Universitat de València

Trabajo de Fin de Máster

Segmentación de los barrios de València para la detección de nichos de mercado

AUTOR
Juan Fernando Morala Girón

TUTOR
Dr. José A. Álvarez Jareño

JOSE ANTONIO ALVAREZ JAREÑO - NIF:18955137D
Firmado digitalmente por JOSE ANTONIO ALVAREZ JAREÑO - NIF:18955137D
Fecha: 2020.09.28 10:27:43 +02'00'

28 de septiembre de 2020

Segmentación de los barrios de València para la detección de nichos de mercado

Autor:
Juan Fernando Morala Girón

Tutor:
Dr. José A. Álvarez Jareño

Resumen

La localización de zonas con características demográficas y económicas homogéneas se presenta como una herramienta de utilidad para el sector asegurador. La edad de las personas y su nivel de renta tienen gran incidencia en la adecuación de las pólizas de seguro y la tarificación de sus primas, así como en el horizonte de ahorro y las opciones de inversión en los distintos productos financieros. Debido a la importancia de ambos factores, conocer su distribución dentro de un territorio permite orientar hacia resultados más positivos la oferta en el mercado y optimizar la relación con los clientes de la compañía. Con estas premisas, en el presente trabajo se utilizan diversas técnicas de análisis multivariante para la segmentación de los barrios de València que posibilitan la detección de zonas envejecidas, rejuvenecidas o de alto poder adquisitivo dentro de la ciudad.

Usando como principal fuente de información la Oficina de Estadística del Ayuntamiento de València y el software estadístico R como herramienta de trabajo para los análisis, se calcularán para cada uno de los barrios series temporales de diversos indicadores demográficos que se proyectarán mediante modelos ARIMA para el año 2021. Combinando estos valores con los de renta proporcionados por el Instituto Nacional de Estadística, construiremos un conjunto de datos sobre el cual se realizará un detallado análisis factorial exploratorio que concluirá con la obtención de tres variables explicativas: una asociada a la composición etaria, otra a los flujos migratorios y una tercera a la capacidad económica. Finalmente, un posterior análisis cluster y de autocorrelación espacial realizado sobre tales factores nos permitirá clasificar los barrios y tener una visión amplia de su potencial asegurable.

Los resultados nos indican que los barrios de los distritos centrales de la ciudad son los que disponen de mayor solvencia económica, pero a su vez se encuentran más envejecidos, mientras que en la periferia las tendencias se invierten. Destacan los barrios de Sant Pau, Ciutat de les Arts i de les Ciències, Penya-roja y Massarrojos, que se posicionan como un nicho de mercado atractivo al poseer una población con rentas elevadas y altamente rejuvenecida.

Palabras clave: barrios de València, nichos de mercado, análisis factorial exploratorio, análisis cluster, autocorrelación espacial.

Índice

Introducción

1. Motivación

Cualquier negocio o empresa necesita información del comportamiento del mercado para poder competir y generar beneficios a corto y a largo plazo. En concreto, de los gustos, preferencias o necesidades de los potenciales consumidores y clientes. Este conocimiento es indispensable para la definición de estrategias comerciales con las que obtener rédito. La existencia de diferentes grupos de consumidores obliga además a orientar la oferta a uno de ellos o a disponer de una mayor gama de productos con el que satisfacer la demanda. En línea con esta idea, surge el concepto de **segmentación**, que se puede definir como "la identificación de un grupo de consumidores que presumiblemente se comporten de un modo similar ante un determinado producto o servicio" [13, p. 100]. Para crear estos grupos o segmentos se requieren bases de segmentación, que son variables que permiten caracterizar el perfil del consumidor, ya sean geográficas, demográficas o conductuales. De este modo, la obtención y el análisis de estos segmentos posibilita determinar cuáles otorgan más rentabilidad al negocio en función del producto ofertado así como llevar a cabo una mejor gestión de las relaciones con los clientes.

Este conocimiento de los clientes aún cobra mayor importancia en el sector asegurador, donde la alta competencia en el mercado y el poco margen de maniobra en cuanto a la exclusividad de productos convierte en una tarea ardua la suscripción de nuevas pólizas. De hecho, las compañías aseguradores han ido evolucionando hacia el seguro personalizado, centrando los esfuerzos en cerrar el círculo con el cliente de por vida ofertando productos individualizados basados en las características propias de cada asegurado. Para ello es crucial tener información del asegurado del mayor número de variables posibles: edad, composición del hogar, salario, situación laboral, motivación del ahorro, propensión al riesgo... Algunas de estas variables, por muy simples que parezcan, se yerguen como fundamentales, como es el caso de la edad, la cual esconde tras ella bastante información. A parte de incidir directamente en aspectos como la esperanza de vida, existen investigaciones que sugieren que las personas mayores tienen menos aversión al riesgo que las personas de mediana edad [18, p. 865].

Así, considerando variables como la edad y la renta de las personas, en este trabajo se ofrece una herramienta de conocimiento del mercado en la ciudad de València con la que disponer de una visión amplia previa a la relación con el cliente. Usando la técnica de segmentación, dividiremos los barrios de la ciudad en zonas que poseen características demográficas y económicas homogéneas, con el fin de que el sector asegurador pueda tomar decisiones a raíz de la clasificación resultante de los barrios. Esta clasificación, con la que se detectan en la ciudad áreas rejuvenecidas o enriquecidas, permitirá interpretar ciertos barrios como pequeños territorios que albergan un grupo de clientes potenciales y comparten características comunes que los hacen especialmente receptivos a un determinado producto o servicio, lo que se puede definir como **nicho de mercado** [12, p. 26].

Cabe destacar que a pesar de que este documento ha sido redactado como Trabajo de Fin de Máster de un Máster en Ciencias Actuariales y Financieras y por lo tanto pensado para que resulte de beneficio o interés en el sector asegurador y financiero, el análisis estadístico desarrollado así como los resultados obtenidos pueden ser aprovechados por personas vinculadas a otros colectivos o estudiantes de otras áreas de conocimiento, por lo que no se ha empleado un lenguaje excesivamente técnico con el fin de facilitar su lectura y comprensión. De esta suerte, con la intención de que en el futuro se puedan realizar nuevas líneas de investigación con ayuda del trabajo aquí ya realizado, se ha implementado una política de código libre posibilitando la réplica de todos los gráficos y resultados incorporados, así como el acceso rápido a las bases de datos construidas.

2. Objetivos

Sin duda, el objetivo principal de este Trabajo de Fin de Máster es establecer una clasificación de los barrios de la ciudad de València basada en las características demográficas y económicas de los habitantes que lo conforman. Esta clasificación pretende servir como una herramienta de soporte para la localización de áreas de la ciudad que pueden ser vistas como nichos de mercado en los que potenciar la oferta aseguradora. De esta forma, el presente TFM ayuda a optimizar la gestión y planificación de los departamentos comerciales al ofrecer una visión de la ciudad que permita organizar las campañas de captación y orientar la venta de productos financieros o aseguradores en función de los perfiles de los barrios. No obstante, este trabajo ha sido redactado también con la idea de ser de utilidad a otros sectores de la sociedad. La detección de zonas envejecidas, rejuvenecidas, con un nivel alto o bajo de renta o con mucha o poca migración es una información de la que aquellas personas ajenas al mundo del seguro pueden sacar provecho de ella igualmente. Así por ejemplo, este trabajo puede ser usado por la administración pública en la toma de decisiones de las políticas sociales al analizar las desigualdades resultantes o por pequeños comerciantes que tengan la intención de abrir un negocio ligado a un público en concreto. No obstante, además de esta idea de segmentación de los barrios en la que indudablemente se basa el trabajo, observamos una serie de objetivos secundarios que planteamos a continuación:

* Tener un mayor conocimiento de nuestra ciudad, tanto en los perfiles económicos y etarios de los ciudadanos como en la división territorial de València a un nivel inframunicipal como el que representan los barrios, pudiendo así repasar o aprender sus nombres y localizaciones.
* Demostrar la habilidad adquirida en el máster en el manejo de bases de datos.
* Mostrar el potencial del software estadístico R, una herramienta informática usada con frecuencia en diferentes asignaturas, lo que me ha permitido profundizar en el gran número de funciones que tiene incorporadas y mejorar la técnica asociada a la programación.
* Dar visibilidad a estudios y publicaciones de organismos dedicados a la estadística que en ocasiones no son tan conocidos como el *Padrón Muncipal de Habitantes de la ciudad de València* de la Oficina de Estadística del Ayuntamiento de València o el *Atlas de Distribución de Renta de los Hogares* del Instituo Nacional de Estadística, los cuales contienen bastante información que puede ser explotada con diferentes metas.
* Diseñar y construir de forma específica para este TFM unas bases de datos de calidad de los barrios de la ciudad.
* Calcular indicadores demográficos, como son las tasas, que transformen los datos en bruto capturados en datos relativos y sea así posible comparar la información de los distintos barrios de la ciudad.
* Plasmar de una forma práctica en los análisis estadísticos los conocimientos aprendidos en los dos años de estudios del máster, como es el caso de los modelos ARIMA, el análisis factorial o el análisis cluster entre otros.
* Dibujar mapas coropléticos de los barios con los que mostrar los resultados de una forma visual más agradable.
* Investigar, estudiar y aplicar a los resultados alcanzados un análisis de autocorrelación espacial que aporte un elemento diferencial extra al trabajo.
* Crear un repositorio en la nube abierto a todo el público con el cual alumnado, profesorado, personal investigador o personas interesadas en el tema puedan acceder de manera rápida al material utilizado en este trabajo, en especial a las bases de datos construidas.

3. Metodología

Este trabajo pretende ser una aplicación puramente práctica de las técnicas estadísticas aprendidas en el máster en el que poder plasmar la destreza en el manejo de datos adquirida, más que una profundización teórica de los conceptos o modelos estudiados. Por este motivo, no nos centraremos en la definición de los métodos empleados ni explicaremos de forma detallada la base matemática que subyace en ellos, sino que nos dedicaremos exclusivamente a comentar de manera breve en qué consisten y con qué finalidad los utilizaremos. Existe una gran cantidad de libros científicos y artículos de investigación que abordan y desarrollan las herramientas de análisis que vamos a manejar, una pequeña parte de la cual se ha consultado e incluido en las referencias bibliográficas al final de este TFM. El nivel académico de tales publicaciones hace posible entender el trasfondo teórico de los mismos con mayor claridad y precisión que la que yo pueda en estos momentos ofrecer.

De este modo, el presente trabajo, como su nombre bien indica, se focaliza en clasificar los barrios de la ciudad de València de forma justificada, identificando así las diferencias etarias y económicas entre ellos y, en consecuencia, las zonas con perfiles de población semejantes. Esta segmentación puede ser aprovechada en diferentes esferas, en particular en el sector asegurador, que puede encontrar en este modelo una herramienta de utilidad para la detección de áreas con atractivos potenciales de asegurabilidad. Las zonas en cuestión pueden interpretarse como nichos de mercado en los que abunda una parte de la población con características adecuadas para productos aseguradores o financieros en concreto. Para poder llevar a cabo esta división, a parte de los conocimientos estadísticos necesarios, es fundamental tener datos de los barrios. Por tal motivo, el trabajo consta de dos partes claramente diferenciadas: la creación de las bases de datos de demografía y renta utilizadas y su manipulación para la obtención de los resultados. A pesar de que el grueso del trabajo se concentra en esta segunda parte donde se exponen los pasos que se han seguido en el análisis para alcanzar la segmentación de los barrios, se ha visto la conveniencia de explicar también el proceso de construcción de la base de datos debido a su relevancia. Si bien es cierto que configurar la base de datos demográfica llevó su tiempo a causa de la recopilación y normalización de todas las variables del conjunto de barrios de la ciudad para el periodo **2004-2018**, lo realmente reseñable es el diseño de la base de datos de renta. La importancia radica en el hecho de que los datos de renta utilizados en el posterior análisis se han tenido que calcular *ad hoc*, como veremos en el apartado correspondiente más adelante, lo que seguramente convierta el conjunto de datos de barrios de València de este trabajo único en la actualidad y, por consiguiente, otorgue a los resultados aquí alcanzados una exclusividad y singularidad a remarcar.

Así, en la primera parte del trabajo se detalla la extracción y estandarización de los datos por barrio, tanto los de demografía como los de renta. Además, previamente a iniciar el análisis, tiene lugar un paso intermedio en el que se calculan usando los datos de demografía 30 indicadores para cada uno de los 15 años de la serie, lo que permite trabajar con valores en términos relativos y facilita la comparación de la información entre territorios con diferentes tamaños de población. Las bases de datos construidas, así como todos los códigos R usados, se han depositado en la nube mediante mi cuenta de GitHub con la finalidad de que cualquier estudiante, profesor u persona interesada pueda acceder a ellos con total libertad y rapidez. De este modo, es posible replicar el análisis aquí practicado, así como llevar a cabo nuevas investigaciones relacionadas.

Cabe destacar que aunque en la práctica la ciudad de València está dividida en 19 distritos municipales y subdividida en 87 barrios, en este estudio se trabajará con un total de 85 barrios. Como se explicará posteriormente, el motivo es que se ha decidido fusionar por un lado los barrios 17.4. Cases de Bárcena y 17.5. Mauella y por otro lado los barrios 19.7. la Torre y 19.8. Faitanar, debido a que Mauella y la Torre cuentan con una población demasiado pequeña como para poder considerar aceptables las conclusiones que de ellos se puedan colegir. Con la intención de dar a conocer el nombre de todos los barrios y la división territorial de València que rige desde el año 2004, se presenta seguidamente la actual composición de la ciudad por distritos y barrios.

División territorial de la ciudad de València por distritos y barrios.

1. Ciutat Vella
    - 1.1. la Seu
    - 1.2. la Xerea
    - 1.3. el Carme
    - 1.4. el Pilar
    - 1.5. el Mercat
    - 1.6. Sant Francesc
2. l'Eixample
    - 2.1. Russafa
    - 2.2. el Pla del Remei
    - 2.3. Gran Via
3. Extramurs
    - 3.1. el Botànic
    - 3.2. la Roqueta
    - 3.3. la Petxina
    - 3.4. Arrancapins
4. Campanar
    - 4.1. Campanar
    - 4.2. les Tendetes
    - 4.3. el Calvari
    - 4.4. Sant Pau
5. la Saïdia
    - 5.1. Marxalenes
    - 5.2. Morvedre
    - 5.3. Trinitat
    - 5.4. Tormos
    - 5.5. Sant Antoni
6. el Pla del Real
    - 6.1. Exposició
    - 6.2. Mestalla
    - 6.3. Jaume Roig
    - 6.4. Ciutat Universitària
7. l'Olivereta
    - 7.1. Nou Moles
    - 7.2. Soternes
    - 7.3. Tres Forques
    - 7.4. la Fontsanta
    - 7.5. la Llum
8. Patraix
    - 8.1. Patraix
    - 8.2. Sant Isidre
    - 8.3. Vara de Quart
    - 8.4. Safranar
    - 8.5. Favara
9. Jesús
    - 9.1. la Raiosa
    - 9.2. l'Hort de Senabre
    - 9.3. la Creu Coberta
    - 9.4. Sant Marcel·lí
    - 9.5. Camí Real
10. Quatre Carreres
    - 10.1. Mont-Olivet
    - 10.2. en Corts
    - 10.3. Malilla
    - 10.4. Fonteta de Sant Lluís
    - 10.5. na Rovella
    - 10.6. la Punta
    - 10.7. Ciutat de les Arts i de les Ciències
11. Poblats Marítims
    - 11.1. el Grau
    - 11.2. el Cabanyal - el Canyamelar
    - 11.3. la Malva-rosa
    - 11.4. Beteró
    - 11.5. Natzaret
12. Camins al Grau
    - 12.1. Aiora
    - 12.2. Albors
    - 12.3. la Creu del Grau
    - 12.4. Camí Fondo
    - 12.5. Penya-roja
13. Algirós
    - 13.1. l'Illa Perduda
    - 13.2. Ciutat Jardí
    - 13.3. l'Amistat
    - 13.4. la Bega Baixa
    - 13.5. la Carrasca
14. Benimaclet
    - 14.1. Benimaclet
    - 14.2. Camí de Vera
15. Rascanya
    - 15.1. Orriols
    - 15.2. Torrefiel
    - 15.3. Sant Llorenç
16. Benicalap
    - 16.1. Benicalap
    - 16.2. Ciutat Fallera
17. Pobles del Nord
    - 17.1. Benifaraig
    - 17.2. Poble Nou
    - 17.3. Carpesa
    - 17.4. Cases de Bàrcena
    - 17.5. Mauella
    - 17.6. Massarrojos
    - 17.7. Borbotó
18. Pobles de l'Oest
    - 18.1. Benimàmet
    - 18.2. Beniferri
19. Pobles del Sud
    - 19.1. el Forn d'Alcedo
    - 19.2. el Castellar-l'Oliverar
    - 19.3. Pinedo
    - 19.4. el Saler
    - 19.5. el Palmar
    - 19.6. el Perellonet
    - 19.7. la Torre
    - 19.8. Faitanar

La segunda parte del trabajo muestra el análisis estadístico que conduce a la segmentación de los barrios. Con la idea de aportar mayor utilidad a los resultados, este sección se inicia con una pequeña proyección a corto plazo de los indicadores demográficos con modelos ARIMA para estimar los valores esperados en el año 2021 y poder realizar el análisis teniendo en cuenta dichos valores. De este modo, el data frame que servirá como input de los algoritmos está compuesto por estos 30 indicadores estimados para el próximo año y por 3 procedentes de los indicadores de renta: la renta media por persona, la renta media por hogar y el porcentaje de poblacion con ingresos por unidad de consumo por debajo del 60 % de la mediana, indicador este último que utiliza el INE como definición del umbral de pobreza. Puesto que de los indicadores de renta solo hay información para el periodo 2015-2017, estos no se proyectarán, sino que se incluirán en el data frame el valor promedio de ellos.

A continuación, se procede a realizar un análisis factorial sobre este conjunto de datos de 33 variables, en el cual tras observar las altas correlaciones entre los diversos indicadores y comprobar con el índice de Kaiser-Meyer-Olkin que podemos factorizar las variables originales de forma eficiente, se busca reducir la dimensionalidad y trabajar con un grupo más pequeño de variables. Basándonos en el gráfico de sedimentación de los autovalores de las 33 componentes decidimos trabajar con 3 factores, calculando entonces múltiples soluciones factoriales y seleccionando aquella que mejor cantidad de varianza explica, resultando ser en nuestro caso la hallada por el método de las componentes principales con rotación simplimax. Decidido el número de factores y el método de factorización, seguidamente se desarrolla lo que se conoce como análisis factorial exploratorio, con el cual vamos quitando variables del conjunto inicial que no aportan información de manera significativa y recalculando los factores hasta alcanzar una solución factorial aceptable. Además, se computa el valor del α de Cronbach que nos sirve para evaluar la consistencia y fiabilidad de los factores. Llegados a este punto, se analizan los "loadings" o cargas factoriales para identificar los indicadores a los que hace referencia cada componente transformada. Observamos que los 3 factores se pueden asociar con los conceptos de *Vejez*, *Migración* y *Riqueza*.

En el siguiente apartado del análisis se realiza una clusterización por el método de las K-Medias para una primera clasificación en 5 grupos de los 85 barrios, lo que nos permite por medio de las puntuaciones factoriales saber las características de cada clase y poder así crear mapas coropléticos con los que localizar zonas envejecidas o zonas con alto poder adquisitivo. En una segunda clasificación de los barrios más profunda, compararemos distintos métodos de cluster jerárquicos para elegir el más óptimo, usando como medida de bondad de ajuste el coeficiente de correlación cofenética. Nos decantaremos por el método Average-Linkage considerando la distancia máxima y dividiremos los barrios en 11 grupos, con los que podremos identificar unos perfiles más precisos en las diferentes zonas de la ciudad. En esta sección nos apoyaremos en gráficos de dispersión de las puntuaciones factoriales para entender mejor el resultado devuelto por el algoritmo de particiones interno del método de las K-Medias y en un dendograma circular para el algoritmo de clasificación del método jerárquico Average-Linkage.

Finalmente, se incluye un análisis espacial tras percibirse en los mapas cierta correlación espacial. Como veremos, la distribución geográfica es importante y aquellos barrios colindantes tienen características más similares que aquellos barrios más distantes. Se utilizarán los estadísticos Global Moran's I y Geary's C para determinar que efectivamente existe correlación espacial y a continuación los estadísticos Getis-Ord Gi y Local Moran's I para encontrar las zonas en las que se da lugar de forma significativa esta correlación. En este apartado además se usarán unas funciones basadas en el Método Monte Carlo para poder aseverar que los valores de estos estadísticos no son casuales.

Bases de datos

Uno de los aspectos más importantes en cualquier trabajo de modelización y análisis es tener a disposición datos de calidad, puesto que estos representan la piedra angular del proyecto. Por muy buenos que sean los modelos construidos y por muy potentes que sean las herramientas de análisis disponibles, estos carecen de utilidad si no es posible aportarle unos inputs decentes. Dichos inputs son indispensables para la obtención de buenos resultados. Por tal motivo, es habitual que la verdadera dificultad del trabajo resida en hallar los datos más que en la técnica para manipularlos. De hecho, en el presente TFM se tuvo que dedicar una parte importante del tiempo a la búsqueda, captura, normalización y elaboración del conjunto de datos final sobre el que realizar todo el análisis posterior. Para remarcar la relevancia de este primer paso consistente en la construcción de unas bases de datos de calidad, se ha decidido incluir este apartado de *Bases de datos* antes de adentrarse en lo que es propiamente el trabajo estadístico en sí. De este modo, explicaremos ahora el proceso llevado a cabo para la recopilación de los datos sobre los que se sustenta el trabajo. Cabe destacar que la información a nivel inframunicipal, como es el caso de los barrios de València, no se encuentra en muchos sitios, por lo que es importante tener claro qué buscar y dónde buscar. En cuanto a *qué buscar*, si lo que queremos es localizar zonas envejecidas o empobrecidas, obviamente necesitaremos datos demográficos y datos de renta. Algo más complicado como ya hemos mencionado es contestar adecuadamente a la pregunta *dónde buscar*. Una buena forma de proceder es pensar en primer lugar en los organismos oficiales especializados en la estadística que trabajen en el ámbito geográfico deseado. En València por ejemplo existe desde 1985 la Oficina de Estadística del Ayuntamiento de València encargada de producir y difundir estadísticas asociadas a la ciudad. Una de sus tareas es publicar anualmente el *Padrón Municipal de Habitantes de la Ciudad de València* [39], en el cual se puede encontrar información demográfica a nivel de barrio o incluso de sección censal. Por lo tanto, este es sin duda el mejor lugar para encontrar datos demográficos de los barrios de València. Sin embargo, los datos de renta son más complicados de hallar y en la mencionada Oficina de Estadística no se trabaja con ellos. Descartada esta opción, siempre es conveniente recurrir al Instituto Nacional de Estadística (INE), la institución estadística de mayor nivel dentro de España. Recientemente, a finales de 2019 el INE dio a conocer un proyecto experimental en el que colaborando con la Agencia Estatal de Administración Tributaria (AEAT) había elaborado indicadores estadísticos asociados a la distribución de renta de los hogares a nivel municipal e inframunicipal. Así, distinguiremos a continuación la captura de los datos demográficos tomando como fuente la Oficina de Estadística y la de los datos de renta tomando como fuente al INE.

4. Datos demográficos

Como ya se ha indicado, la fuente de información utilizada para la extracción de los datos de demografía relacionados con los barrios de València es la Oficina de Estadística del Ayuntamiento de València. El principal medio de difusión que utiliza este organismo es su página web (`www.valencia.es/estadistica`), donde periódicamente va actualizando las estadísticas que ofrece. Dentro de ella, es sencillo acceder al "Catálogo de publicaciones", donde podemos hallar todas las series de publicaciones disponibles. Aquí vamos a hacer uso de dos de ellas: *Padrón Municipal de Habitantes. Características de la Población de la Ciudad de Valencia* [39] y *Padrón Municipal de Habitantes. Dinámica demográfica de la Ciudad de Valencia* [37]. El primero contiene tablas de contingencia con los datos de la población de la ciudad a 1 de enero de cada año registrados en el **Padrón Municipal de Habitantes**. El segundo, llamado también Altas y Bajas del Padrón, contiene datos relacionados con las componentes demográficas, es decir, de los nacimientos, las defunciones y los flujos migratorios contabilizados, acaecidos en el año anterior. Ambas publicaciones incluyen un archivo Excel con los datos de las tablas, que son los que vamos a descargar para explotar.

Debido a la reestructuración de la división territorial de la ciudad llevada a cabo en 2003 que impide la comparación de la información inframunicipal anterior a esa fecha con la posterior, nos centraremos en construir una base de datos referida al periodo **2004-2018**. Además, a pesar de que en València existan 87

barrios, el hecho de que tanto en Mauella como en La Torre haya una población considerablemnte pequeña nos motiva a juntar ambos núcleos de población con la de barrios vecinos con el fin de conseguir una mayor significación estadística en los resultados, sobre todo en los datos de mortalidad, natalidad y movimientos de migración donde se podrían producir importantes variaciones de un año para otro, consiguiendo así unas series más consistentes. De este modo, sumaremos por un lado la información descargada del barrio "17.5. Mauella" con la del "17.4. Cases de Bàrcena" y por otro lado la del barrio "19.7. La Torre" con la del "19.8. Faitanar", trabajando por tanto con un total de **85 barrios**. Codificaremos los barrios manteniendo la ordenación oficial que podemos observar en las publicaciones descargadas, con la salvedad de las fusiones mencionadas en los distritos 17 y 19.

Para la construcción de la base de datos de demografía haremos uso de una hoja Excel. El conjunto de datos resultante tras la explotación contará con 222 variables, siendo las 5 primeras las vinculadas con la identificación del año y el barrio en cuestión. Las siguientes 147 obtenidas de la publicación del Padrón contendrán información de la estructura de la población de cada barrio del siguiente tipo:

- Población según sexo y tres grupos de edad: 0-15, 16-64 y 65 años o más.
- Población según sexo y grupos quinquenales de edad: 0-4, 5-9, 10-14, . . . , 84-89 y 90 años o más.
- Población según sexo y lugar de nacimiento: València, resto de l'Horta, resto de la Comunidad Valenciana, resto de España y Extranjero.
- Población extranjera según sexo y tres grupos de edad: 0-15, 16-64, 65 años o más.
- Población extranjera según sexo y continente de nacionalidad: Unión Europea, resto de Europa, África, América del Norte, América Central, América del Sur, Asia y Otros (incluye Oceanía y Apátridas).
- Hojas padronales según tipo: familiares y colectivas.
- Población en hojas padronales según tipo: familiares y colectivas.
- Hojas familiares según tamaño: 1, 2, 3, 4, 5, 6 y 7 personas o más.
- Hojas familiares según composición: con alguien de 0-15 años, con alguien de 16-24 años, con alguien de 25-64 años, con alguien de 65 años o más, con alguien de 80 años o más, con solo personas de 0-24 años, con solo personas de 65 años o más y con solo personas de 80 años o más.
- Hojas familiares según menores de 18 años: 0, 1, 2, 3, 4, 5 y 6 o más menores.

Las 70 variables restantes corresponderán a datos asociados a las componentes demográficas extraídos del documento de Altas y Bajas del Padrón y contendrán el siguiente tipo de información:

- Nacimientos según sexo.
- Defunciones según sexo.
- Defunciones según cuatro grupos de edad: 0-15, 16-64, 65-79 y 80 años o más.
- Inmigración interurbana según sexo y nacionalidad: españoles y extranjeros.
- Inmigración interurbana según sexo y tres grupos de edad: 0-15, 16-64 y 65 o años y más.
- Emigración interurbana según sexo y nacionalidad: españoles y extranjeros.
- Emigración interurbana según sexo y tres grupos de edad: 0-15, 16-64 y 65 años o más.
- Altas por cambio de domicilio según sexo y nacionalidad: españoles y extranjeros.

- Altas por cambio de domicilio según tres grupos de edad: 0-15, 16-64 y 65 años o más.
- Bajas por cambio de domicilio según sexo y nacionalidad: españoles y extranjeros.
- Bajas por cambio de domicilio según tres grupos de edad: 0-15, 16-64 y 65 años o más.

Es importante mencionar que por motivos de secreto estadístico y protección de datos, cuanto más pequeño es el ámbito geográfico deseado hay menos información disponible o esta se encuentra de manera más agregada. Así, ambas publicaciones contienen tablas más detalladas para la ciudad o los distritos. No obstante, consideramos que los datos demográficos a nivel de barrio extraídos contienen el suficiente nivel de información para hacer un adecuado análisis estadístico de los mismos.

Una vez finalizada la extracción y guardado en el libro Excel todas las variables, el propósito ahora es condensar toda esta información en un conjunto de indicadores que nos permita tener una visión más precisa de la estructura poblacional intrínseca de cada barrio, así como de sus componentes demográficas. Existe mucha bibliografía y metodología para el cálculo de indicadores demográficos básicos como [24], [36], [40], [41] o [43]. Estos indicadores facilitan tener una perspectiva general de la dinámica de la población y permiten conocer mejor su evolución al explicar unas características más precisas de la misma, como es el caso del Índice de Envejecimiento o la Tasa de Mortalidad. Además, en nuestro caso, el uso de indicadores como son las tasas (datos en términos relativos) en vez de los datos en bruto recopilados (datos en términos absolutos) se antoja como necesario para la correcta interpretación de las series, ya que así es posibe comparar territorios con tamaños diferentes. Por tal motivo, el siguiente paso tras la normalización de los datos fue la elaboración de indicadores demográficos, que se realizó en el mismo libro Excel debido a lo práctico que resulta el uso de fórmulas con celdas referenciadas y su emulación para filas posteriores. En total se calcularon las series de 30 indicadores distintos para cada uno de los 85 barrios. La definición de dichos indicadores se encuentra disponible en el Anexo I al final de este trabajo. Enumeramos a continuación los 30 indicadores, estando los primeros 11 vinculados a la mortalidad, natalidad y flujos migratorios y los siguientes 19 a la estructura de la población y de las hojas familiares:

- Indicador 1: Crecimiento Vegetativo.
- Indicador 2: Saldo Migratorio.
- Indicador 3: Saldo de Movimientos Intraurbanos.
- Indicador 4: Tasa de Natalidad.
- Indicador 5: Tasa General de Fecundidad.
- Indicador 6: Tasa de Mortalidad.
- Indicador 7: Tasa de Inmigración.
- Indicador 8: Tasa de Emigración.
- Indicador 9: Tasa de Llegadas por Cambio de Domicilio.
- Indicador 10: Tasa de Salidas por Cambio de Domicilio.
- Indicador 11: Relación de Masculinidad al Nacimiento.
- Indicador 12: Índice de Sundbarg.
- Indicador 13: Índice de Friz.
- Indicador 14: Índice de Burgdofer.

- Indicador 15: Índice Generacional de Ancianos.
- Indicador 16: Índice de Envejecimiento.
- Indicador 17: Índice de Sobreenvejecimiento.
- Indicador 18: Índice Demográfico de Dependencia.
- Indicador 19: Índice de Estructura de la Población Activa.
- Indicador 20: Índice de Reemplazamiento de la Población Activa.
- Indicador 21: Índice de Carga Preescolar.
- Indicador 22: Razón de Progresividad Demográfica.
- Indicador 23: Relación de Masculinidad.
- Indicador 24: Porcentaje de población extranjera.
- Indicador 25: Porcentaje de población de 65 años o más.
- Indicador 26: Porcentaje de población de 15 años o menos.
- Indicador 27: Porcentaje de población nacida en la ciudad de València.
- Indicador 28: Porcentaje de hojas familiares con solo personas de 80 años o más.
- Indicador 29: Porcentaje de hojas familiares sin menores.
- Indicador 30: Media de personas por hoja familiar.

El archivo Excel final que contiene todos los datos demográficos normalizados y los indicadores calculados se ha depositado bajo el nombre `DatosBarrioDemografia.xlsx` en un repositorio abierto al público de mi cuenta de GitHub asociado a este Trabajo de Fin de Máster. Se puede descargar de forma rápida desde el siguiente enlace: `https://raw.githubusercontent.com/JuanferMG/TFM/master/DatosBarrioDemografia.xlsx`

Una vez tenemos a nuestra disposición los indicadores calculados en Excel, podemos capturar los datos con R para llevar a cabo un mayor procesamiento. Así, ya sería posible observar la distribución de cualquiera de ellos en los diferentes barrios. Una herramienta muy eficaz para ello son los mapas de coropletas, los cuales con un simple vistazo nos permiten averiguar en qué lugares una variable presenta valores más altos o más bajos según el sombreado de las diferentes áreas. Con el paquete *leaflet* de R podemos crear estos mapas, haciendo uso de cualquier criterio de segmentación para observar la distribución, como pueden ser los quintiles en su vertiente más sencilla. A modo de ejemplo, mostramos los construidos para los indicadores 24 y 25 referidos al porcentaje de población extranjera y al porcentaje de población de 65 años o más respectivamente, considerando como año de referencia el 2019. Podemos darnos cuenta del potencial de tales mapas temáticos en el hecho de que es inmediato observar en qué barrios tiene mayor incidencia la población de la tercera edad o la extranjera, lo que ya nos da una primera idea de las zonas que pueden estar más envejecidas o las que pueden manifestar un mayor número de migraciones.

Aquellos usuarios más avanzados de R habituados a la creación de mapas que quieran replicar estos u otros semejantes trabajando con los polígonos de los barrios de València sabrán que necesitarán el shapefile asociado. En el repositorio de GitHub ya nombrado se encuentra también un archivo que se puede obtener en la página web de la Oficina de Estadística indicada anteriormente llamado `barrios.rda`, el cual contiene un objeto de la clase *SpatialPolygonsDataFrame* que incluye la lista de los polígonos requeridos. Se puede descargar rápidamente desde el siguiente enlace: `https://raw.githubusercontent.com/JuanferMG/TFM/master/barrios.rda`

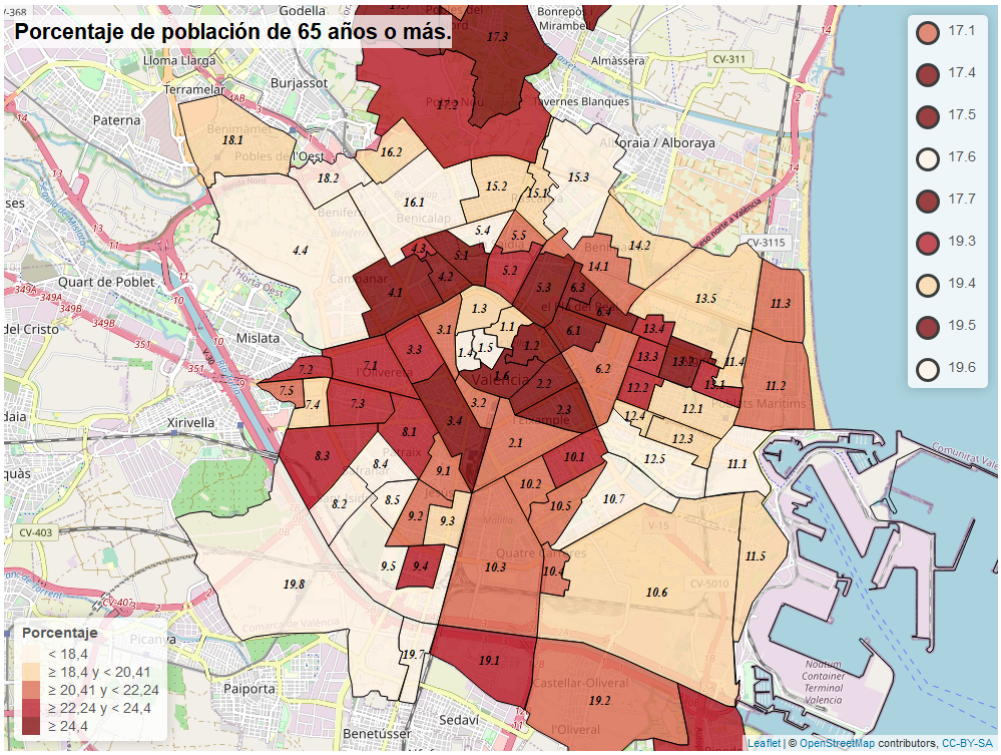

Figura 1: Porcentaje de población de 65 años o más en los barrios de València. 2019.

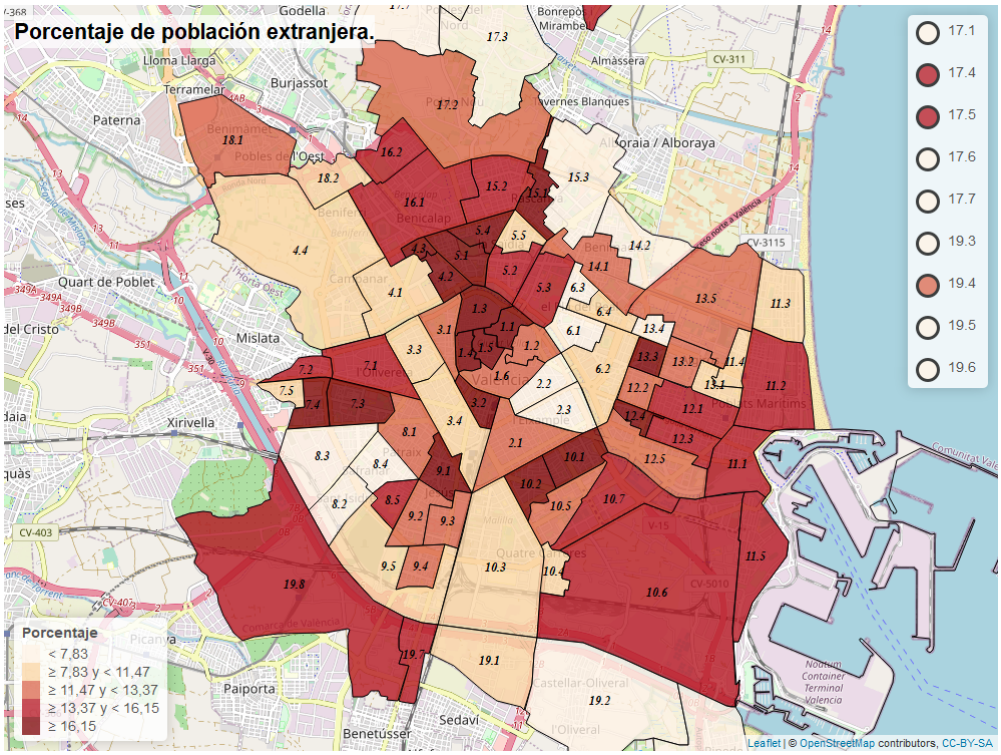

Figura 2: Porcentaje de población extranjera en los barrios de València. 2019.

5. Datos de renta

Construido nuestro conjunto de datos demográficos, el siguiente paso llevado a cabo se centró en la búsqueda de variables económicas que nos permitieran conocer el poder adquisitivo de los diferentes barrios de la ciudad. Para ello, recurrimos a un reciente proyecto del Instituto Nacional de Estadística que salió a la luz mediante una nota de prensa en septiembre de 2019: el **Atlas de Distribución de Renta de los Hogares** (ADRH) [23]. Este ambicioso proyecto, aún en fase experimental, ofrece datos estadísticos de renta neta media para todos los municipios del país, así como para los distritos municipales y las secciones censales, cuyas áreas geográficas alberguen al menos 500 habitantes. El estudio pudo realizarse gracias al enlace de datos tributarios proporcionados por la Agencia Tributaria (AEAT), así como de las haciendas forales de País Vasco y Comunidad Foral de Navarra. Inicialmente publicado con información para los años 2015 y 2016, en abril de 2020 se incluyeron los resultados del año 2017, asociados estos últimos con la población de 2018. En total, el ADRH hace referencia a 55.087 territorios y se puede acceder a todo su contenido desde el siguiente enlace: `https://www.ine.es/experimental/atlas/exp_atlas_tab.htm`. Los resultados de este proyecto, así como el resto de información estadística de la que dispone el INE, se almacena en la red mediante su sistema INEbase. Gracias a ello, se facilita la consulta de sus indicadores por medio de tablas con filtros para cada una de las 50 provincias y 2 ciudades autónomas de España de manera que se puede encontrar rápidamente el dato deseado de un territorio en concreto. No obstante, para la explotación estadística que vamos a realizar, trabajaremos con los 9 ficheros a nivel nacional que se ofrecen para su descarga:

- 0.1 Indicadores de renta media
- 0.2 Distribución por fuente de ingresos
- 0.3 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo
- 0.4 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y tramos de edad
- 0.5 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales fijos por sexo y nacionalidad
- 0.6 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales relativos por sexo
- 0.7 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales relativos por sexo y tramos de edad
- 0.8 Porcentaje de población con ingresos por unidad de consumo por debajo de determinados umbrales relativos por sexo y nacionalidad
- 0.9 Indicadores demográficos

Existen varios formatos de descarga posible, en nuestro caso hemos optado por el formato *CSV separado por ;* guardando y renombrando los archivos como *0.1.csv*, *0.2.csv*, ... , *0.9.csv* respectivamente. El motivo de trabajar con los datos a nivel nacional es que el INE no utiliza la división territorial de los barrios de València en su jerarquía geográfica, por lo cual nos veremos obligados a calcular los valores para este ámbito, programando para tal fin un código en R. La idea es utilizar la información de las secciones censales de València incluidas en el ADRH para construir la de los barrios. Así por ejemplo, para obtener los datos de renta del barrio 1.1. la Seu, el cual está integrado por las secciones censales 1001, 1002 y 1003, calcularíamos el valor promedio de cada indicador registrado en las correspondientes tres secciones para imputárselo al barrio.

A modo resumen y para entender mejor este proceso, vamos a mostrar de ejemplo los pasos a seguir con R para la explotación del archivo que hemos denominado 0.1.csv tras su descarga vinculado a la tabla 0.1 Indicadores de renta media. Es importante destacar que antes de capturar con R los datos se han eliminado las primeras 5 filas del CSV, así como las últimas 5 filas, ya que estas se dedicaban a definir la operación estadística y otras anotaciones, de tal modo que tras quitarlas el archivo solo contiene datos de los indicadores en cuestión. Después de esta breve depuración, procedemos a leer el archivo con el software estadístico,

V1	V2	V3	V4	V5	V6	V7
	Renta media por persona			Renta media por hogar		
	2017	2016	2015	2017	2016	2015
01001 Alegría-Dulantzi	13.281	13.086	12.936	34.618	34.373	33.702
0100101 Alegría-Dulantzi distrito 01						
0100101001 Alegría-Dulantzi sección 01001						
0100101002 Alegría-Dulantzi sección 01002						
01002 Amurrio	13.862	13.691	13.800	34.411	33.936	34.421

Tabla 1: Vista inicial de los datos de renta de toda España.

Observamos que hay zonas con valores perdidos, generalmente de secciones y lugares poco poblados. A continuación, ajustamos las cabeceras de modo que seamos capaces de identificar cada columna de datos con el indicador y el año correspondiente,

V1	V2	V3	V4	V5	V6	V7
	Renta media por persona 2017	Renta media por persona 2016	Renta media por persona 2015	Renta media por hogar 2017	Renta media por hogar 2016	Renta media por hogar 2015
01001 Alegría-Dulantzi	13.281	13.086	12.936	34.618	34.373	33.702
0100101 Alegría-Dulantzi distrito 01						
0100101001 Alegría-Dulantzi sección 01001						
0100101002 Alegría-Dulantzi sección 01002						
01002 Amurrio	13.862	13.691	13.800	34.411	33.936	34.421

Tabla 2: Datos de renta de toda España con la cabecera ajustada.

Aplicamos entonces un filtro sobre el conjunto para quedarnos con los datos referentes al municipio de València, seleccionando solo aquellas filas que contengan el caracter ≪ València ≫ en la primera columna,

V1	V2	V3	V4	V5	V6	V7
	Renta media por persona 2017	Renta media por persona 2016	Renta media por persona 2015	Renta media por hogar 2017	Renta media por hogar 2016	Renta media por hogar 2015
46250 València	12.453	12.133	11.865	31.456	30.725	29.986
4625001 València distrito 01	17.076	17.157	16.374	38.169	38.300	36.411
4625001001 València sección 01001	16.800	16.803	14.918	37.731	37.140	32.795
4625001002 València sección 01002	17.038	17.238	16.769	38.276	39.265	36.247
4625001003 València sección 01003	17.864	18.087	16.668	40.047	40.578	36.559

Tabla 3: Datos de renta del municipio de València.

Ahora normalizaremos el conjunto de datos. Para ello, primero hemos de sustituir los "." por un espacio vacío y a continuación las "," por un "." ya que en R no existe separador de millares y los puntos se usan como separador decimal. El conjunto transformado tendrá 3 columnas: Nombre, Anyo y Valor, siendo la columna Nombre la combinación del territorio sin el código numérico más el indicador,

Nombre	Anyo	Valor
València. Renta media por persona .	2017	12453
València distrito 01. Renta media por persona .	2017	17076
València sección 01001. Renta media por persona .	2017	16800
València sección 01002. Renta media por persona .	2017	17038
València sección 01003. Renta media por persona .	2017	17864

Tabla 4: Datos de renta normalizados del municipio de València.

Para quedarnos ahora con los datos de renta de las secciones censales, filtraremos el anterior conjunto seleccionando solo aquellos registros que contengan el caracter ≪ sección ≫ en la columna Nombre,

Nombre	Anyo	Valor
València sección 01001. Renta media por persona .	2017	16800
València sección 01002. Renta media por persona .	2017	17038
València sección 01003. Renta media por persona .	2017	17864
València sección 01005. Renta media por persona .	2017	17289
València sección 01007. Renta media por persona .	2017	21431

Tabla 5: Datos de renta de las secciones censales de la ciudad de València.

De este modo hemos conseguido reducir la base de datos inicial de España hasta tener únicamente datos de las secciones de València. Ahora bien, para poder trabajar de forma óptima es necesario tener la información de la variable Nombre mejor dividida. Para hacerlo, incorporaremos dos variables más: Indicador y SC. Ambas serán códigos que simplifiquen la manipulación de los datos. En Indicador codificamos como el 1 la *Renta media por persona* y como el 2 la *Renta media por hogar*, mientras que en SC incluimos el código numérico propio de la sección que aparece en la primera columna Nombre,

Nombre	Anyo	Valor	Indicador	SC
València sección 01001. Renta media por persona	2017	16800	1	1001
València sección 01002. Renta media por persona	2017	17038	1	1002
València sección 01003. Renta media por persona	2017	17864	1	1003
València sección 01005. Renta media por persona	2017	17289	1	1005
València sección 01007. Renta media por persona	2017	21431	1	1007

Tabla 6: Datos de renta codificados de las secciones censales de la ciudad de València.

Llegados a este punto, ahora la dificultad radica en asociar cada sección a uno de los 85 barrios con los que trabajamos. Haremos uso de otra publicación de la Oficina de Estadística del Ayuntamiento de València titulada *Evolución del Seccionado Censal* [38], en la cual podemos ver a qué barrio corresponde cada sección. Construir esta matriz de asociación que relacione cada sección con un barrio es una tarea delicada y algo costosa (existen 590 secciones censales en la ciudad), pero es necesaria para poder avanzar en el análisis por barrio. Tras haber realizado esta asignación, añadimos al conjunto de datos una última variable llamada BA con el número del barrio al que pertenece. Cabe destacar que una sección censal incluye por ley un máximo de 2.000 electores y un mínimo de 500 [29, art. 23]. Por tal motivo, los barrios 17.1 Benifaraig, 17.4 Cases de Bàrcena y 17.5 Mauella que tienen poca población se ven obligados a compartir la misma sección censal, lo que hará que tengan imputados los mismos datos de renta. La misma situación se repite con los barrios 19.7 la Torre y 19.8. Faitanar. No obstante, puesto que en nuestro trabajo habíamos fusionado los barrios 19.7 y 19.8 y también los barrios 17.4 y 17.5, los únicos datos de renta que se verán repetidos en nuestro conjunto de datos serán los de estos últimos con los del barrio 17.1. Benifaraig.

Nombre	Anyo	Valor	Indicador	SC	BA
València sección 01001. Renta media por persona	2017	16800	1	1001	1
València sección 01002. Renta media por persona	2017	17038	1	1002	1
València sección 01003. Renta media por persona	2017	17864	1	1003	1
València sección 01005. Renta media por persona	2017	17289	1	1005	2
València sección 01007. Renta media por persona	2017	21431	1	1007	2

Tabla 7: Datos de renta de las secciones censales con el barrio asociado.

Finalmente, obtendríamos el valor de los indicadores de renta de los barrios para cada año calculando la media del valor de las secciones asociadas, como ya habíamos anticipado. Si repitiéramos este proceso con cada uno de los 9 archivos CSV descargados, veríamos que en total el INE ofrece 175 indicadores distintos, aunque este número se ve reducido a 40 para nuestro conjunto de datos debido a que aquellos que cruzan las variables de sexo, edad o nacionalidad requieren áreas con más habitantes. Entre los indicadores calculados para los barrios, además de los dos ya mencionados de renta media por persona y hogar, destacan la distribución por fuente de ingresos o diferentes umbrales de pobreza. El código R usado para la obtención de los indicadores del ADRH para los barrios de València se puede encontrar en la siguiente ruta: `https://raw.githubusercontent.com/JuanferMG/TFM/master/RentaBarrios.R` El conjunto final de datos de renta, junto con la tabla de codigos creada que permite identificar cada indicador del ADRH, se ha incorporado también con el nombre `DatosBarrioRenta.rda` al repositorio de GitHub creado ad hoc para este TFM y se puede descargar fácilmente desde el siguiente enlace: `https://raw.githubusercontent.com/JuanferMG/TFM/master/DatosBarrioRenta.rda`

La creación de la base de datos de renta es más complicada y a diferencia de la de demografía en esta se ha utilizado un software estadístico al requerirse un proceso de depuración y normalización más sofisticado, teniendo además que codificar 312.900 registros, lo cual también conlleva su coste computacional. No obstante, al igual que antes con los indicadores demográficos, ya sería posible crear por ejemplo un mapa de coropletas que nos permitiera conocer rápidamente aquellos barrios con la renta media por persona más alta, como el que se muestra en la figura 3 inferior. A pesar del potencial de este tipo de mapas temáticos, están basados en una técnica de segmentación muy básica como son los percentiles. Por este motivo, en este TFM se aplicarán técnicas estadísticas de mayor relevancia que he ido aprendiendo en estos dos años de estudios en el máster y que comentaremos a continuación en el siguiente apartado de **Análisis**, lo que nos permitirá obtener una clasificación de los barrios más notoria apoyada en una base científica destacable.

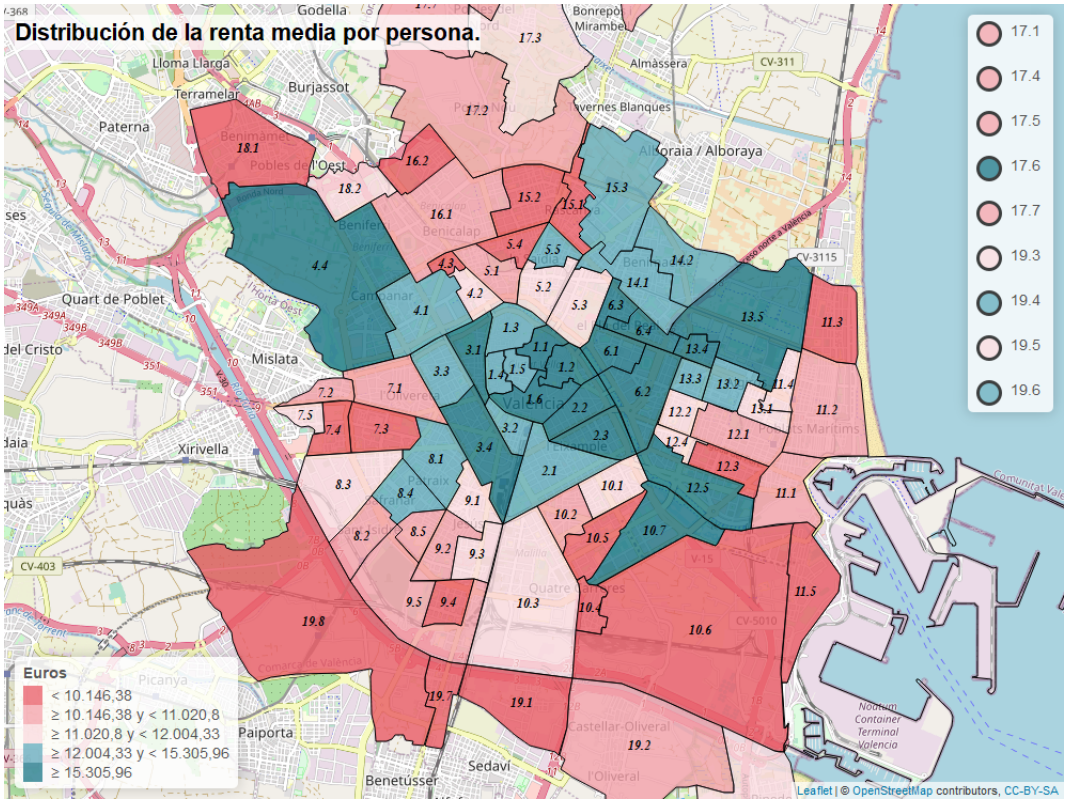

Figura 3: Renta media por persona en los barrios de València. 2017.

Análisis

6. Proyecciones demográficas mediante modelos ARIMA

Con el objetivo de aportar un valor añadido a los resultados alcanzados en este trabajo, realizaremos una proyección de los treinta indicadores demográficos calculados para el año **2021**, efectuando de este modo los posteriores análisis asociados a la segmentación de los barrios teniendo en cuenta la estructura poblacional esperada para el próximo año. El método predictivo que aplicaremos sobre los indicadores será uno de los procesos lineales estocásticos más conocidos, el de los procesos integrados autorregresivos y de medias móvil, denominado de forma abreviada **ARIMA**. Como se indica en el capítulo 8 de [42], los modelos ARIMA parten de los valores pasados de la serie temporal para establecer previsiones de los valores futuros sin considerar variables exógenas para ello y cuentan en su definición con una componente llamada *ruido blanco* que representa los errores del ajuste y proporciona la aleatoriedad del proceso. Así, considerando la serie temporal de cada indicador para cada barrio como una familia de variables aletorias parametrizadas por el tiempo las ajustaremos una a una a diferentes ARIMA identificando los modelos y estimando los parámetros mediante las funciones incorporadas en R en el paquete *forecast* que se dedican a ello. Una vez construidos los modelos asociados a cada serie, calcularemos los residuos de los ajustes. Comprobaremos por un lado que los residuos están normalmente distribuidos con el test de Shapiro y en segundo lugar que no están correlacionados entre sí haciendo uso de la función de autocorrelación. Ambas condiciones son indispensables para poder afirmar que el término del error es efectivamente ruido blanco y nos servirán como validación de los modelos. Finalmente, predeciremos los valores a corto plazo para los años 2019, 2020 y 2021, quedándonos con los de 2021 para conformar las variables demográficas del conjunto de datos sobre el cual aplicaremos el posterior análisis factorial.

A modo de ejemplo y para visualizar los pasos que acabamos de describir, mostraremos el ajuste llevado a cabo para el Indicador 25 correspondiente al porcentaje de población de 65 años o más en el barrio 13.5 la Carrasca, barrio en el que se encuentra la Facultad de Economía de la Universitat de València. Para ello, en primer lugar observamos los valores recopilados y calculados en la base de datos demográficos de la serie en cuestión:

Año	Valor
2004	9.38
2005	9.44
2006	9.87
2007	10.44
2008	10.93
2009	11.89
2010	12.56
2011	13.26
2012	14.02
2013	14.91
2014	15.59
2015	16.17
2016	17.04
2017	17.88
2018	19.08

Tabla 8: Evolución del porcentaje de población de 65 años o más en el barrio de la Carrasca.

Una vez capturados los valores en R, los cuales utilizaremos como predictores, calibraremos las componentes del modelo que mejor se ajusta a la serie con la función auto.arima() y construiremos el modelo ARIMA asociado con la función de nombre análogo arima(). A continuación, calularemos los residuos de la regresión modelizada, los cuales se pueden hallar de forma sencilla con la función residuals(). Sobre este conjunto de residuos plantearemos el test de normalidad de Shapiro-Wilk con la función shapiro.test(), obteniendo por pantalla lo siguiente:

```
Shapiro—Wilk normality test

data:  modelo.res[[k]]
W = 0.96783, p—value = 0.8248
```

El test de Shapiro-Wilk contrasta la hipótesis nula de normalidad de una variable, en nuestro caso, de si los residuos del modelo ajustado están normalmente distribuidos. Un valor tan alto del *p-value* nos impide rechazar dicha hipótesis y nos permite aceptar la normalidad de los residuos.

Para comprobar ahora que los residuos no presentan relaciones entre sí, ejecutaremos sobre ellos la función de autocorrelación con el comando acf(). Esta función proporciona la correlación de una serie consigo misma para diferentes retardos en el tiempo y nos devuelve un correlograma con el que poder analizar si los valores de los sucesivos residuos influyen en los posteriores.

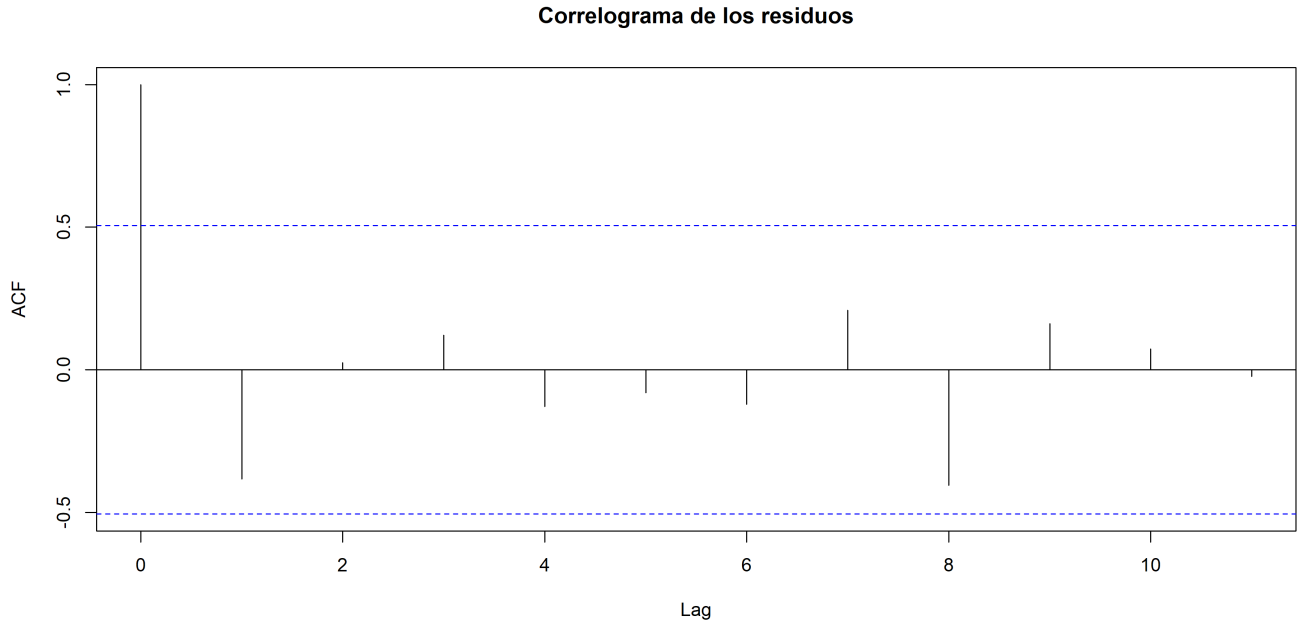

Figura 4: ACF de los residuos del modelo ARIMA ajustado a la serie de la Tabla 8.

Las bandas horizontales de color azul que aparecen en el gráfico indican los límites para considerar significativa una observación, pudiendo afirmar con un nivel de confianza del 95 % que los valores contenidos entre ambas bandas no son significativos. Como podemos ver en el correlograma, sin contar el retardo en 0 en el que obviamente la correlación de la serie temporal de datos con ella misma siempre es 1, podemos afirmar que no se aprecia autocorrelación y que por tanto los residuos son independientes entre sí al carecer completamente de estructura, lo que nos permite afirmar que son ruido blanco.

Una vez obtenido el modelo, predeciremos los próximos valores a corto plazo de la serie con la función forecast(), la cual nos muestra por pantalla los porcentajes esperados para los años 2019, 2020 y 2021, así como los intervalos en los que se espera se encuentre el valor tanto con un nivel de confianza del 80 % como del 95 %, como se puede ver a continuación,

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2019	20.28	19.97	20.60	19.81	20.76
2020	21.48	20.79	22.18	20.42	22.55
2021	22.68	21.52	23.85	20.91	24.46

Tabla 9: Predicciones a corto plazo de la serie de la Tabla 8.

Si nos fijamos en los intervalo de confianza del 95 % de la tabla, vemos que para 2019 se predijo que el porcentaje de población de 65 años o más en el barrio de la Carrasca estuviera entre el 19.81 % y el 20.76 %, mientras que en 2020 se estimó fuera mayor o igual que 20.42 % y menor o igual que 22.55 %. Así, debido a que en el transcurso de la realización de este trabajo se ha actualizado la publicación del Padrón Muncipal de Habitantes para la Ciudad de València, ya es posible comparar el valor observado para los años 2019 y 2020 con los esperados según el modelo. En 2019 el porcentaje registrado ha sido del 19.85 %, mientras que en 2020 ha subido hasta el 21.50 %, por lo que en ambos casos los valores observados caen dentro de los respectivos intervalos de confianza de sus estimaciones.

De este modo, programando una función que permita repetir de forma automática este proceso para cada serie y cada barrio, en total obtenemos 2.550 modelos ajustados. En cuanto a la bondad del ajuste de todos ellos, cabe señalar que la gran mayoría cumplen los requisitos de normalidad e independencia de los residuos, pero que hay un número reducido de modelos que no se antojan del todo válidos. No obstante, puesto que representan un porcentaje bajo respecto del total, no afectará de forma significativa a los resultados de este trabajo. En concreto, de las 2.550 series de residuos obtenidas, 93 de ellas no superan el test de Shapiro considerando para los p-valores un $\alpha = 0.01$, mientras que hay 10 de ellas que presentan autocorrelación en 2 o 3 retardos.

Es importante destacar que a corto plazo es más probable que no haya variaciones importantes en las series temporales y la precisión de la estimación sea más alta que a medio y a largo plazo. Por tal motivo, se ha decidido llevar a cabo las proyecciones con un horizonte de 3 años con la finalidad de reducir los inevitables errores en la predicción que se pueden obtener considerando un mayor número de años. No obstante, aunque existen manuales de referencia orientados a pequeñas poblaciones como [19], es difícil realizar proyecciones demográficas y aun más difícil acertar porque siempre pueden suceder acontecimientos imprevistos que adulteren las series, como es el caso de una pandemia.

Ahora, una vez calculadas las proyecciones de los 30 indicadores demográficos para 2021 que conformarán el conjunto de datos de partida para la realización del análisis factorial que veremos a continuación, añadiremos al conjunto 3 variables más provinientes de la base de datos de renta. En este caso, puesto que la serie histórica de datos de renta disponible solo incluye tres años y presenta valores bastante estables, no realizaremos proyecciones, sino que computaremos los indicadores como medias ponderadas por las poblaciones de los barrios en los distintos años. Los 3 indicadores en cuestión serán:

- Indicador 31: Renta media por persona.
- Indicador 32: Renta media por hogar.
- Indicador 33: Población con ingresos por unidad de consumo por debajo del 60 % de la mediana.

7. Análisis factorial exploratorio (*EFA*)

El análisis factorial es una técnica de reducción de dimensionalidades muy conocida y ampliamente utilizada en el campo de la estadística que "sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables" [16, p. 1]. El objetivo principal de aplicar dicha técnica es reducir el conjunto de datos inicial de 33 variables a uno que tenga pocas, pero que sean capaces de explicar la variabilidad original de los datos. Tal cual se explica en [27], estas nuevas componentes, que llamaremos **factores**, nos permitirán simplificar la información extrayendo los conceptos subyacentes entre las variables correlacionadas. Por tanto, para llevar a cabo el análisis factorial y asegurarnos de que podemos factorizar las variables de forma eficiente, es importante asegurarse de que en las variables iniciales la correlación sea latente. Para ello, graficaremos la matriz de correlaciones donde se puede apreciar fácilmente la dependencia entre los indicadores. Es importante señalar que previamente a dar comienzo al análisis multivariante con R todas las variables han sido estandarizadas con la función scale(), procedimiento recomendable para deshacerse de las variaciones sistemáticas presentes en los datos [35, p. 9].

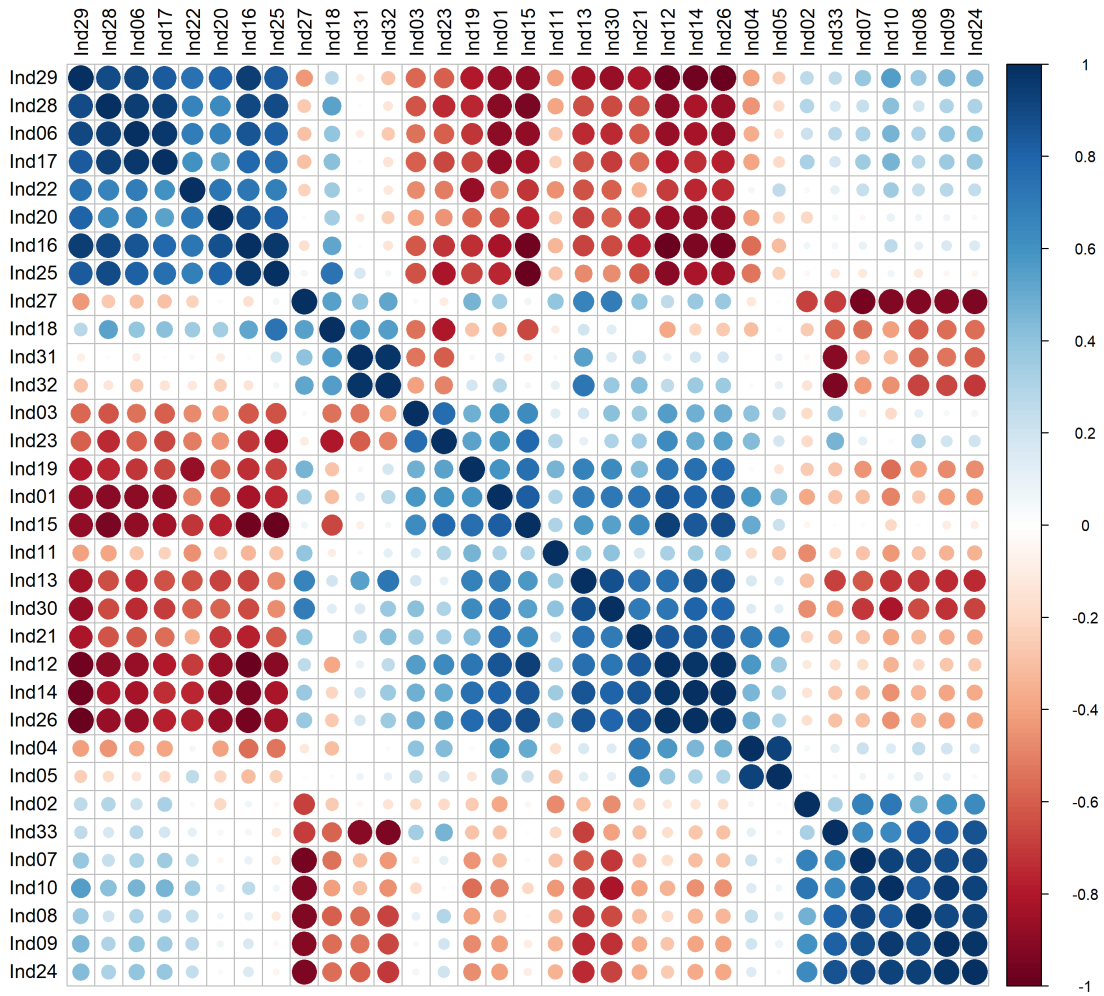

Figura 5: Representación gráfica de la matriz de correlaciones del conjunto de indicadores elaborados.

Si observamos la estructura de la matriz nos percatamos de que existen indicadores entre los cuales la correlación es muy alta, pero esta no se da en todo el conjunto. Por tanto, si queremos asegurarnos de que la correlación intrínseca en los datos es suficiente, podemos usar dos métodos válidos para tal fin: comprobar si el determinante de la matriz de correlaciones es aproximadamente 0 o estudiar si el índice de Kaiser-Meyer-Olkin (KMO) es mayor que 0,7 [17], [3, p. 8]. En nuestro caso, computando ambos parámetros con R obtenemos que el determinante de la matriz es prácticamente 0 y el índice KMO es aproximadamente 0,75, por lo que asumimos que podemos factorizar las variables de manera eficiente [34, p. 284]. Estas consideraciones previas son relevantes para garantizar un análisis factorial sólido y con fundamento, puesto que si todas las variables fueran independientes entre sí la aplicación de esta técnica carecería de sentido.

Ahora, el objetivo se centra en determinar el número de factores óptimo a extraer. Existen mútiples métodos creados para tal fin, muchos de ellos programados para su uso en R como son el criterio del análisis paralelo, la regla de Kaiser o el método de las coordenadas óptimas, los cuales pueden ser ejecutados a la vez con la función nScree(). La idea sigue siendo escoger un número bajo de factores, pero que conserven en la medida de lo posible la mayor información del conjunto original. Este paso del análisis factorial es por tanto más subjetivo y determinar el número exacto de factores depende en gran parte del investigador. Si bien los métodos mencionados sirven para tener una orientación de la adecuación en la elección, el hecho de que según el que se use se nos aconseje un número óptimo de factores distinto nos indica que no existe una única solución factorial buena, sino que hay un abanico de posibilidades en el cual elegir. La herramienta que vamos a utilizar en este trabajo para decidir el número de factores será la gráfica de sedimentación, que muestra los autovalores según el número de factores. Los autovalores están asociados al concepto de la varianza y nos dan una visión del peso que tiene un factor en comparación con una variable original. Los autovalores superiores a 1 nos indican que el factor logra explicar más varianza que una variable por sí sola, por lo que en nuestro caso escogeremos como máximo 7 factores.

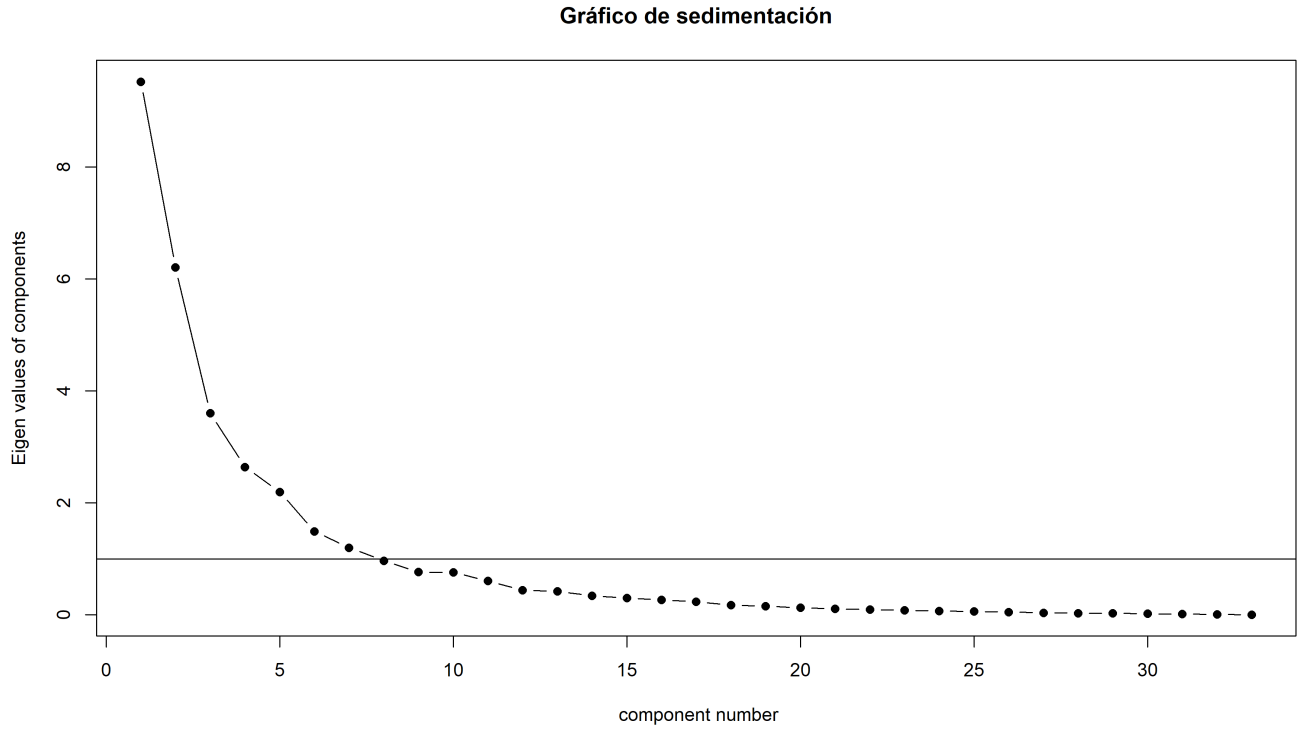

Figura 6: Autovalores del análisis factorial calculados con la función scree() del paquete *psych* de R.

A la vista del gráfico, puesto que podemos apreciar que a partir del tercer factor la ganancia en la variabilidad explicada alcanzada al considerar otra componente ya no es tan notable, vemos conveniente fijar en 3 el número de factores a extraer.[1] Así, tomada esta decisión, el siguiente paso es decidir cómo factorizar las variables. Para ello, usaremos el método cuyos factores resultantes expliquen mayor porcentaje de la variablidad del total del conjunto de datos. Probaremos con 11 procedimientos de factorización diferentes integrados en R basados en tres métodos:

- El método de **Máxima Verosimilitud** (que denotaremos **MLE**, del inglés *maximum likelihood estimation*). Se puede ejecutar con la función factanal() y entre sus argumentos existe la opción de incluir el tipo de rotación, con el cual poder cambiar la forma en la que se relacionan los factores con las variables y obtener así distintas soluciones factoriales. Computaremos la factorización sin rotación y con las rotaciones varimax y promax admitidas en la función.
- El método de las **Componentes Principales** (que abreviaremos como **PC**, del inglés *principal components*). Con la instrucción principal() obtenemos los factores y al igual que antes podemos especificar la rotación de los ejes asociados con la descomposición del valor propio. Cabe destacar que los factores extraídos se etiquetan como PC_i si no se establece rotación, como RC_i si las componentes principales se rotan de forma ortogonal (como es el caso de la rotación varimax, promax, quartimax o cluster) y como TC_i si las componentes principales son transformadas oblicuamente (como es el caso de rotar con la rotación simplimax u oblimin).
- El método de los **Ejes Principales** (que llamaremos **PA**, del inglés *Principal Axis*). Para este método usaremos la función principalAxis().

Como ya hemos mencionado, el criterio que seguiremos para decantarnos por un método de factorización será el de escoger aquella solución factorial que de entre todas sea capaz de explicar un mayor porcentaje de la varianza. De este modo, extrayendo los factores con R y analizando los outputs podemos ver que la mejor opción es utilizar la solución factorial devuelta por el método de las Componentes Principales con la rotación **simplimax**, en el cual la varianza explicada por cada factor es más alta. En su conjunto la solución explica el 64 % de la varianza total, un porcentaje satisfactorio a la vista de que en la literatura se suele establecer como umbral mínimo el 60 % [31, p. 77]. En la tabla 10 podemos ver de forma más detallada la proporción de varianza acumulada entre los factores para cada uno de los métodos aplicados.

	Factor 1	Factor 2	Factor 3
MLE None	0.2657	0.4099	0.5435
MLE Varimax	0.2648	0.4241	0.5435
MLE Promax	0.2651	0.4312	0.5456
PC None	0.2886	0.4766	0.5857
PC Varimax	0.2741	0.4508	0.5857
PC Promax	0.2754	0.4568	0.5903
PC Quartimax	0.2773	0.4616	0.5857
PC Oblimin	0.2741	0.4544	0.5800
PC Simplimax	0.3079	0.5218	0.6399
PC Cluster	0.2823	0.4498	0.5841
PA	0.2781	0.4533	0.5433

Tabla 10: Proporción de la varianza total explicada por los fatores según método.

[1]Para determinar el número de componentes exactas con más solidez, en este trabajo se llevó a cabo la realización del análisis factorial y su posterior exploración de las cargas factoriales mediante el EFA definido más adelante con 2, 3, 4 y 5 factores. La solución factorial que permitía una interpretación de los factores más sencilla y consistente fue la implementada con 3 factores que se muestra en este apartado.

Llegados a este punto, se lleva a cabo con la función principal() del paquete `psych` de R la factorización del conjunto de 33 indicadores con 3 factores mediante el método de las Componentes Principales considerando la transformación oblicua simplimax como rotación. Como ya hemos dicho, estos factores condensan la información subyacente de todas las variables facilitando la comprensión de los datos. Ahora bien, como se indica en [21], el aspecto complicado del análisis factorial es interpretar los factores en sí y deberemos entender su significado para poder darles un nombre adecuado. Para ello, nos interesa que en la medida de lo posible cada factor haga referencia a un grupo distinto de variables. Los factores estarán relacionados con todas los indicadores, pero cada factor incidirá más en algunos que en otros. Para estudiar cómo están relacionados los factores con las variables estudiaremos los *loadings* o las cargas factoriales.[II] Mostramos a continuación las cargas superiores a 0,5 resultantes de la solución, con el fin de fijarnos solamente en aquellas variables que condicionan en mayor medida a los factores.

Listing 1: Cargas factoriales iniciales del conjunto de indicadores

```
Standardized loadings (pattern matrix) based upon correlation matrix
         TC1    TC2    TC3     h2     u2  com
Ind01  0.602                0.3513 0.6487 1.05
Ind02                       0.1964 0.8036 1.91
Ind03                       0.1933 0.8067 2.77
Ind04                0.512  0.4503 0.5497 2.68
Ind05                0.557  0.3613 0.6387 1.54
Ind06 -0.638                0.4970 0.5030 1.19
Ind07         0.753         0.6663 0.3337 1.22
Ind08         0.775         0.5784 0.4216 1.01
Ind09         0.827         0.7036 0.2964 1.01
Ind10         0.729         0.7133 0.2867 1.26
Ind11                       0.0542 0.9458 2.83
Ind12  0.994                0.9222 0.0778 1.27
Ind13  0.544 -0.534         0.7785 0.2215 2.80
Ind14  0.847                0.7457 0.2543 1.34
Ind15  0.921                0.8048 0.1952 1.38
Ind16 -1.010                0.9064 0.0936 1.24
Ind17 -0.517                0.4154 0.5846 1.62
Ind18        -0.677  0.501  0.6374 0.3626 2.53
Ind19                       0.3218 0.6782 1.79
Ind20 -0.628                0.4280 0.5720 1.81
Ind21  0.556         0.600  0.5674 0.4326 1.99
Ind22                       0.2538 0.7462 1.69
Ind23  0.538                0.5732 0.4268 2.95
Ind24         0.926         0.8579 0.1421 1.00
Ind25 -0.922 -0.531         0.8386 0.1614 1.62
Ind26  0.959                0.9189 0.0811 1.22
Ind27        -0.746         0.5706 0.4294 1.00
Ind28 -0.758                0.6646 0.3354 1.31
Ind29 -0.927                0.8746 0.1254 1.01
Ind30                       0.4755 0.5245 1.99
Ind31        -0.503  0.731  0.6542 0.3458 1.78
Ind32        -0.583  0.731  0.7361 0.2639 1.93
Ind33         0.752         0.6165 0.3835 1.56
```

[II] Las cargas factoriales nos indican el peso que tienen las variable en cada factor. Como se hace constar en [21], una buena solución factorial será aquella en que cada variable cargue alto en un factor y bajo en los demás, por lo que nuestro objetivo será producir factores con una mezcla de cargas altas y bajas y pocas cargas de tamaño moderado.

Las tres primeras columnas que se muestran llamadas $TC1$, $TC2$ y $TC3$ contienen las cargas factoriales para cada uno de los 33 indicadores. Las últimas 3 columnas hacen referencia a las comunalidades ($h2$), a las unicidades ($u2$) y a la complejidad de las cargas (com). Como bien se detalla en [30], la comunalidad nos muestra la proporción de variabilidad de cada variable que es explicada por los factores, la unicidad se puede interpretar como un error o residuo de cada variable de lo que no se explica por los factores comunes (nótese que $h2 + u2 = 1$) y finalmente la complejidad nos habla de las cargas cruzadas o *cross-loadings*, siendo más bajo cuanto más específico es un indicador de un factor. A la vista de los valores obtenidos, nos damos cuenta de que la calidad de la solución factorial está lejos de considerarse óptima. Para ello, nos vamos a dedicar a mejorar la calidad de los factores realizando un análisis factorial exploratorio, más conocido como *EFA* (del inglés Exploratory Factor Analysis), siguiendo los pasos expuestos en [3]. La idea es eliminar aquellos indicadores del conjunto de datos que añaden complejidad al sistema y aportan poco valor a la solución factorial. Los indicadores se deben descartar uno a uno hasta obtener un grupo más reducido que consideremos suficientemente bueno, por lo que se debe repetir el proceso del EFA cada vez se elimine una variable. Los criterios a seguir para eliminar una variable serán los siguientes:

1. Quitar el indicador que menos comunalidad tiene de entre aquellos que tienen una comunalidad inferior a 0,25.

2. En caso de que todos los indicadores tengan una comunalidad superior o igual a 0,25 quitar aquel indicador cuyas cargas factoriales sean inferior a 0,5 en los 3 factores. Si hay varios, quitar aquel que tenga una comunalidad más baja.

3. En caso de que todos los indicadores tengan una comunalidad superior o igual a 0.25 y loadings superiores o iguales a 0,5 en algún factor, quitar aquel indicador que presente mayor complejidad de entre aquellos que tengan una complejidad superior a 1,9.

En el momento en que ningún indicador sea susceptible de ser eliminado en función de los criterios anteriores, finalizaremos con el análisis factorial exploratorio. Con la intención de que se entienda mejor el proceso descrito, adjuntaremos al final del trabajo en el Anexo II los outputs del EFA, mencionando a continuación el criterio que ha conducido a las variables descartadas a su eliminación. Comenzamos a quitar por tanto variables que aportan poca información de forma escalonada, paso a paso y del siguiente modo:

1. Empezamos quitando el Indicador 11, puesto que es el que menos comunalidad presenta: 0,0542 ($<$0,25).

2. Quitamos después el Indicador 2 al ser ahora el que menos comunalidad tiene: 0,189 ($<$0,25).

3. A continuación, quitamos el Indicador 3 al poseer menor comunalidad: 0,187 ($<$0,25).

4. Quitamos el Indicador 22, puestos que todas las cargas factoriales son inferiores a 0,5 siendo además de entre aquellos que también presentan loadings bajos el que menos comunalidad tiene (0,289).

5. Quitamos el Indicador 19, puestos que todas las cargas factoriales son inferiores a 0,5 siendo además de entre aquellos que también presentan loadings bajos el que menos comunalidad tiene (0,299).

6. Quitamos el Indicador 4, puestos que todas las cargas factoriales son inferiores a 0,5 siendo además de entre aquellos que también presentan loadings bajos el que menos comunalidad tiene (0,443).

7. Quitamos el Indicador 5 al ser el que menos comunalidad tiene: 0,179 ($<$0,25).

8. Quitamos el Indicador 30 al mostrar cargas factoriales por debajo de 0,5 en todos los factores.

9. Quitamos el Indicador 23 que manifiesta cross-loading por no ser específico de un factor al tener mayor complejidad: 2,70 ($>$1.9).

10. Quitamos el Indicador 13 al tener mayor complejidad: 2,98 ($>$1.9).

11. Quitamos el Indicador 18 al tener mayor complejidad: 2,64 (>1,9).

12. Quitamos el Indicador 20 al tener mayor complejidad: 2,25 (>1,9).

13. Quitamos el Indicador 21 al tener mayor complejidad: 1,97 (>1,9).

Tras esta depuración, observamos que los 20 indicadores que quedan tienen todos una comunalidad superior a 0,25, todos tienen loadings por encima de 0,5 en algún factor y presentan un coeficiente de complejidad inferior a 1,9, por lo que damos por conluido el EFA. La calidad de la solución factorial es ahora satisfactoria, lo cual podemos además comprobar con el **alfa de Cronbach**. Dicho instrumento estadístico nos permite medir la fiabilidad de la consistencia interna de los factores [15], es decir, cómo de relacionadas están las variables que los conforman. El alfa de Cronbach varía entre 0 y 1 y cuanto más bajo es menor será la consistencia, indicándonos en tal caso que hay poca relación entre las variables que ponderan en mayor grado a los factores. En [3, p. 6] se nos indica que un valor del alfa de entre 0,65-0,7 es el mínimo aceptable, mientras que en [15, p. 6] se menciona que un coeficiente comprendido entre 0,9-0,95 es considerado excelente. De este modo, haciendo uso de nuevo de la librería `psych`, en concreto de la función alpha(), computamos el alfa de Cronbach para cada factor y concluimos que la solución factorial alcanzada es óptima para continuar con la segmentación de los barrios basada en esta nuevas componentes.

	α
TC1	0.945
TC2	0.897
TC3	0.951

Tabla 11: Valor del α de Cronbach de cada factor.

Antes de dar paso al siguiente apartado referente al análisis de conglomerados, deberemos asignar un nombre a cada uno de los factores que nos ayude a interpretarlos. Para ello, conviene fijarnos en las cargas factoriales asociadas de cada variable. Si analizamos los loadings observamos lo siguiente:

- Los indicadores relacionados con el factor TC1 son el 1, 6, 12, 14, 15, 16, 17, 25, 26, 28 y 29. Si observamos los signos, los que inciden positivamente son la Tasa de Mortalidad, el Índice de Envejecimiento, el Índice de Sobreenvejecimiento, el porcentaje de población de 65 años o más, el porcentaje de hojas familiares con solo personas de 80 años o más y el porcentaje de hojas familiares sin menores. Vemos que todos estos tienen en común que se asocian a una población envejecida. Por contra, los indicadores que inciden con signo negativo son el Crecimiento Vegetativo, el Índice de Sundbarg, el Índice de Burgdofer, el Índice Generacional de Ancianos y el porcentaje de población de 15 años o menos. Estos indicadores se asocian a zonas rejuvenecidas, por lo que el signo negativo nos señala que altas puntuaciones de este factor se relacionarán con zonas poco rejuvenecidas. Por tanto, vemos conveniente denominar este factor TC1 **Vejez**.

- Los indicadores relacionados con el factor TC2 son el 7, 8, 9, 10, 24, 27, 31, 32 y 33. En concreto, inciden positivamente la Tasa de Inmigración, la Tasa de Emigración, la Tasa de Llegadas por Cambio de Domicilio, la Tasa de Salidas por Cambio de Domicilio, el porcentaje de población extranjera y el porcentaje de población con ingresos por unidad de consumo por debajo del 60 % de la mediana. Observamos que estos indicadores van ligados a los movimientos migratorios, tanto interurbanos como intraurbanos, que a su vez se asocian con un incremento de población extranjera. Con signo negativo ponderan el porcentaje de población nacida en la ciudad de València, la renta media por persona y la renta media por hogar. Esto nos indica que a parte de los movimientos migratorios este factor también alberga una componente económica negativa. No obstante, puesto que las cargas más fuertes se dan en las tasas asociadas a los movimientos de la población y a la ciudadanía extranjera, creemos adecuado llamar a este factor TC2 **Migración**.

- Los indicadores relacionados con el factor TC3 son el 31, el 32 y el 33, los tres provinientes de la base de datos de renta. Positivamente influyen la renta media por persona y la renta media por hogar, mientras que con signo negativo incide el porcentaje de población con ingresos por unidad de consumo por debajo del 60 % de la mediana. No cabe duda que estos indicadores se asocian con el poder asquisitivo de una zona, por lo que denominamos al factor TC3 **Riqueza**. Es importante señalar que el factor TC2 puede ser visto también como una componente que ajuste la riqueza, al contrarrestar los efectos del factor TC3.

Con los factores ya calculados y estudiados, ahora podemos obtener los valores que toman los 85 barrios en cada uno de ellos, es decir, podemos calcular las puntuaciones factoriales o *scores*. Estas puntuaciones serán las nuevas variables con las que trabajaremos a partir de ahora y se obtienen como una suma ponderada de los valores originales de los indicadores que conforman los factores. Con las puntuaciones ya podremos sacar conclusiones de las características intrínsecas de los barrios y además hacerlo de forma sencilla gracias a la reducción de dimensionalidades que hemos realizado. Una forma visual de representar las puntuaciones factoriales es mediante diagramas de dispersión, como los que se muestran en las figuras 7, 8 y 9. En ellos, hemos dibujado los 85 barrios en función de los valores que tienen para cada par de factores, coloreando los barrios según el distrito al que pertenecen para facilitar su búsqueda en el gráfico. Estos diagramas de burbujas nos dan bastante información de forma rápida. Así por ejemplo, podemos ver sin grandes complicaciones que el barrio más envejecido corresponde a *4.3 El Calvari*, el barrio con menos flujos migratorios y más autóctono es *19.5 el Palmar* y que el barrio más adinerado es *2.2 El Pla del Remei*. Con estos datos ya es posible analizar los barrios uno a uno. No obstante, lo que nos interesa es intentar clasificarlos en grupos cuyas características respecto a la Vejez, la Migración y la Riqueza sean parecidas. Para tal fin, efectuaremos en el siguiente apartado la técnica del clustering.

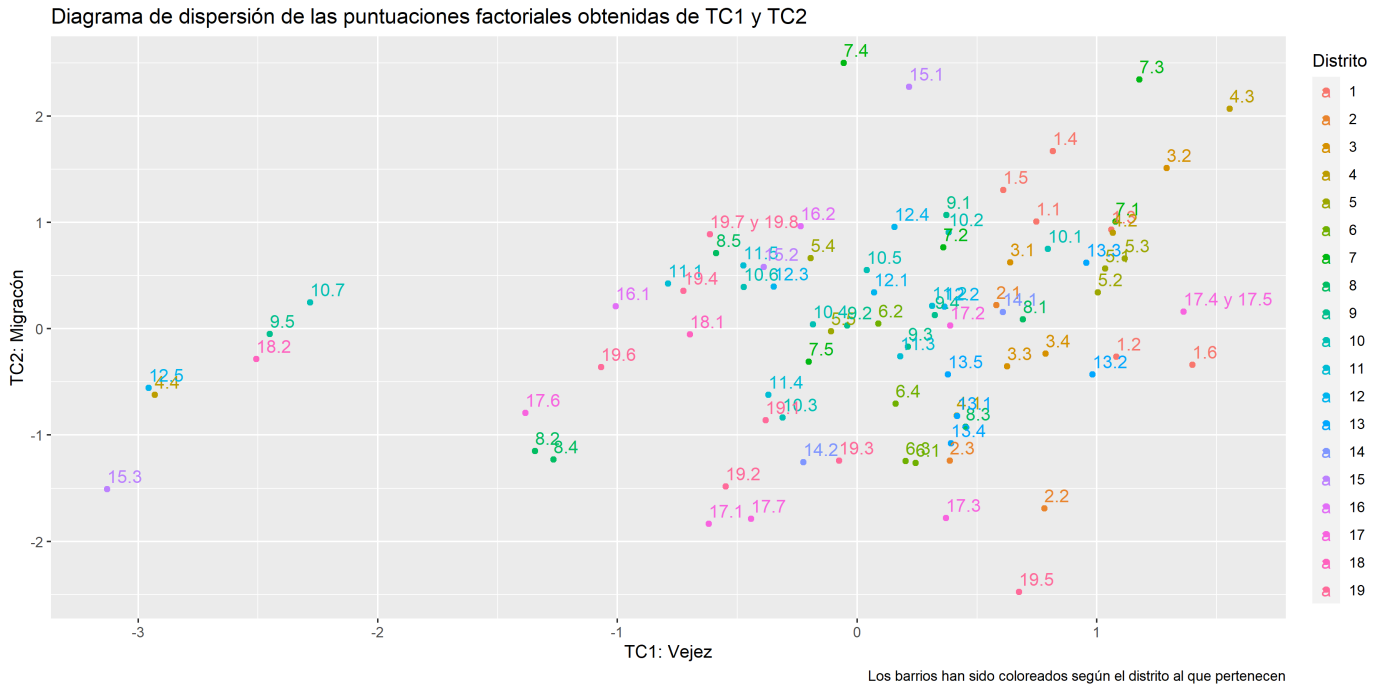

Figura 7: Puntuaciones factoriales TC1×TC2 correspondientes a cada barrio.

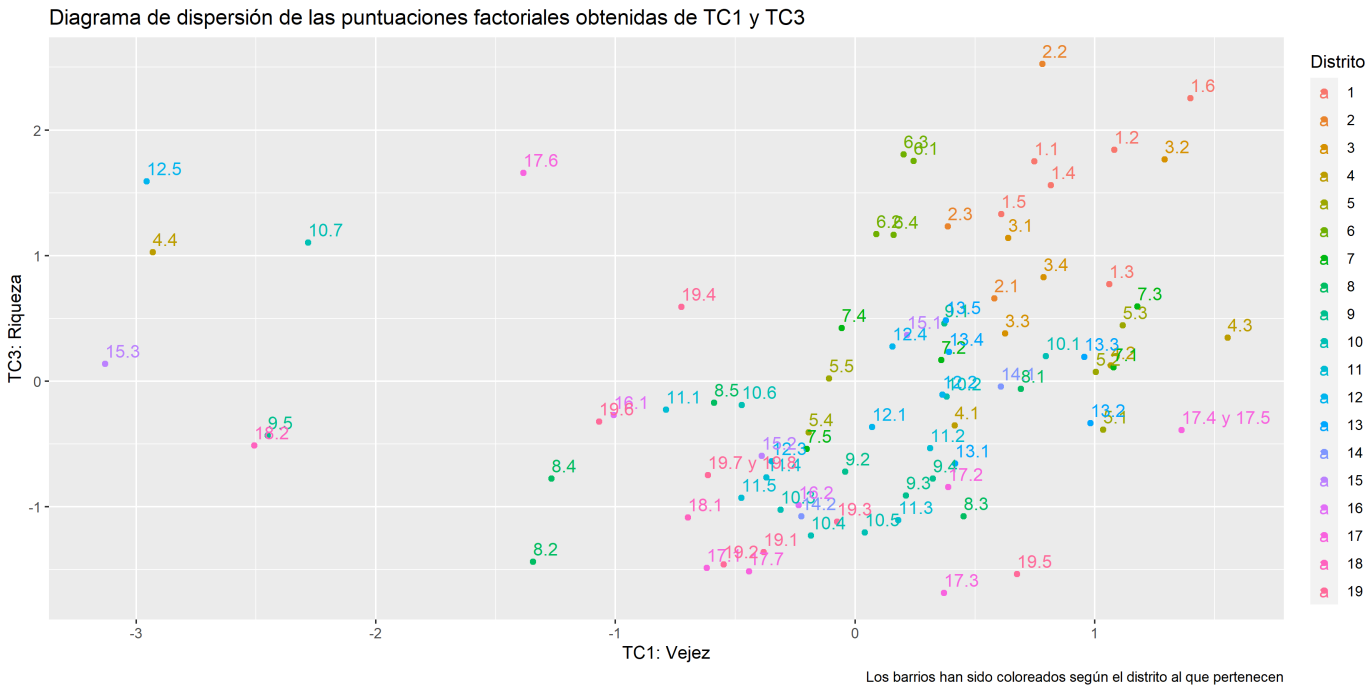

Figura 8: Puntuaciones factoriales TC1×TC3 correspondientes a cada barrio.

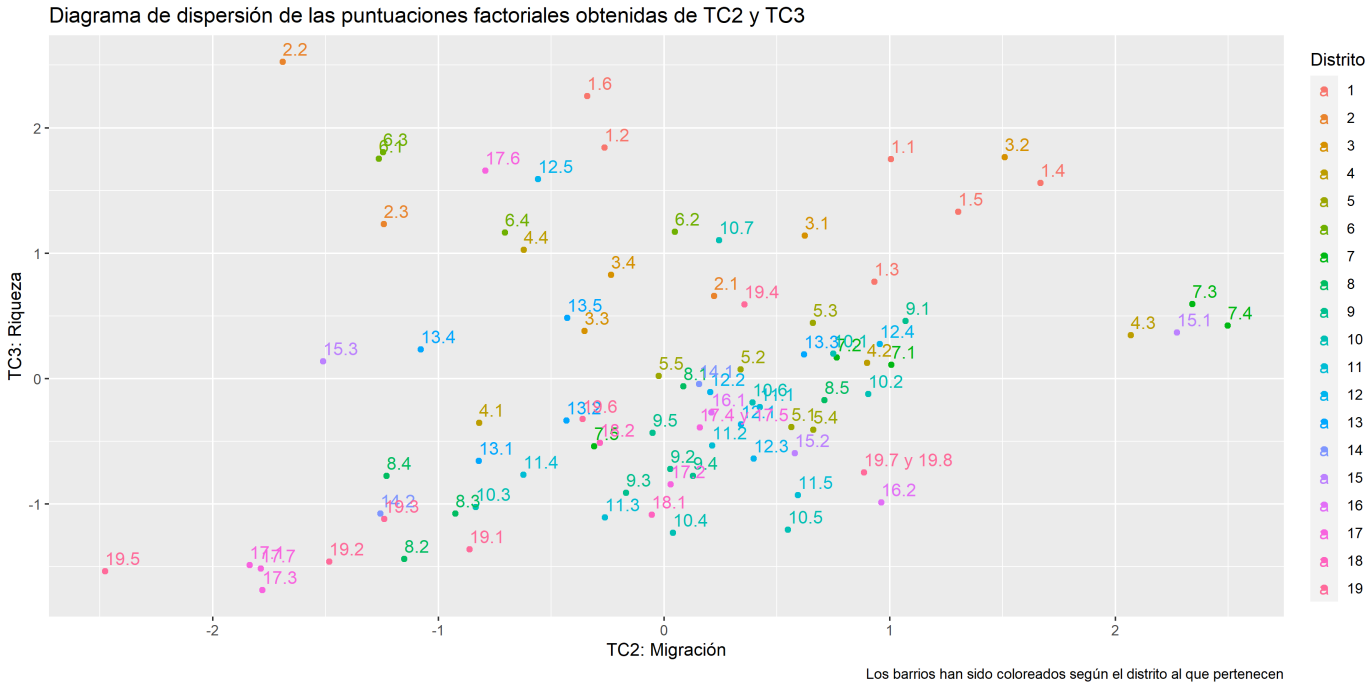

Figura 9: Puntuaciones factoriales TC2×TC3 correspondientes a cada barrio.

8. Análisis de conglomerados o *cluster*

El análisis de conglomerados, análisis de grupos o análisis cluster es un conjunto de técnicas multivariantes que se utilizan para clasificar en grupos homogéneos un conjunto de miembros o individuos y se encuentra encuadrado en los conocidos como métodos de aprendizaje no supervisados [28]. En nuestro caso, nesesitaremos aplicar un análisis cluster sobre los barrios para dividirlos en grupos con características etarias, migratorias y monetarias semejantes. Existen muchos métodos de clusterización y algoritmos de clasificación basados en criterios de similitud, divergencia o distancias y suelen diferenciarse en dos grandes clases: los métodos de clasificación jerárquicos, los cuales no obtienen una única partición, sino que se van haciendo particiones sucesivas minimizando distancias o maximizando medidas de similitud; y los métodos no jerárquicos, los cuales requieren que se les especifique un número fijo de grupos deseado. Para nuestro propósito, usaremos dos métodos de segmentación de los barrios de València. En primer lugar, escogeremos una de las técnicas de clustering más conocidas, el algoritmo particional de las K-Medias, el cual está englobado dentro de los métodos no jerárquicos. Fijaremos un número de grupos o **clusters** pequeño para obtener una clasificación inicial que nos otorgue una visión más general de la ciudad. Después, con el objetivo de profundizar en las diferencias entre los barrios, usaremos un método de clasificación jerárquica y dividiremos los barrios en un mayor número de clusters.

El método de las K-Medias está implementado en R bajo la función kmeans() y su uso está extendido debido a que se trata de un algoritmo rápido y sencillo. En pocas palabras, K-Medias es un proceso iterativo mediante el cual cada observación se asigna al grupo más cercano. Inicialmente se establecen de forma aleatoria las ubicaciones centrales de los grupos, conocidas como *centroides*. Una vez las observaciones han sido clasificadas a aquellos grupos con el centroide más cercano, se vuelven a calcular los centroides como ubicación promedia de los objetos que lo conforman y, a continuación, se reasignan los objetos a los centroides más próximos. Así se procede sucesivamente hasta que los centroides se estabilizan. Su mayor dificultad (al igual que en el resto de métodos de agrupación) es el de determinar el número óptimo de clusters, como bien se explica en [26]. No obstante, en R existe desde 2015 un paquete llamado `NbClust` que contiene una única función de nombre análogo que se dedica a examinar 30 índices distintos para decidir el número más adecuado de grupos a utilizar de entre un rango proporcionado, cuyo uso podemos comprender en profundidad mediante la lectura del artículo [11] escrito por sus autores. Si ejecutamos la función NbClust() estableciendo en sus argumentos que el método de análisis cluster requerido es K-Medias y que el número de clusters deseado está entre 2 y 9, por pantalla se nos muestra el diagrama de barras de la figura 10.

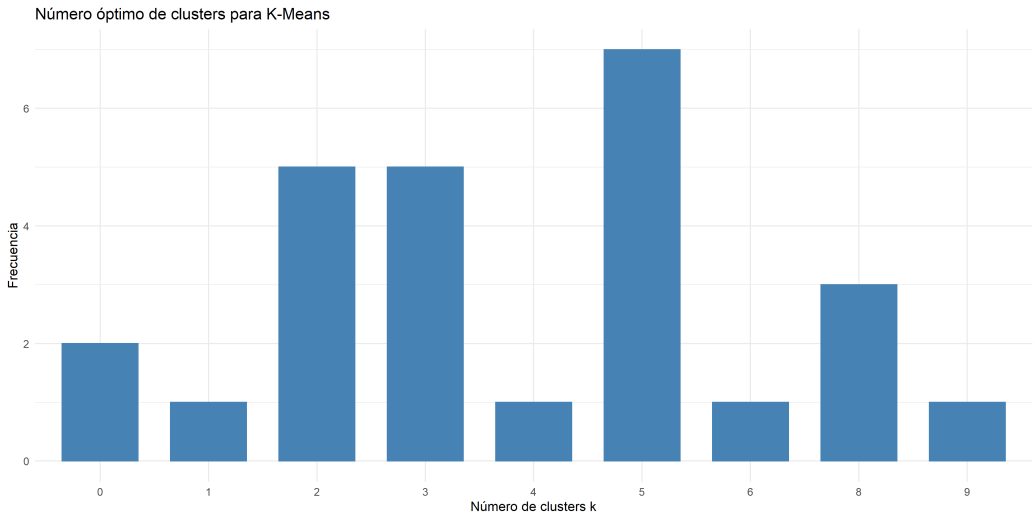

Figura 10: Número óptimo de clusters para K-Means determinado por el paquete de R *NbClust*.

Como podemos ver en el gráfico devuelto por la función NbClust(), se nos indica que la mejor opción para clasificar los barrios por el método de las K-Medias es utilizar **5** clusters. Por lo tanto, si fijamos en 5 el número de particiones deseado e introducimos la matriz de puntuaciones factoriales como el conjunto de variables que definen los barrios en el primer argumento de la función kmeans() ya mencionada, obtenemos de forma inmediata los distintos grupos con los 85 territorios clasificados. Podemos visualizar el resultado del análisis sobre los diagramas de dispersión de las figuras 7, 8 y 9 anteriores para comprender cómo se han agrupado, coloreando los distintos polígonos que albergan los barrios y conforman los clusters. Así, en las figuras 11, 12 y 13 se nos muestra la clasificación de los barrios sobre las puntuaciones factoriales para cada par de factores y en ellas es posible apreciar la ubicación de los centroides. Debemos fijarnos en los ejes de coordenadas de los diagramas para saber en qué cuadrante del gráfico se encuentra cada cluster y así interpretar adecuadamente las cualidades propias de cada grupo.

Si analizamos de forma breve las imágenes, vemos por ejemplo que en la figura 11, centrándonos en el eje del factor asociado a la Vejez, hay dos polígonos en la parte derecha con los barrios con puntuaciones positivas, otros dos que combinan puntuaciones positivas y negativas y un cluster a la izquierda claramente diferenciado con puntuaciones negativas, lo que nos indica que este último polígono contiene barrios singularmente rejuvenecidos. Por otra parte, si en el mismo gráfico nos fijamos ahora en el eje del factor asociado a la Migración, podemos apreciar que existen dos clusters con barrios que presentan casi en su totalidad puntuaciones positivas, mientras que en los otros tres predominan las puntuaciones negativas. No obstante, vemos que existen dos clusters con un carácter migratorio muy marcado, uno con una fuerte componente migratoria y con población extranjera situado en la parte superior del diagrama y otro con barrios en los que se dan pocos flujos y que presentan una índole más autóctona en la parte inferior. Finalmente, para ver las particularidades de cada grupo respecto al factor correspondiente a la Riqueza, nos podemos fijar en el eje vertical tanto de la figura 12 como de la 13. En este caso tenemos tres clusters que contienen barrios que en líneas generales disponen de puntuaciones factoriales positivas en cuanto a nivel adquisitivo y otros dos con valores casi en su totalidad negativos, siendo además estos dos últimos clusters los que a la vista de los polígonos representan un mayor número de barrios, lo que nos lleva a pensar que hay pocos barrios particularmente ricos.

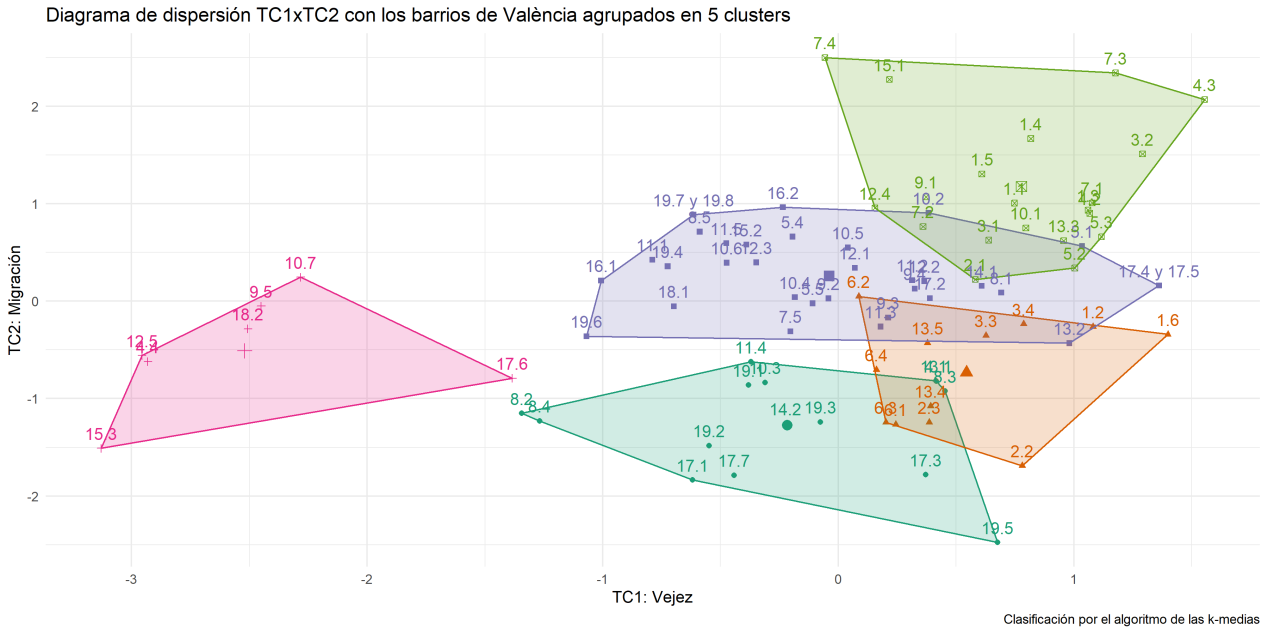

Figura 11: Barrios agrupados en 5 clusters sobre las puntuaciones factoriales TC1×TC2.

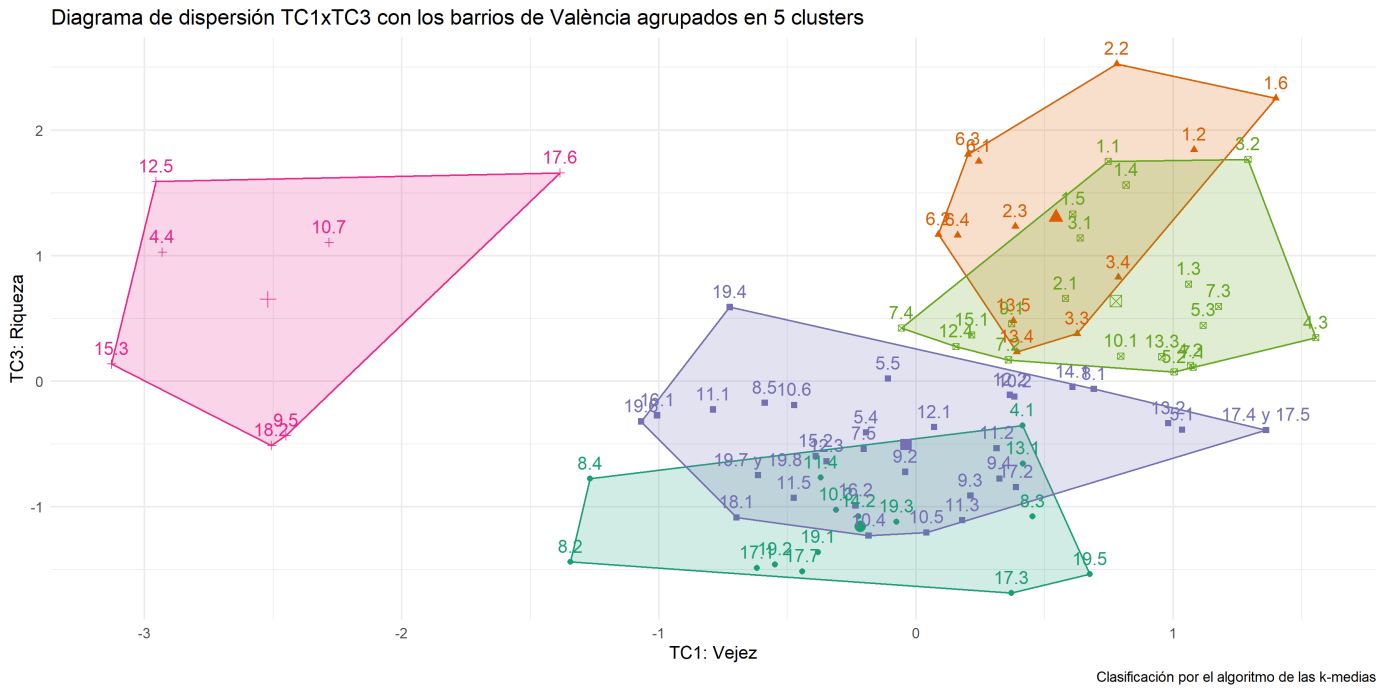

Figura 12: Barrios agrupados en 5 clusters sobre las puntuaciones factoriales TC1×TC3.

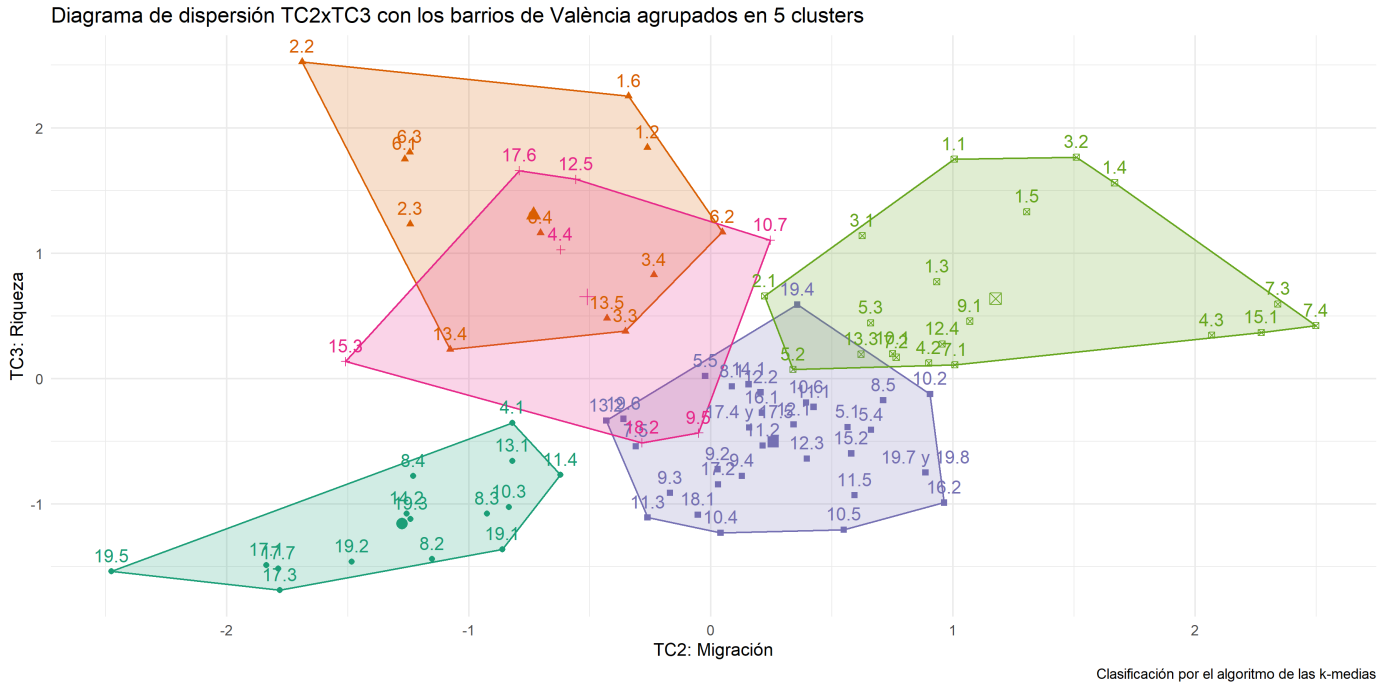

Figura 13: Barrios agrupados en 5 clusters sobre las puntuaciones factoriales TC2×TC3.

De esta forma, analizando las posiciones de los polígonos dentro de los diagramas con más detenimiento podemos tener una visión más precisa de las características propias de cada cluster, los cuales están formados por grupos homogéneos de barrios con un perfil etario, migrante y económico similares. Podemos dibujar los grupos en un mapa, coloreando las áreas de los barrios de València en función del cluster al que pertenecen. Así, manteniendo los colores de los diagramas de dispersión anteriores obtenemos el mapa que se muestra en la figura 14. Los diversos grupos pueden ser vistos ahora como 5 nichos de mercado distintos, ya que cada uno de ellos representan zonas de la ciudad cuyos habitantes de forma mayoritaria comparten unos características semejantes entre sí y al mismo tiempo diferentes a los barrios del restos de clusters.

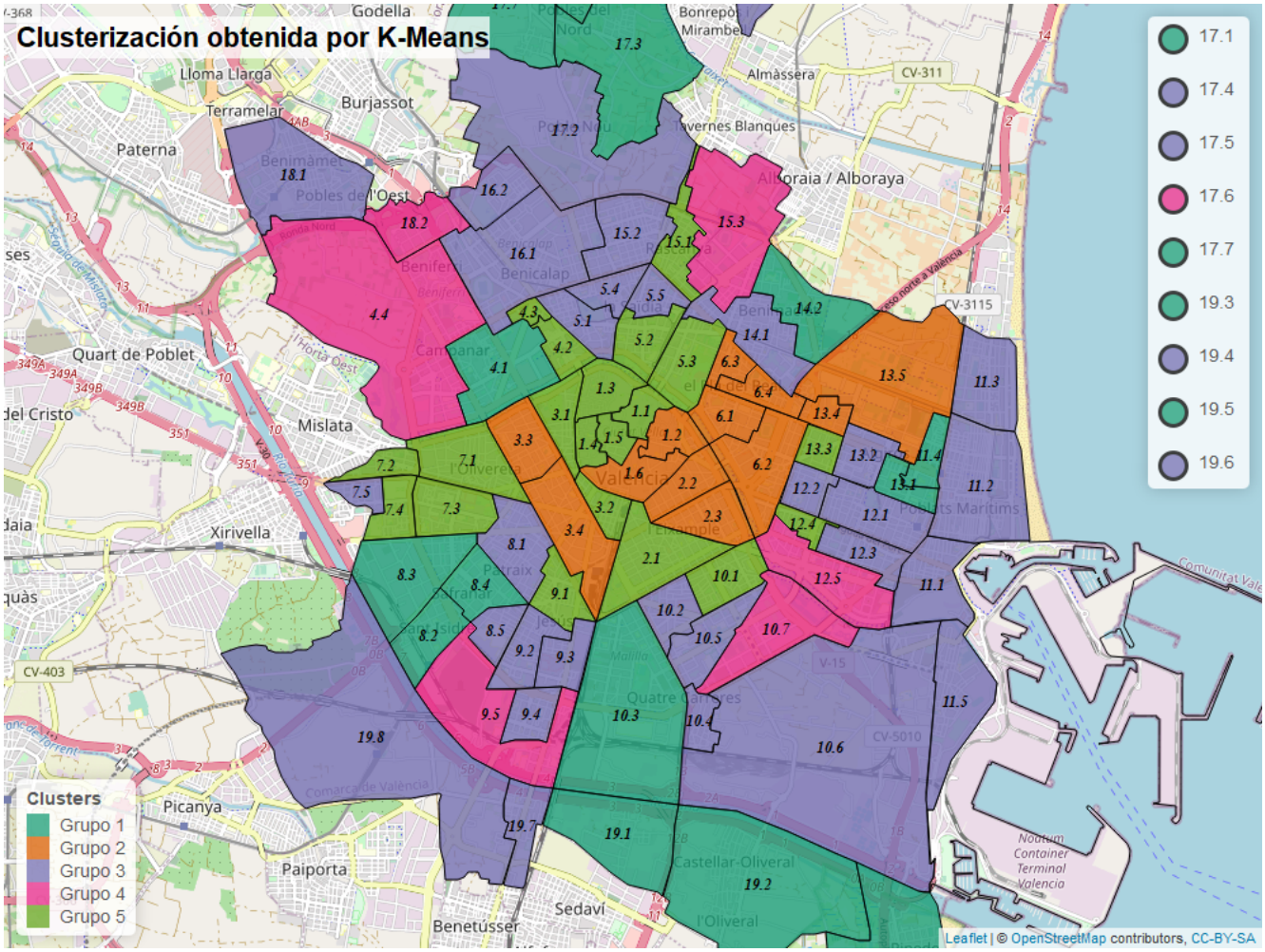

Figura 14: Mapa de València con los barrios agrupados según el algoritmo de las k-medias.

Veamos cuáles son las singularidades de cada grupo o nicho de mercado específico:

- **Grupo 1**: Compuesto por zonas rejuvenecidas en líneas generales, salvo los barrios 4.1, 8.3, 13.1, 17.3 y 19.5. Presenta muy poca migración y poca población extranjera en su conjunto y sus barrios se posicionan como territorios muy empobrecidos en comparación con el resto de la ciudad.
- **Grupo 2:** Formado por zonas envejecidas en su conjunto, con unos flujos migratorios escasos y una población generalmente autóctona. Este grupo se caracteriza sobre todo por estar **muy enriquecido** en su conjunto.

- **Grupo 3:** En este grupo no se aprecian rasgos etarios distintivos, pero sí una componente migratoria bastante elevada en líneas generales, a excepción de los barrios 7.5, 9.3, 11.3, 13.2 y 19.6. Además, exceptuando el barrio 19.4, los barrios que aquí se agrupan se encuentran bastante empobrecidos en su totalidad.

- **Grupo 4**: Integrado por barrios **muy rejuvenecidos** en su conjunto, siendo esta componente etaria la más significativa que subyace en ellos. Presenta poca migración prácticamente en su totalidad, salvo en el barrio 10.7 donde la inmigración incide con más fuerza. Es un grupo bastante enriquecido en líneas generales, a excepción de los barrios 18.2 y 9.5 que disponen de un nivel económico inferior.

- **Grupo 5**: Encontramos en este grupo territorios en los que reside una población envejecida, salvo en el caso del barrio 7.4. Es característico de manera singular en estos lugares que haya **mucha migración** en todo su conjunto. Además, se trata de barrios generalmente adinerados.

A pesar de que esta clasificación ya nos permite hacernos una idea de cuáles son los barrios más rejuvenecidos o enriquecidos, vemos conveniente profundizar más en la clusterización y buscar una partición más fina que concrete más las características expresadas por los factores. Para ello, realizaremos ahora un análisis de conglomerados con un método jerárquico. Como ya hemos avanzado, los métodos jerárquicos no precisan fijar a priori un número de clusters específico y para llevarlos a cabo existen dos estrategias principalmente. La primera es considerar todos los elementos como un único grupo e ir haciendo subgrupos en función de las diferencias entre ellos. La segunda es considerar cada elemento como un grupo e ir juntándolos en grupos más grandes en función de las semejanzas entre ellos. De este modo, dependiendo de la estrategia en la que se base el algoritmo de clasificación, los métodos jerárquicos se suelen separar a su vez en divisivos y aglomerativos respectivamente. En todos ellos se utiliza la distancia entre los elementos como instrumento de medida, pudiendo considerarse la distancia euclídea, la distancia máxima, la distancia de Manhattan, etcétera. Además, el uso de los métodos jerárquicos, en cualquiera de ambos casos, permite obtener un gráfico llamado dendograma que ilustra cómo se van haciendo las subdivisiones o los agrupamientos paso a paso. Por este motivo, el reto al que nos enfrentamos ahora es determinar de forma razonada el método jerárquico con el cual realizar la clusterización. Para ello, nos guiaremos por el coeficiente de correlación cofenético, basado en las distancias entre los elementos cuando estos se unen en un mismo cluster. Aunque su uso como criterio de clusterización se ha puesto en duda [22], este parámetro ha sido ampliamente estudiado en la literatura estadística [14], [44], [46] y aún se considera adecuado para discriminar los métodos jerárquicos. El coeficiente cofenético tiene un valor comprendido entre 0 y 1 y es más elevado cuanto mayor es la precisión de la clasificación propuesta por el proceso. Así, por medio de la función hclust() realizaremos 40 clasificaciones de los barrios combinando 8 método jerárquicos con 5 tipos de distancias. Una vez realizadas todas las clasificaciones, computaremos el valor del coeficiente de correlación cofenético para cada una de ellas y nos quedaremos con aquel método y distancia que devuelvan un valor más alto del coeficiente cofenético. Mostramos en la siguiente tabla los valores de los coeficientes, cuyas filas hacen referencia a los métodos y las columnas a las distancias utilizadas.

	euclidean	maximum	manhattan	canberra	minkowski
ward.D	0.5168	0.5118	0.4930	0.6530	0.5168
ward.D2	0.5671	0.6679	0.5491	0.6882	0.5671
single	0.6650	0.6784	0.6373	0.5562	0.6650
complete	0.5791	0.6965	0.6823	0.6465	0.5791
average	0.7516	0.7547	0.7135	0.7482	0.7516
mcquitty	0.6111	0.6508	0.6786	0.7328	0.6111
median	0.4806	0.5314	0.5392	0.4082	0.4806
centroid	0.7390	0.7493	0.7264	0.5027	0.7390

Tabla 12: Coeficientes de correlación cofenética según metódo jerárquico y distancia.

Como vemos en la tabla 12, entre los métodos ejecutados se encuentran algunos de los más conocidos como el del vecino más próximo o Simple-Linkage, el del vecino más lejano o Complete-Linkage, el de la distancia media o Average-Linkage y el método de Ward entre otros. También se han considerado varias de las distancias más usadas como la euclídea (la norma vectorial 2 desde un punto de vista matemático) y la de Manhattan (la norma 1). A la vista de los resultados, el método que presenta un coeficiente de correlación cofenético más alto es el de la distancia media o Average-Linkage usando como criterio de comparación la distancia máxima (la norma infinito), al ofrecer un coeficiente de 0,7547.

El método Average-Linkage que acabamos de escoger es un método de clasificación jerárquica de tipo aglomerativo. Es decir, el algoritmo empieza considerando que hay 85 grupos, cada uno de ellos correspondiente a un barrio. Además se calcula también la matriz de distancias de tamaño 85×85, distancias calculadas utilizando la norma infinito sobre las puntuaciones factoriales entre los grupos. Al tratarse del método de Average-Linkage, que se traduciría como *enlace promedio*, las distancias entre dos clusters se obtienen como la media de las distancias entre los elementos de ambos clusters. Buscaríamos entonces en la matriz la distancia más pequeña y juntariamos los dos grupos que la producen en un único cluster. A continuación volveríamos a construir la matriz de distancias del nuevo conjunto de grupos, esta vez de tamaño 84×84, y buscaríamos otra vez la distancia más baja en la matriz para volver a juntar dos grupos en uno, procediendo así sucesivamente. El algoritmo finaliza cuando todo los barrios se han agrupado en un único cluster. Este algoritmo está implementado en R y como ya sabemos se puede realizar de forma rápida por medio de la función hclust(). Ahora bien, para decidir el número de clusters en los que dividir el conjunto de barrios, podríamos fijarnos en el dendograma asociado al proceso para discernir el corte en el número de grupos que consideremos apropiado. No obstante, la función NbClust() también nos aconseja sobre cuál es la cantidad óptima de clusters para métodos jerárquicos. De este modo, procediendo como antes con el algoritmo de las K-Medias, deberemos introducir un rango de clusters que nos parezca razonable. Recordemos que la motivación de repetir la clusterización era conseguir una división más precisa, es decir, clasificando los barrios en un mayor número de grupos. Por tanto, indicando en los argumentos de la función que el número mínimo de grupos deseado es 9 y el máximo 15, a parte del método y la distancia escogida, el diagrama resultante nos indica que la mejor opción para clasificar los barrios por el método de la distancia media es utilizar **11** clusters.

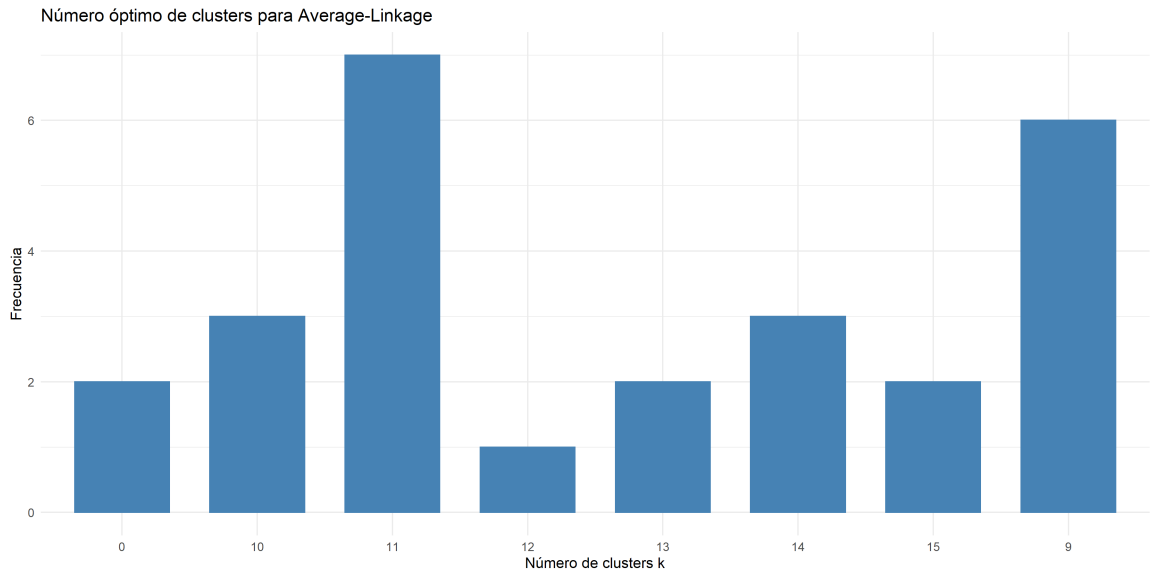

Figura 15: Número óptimo de clusters para Average-Linkage según el paquete de R *NbClust*.

Decidido el número de clusters en el que dividir los barrios, podríamos reproducir inmediatamente el mapa de la ilustración 14 actualizado a la nueva clasificación. Los grupos corresponderían a los resultantes del proceso iterativo si dicho algoritmo finalizara en la décima iteración. No obstante, previamente a dibujar el mapa con los 11 clusters, consideramos necesario mostrar una herramienta tan representativa e importante en los métodos de agrupación jerárquicos como son los **dendogramas**. Este gráfico nos permite ver de forma resumida cómo se han ido agrupando los 85 barrios y nos da información de aquellos que se destacan con una mayor singularidad. El dendograma se muestra en su vertiente circular y coloreado en función de los clusters. Si nos fijamos, aquellos barrios del grupo 4 de la anterior clasifiación por el método de las K-Medias son los que más rápido se distancian del resto.

Figura 16: Dendograma para el método de Average-Linkage considerando la distancia máxima.

Ahora, manteniendo los colores de la paleta utilizada en el dendograma, dibujamos el mapa de la ciudad de València con los 85 barrios clasificados en los 11 grupos según el método de Average-Linkage. Seguidamente, al igual que antes, procederemos a examinar y comentar las peculiaridades de las distintas zonas que hemos detectado como nichos de mercado dentro del municipio.

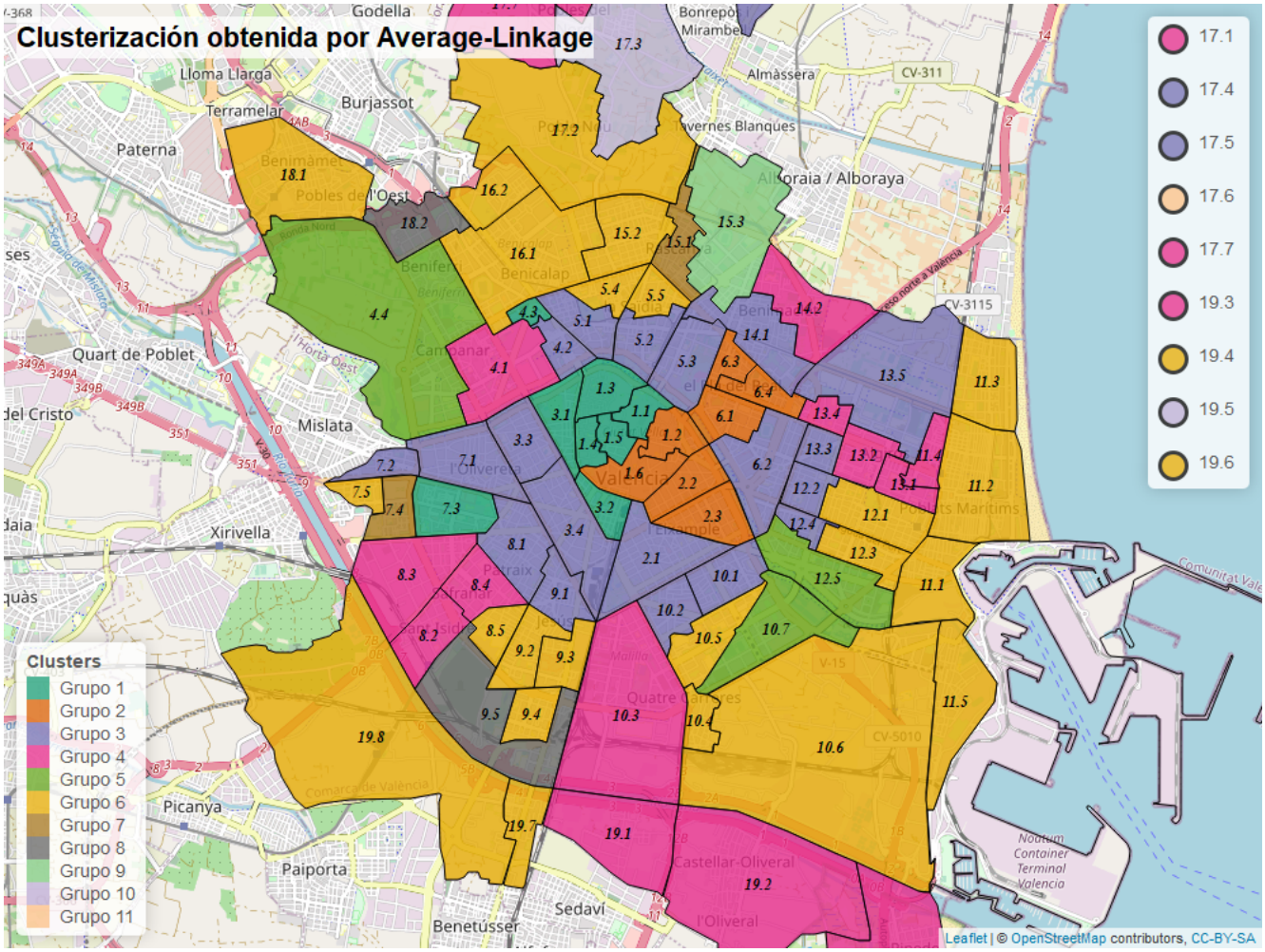

Figura 17: Mapa de València con los barrios agrupados según el método de Average-Linkage.

- **Grupo 1**: Formado por 8 barrios, la mayoría de ellos ubicados en pleno centro de la ciudad. Contiene barrios envejecidos, incluyendo el barrio 4.3. El Calvari, el más envejecido de toda València. Presenta una altísima migración de manera generalizada. Son zonas enriquecidas, a excepción de los barrios 4.3. el Calvari y 7.3. Tres Forques donde la puntuación del factor TC3 (el cual recordemos posee un peso migratorio negativo en sus loadings) está más orientado a suavizar el alto valor del factor TC2.
- **Grupo 2**: Compuesto por 7 barrios, exclusivamente de los distritos centrales de Ciutat Vella, l'Eixample y el Pla del Real. Es un territorio envejecido, que ostenta poca migración y que destaca por alzarse como la zona más adinerada de la ciudad.
- **Grupo 3**: Integrado por 21 barrios del municipio, contando como 2 la fusión 17.4. Cases de Bàrcena y 17.5. Mauella, se presenta como el segundo nicho de mercado más grande de la clasificación. Todos los barrios tienen un perfil demográfico envejecido y es propio de estas zonas que haya altos movimientos migratorios, salvo barrios más autóctonos como 3.3. la Petxina, 3.4. Arrancapins y 13.5. la Carrasca.

En cuanto al nivel de riqueza de este grupo, no podemos hablar ni de barrios muy acaudalados ni poco acaudalados, sino que se encuentran en la media.

- **Grupo 4**: Compuesto por 15 barrios localizados en la periferia. No se trata de una zona envejecida ni rejuvenecida, salvo un par de barrios bastante rejuvenecidos como 8.2. Sant Idsidre y 8.4. Safranar y un barrio envejecido como 13.2. Ciutat Jardí. Este grupo se caracteriza porque todos presentan una migración bastante baja y porque, a excepción del barrio 13.4. la Bega Baixa, se vislumbran como lugares muy poco adinerados.

- **Grupo 5**: Este cluster está conformado por tan solo 3 barrios: 4.4. Sant Pau, 10.7. Ciutat de les Arts i de les Ciències y 12.5. Penya-roja. Son barrios muy rejuvenecidos donde los movimientos migratorios y la incidencia de la población extranjera no son característicos ni particularmente relevantes. Además están muy enriquecidos. Lo más notorio de este grupo es que combinan la juventud con la riqueza de forma llamativa, ya que junto con el barrio 17.6. Massarrojos estos tres barrios son los más acaudalados de entre los demográficamente más jóvenes.

- **Grupo 6**: Este grupo es el más grande de todos al albergar en total 25 barrios situados en la periferia, contando como 2 la fusión de los barrios 19.7. la Torre y 19.8. Faitanar. Estos barrios no destacan ni por estar envejecidos ni rejuvenecidos, aunque en este cluster se agrupan barrios de corte más joven, en particular los barrios 8.5. Favara, 11.1. el Grau, 16.1. Benicalap, 18.1. Benimàmet, 19.4. el Saler, 19.6. el Perellonet, 19.7. la Torre y 19.8. Faitanar. Se aprecia un perfil migratorio más bien positivo pero no muy marcado, salvo en los barrios 7.5. la Llum, 9.3. la Creu Coberta, 11.3. la Malva-rosa y 19.6. el Perellonet en los que se da una ponderación negativa en el factor asociado a la migración. A excepción del barrio 19.4. el Saler, todos los demás están bastante empobrecidos en comparación con el resto de la ciudad, siendo el factor ligado a la riqueza de forma negativa el que más incide en este grupo, puesto que los otros de vejez y migración no son tan significativos.

- **Grupo 7**: En este grupo se incluyen únicamente 2 barrios: 7.4. la Fontsanta y 15.1 Orriols. En lo que respecta a la componente etaria, no se trata de barrios ni envejecidos ni rejuvenecidos sino que ambos se sitúan en la media. Es reseñable en ellos la altísima migración y población extranjera que manifiestan, siendo el barrio 7.4. la Fontsanta el que tiene un factor migratorio más fuerte de toda la ciudad y el barrio 15.1. Orriols el tercero, separados solo por el barrio 7.3. Tres Forques. El factor de riqueza es ligeramente positivo, ya que incluye una componente negativa de población extranjera que contrarresta la elevada puntuación migratoria, pero en todo caso se trata de barrios poco adinerados.

- **Grupo 8**: Al igual que en el cluster anterior, está formado solo por 2 barrios: 9.5. Camí Real y 18.2. Beniferri. Lo más destacado de ellos es que tienen una población muy rejuvenecida, siendo el barrio 18.2. Beniferri el cuarto y 9.5. Camí Real el quinto más joven de todo el municipio. Presentan una componente migratoria prácticamente nula y son dos barrios poco adinerados.

- **Grupo 9**: Compuesto por un único barrio: 15.3. Sant Llorenç. Es sin duda el barrio más rejuvenecido de toda la ciudad, motivo por el cual se ha posicionado como un nicho de mercado aparte. La migración registrada es bastante baja y aunque no es una zona de poca riqueza tampoco es un barrio muy adinerado.

- **Grupo 10**: Integrado por 2 barrios de las afueras de la ciudad: 17.3. Carpesa y 19.5 el Palmar. La población en estas pedanías está algo envejecida y ambos barrios cuentan con una bajísima migración, siendo el 19.5 el más autóctono de toda la ciudad. Además se trata de zonas poco adineradas. La mayor singularidad de este grupo es que alberga los dos barrios con puntuaciones en el factor de riqueza más bajas, si bien es cierto que en este caso se debe en parte a su finalidad como ajuste del factor migratorio.

- **Grupo 11**: Este último grupo consta de un solo barrio: 17.6. Massarrojos. Este barrio está bastante rejuvenecido, presenta poca migración y se posiciona como uno de los territorios más enriquecidos de la ciudad.

Como hemos visto, los grupos son bastante heterogéneos entre sí y cada uno de ellos cuenta con un perfil de población singular. La clasificación aquí expuesta sirve como referencia para identificar el carácter subyacente en los distintos barrios a partir de los tres factores calculados. Ahora bien, los factores extraídos y sobre los que se sustenta el análisis cluster también pueden ser interpretados desde un punto de vista más financiero, entendiendo las repercusiones económicas y posibilidades de inversión que se desprenden de ellos, lo que nos permite visionar los barrios como nichos de mercado. Así, el factor de Riqueza puede ser visto como un indicador de solvencia, el factor de Migración como un indicador de estabilidad y el factor de Vejez como un indicador de proyección de cara al futuro, puesto que aquellos territorios más rejuvenecidos ofrecen la posibilidad de explotar durante más tiempo y a largo plazo los diferentes productos financieros. Por tanto, podemos resumir la información proporcionada por el análisis en la siguiente tabla, donde pintaremos en verde, amarillo o rojo las casillas asociadas a cada nicho de mercado en función de si en los barrios tales indicadores se manifiestan de forma significativa, moderada o baja respectivamente.

Cluster	**Proyección**	**Estabilidad**	**Solvencia**
Grupo 1			
Grupo 2			
Grupo 3			
Grupo 4			
Grupo 5			
Grupo 6			
Grupo 7			
Grupo 8			
Grupo 9			
Grupo 10			
Grupo 11			

Tabla 13: Interpretación financiera de la clusterización ilustrada en la figura 17.

A la vista de la tabla 13, percibimos rápidamente que los barrios con mayores atractivos de inversión son los que integran los clusters 5, 9 y 11. En total, estos grupos representan 5 barrios de la ciudad, los cuales se vislumbran como zonas en las que predomina una población joven y adinerada potencialmente propicia para determinados productos financieros a los que puede ser receptiva. Los barrios en cuestión son 4.4. Sant Pau, 10.7. Ciutat de les Arts i de les Ciències, 12.5. Penya-roja, 15.3. Sant Llorenç y 17.6. Massarrojos. Por otro lado, aquellos barrios que presentan unas características menos deseables económicamente hablando son los de los cluster 6 y 7 en mayor medida, en especial los barrios 7.4. la Fontsanta y 15.1. Orriols. De este modo, una adecuada interpretación de los resultados permitiría la detección de las mejores zonas de la ciudad a las que orientar de forma beneficiosa las campañas de captación y la oferta, pudiendo considerar esta clasifiación como una herramienta que mide cuantitativa y objetivamente el potencial asegurable de los diferentes barrios.

Antes de finalizar este trabajo, vamos a tratar un tema que puede aportar algo más de valor al estudio estadístico aquí realizado. Si en el mapa 17 nos fijamos en la localización de los clusters dentro de la ciudad, podemos apreciar que en diferentes zonas barrios del mismo cluster se encuentran colindantes unos con otros, lo que nos lleva a preguntarnos si la distribución espacial de los datos influye de manera relevante en el valor de las variables y de los factores. La herramienta que se encarga de abordar las relaciones territoriales y las posibles dependencias geográficas de los datos es el **análisis espacial**. Para ello, decidimos incorporar de manera breve un último apartado que complementa los métodos estadísticos ya estudiados donde comprobaremos la existencia de la autocorrelación espacial percibida, tanto a nivel global como a nivel local.

9. Autocorrelación espacial e indicadores locales de asociación espacial (*LISA*)

La autocorrelación espacial es un instrumento que permite medir "el grado de asociación que una variable desarrolla a través de un espacio definido como marco geográfico" y cuyo objetivo es "comprender cómo se distribuye el fenómeno en el espacio analizado y en qué grado los elementos locales pueden verse afectados por sus vecinos" [45, p. 2]. Es decir, es un procedimiento mediante el cuál podemos averiguar si las relaciones espaciales entre dos territorios determinan de manera no aleatoria características similares entre ambos. En línea con esta idea, existen los indicadores locales de asociación espacial (*LISA* del inglés Local Indicators of Spatial Association) que son los estadísticos que cuantifican los patrones espaciales y evalúan la autocorrelación espacial en un conjunto de datos. En nuestro caso, el conjunto de datos a testar es la distribución de los valores de los tres factores entre los diferentes barrios. Para ello, nos basaremos en los estadísticos definidos en [1], [10] y [33] siguiendo los pasos para su interpretación propuestos en el apartado 7 de [32] y particularmente en el artículo [2] escrito por uno de los mayores expertos en el tema, Luc Anselin. Estos estadísticos tienen en cuenta los valores en todas las localizaciones y se basan en las desviaciones respecto a la media de las variables en cada unidad espacial así como en una matriz W de pesos espaciales que informa sobre cuáles son aquellas observaciones vecinas y que tiene como fin definir la estructura espacial de los barrios. Podemos hablar de dos tipos de estadísticos: unos con los que estudiar la existencia de la autocorrelación espacial a nivel global (i.e., si existe o no) y otros que nos permiten estudiar la localización y el signo de la autocorrelación espacial localmente (i.e., dónde existe). Estos estadísticos podrán ser calculados fácilmente gracias a las funciones implementadas en el paquete de R `spdep`, paquete creado por otro de los grandes conocedores de la materia, Roger Bivand, sobre el cual nos apoyaremos para realizar el siguiente análisis espacial. En concreto, los estadísticos que utilizaremos serán:

- ◇ **Índice *I* Global de Moran**: Se trata de un estadístico que estudia la existencia de la autocorrelación espacial. Cuando la matriz de pesos espaciales está estandarizada por filas, el valor del estadístico oscila entre -1 y 1. Si los valores en el conjunto de datos tienden a agruparse espacialmente, ya sea porque valores altos de una variable están cerca de otros valores altos o bien porque valores bajos de la variable se hayan próximos de otros valores bajos, el índice será positivo. Por otro lado, si valores altos de una variable suelen estar cerca de valores bajos, el índice será negativo. Finalmente, el índice se aproximará a 0 cuando variables altos y bajos se encuentran de forma aletaoria y equilibrada entre las unidades espaciales. La función de R que computa este índice es moran().

- ◇ ***C* de Geary**: Al igual que la I Global de Moran, se trata de un indicador que también estudia a nivel global la existencia de la autocorrelación espacial, pero que en este caso oscila entre 0 y 2. Este índice valdrá 1 cuando no se aprecie autocorrelación espacial, lo que nos indicará que los valores se distribuyen a través del espacio sin dependencias entre sí, aleatoriamente. Valores del índice C próximos a 0 nos avisan de que existe autocorrelación positiva en los datos, mientras que valores cercanos a 2 denotan un alto nivel de autocorrelación de signo negativo. En R se calculará mediante la función geary().

- ◇ **Índice *I* Local de Moran**: Este índice estudia a nivel local y para cada territorio la autocorrelación espacial con sus regiones limítrofes, por lo que se obtendrán tantos índices como unidades espaciales consideradas. Al igual que en el Índice I Global de Moran, con matrices de pesos estandarizadas su valor oscila entre 1 y -1[III], siendo 0 lo equivalente a un patrón espacial completamente aleatorio. Índices positivos indicarán vecindades con valores semejantes e índices negativos nos informarán sobre vecindades con valores diferentes. La función que computa el Índice I Local de Moran es localmoran(). Es importante señalar que este indicador es en esencia una medida relativa y que debe

[III]Debido a su definición matemática, el rango del valor del Índice I Local de Moran está limitado por la matriz W de ponderaciones, por lo que matrices de pesos con filas y columnas de ceros pueden distorsionar sus valores. En nuestro caso, algunos barrios del distrito 17. Pobles del Nord no cuentan con barrios fronterizos, lo que puede hacer que el estadísitico sobrepase el límite estándar de 1.

ser interpretado en función de los p-valores que testan la significación estadística de los resultados. Es decir, la función localmoran() devuelve también un p-valor para cada observación asociado a la hipótesis nula de aletoriedad espacial de los valores. De este modo, solo cuando el p-valor sea estadísticamente significativo diremos que en ese territorio existe autocorrelación espacial con los territorios colindantes.

- ◇ **El estadístico G_i de Getis-Ord**: Se trata de un indicador de autocorrelación espacial local y sirve como complemento del Índice I local de Moran. Si bien el índice I encuentra solo algunos barrios donde la autocorrelación espacial es significativa, el estadístico G_i permite tener una visión global del grado de asociación y es muy útil para la localización de *hot-spots* en el análisis de puntos calientes [1, p. 5]. Un valor positivo de G_i indica que las unidades espaciales colindantes tienen valores altos similares, mientras que $G_i < 0$ para un territorio lo deberemos traducir a que tiene zonas alrededor con valores bajos de la variable. La función disponible en R para computar este estadístico es localG().

De este modo, procedemos sin más dilación a calcular en R estos estadísticos basándonos en las funciones y documentación adjunta de la librería `spdep`. Usaremos el objeto *barrios* de clase SpatialPolygonsDataFrame mencionado en el apartado de Bases de Datos para el cálculo de la matriz de pesos espaciales, el cual recordemos contiene los polígonos georreferenciados correspondientes a los barrios de la ciudad con los que definir la estructura espacial de los datos. En primer lugar, computaremos los dos asociados a la existencia de la autocorrelación espacial a nivel global. Como podemos ver en la tabla 14, sí que se aprecia autocorrelación espacial positiva en los tres factores, sobretodo en la componente TC3 de Riqueza, donde la I de Moran está más próxima a 1 que a 0 y la C de Geary más próxima a 0 que a 1, mientras que en las otros dos variables la incidencia no es tan notable.

	TC1: Vejez	TC2: Migración	TC3: Riqueza
I Global de Moran	0.351035	0.268106	0.513653
C de Geary	0.644034	0.711388	0.470911

Tabla 14: Estadísticos de autocorrelación espacial global obtenidos para los factores.

Para considerar de forma estricta que los valores obtenidos son significativos y proceder así a estudiar la autocorrelación espacial a nivel local, podemos usar dos funciones integradas en R: moran.mc() y geary.mc(). Estas funciones se basan en el Método Monte Carlo[IV] permutando de forma aleatoria la estructura espacial y tienen entre sus outputs unos p-valores que sirven para contrastar la hipótesis de nula de que los datos se distribuyen espacialmente de forma aleatoria. En los dos casos y para los tres factores, los p-valores son inferiores a 0,001. Es decir, la probabilidad de observar los valores de la I de Moran y la C de Geary de la tabla 14 si hubiera aleatoriedad espacial completa está por debajo de 0,001, por lo que podemos asumir sin lugar a dudas que en nuestro conjunto de datos tiene lugar el fenómeno de la autocorrelación espacial. Así, tras este análisis global, nos centraremos en encontrar los barrios en los que tiene lugar esta dependencia geográfica de forma significativa mediante el Índice I Local de Moran y el estadístico G_i de Getis-Ord. En cuanto al Índice I Local de Moran, la autocorrelación espacial a nivel local medida se puede manifestar (en caso de existir de forma significativa) de 4 formas diferentes, lo que desembocará en la construcción de cuatro posibles clusters distintos de asociación geográfica entre los barrios de la ciudad. Estos clusters solo se mostrarán cuando podamos asumir que la autocorrelación espacial observada en los valores no es fruto del azar, para lo cual deberemos fijar un umbral para los p-valores que devuelve la función localmoran() y determinar así la significación estadística de los resultados. De esta forma, los clusters se definirán en función de las relaciones entre las unidades espaciales y sus regiones colindantes, pudiendo ser del tipo Alto-Alto, Alto-Bajo, Bajo-Alto y Bajo-Bajo. El primer adjetivo corresponderá al valor de la variable en el barrio y el segundo adjetivo al valor medio de la variable en sus barrios fronterizos. Este valor medio de la variable

[IV] En [7, p. 271] y [20, p. 225] podemos entender de manera más precisa el funcionamiento paso a paso del Método Monte Carlo y la interpretación de los pseudo p-valores asociados. Además, en [7] incluso se nos adjunta el código para aprender a programarlo con R.

registrado en los barrios limítrofes es lo que se conoce como *spatial lag*. Así, si dibujáramos un diagrama de dispersión donde el eje X esté asociado al valor de los barrios y el eje Y al valor de los spatial lags de cada barrio y dividiéramos dicho gráfico en 4 zonas separadas por las rectas $X = 0$ e $Y = 0$, obtendríamos las siguientes regiones vinculadas a los clusters:

- **Alto-Alto**: Zonas con un índice I local de Moran estadísticamente significativo de la región comprendida en el cuadrante $X > 0$, $Y > 0$. Es decir, barrios con un valor positivo de la variable de estudio y con un valor positivo del spatial lag asociado, en los que se asuma una asociación espacial no aleatoria.

- **Alto-Bajo**: Zonas con un índice I local de Moran estadísticamente significativo de la región comprendida en el cuadrante $X > 0$, $Y <= 0$. Es decir, barrios con un valor positivo de la variable de estudio y con un valor no positivo del spatial lag asociado, en los que se asuma una asociación espacial no aleatoria.

- **Bajo-Alto**: Zonas con un índice I local de Moran estadísticamente significativo de la región comprendida en el cuadrante $X <= 0$, $Y > 0$. Es decir, barrios con un valor no positivo de la variable de estudio y con un valor positivo del spatial lag asociado, en los que se asuma una asociación espacial no aleatoria.

- **Bajo-Bajo**: Zonas con un índice I local de Moran estadísticamente significativo de la región comprendida en el cuadrante $X <= 0$, $Y <= 0$. Es decir, barrios con un valor no positivo de la variable de estudio y con un valor no positivo del spatial lag asociado, en los que se asuma una asociación espacial no aleatoria.

Una vez explicado el procedimiento para la localización de barrios espacialmente autocorrealcionados, ahora el objetivo es determinar el umbral α de significación sobre los p-valores devueltos por localmoran() para afrontar los errores de tipo I. Es decir, necesitamos fijar un valor de corte que nos permita rechazar la hipótesis nula de aleatoriedad espacial de los datos y juzgar así la autocorrelación espacial en los barrios adecuadamente. Uno de los niveles de significación más empleados en los constrastes de hipótesis es el de $\boldsymbol{\alpha = 0,01}$. No obstante, en la literatura vinculada a la estadística espacial se suelen usar niveles de significación más pequeños, sobretodo en ámbitos en los que se dispone de muchas más observaciones y en contextos de comparación múltiple donde aumenta la probabilidad de cometer un error de tipo I y rechazar de forma errónea la hipótesis nula. Si bien es cierto que en nuestro conjunto de datos el número de observaciones es ostensiblemente bajo (85 barrios), vemos instructivo mostrar los resultados con diferentes umbrales y reflejar de este modo el aumento del número de clusters identificados como significativamente estadísticos cuanto mayor es el nivel de significación. Por lo tanto, consideraremos además del valor $\alpha = 0,01$ dos umbrales siguiendo el ejemplo de Anselin en [2]. Esto niveles de significación vendrán determinados por la Correción de Bonferroni o *Bonferroni Bound* y por la Tasa de Descubrimientos Falsos o *False Discovery Rate*, popularmente llamada *FDR*. La correción de Bonferroni consiste en usar como umbral el cociente del p-valor base usado (en nuestro caso 0,01) entre el número de observaciones. Es decir, en nuestro caso se propone un nivel de significación de $\alpha = \frac{0,01}{85} = 0,0001176471$, por lo que se trata de un ajuste muy restrictivo. Con la finalidad de suavizar el nivel tan bajo que computa Bonferroni, en la literatura tiene más aplicación el False Discovery Rate. La idea es crear una vector de 85 posiciones llamado FDR cuya posición i contenga el valor $i \cdot \frac{0,01}{85}$. A continuación, se ordenan los p-valores devueltos por localmoran() de menor a mayor. Se comparan entonces el primer valor de la variable FDR, que para $i = 1$ coincide con el Bonferroni Bound, con el p-valor más bajo de los ordenados. Si el p-valor es inferior, se pasa a escoger el segundo valor de la FDR, que en este caso es $2 \cdot \frac{0,01}{85}$, y ahora se compara con el segundo p-valor más bajo. Así tiene lugar la comparación posición por posición mientras el p-valor sea inferior al valor FDR. Tomaremos aquí como umbral el valor más alto de los contenidos en el FDR de entre aquellos cuyos p-valores asociados sean inferiores. Para entenderlo mejor, veamos en la siguiente tabla la comparación entre las primeras 6 posiciones de los p-valores ordenados de menor a mayor devueltos por la función localmoran() evaluada sobre el factor de Vejez con los valores FDR.

	P-valores	FDR
1	2.827054e-11	0.0001176471
2	0.0000176396	0.0002352941
3	0.0000398360	0.0003529412
4	0.0004468589	0.0004705882
5	0.0008194053	0.0005882353
6	0.0012931555	0.0007058824

Tabla 15: Método FDR para la determinación del nivel de significación para el factor TC1.

A la vista de la tabla 15, observamos que para los 4 primeros casos el valor FDR es superior al p-valor correspondiente, pero que en la quinta posición el p-valor está por encima del valor FDR. Por lo tanto, tal cual hemos explicado, tomaremos como nivel de significación $\alpha = 0.0004705882$. Este mismo umbral de $\alpha = 0.0004705882$ será el que se usará para el factor TC2, mientras que el umbral para el factor TC3 por esta técnica será $\alpha = 0.001647059$. De este modo, ya podemos dibujar con las funciones del paquete de R `tmap` los mapas LISA con los clusters resultantes del análisis para cada uno de los factores. Además los dibujaremos considerando para la significación estadística los distintos umbrales mencionados, obteniendo los mapas de las figuras 18, 19 y 20. En la figura 18 se muestran más zonas con autocorrelación espacial local, al haber escogido un un nivel de significación más grande, mientras que en la figura 19 desaparecen algunos barrios y en la figura 20 el número se reduce un poco más. En cualquier caso, podemos sacar varias conclusiones claras de las imágenes. Si nos fijamos en los mapas correspondientes al factor TC1, vemos que la autocorrelación espacial es predominante del tipo Bajo-Bajo, esto es, tienen más dependencia los barrios rejuvenecidos que los envejecidos, visualizándose en la ciudad zonas jóvenes correlacionadas entre los barrios 18.1. Benimàmet y 18.2. Beniferri y 4.4. Sant Pau, entre los barrios 8.2. Sant Isidre y 9.5. Camí Real y entre los barrios 10.7. Ciutat de les Arts i de les Ciències y 12.5. Penya-roja. Por otro lado, fijándonos en el factor TC2 vemos que claramente hay un patrón migratorio negativo en los barrios del norte de la ciudad 17.1. Benifaraig, 17.3. Carpesa y 17.7. Borbotó. Finalmente, hay una zona bastante amplia en los barrios centrales de València con una autocorrealción espacial de tipo Alto-Alto de gran relevancia en cuanto al factor TC3 de Riqueza, la cual además se muestra para todos los niveles de significación.

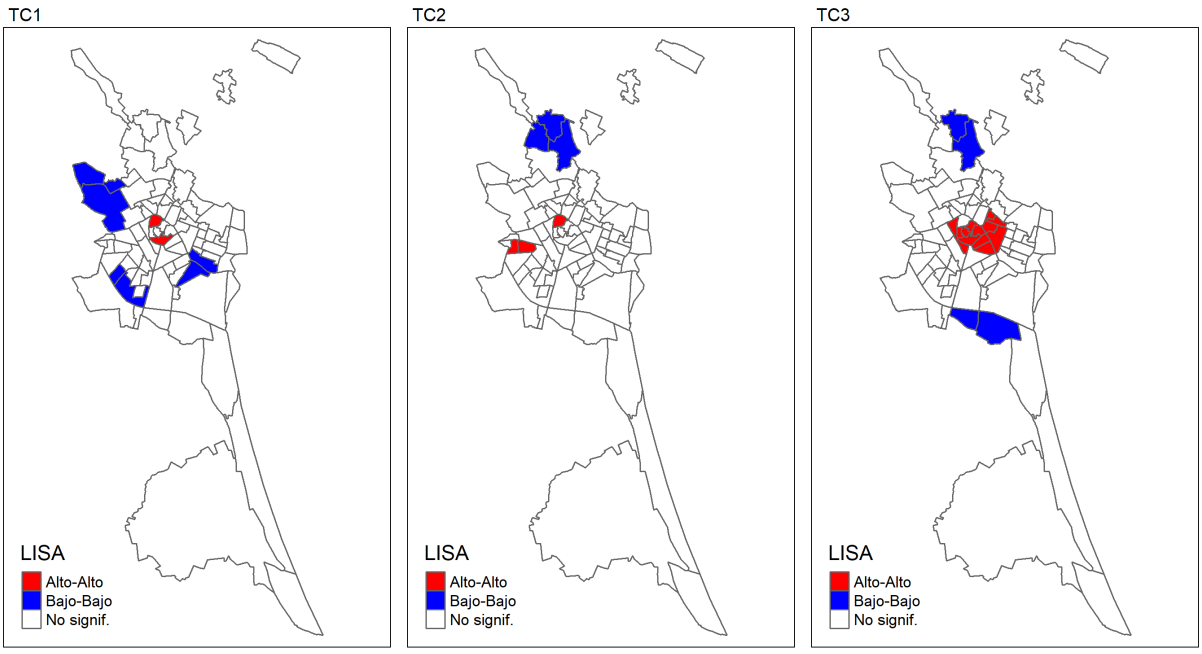

Figura 18: Clusters asociados al índice Local I de Moran con un nivel de significación básico $\alpha = 0,01$.

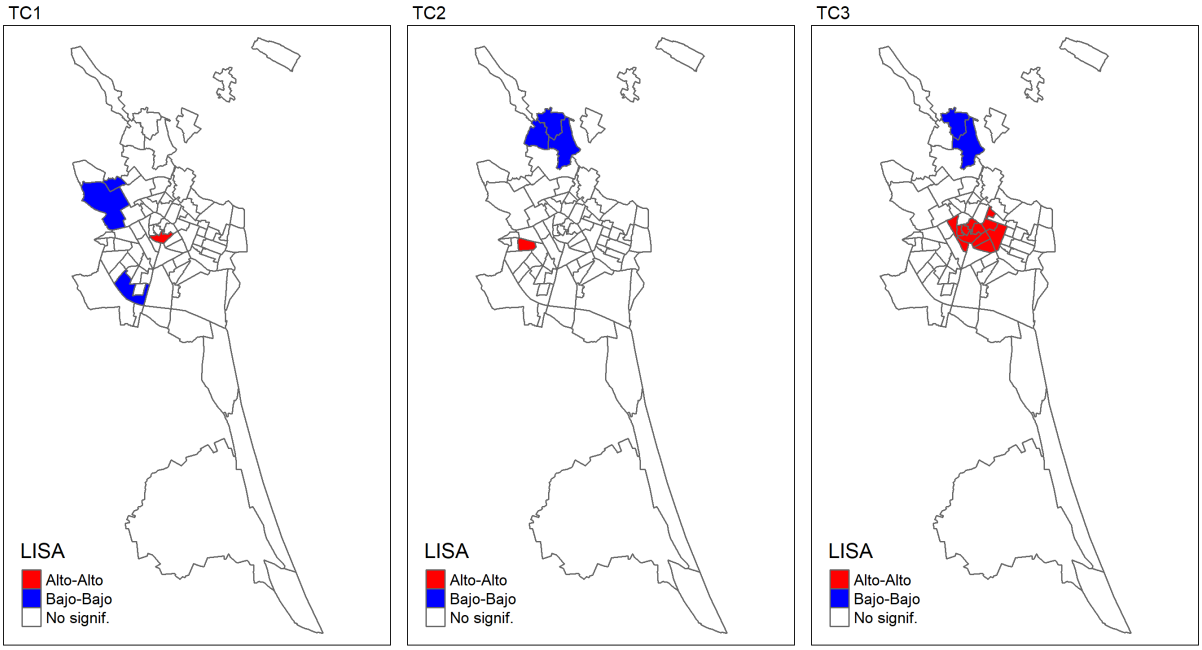

Figura 19: Clusters asociados al índice Local I de Moran un nivel de significación basado en las FDR.

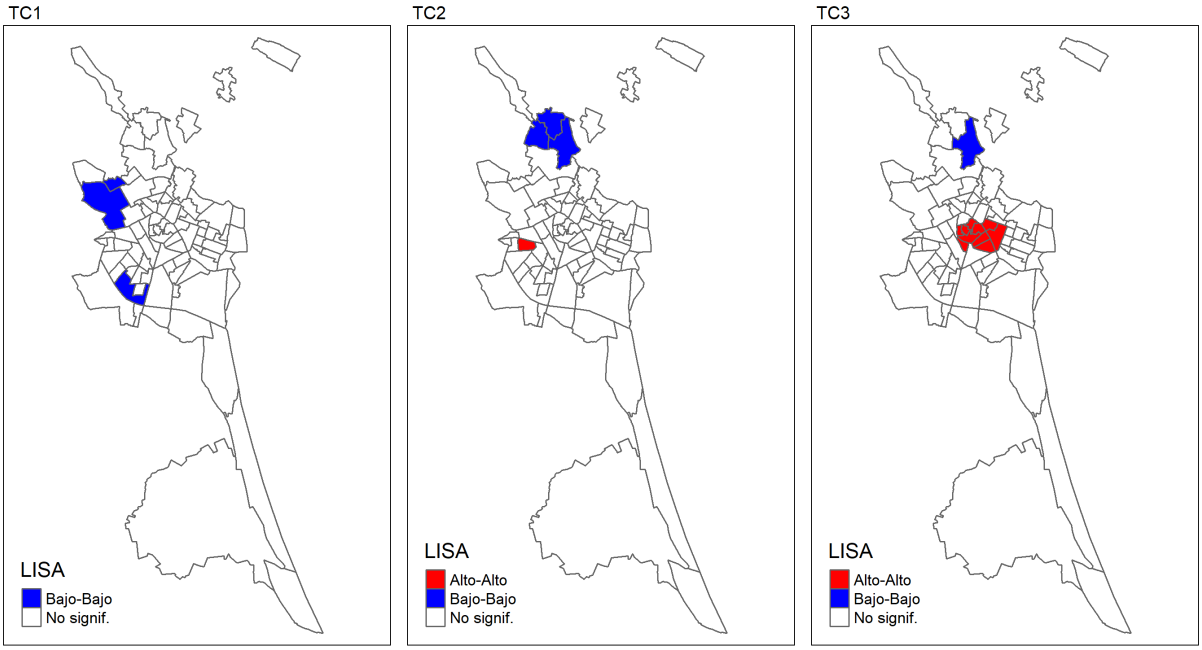

Figura 20: Clusters asociados al índice Local I de Moran con un nivel de significación basado en la corrección de Bonferroni.

Finalmente, para completar el análisis espacial, el estadístico G_i nos permite localizar de forma significativa barrios que tienen puntuaciones altas en los factores y que además tienen como vecinos a otros barrios con puntuaciones altas, lo que se conoce como puntos calientes o *hot-spots*. El mismo estadístico sirve para visualizar *cold-spots* de forma análoga a los puntos calientes pero con puntuaciones negativas. De esta forma, podemos obtener clusters en todo el municipio que representan agrupaciones locales tanto a niveles extremos como clusters con un grado de autocorrelación espacial media y baja, lo que le diferencia de manera sustancial del Índice I Local de Moran. Además, los clusters se colorean en función de los valores de los Z-Scores que devuelve como output la función localG() que computa este estadístico, puntuaciones Z basadas en la misma hipótesis nula de aleatoriedad espacial y que miden en términos de desviaciones típicas lo lejos que están los valores del barrio alejados de la media.

De este modo, si nos fijamos en los clusters de los mapas de la figura 21 y en sus escalas concluimos que en el factor TC3 de Riqueza es donde se localizan los puntos calientes de manera más intensa, concretamente en los barrios del centro de la ciudad 1.2. la Xerea, 1.6. Sant Francesc y 2.2. el Pla del Remei. Por contra, los cold-spots cobran más relevancia en el factor TC1 de Vejez, es decir, hay zonas con población rejuvenecida correlacionada espacialmente de forma importante en el sur de la ciudad y en especial en los barrios del noroeste del distrito 18. En cuanto a la componente TC2 de Migración, encontramos hot-spots en los barrios del distrito 7 ubicado al oeste de la ciudad, destacando 7.2. Soternes y 7.5. la Llum, además de una conexión significativa entre los barrios 1.3. el Carme y 5.1. Marxalenes con los de su alrededor. En este factor TC2 también se vislumbra una asociación espacial de autoctonía en las regiones del norte en modo de cold-spots, correspondientes a los barrios 17.1. Benifaraig, 17.3. Carpesa y 17.7. Borbotó.

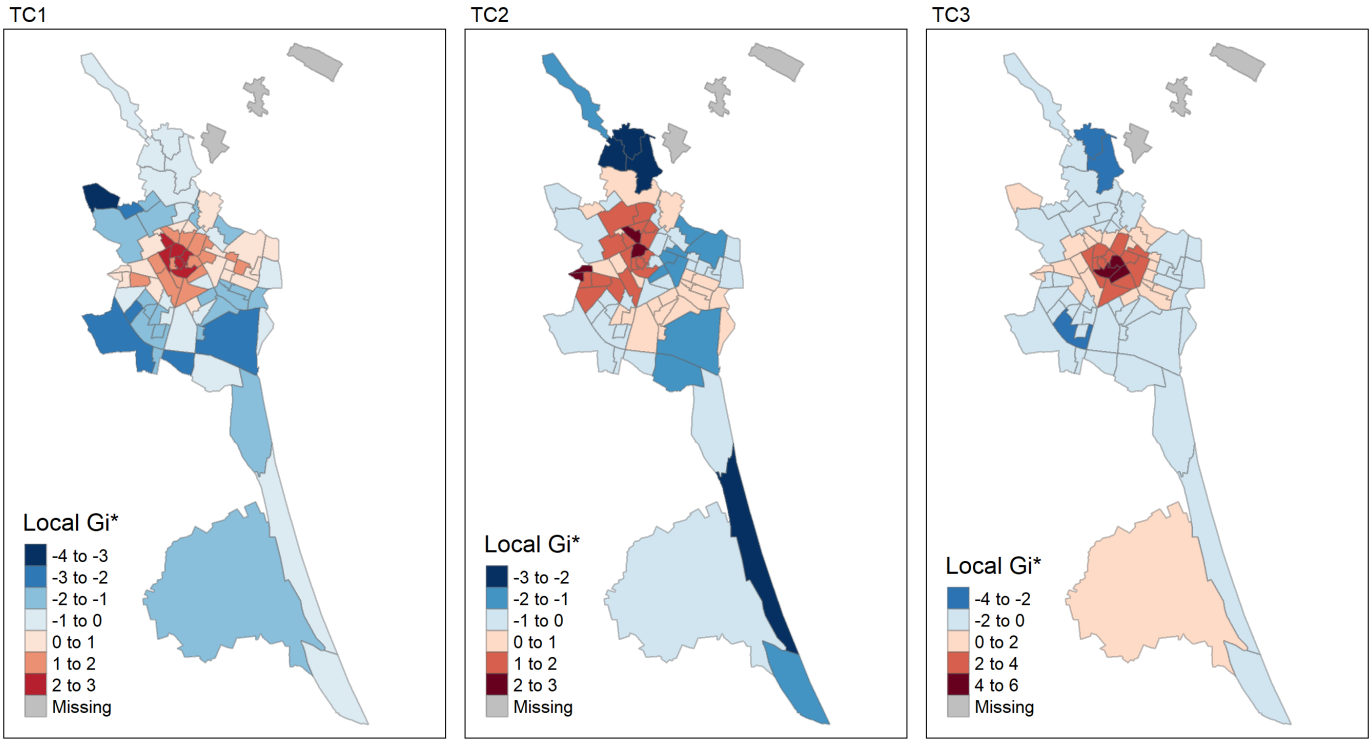

Figura 21: Clusters asociados al estadístico G_i de Getis-Ord.

Conclusiones

La redacción de este Trabajo de Fin de Máster ha sido posible gracias a la combinación de tres causas que se antojan como fundamentales para poder llevar a cabo cualquier estudio de características similares. En primer lugar, a la labor previa de investigación en artículos científicos, libros de referencia y páginas web especializadas en la materia como los que se incluyen en la bibliografía y con los que asentar los conocimientos adquiridos en las asignaturas del máster. La consulta de tales publicaciones ha hecho viable profundizar en los conceptos aprendidos estudiando nuevas técnicas y modelos estadísticos. Se trata además de una documentación en la que encontramos una base teórica sólida con la cual poder diseñar con garantías una metodología como la que aquí se ha desarrollado. En segundo lugar, es esencial disponer de datos provinientes de fuentes fiables que den valor al trabajo, datos de calidad y bien estructurados para utilizarlos de forma eficiente y evitar posibles dificultades durante su procesamiento. Para ello, debemos conocer y saber explotar los recursos que ofrecen los organismos públicos - en nuestro caso la Oficina de Estadística del Ayuntamiento de València y el Instituto Nacional de Estadística - u otras entidades que comparten datos con los que poder sacar información de utilidad si los aprovechamos de forma adecuada. Como ya hemos comentado al principio, la normalización, estandarización y depuración de las bases de datos empleadas ha sido tan importante para la consecución de los resultados mostrados en las páginas previas como la ejecución de los análisis posteriores. Finalmente, el tercer pilar sobre el que se ha sustentado este estudio han sido las herramientas informáticas que agilizan los procesos y que incorporan algoritmos y funciones que nos permiten analizar los datos a un nivel que de otra forma no sería posible. En nuestro caso en particular, del potencial estadístico que alberga el software R y del entorno R Studio con el que hemos gestionado de forma más organizada el espacio de trabajo.

De este modo, hemos podido construir conjuntos de datos de confianza que nos han servido como base para extraer información que puede ser útil para la sociedad. Hemos comprobado que la distribución en la ciudad de variables como la renta, la población extranjera o la población más envejecida es desigual en según qué zonas dentro del municipio. Fuimos capaces de condensar toda esta información en tres componentes con las que hemos conseguido clasificar los barrios en grupos con características similares en cuanto a estos factores, los cuales como ya hemos explicado se manifiestan como una gama de nichos de mercado heterogéneos entre ellos. Además, hemos analizado las relaciones espaciales subyacentes para averiguar qué barrios están más conectados y muestran mayor dependencia entre ellos. Durante todo el trabajo nos hemos apoyado en tablas, gráficos y mapas que facilitan la comprensión del desarrollo del mismo y que nos han ayudado de manera destacable a visualizar los resultados.

Dichos resultados ofrecen una gran información de la composición de la ciudad incidiendo en las diferencias y distribución de las características demográficas y económicas de los ciudadanos en el conjunto de los barrios. De entre todos ellos, conviene destacar los siguientes resultados:

- Los barrios periféricos del municipio se asemejan entre ellos y se diferencian claramente de los barrios del centro de València, principalmente en que los de la periferia son zonas con poco poder adquisitivo, mientras que los barrios centrales se posicionan como las zonas más ricas de la ciudad. Concretamente, aquellos que integran el núcleo urbano en los distritos 2. l'Eixample y 6. el Pla del Real son los más adinerados, percibiéndose además en ellos una autocorrelación espacial positiva importante ligada al factor de Riqueza, la cual se intensifica en los barrios 1.2. la Xerea, 1.6. Sant Francesc y 2.2. el Pla del Remei. Por contra, los barrios 7.4. la Fontsanta y 15.1 Orriols del exterior son los económicamente más desfavorables.

- Existen barrios ostensiblemente rejuvenecidos que hace que los algoritmos de clasificación enseguida los presenten como pequeños nichos de mercado específicos. De hecho, la autocorrelación espacial existente asociada al factor de Vejez es predominantemente de tipo Bajo-Bajo, habiendo así más dependencia geográfica entre los barrios rejuvenecidos que los envejecidos. Los barrios más jóvenes en cuestión son

4.4. Sant Pau, 9.5. Camí Real, 10.7. Ciutat de les Arts i de les Ciències, 12.5. Penya-roja, 15.3. Sant Llorenç, 17.6. Massarrojos y 18.2. Beniferri. Cabe mencionar que salvo los barrios 9.5. Camí Real y 18.2. Beniferri, el resto de zonas rejuvenecidas se presentan como territorios adinerados.

- Los movimientos migratorios y la población de nacionalidad extranjera ligados al factor de Migración no influyen de manera determinante ni en el factor de Vejez ni en el factor de Riqueza, combinándose estas características de manera indistinta entre ellas en los diferentes clusters alcanzados. Respecto al factor de Migración, este incide con fuerza en los barrios del distrito 1. Ciutat Vella, 3. Extramurs y 7. l'Olivereta. Por contra, los dos barrios más autóctonos de la ciudad corresponden a 17.3. Carpesa y a 19.5 el Palmar. Precisamente en los barrios 17.1. Benifaraig, 17.3. Carpesa y 17.7. Borbotó del distrito 17. Pobles del Nord se vislumbra claramente una autocorrelación espacial negativa vinculada a la autoctonía de dicha región.

Como ya hemos mencionado al principio del trabajo, una mirada prospectiva de la segmentación de los barrios que acabamos de desarrollar y de los resultados aquí obtenidos puede ser de provecho y utilidad para los distintos sectores de la sociedad, en concreto para los servicios aseguradores y financieros a los que va especialmente dirigido. Sin duda, la búsqueda en el mercado de clientes potenciales se agiliza focalizando los esfuerzos de la oferta en aquellas regiones con perfiles más propensos a la demanda de los productos ofertados. Así por ejemplo, aquellos clusters que incluyan los barrios más jóvenes y adinerados, como el 5 y el 11 coloreados en el mapa de la figura 17, son un buen nicho de mercado para la venta de productos de riesgo y con altas rentabilidades. Del mismo modo, residirán en dichos barrios consumidores con horizontes de inversión más cortos que manifiesten una mayor necesidad de liquidez para pagar deudas como puede ser una hipoteca, siendo más probable en tales zonas la captación de clientes para fondos de inversión con carteras en renta variable que les otorguen mayores beneficios. Por otra parte, aquellos clusters con barrios adinerados pero con poblaciones más envejecidas, como puden ser los de los grupos 1 y 2 coloreados en el mismo mapa, serán más proclives a aceptar productos de ahorro a largo plazo vinculados a la jubilación. Estos barrios serán por lo tanto un buen nicho de mercado en los que suscribir planes de pensiones u otros instrumentos financieros que permitan complementar la pensión pública.

Todo el material y código usado para la realización de este trabajo se encuentra depositado en mi cuenta de GitHub en el siguiente repositorio: `https://github.com/JuanferMG/TFM`. Este puede ser utilizado libremente para su réplica en futuras líneas de investigación. En concreto, en [39] y [23] podemos encontrar datos como los aquí empleados pero para otros ámbitos territoriales de la ciudad, como es el caso de los distritos municipales o las secciones censales, mediante las cuales podríamos llevar a cabo un análisis más estricto y profundo. Además, se pueden incorporar un mayor número de variables, en especial de [23], como puede ser la distribución según fuente de ingresos (sueldo, pensiones, prestaciones por desempleo, etcétera) o de la población con ingresos por unidad de consumo según distintos umbrales y sexo, pudiendo incluso realizar análisis similares para otros municipios del país. En el caso de querer realizar proyecciones, existen otras técnicas que pueden resultar más efectivas para determinados indicadores, como usar el modelo Lee-Carter para el caso de la Tasa de Mortalidad [6]. Otra opción sería trabajar con pirámides de población en vez de con indicadores demográficos, usando estas pirámides en R como objetos simbólicos sobre los cuales efectuar el análisis cluster, tal cual se detalla en [8]. Asimismo, debido a que hemos comprobado la existencia de autocorrelación espacial en los datos, optimizaríamos los resultados si tuviéramos en cuenta este hecho mediante modelos espacio temporales, como por ejemplo el uso de STARIMA [25], SD-STARIMA [48], INLA [5] - método que se encuentra integrado en R en el paquete R-INLA, existiendo en la literatura ejemplos muy instructivos de su uso como en [9, p. 46] - u otros modelos espacio temporales de índole bayesiana como los métodos Monte Carlo basados en Cadenas de Markov (MCMC) [4], todos ellos disponibles en R [47].

Anexo I. Definición de los indicadores demográficos calculados

- Ind1: El **Crecimiento Vegetativo** del barrio i en el año t (CV_t^i) se define como la diferencia entre el número total de nacimientos (N_t^i) y el número total de defunciones (D_t^i),

$$\boxed{\mathrm{CV}_t^i = \mathrm{N}_t^i - \mathrm{D}_t^i} \tag{1}$$

- Ind2: El **Saldo Migratorio** del barrio i en el año t (SM_t^i) se define como la diferencia entre el número total de inmigraciones (I_t^i) y el número total de emigraciones (E_t^i),

$$\boxed{\mathrm{SM}_t^i = \mathrm{I}_t^i - \mathrm{E}_t^i} \tag{2}$$

- Ind3: El **Saldo de Movimientos Intraurbanos** del barrio i en el año t (SMI_t^i) se define como la diferencia entre el número total de llegadas procedentes de otro barrio por cambio de domicilio (LCD_t^i) y el número total de salidas hacia otro barrio por cambio de domicilio (SCD_t^i),

$$\boxed{\mathrm{SMI}_t^i = \mathrm{LCD}_t^i - \mathrm{SCD}_t^i} \tag{3}$$

- Ind4: La **Tasa de Natalidad** del barrio i en el año t (TN_t^i) se define como el cociente entre el número total de nacimientos en el año t (N_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TN}_t^i = 1000 \cdot \frac{\mathrm{N}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{4}$$

- Ind5: La **Tasa General de Fecundidad** del barrio i en el año t (TGF_t^i) se define como el cociente entre el número total de nacimientos en el año t (N_t^i) y el promedio del número de mujeres (PobM) entre 15 y 49 años de edad en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TGF}_t^i = 1000 \cdot \frac{\mathrm{N}_t^i}{\left(\sum_{x=15}^{49} \mathrm{PobM}_{x,t}^i + \sum_{x=15}^{49} \mathrm{PobM}_{x,t}^i\right)/2}} \tag{5}$$

- Ind6: La **Tasa de Mortalidad** del barrio i en el año t (TM_t^i) se define como el cociente entre el número total de defunciones en el año t (D_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TM}_t^i = 1000 \cdot \frac{\mathrm{N}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{6}$$

- Ind7: La **Tasa de Inmigración** del barrio i en el año t (TI_t^i) se define como el cociente entre el número total de inmigraciones en el año t (I_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TI}_t^i = 1000 \cdot \frac{\mathrm{I}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{7}$$

- Ind8: La **Tasa de Emigración** del barrio i en el año t (TE_t^i) se define como el cociente entre el número total de emigraciones en el año t (E_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TE}_t^i = 1000 \cdot \frac{\mathrm{E}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{8}$$

- Ind9: La **Tasa de Llegadas por Cambio de Domicilio** del barrio i en el año t (TLCD_t^i) se define como el cociente entre el número total de llegadas procedentes de otro barrio por cambio de domicilio en el año t (LCD_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TLCD}_t^i = 1000 \cdot \frac{\mathrm{LCD}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{9}$$

- Ind10: La **Tasa de Salidas por Cambio de Domicilio** del barrio i en el año t (TSCD_t^i) se define como el cociente entre el número total de salidas hacia otro barrio por cambio de domicilio en el año t (SCD_t^i) y el promedio de la población total (PobT) en los años t y $t+1$ multiplicado por 1000,

$$\boxed{\mathrm{TSCD}_t^i = 1000 \cdot \frac{\mathrm{SCD}_t^i}{(\mathrm{PobT}_t^i + \mathrm{PobT}_{t+1}^i)/2}} \tag{10}$$

- Ind11: La **Relación de Masculinidad al Nacimiento** del barrio i en el año t (RMN_t^i) se define como el cociente entre el número total de nacimientos de sexo masculino (NH_t^i) y el número total de nacimientos de sexo femenino (NM_t^i) multiplicado por 100,

$$\boxed{\mathrm{RMN}_t^i = 100 \cdot \frac{\mathrm{NH}_t^i}{\mathrm{NM}_t^i}} \tag{11}$$

- Ind12: El **Índice de Sundbarg** del barrio i en el año t (IS_t^i) se define como el cociente entre la población de 0 a 14 años de edad y la población de 50 años o más,

$$\boxed{\mathrm{IS}_t^i = \frac{\sum_{x=0}^{14} \mathrm{PobT}_{x,t}^i}{\sum_{x\geq 50} \mathrm{PobT}_{x,t}^i}} \tag{12}$$

- Ind13: El **Índice de Friz** del barrio i en el año t (IF_t^i) se define como el cociente entre la población de 0 a 19 años de edad y la población entre 30 y 49 años de edad multiplicado por 100,

$$\boxed{\mathrm{IF}_t^i = 100 \cdot \frac{\sum_{x=0}^{19} \mathrm{PobT}_{x,t}^i}{\sum_{x=30}^{49} \mathrm{PobT}_{x,t}^i}} \tag{13}$$

- Ind14: El **Índice de Burgdofer** del barrio i en el año t (IB_t^i) se define como el cociente entre la población de 5 a 14 años de edad y la población entre 45 y 64 años de edad multiplicado por 100,

$$\boxed{\mathrm{IB}_t^i = 100 \cdot \frac{\sum_{x=5}^{14} \mathrm{PobT}_{x,t}^i}{\sum_{x=45}^{64} \mathrm{PobT}_{x,t}^i}} \tag{14}$$

- Ind15: El **Índice Generacional de Ancianos** del barrio i en el año t (IGA_t^i) se define como el cociente entre la población de 35 a 64 años de edad y la población de 65 años o más,

$$\boxed{\mathrm{IGA}_t^i = \frac{\sum_{x=35}^{64} \mathrm{PobT}_{x,t}^i}{\sum_{x\geq 65} \mathrm{PobT}_{x,t}^i}} \tag{15}$$

- Ind16: El **Índice de Envejecimiento** del barrio i en el año t (IE_t^i) se define como el cociente entre la población de 65 años o más y la población de 0 a 15 años de edad multiplicado por 100,

$$\boxed{\mathrm{IE}_t^i = 100 \cdot \frac{\sum_{x\geq 65} \mathrm{PobT}_{x,t}^i}{\sum_{x=0}^{15} \mathrm{PobT}_{x,t}^i}} \tag{16}$$

- Ind17: El **Índice de Sobreenvejecimiento** del barrio i en el año t (ISE_t^i) se define como el cociente entre la población de 85 años o más y la población de 65 años o más multiplicado por 100,

$$\boxed{\mathrm{ISE}_t^i = 100 \cdot \frac{\sum_{x\geq 85} \mathrm{PobT}_{x,t}^i}{\sum_{x\geq 65} \mathrm{PobT}_{x,t}^i}} \tag{17}$$

- Ind18: El **Índice Demográfico de Dependencia** del barrio i en el año t (IDD_t^i) se define como el cociente entre la suma de la población de 0 a 15 años de edad más la población de 65 años o más y la población de 16 a 64 años multiplicado por 100,

$$\boxed{\mathrm{IDD}_t^i = 100 \cdot \frac{\left(\sum_{x=0}^{15} \mathrm{PobT}_{x,t}^i + \sum_{x\geq 65} \mathrm{PobT}_{x,t}^i\right)}{\sum_{x=16}^{64} \mathrm{PobT}_{x,t}^i}} \tag{18}$$

- Ind19: El **Índice de Estructura de la Población Activa** del barrio i en el año t (IEPA_t^i) se define como el cociente entre la población de 16 a 39 años de edad y la población de 40 a 64 años multiplicado por 100,

$$\boxed{\mathrm{IEPA}_t^i = 100 \cdot \frac{\sum_{x=16}^{39} \mathrm{PobT}_{x,t}^i}{\sum_{x=40}^{64} \mathrm{PobT}_{x,t}^i}} \tag{19}$$

- Ind20: El **Índice de Reemplazamiento de la Población Activa** del barrio i en el año t (IRPA_t^i) se define como el cociente entre la población de 60 a 64 años de edad y la población de 15 a 19 años multiplicado por 100,

$$\boxed{\mathrm{IRPA}_t^i = 100 \cdot \frac{\sum_{x=60}^{64} \mathrm{PobT}_{x,t}^i}{\sum_{x=15}^{19} \mathrm{PobT}_{x,t}^i}} \tag{20}$$

- Ind21: El **Índice de Carga Preescolar** del barrio i en el año t (ICP_t^i), más conocido como "Índice del número de niños por mujer fecunda", se define como el cociente entre la población de 0 a 4 años de edad y el número de mujeres (PobM) entre 15 y 49 años de edad multiplicado por 100,

$$\boxed{\mathrm{ICP}_t^i = 100 \cdot \frac{\sum_{x=0}^{4} \mathrm{PobT}_{x,t}^i}{\sum_{x=15}^{49} \mathrm{PobM}_{x,t}^i}} \tag{21}$$

- Ind22: La **Razón de Progresividad Demográfica** del barrio i en el año t (RPD_t^i) se define como el cociente entre la población de 0 a 4 años de edad y la población de 5 a 9 años multiplicado por 100,

$$\boxed{\mathrm{RPD}_t^i = 100 \cdot \frac{\sum_{x=0}^{4} \mathrm{PobT}_{x,t}^i}{\sum_{x=5}^{9} \mathrm{PobT}_{x,t}^i}} \tag{22}$$

- Ind23: La **Relación de Masculinidad** del barrio i en el año t (RM_t^i) se define como el cociente entre el número total de hombres (PobH) y el número total de mujeres (PobM) multiplicado por 100,

$$\boxed{\mathrm{RM}_t^i = 100 \cdot \frac{\mathrm{PobH}_t^i}{\mathrm{PobM}_t^i}} \tag{23}$$

- Ind24: **Porcentaje de población extranjera**,

$$\boxed{\% = 100 \cdot \frac{\text{Población extranjera}_t^i}{\text{Población total}_t^i}} \tag{24}$$

- Ind25: **Porcentaje de población de 65 años o más**,

$$\boxed{\% = 100 \cdot \frac{\text{Población de 65 años o más}_t^i}{\text{Población total}_t^i}} \tag{25}$$

- Ind26: **Porcentaje de población de 15 años o menos**,

$$\boxed{\% = 100 \cdot \frac{\text{Población de 15 años o menos}_t^i}{\text{Población total}_t^i}} \tag{26}$$

- Ind27: **Porcentaje de población nacida en la ciudad de València**,

$$\boxed{\% = 100 \cdot \frac{\text{Población nacida en la ciudad de València}_t^i}{\text{Población total}_t^i}} \tag{27}$$

- Ind28: **Porcentaje de hojas familiares con solo personas de 80 años o más**,

$$\boxed{\% = 100 \cdot \frac{\text{Hojas familiares con solo personas de 80 años o más}_t^i}{\text{Número total de hojas familiares}_t^i}} \tag{28}$$

- Ind29: **Porcentaje de hojas familiares sin menores**,

$$\boxed{\% = 100 \cdot \frac{\text{Hojas familiares sin menores}_t^i}{\text{Número total de hojas familiares}_t^i}} \tag{29}$$

- Ind30: **Media de personas por hoja familiar**,

$$\boxed{\overline{x} = \frac{\text{Población en hojas familiares}_t^i}{\text{Número total de hojas familiares}_t^i}} \tag{30}$$

Anexo II. Resultados del *EFA*

Listing 2: Cargas factoriales tras eliminar el Indicador 11

```
Standardized loadings (pattern matrix) based upon correlation matrix
         TC1    TC2    TC3    h2     u2  com
Ind01  0.595                0.351 0.6491 1.05
Ind02                       0.189 0.8110 1.96
Ind03                       0.196 0.8036 2.83
Ind04                0.514  0.442 0.5583 2.47
Ind05                0.558  0.351 0.6488 1.41
Ind06 −0.636                0.507 0.4926 1.23
Ind07          0.729        0.677 0.3234 1.26
Ind08          0.773        0.581 0.4193 1.00
Ind09          0.821        0.707 0.2930 1.01
Ind10          0.707        0.712 0.2881 1.31
Ind12  0.980                0.923 0.0775 1.24
Ind13  0.538 −0.582         0.779 0.2208 2.74
Ind14  0.834                0.747 0.2532 1.36
Ind15  0.912                0.806 0.1942 1.36
Ind16 −0.996                0.906 0.0939 1.20
Ind17 −0.517                0.431 0.5692 1.70
Ind18        −0.706  0.506  0.637 0.3628 2.45
Ind19                       0.319 0.6809 1.78
Ind20 −0.617                0.431 0.5691 1.76
Ind21  0.544         0.605  0.566 0.4336 2.02
Ind22                       0.249 0.7513 1.69
Ind23  0.534  0.518         0.570 0.4304 2.96
Ind24          0.923        0.858 0.1419 1.00
Ind25 −0.911 −0.510         0.838 0.1617 1.60
Ind26  0.945                0.919 0.0808 1.24
Ind27        −0.740         0.567 0.4335 1.00
Ind28 −0.753                0.666 0.3345 1.33
Ind29 −0.917                0.873 0.1267 1.02
Ind30                       0.476 0.5237 2.01
Ind31        −0.563  0.741  0.657 0.3434 1.87
Ind32        −0.646  0.739  0.738 0.2625 1.98
Ind33          0.784        0.616 0.3836 1.53
```

Listing 3: Cargas factoriales tras eliminar el Indicador 2

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.576                0.354 0.6460 1.10
Ind03                       0.187 0.8134 2.60
Ind04                0.569  0.509 0.4908 1.94
Ind05                0.613  0.422 0.5784 1.12
Ind06 -0.674                0.520 0.4799 1.24
Ind07         0.705         0.653 0.3471 1.26
Ind08         0.771         0.600 0.4001 1.00
Ind09         0.809         0.706 0.2939 1.02
Ind10         0.682         0.680 0.3197 1.38
Ind12  0.918                0.924 0.0759 1.27
Ind13        -0.601         0.776 0.2236 2.60
Ind14  0.771                0.742 0.2576 1.41
Ind15  0.921                0.808 0.1924 1.34
Ind16 -0.961                0.904 0.0963 1.19
Ind17 -0.574                0.424 0.5755 1.56
Ind18        -0.711  0.509  0.645 0.3545 2.59
Ind19                       0.339 0.6611 1.74
Ind20 -0.551                0.403 0.5975 1.72
Ind21                0.642  0.610 0.3899 1.82
Ind22                       0.280 0.7196 1.75
Ind23  0.594  0.538         0.549 0.4505 2.78
Ind24         0.907         0.845 0.1550 1.00
Ind25 -0.909                0.837 0.1625 1.58
Ind26  0.887                0.915 0.0847 1.26
Ind27        -0.722         0.542 0.4575 1.01
Ind28 -0.797                0.657 0.3428 1.27
Ind29 -0.908                0.874 0.1262 1.03
Ind30                       0.472 0.5283 1.98
Ind31        -0.589  0.707  0.631 0.3688 2.04
Ind32        -0.672  0.704  0.715 0.2853 2.00
Ind33         0.795         0.613 0.3867 1.47
```

Listing 4: Cargas factoriales tras eliminar el Indicador 3

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.590                0.351 0.6493 1.09
Ind04                0.583  0.533 0.4666 2.23
Ind05                0.645  0.462 0.5381 1.28
Ind06 -0.634                0.532 0.4677 1.37
Ind07         0.726         0.645 0.3546 1.21
Ind08         0.781         0.601 0.3994 1.01
Ind09         0.816         0.704 0.2955 1.01
Ind10         0.703         0.672 0.3281 1.24
Ind12  0.975                0.925 0.0746 1.20
Ind13  0.532 -0.586         0.772 0.2280 2.56
Ind14  0.827                0.743 0.2574 1.29
Ind15  0.910                0.822 0.1779 1.42
Ind16 -0.991                0.906 0.0936 1.18
Ind17 -0.518                0.428 0.5717 1.75
Ind18        -0.712  0.547  0.669 0.3315 2.51
Ind19                       0.342 0.6578 1.98
Ind20 -0.610                0.399 0.6007 1.54
Ind21  0.541         0.641  0.630 0.3702 1.98
Ind22                       0.289 0.7108 1.95
Ind23  0.537  0.532         0.536 0.4636 2.89
Ind24         0.916         0.845 0.1546 1.00
Ind25 -0.910 -0.514         0.851 0.1485 1.65
Ind26  0.939                0.916 0.0841 1.17
Ind27        -0.735         0.543 0.4566 1.00
Ind28 -0.753                0.665 0.3346 1.40
Ind29 -0.910                0.872 0.1283 1.02
Ind30                       0.467 0.5329 2.01
Ind31        -0.560  0.680  0.586 0.4136 1.94
Ind32        -0.644  0.675  0.676 0.3240 2.01
Ind33         0.783         0.597 0.4034 1.42
```

Listing 5: Cargas factoriales tras eliminar el Indicador 22

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.608                0.357 0.6428 1.09
Ind04                0.555 0.467 0.5329 2.38
Ind05                0.603 0.389 0.6108 1.37
Ind06 -0.645                0.541 0.4593 1.25
Ind07         0.728         0.661 0.3387 1.25
Ind08         0.782         0.602 0.3978 1.01
Ind09         0.817         0.712 0.2881 1.02
Ind10         0.704         0.695 0.3052 1.27
Ind12  0.991                0.926 0.0737 1.31
Ind13  0.532 -0.588         0.776 0.2245 2.79
Ind14  0.831                0.744 0.2555 1.50
Ind15  0.927                0.827 0.1729 1.34
Ind16 -1.005                0.904 0.0959 1.23
Ind17 -0.530                0.457 0.5426 1.67
Ind18        -0.710  0.512  0.670 0.3301 2.48
Ind19                       0.299 0.7011 1.70
Ind20 -0.610                0.418 0.5820 1.84
Ind21  0.558         0.657  0.598 0.4016 1.98
Ind23  0.548  0.531         0.559 0.4414 2.83
Ind24         0.917         0.854 0.1460 1.00
Ind25 -0.925 -0.512         0.851 0.1489 1.58
Ind26  0.950                0.915 0.0845 1.32
Ind27        -0.735         0.553 0.4473 1.00
Ind28 -0.773                0.703 0.2965 1.30
Ind29 -0.925                0.874 0.1255 1.04
Ind30                       0.473 0.5273 1.99
Ind31        -0.558  0.702  0.614 0.3860 1.91
Ind32        -0.643  0.709  0.704 0.2964 2.00
Ind33         0.782         0.600 0.4003 1.44
```

Listing 6: Cargas factoriales tras eliminar el Indicador 19

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.599                0.366 0.6337 1.01
Ind04                       0.443 0.5575 2.59
Ind05                0.545 0.364 0.6365 1.44
Ind06 -0.632                0.568 0.4318 1.53
Ind07         0.733         0.666 0.3344 1.19
Ind08         0.791         0.604 0.3956 1.01
Ind09         0.827         0.713 0.2871 1.01
Ind10         0.713         0.701 0.2994 1.24
Ind12  0.966                0.924 0.0756 1.14
Ind13  0.523 -0.610         0.777 0.2233 2.63
Ind14  0.809                0.743 0.2569 1.34
Ind15  0.896                0.811 0.1886 1.46
Ind16 -0.978                0.900 0.1004 1.15
Ind17 -0.520                0.493 0.5067 1.93
Ind18        -0.715  0.596  0.666 0.3337 2.53
Ind20 -0.598                0.422 0.5776 1.61
Ind21  0.556         0.568  0.585 0.4146 2.06
Ind23  0.526  0.531         0.555 0.4450 2.99
Ind24         0.927         0.851 0.1494 1.00
Ind25 -0.898                0.846 0.1540 1.70
Ind26  0.925                0.911 0.0887 1.18
Ind27        -0.743         0.548 0.4525 1.00
Ind28 -0.753                0.717 0.2826 1.56
Ind29 -0.901                0.874 0.1259 1.03
Ind30                       0.500 0.5005 2.06
Ind31        -0.570  0.746  0.626 0.3738 1.88
Ind32        -0.659  0.737  0.709 0.2906 1.99
Ind33         0.794         0.602 0.3977 1.54
```

Listing 7: Cargas factoriales tras eliminar el Indicador 4

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.599                0.359 0.6415 1.01
Ind05                       0.179 0.8213 1.62
Ind06 -0.660                0.556 0.4439 1.23
Ind07         0.765         0.713 0.2871 1.26
Ind08         0.786         0.596 0.4038 1.01
Ind09         0.828         0.715 0.2852 1.01
Ind10         0.733         0.728 0.2722 1.26
Ind12  0.990                0.919 0.0810 1.33
Ind13  0.550 -0.532         0.804 0.1959 2.97
Ind14  0.834                0.760 0.2404 1.54
Ind15  0.929                0.809 0.1905 1.34
Ind16 -1.009                0.907 0.0934 1.28
Ind17 -0.535                0.510 0.4901 1.78
Ind18        -0.689         0.645 0.3549 2.51
Ind20 -0.633                0.510 0.4899 2.14
Ind21  0.545         0.553  0.498 0.5022 2.01
Ind23  0.539                0.583 0.4174 2.95
Ind24         0.929         0.868 0.1316 1.00
Ind25 -0.933 -0.532         0.851 0.1492 1.59
Ind26  0.954                0.928 0.0723 1.35
Ind27        -0.753         0.582 0.4180 1.00
Ind28 -0.778                0.718 0.2816 1.31
Ind29 -0.932                0.876 0.1239 1.03
Ind30                       0.497 0.5029 1.97
Ind31                0.755  0.670 0.3302 1.74
Ind32        -0.579  0.765  0.755 0.2446 1.88
Ind33         0.750         0.612 0.3882 1.60
```

Listing 8: Cargas factoriales tras eliminar el Indicador 5

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01  0.590                0.359 0.6414 1.00
Ind06 -0.666                0.535 0.4650 1.15
Ind07         0.788         0.742 0.2583 1.27
Ind08         0.782         0.589 0.4111 1.01
Ind09         0.830         0.709 0.2910 1.00
Ind10         0.751         0.739 0.2612 1.24
Ind12  0.984                0.906 0.0944 1.30
Ind13  0.551         0.514  0.820 0.1795 2.98
Ind14  0.829                0.748 0.2516 1.50
Ind15  0.934                0.808 0.1925 1.33
Ind16 -1.007                0.906 0.0945 1.28
Ind17 -0.537                0.500 0.5004 1.75
Ind18        -0.675         0.615 0.3846 2.48
Ind20 -0.636                0.537 0.4629 2.24
Ind21  0.527                0.419 0.5808 1.97
Ind23  0.536                0.609 0.3914 2.96
Ind24         0.932         0.867 0.1329 1.00
Ind25 -0.938 -0.536         0.854 0.1457 1.59
Ind26  0.950                0.924 0.0764 1.34
Ind27        -0.766         0.606 0.3940 1.02
Ind28 -0.784                0.700 0.3000 1.25
Ind29 -0.931                0.877 0.1228 1.03
Ind30                       0.504 0.4957 1.98
Ind31                0.792  0.720 0.2799 1.58
Ind32        -0.529  0.801  0.801 0.1990 1.75
Ind33         0.718         0.634 0.3662 1.76
```

Listing 9: Cargas factoriales tras eliminar el Indicador 30

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01 -0.575                0.364 0.6357 1.11
Ind06  0.648                0.548 0.4517 1.47
Ind07          0.682  0.619 0.715 0.2852 2.00
Ind08          0.749        0.589 0.4107 1.37
Ind09          0.780        0.706 0.2943 1.54
Ind10          0.656  0.587 0.708 0.2920 2.01
Ind12 -0.955                0.908 0.0921 1.15
Ind13 -0.527 -0.580         0.807 0.1930 2.17
Ind14 -0.801                0.751 0.2494 1.30
Ind15 -0.917                0.825 0.1746 1.28
Ind16  0.980                0.909 0.0910 1.18
Ind17  0.528                0.527 0.4726 1.98
Ind18         -0.700        0.623 0.3769 1.87
Ind20  0.608        -0.502  0.554 0.4464 2.08
Ind21 -0.506                0.429 0.5709 1.95
Ind23 -0.526  0.527         0.607 0.3933 2.70
Ind24          0.895        0.901 0.0988 1.48
Ind25  0.918                0.860 0.1401 1.52
Ind26 -0.920                0.922 0.0777 1.15
Ind27         -0.713        0.600 0.4001 1.61
Ind28  0.771                0.743 0.2574 1.49
Ind29  0.897                0.850 0.1498 1.02
Ind31         -0.602  0.500 0.689 0.3115 1.94
Ind32         -0.675        0.783 0.2172 1.80
Ind33          0.799        0.651 0.3486 1.02
```

Listing 10: Cargas factoriales tras eliminar el Indicador 23

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01 -0.584                0.360 0.6396 1.01
Ind06  0.681                0.595 0.4055 1.31
Ind07          0.775        0.704 0.2960 1.24
Ind08          0.776        0.585 0.4154 1.00
Ind09          0.825        0.705 0.2952 1.01
Ind10          0.736        0.699 0.3014 1.23
Ind12 -0.970                0.909 0.0908 1.30
Ind13 -0.536                0.800 0.1999 2.98
Ind14 -0.809                0.760 0.2400 1.55
Ind15 -0.943                0.837 0.1631 1.33
Ind16  1.000                0.909 0.0914 1.26
Ind17  0.552                0.562 0.4380 1.89
Ind18         -0.664        0.656 0.3438 2.68
Ind20  0.620                0.553 0.4467 2.32
Ind21 -0.500         0.511  0.459 0.5410 2.01
Ind24          0.942        0.902 0.0984 1.00
Ind25  0.944 -0.527         0.869 0.1306 1.58
Ind26 -0.933                0.926 0.0739 1.34
Ind27         -0.779        0.600 0.4001 1.00
Ind28  0.795                0.757 0.2425 1.35
Ind29  0.917                0.849 0.1506 1.03
Ind31                0.726  0.653 0.3474 1.76
Ind32         -0.558  0.744 0.750 0.2495 1.87
Ind33          0.751        0.626 0.3738 1.54
```

Listing 11: Cargas factoriales tras eliminar el Indicador 13

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2     u2  com
Ind01 -0.596                0.365 0.6354 1.02
Ind06  0.692                0.594 0.4058 1.21
Ind07          0.770        0.704 0.2965 1.24
Ind08          0.770        0.580 0.4196 1.00
Ind09          0.822        0.705 0.2947 1.01
Ind10          0.733        0.698 0.3022 1.22
Ind12 -0.974                0.922 0.0782 1.43
Ind14 -0.805         0.512  0.757 0.2432 1.70
Ind15 -0.954                0.839 0.1612 1.28
Ind16  1.003                0.910 0.0897 1.34
Ind17  0.567                0.567 0.4331 1.77
Ind18         -0.659        0.655 0.3451 2.64
Ind20  0.600                0.525 0.4751 2.43
Ind21                0.571  0.481 0.5193 1.97
Ind24          0.939        0.903 0.0974 1.00
Ind25  0.949 -0.522         0.870 0.1304 1.55
Ind26 -0.933                0.926 0.0744 1.48
Ind27         -0.782        0.609 0.3909 1.00
Ind28  0.816                0.757 0.2429 1.19
Ind29  0.925                0.846 0.1543 1.08
Ind31                0.728  0.660 0.3403 1.75
Ind32         -0.551  0.747 0.744 0.2560 1.85
Ind33          0.752        0.641 0.3588 1.57
```

Listing 12: Cargas factoriales tras eliminar el Indicador 18

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2      u2  com
Ind01 -0.589                0.363 0.6367 1.01
Ind06  0.678                0.557 0.4426 1.25
Ind07          0.733        0.720 0.2795 1.35
Ind08          0.779        0.589 0.4112 1.00
Ind09          0.821        0.707 0.2932 1.01
Ind10          0.703        0.712 0.2879 1.34
Ind12 -0.965                0.919 0.0814 1.24
Ind14 -0.809                0.727 0.2731 1.39
Ind15 -0.924                0.792 0.2080 1.26
Ind16  0.983                0.904 0.0959 1.23
Ind17  0.562                0.533 0.4669 1.83
Ind20  0.592                0.507 0.4934 2.25
Ind21 -0.506                0.418 0.5818 1.98
Ind24          0.951        0.912 0.0884 1.00
Ind25  0.916                0.819 0.1814 1.52
Ind26 -0.926                0.913 0.0870 1.27
Ind27         -0.772        0.610 0.3903 1.00
Ind28  0.791                0.694 0.3059 1.21
Ind29  0.924                0.860 0.1402 1.00
Ind31         -0.571  0.839 0.819 0.1806 1.78
Ind32         -0.626  0.826 0.862 0.1382 1.87
Ind33          0.820 -0.525 0.752 0.2479 1.71
```

Listing 13: Cargas factoriales tras eliminar el Indicador 20

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2      u2  com
Ind01 -0.606                0.372 0.6282 1.02
Ind06  0.687                0.540 0.4605 1.15
Ind07          0.731        0.754 0.2461 1.39
Ind08          0.788        0.604 0.3959 1.00
Ind09          0.825        0.712 0.2884 1.01
Ind10          0.698        0.725 0.2750 1.35
Ind12 -0.969                0.916 0.0837 1.26
Ind14 -0.801                0.688 0.3121 1.34
Ind15 -0.939                0.805 0.1954 1.27
Ind16  0.984                0.894 0.1064 1.25
Ind17  0.574                0.502 0.4985 1.68
Ind21 -0.511                0.412 0.5881 1.97
Ind24          0.953        0.907 0.0928 1.00
Ind25  0.923 -0.503         0.823 0.1773 1.55
Ind26 -0.923                0.888 0.1120 1.26
Ind27         -0.775        0.625 0.3755 1.01
Ind28  0.814                0.683 0.3173 1.12
Ind29  0.921                0.851 0.1494 1.00
Ind31         -0.592  0.892 0.875 0.1248 1.76
Ind32         -0.649  0.867 0.895 0.1047 1.85
Ind33          0.832 -0.588 0.781 0.2194 1.80
```

Listing 14: Cargas factoriales tras eliminar el Indicador 21

```
Standardized loadings (pattern matrix) based upon correlation matrix
        TC1    TC2    TC3    h2      u2  com
Ind01 -0.609                0.374 0.6256 1.11
Ind06  0.709                0.524 0.4763 1.01
Ind07          0.718        0.782 0.2175 1.52
Ind08          0.782        0.608 0.3924 1.00
Ind09          0.821        0.711 0.2888 1.01
Ind10          0.689        0.744 0.2564 1.43
Ind12 -0.960                0.875 0.1249 1.29
Ind14 -0.792                0.641 0.3587 1.34
Ind15 -0.956                0.827 0.1735 1.27
Ind16  0.981                0.876 0.1244 1.26
Ind17  0.593                0.450 0.5495 1.27
Ind24          0.951        0.905 0.0953 1.01
Ind25  0.934                0.834 0.1664 1.52
Ind26 -0.915                0.848 0.1516 1.29
Ind27         -0.763        0.651 0.3494 1.06
Ind28  0.833                0.675 0.3249 1.01
Ind29  0.919                0.844 0.1562 1.01
Ind31         -0.609  0.924 0.951 0.0492 1.75
Ind32         -0.665  0.893 0.949 0.0510 1.85
Ind33          0.845 -0.639 0.847 0.1525 1.87
```

Anexo III. Código *R* utilizado

Instalación y carga de librerías

https://raw.githubusercontent.com/JuanferMG/TFM/master/TFM%201.%20librerias.R

Lectura de datos y proyecciones ARIMA

https://raw.githubusercontent.com/JuanferMG/TFM/master/TFM%202.%20arima.R

Análisis Factorial Exploratorio

https://raw.githubusercontent.com/JuanferMG/TFM/master/TFM%203.%20factorial.R

Análisis Cluster

https://raw.githubusercontent.com/JuanferMG/TFM/master/TFM%204.%20cluster.R

Análisis de Autocorrelación Espacial

https://raw.githubusercontent.com/JuanferMG/TFM/master/TFM%205.%20espacial.R

Referencias

[1] Anselin, L. (2018). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. *Geographical Analysis, 51, 133-150.* doi: 10.1111/gean.12164

[2] Anselin, L. (2019). Local Spatial Autocorrelation. *GeoDa: An Introduction to Spatial Data Analysis.* Recuperado de: `https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html#significance`

[3] Arifin, W.N. (2017). *Exploratory factor analysis and Cronbach's alpha. Questionnaire Validation Workshop, 10/10/2017, USM Health Campus, Universiti Sains Malaysia.* Recuperado de `https://wnarifin.github.io/workshop/qvw2017/efa.pdf`

[4] Bakar, K.S. & Sahu, S. (2015). spTimer: Spatio-Temporal Bayesian Modeling Using R. *Journal of Statistical Software, 63(15).* doi: 10.18637/jss.v063.i15

[5] Bakka, H., Rue, H., Fuglstad, G.A., Riebler, A., Bolin, D., Illian, J., ... Lindgren, F. (2018). Spatial modelling with R-INLA: A review. *WIREs Computational Statistics, 10(6), e1443.* doi: 10.1002/wics.1443

[6] Betzuen, A. (2010). Un Análisis sobre las posibilidades de prediccion de la mortalidad futura aplicando el modelo Lee-Carter. *Anales del Instituto de Actuarios Españoles 2010, 111-140.* Recuperado de `https://app.mapfre.com/documentacion/publico/en/catalogo_imagenes/grupo.do?path=1062078`

[7] Bivand, R., Pebesma, E. & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R.* Nueva York, NY: Springer.

[8] Bivand, R., Wilk, J. & Kossowski, T. (2017). Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests. *Spatial Statistics, 21, 339–361.* doi: 10.1016/j.spasta.2017.03.003

[9] Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology, 4, 33-49.* doi: 10.1016/j.sste.2012.12.001

[10] Celemín, J.P. (2009). Autocorrelación espacial e indicadores locales de asociación espacial. Importancia, estructura y aplicación. *Revista Universitaria de Geografía, 18, 11-31.* Recuperado de `https://www.redalyc.org/articulo.oa?id=383239099001`

[11] Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software, 61(6).* doi: 10.18637/jss.v061.i06

[12] Dalgic, T. (2006). *Handbook of Niche Marketing: Principles and Practice.* Nueva York, NY: The Haworth Reference Press.

[13] Dvoskin, R. (2004). *Fundamentos de marketing: teoría y experiencia.* Buenos Aires, Argentina: Granica.

[14] Farris, J. (1969). On the Cophenetic Correlation Coefficient. *Systematic Biology, 18(03), 279-285.* doi: 10.2307/2412324

[15] Frias-Navarro, D. (2019). *Apuntes de consistencia interna de las puntuaciones de un instrumento de medida.* Valencia, España: Universidad de Valencia. España. Recuperado de `https://www.uv.es/friasnav/AlfaCronbach.pdf`

[16] de la Fuente, S. (2011). *Análisis Factorial.* Madrid, España: Universidad Autónoma de Madrid. Recuperado de `http://www.fuenterrebollo.com`

[17] Glen, S. (2016). Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy. *StatisticsHowTo.com: Elementary Statistics for the rest of us!*. Recuperado de: `https://www.statisticshowto.com/kaiser-meyer-olkin/`

[18] Gordon, C. & Strauss, K. (2008). Individual pension-related risk propensities: the effects of socio-demographic characteristics and a spousal pension entitlement on risk attitudes. *Ageing & Society, 28(06), 847-874*. doi: 10.1017/S0144686X08007083

[19] Granados, M.P. (1986). *Técnicas de proyección de población de áreas menores: aplicación y evaluación*. Santiago, Chile: Celade.

[20] Grekousis, G. (2020). *Spatial Analysis Methods and Practice: Describe – Explore – Explain through GIS*. Nueva York, NY: Cambridge University Press.

[21] Hartmann, K., Krois, J. & Waske, B. (2018). A simple example of factor analysis in R. *E-Learning Project SOGA: Statistics and Geospatial Data Analysis*. Berlín, Alemania: Department of Earth Sciences, Freie Universitaet Berlin.

[22] Holgersson, M. (1978). The limited value of cophenetic correlation as a clustering criterion. *Pattern Recognition, 10(4), 287-295*. doi: 10.1016/0031-3203(78)90038-9

[23] Instituto Nacional de Estadística. (2020). *Atlas de Distribución de la Renta de los Hogares*. Madrid, España: INE.

[24] Instituto Nacional de Estadística. (2020). *Indicadores Demográficos Básicos.*. Madrid, España: INE.

[25] Kamarianakis, Y. & Prastacos, P. (2006). Spatial Time-Series Modeling: A review of the proposed methodologies. *Working Papers 0604, University of Crete, Department of Economics*. Recuperado de `http://economics.soc.uoc.gr/wpa/docs/103_YK_AGILE_05.pdf`

[26] Kodinariya, T. & Makwana, P. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95*. Recuperado de `https://www.researchgate.net`

[27] Mtz. de Lejarza, I. (2007). *"Análisis Factorial" en Mtz. de Lejarza, Juan (Coord.) y otros: "PROYECTO CEACES: Contenedor Hipermedia de Estadística Aplicada a las Ciencias Económicas y Sociales"*. Valencia, España: Universidad de Valencia. Recuperado de `http://www.uv.es/ceaces`

[28] Mtz. de Lejarza, I. (2007). *"Análisis Cluster" en Mtz. de Lejarza, Juan (Coord.) y otros: "PROYECTO CEACES: Contenedor Hipermedia de Estadística Aplicada a las Ciencias Económicas y Sociales"*. Valencia, España: Universidad de Valencia. Recuperado de `http://www.uv.es/ceaces`

[29] Ley Orgánica 5/1985, de 19 de junio, del Régimen Electoral General. *Boletín Oficial del Estado*. Referencia: BOE-A-1985-11672.

[30] Llopis, J. (2013). Tema 18: ANÁLISIS FACTORIAL. *LA ESTADÍSTICA: UNA ORQUESTA HECHA INSTRUMENTO*. Recuperado de `https://jllopisperez.com/`

[31] Mavrou, I. (2015). Análisis factorial exploratorio: cuestiones conceptuales y metodológicas. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas, 19, 71-80*. Recuperado de: `https://www.nebrija.com/revista-linguistica/analisis-factorial-exploratorio.html`

[32] Medina, J. & Solymosi, R. (2019). Crime Mapping in R. *Crime Mapping, University of Manchester*. Recuperado de `https://maczokni.github.io/crimemapping_textbook_bookdown/`

[33] Mencken, F.C. & Barnett, C. (1999). Murder, Nonnegligent Manslaughter, and Spatial Autocorrelation

in Mid-South Counties. *Regional Research Institute Publications and Working Papers, 171.* Recuperado de https://researchrepository.wvu.edu/rri_pubs/171

[34] Montoya, O. (2007). Aplicación del análisis factorial a la investigación de mercados. Caso de estudio. *Scientia Et Technica, 1(35), 281-286.* doi: 10.22517/23447214.5443

[35] Muralidharan, K. (2010). A note on transformation, standardization and normalization. *The IUP Journal of Operations Management, 9(1-2), 116-122.* Recuperado de https://www.researchgate.net

[36] Naciones Unidas. (1986). *Manual X. Técnicas indirectas de estimación demográfica.* Nueva York, NY: United Nations.

[37] Oficina de Estadística del Ayuntamiento de València. (2019). *Altas y Bajas en el Padrón Municipal. Dinámica demográfica de la Ciudad de Valencia.* València, España: Ayuntamiento de València.

[38] Oficina de Estadística del Ayuntamiento de València. (2020). *Evolución del Seccionado Censal en la Ciudad de València.* València, España: Ayuntamiento de València.

[39] Oficina de Estadística del Ayuntamiento de València. (2019). *Padrón Municipal de Habitantes. Características de la Población de la Ciudad de Valencia.* València, España: Ayuntamiento de València.

[40] Organización Panamericana de la Salud. (2010). *Bases conceptuales Demográficas.* Recuperado de https://www.paho.org/es/nicaragua

[41] Palladino, A. (2010). *Introducción a la demografía.* Corrientes, Argentina: Universidad Nacional del Nordeste. Recuperado de https://med.unne.edu.ar/web/

[42] Parra, F. (2019). *Estadística y Machine Learning con R.* Recuperado de https://bookdown.org/content/2274/portada.html

[43] PATECO - Comercio y Territorio del Consejo de Cámaras de la C.V. (2009). *Atlas Sociocomercial de la Comunitat Valenciana 2009.* Recuperado de http://www.pateco.org/administracion/ficheros/Capitulo2_indicadores_demograficos.pdf

[44] Saraçli, S., Doğan, N. & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications, 203.* Recuperado de https://journalofinequalitiesandapplications.springeropen.com/track/pdf/10.1186/1029-242X-2013-203

[45] Siabato, W. & Guzmán-Manrique, J. (2019). La autocorrelación espacial y el desarrollo de la geografía cuantitativa. *Cuadernos de Geografía: Revista Colombiana de Geografía 28 (1): 1-22.* doi: 10.15446/rcdg.v28n1.76919

[46] da Silva, A.R. & Dias, C.T. (2013). A cophenetic correlation coefficient for Tocher's method. *Pesquisa Agropecuária Brasileira, 48(6), 589-596.* doi: 10.1590/S0100-204X2013000600003

[47] Wikle, C. K., Zammit-Mangion, A. & Cressie, N. (2019). *Spatio-Temporal Statistics with R.* Boca Raton, FL: Chapman & Hall/CRC.

[48] Zhao, Y., Ge, L., Zhou, Y., Sun, Z., Zheng, E., Wang, X., ...Cheng, H. (2018). A new Seasonal Difference Space-Time Autoregressive Integrated Moving Average (SD-STARIMA) model and spatiotemporal trend prediction analysis for Hemorrhagic Fever with Renal Syndrome (HFRS). *PLoS ONE 13(11): e0207518.* doi: 10.1371/journal.pone.0207518