Segregación residencial en la Ciudad de Buenos Aires: un enfoque de clústers dinámicos, 1980-2022.

Propuesta Final - Grupo N°2
Valentina Fuentes Mortensen & Guadalupe Gorostiaga
Big Data y Aprendizaje Automático para Economistas 2024

1. Introducción

La segregación es la distribución desigual de la población en el espacio físico en función de una o más características compartidas. La segregación no es necesariamente un problema en sí misma; según Linares et al. (2016) en su revisión de literatura, Durkheim (1967) sugiere que puede ser una forma de integración social, en la medida en que la separación espacial de los grupos se asocia a vínculos definidos por los individuos de una comunidad. En esa línea, Sabatini (2003) señala que la segregación étnica puede ser positiva tanto para preservar culturas minoritarias como para enriquecer las ciudades, que se vuelven más cosmopolitas. Sin embargo, la segregación también puede tener efectos negativos cuando conduce al aislamiento de poblaciones vulnerables, lo que reduce sus posibilidades de movilidad social al limitar la interacción con otros grupos sociales, el acceso a empleos y, en particular a servicios clave como eduación y salud de calidad, transporte y seguridad (Suárez, 2007). En ese sentido, una población segregada intensifica o agrava diferencias, dificultando el camino hacia una sociedad más integrada y equitativa.

La segregación residencial ha despertado un gran interés académico, aunque los países han puesto su atención en distintas dimensiones de acuerdo a sus propios contextos. En los países desarrollados, el foco se ha puesto en la segregación étnica y racial (como en Estados Unidos), y en la concentración espacial de inmigrantes (en países europeos). En América Latina, predominó la segregación en función del nivel socioeconómico. Esta propuesta de investigación busca

integrar ambas perspectivas en el caso de la Ciudad Autónoma de Buenos Aires (CABA), una de las principales ciudades de Latinoamérica e histórica receptora de grandes flujos migratorios, especialmente de países limítrofes durante las últimas décadas. Dado el alto nivel de dinamismo de la ciudad, a la hora de explorar la segregación resulta relevante tomar en cuenta no sólo la dimensión puramente socioeconómica, sino también cómo esta se vincula con otros factores sociodemográficos.

Esta propuesta se centra en estudiar la presencia de segregación en los sectores residenciales de CABA a través de diversas dimensiones, tales como educación, salud, calificación de la ocupación, edad, género, condición de migrante y planificación familiar, y analizar su evolución en el tiempo entre 1980 y 2022.

Entre las contribuciones de esta propuesta se destacan, en primer lugar, la extensión de la evidencia existente tanto en el marco temporal como en las dimensiones exploradas[1]. En segundo lugar, la aplicación de una metodología innovadora de clústers dinámicos, que involucra técnicas de Machine Learning, la cual fue implementada previamente en un estudio de segregación residencial en Berlín (Masías et al. 2024). Esta metodología permitirá analizar la existencia y las características de clústers en CABA, así como los cambios en cantidad y composición de dichos clústers a lo largo del tiempo y el espacio.

2. Revisión de Literatura

La segregación, vista como la relación entre sociedad y espacio, es un concepto vigente y ampliamente estudiado para caracterizar a las grandes ciudades (Cuenya, 2018). Sin embargo, en

[1] Ver más detalle en revisión de literatura.

el caso particular de CABA, aún existen brechas en la investigación, especialmente en cuanto a la posibilidad de extender los períodos analizados y aplicar nuevas técnicas de clustering más avanzadas. El Cuadro 1 resume la literatura empírica existente sobre la organización espacial de entornos urbanos en ciudades de Buenos Aires, señalando el marco temporal de su análisis, los factores que cada uno identifica como determinantes de la segregación, la metodología empleada y sus principales aportes.

La literatura existente se concentró en el análisis del período 1991-2001, presentándose la oportunidad de ampliar el estudio tanto hacia atrás, con los datos del Censo de 1980, como hacia adelante en el tiempo, incluyendo los de 2010 y 2022. Por otro lado, las dimensiones exploradas se han centrado principalmente en el nivel educativo o la calificación de los jefes de hogar, utilizados como proxy de ingresos. Sin embargo, Groisman y Suárez (2006) amplían estas dimensiones al incluir otras características, como la condición de migrante de países limítrofes y la cobertura de salud, ambas vinculadas también al jefe de hogar. En este sentido, se propone agregar las dimensiones de edad, género y planificación familiar al análisis, así como ampliar la consideración de las características previamente analizadas, por ejemplo, expandiendo el universo de migrantes para incluir no solo a los provenientes de países limítrofes, sino a aquellos de otros países de origen. Además, el análisis no se limitaría al nivel del hogar, centrándose únicamente en el jefe de hogar, sino que se ampliaría al nivel individual. Estas consideraciones se detallan en la sección 3.

Cuadro 1: Resumen de la evidencia existente para los estudios de segregación espacial en CABA.

Autores	Marco Temporal	Dimensiones	Método	Fuente de datos	Principales hallazgos
Rodríguez (2008)	1991, 2001	Máximo Nivel de Instrucción del Jefe de Hogar	Tres tipos de Índices: de Disimilitud, de Segregación, de Aislamiento y de Interacción. Coeficiente de Localización (QL) e Índice de Continuidad (IC)	Censo Nacional de Población y Vivienda 1991. Censo Nacional de Población, Hogares y Viviendas 2001	Segregación de los jefes de hogar con nivel de instrucción bajo (y medio bajo) en la zona sur de la ciudad que aumenta entre 1991 y 2001. Disminuye la segregación de los grupos de nivel alto (y medio alto).
Groisman & Suárez (2006)	1991, 2001	-Nivel educativo del Jefe de Hogar - Proporción de Jefes de Hogar migrantes de países limítrofes - Cobertura de salud del Jefe de Hogar	Tres tipos de Índices: de Disimilitud de Duncan, de Aislamiento o exposición de Bell y de autocorrelación espacial (Moran Local y Moran Global)	Censo Nacional de Población y Vivienda 1991. Censo Nacional de Población, Hogares y Viviendas 2001	Se verifica un aumento de la segregación entre 1991 y 2001 en las dimensiones de cobertura de salud de los jefes de hogar y a la condición migratoria del jefe de hogar; aumentó el aislamiento de los jefes con elevado nivel de educación –con estudios universitarios completos–. Aun cuando no hubo un incremento en los índices de segregación, existe una marcada polarización territorial (norte-sur), particularmente en el nivel educativo del jefe.
Linares et al. (2016)[2]	2001	Calificación ocupacional en base al Clasificador Nacional de Ocupaciones de INDEC (2001)	Índice H de teoría de la información (Theil, 1972)	Censo Nacional de Población, Hogares y Viviendas 2001	Cuanto mayor es la población urbana, mayor es la segregación espacial y la calidad de vida disminuye.

Fuente: Elaboración propia.

Por último, en todos los estudios previos se emplearon métodos tradicionales y estáticos de clustering y segregación. En contraste, el estudio reciente de Masías et al. (2024) desarrolla una metodología secuencial de clústers dinámicos para analizar la segregación residencial en Berlín, considerando las dimensiones de condición migratoria, edad, sexo y características socioeconómicas durante el período 2009-2020. Para esta propuesta de investigación, se adoptará esta metodología, la cual se detalla en la sección 4.

[2] Este trabajo no mide la segregación espacial en CABA, sino en las ciudades de Tandil y Mar del Plata, ambas pertenecientes a la provincia de Buenos Aires; sin embargo, resulta pertinente presentar este antecedente metodológico por su cercanía con CABA.

3. Datos

Para llevar adelante este análisis, necesitamos datos que cubran todas las dimensiones de interés, se puedan georreferenciar y presenten un nivel adecuado de desagregación. Es por esto que se propone emplear datos provenientes de los censos de población a nivel de radio censal, para los siguientes años: 1980, 1991, 2001, 2010 y 2022.

La Figura 1 presenta los radios censales de CABA en 2010. En Argentina, la digitalización cartográfica comenzó a implementarse a partir del Censo de 1991. Sin embargo, estos límites censales no son homogéneos en el tiempo, y no existen registros digitales para el censo de 1980. El trabajo de Rodríguez (2021) resuelve este problema mediante la digitalización de los límites censales de 1970 y 1980 para CABA, además de homogeneizar la cartografía digital de los censos de 1970, 1980, 1991, 2001 y 2010. Esta contribución será de gran utilidad para esta propuesta de investigación.

Figura 1: Radios censales de CABA en 2010

Fuente: Elaboración propia en base a Ministerio de Economía y Finanzas y DG de Estadística y Censos.

En el Cuadro 2 se detallan las dimensiones a analizar, junto con las variables correspondientes del censo, la población alcanzada, la clasificación de cada dimensión en categorías y consideraciones adicionales.

Cuadro 2: Descripción de las dimensiones consideradas para la segregación residencial en CABA.

Dimensión	Variable censal	Universo	Categorías	Comentario
Educación	Máximo nivel educativo alcanzado	PEA y personas de 65 años o más	- Bajo (primaria incompleta, primaria completa) - Medio-bajo (secundaria incompleta) - Medio-alto (secundaria completa, universitaria incompleta) - Alto (universitaria completa)	Esta variable sirve como una proxy de ingresos o nivel socioeconómico, por lo que se excluyen niños, adolescentes y adultos jóvenes que no trabajan ni están buscando trabajo. Sin embargo, no se excluye a adultos mayores porque, pese a que pueden ser parcialmente dependientes, se considera que su nivel educativo puede aproximar su nivel socioeconómico actual en función de lo generado durante su etapa de actividad.
Calificación de la ocupación	Descripción de la ocupación	Ocupados	Según Clasificador Nacional de Ocupaciones de INDEC (2017): - Calificación profesional - Calificación técnica - Calificación operativa - No calificados	Se considera una medición objetiva. Esta es otra forma de aproximar los ingresos o el nivel socioeconómico y se puede contrastar si existen diferencias con los resultados obtenidos utilizando la dimensión de Educación.
Salud	Cobertura de salud	Toda la población	- Cobertura privada: obra social o prepaga (incluye PAMI) - Cobertura pública	Esta variable no se encuentra disponible en el censo de 1980.
Condición de migrante	País de nacimiento	Toda la población	Según país de nacimiento	
Edad	Años cumplidos	Toda la población	- de 0 a 10 - de 10 a 20 - de 20 a 30 - de 30 a 40 - de 40 a 50 - de 50 a 60 - de 60 a 70 - de 70 de 80 - de 80 a 90 - más de 90	
Sexo	Hombre o mujer	Toda la población	- Hombre - Mujer	
Planificación familiar	Cantidad de hijos nacidos vivos	Toda la población	- Sin hijos (0 hijos nacidos vivos) - Pocos hijos (1-2 hijos nacidos vivos) - Familia mediana (3-4 hijos nacidos vivos) - Familia numerosa (5-6 hijos nacidos vivos) - Familia muy numerosa (7 o más hijos nacidos vivos)	

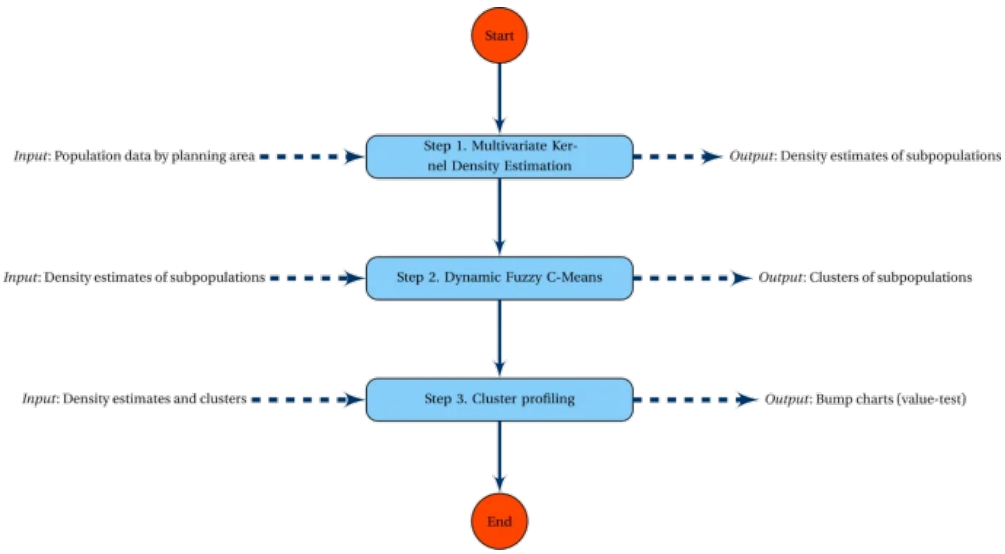
Fuente: Elaboración propia.

En Argentina, las bases de microdatos completas de los censos no son de acceso público, por lo que es necesario solicitar la información al Instituto de Estadística y Censos de la Ciudad Autónoma de Buenos Aires (IDECBA) o al INDEC. La única excepción es la base de microdatos del censo de 1980, que ha sido recuperada y puesta a disposición en acceso libre por Rodríguez (2023).

4. Metodología

La metodología propuesta se basa en la aplicación del enfoque de clústers dinámicos, desarrollado por Masías et al. (2024), para el caso de CABA. Este enfoque se divide en tres etapas, las cuales se resumen en la Figura 2.

Figura 2: Etapas de la metodología de análisis dinámico para la segregación residencial.



Fuente: Masías et al. (2024)

La primera etapa consiste en estimar densidades poblacionales para cada dimensión considerada, para lo que se usa un método de estimación no paramétrico de densidad de kernel multivariante. Este método está diseñado para estimar densidades poblacionales en áreas de formas y tamaños

arbitrarios, como lo son los radios censales, cuando sólo se dispone de pocas observaciones. La idea es estimar la densidad en cada punto (dado por coordenadas) de un área sumando las "contribuciones" de todos los puntos de datos cercanos; estas contribuciones son funciones kernel que asignan valores más altos a los puntos cercanos y más bajos a los lejanos, logrando una estimación suavizada de la densidad. Dado que contamos con datos agregados a nivel de radio censal en lugar de las coordenadas exactas de cada observación, este puede modelarse como un error de medición. Es por esto que se sigue el enfoque propuesto por Groß et al. (2017), que permite estimar las densidades asignando a cada observación que pertenece a un determinado radio las coordenadas del centroide correspondiente. Este proceso se realiza mediante un método iterativo de densidad de kernel multivariante.[3] En cada iteración, las pseudo-muestras ayudan a reconstruir la información perdida por el redondeo. Así, se obtienen estimaciones más precisas de la distribución espacial de las subpoblaciones, según cada dimensión, permitiendo identificar zonas con, por ejemplo, alta densidad de migrantes o personas en situación de vulnerabilidad.

En una segunda etapa, se utiliza el algoritmo dinámico de Fuzzy C-Means para clasificar los clústers en base a las estimaciones de densidad obtenidas en la primera fase. A diferencia de los métodos de agrupamiento convencionales, como k-means, que asignan cada punto de datos a un único cluster, Fuzzy C-Means permite que un punto pertenezca a múltiples clústers con distintos grados de pertenencia (valores entre 0 y 1). Este método adapta la estructura de los clústers a medida que se incorporan nuevos datos, como nuevos censos, permitiendo crear, eliminar o ajustar clústers según cambios relevantes en la distribución de los datos. El proceso se realiza en

[3] El autor explica el algoritmo propuesto para su aplicación en R (Groß et al. (2017), Sección 3.4).

ciclos, donde en cada uno se actualizan los clústers en función de los datos añadidos, hasta que ya no se incorporan datos adicionales.

Finalmente, se genera información sobre la composición de los clústers obtenidos para su posterior interpretación. En primer lugar, se propone caracterizar los clústers utilizando un indicador de medias en las distintas categorías de cada dimensión y un test-v[4], el cual permite evaluar qué tan bien una categoría caracteriza al clúster y si esta está sobrerrepresentada o subrepresentada en el clúster en comparación con el total de la Ciudad. Luego, para visualizar los cambios y tendencias intra-clústers a lo largo del tiempo se presentarán Bump Charts, que muestran graficamente cómo han cambiado las clasificaciones de las variables a lo largo del tiempo. Este gráfico se elabora para cada clúster y se estructura típicamente en torno a un eje x temporal con intervalos iguales desde el momento inicial hasta el más reciente. En el eje y se muestran los nombres de las variables, ordenados de forma descendente según sus valores del test-v para cada año. Además, cuenta con líneas de conexión que ilustran cómo cambia la clasificación de las distintas categorías a lo largo del tiempo, lo cual resulta muy útil para resumir estas dinámicas. La ilustración se complementará con figuras que muestren la ubicación de los distintos clústers en el tiempo en el mapa de radios de CABA.

[4] El test-v se computa de la siguiente manera $\frac{\overline{X_c}-\bar{X}}{\sqrt{\left(1-\frac{n_c}{n}\right)\frac{s^2}{n_c}}} \sim t_{n_c-1}$, donde $\bar{X}$ es la media de la variable X en todo el conjunto de datos, $\bar{X}_c$ es la media de X dentro del clúster C, n_c es el número de objetos en C, y s^2 es la varianza global de X. La estadística sigue una distribución t de Student con $n_c - 1$ grados de libertad. Si el valor del estadístico para una variable X dentro del clúster C es mayor que 1.96, se interpreta que la variable caracteriza al clúster. Además, cuanto mayor sea el valor del estadístico, mejor caracteriza esa variable al clúster. El signo del test indica si la variable está subrepresentada (signo negativo) o sobre-representada (signo positivo) en el clúster dado, en comparación con todos los datos disponibles para un año determinado.

5. Resultados y limitaciones

En base a la evidencia expuesta en la revisión de literatura, se espera que persista, en alguna medida, la histórica división norte-sur en CABA, donde el sur concentra una mayor proporción de población vulnerable o de menor nivel socioeconómico. Sin embargo, este análisis pretende enriquecer la caracterización de la segregación residencial a partir de la exploración de nuevas dimensiones. También se espera que la metodología propuesta permita captar y describir los cambios ocurridos en las características de la segregación a lo largo del tiempo e identificar posibles quiebres estructurales como la aparición o desaparición de clústers. Además, el alcance de este análisis pretende no solo enriquecer la caracterización de la segregación socioeconómica, sino también revelar patrones de agrupación social según distintos factores que pueden ser de interés a la hora de comprender las dinámicas urbanas.

Una de las principales limitaciones de esta propuesta es la posible falta de acceso a los datos necesarios, lo que dificultaría la implementación del análisis. Por otro lado, el algoritmo dinámico de Fuzzy C-Means actualiza los clústers incorporando los datos evaluados previamente en cada ciclo como si fueran completamente nuevos. Si bien permite detectar cambios estructurales, no proporciona el tamaño exacto de los clústers en cada período, lo que puede considerarse como otra limitación. No obstante, esta estrategia permite obtener una evaluación robusta de los cambios dinámicos, la cual representa una de las principales contribuciones de esta propuesta.

6. Bibliografía

Linares, S., Mikkelsen, C. A., Velázquez, G. A., & Celemín, J. P. (2016). Spatial segregation and quality of life: empirical analysis of medium-sized cities of Buenos Aires Province. Indicators of Quality of Life in Latin America, 201-218.

Suárez, A. L. (2007). Structure and consequences of socioeconomic segregation in poor Buenos Aires settlements (Doctoral dissertation, UC San Diego).

Masías H, V. H., Stier, J., Navarro R, P., Valle, M. A., Laengle, S., Vargas, A. A., & Crespo R, F. A. (2024). Evolving demographics: a dynamic clustering approach to analyze residential segregation in Berlin. EPJ Data Science, 13(1), 1-41.

Cuenya, B. E. (2018). Consensos y puntos de debate en torno a los conceptos de segregación y fragmentación urbanas. Revista Iberoamericana de Urbanismo, 14, 1-4. https://notablesdelaciencia.conicet.gov.ar/handle/11336/99009.

Groisman, F., & Suárez, A. L. (2006). Segregación residencial en la Ciudad de Buenos Aires. Población de Buenos Aires, 3(4), 27-37.

Rodríguez, G. M. (2008). Segregación residencial socioeconómica en la Ciudad Autónoma de Buenos Aires. Dimensiones y cambios entre 1991–2001. Población de Buenos Aires, 5(8).

Instituto Nacional de Estadística y Censos (INDEC). (2017). Clasificador Nacional de Ocupaciones (CNO) 2017. Disponible en https://www.indec.gob.ar//CNO_2017.pdf

Rodríguez, G. M. (2021). Comparabilidad retrospectiva en la cartografía censal digital del INDEC. Estado actual, avances y desafíos en Argentina y la Ciudad de Buenos Aires. Población de Buenos Aires, 18(30), 22-33.

Rodriguez, G. M. (2023): Bases de microdatos completas del Censo Nacional de Población y Vivienda de 1980 – Argentina. Consejo Nacional de Investigaciones Científicas y Técnicas. (dataset). http://hdl.handle.net/11336/196594

Groß, Marcus; Rendtel, Ulrich; Schmid, Timo; Schmon, Sebastian; Tzavidis, Nikos (2015) : Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error, Diskussionsbeiträge, No. 2015/7, Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin.