



UNIVERSIDAD DE BUENOS AIRES
FCEN - FI
Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Trabajo Final Análisis Inteligente de Datos

Aplicación de Técnicas Multivariadas al Estudio del Mercado Inmobiliario de la Ciudad de Buenos Aires.

Federico Baylé

Profesora: Dra. Ana Silvia Haedo

10 de octubre de 2013

1. Introducción

El objetivo del presente trabajo será el de aplicar los métodos estudiados en la materia Análisis Inteligente de Datos en el ámbito del mercado inmobiliario de la Ciudad Autónoma de Buenos Aires (de aquí en más mencionada como CABA). Más en detalle, se intentará ver si pueden, a partir de las variables seleccionadas, reproducirse las diferencias socio-demográficas de las distintas zonas de la Ciudad (entiendo por zona a un barrio o una comuna). El marco teórico y rector de este trabajo delineará como objetivo analizar la existencia de cuatro ejes dentro de la ciudad, los cuales se intentará reproducir.

Como posible aplicación de este análisis, se podrían sacar conclusiones acerca de la diferencia (en término de las variables utilizadas aquí) entre las distintas zonas de la CABA, de modo de tener un marco de referencia a la hora de intentar equiparar las condiciones de vida a lo largo de la ciudad.

El hecho de contar con el precio de cada vivienda, nos permitirá también sentar las bases en este trabajo de un futuro análisis de determinantes de precio, intentando estimar el impacto de cada variable a la hora de valuar una propiedad. A largo plazo podría pensarse en la implementación de un tasador automático de viviendas, lo que queda fuera del alcance del presente trabajo.

Para realizar el presente trabajo, se utilizaron varios conjuntos de datos de distintas fuentes, a continuación se presenta el listado correspondiente.

- Sitio Web Properati:¹: Precio Departamento - Superficie - Cantidad de Ambientes - Barrio - Tipo de Vivienda.
- Base de Datos Públicos Gobierno de la Ciudad de Buenos Aires:²: Comisarías - Hospitales - Estaciones Subterráneo - Bibliotecas - Punto de Wifi abierto - Establecimientos de Educación Privada - Establecimientos de Educación Pública - Espacios Verdes - Zonas de fácil Anegamiento.

El conjunto de datos utilizado tiene 6,413 casos, con 12 variables. Para vincular las bases de datos, se utilizó el hecho de que todos los registros estaban geo-referenciados, de modo de facilitar el cálculo de distancias, reduciéndolo a la aplicación de la fórmula correspondiente (lo cual será explicado posteriormente).

¹www.properati.com.ar

²<http://data.buenosaires.gob.ar/>

2. Descripción de los Datos

Como se mencionó anteriormente, fueron utilizados varios conjuntos de datos a la hora de hacer el análisis. Cabe destacar que todos los conjuntos de datos utilizados son públicos y de libre acceso en internet, lo que garantiza la reproducción de este estudio.

2.1. Conjunto de Datos Inmuebles

La información acerca de las propiedades en venta fue obtenida del sitio web [Properati](#), correspondiente al mes de Agosto del presente año. Este conjunto de datos presenta información acerca de las propiedades en venta para Ciudad de Buenos Aires, Gran Buenos Aires e interior del país, siendo que para el presente análisis solo se tomó en cuenta la información correspondiente a CABA.

El subconjunto de variables³ que serán utilizados de dicha fuente es: Tipo de Propiedad, Precio USD/ m^2 , Cantidad de Ambientes, Latitud y Longitud, siendo que las dos últimas fueron utilizadas para referenciar con los otros conjuntos de datos.

Dada la gran cantidad de datos faltantes, se tomó un subconjunto que posea la totalidad de los campos completos, de modo de poder aplicar los métodos estadísticos sin complicaciones. De este subconjunto, se extrajeron varios casos erróneos (por ejemplo, departamentos cuyo barrio era Barrio Norte, pero cuando se verificaban el aviso, correspondían a departamentos en Pinamar), resultando de este modo los 6.413 casos que analizarán en este trabajo.

2.1.1. Sesgo del Conjunto de Datos

Cabe destacar que se observa un cierto sesgo en el conjunto de datos resultante en favor de departamentos, lo cual tiene sentido si tenemos en cuenta el uso del suelo en la ciudad. Como se menciona en el Relevamiento de Usos del Suelo (2013) elaborado por el Gobierno de la Ciudad, podemos ver que las edificaciones más altas se encuentran en la zona norte de la ciudad⁴, marcadas por el trazado de las grandes avenidas como Santa Fe o Cabildo. Hacia el oeste de la ciudad, siguiendo el trazado de la avenida Rivadavia, podemos ver que en las Comunas 3 (Balvanera),⁵ (Almagro y Boedo),⁶ (Caballito) y 7 (Flores) también se encuentran edificaciones altas. En cambio para el caso de las comunas más cercanas a los límites de la ciudad, esta característica se pierde, cediéndole terreno a las edificaciones bajas (casas o PH). Por el contrario, en las comunas más cercanas a los límites de la Ciudad, tanto por el Norte como por el Oeste y Sur, tienden a predominar las casas y edificios bajos.

Analizando la cantidad de propiedades por comuna en nuestro conjunto de datos, podemos ver que el sesgo hacia los departamentos se corresponde con lo mencionado en el párrafo anterior, dada la distribución de propiedades en venta según el barrio de nuestra base de datos, lo que se muestra en la figura 2. Se presenta debajo un mapa de calor que muestra la cantidad de propiedades según la zona de la ciudad, viendose allí el claro predominio de la zona céntrica y norte dentro de las viviendas incluidas en este conjunto de datos⁵.

2.2. Conjunto de Datos Gobierno de la Ciudad de Buenos Aires

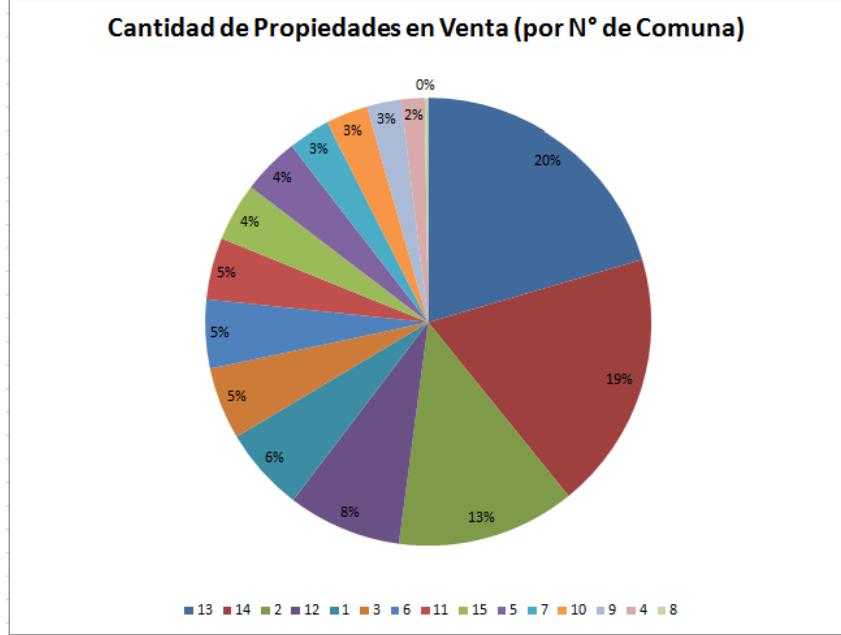
El catálogo de datos abiertos del gobierno de la Ciudad de Buenos Aires comprende una amplia gama de datos de libre acceso: desde información de comisarías hasta de un censo del arbolado de la ciudad, teniendo como principal característica la geolocalización de toda la información. El formato de este catálogo viene

³El total de variables que incluye está formado por Fecha de creación, Tipo de la propiedad (casa, departamento, PH), Operación del aviso (venta, alquiler), Nombre del lugar, Nombre del lugar + nombre de sus 'padres', ID de geonames del lugar (si está disponible), Latitud, Longitud, Precio original del aviso, Moneda original del aviso (ARS, USD), Precio del aviso en moneda local (ARS), Precio aproximado en USD, Superficie en m^2 , Precio en $\frac{USD}{m^2}$, Ambientes, URL en Properati y URL primer foto

⁴Más precisamente en el norte de la Comuna 1(Puerto Madero, San Nicolás y Retiro), 13 (Belgrano y Núñez) y 14 (Palermo) y en la Comuna 2 (Recoleta).

⁵Todos los mapas georeferenciados incluidos en este trabajo fueron realizados con la herramienta open source CartoDB. Para más información ver www.cartodb.com.

Figura 1: Cantidad de Propiedades en Venta según Comuna



dado por el programa CKAN, el cuál es de código abierto,⁶, lo cual lo hace compatible con la mayoría de los datos abiertos de diferentes ciudades del mundo.

De toda la información disponible, cómo se mencionó en la introducción, se utilizaron solamente aquellos datos que se consideraron relevantes a la hora de realizar este análisis, en particular se buscó aquella información que tenga correlato en gran parte de la ciudad, dejando de lado conjuntos de datos como el de Estado de las Autopistas, el cual no está relacionado con los fines de este trabajo.

2.3. Cálculo de Distancias

Para el cálculo de la distancia entre un punto geográfico y otro, se empleó la fórmula del Haversine⁷, la cuál resulta útil para el propósito de este análisis debido a que considera la esfericidad del cuerpo sobre el que se la calcule. Dado que el planeta Tierra no es perfectamente esférico, esta resulta una aproximación, pero dado el escaso territorio de influencia⁸, el error se minimiza, comparando con otras medidas como la distancia euclídea (la fórmula del haversine tiene un error de 3 metros cada 300 km. aproximadamente). La implementación de esta fórmula fue hecha con lenguaje Python, el cual se expone en el apéndice del presente trabajo. Se sugiere ver Sinnott (1984)⁹ para más información sobre el uso y la precisión de esta fórmula.

Dado que la información estaba expresada en formato decimal, (por ejemplo el barrio Chacarita tiene como latitud $-58,4518$) fue necesario pasarlal formato radianes, de modo de poder utilizar la fórmula antes mencionada. Para esto, basta con multiplicar cada dato por $\frac{\pi}{180}$.

Se expone a continuación un ejemplo de un registro aleatorio del conjunto de datos resultante de lo anteriormente expresado, de modo de dar pie al análisis estadístico del mismo:

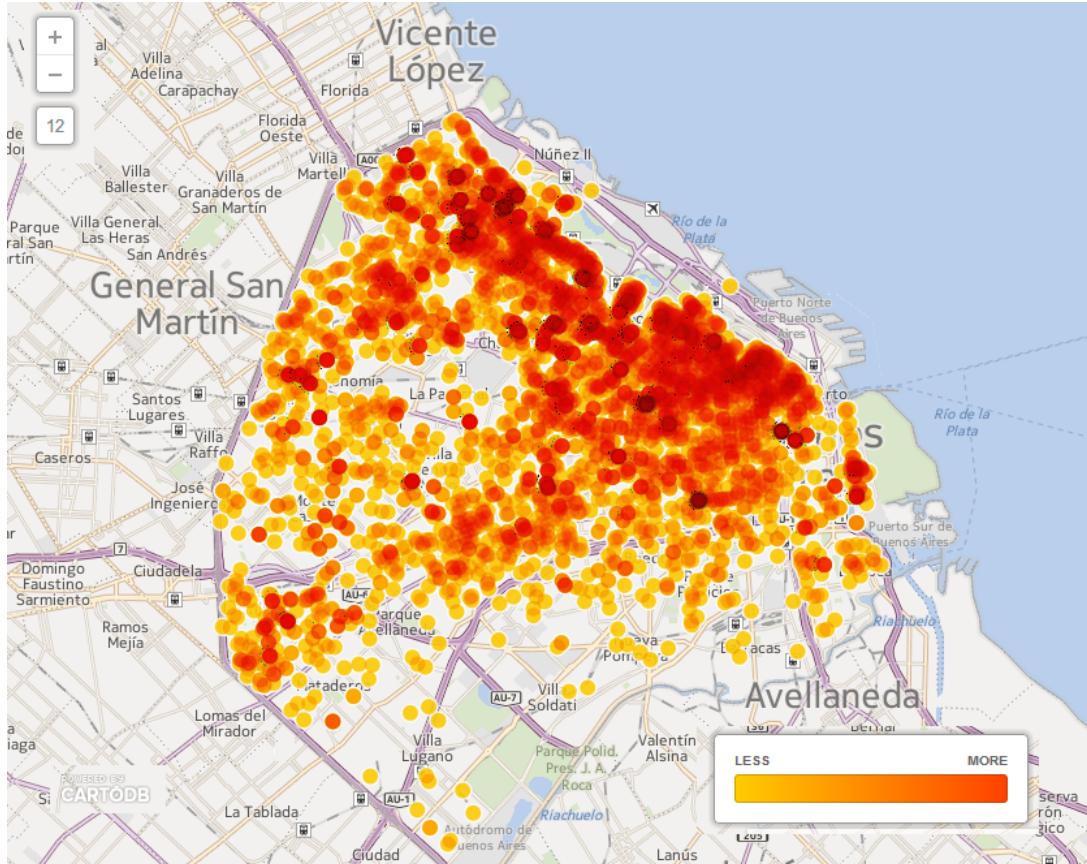
⁶La plataforma CKAN fue desarrollada por la organización Open Knowledge Foundation, en la actualidad es utilizado por más de 40 iniciativas de datos abiertos a lo largo del mundo

⁷La fórmula es $\left(\frac{d}{r}\right) = \text{haversin}(\phi_2 - \phi_1) + \cos(\phi_1) \cos(\phi_2) \text{haversin}(\lambda_2 - \lambda_1)$, donde λ_i representa las longitudes de cada punto y ϕ_i las latitudes, para $i = 1, 2$. Donde el operador haversine tiene la siguiente fórmula: $\text{haversin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2}$

⁸La superficie de la CABA es de 202 km^2

⁹Disponible en http://daimi.au.dk/~dam/thesis/Sky_and_Telescope_1984.pdf

Figura 2: Cantidad de viviendas según localización (mapa de calor).



Cuadro 1: Ejemplo de registro en el conjunto de datos a utilizar

Precio m ² (U\$S)	Comisaría	Hospital Público	Estación Subte	Biblioteca	Wifi
2264,00	0,1704	0,5741	0,1487	0,5325	0,6750

Cuadro 2: Ejemplo de registro en el conjunto de datos a utilizar (continuación)

Escuela Privada	Escuela Pública	Espacio Verde	Calles Anegadas	Superficie	Ambientes
0,1767	0,0290	0,4108	2,2941	170	4

Cabe la aclaración de que las variables referidas a distancia (Comisaría, Hospital Público, Estación Subte, Biblioteca, Wifi, Escuela Privada, Escuela Pública, Espacio Verde y Calles Anegadas), están medidas en kilómetros, y representan la distancia mínima de la vivienda en cuestión a cada lugar en particular, por ejemplo: que el campo Escuela Pública tenga asociado el registro 0,0290, refiere a que la distancia de ese inmueble a la escuela pública más cercana es de 0,0290 km. Las variables ambientes y superficie están expresadas en número enteros, siendo que superficie está medida en m^2 . La variable Precio m^2 (U\$S), como su nombre lo indica, está medida en dólares estadounidenses, representando el valor del metro cuadrado de la propiedad en cuestión en dicha moneda.

3. Métodos Estadísticos Considerados

En primer lugar se llevará a cabo un análisis descriptivo de cada variable, intentando de este modo tener un primer acercamiento a los datos. A través de los estadísticos considerados, se obtendrá un panorama inicial para comenzar a entender los resultados que arrojarán los métodos multivariados.

Una vez caracterizadas las variables individualmente, se procederá a aplicar un análisis de componentes principales, de modo de poder ver qué interpretación se puede obtener de cada eje, quedando como precedente para luego hacer un análisis de aglomerados, el cual intentaremos comparar con la distribución original de las zonas de cada vivienda.

Por último, se intentará mediante el método de análisis discriminante, poder ver cuáles son las variables que más discriminan una zona de la ciudad de otra, tomando como referencia un estudio sobre las condiciones demográficas de la ciudad realizado por la Universidad de Buenos Aires y el gobierno de la Ciudad, el cual se comentará posteriormente. Puede ser de sumo interés los resultados de esto último, ya que nos permitiría saber cuáles son las variables críticas a la hora de intentar equiparar las condiciones socioeconómicas de los habitantes de CABA.

3.1. Software Utilizado

Para la preparación de los datos y limpieza de los mismos, se ha utilizado una planilla de cálculo estándar, como el caso de Microsoft Excel® 2007, debido a que el tamaño de los mismos resultó manejable en dicha plataforma.

A la hora de hacer el análisis estadístico descriptivo, de componentes principales, de aglomerados y discriminante, se optó por el programa R (versión 3.0.1).

En el apéndice se colocará el código tanto de Python (para el cálculo de distancias) como de R utilizado, de modo de que pueda ser utilizado para profundizar este trabajo.

4. Desarrollo

En esta sección se presentarán las metodologías y los resultados parciales de los métodos mencionados anteriormente, explicando las cuestiones que se vayan presentando a lo largo del análisis, de modo de sentar las bases para poder, luego, discutir las principales conclusiones que se vayan obteniendo.

4.1. Análisis Descriptivo de las Variables

En la siguiente tabla se presentan los principales estadísticos descriptivos de las variables consideradas para el análisis. Cabe destacar que todas las variables, excepto Cantidad de Ambientes y Superficie, son numéricas continuas, mientras que dichas dos son numéricas discretas.

Cuadro 3: Estadísticos descriptivos para las variables analizadas.

Variable	N	Mínimo	Máximo	Media	Desviación estándar	Coef.de Variación
Comisaria	6.398	0,014	2,349	0,732	0,410	0,561
Hospital	6.398	0,028	5,676	1,246	0,679	0,545
Subte	6.398	0,002	5,793	1,021	1,063	1,041
Biblioteca	6.398	0,001	2,511	0,834	0,407	0,488
Wifi	6.398	0,001	2,460	0,668	0,369	0,552
Privada	6.398	0,005	0,961	0,188	0,123	0,657
Publica	6.398	0,009	1,582	0,255	0,148	0,582
Verde	6.398	0,009	2,061	0,437	0,248	0,568
Anegadas	6.398	0,002	6,747	1,626	1,342	0,825
Superficie	6.398	20,000	950,000	92,315	78,059	0,846
Precio	6.398	125,000	28666,667	2415,190	1151,596	0,477
Ambientes	6.398	1,000	8,000	2,523	1,248	0,495

Un primer análisis puede mostrar que la variable Distancia Mínima a una Estación de Subterráneo (de aquí en más *Subte*) es una de las que más variabilidad relativa presenta (mirando el coeficiente de variación). Esto resulta intuitivo, puesto que el trazado de la red de subterráneos no cubre la totalidad de la ciudad, por lo que en algunos barrios la distancia mínima hacia una estación de este servicio podría ser grande en comparación con aquellas viviendas de zonas céntricas, donde están concentradas gran cantidad de estaciones. Otra de las variables que presenta gran variabilidad relativa es la superficie de la vivienda, lo cual también tiene sentido dado que de acuerdo a la zona de la ciudad que se analice, podría haber predominio de edificios (con departamentos de no más de 3 ambientes) mientras que en otras podría haber predominio de casas de mayor tamaño. Como se mencionó anteriormente, no hay que olvidar cierto sesgo de este conjunto de datos hacia viviendas en zonas de edificios, lo cual tendrá un impacto en el predominio de inmuebles con poca cantidad de ambientes, y alto precio (dada la concentración en comunas de alto poder adquisitivo, como la Comuna 2).

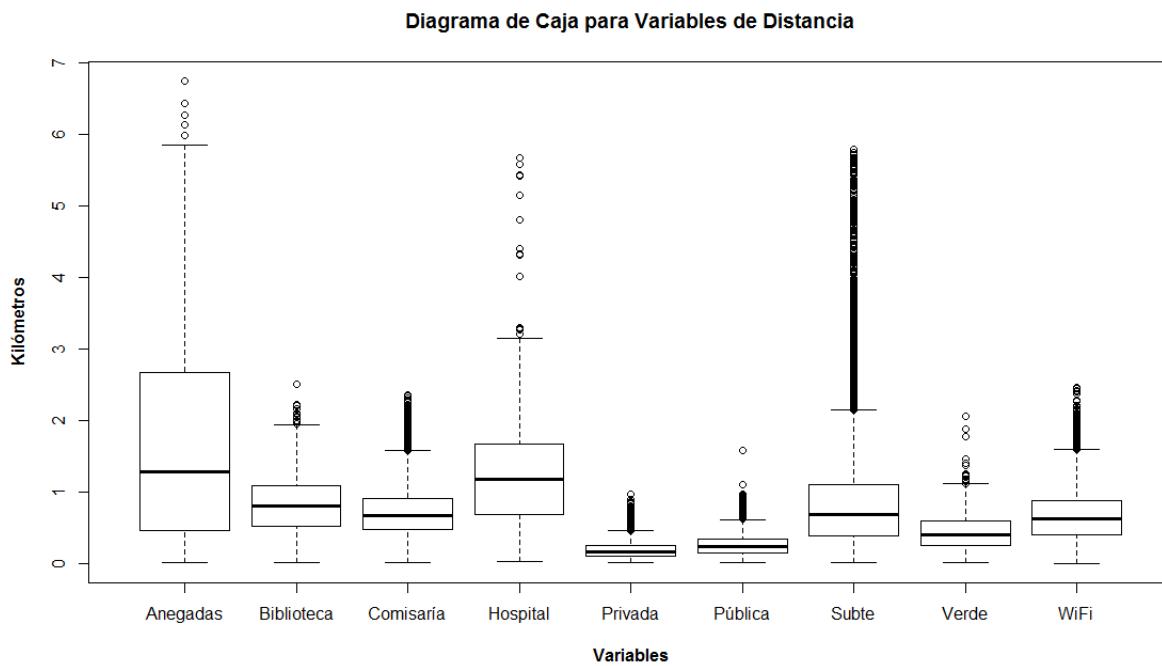
4.1.1. Detección de Valores Extremos

Es de esperar que haya valores extremos en el conjunto de datos seleccionado, dado que resulta mucho más caro una vivienda en el barrio de Puerto Madero que en cualquiera de la zona Oeste. Lo mismo ocurre con la distancia a la estación de subterráneo más cercana, dado que la extensión de este servicio todavía no llega a penetrar tanto en el sur como en el oeste de la ciudad, lo que tendrá su correlato en valores altos de distancia, en comparación con lo que podría suceder para aquellas propiedades de zonas como Barrio Norte o Retiro. Otra variable que a priori uno podría sospechar que va a presentar gran cantidad de datos faltantes, es la que refiere a la distancia mínima a una calle pasible de anegamiento, dado que este tipo de calles suele

ubicarse sobre el entubado de arroyos o canales que recorren la ciudad, como el caso del arroyo Maldonado debajo del trazado de la avenida Juan B. Justo.

Se presenta a continuación los diagramas de caja para las variables de que refieren a distancias.

Figura 3: Diagrama de Caja para las variables que representan distancias



Como puede verse allí, se confirma lo previsto para la variable Subte, mostrando también esta situación para las variables educación privada y comisaría, siendo esto último un rasgo que se intentará explicar una vez aplicados los métodos multivariados que vendrán posteriormente.

Cabe destacar que como resultado de este análisis, se quitaron quince registros, los cuales correspondían a propiedades cuyo precio por metro cuadrado superaba resultada inadmisible dada las características de dichas viviendas, comprobando dichos errores en al página web de Properati. De este modo, el número final de registros a analizar serán 6.398 casos.

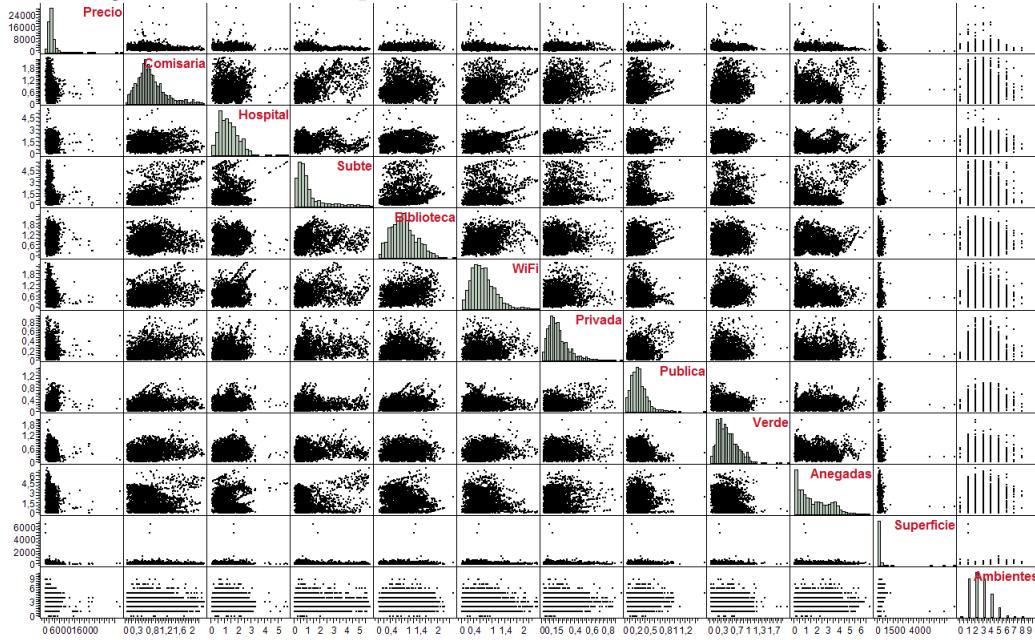
4.2. Análisis Multivariado

Se presenta a continuación la matriz de dispersión de las variables consideradas para el análisis, exponiendo el histograma de cada una en la diagonal principal. Como puede verse en los gráficos de dispersiones, existe poca correlación entre las variables, lo cual será nocivo para aplicar el método de componentes principales en el siguiente apartado. Las variables que aparecen más correlacionadas (con un valor de 0,58) son distancia mínima a una comisaría y a una estación de subterráneo.

Puede verse en los histogramas presentados en dicha matriz, que las variables presentan una forma de campana corrida hacia la derecha, lo que da cuenta de una asimetría positiva en las variables.

Teniendo en cuenta la baja correlación entre cada una de las variables, se procederá con un análisis de componentes principales (de aquí en más ACP), de modo de ver en este caso la relación entre las variables en más de dos dimensiones, para continuar luego con el análisis de aglomerados. Cabe destacar que dicha baja correlación producirá que las componentes halladas no acumulen mucha varianza a medida que crezca la cantidad que se considere de estas.

Figura 4: Matriz de Dispersion para las variables consideradas en el análisis



4.2.1. Análisis de Componentes Principales

Se realizará dicho análisis considerando todas las variables que se vienen trabajando hasta ahora. Como puede verse allí, tomando solamente las dos primeras componentes solamente se obtiene un 34 % de varianza acumulada, lo cual resulta poco teniendo como objetivo resumir las dimensiones en solamente dos componentes (concentran poca varianza respecto del total). Por estas razones, a la hora de hacer el análisis de aglomerados, nos estamos garantizando que las variables no están correlacionadas entre si, cuestión que se había mencionado a la hora de presentar la matriz de dispersiones. La fórmula de las dos primeras componentes es presentada en la tabla 5.

Cuadro 4: Resultados y estadísticos de la metodología de componentes principales.

Número	Autovalor	% Varianza	% Var. Acumulada	Valor Chi	gl	p - valor
1	2,2115	18,429	18,429	12583,784	65,36	<,0001
2	1,8767	15,639	34,068	9450,911	57,327	<,0001
3	1,4264	11,886	45,955	6706,397	49,305	<,0001
4	1,242	10,35	56,305	5097,896	41,09	<,0001
5	1,0731	8,942	65,247	3729,6	33,431	<,0001
6	0,8626	7,188	72,436	2610,206	26,286	<,0001
7	0,7379	6,149	78,585	2000,494	19,734	<,0001
8	0,7137	5,948	84,532	1623,686	14,097	<,0001
9	0,6243	5,203	89,735	1109,195	9,193	<,0001
10	0,5428	4,524	94,259	660,283	5,226	<,0001
11	0,4051	3,376	97,634	200,656	2,219	<,0001
12	0,2839	2,366	100	-	-	-

Se grafica a continuación el biplot correspondiente, de modo de poder obtener alguna conclusión relevante, a pesar de la poca concentración de varianza anteriormente dicha.

Cuadro 5: Primeras dos componentes resultantes del análisis.

Variable	Componente 1	Componente 2
Comisaria	0,50175	0,08379
Hospital	0,10416	-0,0169
Subte	0,49771	0,11273
Biblioteca	0,2668	0,07164
Wifi	0,41408	-0,12328
Privada	0,32619	0,20589
Publica	0,05489	0,30126
Verde	0,2936	-0,21333
Anegadas	-0,06798	0,1842
Superficie	-0,0206	0,62376
Precio	-0,19263	0,01346
Ambientes	-0,09621	0,59818

Puede verse en dicha figura que la primera componente caracteriza a las zonas de la ciudad de mayor poder adquisitivo, dado que tienen mayor importancia en ella las variables Comisaría y Subte, cuyas dependencias se encuentran más concentradas en dicha zona que en el resto de la ciudad. Cabe destacar que también tienen importancia para dicha componente las variables Wifi, Educación Privada y Espacios Verdes, siendo que el Wifi libre está asociado con centros culturales o tecnológicos, lo que tiene sentido con la interpretación anterior. Para el caso de la segunda componente, esta podría caracterizar a las zonas de edificaciones más bajas, dado que las variables que , más importancia tienen allí son Superficie y Ambientes.

Para cerrar este análisis, si bien se le dió una interpretación intuitiva a las dos primeras componentes, no hay que olvidar que la proporción de varianza explicada no resulta relevante a la hora de reducir las dimensiones del problema, por lo que la verdadera utilidad que se le puede dar aquí a esta metodología es la de confirmar la escasa correlación que se había manifestado a la hora de presentar la matriz de dispersiones.

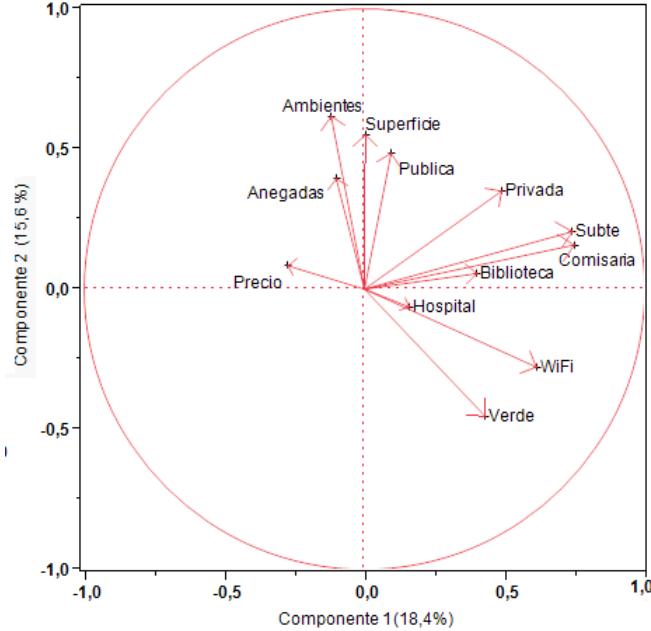
4.2.2. Análisis de Aglomerados

Se buscará mediante este método, poder analizar y ver variaciones en los cuatro ejes que se mencionan en Diagnóstico Socioterritorial de la Ciudad de Buenos Aires, elaborado por el Gobierno de la Ciudad de Buenos Aires y FADU - UBA en 1999.

Dichos ejes son:

- Eje Norte (color rojo)
 - Retiro y Recoleta
 - Centro y San Nicolás
 - Palermo
 - Belgrano, Colegiales, Coghlan, Nuñez y Saavedra.
- Eje Oeste (color verde)
 - Balvanera y San Cristóbal
 - Caballito, Flores y Parque Chacabuco
 - Floresta, Vélez Sarsfield y Parque Avellaneda
 - Villa Luro, Versalles y Liniers.
 - Mataderos

Figura 5: Biplot correspondiente al análisis de componentes principales



- Eje Noroeste (color azul)
 - Chacarita, Villa Crespo, Villa Ortúzar, Agronomía y Paternal.
 - Villa del Parque, Villa Devoto, Villa Urquiza y Villa Pueyrredón.
- Eje Sud (color naranja)
 - Montserrat, San Telmo, Constitución y La Boca.
 - Barracas
 - Parque Patricios y Nueva Pompeya
 - Villa Soldati, Villa Lugano y Villa Riachuelo

Se intentará estudiar en qué medida pueden ser encontrados estos ejes mediante el análisis de conglomerados. Los métodos utilizados serán k - medias y PAM (con distancia euclídea), siendo que la diferencia principal entre estos es que el último de estos busca medoides en lugar de centroides, lo que refiere a la búsqueda de objetos representativos del agrupamiento que tienen una distancia mínima al resto de los miembros de dicho conglomerado, lo cual puede ser útil debido a la centralidad de algunos barrios dentro de los ejes mencionados (como el caso de Retiro, Recoleta y Belgrano en el eje Norte).

Otro aspecto a destacar es que el método de k-medias busca minimizar la suma del cuadrado de las distancias euclídeas entre los miembros de un agrupamiento y su respectiva media. En el método de PAM el objetivo es la minimización de la suma de disimilitudes entre los miembros del agrupamiento y su medoide. Esta diferencia en cuanto a la función a optimizar hace que PAM tienda a ser más robusto que k-medias en conjuntos de datos con valores extremos.

La estrategia de búsqueda consistirá en intentar replicar con ambos métodos la distribución de los cuatro ejes mencionados anteriormente. Luego, se intentará poder obtener las agrupaciones de los barrios dentro de los ejes, teniendo en cuenta los matices mencionados en el trabajo de referencia. Tal es el caso de la clara

pertenencia de barrios como Belgrano a su eje de referencia, en oposición situaciones como las de Palermo o Villa Crespo, en los cuales no resulta homogénea la estructura social del barrio, sino que se presentan, a su vez, diferentes situaciones dentro de cada uno de ellos (como el caso de la diferenciación comercial entre Palermo Hollywood, Las Cañitas o Pacífico).

Se evaluarán dos criterios gráficos para la evaluación de los agrupamientos, como el estudio de las siluetas de los grupos (para el caso del método PAM) y la comparación de la suma de errores al cuadrado versus la cantidad de clusters (para el caso de k - medias). Se sugiere ver el octavo capítulo del libro “Introduction to Data Mining” de Tan, Steinbach y Kumar (2006) para ampliar sobre los diferentes métodos de validación.

Búsqueda de Ejes

En primer lugar se aplicarán los métodos en función de intentar obtener los cuatro ejes mencionados anteriormente, es decir, tanto para el k - medias como para el PAM, el parámetro k tomará en valor 4.

En las tablas 6 y 7 se presentan los resultados obtenidos para cada caso. Como puede verse allí, la cantidad de viviendas que integran cada aglomerado difiere entre uno y otro método. En las figuras 7 se presenta la distribución geográfica de cada aglomerado según k - medias, de modo de poder ver con mayor claridad el resultado de la asignación.

Cuadro 6: Resumen aplicación método k - medias (seleccionando cuatro agrupamientos).

Aglomerado	Color Mapa	SSE	Cantidad	%
1	Verde	6635,96	495	8 %
2	Azul	13438,168	1591	25 %
3	Naranja	18737,77	3032	47 %
4	Rojo	17184,606	1280	20 %

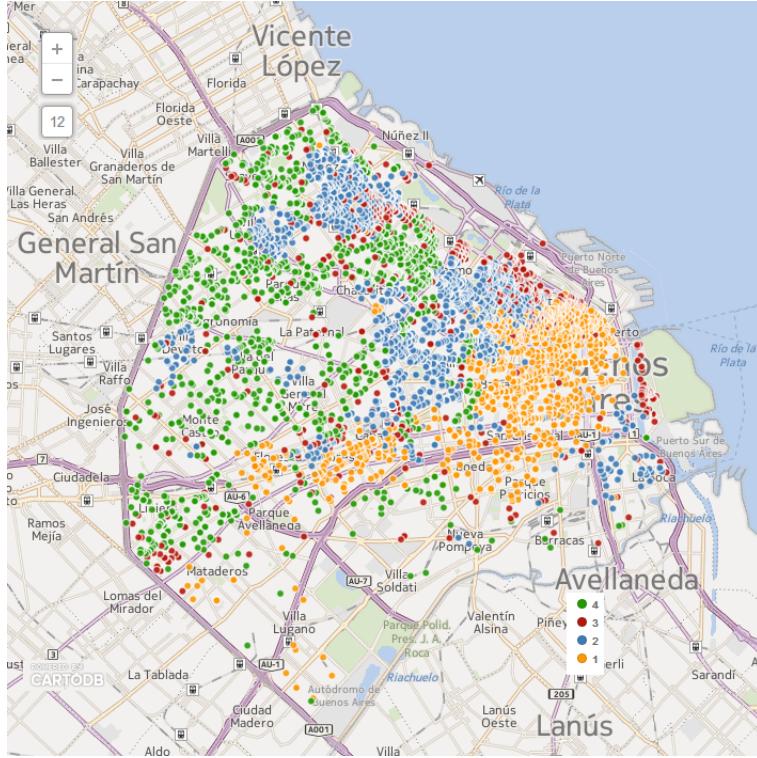
Cuadro 7: Resumen aplicación método PAM (seleccionando cuatro agrupamientos).

Aglomerado	Color Mapa	Tamaño	Max. Dist.	Dist. Promedio	Silueta Promedio	%
1	Naranja	1433	8,935	2,68	0,103	22 %
2	Azul	2255	14,414	2,494	0,157	35 %
3	Rojo	904	22,71	0,196	-0,005	14 %
4	Verde	1806	11,44	3,301	0,017	28 %

Para el caso del k - medias, se obtuvo captó gran parte de las viviendas correspondientes al Eje Oeste (color verde) en el aglomerado 1, con la particularidad de capturar también parte de aquellas pertenecientes al Eje Noroeste, en particular a Villa del Parque, Villa Devoto, Villa Urquiza y Villa Pueyrredón. Para el Eje Noroeste (color azul), podemos ver que este está capturado en gran parte por el aglomerado 2, siendo que también capta parte del Eje Norte, más precisamente la zona correspondiente a Belgrano, Colegiales, Coghlan, Nuñez, Saavedra y parte de Palermo. Cabe destacar la presencia de algunas viviendas fuera de rango captadas en el sur de la ciudad. En cambio, el aglomerado 3, consigue captar la totalidad de la zona céntrica de la ciudad, combinando parte del Eje Sud (naranja) con el Eje Norte y perimetrales del Eje Oeste (como Flores y Balvanera). Por último, el aglomerado 4, consigue acaparar gran parte del Eje Norte (color rojo), poniendo el énfasis en los barrios de Recoleta Y palermo, aunque captando también una porción del eje Oeste, en los barrios de Caballito y Flores. También se mezcla con el Eje Noroeste, particularmente la zona de los barrios Chacarita, Villa Crespo, Villa Ortúzar, Agronomía y Paternal.

Como resumen de la aplicación del k - medias, se puede afirmar que el criterio de referencia en este caso no fue respetado en gran parte. A la vista de los resultados obtenidos, se obtuvo un criterio de agrupamiento de carácter geográfico, siendo la zona oeste de la ciudad la mejor captada. Cabe destacar que más allá de

Figura 6: Distribución geográfica de los agrupamientos (método k - medias, $k = 4$)



los límites geográficos de cada barrio, suele ocurrir que, por ejemplo, determinadas manzanas del barrio de Palermo son similares a las de Chacarita o Villa Crespo, mientras que otras son más cercanas a las de Belgrano o Recoleta. Esta cuestión resulta de interés para poder lograr una mejor separación, lo que constituye tema de futura investigación.

Analizando los resultados obtenidos por el método PAM, podemos ver que resultan peores que para el anterior método, mostrando una mayor dispersión de cada uno de los aglomerados a lo largo de la ciudad. En este caso, no resulta posible hacer una aproximación a los ejes mencionados anteriormente, lo que da la pauta del pobre desempeño de este algoritmo para este caso. Esto era previsible analizando el gráfico de siluetas que arrojó este método, dado el bajo valor de la silueta promedio para cada aglomerado y el promedio general. Sin embargo, resulta interesante analizar el agrupamiento ofrecido por este método por fuera del esquema teórico tomado como referencia aquí, puesto que los aglomerados resultantes podrían aportar otra visión de la trama urbana de la ciudad, identificando cuestiones intrínsecas comunes entre diferentes barrios, más allá de la ubicación de cada uno.

Se deja como futuro trabajo el hecho de discriminar según el tipo de vivienda, dado que el conjunto de datos que se está utilizando tiene un 92 % de propiedades de tipo departamento.

Cantidad de Aglomerados según Métodos Gráficos y Jerárquicos

Mas allá de lo que la teoría a priori propone, veamos cuál sería el número óptimo de aglomerados según los criterios de validación y el uso de métodos jerárquicos de agrupamiento.

Observando los métodos gráficos, se presenta a continuación el gráfico de suma de los errores al cuadrado (el cual es una medida de la cohesión de cada agrupamiento) versus la cantidad de aglomerados seleccionados. Este tipo de gráfico es similar al gráfico de autovalores realizado para componentes principales.

En dicho gráfico puede verse que no hay un punto de inflexión que sea claramente el indicado, podría

Figura 7: Distribución geográfica de los agrupamientos por separado (método k - medias, $k = 4$).

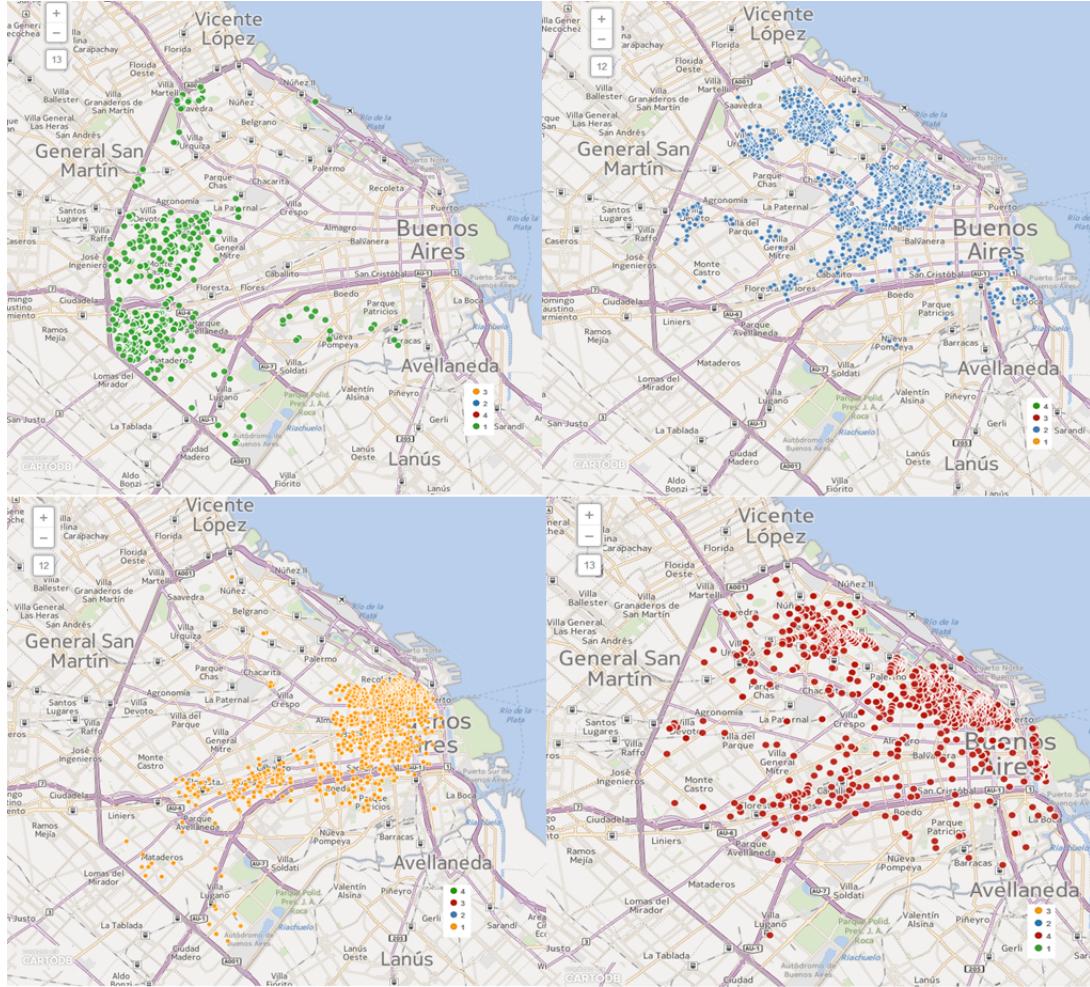
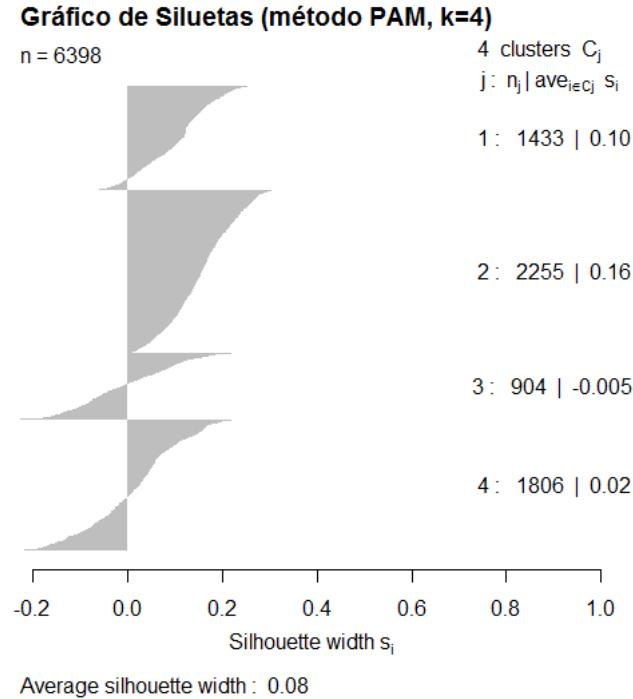


Figura 8: Gráfico de Siluetas (método PAM, $k = 4$).


decirse que entre 4 y 7 agrupamientos se encuentra el mayor descenso de la suma de errores al cuadrado, además de ser coherente con la mayor parte de la bibliografía acerca de zonificación de la Ciudad. De modo que se intentará con estas 4 alternativas, de modo de comparar los resultados entre ambas.

Cuadro 8: Resumen aplicación método k - medias (seleccionando cinco agrupamientos).

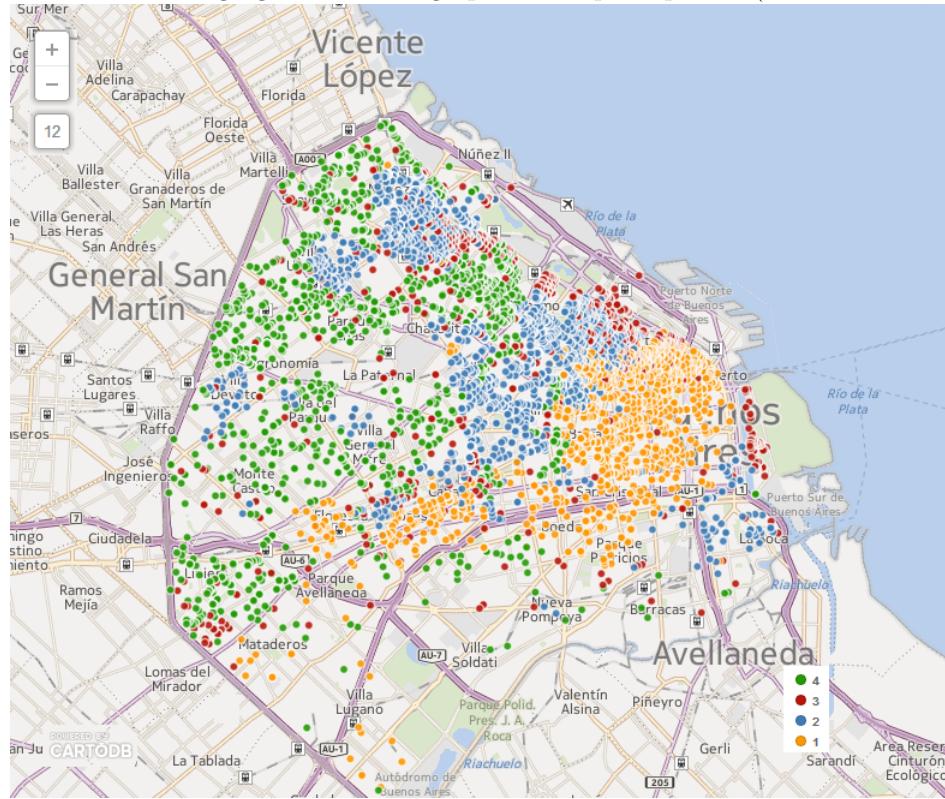
Aglomerado	SSE	Cantidad	%
1	5644	422	7 %
2	8644	1585	25 %
3	12747	1083	17 %
4	15094	862	13 %
5	10961	2466	39 %

Como puede verse allí, el agrupamiento correspondiente al Eje Oeste permanece constante a pesar de modificar la cantidad de aglomerados deseada, lo que da la pauta de estabilidad en dicho agrupamiento. Esto proviene de la homogeneidad de esas viviendas respecto de variables como la distancia a una estación de subterráneo, dado que en el Eje Oeste, la primer estación se encuentra en el barrio de Flores, lo cual hace que para los barrios de Liniers, Villa Luro, Versalles y Mataderos, dicha distancia sea homogénea.

Para los demás agrupamientos, no se observa una gran mejora respecto de los cuatro ejes postulados inicialmente, en término de lograr una definición más clara de los ejes.

Como pasaba anteriormente, no se logra una clara diferenciación entre el eje Norte y Noroeste, dado que se sigue evidenciando un núcleo del eje Norte y Sud dentro del área del Eje Noroeste. Sin embargo, cabe destacar que esto no resulta un aspecto negativo en este análisis, sino que permite comenzar a captar la complejidad de la trama urbana de la ciudad, mostrando que a medida que se avanza en sentido Norte, la homogeneidad

Figura 9: Distribución geográfica de los agrupamientos por separado (método PAM, $k = 4$).



Cuadro 9: Resumen aplicación método k - medias (seleccionando seis agrupamientos).

Aglomerado	SSE	Cantidad	%
1	12464	708	11 %
2	8745	1738	27 %
3	5008	820	13 %
4	9775	1703	27 %
5	5661	871	14 %
6	8446	558	9 %

Cuadro 10: Resumen aplicación método k - medias (seleccionando siete agrupamientos).

Aglomerado	SSE	Cantidad	%
1	4701	1235	19 %
2	7896	380	6 %
3	6840	1718	27 %
4	8818	293	5 %
5	7685	1282	20 %
6	4829	797	12 %
7	7128	693	11 %

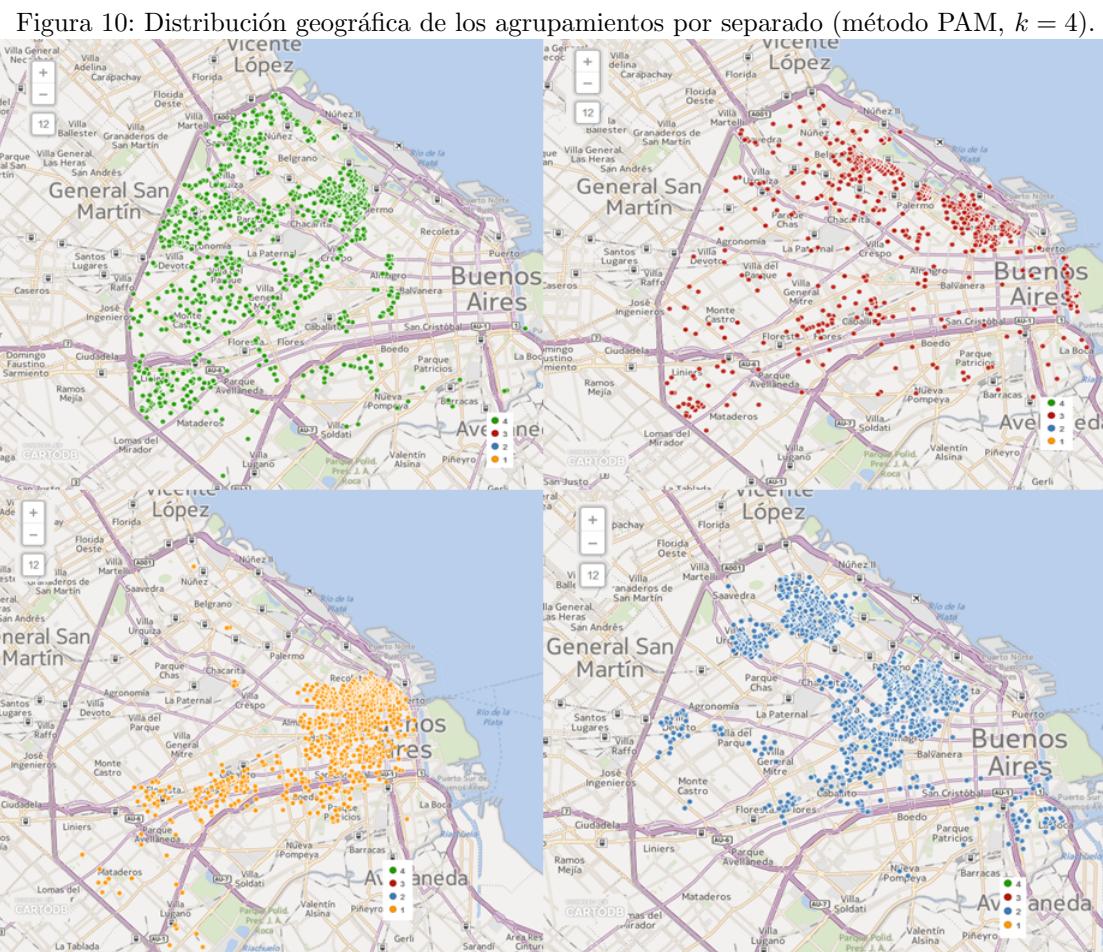
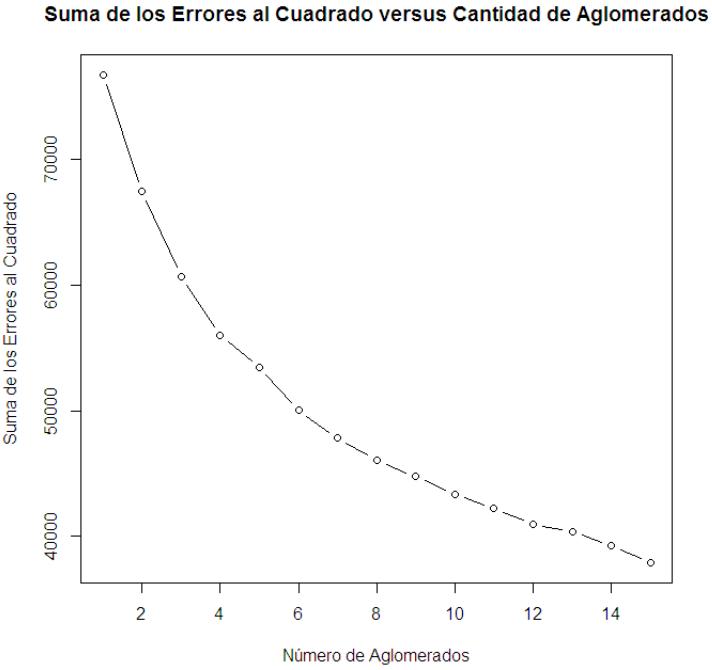


Figura 11: Distribución geográfica de los agrupamientos por separado.



de los barrios del Oeste y del Sur comienza a desaparecer, dando lugar a zonas más heterogéneas en materia de servicios y tipos de viviendas.

Por último, veamos cuál es la cantidad de agrupamientos que sugieren los métodos de agrupamiento jerárquicos, haciendo primero una distinción entre ellos para elegir el adecuado.

Analizando entre los métodos de Ward, mínima distancia y máxima distancia, podemos ver que dada la estructura del conjunto de datos, se requerirá un método que soporte el ruido y la presencia de valores extremos¹⁰. De este modo, el método de distancia máxima será preferido al de distancia mínima, por el primero más robusto a la presencia de dicho tipo de valores, aunque cabe destacar que el método de distancia mínima tiene la tendencia de quebrar los agrupamientos favoreciendo las formas globulares, siendo una alternativa intermedia el uso de distancia promedio. Cabe destacar que para cualquiera de este tipo de métodos, la cantidad de datos con la que se está trabajando aquí no dificulta el desempeño de los algoritmos, pero si se quisiera elevar la cantidad de datos, se comenzaría a tener problemas de desempeño, dada la necesidad de cómputo de la matriz de distancia en cada paso de ejecución.

En el caso del método de Ward, este tiene la misma función objetivo que el k - medias, definiendo la proximidad entre dos agrupamientos como el incremento en la suma de los errores al cuadrado que surge de unir dos aglomerados.

Se presenta a continuación el dendrograma correspondiente para los métodos de Ward y distancia máxima, de modo de analizar las diferencias que arrojan cada uno de ellos.

Del análisis de estos se desprende el hecho de que el método de Ward fue el que mejores resultados proporcionó, dado que puede verse como el dendrograma resulta más estable, marcando entre cuatro y cinco agrupamientos posibles, lo cual está destacado en el gráfico con las líneas negra (punto de corte para cuatro agrupamientos) y la línea violeta (punto de corte para cinco agrupamientos). Cabe destacar que el resultado

¹⁰Dado que resulta interesante para el presente análisis la presencia de valores extremos (como algunas viviendas en Recoleta y Puerto Madero) y ruido (como el caso de propiedades de gran tamaño o departamentos de gran superficie) debido a que permiten caracterizar mejor cada zona de la ciudad, se decidió no quitar del conjunto de datos esa información

Figura 12: Distribución geográfica de los agrupamientos (método k - medias, $k = 5, 6, 7$).

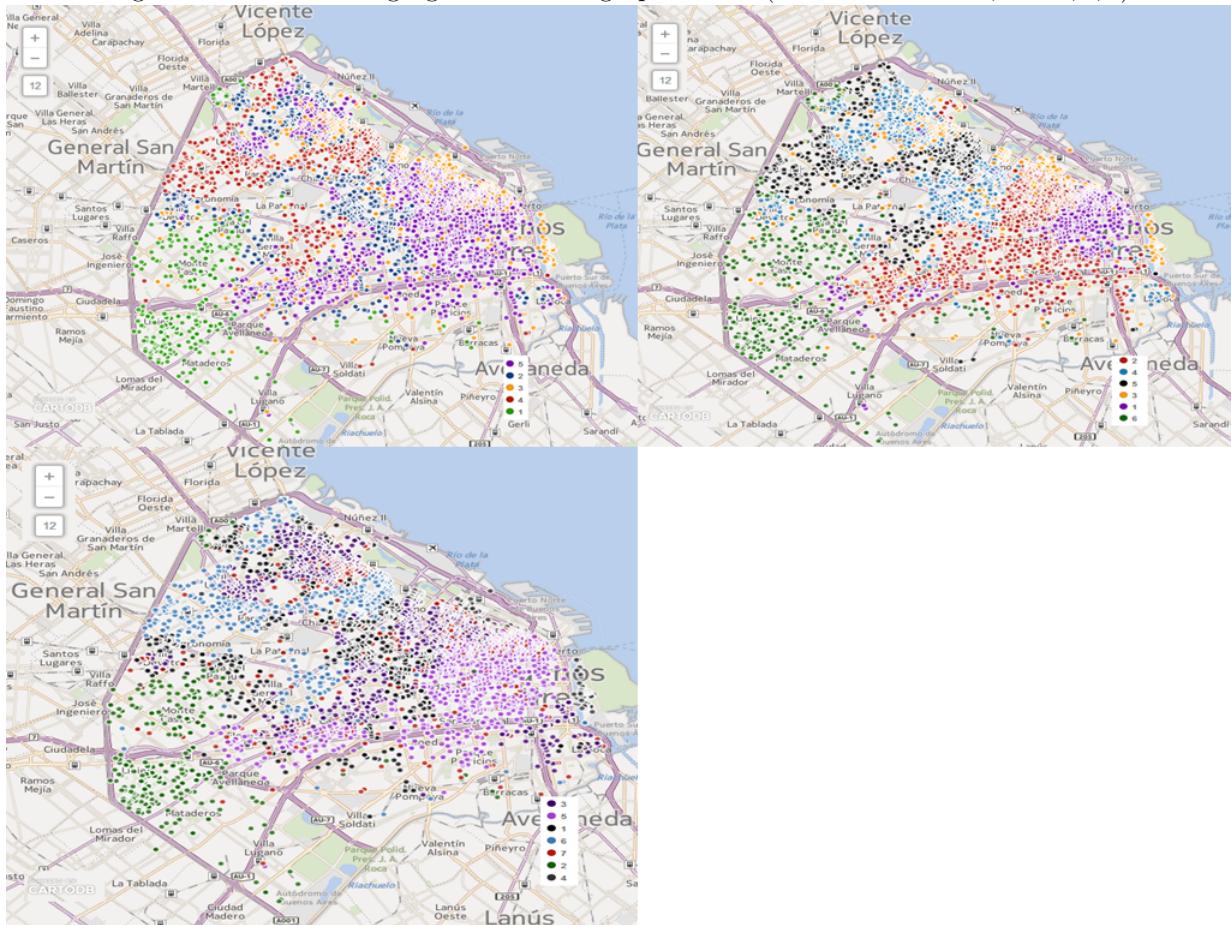


Figura 13: Dendrograma correspondiente al método de Ward.

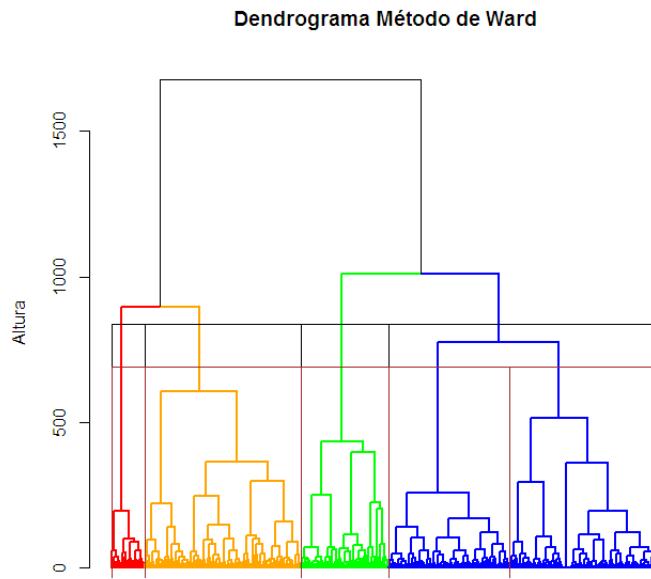
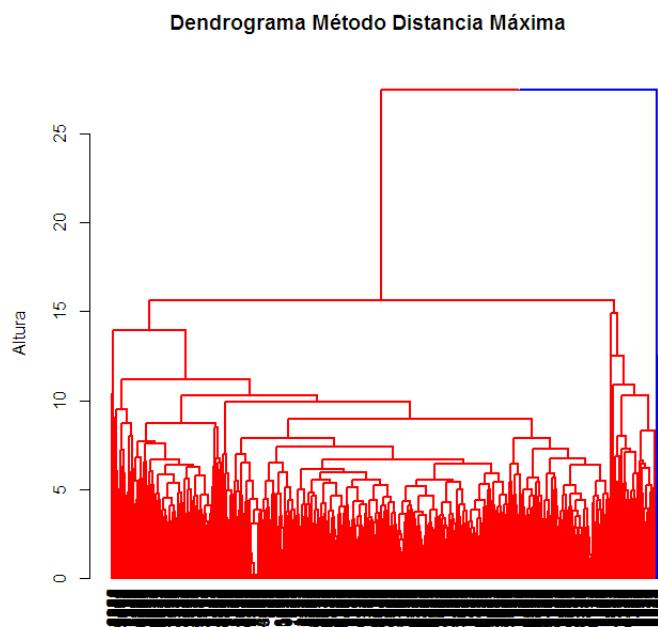


Figura 14: Dendrograma correspondiente al método de máxima distancia



de la aplicación de este método es consistente con lo obtenido anteriormente para el método k - medias.

4.2.3. Análisis Discriminante

Una vez estudiada la pertinencia de intentar agrupar los casos en determinadas categorías (ejes), se intentará ver en esta sección cuáles son las variables que más contribuyen a la hora de discriminar entre cada una de las clases planteadas, de modo de dejar abierta la cuestión de intentar clasificar las viviendas según características de estas.

El esquema de trabajo será comenzar por verificar el cumplimiento de los supuestos de esta metodología por parte de las variables, para luego analizar los resultados obtenidos. A la hora de clasificar, nuevamente se utilizarán los ejes considerados para el análisis de agrupamientos.

En primer lugar, analicemos si las matrices de varianzas - covarianza poblacionales correspondientes a cada grupo son iguales. La idea de verificar este supuesto radica en verificar si todos los grupos pertenecen a la misma población. Para esto acudiremos a la prueba de Box de igualdad de matrices de covarianza. Se presenta a continuación el resultado de este contraste.

Cuadro 11: Prueba M de Box sobre igualdad de matrices de covarianza.

Resultados de pruebas

M de Box	12510,12
F	Aprox. 52,87715
	gl1 234
	gl2 1188779
	Sig. 0,0000

Como puede verse allí, el p - valor es cercano a 0, lo que da la pauta de rechazar la hipótesis nula de covarianzas poblacionales iguales. De este modo, corresponde continuar el análisis discriminante de tipo cuadrático. Sin embargo, la estrategia será continuar con este método para luego comparar con los resultados que arroje el cuadrático.

A continuación, veamos si las medias entre los grupos se encuentran lo suficientemente separadas como para garantizar que tiene sentido realizar el análisis discriminante. Para evaluar si los centros de los 4 grupos están separados, utilizaremos otra de las pruebas provenientes de MANOVA, la cual tiene como hipótesis nula que todas las medias de los grupos son iguales y como alternativa que existe al menos una media distinta. Para llevar a cabo esta prueba, contamos con 4 distintos estadísticos de prueba: Lamba de Wilks, Traza de Pillai, traza de Hotelling-Lawley y máxima raíz de Roy. Se presenta a continuación una tabla que presenta el resultado para cada uno de estos.

Cuadro 12: Resultados prueba de hipótesis sobre medias de los grupos

Prueba	Valor	F	gl	p - valor
Lambda de Wilks	0,3850	199,7582	36	< 0,0001
Traza de Pillai	0,7664	182,5754	36	< 0,0001
Hotelling-Lawley	1,2181	215,9281	36	< 0,0001
Máxima raíz de Roy	0,7662	407,6808	12	< 0,0001

Analizando el resultado de las pruebas, en todos los casos podemos rechazar la hipótesis nula de igualdad de medias (es decir, los grupos no están completamente solapados), dado que el p-valor es menor a 5 %. Este resultado es alentador en vista de la calidad de los resultados del análisis discriminante, dado que al no ser iguales las medias, al menos podemos decir que tienen algún grado de separación (cuanto más separadas estén mejor). El hecho de que el valor de lambda de Wilks no sea cercano a 1, da la pauta que dicha discriminación

no será tan buena. Cabe destacar que la traza de Pillai suele ser considerado el mejor de estos estadísticos al ser el más robusto de los cuatro.

Resultados Análisis Discriminante

Para analizar cuál es la variable de mayor importancia en el análisis, se tendrá en cuenta los Coeficientes estandarizados de las funciones discriminantes canónicas. Esto se debe a que, desde el punto de vista técnico, al considerar las variables estandarizadas, se pierde el peso que pueda tener cada una debido a su escala de medición y varianza, de modo que se puede comparar partiendo todas de la misma base. Se presenta también el valor de cada autovalor asociado a cada función discriminante, de lo cual se ve que con solo considerar las primeras dos funciones, se capta el 96,2 % de la varianza total.

Cuadro 13: Autovalores de cada función discriminante

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	0,766	62,9	62,9	0,659
2	0,406	33,3	96,2	0,537
3	0,046	3,8	100,0	0,209

Cuadro 14: Coeficientes de función discriminante canónica estandarizadas

Variable	Función		
	1	2	3
Comisaria	0,117	-0,439	0,694
Hospital	-0,528	-0,055	0,096
Subte	0,057	0,779	-0,231
Biblioteca	0,174	-0,034	0,287
Wifi	0,107	0,370	0,379
Privada	0,168	0,123	-0,646
Publica	-0,135	-0,104	-0,039
Verde	0,284	0,176	0,109
Anegadas	0,739	-0,570	0,169
Superficie	-0,192	-0,050	0,165
Precio	-0,449	-0,179	0,326
Ambientes	-0,100	0,005	-0,147

Analizando esta información, podemos ver que las variables más importantes para el cálculo de las dos primeras funciones son distancia mínima a estación de subterráneo (Subte), a una calle pasible de anegamiento (Anegadas), a un hospital (Hospital) y el precio del m^2 en dólares (Precio). Resulta interesante destacar de esto el hecho de que para el caso de Subte, puede verse como la separación entre los ejes de mayor nivel socioeconómico (sector norte de la ciudad) está explicada por cuestiones como el acceso al transporte más veloz, cuestión que resulta intuitiva. El hecho de que el precio también presente importancia a la hora de discriminar era esperable, mientras que sorprende la aparición de la variable Anegadas, lo cual puede estar relacionado con que gran parte de las calles pasibles de anegamiento se encuentran en el norte de la ciudad, con fuerte presencia en los barrios de Belgrano y Palermo (corresponden al eje Norte). Se presenta también la matriz de estructura para cada función, de modo de buscar ampliar lo argumentado anteriormente.

De analizar dicha matriz de estructura se refuerza lo mencionado anteriormente, respecto de la importancia de las variables a la hora de poder discriminar las diferentes propiedades en base a su eje de pertenencia.

Por último se presentan las funciones discriminantes propiamente dichas, las cuales nos permitirán clasificar luego propiedades en los ejes de acuerdo a las variables que se han analizado.

Cuadro 15: Matriz de estructuras para las funciones discriminantes.

	Función		
	1	2	3
Anegadas	0,58	-0,55	-0,03
Precio	-0,45	-0,24	0,24
Hospital	-0,35	0,03	0,12
Publica	-0,20	-0,14	0,02
Ambientes	-0,12	-0,11	-0,11
Superficie	-0,11	-0,07	-0,04
Subte	0,27	0,55	0,10
Verde	0,13	0,36	0,23
Wifi	0,13	0,51	0,52
Comisaria	0,22	0,14	0,48
Privada	0,13	0,16	-0,42
Biblioteca	0,05	0,20	0,37

Cuadro 16: Coeficientes de las funciones de clasificación.

	Ejes			
	Norte	Noroeste	Oeste	Sur
Comisaria	3,56	1,88	4,38	1,24
Hospital	2,42	1,72	0,72	1,02
Subte	-0,92	0,54	-0,86	-0,16
Biblioteca	3,53	3,72	4,51	3,20
Wifi	3,46	5,65	4,13	2,98
Privada	3,86	6,31	6,29	13,15
Publica	6,25	4,31	4,31	4,93
Verde	7,97	10,11	10,46	9,42
Anegadas	1,60	1,15	3,00	2,11
Superficie	-0,01	-0,01	-0,01	-0,01
Precio	0,00	0,00	0,00	0,00
Ambientes	2,17	2,11	1,99	2,21
(Constante)	-16,59	-16,97	-17,66	-14,01

Una vez que contamos con esta información, se presenta la matriz de clasificación que arrojó este método, de modo de poder medir el desempeño. Cabe destacar que existen otras metodologías para evaluar la precisión del modelo (como el estudio de la curva ROC), pero que quedan fuera del presente trabajo.

Del análisis de dicha tabla se desprende que el 68 % de las viviendas fueron clasificadas correctamente en uno de los cuatro posibles ejes, siendo que los ejes mejor clasificados son el Eje Norte y el Eje Oeste, lo cual luego del análisis de aglomerados resulta intuitivo puesto que estas zonas de la ciudad fueron las que mejores agrupamientos presentaron.

Cuadro 17: Tabla de clasificación

		Pronosticada					
		Ejes	Norte	Noroeste	Oeste	Sud	Total
Original (%)	1	69,2		12,1	10,9	7,9	100,0
	2	17,4		63,5	9,5	9,6	100,0
	3	12,6		5,2	69,1	13,1	100,0
	4	7,6		7,1	27,6	57,6	100,0

Recordando que el anterior análisis no cumple el supuesto de homogeneidad de matriz de covarianzas, se muestra a continuación la tabla de clasificación resultante de hacer análisis discriminante cuadrático, en la cual se verifica un mayor nivel de clasificaciones correctas, el cual alcanza el 77 % de las viviendas.

Cuadro 18: Tabla de clasificación análisis discriminante cuadrático

		Pronosticada					
		Ejes	Norte	Noroeste	Oeste	Sud	Total
Original (%)	1	78,3		11,2	5,7	4,8	100,0
	2	12,4		73,6	8,4	5,6	100,0
	3	8,2		4,1	77,8	9,8	100,0
	4	10,6		1,8	12,4	75,3	100,0

5. Resultados y Conclusión

Luego de analizar bajo la óptica de diferentes métodos univariados y multivariados el conjunto de datos creado, se observó que la trama urbana de la ciudad de buenos aires resultó ser más compleja de lo que se esperaba. Al momento de realizar el análisis de componentes principales, se vió que las variables seleccionadas explican un bajo porcentaje de la varianza total, lo cual se tradujo en una interpretación cuidadosa de cada componente, dado que no tenían el nivel de representatividad adecuado como para sacar alguna conclusión sobre los coeficientes de cada variable. Sin embargo, la baja proporción de varianza acumulada a lo largo de las componentes (junto el cómputo del coeficiente de correlación para cada caso), reforzó la idea de no dependencia entre las variables, lo cual permitió obtener conclusiones confiables del análisis de agrupamientos. A la hora de efectuar el análisis de aglomerados, los mejores resultados se observaron para el k - medias con 4 agrupamientos, teniendo en cuenta los criterios de validación y la distribución geográfica de cada vivienda a cada aglomerado. El hecho de aumentar la cantidad de agrupamientos empeoró los resultados anteriores, debido a que comenzaban a producirse distorsiones y entrecruzamientos aún mayores entre los ejes seleccionados (por ejemplo, se distorsionaba la asignación al eje oeste, asignando viviendas de zona norte a este eje). Al intentar analizar estos datos con el método de agrupamiento PAM, no se encontraron resultados satisfactorios, siendo estos peores, en todos los casos, respecto de lo arrojado por el k - medias, aunque dejan abierta la posibilidad de tratar de interpretar la trama urbana de la ciudad desde un enfoque distinto del estrictamente geográfico.

Dado que para estos dos métodos la cantidad de agrupamientos a detectar es arbitraria, se buscó mediante métodos jerárquicos la cantidad de aglomerados que resulte más estable, tomando como criterio entre objetos pertenecientes a cada aglomerado la distancia máxima y el método de Ward. De la aplicación de estos surgió que el método de Ward fue el que dio los resultados más cercanos a los esperado, dado que se sugería a partir de estos que la cantidad de agrupamientos que más se adaptaba a este conjunto de datos es de cuatro o cinco aglomerados, lo cual coincide con el marco teórico, lo obtenido en k - medias y la distribución geográfica de los datos.

Para concluir, se puede afirmar que en alguna medida se han corroborado los resultados provistos por el marco teórico, logrando tanto para el Eje Oeste como para el Sud una precisión interesante en la asignación a los aglomerados. Sin embargo, el hecho de que el estudio sea de hace más de una década, plantea la posibilidad de una modificación del entramado discutido durante el trabajo, presentando un desarrollo y movilización de barrios cada vez más heterogéneos como Palermo, marcado también por el desarrollo de otros polos inmobiliarios traccionados por el saturamiento de zonas como el centro o microcentro, dando lugar a nuevas valorizaciones y usos del suelo en zonas que antes eran destinadas a otro tipo de viviendas.

Como futuro trabajo, se podría avanzar aún más en la descripción interna de cada eje, intentando aplicar nuevamente los algoritmos de agrupamiento dentro de cada una de estas zonas, de modo de lograr una separación mucho más precisa e interesante. Además, el hecho de haber detectado mediante el análisis discriminante cuáles son las variables que más peso tienen en la diferenciación de las zonas, nos da el puntapié para seguir profundizando y poder determinar e implementar acciones concretas que permitan revalorizar zonas y fomentar su desarrollo, de modo de mejorar las condiciones de vida dentro de aquellos aglomerados más perjudicados dentro de la ciudad.