FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADÍSTICA APLICADA

CURSO 2020/2021

TRABAJO FIN DE GRADO

Título: Estratificación de las secciones censales de la ciudad de Madrid a partir de datos sociodemográficos.

Alumno: Valentina Estephanía Crameri Ramírez

Tutor UCM: Eduardo Ortega Castelló

Tutor Ayuntamiento de Madrid: Antonio Bermejo Aguña

Junio 2021

UNIVERSIDAD COMPLUTENSE
MADRID

Tras cuatro años de esfuerzo y dedicación, quiero dar las gracias a mi familia por su apoyo, por ayudarme a conseguir todos mis propósitos y confiar en mí, porque sin ellos no hubiera sido posible finalizar esta etapa de estudios.

Expresar a los profesores mi gratitud por su dedicación y tantas enseñanzas, que me han permitido alcanzar mis objetivos y hacer posible convertirme en un profesional en el campo de la estadística.

Agradecer a mis compañeros por haber hecho de esta etapa universitaria una gran experiencia. A mis amigos, por tantas horas de estudio y por tanta dedicación y esfuerzo, gracias por haber estado ahí y haber hecho de estos años únicos e inolvidables.

Finalmente, agradecer al Ayuntamiento de Madrid por su colaboración, ayuda y apoyo en la elaboración de este trabajo; haciendo posible la realización de este proyecto que tanto me ha cautivado.

ÍNDICE

ILUSTRACIONES

TABLAS

RESUMEN

El propósito de este estudio es agrupar las 2443 secciones censales que dividen la ciudad de Madrid en diferentes grupos lo más parecido posible, en base a 23 variables sociodemográficas que se redujo a 4 variables no observables (factores). Posteriormente aplicar un análisis de conglomerados, del cual se obtuvo un total de 5 grupos más un grupo cuyas secciones se consideran atípicas. Todo esto para su posterior uso en diseños muestrales multietápicos.

ABSTRACT

The purpose of this study is to group the 2443 census sections that divide the city of Madrid into different groups as similar as possible, based on 23 sociodemographic variables that we reduce into 4 unobservable variables (factors), and then apply a cluster analysis, with which we obtain a total of 5 groups plus a group which sections are considered atypical. All this in order that they can then be used in multi-stage sample designs.

1. INTRODUCCIÓN

Las secciones censales son un referente geográfico de carácter tanto administrativo como estadístico, las cuales constituyen una unidad territorial clave para el Censo Electoral y también para trabajos de los Censos de Población y Vivienda o para las investigaciones por muestreo.

Esencialmente se trata de un área de terreno del término municipal de forma que, cada vivienda o habitante ha de pertenecer a una y sólo una sección; merece especial mención el tema de su tamaño en términos poblacionales, por cuanto la Ley de Régimen Electoral General asigna unos tamaños mínimos y máximos medidos en número de electores (ha de estar entre 500 y 2000 electores). Dentro de tales límites y pensando en su uso como unidad de trabajo en los censos, existe un segundo condicionante en términos de tamaño y esto no debe superar los 2.500 habitantes.

Como consecuencia de lo anterior, el seccionado es algo vivo y cambiante, de manera que será objeto de revisión generalmente anual, debiendo fusionarse aquellas secciones que tengan menos de 500 electores (excepto en municipios de sección única) y dividirse aquellas otras que sobrepasen los 2.000 electores o los 2.500 habitantes (las secciones resultantes de la partición deberán tener un mínimo de 500 electores).

Todo municipio se divide en uno o más secciones censales y no hay ninguna parte del municipio que no pertenezca a una sección censal. Las secciones de un municipio se agrupan en uno o más distritos censales y toda sección pertenece a un distrito censal, por lo tanto, todas las secciones deben ajustarse a 3 normas obligatorias:

1) Deben estar definidas mediante límites fácilmente identificables.
2) La división debe comprender todo el territorio del término municipal.
3) Las secciones pertenecientes a un núcleo urbano estarán formadas por manzanas completas de edificios. Excepcionalmente, una manzana podrá subdividirse por fachadas completas o por portales (solo en el caso de que supere el número máximo de electores).

Centrándonos en el municipio de Madrid he de señalar que la división administrativa vigente divide la ciudad en 21 distritos, estas a su vez constan de barrios que se componen de secciones censales. El seccionado actualmente vigente, del 1 de noviembre de 2017, está formado por un total de 2443 secciones censales.

La tabla 1.1, obtenida de la Subdirección General de Estadística, recopila toda la información referente a las secciones censales de la ciudad de Madrid. En ella se puede observar el nombre, código y número total de distritos, el nombre y código de los barrios en los que se divide, el código de las secciones que se encuentran en cada barrio y el número total de secciones censales que hay en los distritos.

En las siguientes imágenes, ilustración 1 y 2, se puede observar, por un lado, la división del distrito 1 (distrito Centro) en los 6 barrios que la conforman (Palacio, Embajadores, Cortes, Justicia, Universidad y Sol), y por el otro, sus 111 secciones censales.

Ilustración 1: Mapa del Distrito 1 de la ciudad de Madrid dividido en barrios.

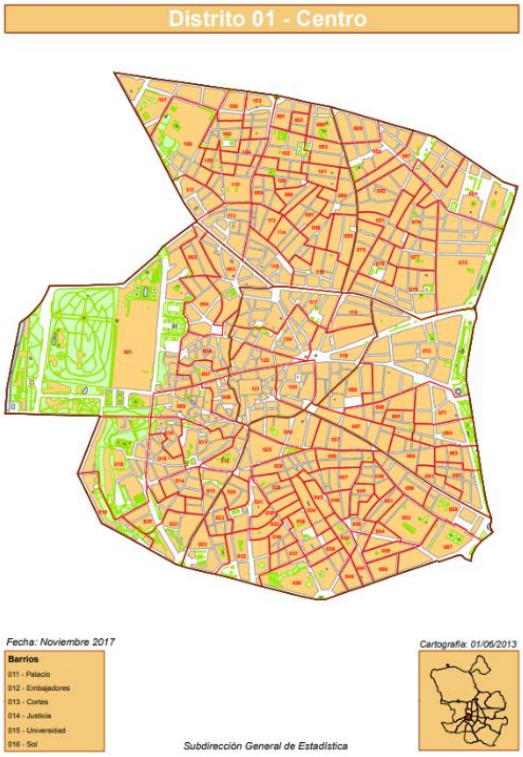

Ilustración 2: Mapa del Distrito 1 de la ciudad de Madrid dividido en secciones censales.

2. RECOGIDA DE LA INFORMACIÓN Y ESTRUCTURA DE LA BASE DE DATOS

Los datos utilizados para la elaboración de este proyecto proceden de distintas fuentes de información:

Del Padrón Municipal de Habitantes de la ciudad de Madrid revisado a 1 de enero de 2020 (Subdirección General de Estadística) se tiene información sobre:

- Indicadores de la Estructura Demográfica: La proporción de jóvenes menores de 16 años, la proporción de personas mayores de 65 años en adelante, la proporción de personas extranjeras según nacionalidad, el porcentaje de la población con nacionalidad no española y edad inferior a 16 años, la proporción de individuos nacidos fuera de España, el índice de dependencia (razón entre la suma de los grupos de 0 a 15 y de 65 y más años sobre el grupo de 16 a 64 años), el índice de reemplazo de la población activa (razón entre el grupo de 16 a 19 años y el grupo de 60 a 64 años) y el índice de estructura de la población activa (razón entre el grupo de 16 a 19 años y el grupo de 40 a 64 años).
- Población de 25 y más años clasificada por Nivel de estudios y Sexo según Distrito y Sección: El porcentaje de personas que han obtenido un título de estudio superior (considerando estudios superiores las siguientes la variables: Diplomado Escuela Universitaria, Arquitecto o Ingeniero Técnico, Licenciado Universitario, Titulado Estudios Superiores no Universitarios, Doctorado o Estudios Postgraduados), el porcentaje de personas que han obtenido un título correspondiente a estudios obligatorios (se considera estudios obligatorios las siguientes variables: No sabe leer ni escribir, Sin estudios, Enseñanza primaria incompleta, Bachiller elemental, Graduado escolar o ESO).
- Población clasificada por Sexo y Edad (grandes grupos), según Distrito y Sección, para cada Nacionalidad (españoles y extranjeros): Proporción de extranjeros en edad escolar.
- Hogares por Composición del hogar según Distrito y Sección: La proporción de hogares con un único habitante con edad igual o superior a 65 años, la proporción de hogares con un único habitante con edad inferior a 65 años, la proporción de hogares compuestos por una persona de entre 16 años y 64 años, y uno o más menores de 16 años, la proporción de hogares con más de una persona adulta con más de un menor de edad y la proporción de hogares con más de un adulto sin menores.

Del Atlas de Distribución de la Renta de los Hogares del 2016 al 2017 elaborado por el INE dentro de su apartado de “Estadística Experimental”, tenemos información acerca de:

- Indicadores de la Renta Media según Distritos, Barrios y Secciones Censales: La renta media por persona y la renta media por hogar (Se ha utilizado el año 2017).

Del Impuesto de Vehículos de Tracción Mecánica (IVTM). Agencia Tributaria de Madrid. Ayuntamiento de Madrid. Fecha de referencia padrón del impuesto del año 2019:

- Parque de vehículos existentes de Personas físicas por Sección censal según Tipo de vehículo y Potencia fiscal de los turismos: La proporción de turismos de persona física con 16 y más caballos fiscales.

Del Registro de demandantes de empleo. Servicio Público de Empleo Estatal (SEPE). Fichero de microdatos del municipio de Madrid. Fecha de referencia 31 de diciembre de 2019:

- Paro registrado en la ciudad de Madrid por secciones censales: El porcentaje de parados con edad comprendida entre los 16 y 64 años

De los Ficheros de Afiliación a la Seguridad Social (Tesorería General de la Seguridad Social -Ministerio de Empleo y Seguridad Social). Microdatos anonimizados. Fecha de referencia 1 de enero de 2020:

- Afiliados que trabajan en la ciudad de Madrid por Distritos, Barrios y Sección: La proporción de afiliados con edad comprendida entre los 16 y 64 años, la proporción de totales afiliados de los grupos 1 y 2, la proporción de total de afiliados del grupo 10.

Por tanto, la base de datos cuenta con 24 variables, de las cuales una de ellas es el identificador de la sección censal y el resto son variables cuantitativas, que aportan información de cada una de las 2443 secciones censales.

A continuación, se explicará brevemente que información proporciona cada variable:

- Edad promedio: Edad promedio de los habitantes de cada sección.
- Proporción de juventud: Porcentaje de la población residente en la sección censal con una edad inferior a 16 sobre la población total.
- Proporción de envejecimiento: Porcentaje de la población con edad igual o superior a 65 años, pertenecientes a una sección censal, sobre la población total.
- Proporción de extranjeros: Porcentaje de la población de nacionalidad no española, perteneciente a una sección censal, sobre la población total.
- Proporción de nacidos fuera de España: Porcentaje de la población nacida fuera de España, perteneciente a una sección censal, sobre la población total.
- Proporción de extranjeros en edad escolar: Porcentaje de la población de nacionalidad no española menor de 16 años, perteneciente a una sección censal, sobre la población total.
- Proporción de estudios superiores: Porcentaje de la población con un nivel de estudios de grado medio o superior sobre la población de 25 años o más, pertenecientes a una sección censal, sobre la población total.
- Proporción de estudios obligatorios e inferior: Porcentaje de la población con un nivel de estudios básicos o inferior sobre la población de 25 años o más, pertenecientes a una sección censal, sobre la población total.
- Proporción de hogares unipersonales de 65 años y más: Porcentaje de hogares en los que solo habita una persona de 65 años o más, pertenecientes a una sección censal, sobre el total de hogares.

- Proporción de hogares monoparentales: Porcentaje de hogares compuesto por una persona entre 16 años y 64 años, y uno o más menores de 16 años, pertenecientes a una sección censal, sobre el total de hogares.
- Proporción de hogares con más de una persona adulta con menores: Porcentaje de hogares en los que habitan más de una persona entre 16 y 64 años y con más de un menor de 16 años, pertenecientes a una sección censal, sobre el total de hogares.
- Proporción de turismo de PF con 16 y más CF: Porcentaje de turismos con 16 y más caballos fiscales (CF), pertenecientes a una sección censal, sobre el total de turismos obtenidos a partir del Impuesto de Vehículos de Tracción Mecánica (IVTM).
- Renta media por persona: Renta media anual por persona, pertenecientes a una sección censal.
- Renta media por hogar: Renta media anual por hogar, pertenecientes a una sección censal.
- Proporción de afiliados con edad entre 16 y 64 años: Porcentaje de personas dadas de alta en la Seguridad Social con una edad comprendida entre los 16 años y los 64 años.
- Proporción de afiliados totales de los grupos 1 y 2: Porcentaje de personas que cotizan en los grupos de cotización 1 y 2 sobre el total de afiliados al Régimen General. Los grupos 1 y 2 se corresponden con los de mayor base de cotización:
  - Grupo 1: Ingenieros y licenciados. Personal de alta dirección.
  - Grupo 2: Ingenieros técnicos, peritos y ayudantes titulados.
- Proporción de afiliados totales del grupo 10: Porcentaje de personas que cotizan en el grupo 10 sobre el total de afiliados al Régimen General.
- Proporción de parados con edad entre 16 y 64 años: Porcentaje de parados sobre la población potencialmente activa, esto es, individuos con una edad comprendida entre los 16 a 64 años.
- Índice de dependencia: Número de personas dependientes (menores de 16 años y mayores de 65 años y más) por cada 100 personas de la población activa de una sección censal.
- Índice de reemplazo población activa: Número de personas previstas que abandonen la edad activa por cada 100 personas de la población que se prevé que van a entrar.
- Índice de estructura de población activa: Número de personas activas con edad entre 40 a 64 años por cada 100 personas de la población activa de entre 16 a 39 años.
- Proporción de hogares unipersonales con edad inferior a 65 años: Porcentaje de hogares donde solo habita una persona con una edad inferior a los 65 años, pertenecientes a una sección censal, sobre el total de hogares.
- Proporción de hogares con más de un adulto sin menores: Porcentaje de hogares donde habitan más de una persona adulta sin menores, pertenecientes a una sección censal, sobre el total de hogares.

3. OBJETIVOS Y METODOLOGÍA

El objetivo de este estudio consiste en llevar a cabo un análisis que permita establecer una tipología de las secciones censales vigentes de la ciudad de Madrid, agrupando éstas en estratos que, puedan luego ser utilizados en diseños muestrales multietápicos.

Para la realización de este proyecto se utilizó en primer lugar la Matriz de Correlaciones Bivariadas para determinar si hay alguna variable que no esté muy relacionada con el resto para descartarla del estudio. Tras comprobar la correlación entre nuestras variables, se aplica un Análisis Factorial para reducir las dimensiones (reducir el número de variables) y, para finalizar, se emplea un Análisis Clúster para la agrupación de las secciones censales según las dimensiones obtenidas.

Se usa el software SAS para todo lo referente al análisis, el software R para realizar el gráfico de correlaciones para una interpretación más sencilla y, por último, el software QGIS para plasmar en un mapa de la ciudad de Madrid los grupos obtenidos en el estudio.

3.1. ANÁLISIS DE CORRELACIONES BIVARIADA

La correlación bivariada es una técnica estadística cuyo objetivo determinar la relación lineal entre dos variables, así como la intensidad de esta relación (alta, media o baja) y su dirección (positiva o negativa). Existen varios tipos de coeficientes de correlación, sin embargo, al tratarse de un estudio sobre toda la población, y no sobre una muestra, se utiliza el coeficiente de Pearson.

3.1.1. COEFICIENTE DE PEARSON

El coeficiente de correlación de Pearson mide la relación lineal entre dos variables continuas que se distribuyan de forma normal. En el caso de tratarse de una población, la fórmula se define como:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Cuando nos referimos a un estadístico muestral, la fórmula es la siguiente:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Este coeficiente toma valores entre -1 y 1, siendo 1 una correlación positiva perfecta, -1 una correlación negativa perfecta y 0 una ausencia de correlación.

3.2. ANÁLISIS FACTORIAL

El Análisis Factorial es un modelo de regresión múltiple cuyo objetivo es reducir la dimensión y buscar nuevas variables (factores) que expliquen nuestras variables originales. Se agrupan mediante sus correlaciones, de manera que, todas las variables dentro de un grupo estén altamente correlacionadas entre ellas, pero tengan una correlación relativamente baja con las otras variables que a su vez se encuentran en grupos diferentes.

Este modelo postula que nuestra variable de estudio (X) es linealmente dependiente de unas pocas variables aleatorias inobservables $(F_1, \dots, F_m)$ llamadas Factores Comunes, junto con p fuentes de variación $(\varepsilon_1, \dots, \varepsilon_p)$ llamadas Factores Específicos, únicos, o errores.

- Comunalidad: Expresa la parte de cada variable (su variabilidad) que puede ser explicada por los factores comunes a todas ellas.
- Especificidad: Es el término opuesto a comunalidad ya que expresa la parte específica de cada variable que escapa a los factores comunes.

$$\begin{aligned} X_1 &= l_{11}F_1 + \cdots + l_{1m}F_m + \varepsilon_1 \\ &\vdots \\ X_p &= l_{pm}F_1 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

Donde l_{ij} se denomina carga de la i-ésima variable del j-ésimo Factor, es decir, el peso que tiene cada variable en cada Factor.

3.2.1. ÍNDICE KMO DE KAISER-MEYER-OLKIN

Este índice es una medida global del grado de correlación entre las variables del estudio. Según el valor que tome este índice podemos obtener diferentes conclusiones:

- Si KMO es inferior a 0.5, se desaconseja realizar el Análisis Factorial.
- Si KMO se encuentra entre 0.5 y 0.6 se dice que la correlación es muy baja.
- Si KMO está entre 0.6 y 0.8 el índice se considera aceptable.
- Si KMO es superior a 0.8, el grado de correlación es muy bueno.

$$KMO = \frac{\sum_{i \neq j} \sum_{j=1}^{p} r_{ij}^2}{\sum_{i \neq j} \sum_{j=1}^{p} r_{ij}^2 + \sum_{i \neq j} \sum_{j=1}^{p} rp_{ij}^2}$$

3.2.2. MEDIDA DE LA ADECUACIÓN DE LA MUESTRA MSAj

Esta medida tiene como fin evaluar la correlación entre nuestras variables en función del índice KMO de Kaiser-Meyer-Olkin obtenido. Si nuestro KMO toma valores inferiores a 0.6 se eliminarán todas aquellas variables con un valor de MSA_j inferior a 0.5.

$$MSA_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} rp_{ij}^2}$$

3.2.3. ROTACIÓN VARIMAX

La intención fundamental al realizar una rotación es conseguir una sencilla interpretación de los Factores.

$$V = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \left(\frac{\hat{l}_{i,j}^*}{\hat{h}_i} \right)^4 - \frac{\left(\sum_{i=1}^{p} \left(\frac{\hat{l}_{i,j}^*}{\hat{h}_i} \right)^2 \right)}{p} \right]$$

Este método simplifica las columnas de la matriz de Factores, en el sentido de conseguir que cada Factor rotado tenga unas cargas Factoriales altas sólo con unas pocas variables. Las demás deben tener correlaciones próximas a 0 con el Factor.

3.3. ANÁLISIS CLÚSTER

El Análisis Clúster es una técnica multivariante cuya idea básica consiste en agrupar un conjunto de observaciones en un número dado de clústeres o grupos que sean lo más homogéneos posible dentro de sí mismos y heterogéneos entre sí. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones y la obtención de dichos clústeres depende del criterio o distancia considerados.

3.3.1. CLÚSTER JERÁRQUICO

Este procedimiento busca identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas. Permite trabajar con variables de tipo mixto (cualitativas y cuantitativas), siendo posible analizar las variables

brutas o elegir entre una variedad de transformaciones de estandarización. Se utiliza cuando no se conoce el número de clústeres a priori y cuando el número de objetos no es muy grande.

3.3.2. CLÚSTER NO JERÁRQUICO

Se usan para agrupar objetos, pero no variables, en un conjunto de k clústeres ya predeterminado. No se tiene que especificar una matriz de distancias ni almacenar las iteraciones. Todo esto permite trabajar con un número de datos mayor que en el caso de los métodos jerárquicos. Se parte de un conjunto inicial de clústeres elegidos al azar, que son los representantes de todos ellos; luego se van cambiando de modo iterativo. Se usa habitualmente el método de las k-medias.

4. ANÁLISIS DE VARIABLES UNIVARIANTE

El análisis univariante permite realizar un análisis descriptivo de cada variable del estudio para extraer sus características más destacables, así como recolectar información mediante gráficos. Sin embargo, debido a la gran cantidad de variables, solo se realiza dicho análisis para aquellas que resulten más relevantes.

Se utiliza la media, mediana, mínimo, máximo, desviación típica, el primer cuartil, el tercer cuartil y el histograma para conocer cómo se comporta la variable y ver si existen grandes diferencias entre las secciones. Se realiza un gráfico de caja y bigotes y un gráfico de dispersión de los datos para detectar secciones atípicas.

4.1. EDAD PROMEDIO

La variable *Edad Promedio* toma valores entre 26.69 años y 65.43 años, en media, tabla 4.1. La media y la mediana no son muy distintas, ambas rondan una edad media de 45 años aproximadamente, pero esta diferencia alerta de un sesgo lateral izquierdo, es decir, secciones con una edad media muy baja. También se puede intuir, por la diferencia entre el tercer cuartil y el máximo, secciones atípicas por tener una edad promedio muy alto.

TABLA 4.1

Análisis descriptivo: Edad promedio

Variable	Mínimo	Q1	Mediana	Media	Q3	Máximo	Dev std
Edad promedio	26.69	42.88	45.20	44.87	47.40	65.43	4.14

En la ilustración 3, se puede percibir que hay una mayor proporción de secciones censales en edades comprendidas entre los 40 y 50 años. Sin embargo, no se aprecia a simple vista casos atípicos.

Mediante el gráfico de caja y bigotes de la ilustración 4, se observa 128 secciones atípicas de las 2443 que tiene la base de datos. Dichas secciones censales tienen una edad media inferior a 36.12 años o superior a 54.14 años.

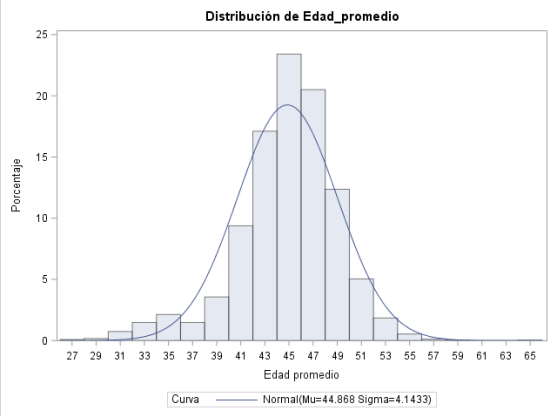

Ilustración 3: Histograma de la variable Edad Promedio.

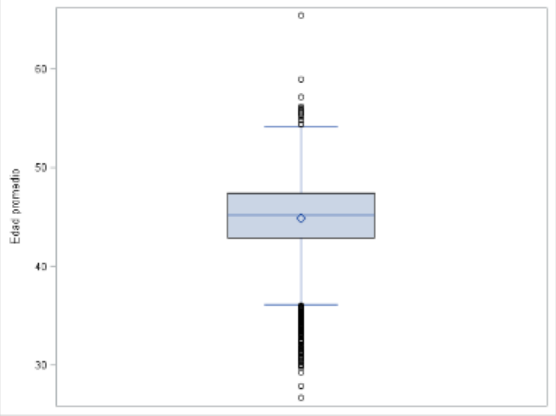

Ilustración 4: Caja y bigotes de la Edad Promedio.

Aplicando la distancia a la media de la variable estandarizada *Edad Promedio*, en la cual se considera atípicos aquellos valores, en valor absoluto, superiores a 3, se obtienen un total de 32 secciones censales con una edad media atípica. En este caso se consideran atípicas aquellas secciones con una edad media inferior a 32.45 años y mayor a 57.16 años.

En la ilustración 5, se observa en color rojo las 32 secciones censales atípicas. De estas 32 secciones, solo dos sobrepasan la edad máxima de 57.16 años en media, estas dos secciones se corresponden a: El Goloso (Fuencarral El Pardo) y Puerta Bonita (Carabanchel).

Las 30 secciones censales restantes corresponden a: Valverde, Mirasierra y El Goloso (Fuencarral El Pardo), Valdefuentes (Hortaleza), Butarque (Villaverde), Casco Histórico de Vallecas y Ensanche de Vallecas (Villa de Vallecas) y Timón (Barajas).

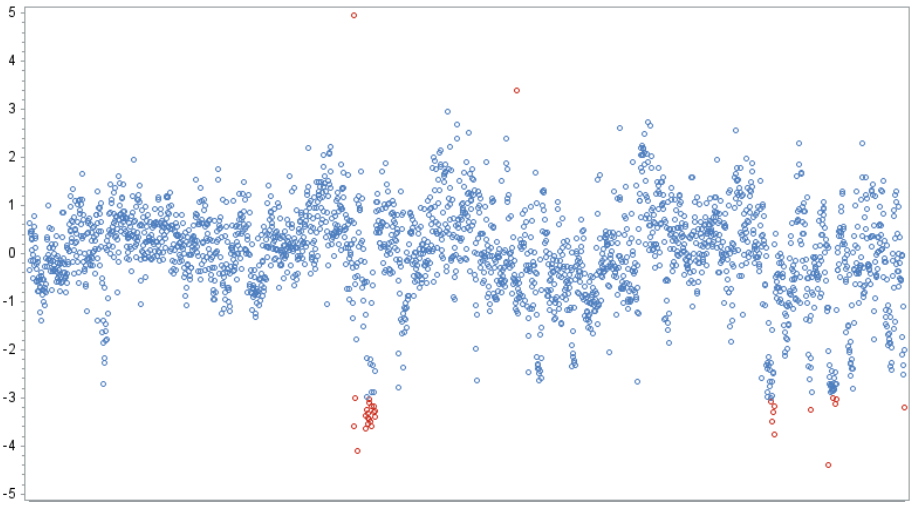

Ilustración 5: Gráfico de dispersión de la variable tipificada Edad Promedio.

4.2. ÍNDICE DE DEPENDENCIA

La variable *Índice de dependencia* toma valores entre 18.30 y 252.1 personas dependientes por cada 100 personas de la población activa. Existe una gran diferencia entre la media y la mediana, lo que indica que existe un sesgo lateral derecho, es decir, hay secciones con una alta dependencia de la tercera edad, equivale a decir que hay un gran número de personas con edad igual o superior a los 65 años. Esto último también se puede comprobar observando el salto existente entre el tercer cuartil y el máximo.

TABLA 4.2
Análisis descriptivo: Índice de dependencia

Variable	Mínimo	Q1	Mediana	Media	Q3	Máximo	Dev std
Índice de dependencia	18.30	44.16	51.82	54.24	62.03	252.10	15.91

En la ilustración 6 se puede observar una mayor frecuencia de secciones entorno a 50 personas dependientes por cada 100 personas de la población activa. También se aprecia una clara asimetría a la derecha lo que hará que se detecten como atípicos valores de esa cola.

Mediante el gráfico de caja y bigotes de la ilustración 7, de las 2443 secciones censales se considera 79 secciones atípicas, todas estas por encima de 88.49105 personas dependientes por cada 100 personas de la población activa.

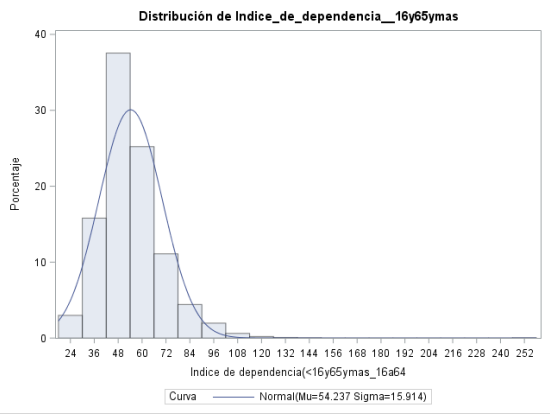

Ilustración 6: Histograma de la variable Índice de Dependencia.

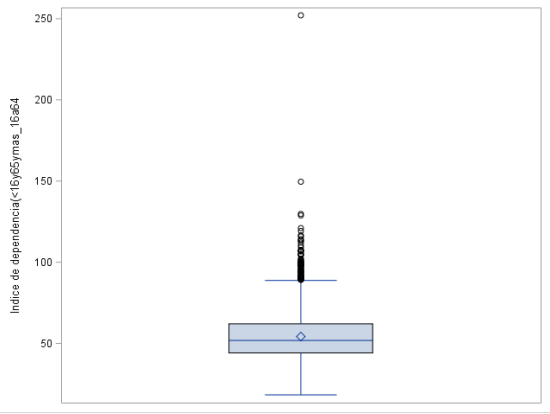

Ilustración 7: Caja y bigotes de la variable Índice de Dependencia.

Aplicando la distancia a la media de la variable estandarizada se obtiene 24 secciones censales atípicas.

En la ilustración 8 se puede observar en color rojo dichas secciones. Estas secciones corresponden a: Adelfas (Retiro), La Paz, Del Pilar y El Goloso (Fuencarral-El Pardo), Aluche y Campamento (Latina), Vista Alegre y Puerta Bonita (Carabanchel), Portazgo (Puente de Vallecas), Marroquina y Media Legua (Moratalaz), Canillas y Pinar

del Rey (Hortaleza) y Los Ángeles (Villaverde). Se debe mencionar que, de todas estas secciones, destaca sobre las demás la 08129 y se podrá ver que será decisiva a la hora de tomar decisiones más adelante.

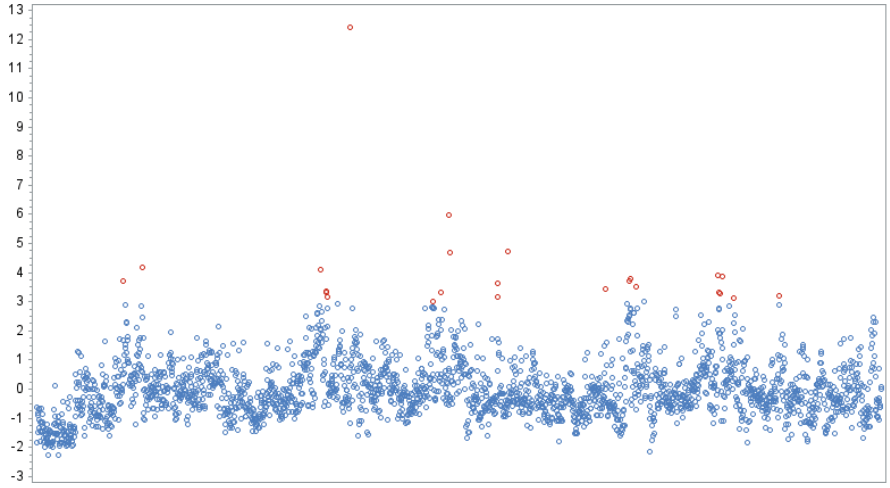

Ilustración 8: Gráfico de dispersión de la variable tipificada Índice de Dependencia.

4.3. PROPORCIÓN DE EXTRANJEROS

La variable *Proporción extranjeros* toma valores entre el 0.83% y el 49.92% de extranjeros. Debido a la diferencia entre la mediana y la media (más de un 1%), se intuye existe un sesgo o asimetría hacia la derecha, es decir, existen secciones con alta proporción de extranjeros. Este sesgo también se puede ver por la diferencia que hay entre el tercer cuartil y el máximo.

TABLA 4.3							
Análisis descriptivo: Proporción de extranjeros							
Variable	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**	**Dev std**
Proporción_de_extranjeros	0.83	8.72	14.09	15.41	20.67	49.92	8.44

En la ilustración 9 se puede apreciar un gran porcentaje de secciones censales comprendidas entre el 5% y 20% de personas extranjeras. Igualmente, se aprecia de forma clara un sesgo lateral derecho, lo que significa que existen valores atípicos.

De las 2443 secciones censales, en la ilustración 10, se considera 19 secciones atípicas; todas aquellas secciones con una proporción de personas extranjeras superior al 38.51% (que corresponden al total de secciones atípicas) y secciones censales que tienen una proporción inferior a 0.83%. Estas 19 secciones censales corresponden a: Embajadores (Centro), Puerta Bonita (Carabanchel), Almendrales, Moscardó y Pradolongo (Usera), San Diego (Puente de Vallecas), San Cristóbal (Villaverde) y Simancas (San Blas).

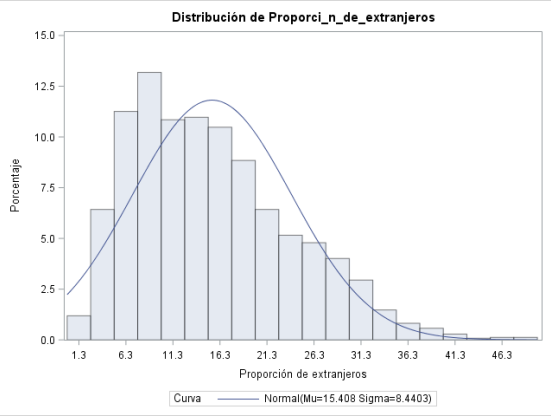

Ilustración 9: Histograma de la variable Proporción de Extranjeros.

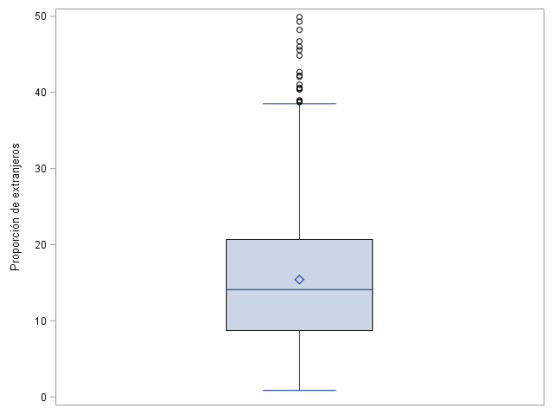

Ilustración 10: Caja y bigotes de la variable Proporción de Extranjeros.

Aplicando la distancia a la media de la variable estandarizada *Proporción Extranjeros*, se tiene 11 secciones censales con una proporción atípica. En este caso se consideran atípicas aquellas secciones con una proporción inferior a 0.83% y mayor a 40.58%.

En la ilustración 11, podemos observar en color rojo las 11 secciones censales atípicas, destacando que todas las secciones consideradas atípicas, en este caso, tienen una proporción muy elevada. Estas secciones censales correspondes a: Embajadores (Centro), Almendrales, Moscardó y Pradolongo (Usera), San Diego (Puente de Vallecas) y Simancas (San Blas).

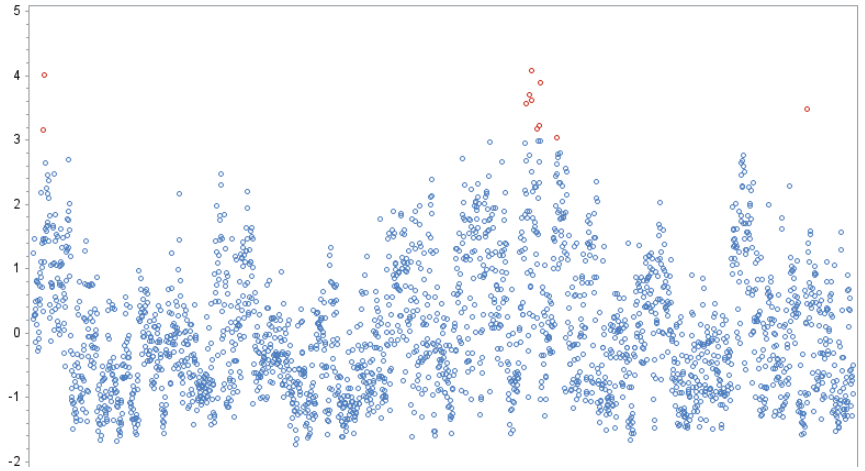

Ilustración 11: Gráfico de dispersión de la variable tipificada Proporción de Extranjeros.

4.4. RENTA MEDIA POR HOGAR

La variable *Renta media por hogar*, toma valores entre 10594 euros y 89215 euros. Como existe una gran diferencia entre la mediana y la media se puede decir que esta variable no está centrada, es decir, existe un sesgo lateral derecho y, por lo tanto, hay secciones con una renta media por hogar muy elevada. Esto mismo se puede apreciar debido a la diferencia que existe entre el tercer cuartil y el máximo.

TABLA 4.4 Análisis descriptivo: Renta media por hogar							
Variable	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**	**Dev std**
Renta media por hogar (€)	10594.00	26285	33592.00	38817.65	46525	89215.00	16681.46

En la ilustración 12 se puede observar una gran frecuencia de secciones censales entre lo 20000 euros y los 40000 euros de renta media por hogar. Tambien se ve claramente una asimetría lateral derecha, significando la existencia de valores atípicos.

En la ilustración 13, se considera 123 secciones atípicas, todas ellas con una renta media por hogar por encima de los 76628 euros.

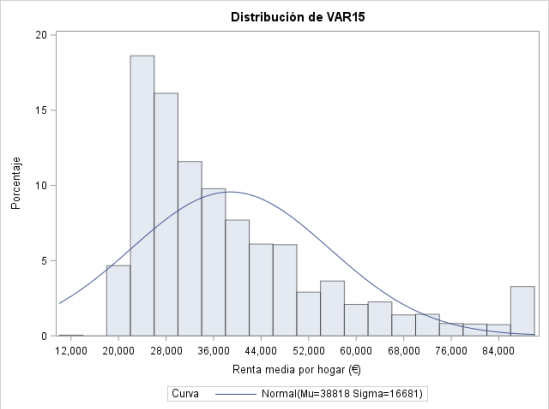

Ilustración 12: Histograma de la variable Renta Media por Hogar.

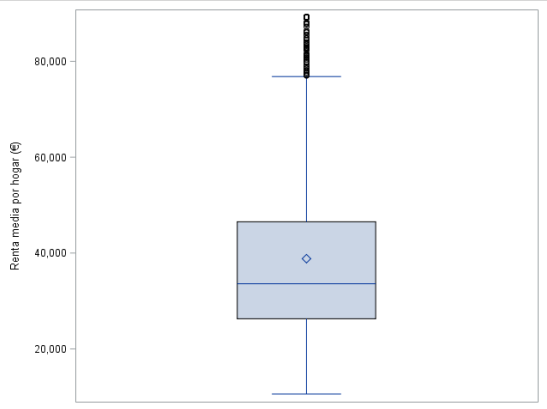

Ilustración 13: Caja y bigotes de la variable Renta Media por Hogar.

Aplicando la distancia a la media de la variable estandarizada *Renta Media Por Hogar*, se obtiene 71 secciones censales atípicas, todas con un valor de 89215 euros (corresponde a la última columna del histograma).

En la ilustración 14, se puede observar en color rojo las 71 secciones censales atípicas. Estas secciones corresponden a los barrios: Los Jerónimos y Niño Jesús (Retiro), Recoletos y Castellana (Salamanca), El Viso, Hispanoamérica y Nueva España (Chamartín), Castillejos (Tetuán), Almagro y Ríos Rosas (Chamberí), Fuentelarreina y Mirasierra (Fuencarral-El Pardo), Argüelles, Ciudad Universitaria, Valdemarín, El Plantío y Aravaca (Moncloa-Aravaca), Paloma, Piovera, Valdefuentes, Los Ángeles y Los Rosales (Hortaleza) y Corralejos (Barajas).

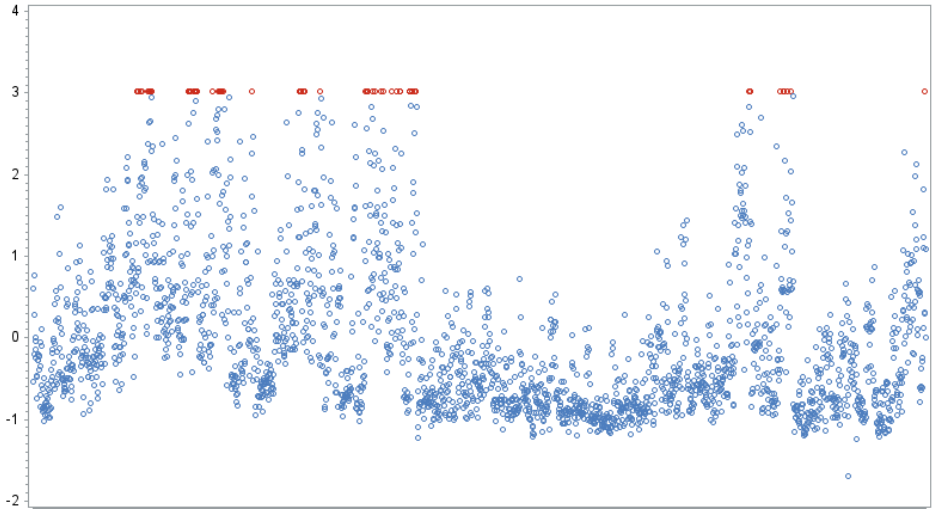

Ilustración 14: Gráfico de dispersión de la variable tipificada Renta Media por Hogar.

5. ANÁLISIS DE RELACIONES ENTRE VARIABLES

Conocidas las variables, se procede a realizar un análisis de relaciones por pares de variables a partir de la matriz de correlaciones para determinar si todas las variables van a ser empleadas en el análisis o se debe prescindir de alguna de ellas. Al tratarse de un estudio sobre toda la población, se utilizará el coeficiente de correlación de Pearson.

En la ilustración 15 y 16 se tiene la correlación de Pearson para cada par de variables. En la matriz triangular inferior tenemos el valor numérico de la correlación y en la matriz triangular superior el grado y signo de la correlación. La X marca los pares de variables que están incorreladas a una significación de 0.05.

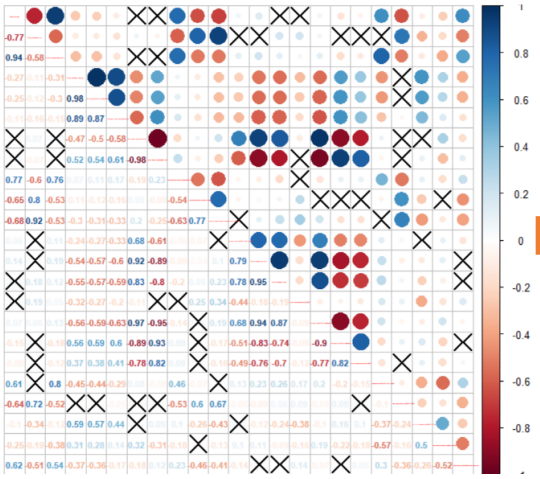

Ilustración 15: Matriz de Correlaciones Bivariadas con las 23 variables iniciales.

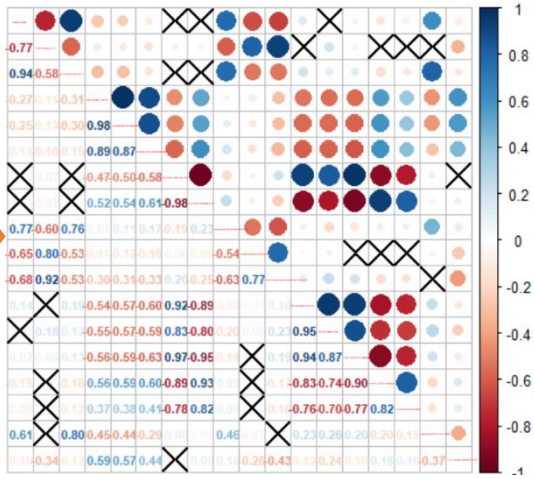

Ilustración 16: Matriz de Correlaciones Bivariadas con las 18 variables finales.

Se observa la existencia de variables como la edad o la proporción de jóvenes que están incorreladas con 4 y 5 variables, las cuales están fuertemente relacionadas con muchas otras. Sin embargo, se destaca la variable Índice de reemplazo población activa, la Proporción de Hogares con más de un Adulto sin Menores, la Proporción de turismo de PF con 16 y más CF, la Proporción de Hogares Unipersonales con edad inferior a 65 años y Proporción de Afiliados con Edad entre 16 y 64 Años que, aunque el número de variables incorreladas es menor, exceptuando la variable Índice de reemplazo población activa que está incorrelada con 5 variables, la mayoría de correlaciones son muy bajas, es decir, todas se encuentran, en su mayoría, por debajo del 50% de correlación, siendo estas cinco variables las que serán excluidas del análisis. Por tanto, el estudio se procederá a realizar solo con un total de 18 variables.

6. REDUCCIÓN DE DIMESIONES

Para realizar una reducción de dimensiones se va a utilizar un Análisis Factorial, en el cual es fundamental que las variables de estudio estén fuertemente correladas, es decir, habrá una gran relación entre ellas. Anteriormente, se observó la correlación por pares de nuestras variables y se pudo ver que las 18 variables finales tienen una alta correlación.

Una vez comprobada la correlación por parejas de variables, se debe determinar la correlación global del modelo mediante el Índice KMO de Kaiser-Meyer-Olkin, en el cual, se obtiene un índice del 82.4%.

Como el valor de este indicador se encuentra por encima del 80%, se determina que los datos son lo suficientemente buenos para realizar la factorización. Además, se comprueba la medida de la adecuación de la muestra MSAj para determinar individualemente si alguna variable no aporta la suficiente información en el modelo, tabla 6.1; en este caso, al tener un KMO superior a 0.6, no sería posible eliminar ninguna variable, aún así, comprobamos que todas ellas se encuentran por encima del 0.5.

TABLA 6.1	
Medida de Kaiser de suficiencia muestral: MSA total= 0.82409046	
Edad promedio	0.70914707
Proporción de juventud	0.63268741
Proporción de envejecimiento	0.75833814
Proporción de extranjeros	0.78780873
Proporción de nacidos fuera de España	0.82136326
Proporción de extranjeros en edad escolar	0.92744387
Proporción estudios superiores	0.83220726
Proporción estudios obligatorios	0.85473756
Proporción de hogares unipersonales de 65 y más	0.76413917
Proporción de hogares monoparentales	0.95017716
Proporción de hogares con más de una persona adulta con menores	0.86413375
Renta media por persona (€)	0.86108775
Renta media por hogar (€)	0.82522372
Proporción grupos 1 y 2 s/total afiliados	0.91475495
Proporción grupo 10 s/total afiliados	0.92763851
Proporción de parados s/pob16a64	0.95253336
Índice de dependencia (<16y65ymas_16a64	0.66145570
Índice estructura pob activa 16a39_40_64	0.60627927

A continuación, se debe determinar el número de factores en los cuales recogeremos la mayor información posible de las 18 variables. Para ello, se utilizará los autovalores de la matriz de correlaciones, tabla 6.2.

TABLA 6.2

Autovalores de la matriz de correlación: Total = 18 Promedio = 1

	Autovalor	Diferencia	Proporción	Acumulada
1	7.83588608	3.11072053	0.4353	0.4353
2	4.72516555	2.18739713	0.2625	0.6978
3	2.53776841	1.59348574	0.1410	0.8388
4	0.94428267	0.43458428	0.0525	0.8913
5	0.50969839	0.07699557	0.0283	0.9196
6	0.43270282	0.14984628	0.0240	0.9436
7	0.28285654	0.04024998	0.0157	0.9594
8	0.24260656	0.05550918	0.0135	0.9728
9	0.18709739	0.07641134	0.0104	0.9832
10	0.11068604	0.04136384	0.0061	0.9894
11	0.06932221	0.02589556	0.0039	0.9932
12	0.04342665	0.01600145	0.0024	0.9956
13	0.02742520	0.01081606	0.0015	0.9972
14	0.01660915	0.00382423	0.0009	0.9981
15	0.01278492	0.00427877	0.0007	0.9988
16	0.00850615	0.00106766	0.0005	0.9993
17	0.00743850	0.00170172	0.0004	0.9997
18	0.00573677		0.0003	1.0000

Mediante el método Kaiser, se estipula que se deben seleccionar tantos factores como autovalores superiores a la unidad, lo que significa reducir las variables a 4 factores (se incluye la componente 4 ya que, aunque no es superior a 1, se encuentra muy próxima a la unidad), ajustandose así a las variables en un 89.13%. Tambien se puede utilizar el siguiente gráfico de sedimentación, ilustración 17, para determinar el número de factores, al igual que en el caso anterior, se decide que el mejor número de factores es 4.

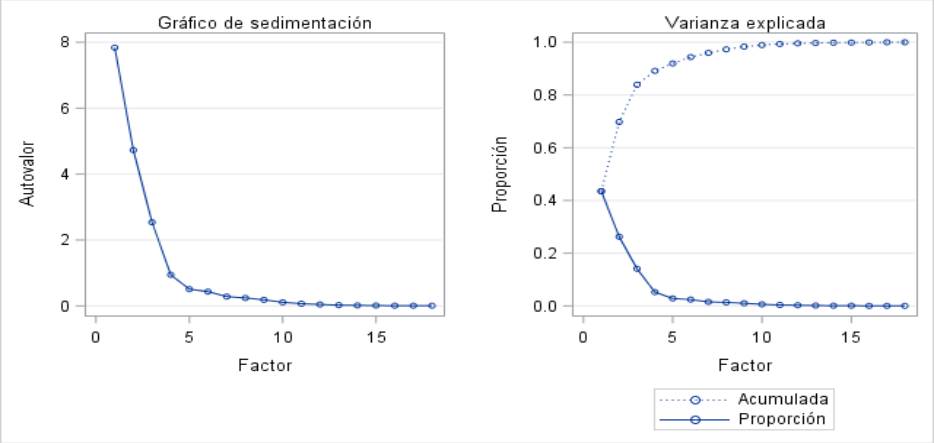

Ilustración 17: Gráfico de Sedimentación y Gráfico de Variabilidad Explicada por cada Factor.

Determinados el número de factores, se procede a obtener las cargas de cada variable para cada factor, tabla 6.3, para su posterior interpretación.

TABLA 6.3				
Modelo factorial				
	Factor 1	Factor 2	Factor 3	Factor 4
Edad promedio	-0.09948	0.96228	-0.19614	-0.00782
Proporción de juventud	-0.15700	-0.85220	-0.31829	0.31422
Proporción de envejecimiento	-0.17252	0.90059	-0.30060	0.21547
Proporción de extranjeros	0.76712	-0.09829	0.51928	0.31143
Proporción de nacidos fuera de España	0.78395	-0.07970	0.48776	0.30223
Proporción de extranjeros en edad escolar	0.79373	0.02194	0.33800	0.34845
Proporción estudios superiores	-0.90007	-0.03536	0.38477	0.05181
Proporción estudios obligatorios	0.91536	0.06249	-0.34177	-0.00190
Proporción de hogares unipersonales de 65 y más	0.18103	0.82929	-0.09503	0.25718
Proporción de hogares monoparentales	-0.13023	-0.78422	-0.28837	0.25667
Proporción de hogares con más de una persona adulta con menores	-0.32979	-0.81076	-0.35184	0.19174
Renta media por persona (€)	-0.91341	0.11292	0.26484	0.08176
Renta media por hogar (€)	-0.88555	-0.01537	0.14484	0.12473
Proporción grupos 1 y 2 s/total afiliados	-0.93628	0.03741	0.27222	0.06039
Proporción grupo 10 s/total afiliados	0.89901	-0.09390	-0.23382	-0.03069
Proporción de parados s/pob16a64	0.77538	-0.05556	-0.33197	-0.20448
Índice de dependencia (<16y65ymas_16a64	-0.32317	0.49091	-0.58647	0.49596
Índice estructura pob activa 16a39_40_64	0.35069	0.11353	0.72866	0.05671

Con estas cargas se determina que el Factor 1 hace referencia a aquella población extranjera con una renta baja, el Factor 2 a la población de más de 65 años, el Factor 3 y el Factor 4 no se puede apreciar con claridad lo que representa debido a que ninguna carga es relevante. Para solventar este problema, se utiliza una rotación Varimax, tabla 6.4, para conseguir una interpretación más sencilla y clara de los Factores.

TABLA 6.4

Modelo factorial de rotación

	Factor 1	Factor 2	Factor 3	Factor 4
Edad promedio	0.02729	-0.75264	-0.21878	0.59945
Proporción de juventud	0.01223	0.96925	-0.08928	-0.05812
Proporción de envejecimiento	0.07783	-0.56785	-0.19927	0.78064
Proporción de extranjeros	-0.37606	-0.02946	0.89398	-0.15267
Proporción de nacidos fuera de España	-0.40623	-0.04157	0.87459	-0.13609
Proporción de extranjeros en edad escolar	-0.47293	-0.06518	0.79866	0.01953
Proporción estudios superiores	0.96809	0.03829	-0.13175	-0.07804
Proporción estudios obligatorios	-0.95204	-0.05529	0.19512	0.10539
Proporción de hogares unipersonales de 65 y más	-0.12927	-0.59145	0.13419	0.64120
Proporción de hogares monoparentales	-0.00450	0.87547	-0.09029	-0.07981
Proporción de hogares con más de una persona adulta con menores	0.12750	0.91274	-0.26308	-0.08997
Renta media por persona (€)	0.93418	-0.03810	-0.20877	0.07858
Renta media por hogar (€)	0.85708	0.12177	-0.24967	0.09603
Proporción grupos 1 y 2 s/total afiliados	0.95125	0.01770	-0.22374	0.02285
Proporción grupo 10 s/total afiliados	-0.89798	0.03545	0.25081	-0.04611
Proporción de parados s/pob16a64	-0.86419	-0.02594	0.02769	-0.08970
Índice de dependencia (<16y65ymas_16a64)	0.10920	0.00014	-0.29718	0.91385
Índice estructura pob activa 16a39_40_64	0.04720	-0.33113	0.69254	-0.28027

La interpretación de los 4 factores es la siguiente:

- Factor 1: Renta alta, profesiones muy cualificadas y poco paro.
- Factor 2: Familias jóvenes con hijos.
- Factor 3: Población extranjera y no nacida en España.
- Factor 4: Población con edad igual o superior a los 65 años.

En la ilustración 18 se puede observar la relación entre cada una de las variables y cada uno de los Factores del modelo factorial rotado, indicando la carga o variabilidad explicada por dicho factor, así como la variabilidad que no queda explicada por los factores comunes a cada variable.

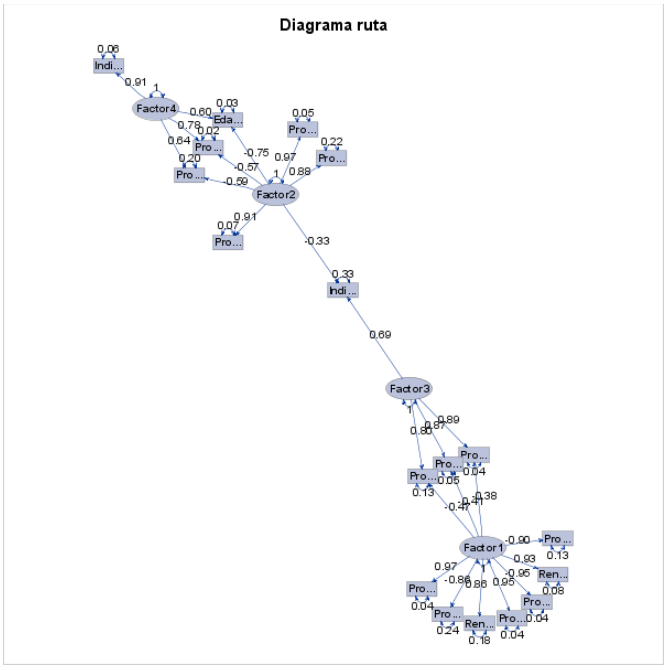

Ilustración 18: Diagrama de Ruta.

Antes de pasar al análisis de conglomerados, se estudiará el ajuste de las secciones a los factores creados anteriormente. Para ello, en las siguientes imágenes, se representa gráficamente las puntuaciones de las 2443 secciones para cada par de factores e individualmente. Se puede observar la existencia de tres secciones censales, 08129, 08130 y 18045, que presentan un coeficiente de puntuación estandarizado muy distinto del resto, significando que estas secciones no se ajustan bien a dichos factores. Esto último se podrá ver con más detalle a la hora de realizar el Análisis Clúster. Estas 3 secciones corresponden a los 3 puntos rojos de la ilustración 19, que coinciden con los datos atípicos inferiores de los gráficos de caja y bigotes del Factor 1 y Factor 2, y al dato atípico superior del Factor 4 de la ilustración 20.

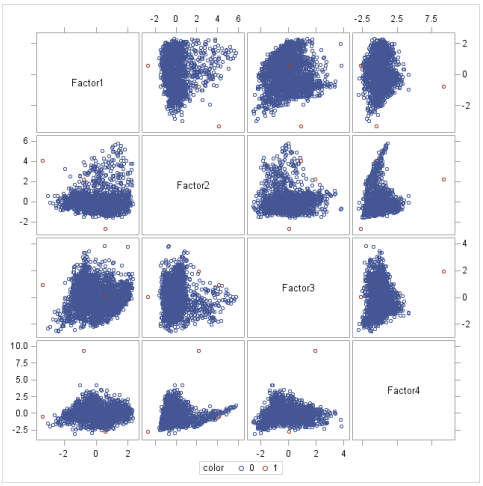

Ilustración 19: Gráfico de dispersión para cada par de factores.

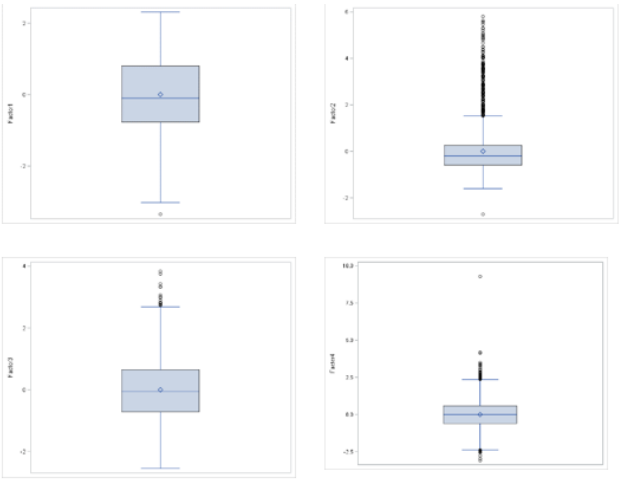

Ilustración 20: Gráfico de Caja y Bigotes para cada factor.

7. CLASIFICACIÓN DE LAS SECCIONES CENSALES

Determinados e interpretados los factores en los cuales se recogió la información de las variables iniciales, se procedió a agrupar todas las secciones mediante el Análisis Clúster, de forma que cada grupo sea lo más parecido dentro de ellos y lo más diferente posible con el resto de los grupos. Al no tener un número de clústeres a priori, se debe empezar utilizando un Análisis Clúster Jerárquico para determinar el número de grupos a realizar. Mediante la tabla de historia del conglomerado, tabla 7.1, se estudia los estadísticos Pseudo F y Pseudo T^2.

TABLA 7.1

Historia de conglomerado

Número de clústeres	Conglomerados unidos		Frec	R-cuadrado semiparcial	R-cuadrado	Estadístico pseudo F	T-cuadrado pseudo	Distancia RMS	Igualdad de rango
2442	13036	13079	2	0.0000	1.00	3907	.	0.0453	
2441	01030	06110	2	0.0000	1.00	3627	.	0.0487	
2440	10114	10164	2	0.0000	1.00	3046	.	0.0589	
2439	10143	10154	2	0.0000	1.00	2816	.	0.0591	
2438	08048	10116	2	0.0000	1.00	2617	.	0.0627	
2437	17049	20020	2	0.0000	1.00	2454	.	0.0655	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
12	CL19	CL23	71	0.0067	.554	275	54.1	2.5174	
11	CL16	08130	74	0.0006	.553	301	6.5	2.5359	
10	CL15	CL14	1875	0.1557	.398	179	835	2.6307	
9	CL12	CL21	108	0.0097	.388	193	54.0	2.6365	
8	CL13	CL17	376	0.0360	.352	189	190	2.8076	
7	CL11	CL28	76	0.0014	.351	219	14.6	2.9422	
6	CL10	CL8	2251	0.1334	.217	135	490	3.0931	
5	CL6	CL7	2327	0.0468	.171	125	144	3.2456	
4	CL9	CL41	110	0.0028	.168	164	10.7	4.1765	
3	CL5	CL36	2332	0.0075	.160	233	21.9	4.334	
2	CL3	CL4	2442	0.1504	.010	24.1	437	4.4921	
1	CL2	08129	2443	0.0098	.000	.	24.1	9.9738	

En el estadístico Pseudo F, se observa un mínimo relativo en 5 clústeres; con el estadístico Pseudo T^2 se tiene un máximo relativo en 6 clústeres. Analizando el R^2 semiparcial, se observa que con 6 clústeres hay una gran diferencia de variabilidad parcial explicada. Se concluye que el número óptimo de grupos a realizar es 6. En la ilustración 21 se tiene la evolución de los estadísticos Pseudo F, Pseudo T^2 y CCC para determinar de manera visual el número de clústeres, y en la ilustración 22 se observa el historial de conglomerados de la tabla 7.1.

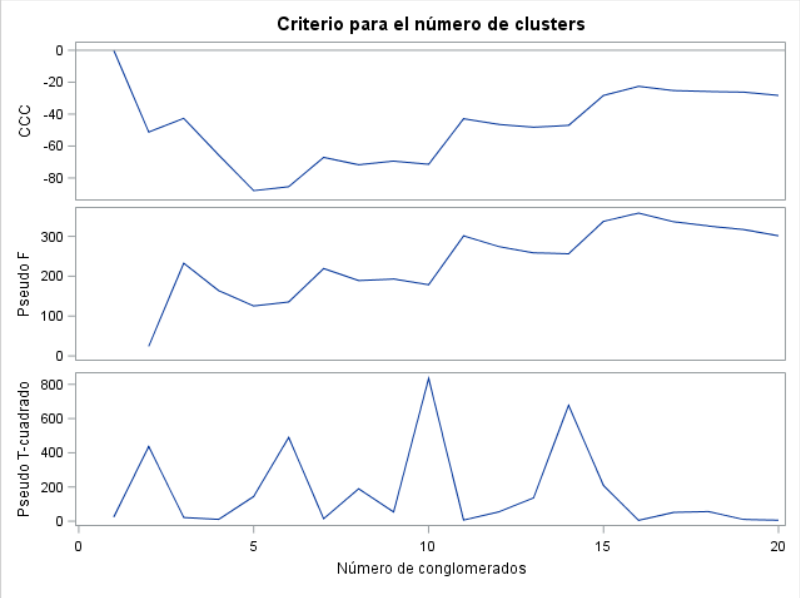

Ilustración 21: Gráfico Pseudo F, Pseudo T^2 y CCC con 2443 secciones censales.

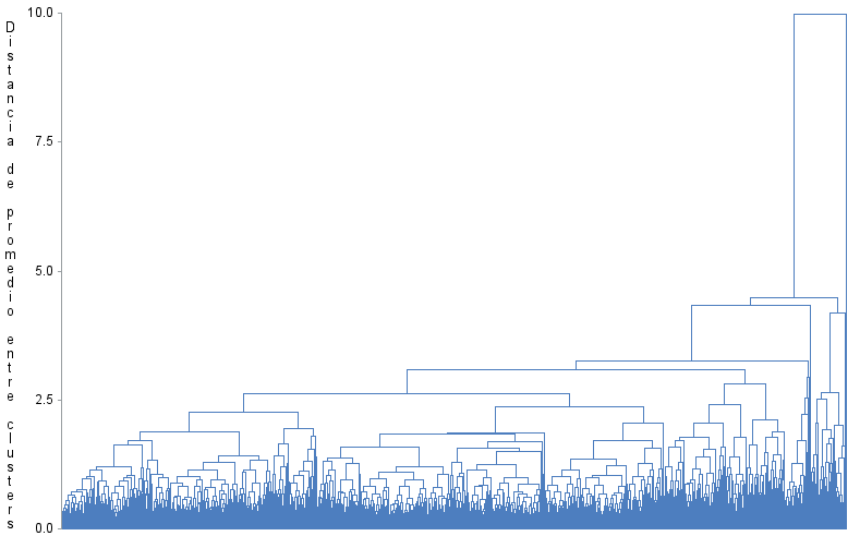

Ilustración 22: Dendograma de todas las Secciones Censales.

En la tabla 7.2 se tiene el número de secciones censales para cada clúster, así como la desviación estándar de la raíz media cuadrática o la distancia entre centroides del clúster.

TABLA 7.2

Resumen de conglomerados

Clúster	Frecuencia	Desviación estándar RMS	Distancia máxima de la semilla a la observación	Radio sobrepasado	Conglomerado más próximo	Distancia entre Centroides del clúster
1	1	.	0		5	8.8739
2	690	0.5826	3.8544		6	2.0556
3	401	0.7296	3.0236		2	2.2352
4	131	0.8462	4.5451		3	3.6959
5	582	0.6334	3.2921		6	1.9643
6	638	0.6415	3.2234		5	1.9643

Cabe señalar que el clúster 1 está compuesto por solo una sección, lo cual no es lo ideal a la hora de querer realizar grupos lo más homogéneos entre ellos, sin embargo, esta sección corresponde a la 08129 (barrio El Goloso, distrito Fuencarral-El Pardo), como se pudo ver tanto en el Análisis Univariante como en el Análisis Factorial, que era una sección atípica. A parte de esta sección, al final del Análisis Factorial, se determina que esta sección y dos más no se ajustaban bien a los factores, por esta razón, se procede a excluir estas 3 secciones censales del Análisis de Conglomerados, dejándolas en un clúster aparte que no aplica a ningún otro grupo y, por tanto, debe ser tratado de forma distinta.

Se realiza nuevamente el análisis obteniendo la tabla 7.3, en la cual, al analizar el estadístico Pesudo F y Pseudo T^2 se tiene un máximo relativo en 6 clústeres y en 5 clústeres, respectivamente. Al analizar el R^2 semiparcial se observa que en 5 clústeres hay una gran diferencia de variabilidad parcial explicada. Se concluye que se debe realizar 5 grupos.

TABLA 7.3									
Historia de conglomerado									
Número de clústeres	Conglomerados unidos		Frec	R-cuadrado semiparcial	R-cuadrado	Estadístico pseudo F	T-cuadrado pseudo	Distancia RMS	Igualdad de rango
2439	13036	13079	2	0.0000	1.00	3856	.	0.0453	
2438	01030	06110	2	0.0000	1.00	3579	.	0.0487	
2437	10114	10164	2	0.0000	1.00	3006	.	0.0589	
2436	10143	10154	2	0.0000	1.00	2779	.	0.0591	
2435	08048	10116	2	0.0000	1.00	2582	.	0.0627	
2434	17049	20020	2	0.0000	1.00	2422	.	0.0655	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
12	CL24	CL18	1081	0.0835	.572	296	678	2.3646	
11	CL16	CL28	256	0.0180	.554	302	136	2.4143	
10	CL17	CL21	71	0.0068	.548	327	54.1	2.5174	
9	CL13	CL12	1875	0.1579	.390	194	835	2.6307	
8	CL10	CL19	108	0.0098	.380	213	54.0	2.6365	
7	CL11	CL15	376	0.0365	.343	212	190	2.8076	
6	CL14	CL26	75	0.0014	.342	253	15.4	2.9136	
5	CL9	CL7	2251	0.1353	.207	159	490	3.0931	
4	CL5	CL6	2326	0.0466	.160	155	142	3.2297	
3	CL8	19001	109	0.0010	.159	231	3.8	3.5501	
2	CL4	CL34	2331	0.0076	.151	435	21.9	4.3324	
1	CL2	CL3	2440	0.1515	.000	.	435	4.4769	

Al igual que en el caso anterior, en la ilustración 23 se tiene la evolución de los estadísticos Pseudo F, Pseudo T^2 y CCC para determinar de manera visual el número de clústeres, y en la ilustración 24 el historial de conglomerados de la tabla 7.3.

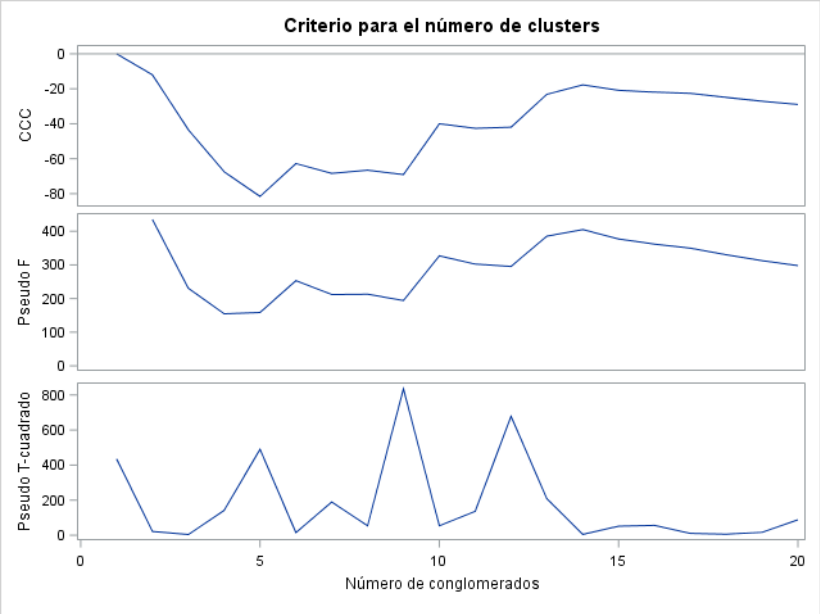

Ilustración 23: **Gráfico Pseudo F, Pseudo T^2 y CCC con 2438 secciones censales.**

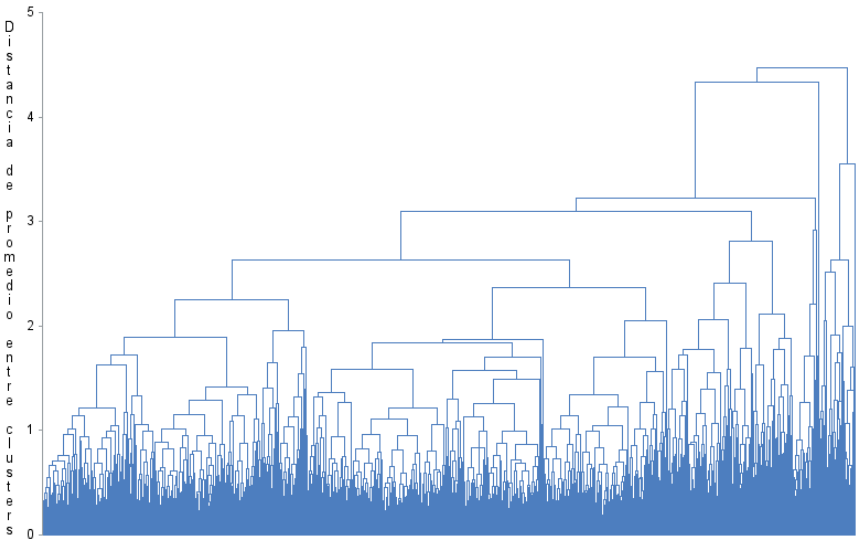

Ilustración 24: **Dendograma sin las Secciones Censales atípicas.**

Determinado el número de clústeres, se realiza un Análisis Clúster no Jerárquico en el cual se obtiene los datos de la tabla 7.4 (frecuencia de secciones en cada clúster) y table 7.5 (media de cada clúster).

TABLA 7.4						
Resumen de conglomerados						
Clúster	**Frecuencia**	**Desviación estándar RMS**	**Distancia máxima de la semilla a la observación**	**Radio sobrepasado**	**Conglomerado más próximo**	**Distancia entre Centroides del clúster**
1	327	0.5647	2.5184		5	2.0209
2	113	0.7872	3.0161		5	3.7817
3	324	0.7169	2.9130		5	2.2097
4	588	0.5732	3.8828		5	1.9870
5	1088	0.7161	3.3768		4	1.9870

TABLA 7.5				
Medias del clúster				
Clúster	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	-0.194972173	-0.311373291	-0.474074409	1.579488334
2	0.798549213	3.433493131	-0.448161221	-0.301620053
3	-1.133955005	-0.043480286	-1.150376916	-0.784430815
4	-0.715213781	0.142748420	1.186787027	-0.038736364
5	0.703142236	-0.330495169	-0.112484389	-0.194352595

En base a los datos de la tabla 7.5, se determina que la interpretación de estos grupos es la siguiente:

- El clúster 1 agrupa aquellas secciones con una población con edad igual o superior a los 65 años.
- El clúster 2 está compuesto por aquellas secciones censales donde viven familias españolas con menores y con una renta muy alta.
- El clúster 3 relaciona aquellas secciones con una población mayoritariamente joven y española con una renta baja.
- El clúster 4 agrupa aquellas secciones censales donde hay una mayor cantidad de extranjeros, con y sin menores, con una renta muy baja.
- El clúster 5 representa aquellas secciones con una población mayoritariamente joven y española sin menores con una renta alta.
- En el caso concreto de este análisis, se tiene un clúster 6 compuesto por las 3 secciones que no encajaron en ninguno de los 5 clústeres anteriores. Este grupo se calificará como ‘Otros’ y debe ser tratado de forma completamente distinta.

En la ilustración 25 se muestra gráficamente el resultado de la estratificación sobre el mapa de la ciudad de Madrid, de forma que cada grupo se representa con un color diferente. A simple vista no se observa una homogeneidad de las secciones, es decir, ni los grupos se encuentran en una zona específica y delimitada, ni todas las secciones de un grupo se encuentran unidas o cercanas.

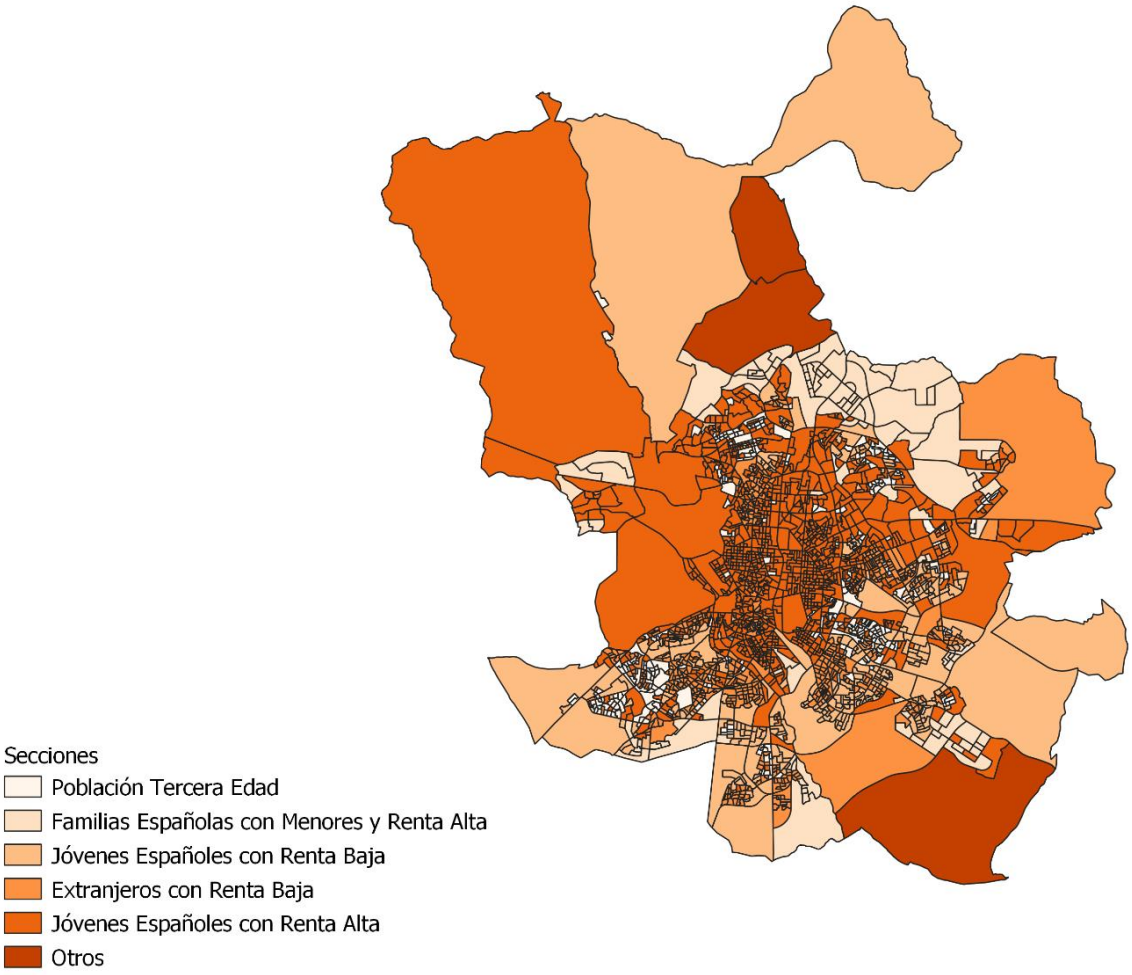

Ilustración 25: Mapa de la ciudad de Madrid, dividida en secciones censales, con su color correspondiente según al grupo que pertenezca.

A continuación, se procede a analizar individualmente cada clúster de forma gráfica para ver mucho más claro el comportamiento de estos grupos entre los distintos distritos que conforman la ciudad de Madrid.

En la ilustración 26 se observa que hay una mayor concentración de secciones censales con una población con edad de 65 años o más en la zona mediocentro de la ciudad, más concretamente en los distritos más envejecidos como: Latina, Carabanchel, Fuencarral-El Pardo, Moratalaz y Hortaleza.

El clúster 2, ilustración 27, que representa aquellas secciones censales con familias de nacionalidad española con menores y una renta muy alta, se encuentra principalmente en la zona norte de la ciudad (Fuencarral-El Pardo, Hortaleza, Barajas y Moncloa-Aravaca). También se encuentra zonas más concretas como Villa de Vallecas o Carabanchel que se incluye en este grupo.

Ilustración 26: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 1.

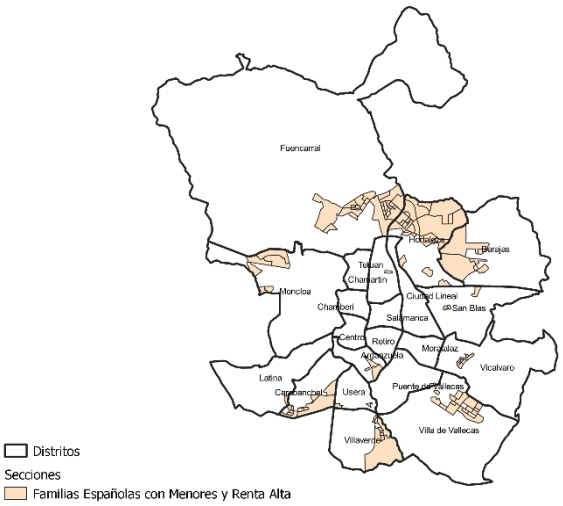

Ilustración 27: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 2.

Al estudiar en mayor profundidad las secciones pertenecientes al grupo 2, se observa que todos los desarrollos urbanísticos de la ciudad de Madrid se encuentran dentro de este clúster.

A continuación, en la ilustración 28 y 29, se muestran los mapas de los desarrollos urbanísticos de Ensanche de Vallecas y Carabanchel, con las respectivas secciones que la conforman, comprobando así que coinciden con las secciones correspondientes vistas en la ilustración 27.

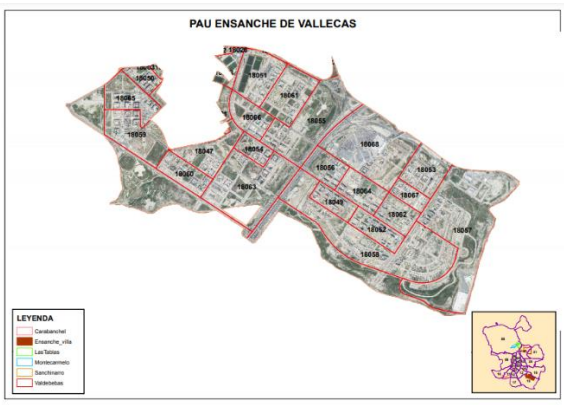

Ilustración 28: Mapa Proyecto de Actuación Urbanística de Ensanche de Vallecas.

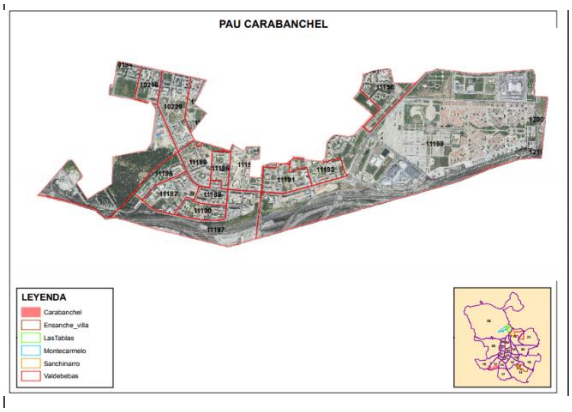

Ilustración 29: Mapa Proyecto de Actuación Urbanística de Carabanchel.

En la ilustración 30 se observa que hay una mayor concentración de secciones censales con jóvenes españoles con una renta baja en los distritos de: Fuencarral-El Pardo, Vicálvaro, Villaverde, Latina y Puente de Vallecas.

En la ilustración 31 se aprecia que la población extranjera con renta baja se concentra mayormente en el sur y a las afueras de la ciudad como los son: Villa de Vallecas, Villaverde, Usera, Carabanchel, Puente de Vallecas y Barajas. También se puede ver una gran cantidad de extranjeros en los distritos de Tetuán y Ciudad Lineal, distritos más céntricos en comparación al resto.

Ilustración 30: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 3.

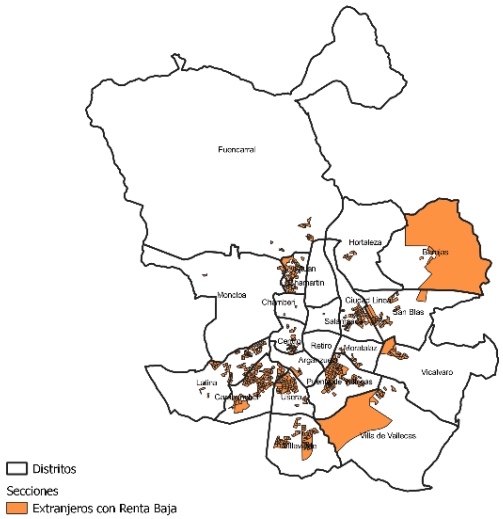

Ilustración 31: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 4.

En la ilustración 32 destaca que, en gran parte de la ciudad de Madrid, particularmente en la zona más céntrica, habitan adultos, sin menores, con una renta muy alta. Esto tiene mucho sentido porque se encuentran en los distritos más lujosos de la ciudad como: Fuencarral-El Pardo, Moncloa-Aravaca, Chamberí, Retiro, Chamartín, Centro y Salamanca.

En este último mapa, ilustración 33, se observa las tres secciones atípicas no incluidas en el análisis de conglomerados. Dos de estas secciones se encuentran en Fuencarral-El Pardo y la tercera al sur de Villa de Vallecas. La sección 08129, barrio El Goloso, es considerada atípica porque presenta una cantidad enorme de personas de la tercera edad debido a la existencia de una residencia de ancianos. La sección 08130, que también forma parte del barrio El Goloso, se encuentra en este grupo por ser una zona meramente adulta, debido a que en esta zona se encuentra la base militar Colonia Militar El Goloso. Por último, la sección 18045, que pertenece al barrio Casco Histórico de Vallecas, es una zona muy atípica por ser una de las zonas más pobres de la ciudad, con habitantes que tienen estudios muy básicos, y, por consiguiente, una renta muy baja.

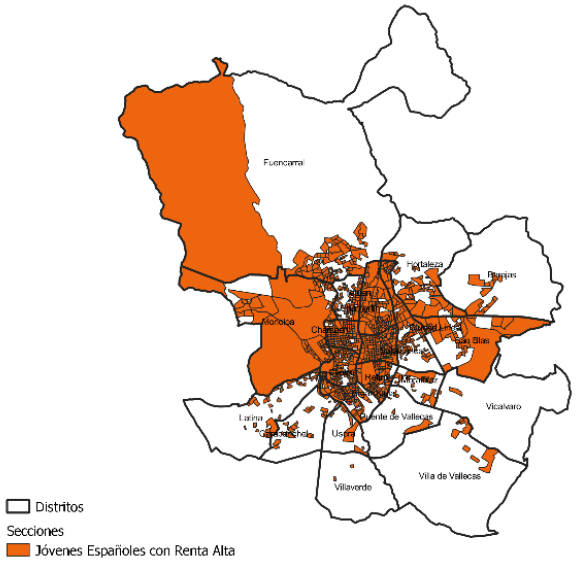

Ilustración 32: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 5.

Ilustración 33: Mapa de la ciudad de Madrid, dividida en distritos, con las secciones censales pertenecientes al grupo 6.

8. CONCLUSIÓN

El propósito de este estudio ha sido segmentar las secciones censales de la ciudad de Madrid en función de datos sociodemográficos para su posterior utilización en diseños muestrales multietápicos. Para ello se utilizó secciones censales, en vez de barrios o distritos, y así tener información de la ciudad de Madrid lo más desagregada posible y, por consiguiente, perder la menor cantidad de precisión ya que, incluso dentro de un mismo barrio, se pueden tener grandes diferencias tanto de renta como de otros factores.

Con este fin, se aplicó un Análisis de Correlaciones Bivariadas para determinar qué tan relacionadas están las 23 variables entre ellas, decidiéndose que se debe descartar 5 variables del modelo. Igualmente, se realizó una reducción de las variables mediante un Análisis Factorial, pasando de tener 18 variables a 4 variables ficticias que recogen la mayor parte de la información de las variables originales: el Factor 1 hace referencia a la renta, el Factor 2 si son familias o no, el Factor 3 la nacionalidad y el Factor 4 si son personas de la tercera edad o no.

En este sentido, se agrupan las secciones en grupos lo más parecido posible utilizando el Análisis Clúster, donde se determina que el mejor número de grupos a formar es 6 (5 más el clúster con las secciones atípicas):

- El clúster 1 está formado por aquellas secciones censales que tienen una gran concentración de personas con una edad igual o superior a los 65 años. Principalmente se encuentran en la zona mediocentro de la ciudad, más concretamente en los distritos de la Latina, Carabanchel, Fuencarral-El Pardo, Moratalaz y Hortaleza.
- El clúster 2 que está compuesto por aquellas secciones formadas por familias de nacionalidad española, que tienen hijos y poseen una renta alta, se encuentran en su mayoría en zonas de desarrollo urbanístico: Montecarmelo, Ensanche de Vallecas o Sanchinarro.
- El clúster 3 lo forman jóvenes adultos de nacionalidad española, que no tienen hijos y poseen una renta muy baja. Se concentran, en su mayoría, en la mitad sur de la ciudad, en distritos como Vicálvaro, Barajas, Carabanchel, Villaverde y Latina.
- El clúster 4 son aquellas secciones censales con una gran cantidad de extranjeros que tienen una renta baja. Similar al clúster 3, estas zonas se encuentran en la mitad sur de la ciudad, así como a las afueras: Villa de Vallecas, Villaverde, Usera, Carabanchel, Puente de Vallecas y Barajas.
- El clúster 5 agrupa aquellas secciones en las cuales sus habitantes son jóvenes adultos de nacionalidad española, sin hijos y con una renta muy alta. Estas secciones pertenecen a las zonas más céntricas y ricas de la ciudad como los son los distritos de Fuencarral-El Pardo, Moncloa-Aravaca, Chamberí, Retiro, Chamartín, Centro, Salamanca, entre otras.

- El clúster 6 está formado por tres secciones censales (08129, 08130 y 18045) que son muy diferentes a las demás. Dos de ellas se encuentra en el barrio El Goloso por tener una gran cantidad de personas de la tercera edad por una residencia de ancianos y por tener una base militar. La otra sección, que se encuentra en el Casco Histórico de Vallecas, es por tener una gran cantidad de infraviviendas.

Al estudiar el comportamiento de estos clústeres y realizar un mapa de la ciudad de Madrid con la representación de cada grupo, se determina que no existe una clara homogeneidad entre barrios ni distritos, es decir, que un grupo no se encuentra en una zona concreta de la ciudad, sino más bien que se encuentran distribuidos por todo Madrid, aunque en algunos casos se puede ver cierta tendencia.

Como corolario a este análisis, los resultados de este estudio pueden ser utilizados con otros fines además del inicial, tanto por entidades públicas como privadas, como por ejemplo en estudios de mercado que permitan vislumbrar la conveniencia de colocar determinadas empresas en función de la renta de los habitantes, colocar colegios en función de si es una zona familiar o dónde construir una nueva residencia de ancianos en función de la concentración de personas de la tercera edad.

9. BIBLIOGRAFÍA

- Ayuntamiento de Madrid, Estadística: https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica?vgnextchannel=8156e39873674210VgnVCM1000000b205a0aRCRD
- Baró Llians: "Estadística descriptiva". Ed. Parramón
- Cabrero Ortega, M. and García Pérez, A., 2020. Análisis estadístico de datos espaciales con QGIS y R. 1st ed. Madrid: Universidad Nacional de Educación a Distancia.
- Junquera González, J. (1997). Resolución de 9 de abril de 1997, de la Subsecretaría, por la que se dispone la publicación, de la Resolución de 1 de abril, de la Presidenta del Instituto Nacional de Estadística y del Director general de Cooperación Territorial, por la que se dictan instrucciones técnicas a los Ayuntamientos sobre la gestión y revisión del padrón municipal. From: https://www.boe.es/eli/es/res/1997/04/09/(4)/dof/spa/pdf
- Valencia Delfa, J., & Vicente Hernanz, M. (2015). Análisis multivariante I. Madrid: Cersa.
- Valencia Delfa, J. (2005). (García Pérez & Cabrero Ortega, 2020). [Madrid]: Compañía Española de Reprografía y Servicios.

10. ANEXO I: TABLAS

TABLA 1.1 División Administrativa Vigente desde 1 de noviembre de 2017		
Distrito/Barrio	**Secciones censales**	**Total de Secciones en el Distrito**
TOTAL SECCIONES		**2.443**
01. CENTRO		**111**
01.1 Palacio	De 1 a 4, 6 a 9, 11 a 16 y 18 a 21	
01.2 Embajadores	De 22 a 43, 45 a 51, 53 a 59	
01.3 Cortes	De 61 a 69, 71 a 72	
01.4 Justicia	De 73 a 77, 79, 81 a 82, 84 y 87 a 89	
01.5 Universidad	De 90 a 116	
01.6 Sol	De 117 a 121, y 123 a 124	
02. ARGANZUELA		**109**
02.1 Imperial	De 1 a 11, 13 a 14, 89 a 90, 95 y 107	
02.2 Las Acacias	De 15 a 31, 85 a 88, 91 a 92, 94, 96 y 108 a 109	
02.3 La Chopera	De 32 a 47	
02.4 Legazpi	De 48 a 49, 93, 99 a 104 y 110 a 111	
02.5 Las Delicias	De 50 a 63, 97 a 98, 105 a 106 y 112	
02.6 Palos de Moguer	De 64 a 75 y 77 a 82	
02.7 Atocha	83	
03. RETIRO		**94**
03.1 Pacífico	De 1 a 23, 89 y 95 a 96	
03.2 Adelfas	De 24 a 30, 92 y 97 a 100	
03.3 La Estrella	De 31 a 42, 44 a 47, 90 a 91 y 93 a 94	
03.4 Ibiza	De 48 a 61 y 63 a 67	
03.5 Los Jerónimos	De 69 a 71, 73 a 74 y 76	
03.6 Niño Jesús	De 77 a 82 y 84 a 88	
04. SALAMANCA		**126**
04.1 Recoletos	De 1 a 14	
04.2 Goya	De 15 a 21, 23 a 42	
04.3 Fuente del Berro	De 43 a 58 y 60	
04.4 Guindalera	De 61 a 65, 67 a 94 y 131	
04.5 Lista	De 95 a 113	
04.6 Castellana	De 114 a 121, 123 a 124, 126 a 130	
05. CHAMARTIN		**101**
05.1 El Viso	De 1 a 13	
05.2 Prosperidad	De 14 a 37, 95, 99 y 102	
05.3 Ciudad Jardín	De 38 a 49 y 96	
05.4 Hispanoamérica	De 51 a 69 y 97 a 98	
05.5 Nueva España	70, 72 a 84	
05.6 Castilla	De 85 a 92, 94, 100 a 101 y 103 a 104	
06. TETUAN		**118**
06.1 Bellas Vistas	De 1 a 19, 21, 23 y 125	

06.2 Cuatro Caminos	De 24 a 29, 31 a 46, 48 a 54	
06.3 Castillejos	De 56 a 71	
06.4 Almenara	De 72 a 81, 84 a 85, 89 y 127 a 129	
06.5 Valdeacederas	De 90 a 99, 101 a 107 y 126	
06.6 Berruguete	De 108 a 124	
07. CHAMBERI		**123**
07.1 Gaztambide	De 1 a 14, 16 a 19, 21 y 23	
07.2 Arapiles	De 24 a 26, 28, 30 a 32, 34 a 36 y 38 a 49	
07.3 Trafalgar	De 50 a 53 y 55 a 73	
07.4 Almagro	De 75 a 76 y 78 a 93	
07.5 Ríos Rosas	De 94 a 103, 105 a 111 y 113 a 117	
07.6 Vallehermoso	De 118 a 135	
08. FUENCARRAL-EL PARDO		**184**
08.1 El Pardo	De 1 a 4	
08.2 Fuentelarreina	De 5 a 7	
08.3 Peñagrande	De 8 a 29, 49 a 52, 140 a 143, 145, 148 a 149, 151, 157 y 182	
08.4 Del Pilar	De 30 a 34, 45 a 48, 53 a 69 y 71 a 89	
08.5 La Paz	De 35 a 44, 90 a 100, 102 a 103 y 144,146 y 150	
08.6 Valverde	De 104 a 117, 119 a 126, 128, 153, 159 a 162, 166 a 168, 170 a 172, 174, 179 y 185 a 188	
08.7 Mirasierra	De 131 a 139, 147, 152, 154 a 156, 158, 163, 175, 177, 180 a 181 y 183	
08.8 El Goloso	De 129 a 130, 164, 165, 169, 173, 176, 178 y 184	
09. MONCLOA-ARAVACA		**85**
09.1 Casa de Campo	De 1 a 2, 4 a 10, 81 y 84	
09.2 Argüelles	De 11 a 26 y 28 a 31	
09.3 Ciudad Universitaria	De 33 a 41 y 43 a 45	
09.4 Valdezarza	De 46 a 47, 49 a 58, 60 a 61, 63 a 70, 82 y 90	
09.5 Valdemarín	71, 86 y 91	
09.6 El Plantío	72	
09.7 Aravaca	De 73 a 80, 83, 85, 87 a 89 y 92	
10. LATINA		**200**
10.1 Los Cármenes	De 1 a 12, 215, 217 y 219	
10.2 Puerta del Angel	De 13 a 18, 20 a 44, 46 a 49, 206 y 211	
10.3 Lucero	De 50 a 63, 65, 67 a 70, 72 a 78, 212 a 213 y 218	
10.4 Aluche	De 81 a 84, 86 a 96, 98 a 103, 106 a 140, 142 a 143 y 207	
10.5 Campamento	De 144 a 154 y 156 a 159	
10.6 Cuatro Vientos	216 y 221	
10.7 Las Águilas	De 161 a 165, 167 a 176, 178 a 185, 187 a 194, 197 a 199, 201 a 205, 209 a 210 y 214	
11. CARABANCHEL		**180**
11.1 Comillas	De 1 a 7, 10 a 13, 15 a 19 y 21	
11.2 Opañel	De 22 a 26, 28 a 46, 181 y 184	
11.3 San Isidro	De 47 a 51, 54 a 63, 65 a 75, 180 y 195	
11.4 Vista Alegre	De 78 a 83, 85 a 89, 91 a 112 y 179	
11.5 Puerta Bonita	De 113 a 126, 128 a 137 y 185	
11.6 Buenavista	De 138 a 143, 145 a 156, 186 a 193 y 196 a 198	
11.7 Abrantes	De 157 a 158, 160 a 163, 165, 167 a 171, 173 a 178, 182 a 183 y 194	
12. USERA		**92**

12.1 Orcasitas	De 1 a 3, 5 a 16, 101 a 102 y 106	
12.2 Orcasur	De 18 a 20, 22 a 24, 99 y 108	
12.3 San Fermín	De 27 a 34, 37, 100, 103 a 105 y 109	
12.4 Almendrales	39, 41 a 52 y 107	
12.5 Moscardó	De 53 a 57, 59 y 61 a 62, 64 a 67, 69 a 71, 73 a 74	
12.6 Zofío	75, 77 a 84 y 98	
12.7 Pradolongo	De 86 a 94 y 96 a 97	
13. PUENTE DE VALLECAS		**177**
13.1 Entrevías	De 2 a 8, 10 a 14, 16 a 19, 21 a 24 y 26 a 35	
13.2 San Diego	De 36 a 50 y 52 a 63, 65 a 66 y 69	
13.3 Palomeras Bajas	74, 77 a 80, 83 a 84, 86 a 92, 94, 103 a 104, 203 a 210 y 213	
13.4 Palomeras Sureste	De 108 a 112, 114 a 129, 131 a 140, 199, 211, 215 y 216	
13.5 Portazgo	De 142 a 148, 150, 153 a 158, 160 a 164, 198, 200 y 214	
13.6 Numancia	166, 168 a 187, 189 a 197, 201 a 202, 212 y 217	
14. MORATALAZ		**88**
14.1 Pavones	De 2 a 6, 79, 88 y 92	
14.2 Horcajo	De 89 a 91 y 93	
14.3 Marroquina	10, 12 a 22, 24 a 29, 80, 81, 85 a 87 y 94 a 95	
14.4 Media Legua	De 30 a 42 y 82 a 84	
14.5 Fontarrón	De 43 a 59	
14.6 Vinateros	De 60 a 77	
15. CIUDAD LINEAL		**170**
15.1 Ventas	De 1 a 3, 5 a 6, 8 a 43	
15.2 Pueblo Nuevo	De 44 a 54, 56 a 62, 65 a 71, 73 a 90, 177, 179 y 181	
15.3 Quintana	De 92 a 106 y 108 a 113	
15.4 La Concepción	De 114 a 131	
15.5 San Pascual	De 132 a 136, 138 y 140 a 147	
15.6 San Juan Bautista	De 148 a 151, 153, 174 a 175, 178 y 180	
15.7 Colina	De 154 a 158	
15.8 Atalaya	159	
15.9 Costillares	De 160 a 168, 170 a 173, 176 y 182	
16. HORTALEZA		**122**
16.1 Palomas	De 1 a 2, 105 y 121	
16.2 Piovera	De 3 a 5, 102 a 104 y 113	
16.3 Canillas	De 6 a 9, 11 a 22, 24 a 35, 98, 100 y 101	
16.4 Pinar del Rey	De 36 a 61, 63 a 67, 69 a 78 y 97	
16.5 Apóstol Santiago	De 79 a 83, 85 a 90	
16.6 Valdefuentes	De 91 a 93, 96, 99,106 a 112, 114 a 120 y 122 a 129	
17. VILLAVERDE		**105**
17.1 Villaverde Alto, Casco Histórico de Villaverde	De 2 a 4, 6 a 29, 98 a 100, 106, 109 y 113	
17.2 San Cristóbal	De 31 a 33, 35 a 39 y 41 a 43	
17.3 Butarque	De 45 a 47, 101, 107, 110 a 111 y 114 a 115	
17.4 Los Rosales	De 48 a 55, 57 a 66, 96 a 97, 103 a 105, 108, 112 y 116	
17.5 Los Ángeles	De 68 a 76, 78 a 80, 82 a 87, 89 a 95 y 102	
18. VILLA DE VALLECAS		**67**
18.1 Casco Histórico de Vallecas	De 2 a 25, 42 a 43, 45 y 46	
18.2 Santa Eugenia	De 26 a 41, 44 y 48	
18.3 Ensanche de Vallecas	47 y 49 a 68	

19. VICALVARO		**47**
19.1 Casco Histórico de Vicálvaro	De 1 a 9, 11 a 12, 14 a 16, 18 a 29 y 32	
19.2 Valdebernardo	De 33 a 39, 43 a 44 y 50	
19.3 Valderrivas	De 40 a 42, 45 a 49 y 51	
19.4 El Cañaveral	52	
20. SAN BLAS		**113**
20.1 Simancas	De 1 a 3, 6 a 20, 100, 118 a 119, 122 y 125	
20.2 Hellín	De 21 a 29	
20.3 Amposta	De 31 a 39	
20.4 Arcos	De 40 a 44, 46 a 51, 53, 55 a 56, 58 a 59, 109 y 111 a 112	
20.5 Rosas	De 62 a 67, 102 a 104, 106 a 107, 110, 113 a 115 y 117	
20.6 Rejas	De 68 a 71, 116, 120 a 121, 123 y 126	
20.7 Canillejas	De 72 a 80, 82 a 83, 85 a 88, 90 a 93 y 101	
20.8 El Salvador	De 95 a 99, 105, 108 y 124	
21. BARAJAS		**31**
21.1 Alameda de Osuna	De 1 a 14	
21.2 Aeropuerto	15	
21.3 Casco Histórico de Barajas	De 17 a 21	
21.4 Timón	De 22 a 23, 26, 28, 30 a 31 y 33	
21.5 Corralejos	25, 27, 29 y 32	

FUENTE: Subdirección General de Estadística. Elaboración propia.

TABLA 6.5

Varianza explicada por cada factor

Factor1	Factor2	Factor3	Factor4
7.8358861	4.7251655	2.5377684	0.9442827

TABLA 6.6

Matriz de transformación ortogonal

	1	2	3	4
1	-0.86871	-0.11326	0.47252	-0.09615
2	0.03457	-0.85154	-0.03432	0.52203
3	0.46433	-0.29956	0.68532	-0.47434
4	0.16898	0.41511	0.55307	0.70231

TABLA 6.7			
Varianza explicada por cada factor			
Factor1	**Factor2**	**Factor3**	**Factor4**
6.4931026	3.9172550	3.2358802	2.3968649

TABLA 6.8				
Coeficientes de puntuación estandarizados				
	Factor 1	Factor 2	Factor 3	Factor 4
Edad promedio	-0.01922	-0.15226	-0.07053	0.13838
Proporción de juventud	0.00916	0.33155	0.09481	0.20097
Proporción de envejecimiento	0.00927	-0.02960	0.02808	0.31806
Proporción de extranjeros	0.06498	0.08224	0.36961	0.11430
Proporción de nacidos fuera de España	0.05583	0.07832	0.35659	0.11520
Proporción de extranjeros en edad escolar	0.03636	0.09786	0.34307	0.18867
Proporción estudios superiores	0.17920	-0.00326	0.08023	-0.02625
Proporción estudios obligatorios	-0.16389	0.01502	-0.03866	0.05814
Proporción de hogares unipersonales de 65 y más	0.01463	-0.02779	0.12986	0.29843
Proporción de hogares monoparentales	0.00187	0.29008	0.07030	0.15976
Proporción de hogares con más de una persona adulta con menores	0.00057	0.27670	0.00329	0.12285
Renta media por persona (€)	0.16518	-0.00247	0.06351	0.03499
Renta media por hogar (€)	0.14688	0.05331	0.05888	0.07487
Proporción grupos 1 y 2 s/total afiliados	0.16469	0.00121	0.05215	0.00966
Proporción grupo 10 s/total afiliados	-0.14863	0.01804	-0.02623	-0.00053
Proporción de parados s/pob16a64	-0.18370	-0.05190	-0.16225	-0.10568
Índice de dependencia (<16y65ymas_16a64	0.02086	0.20346	0.10905	0.53669
Índice estructura pob activa 16a39_40_64	0.10542	-0.08661	0.25031	-0.08578

TABLA 7.6				
Semillas iniciales				
Clúster	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	-0.987991750	0.590934199	-1.211563208	4.129297799
2	1.371623115	5.805752860	0.002480034	1.116556269
3	-3.025473191	0.571256032	-2.318908799	-1.576161496
4	-0.794454228	1.083703895	3.335893365	0.040709366
5	1.652633397	0.626432252	-0.667295294	-0.797812674

TABLA 7.7				
Estadísticos para variables				
Variable	**Total STD**	**STD interior**	**R-cuadrado**	**RSQ/(1-RSQ)**
Factor1	0.99811	0.66906	0.551404	1.229175
Factor2	0.99468	0.61395	0.619646	1.629127
Factor3	0.99965	0.66302	0.560825	1.276995
Factor4	0.98113	0.72514	0.454641	0.833653
OVER-ALL	0.99342	0.66895	0.547297	1.208951

TABLA 7.8	
Estadístico Pseudo F	735.95

TABLA 7.9				
Desviaciones estándar del clúster				
Clúster	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	0.6783915751	0.3839472842	0.5183024133	0.6317187747
2	0.8063325427	0.9334118246	0.7310327693	0.6503090907
3	0.6781086934	0.7221120514	0.6381719522	0.8168172002
4	0.4988866158	0.4763919860	0.6756192193	0.6179331025
5	0.7257420434	0.6579453650	0.6939638849	0.7811556412

11. ANEXO II: CÓDIGO

11.1. CÓDIGO SAS

```
/*Trabajo fin de grado*/
/*Valentina Estephanía Crameri Ramírez*/
/*Elaboración de una estratificación de secciones censales de la ciudad de Madrid a partir de datos sociodemográficos*/

/*Creación de librería e importación de datos*/
libname a 'D:\AYUNTAMIENTODEMADRID\ARCHIVOS';
proc import datafile='D:\AYUNTAMIENTODEMADRID\secciones.xlsx' dbms=xlsx out=a.datos;run;
/*Análisis descriptivo*/
/*Medidas de posición e Histogramas*/
proc univariate data=a.datos outtable=a.analisis;
	var Edad_promedio Indice_de_dependencia__16y65ymas
	    Proporci_n_de_extranjeros VAR15;
	histogram/normal;
run;
/*Gráficos de dispersión*/
%macro dispersion(var);
	proc gplot data=a.desvi;
		plot &var*Distrito_seccion=color_&var;
		symbol v=circle h=0.70 cv=VIYPK ;
		symbol2 v=circle h=0.70 cv=bigb;
	run;
%mend;
data a.desvi;
	set a.datos;
	a= (Edad_promedio-44.8678078) /4.1432653713;
	b= (Indice_de_dependencia__16y65ymas-54.237176836) /15.914332617;
	c= (Proporci_n_de_extranjeros-15.407801907) /8.4403106145;
	d= (VAR15-38817.647564)/16681.461011;
run;
data a.desvi;
	set a.desvi;
	if abs(a)>3 then color_a='Atípico';else color_a='No Atípico';
	if abs(b)>3 then color_b='Atípico';else color_b='No Atípico';
	if abs(c)>3 then color_c='Atípico';else color_c='No Atípico';
	if abs(d)>3 then color_d='Atípico';else color_d='No Atípico';
run;
%dispersion(a);
%dispersion(b);
%dispersion(c);
%dispersion(d);
/*Gráfico de caja y bigotes*/
%macro cbigotes(var);
	proc sgplot  data=&data;
		vbox &var;
	run;
%mend;
```

```
%cbigotes(a.datos,Edad_promedio);
%cbigotes(a.datos,Proporci_n_de_extranjeros);
%cbigotes(a.datos,VAR15);
%cbigotes(a.datos,Indice_de_dependencia__16y65ymas);
/*Análisis de la correlación entre pares de variables*/
proc corr data=a.datos out=a.correlacion;
        var Edad_promedio--Proporcion_hogares_con_mas_de_1;
run;
/*Eliminamos las variables poco correlacionadas*/
data a.datos2;
        set a.datos;
        drop Indice_reemplazo_pob_activa_16a1 Proporci_n_de_afiliados_s_pob16a
                Proporcion_hogares_con_mas_de_1         Proporci_n_de_turismos_de_PF_con
                Proporcion_hogares_unipersonales;
run;
/*Análisis factorial*/
proc factor data=a.datos2 nfactors=4 out=a.datos_factor rotate=varimax outstat=coefic msa
plot=all score;run;
/*Gráfico de caja y bigotes para cada factor para determinar atípicos*/
%cbigotes(a.datos_factor,Factor1);
%cbigotes(a.datos_factor,Factor2);
%cbigotes(a.datos_factor,Factor3);
%cbigotes(a.datos_factor,Factor4);
/*Añadimos colores a las secciones consideradas atípicas para el gráfico de dispersión*/
data a.colores;
        set a.datos_factor;
        if Distrito_seccion='08129' then color=1;
                else if Distrito_seccion='08130' then color=1;
                        else if Distrito_seccion='18045' then color=1;
                                else color=0;
run;
/*Matriz de dispersión para cada par de factores*/
proc sgscatter data=a.colores; matrix Factor1-Factor4/group=color; run;
/*Clúster jerárquico */
proc cluster data=a.datos_factor method=average ccc pseudo nonorm
pseudo RSQUARE out=a.arbol print=20;
var Factor1 Factor2 Factor3 Factor4;
id Distrito_seccion;
run;
proc tree data=a.arbol lineas=(color=BIGB);id Distrito_seccion;run;
/*Cluster no jerárquico*/
proc fastclus data=a.datos_factor out=a.clus maxcluster=6; var Factor1--Factor4;run;
/*Tenemos un clúster con 1 observación (08129), como es atípico, se tratará de forma
distinta junto con las secciones 08130 y 18045 determinadas anteriormente*/
/*Descartamos secciones*/
data a.datos_reducido;
        set a.datos_factor;
        if Distrito_seccion='08129' then delete;
        if Distrito_seccion='08130' then delete;
        if Distrito_seccion='18045' then delete;
run;
/*Realizamos análisis clúster jerárquico nuevamente*/
proc cluster data=a.datos_reducido method=average ccc pseudo nonorm
pseudo RSQUARE outtree=a.arbol2 print=20;
```

```
var Factor1 Factor2 Factor3 Factor4;
id Distrito_seccion;
run;
proc tree data=a.arbol2 lineas=(color=BIGB );id Distrito_seccion;run;
/*Clúster no jerárquico*/
proc fastclus data=a.datos_reducido out=a.clus2 maxcluster=5; var Factor1--Factor4; run;
/*A las secciones atípicas se les asigna clúster*/
data a.cluster;
        merge a.datos2 a.clus2;
        by Distrito_seccion;
        if Distrito_seccion='08129' then CLUSTER=6;
        if Distrito_seccion='08130' then CLUSTER=6;
        if Distrito_seccion='18045' then CLUSTER=6;
        keep Distrito_seccion CLUSTER;
run;
/*Con la base de datos a.cluster podemos proceder a realizar el mapa de la ciudad de Madrid con el software QGIS*/
```

11.2. CÓDIGO R

```
#Librerias

library(haven)

library(corrplot)

#Importamos los datos

Datos_seccion_1ene2020 <- read_sav("D:/AYUNTAMIENTODEMADRID/Datos seccion 1ene2020.sav")

#Análisis de correlación entre pares de variables

correlacion<-round(cor(Datos_seccion_1ene2020[,2:24]), 2)

res<-cor.mtest(Datos, conf.level = .95)

corrplot.mixed(correlacion, number.cex = .7,tl.cex = 0.02,

        sig.level=0.05,p.mat=res$p,cex.main=1)

#Eliminamos variables poco correlacionadas y volvemos a realizar el gráfico.

Datos<-Datos[,-c(12,15,20,22,23)]

correlacion<-round(cor(Datos), 2)

res<-cor.mtest(Datos, conf.level = .95)

corrplot.mixed(correlacion, number.cex = .7,tl.cex = 0.02,

        sig.level=0.05,p.mat=res$p,cex.main=1)
```