclusivas y concurrentes con el Gobierno de la Ciudad. Entre las primeras, se encuentran el mantenimiento de las vías secundarias y los espacios verdes, la administración de su patrimonio, la iniciativa legislativa y la elaboración de su presupuesto y programa de Gobierno.

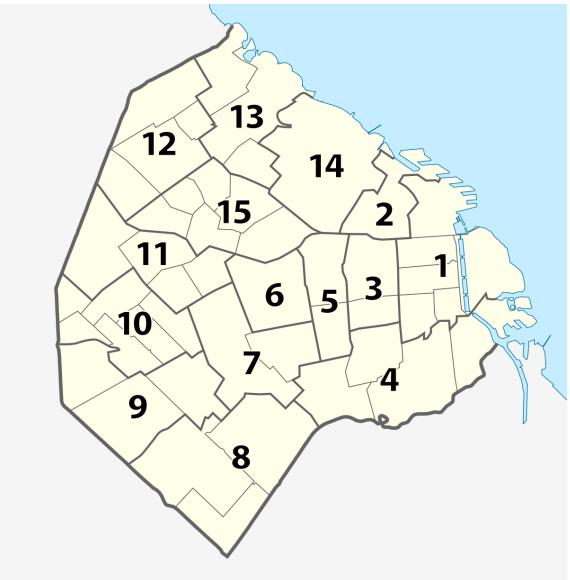

Figura 1: Comunas de CABA [14]

Entre los servicios básicos disponibles en CABA, y los que son de interés para este trabajo, se encuentran los servicios de seguridad pública (Policía y bomberos), de asistencia médica (pública y privada) y de educación (pública y privada) distribuidos por toda la ciudad. Otra característica de CABA es su alta tasa de delitos diarios con un promedio de 169 robos y 113 hurtos por día en los últimos 5 años [7]. Si bien las ubicaciones geográficas de las comisarias, cuarteles de bomberos, hospitales, centros de salud y establecimientos de educación (denominados en este trabajo como *centros de servicios básicos*) y de los delitos cometidos se encuentran repartidos en toda la ciudad, su distribución no es homogénea. La identificación de zonas dentro de CABA con diferentes niveles de accesibilidad a los servicios básicos y de delincuencia, es el primer paso para la implementación de medidas a fin de reducir las desigualdades dentro de la ciudad.

El objetivo de este trabajo fue determinar si la subdivisión de CABA en comunas se encuentra en concordancia con la accesibilidad a los servicios básicos y la incidencia de delitos. Además, se buscó si existe una subdivisión más acorde a estas características.

3. Datos

En este trabajo se utilizaron distintas bases de datos de acceso público, todas con información georeferenciada. A continuación se enumeran las bases de datos utilizadas:

- Límites intercomunales de la ciudad [10].
- Límites de las manzanas de la ciudad [19]
- Posición geográfica de los hospitales públicos de la ciudad [18].
- Posición geográfica de los cuarteles y destacamentos de bomberos de la ciudad (a partir de ahora *cuarteles*) [16].
- Posición geográfica de los establecimientos educativos (educación primaria, secundaria, terciaria, superior o especial) [17].
- Posición geográfica de los centros de salud privados de la ciudad [21].
- Posición geográfica de las comisarías de la ciudad [15].
- Posición geográfica de los delitos ocurridos en la ciudad entre el 2016 y el 2020 [13]

Además de las posiciones geográficas, cada base de datos contó con información descriptiva de cada una de las observaciones. Para cada manzana de la ciudad se calcularon las variables, descriptas en la siguiente sección, utilizando diferente información de todas las bases de datos enumeradas.

4. Materiales y métodos

En primer lugar se realizó un tratamiento de las áreas de las manzanas de la ciudad. Siendo que las calles no pertenecían a ninguna manzana, se determinaron las coordenadas de los centroides de cada manzana y a cada posición en el mapa se le asoció su centroide de manzana más cercano (a partir de la construcción de polígonos de Voronoi). Se obtuvo un polígono por manzana de la ciudad y con alta superposición entre el polígono y su manzana correspondiente, por lo que se consideró que los polígonos son una buena representación de sus manzanas y se los nombrará *manzanas*.

En segundo lugar, se acotó el área de estudio a CABA a partir de la base de los límites intercomunales. Además a cada manzana se le asoció una comuna según su mayor superposición geográfica.

En tercer lugar, a partir de la posición geográfica de los hospitales públicos de la ciudad, se determinó para cada manzana, la distancia semiverseno [20] al hospital más cercano. Para lograrlo se construyeron los polígonos de Voronoi para cada hospital y a cada centroide de manzana se le asoció un hospital según la superposición del punto sobre tal polígono. Luego se calculó la distancia del centroide de cada manzana a su hospital más cercano asociado. Se realizó el mismo procedimiento para los cuarteles, los establecimientos educativos, los centros de salud privados y las comisarías. Dada la importante asimetría hacia el 0 Km (sin alcanzarlo) observada en los valores de distancia a los centros educativos, esta variable fue transformada con con la función logaritmo en base 10.

En último lugar, utilizando la base de datos de delitos, se contaron el número de homicidios (siniestros viales, femicidios, dolosos o trasvesticidios/transfemicidios), hurtos (sin violencia), robos (con violencia) y lesiones (por siniestros viales) ocurridos en cada manzana. Todas las variables utilizadas en las clasificaciones se muestran en la Tabla 1

	Variable	Media	Mediana	IQR	Mínimo	Máximo
1	Distancia a Est. Salud Privado	1.77	1.20	[0.64 ; 2.37]	0.01	7.40
2	Distancia a Hospital Público	1.33	1.22	[0.78 ; 1.79]	0.03	3.89
3	Distancia a Comisaría	0.94	0.88	[0.56 ; 1.26]	0.01	2.80
4	Distancia a Cuartel	1.19	1.15	[0.77 ; 1.53]	0.01	3.47
5	Distancia a Est. Educativo (Log)	-1.92	-1.85	[-2.35 ; -1.46]	-5.20	0.37
6	Homicidios	0.02	0.00	[0 ; 0]	0.00	6.00
7	Hurtos	2.05	1.00	[0 ; 3]	0.00	54.00
8	Lesiones	0.48	0.00	[0 ; 1]	0.00	43.00
9	Robos	2.75	1.00	[0 ; 4]	0.00	139.00

Tabla 1: Métricas resumen de las variables utilizadas. Todas las distancias se informan en kilómetros.

Se calcularon métricas resumen de las variables utilizadas estratificadas por comuna y se presentaron las observaciones en formato mapas.

El abordaje de los objetivos se realizó a partir de dos clasificaciones (una supervisada y otra no supervisada) de las manzanas de CABA. En el caso de la clasificación supervisada, su objetivo fue determinar si las manzanas de la ciudad pueden ser clasificadas correctamente en sus respectivas comunas. Se utilizó el algoritmo clasificador de Máquinas de Soporte Vectorial (SVM) [3] y se optimizaron los hiperparámetros con una validación cruzada de 10 iteraciones sobre el conjunto de entrenamiento (75 % de las manzanas) y se estudió la performance del clasificador sobre el conjunto de validación. La métrica que se buscó maximizar en la determinación de los hiperparámetros fue el promedio de la exactitud de la clasificación de cada nivel de la variable respuesta (comuna) ponderada por la inversa de la frecuencia manzanas en la comuna, a fin de contrarrestar a influencia del desbalance de los niveles de la variable respuesta observados.

En el caso de la clasificación no supervisada, su objetivo fue el de realizar una nueva subdivisión de las manzanas de CABA según las variables utilizadas. Para lograrlo se realizó la clasificación con la totalidad de los datos a partir del algoritmo PAM [11]. Se determinó el número de clusters óptimo a partir del método de Silhouette [12].

5. Resultados

Se estudiaron 12487 manzanas en CABA. La comuna con menos manzanas fue la Comuna 2 con 325 observaciones y aquella con más fue la 15 con 955. En la Figura 2 se presentaron las cuatro variables que, a simple vista, parecieran ser las que presentan mayores diferencias entre comunas. Se pudo observar que las distancias a los centros de servicios básicos suelen ser relativamente grandes en las comunas 8 y 9 (las más australes de la ciudad) mientras que aquellas con menores distancias son las comunas 2 a 6 (posicionadas en el centro-este de la ciudad). Por otro lado, las comunas cuyas manzanas presentaron la menor cantidad de hurtos fueron las comunas 4, 9, 10, 11 y 12. Las manzanas con mayor cantidad de hurtos parecieran agruparse en el noreste de la ciudad. Sin embargo, en todas las comunas se destacaron manzanas con cantidad de hurtos muy superiores a los valores más comunes observados.

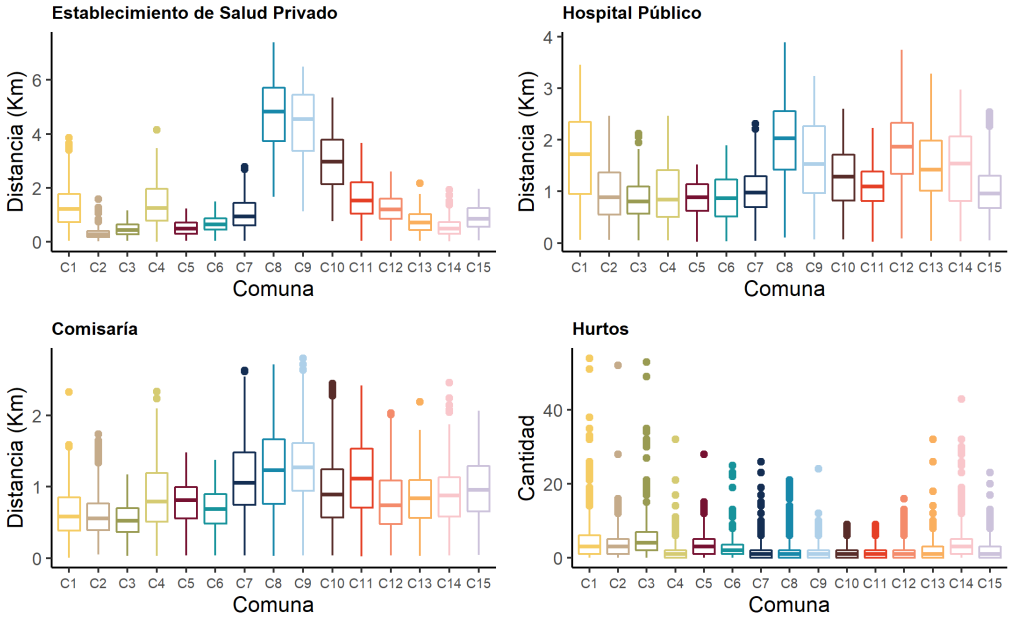

Figura 2: Boxplots de los valores de cuatro de las variables utilizadas en el trabajo

5.1. Clasificación Supervisada

Se determinó que los hiperparámetros óptimos para la inicialización del algoritmo SVM fueron un *Kernel radial* con un *Gamma* de 0.3. La validación sobre el conjunto de test presentó una exactitud de 0.67 ($IC_{95\,\%}$: [0.66 ; 0.69]) y un índice Kappa de 0.65. La comuna que mejor fue identificada fue la 8 (sensibilidad: 0.87) mientras que la peor identificada fue la 3 (sensibilidad: 0.41). Los resultados de la clasificación de todas las manzanas se muestran en la Figura 3.

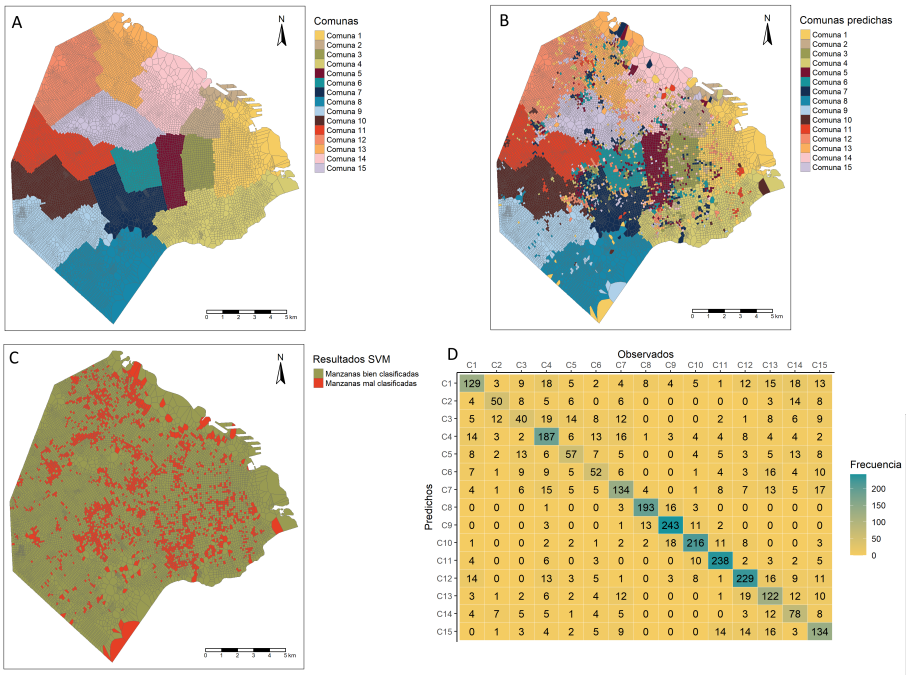

Figura 3: Resultados de la clasificación supervisada. A: Comunas observadas para cada manzana. B: Comunas predichas para cada manzana. C: Manzanas cuyas comunas fueron bien y mal clasificadas. D: Matriz de confusión de la clasificación

5.2. Clasificación No Supervisada

Se determinó que el número óptimo de clusters fue de dos . Las clasificaciones de las manzanas en dos clusters se presentaron en la Figura 4 y las diferencias observadas de los valores de las variables estudiadas entre clusters se mostraron en la Tabla 2.

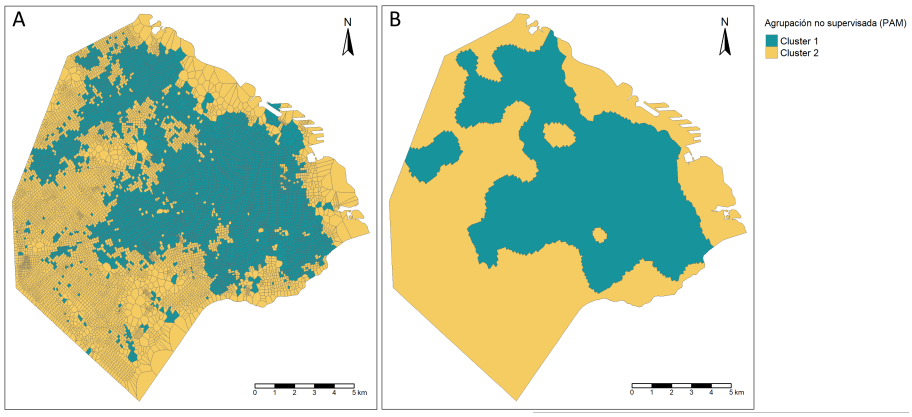

Figura 4: Clusterización realizada por el algoritmo PAM. A: Resultados obtenidos. B: Resultados obtenidos tras suavizado espacial

El Cluster 1 fue caracterizado por presentar manzanas con mayor cantidad de hurtos y robos y menores distancias a los establecimientos de salud privados, a los hospitales públicos, a las comisarias y a los establecimientos educativos más cercanos en comparación al Cluster 2.

	Cluster 1	Cluster 2
n	6008	6479
Homicidios (Mediana [IQR])	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Hurtos (Mediana [IQR])	2.00 [1.00, 4.00]	1.00 [0.00, 2.00]
Lesiones (Mediana [IQR])	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]
Robos (Mediana [IQR])	3.00 [1.00, 5.00]	1.00 [0.00, 2.00]
Distancia a Est. Salud Privado (Mediana [IQR])	0.74 [0.43, 1.21]	2.10 [1.15, 3.76]
Distancia a Hospital Público (Mediana [IQR])	0.99 [0.62, 1.46]	1.47 [0.99, 2.09]
Distancia a Comisaría (Mediana [IQR])	0.61 [0.41, 0.86]	1.19 [0.87, 1.52]
Distancia a Cuartel (Mediana [IQR])	1.02 [0.66, 1.36]	1.30 [0.89, 1.73]
Distancia a Est. Educativo (Log) (Mediana [IQR])	-2.19 [-2.67, -1.83]	-1.54 [-1.91, -1.25]

Tabla 2: Métricas resumen de las variables utilizadas estratificadas por los clusters generados por el método PAM. Todas las distancias se informan en kilómetros.

6. Discusión

Este trabajo presenta una visión innovadora frente al abordaje de la problemática de las desigualdades al acceso de los servicios básicos dentro de CABA. Entendiendo que la desigualdad socio económica es un proceso multifactorial, la simplificación del análisis de la problemática puede generar conclusiones sesgadas. Sin embargo, a través del estudio de la distancias a los centros de servicios básicos más cercanos y de la cantidad de delitos por cuadra, se obtuvieron resultados interesantes.

En cuanto a la clasificación supervisada, se pudo obtener una performance de la clasificación de las manzanas a sus respectivas comunas muy alta, con una exactitud cercana al 70 %, mejorando a un clasificador por azar, de exactitud cercana a 7.5 % para este conjunto de datos, en un 800 %. Este resultado indicaría que las variables estudiadas tienen la capacidad de diferenciar en gran medida a las comunas entre sí. Sin embargo, la performance tan alta alcanzada en la clasificación tuvo como costo la pérdida de poder explicativo. Es decir que las características distintivas de cada comuna detectadas por el algoritmo SVM no son fácilmente describibles en las escalas reales de las variables. Esto se debe a las transformaciones de los valores intrínsecas al algoritmo utilizado para la clasificación que permiten una alta performance a costa de la interpretabilidad de los resultados [3]. Por otro lado, se pudo observar que existen comunas más fácilmente identificables que otras. La más reconocible, mediante las variables estudiadas, pareciera ser las Comunas 8 y 9. Esta observación fue advertida a simple vista en la Figura 2 y detectada por la clasificación por SVM (Figura 3). Estas comunas pertenecen a la zona sur de CABA, históricamente asociada a un menor nivel socio económico en comparación al resto de la ciudad [2]. Las manzanas de estas comunas presentaron mayores distancias a los centros de servicios más cercanos y una baja cantidad de delitos en comparación a las manzanas de las otras comunas. Esta observación se encuentra en concordancia con la bibliografía, explicando que, si bien las zonas más carenciadas se encuentran asociadas a una mayor prevalencia de incidencia en al criminalidad, los actos delictivos suelen no ser cometidos en las inmediaciones vecinales del perpetrador, sino a una distancia mayor [22]. Otro factor que podría influir en este resultado es la densidad poblacional, la cual no es homogénea en el área de CABA. Ésta es mayor en las comunas 2, 3, 5 y 6 y menor en las 8 y 9 [4]. Si bien este aspecto no fue abordado en este trabajo, sería interesante la incorporación de la densidad poblacional en trabajos próximos, reconociendo la asociación previamente documentada de la densidad poblacional con la incidencia de delitos [1].

En cuanto a la clasificación no supervisada, se determinó que la mejor clusterización esta compuesta por dos clusters. Si bien se pudo observar una importante agrupación espacial de las manzanas pertenecientes al mismo cluster, esta no es perfecta (Figura 4). Sin embargo, la clasificación pareciera indicar, a grandes rasgos, una diferenciación entre la zona norte-noreste (Cluster 1) de la ciudad contra la zona sur y costera (Cluster 2). Este resultado se encuentra en concordancia con lo observado en la clasificación supervisada: El Cluster 1 se caracteriza por presentar manzanas con mayor cantidad de delitos y menores distancias a los centros de servicios básicos en comparación al Cluster 2. Estos resultados podrían explicarse por los mismos factores que los propuestos en el párrafo anterior. Sin embargo, la mayor diferencia entre ambas clasificaciones es el poder explicativo de la no supervisada. Esta permite identificar de manera más sencilla los problemas más comunes para cada manzana y es de mayor ayuda al momento de plantear políticas dirigidas a combatir las problemáticas únicas de cada cuadra.

Si bien los datos utilizados en este trabajo permitieron un desarrollo adecuado para la problemática,

la información de los delitos utilizada representa únicamente los delitos reportados y no los ocurridos realmente. Se describió que en CABA la tasa de no denuncia de delitos supera el 70 % [8]. Sin embargo, al no haber encontrado bibliografía que defienda que la no denuncia de delitos se ve afectada por factores externos, en este trabajo se considera que los delitos reportados conforman una muestra aleatoria de los delitos reales y, por lo tanto, los resultados extraídos del análisis de los mismos no se encuentran severamente sesgados.

7. Conclusión

Los resultados de este trabajo evidenciaron inequidades en el acceso a servicios básicos e incidencia de delitos en las manzanas de CABA. Si bien la división actual de la ciudad en comunas podría ser explicada en gran medida por la lejanía a los centros de servicios básicos y la cantidad de delitos cometidos por manzana, su relación no es sencilla de explicar y, en consecuencia, se dificulta la implementación de medidas cuyo objetivo es abordar las problemáticas particulares y combatir la desigualdad dentro el territorio. Por esta razón, se propuso una nueva división interna de la ciudad, la cual agrupa de manera más adecuada a las manzanas en función de las dimensiones estudiadas y, en consecuencia, de las diferencias a tener en cuenta para reducir la desigualdad evidenciada en las variables estudiadas.

Entendiendo que las desigualdad social es un proceso multifactorial y que en este trabajo solo se abordaron algunos de tales factores relacionados, no se asegura que la subdivisión propuesta sea la alternativa más eficiente para erradicar la desigualdad socio económica, sino que propone una forma de división más adecuada que podría ayudar a abordar la problemática de una forma diferente. A fin de obtener divisiones más eficientes para la problemática propuesta, es necesario aumentar la complejidad del análisis utilizando otros factores tales como la densidad poblacional.

Referencias

[1] Joshua R Battin y Justin N Crowl. "Urban sprawl, population density, and crime: an examination of contemporary migration trends and crime in suburban and rural neighborhoods". En: *Crime prevention and community safety* 19.2 (2017), págs. 136-150.

[2] Centro de Estudios Metropolitanos (CEM). *Las Desigualdades en la Ciudad de Buenos Aires*. 2020. URL: http://estudiosmetropolitanos.com.ar/wp-content/uploads/2020/04/Radiograf%7B%5C'%7Bi%7D%7Da-de-las-desigualdades-en-la-Ciudad-de-Buenos-Aires.pdf.

[3] Corinna Cortes y Vladimir Vapnik. "Support-vector networks". En: *Machine learning* 20.3 (1995), págs. 273-297.

[4] Gerencia Operativa de Ingeniería de Datos. DG Arquitectura de Datos. SS de Políticas Públicas Basadas en Evidencia. Secretaría de Innovación y Transformación Digital. Jefatura de Gabinete de Ministros. *Estructura de la Población*. 2021. URL: https://data.buenosaires.gob.ar/dataset/estructura-poblacion.

[5] Gobierno de la Ciudad Autónoma de Buenos Aires. *Ciudad de Buenos Aires*. Inf. téc. 2021. URL: https://www.buenosaires.gob.ar/laciudad/ciudad.

[6] Gobierno de la Ciudad Autónoma de Buenos Aires. *Comunas*. Inf. téc. 2021. URL: https://www.buenosaires.gob.ar/comunas.

[7] Gobierno de la Ciudad Autónoma de Buenos Aires. *Robos y hurtos registrados, y distribución porcentual y promedio diario por comuna de ocurrencia. Ciudad de Buenos Aires. Años 2016/2020*. 2021. URL: https://www.estadisticaciudad.gob.ar/eyc/?p=101227.

[8] INDEC. *Encuesta Nacional de Victimización 2017*. Inf. téc. 2018. URL: https://www.indec.gob.ar/uploads/informesdeprensa/env20170218.pdf.

[9] INDEC. *Población - Proyecciones y Estimaciones*. Inf. téc. 2010. URL: https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-24-85.

[10] Instituto Geográfico Nacional. *Departamentos de la República Argentina*. 2019. URL: https://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG.

[11] Leonard Kaufman y Peter J Rousseeuw. "Clustering large applications (Program CLARA)". En: *Finding groups in data: an introduction to cluster analysis* (2008), págs. 126-146.

[12] Leonard Kaufman y Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[13] Ministerio de Justicia y Seguridad. *Delitos*. 2021. URL: https://data.buenosaires.gob.ar/dataset/delitos.

[14] NordNordWest. *Mapa de las Comunas de la Ciudad de Buenos Aires*. Inf. téc. Wikipedia, 2011. URL: http://commons.wikimedia.org/wiki/User:NordNordWest.

[15] Secretaría General y Relaciones Internacionales. Subsecretaría Gestión Estratégica y Calidad Institucional. Dirección General Calidad Institucional y Gobierno Abierto. *Comisarías Policía de la Ciudad*. 2021. URL: https://data.buenosaires.gob.ar/dataset/comisarias-policia-ciudad.

[16] Secretaría General y Relaciones Internacionales. Subsecretaría Gestión Estratégica y Calidad Institucional. Dirección General Calidad Institucional y Gobierno Abierto. *Cuarteles y Destacamentos de Bomberos*. 2021. URL: https://data.buenosaires.gob.ar/dataset/cuarteles-destacamentos-bomberos.

[17] Secretaría General y Relaciones Internacionales. Subsecretaría Gestión Estratégica y Calidad Institucional. Dirección General Calidad Institucional y Gobierno Abierto. *Establecimientos Educativos*. 2021. URL: https://data.buenosaires.gob.ar/dataset/establecimientos-educativos.

[18] Secretaría General y Relaciones Internacionales. Subsecretaría Gestión Estratégica y Calidad Institucional. Dirección General Calidad Institucional y Gobierno Abierto. *Hospitales*. 2021. URL: https://data.buenosaires.gob.ar/dataset/hospitales.

[19] Secretaría General y Relaciones Internacionales. Subsecretaría Gestión Estratégica y Calidad Institucional. Dirección General Calidad Institucional y Gobierno Abierto. *Manzanas*. 2021. URL: https://data.buenosaires.gob.ar/dataset/manzanas.

[20] R.∼W. Sinnott. "Virtues of the Haversine". En: 68.2 (1984), pág. 158.

[21] Subgerencia de Información Geoespacial. DG Ciencias de la Información. SS de Políticas Públicas Basadas en Evidencia. Secretaría de Innovación y Transformación Digital. Jefatura de Gabinete de Ministros. "Centros de Salud Privados". En: (2021). URL: https://data.buenosaires.gob.ar/dataset/centros-salud-privados.

[22] Matt Vogel y Scott J South. "Spatial dimensions of the effect of neighborhood disadvantage on delinquency". En: *Criminology* 54.3 (2016), págs. 434-458.