FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2017/2018

Trabajo de Fin de Grado

TITULO: **ESTUDIO SOCIODEMOGRÁFICO DE LOS DISTRITOS DE MADRID**

***Alumno*: NEREA DOMINGO ASENJO**

***Tutor*: JUANA MARÍA ALONSO REVENGA**

Junio de 2018

UNIVERSIDAD COMPLUTENSE
MADRID

ÍNDICE

1. Introducción

El objetivo principal de este trabajo es el estudio sociodemográfico de los distritos de Madrid. La base de datos de la que se parte para realizar este estudio ha sido obtenida del Portal de Datos Abiertos del Ayuntamiento de Madrid (Ayuntamiento de Madrid, 2017a). Este conjunto de datos contiene información tanto de los distritos como de los barrios de Madrid, pero en este caso únicamente nos hemos quedado con la información correspondiente a los 21 distritos. Además, para cada uno de estos distritos se dispone de la información sobre una serie de indicadores. Estos indicadores de forma general hacen referencia a las características generales del distrito, información relativa a la población, indicadores económicos, indicadores de desempleo, información sobre la educación, la salud, los servicios sociales, la vivienda, la calidad de vida, la seguridad, y por último, sobre los resultados de las elecciones locales.

Finalmente, disponemos de 118 variables correspondientes a los diferentes indicadores y 21 observaciones que hacen referencia a los diferentes distritos de Madrid. Debemos tener en cuenta que la mayor parte de estas variables están expresadas en porcentaje, por lo que esto va a ocasionar que muchos grupos de variables sumen el 100%, por representar el total de la población. Por este motivo, tendremos problemas de multicolinealidad, que se produce cuando una o varias variables regresoras son combinación lineal de otras, pero este problema lo analizaremos más adelante a lo largo del trabajo. Además de estas variables expresadas en porcentaje, también vamos a disponer de otras variables numéricas.

Otro aspecto a tener en cuenta de estas variables es que son indicadores correspondientes al año 2017, a excepción de algunas de ellas que corresponden al año 2015 o 2016.

A partir del conjunto de variables utilizadas en este trabajo, el Servicio de Estudios del Ayuntamiento de Madrid publicó un estudio gráfico con motivo de la actualización del Panel de Indicadores 2017, en el que se puede ver de forma visual el conjunto básico de variables (Ayuntamiento de Madrid, 2017b).

Una vez preparado el conjunto de datos, realizaremos el estudio sociodemográfico de los distritos de Madrid en dos partes. En la primera de ellas, debido al gran número de variables que tenemos, vamos a analizarlos en función de distintos grupos de variables, con el fin de agrupar los distritos similares a partir de dichos grupos. Y en segundo lugar, vamos a explicar algunas de las variables de nuestro fichero, en este caso nos centraremos en el porcentaje de votos de los cuatro partidos políticos más importantes mediante una serie de variables, que dividiremos en dos grupos: económicas y demográficas; con el objetivo de averiguar cuáles son las que más influyen en estos porcentajes.

Por otra parte, para llevar a cabo el análisis de este conjunto de datos a partir de las técnicas estadísticas adecuadas, y de esta manera poder obtener las conclusiones oportunas, vamos a utilizar las hojas de cálculo Microsoft Excel y los softwares estadísticos SPSS y SAS.

2. Análisis descriptivo de las variables

Para conocer más en profundidad las variables de las que disponemos, vamos a realizar un análisis descriptivo de ellas. Pero debido a que tenemos 118 variables, únicamente vamos a realizarlo de aquellas que nos parecen más relevantes.

En primer lugar, se analizan las dos variables que describen de forma general el distrito. Estas dos variables son la superficie y la densidad de población.

- **SUPERFICIE**: esta variable representa la superficie del distrito y se mide en hectáreas. A continuación, vamos a representar el diagrama de caja y bigotes, así como una tabla con los principales descriptivos de esta variable:

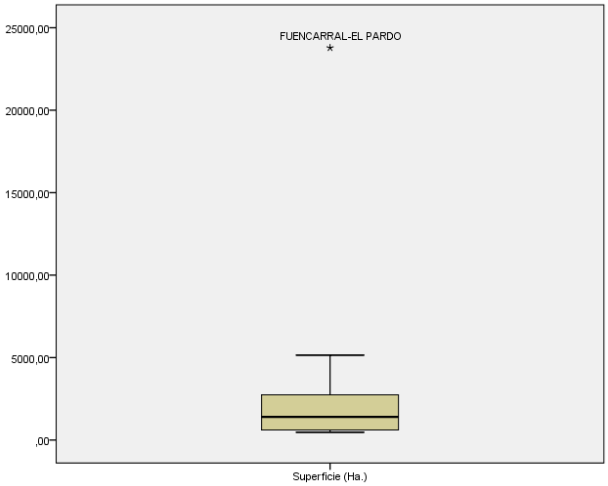

Figura 2.1

	Media			Desviación estándar	
Superficie (Ha.)	2878,3576			5009,97036	
	Mínimo	**Q1 (25%)**	**Mediana**	**Q3 (75%)**	**Máximo**
	467,92	578,4700	1404,8300	3134,3250	23783,84

Tabla 2.1

Tal y como se puede observar en el diagrama de caja y bigotes *(figura 2.1)*, el distrito Fuencarral-El Pardo se considera un valor atípico, ya que este distrito tiene una superficie muy superior al resto de distritos. Esto se puede observar en la *tabla 2.1*, en la cual vemos que tanto la media como la mediana se encuentran muy alejadas del valor máximo, que corresponde a este distrito.

- **DENSIDAD**: esta variable representa la densidad de población de cada distrito, y se mide en habitantes/hectárea. A continuación, vamos a representar un gráfico de barras para ver la densidad de los diferentes distritos, un diagrama de caja y bigotes, y una tabla con los principales descriptivos de esta variable:

	Media			Desviación estándar	
Densidad (hab./Ha.)	139,5514			96,59989	
	Mínimo	**Q1 (25%)**	**Mediana**	**Q3 (75%)**	**Máximo**
	10,04	45,4650	154,3400	225,9900	293,64

Tabla 2.2

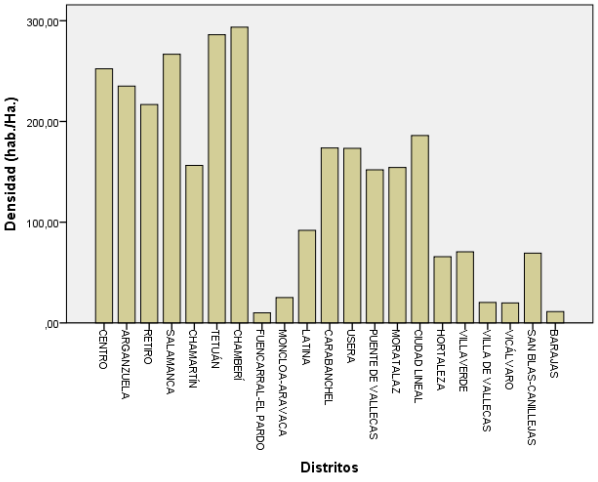

Figura 2.2

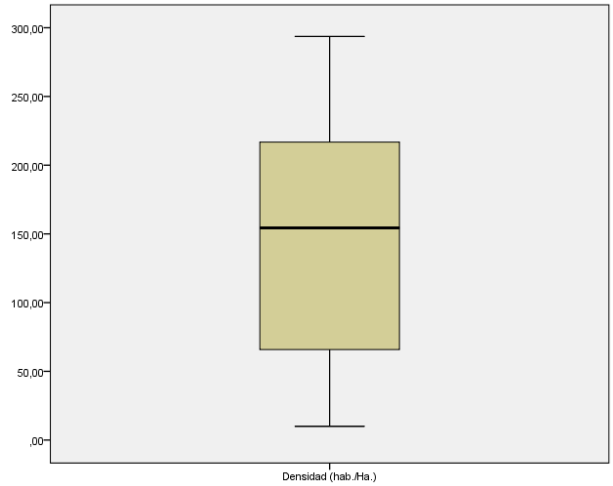

Figura 2.3

Respecto a la densidad de población, en el gráfico de barras *(figura 2.2)* podemos observar claramente cuáles son los distritos con mayor densidad, entre los que destacan Chamberí, Tetuán y Salamanca, y cuáles son aquellos con menor densidad, entre los que podemos destacar Fuencarral-El Pardo y Barajas. Además, si observamos la representación del diagrama de caja y bigotes *(figura 2.3)* y la *tabla 2.2*, podemos ver que hay una gran dispersión entre las densidades de población de los diferentes distritos, y también podemos observar que no tenemos ningún valor atípico.

A continuación, realizaremos el análisis descriptivo de las variables más relevantes referidas a la población del distrito. Las variables que hemos seleccionado son la proporción de población de cada distrito, la proporción de personas con nacionalidad española y extranjera, y por último, atendiendo a la estructura de los hogares, vamos a analizar el tamaño medio del hogar.

- **PROPORCIÓN DE POBLACIÓN**: esta variable expresa la proporción de población de cada distrito respecto del total de población de todos los distritos. Seguidamente, mostraremos el gráfico de barras de esta variable *(figura 2.4)*, el diagrama de caja y bigotes *(figura 2.5)*, y además, una tabla resumen con los descriptivos más importantes *(tabla 2.3)*:

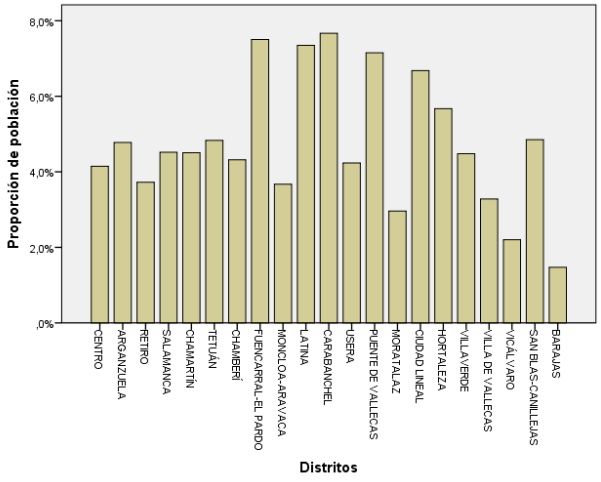

Figura 2.4

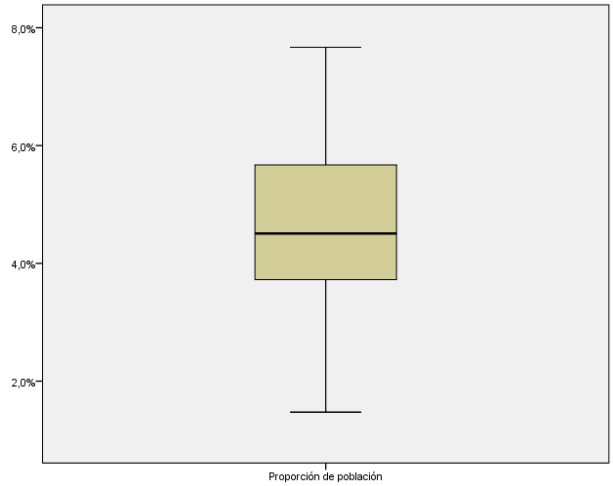

Figura 2.5

Proporción de población	Media			Desviación estándar	
	4,762%			1,7175%	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	1,5%	3,699%	4,507%	6,175%	7,7%

Tabla 2.3

Atendiendo a la proporción de población de cada distrito, podemos observar que hay una serie de distritos que sobresalen frente al resto por tener mayor proporción de población, entre los que podemos destacar Carabanchel, Fuencarral-El Pardo, Latina y Puente de Vallecas con una proporción superior al 7%. Sin embargo, la proporción del resto de distritos es similar, por lo que no hay una gran variabilidad entre ellos. Y, por último, podemos destacar Barajas con una proporción del 1.5%, siendo el distrito con menor proporción de población.

- **NACIONALIDAD ESPAÑOLA Y EXTRANJERA**: en este caso, tenemos dos variables, una correspondiente a la proporción de personas con nacionalidad española y otra con la proporción de personas con nacionalidad extranjera en los diferentes distritos de Madrid. A continuación, vamos a representar estas dos variables mediante gráficos circulares para cada uno de los distritos:

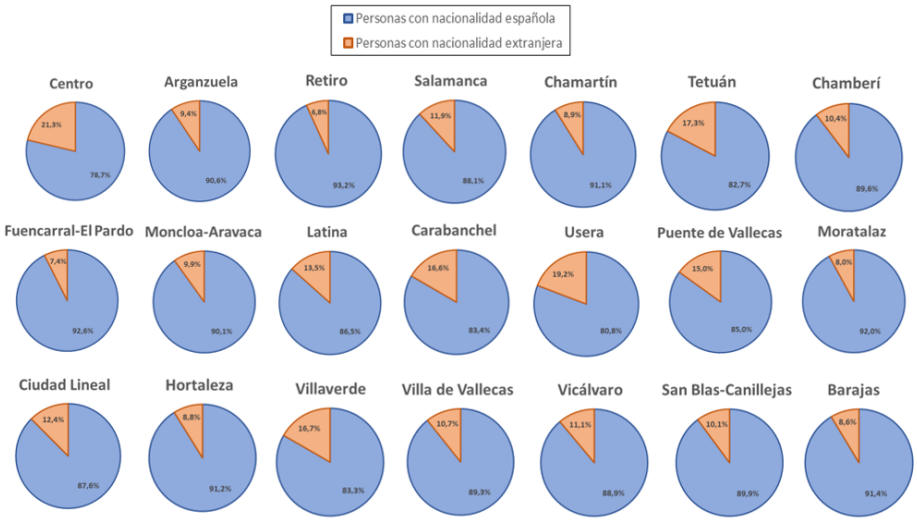

Figura 2.6

En los gráficos circulares *(figura 2.6)*, podemos observar que los distritos con mayor proporción de personas con nacionalidad extranjera son Centro, Usera, Tetuán, Villaverde, Carabanchel y Puente de Vallecas, con una proporción de extranjeros entre el 20% y el 15%. Por otro lado, los distritos con menor proporción de personas con nacionalidad extranjera son Retiro, Fuencarral-El Pardo, Moratalaz, Barajas, Hortaleza y Chamartín, con una proporción en torno al 7-8%.

- **TAMAÑO MEDIO DEL HOGAR**: esta variable indica el tamaño medio del hogar de cada uno de los distritos, es decir, el número medio de personas que viven en los hogares en cada distrito. A continuación, vamos a representar el diagrama de caja y bigotes, así como una tabla con los descriptivos más importantes:

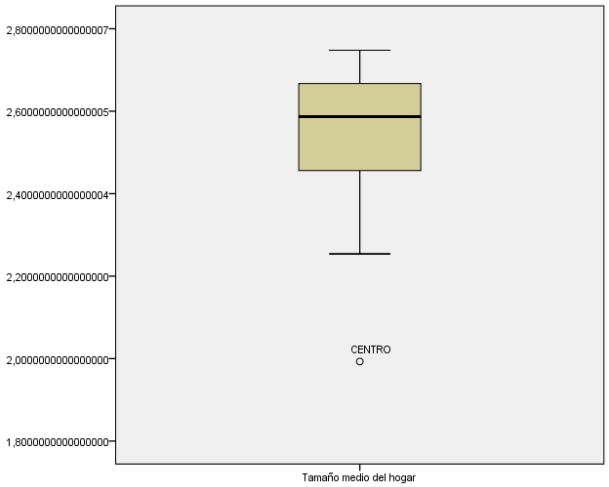

Figura 2.7

Tamaño medio del hogar	Media			Desviación estándar	
	2,526413539			,1892854971	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	1,99290019	2,41233461	2,58698345	2,67403582	2,74796014

Tabla 2.4

Como vemos en la *tabla 2.4* la media se encuentra en torno a 2.5 personas, sin embargo, si observamos el diagrama de caja y bigotes *(figura 2.7)*, vemos que existe un valor atípico, correspondiente a Centro con un tamaño medio del hogar de 1.99 personas. Este valor se considera atípico, ya que tiene un tamaño medio muy inferior al resto de distritos. Además, la explicación que se puede encontrar de este valor atípico, es que en el Centro no suelen vivir grandes familias, ya que las casas suelen ser más pequeñas que en otros distritos, lo que ocasiona que el tamaño medio del hogar sea mucho más pequeño que en el resto de distritos.

Posteriormente, vamos a hacer el análisis descriptivo de dos variables numéricas referidas a indicadores económicos. Estas dos variables son la renta neta media anual de los hogares y la pensión media mensual del distrito, tanto en hombres como en mujeres, para realizar una comparación de ambas pensiones.

- **RENTA NETA MEDIA ANUAL DE LOS HOGARES**: esta variable representa la renta neta media anual de los hogares de cada distrito y se mide en euros. A continuación, para conocer su distribución vamos a realizar un gráfico de barras *(figura 2.8)*, un diagrama de caja y bigotes *(figura 2.9)*, y una tabla con los principales descriptivos *(tabla 2.5)*:

Renta neta media anual de los hogares	Media			Desviación estándar	
	40440,0548			14370,8123	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	23405,0321	28872,0999	34191,2476	54364,5068	69915,6399

Tabla 2.5

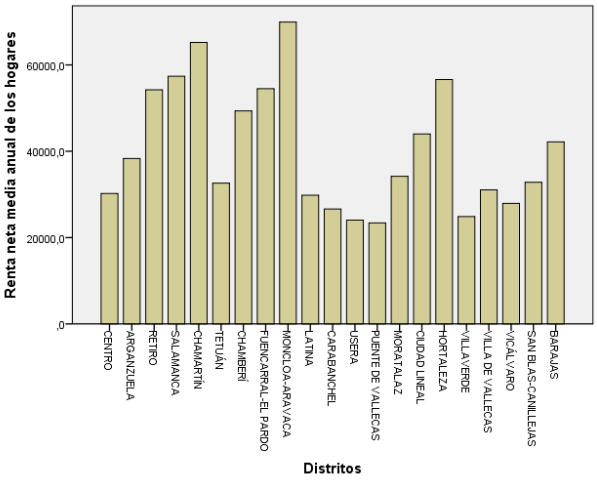

Figura 2.8

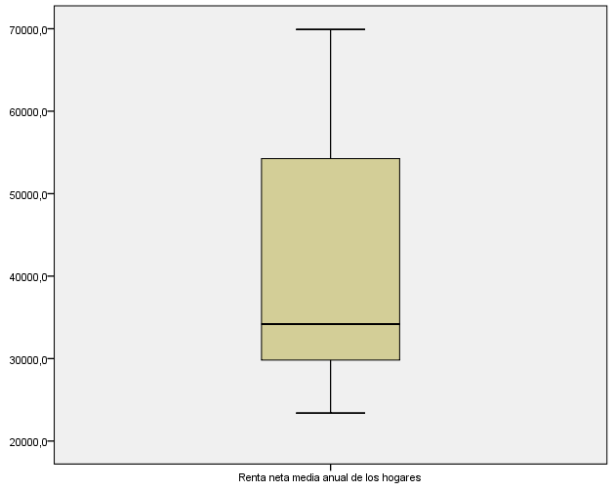

Figura 2.9

En la *tabla 2.5* podemos observar que la renta neta media anual de los hogares es de 40440.0548€, pero este valor está muy influenciado por los valores extremos, ya que como vemos el máximo se encuentra alrededor de 70000€. Por este motivo, una medida resumen más realista sería la mediana, tomando un valor de 34191.2476€. Además, observando el gráfico de barras *(figura 2.8)* podemos ver que el distrito con mayor renta neta media anual es Moncloa-Aravaca, seguido de Chamartín. Y, por último, podemos decir que en esta variable no tenemos ningún valor que se considere atípico, y además, existe una variabilidad bastante grande.

- **PENSIÓN MEDIA MENSUAL DE HOMBRES Y DE MUJERES**: en este caso, tenemos dos variables que miden la pensión media mensual del distrito tanto de los hombres como de las mujeres, para así de este modo poder hacer una comparación entre ellas. Además, estas variables están medidas en euros. Seguidamente, vamos a realizar un gráfico de barras *(figura 2.10)* conjunto de las dos variables para ver la diferencia entre ambos sexos y entre los diferentes distritos de Madrid, y además mostraremos una tabla con los descriptivos más importantes para cada sexo *(tablas 2.6 y 2.7)*:

	Media			Desviación estándar	
Pensión media mensual del Distrito HOMBRES	1399,114762			174,898285	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	1124,2400	1239,4250	1398,7000	1553,9800	1667,0000

Tabla 2.6

	Media			Desviación estándar	
Pensión media mensual del Distrito MUJERES	863,994762			121,663757	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	671,0100	757,6800	854,5100	985,9350	1052,8600

Tabla 2.7

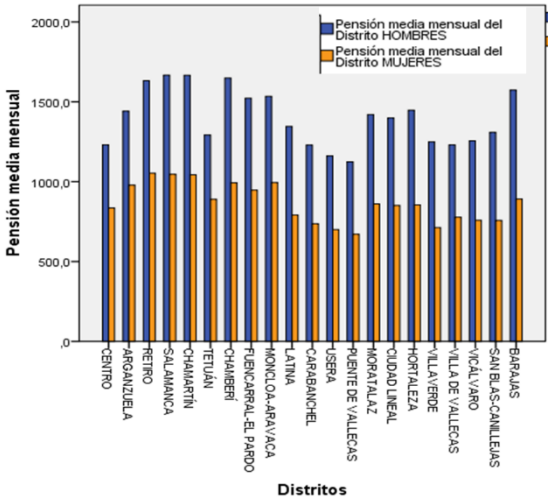

Figura 2.10

En primer lugar, cabe destacar que la pensión media mensual de los hombres es mayor que la de las mujeres en todos los distritos. En el caso de los hombres la media se encuentra en 1399.11€, y en el caso de las mujeres esta media se encuentra en 863.99€. Por lo tanto, la diferencia en media entre hombres y mujeres es de 535.12€. Además, también podemos observar que, aunque las pensiones medias mensuales de las mujeres son más bajas que las de los hombres, ambas siguen la misma distribución en los diferentes distritos.

En relación a los indicadores de desempleo, únicamente vamos a hacer el análisis descriptivo de la variable: tasa absoluta de paro registrado.

- **TASA ABSOLUTA DE PARO REGISTRADO**: esta variable mide la tasa absoluta de paro registrado en cada distrito, y para verla más claramente vamos a representar un gráfico de barras *(figura 2.11)*, un diagrama de caja y bigotes *(figura 2.12)*, y una tabla *(tabla 2.8)* con los descriptivos más relevantes:

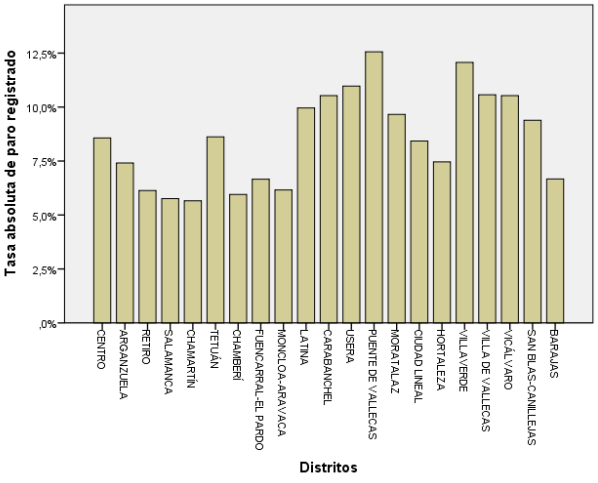

Figura 2.11

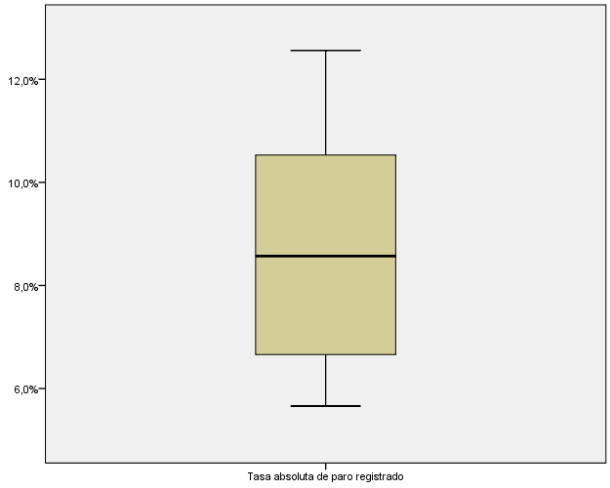

Figura 2.12

	Media			Desviación estándar	
a absoluta de paro registrado	8,558%			2,1708%	
	Mínimo	Q1 (25%)	Mediana	Q3 (75%)	Máximo
	5,7%	6,410%	8,570%	10,530%	12,6%

Tabla 2.8

En este caso, podemos ver que la tasa absoluta media de paro registrado se encuentra en torno al 8.5%. Además, también se puede destacar que los distritos Puente de Vallecas y Villaverde son aquellos que tienen una mayor tasa absoluta de paro registrado, tomando valores alrededor del 12%. Y por último, tal y como vemos en el diagrama de caja y bigotes no hay ningún distrito que se considere atípico.

Si nos centramos en las variables asociadas a la educación, vamos a estudiar la escolarización de alumnos por tipo de centro en el año escolar 2015/2016.

- **ESCOLARIZACIÓN DE ALUMNOS POR TIPO DE CENTRO**: en este caso contamos con tres variables asociadas a la proporción de alumnos escolarizados: en centros privados concertados, en privados sin concierto y en públicos. A continuación, vamos a representar estas tres variables en un gráfico circular para cada distrito:

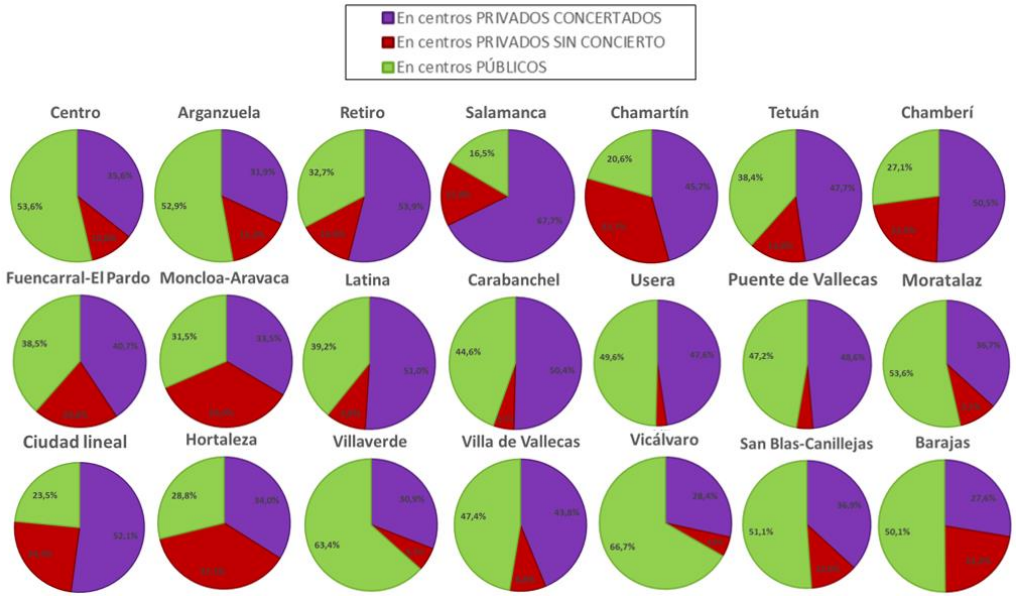

Figura 2.13

Observando estos gráficos circulares *(figura 2.13)*, podemos destacar que Salamanca es el distrito en el que hay mayor proporción de alumnos escolarizados en centros privados concertados, seguido de Retiro y Ciudad Lineal. Por otra parte, Hortaleza es el distrito con mayor proporción de alumnos en centros privados sin concierto, seguido de Moncloa-Aravaca y Chamartín. Y por último, en relación con los centros públicos, podemos decir que los distritos con mayor proporción de alumnos en este tipo de centros son Vicálvaro y Villaverde.

A continuación, vamos a analizar otra de las variables numéricas de las que disponemos en nuestra base de datos. Esta variable está relacionada con las viviendas y mide el valor catastral medio de los bienes inmuebles: personas físicas.

- **VALOR CATASTRAL MEDIO DE LOS BIENES INMUEBLES: PERSONAS FÍSICAS**: esta variable mide el valor catastral medio de los bienes inmuebles de las personas físicas de cada distrito, y está expresado en euros. Seguidamente, vamos a mostrar el gráfico de barras *(figura 2.14)* y el diagrama de caja y bigotes de esta variable *(figura 2.15)*, junto con una tabla con los descriptivos más importantes *(tabla 2.9)*:

	Media			Desviación estándar	
Valor catastral medio de los bienes inmuebles: personas físicas	94368,6305			33760,5581	
	Mínimo	**Q1 (25%)**	**Mediana**	**Q3 (75%)**	**Máximo**
	53729,1888	64950,6733	85687,5064	126857,5053	162041,3047

Tabla 2.9

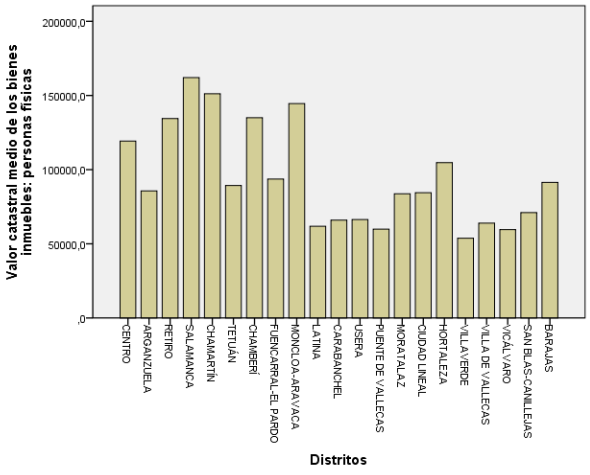

Figura 2.14

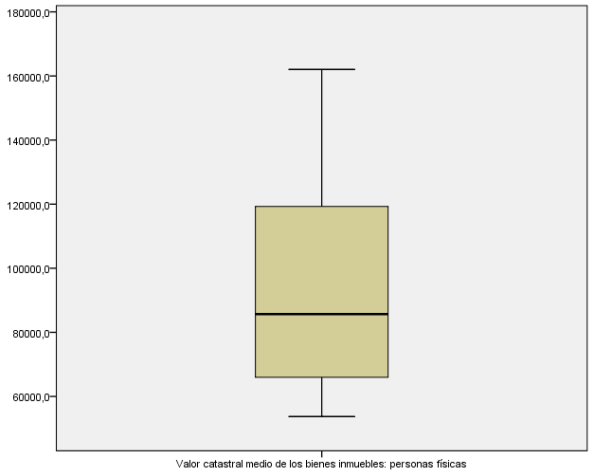

Figura 2.15

Como vemos en la *tabla 2.9*, el valor catastral medio es de 94368.6305€. Sin embargo, al igual que pasaba con la renta neta media anual de los hogares, la media está influida por los valores extremos, por lo que es más representativa la mediana, que toma un valor de 85687.5064€. Por otra parte, los distritos en los que el valor catastral medio es mayor son Salamanca y Chamartín, tomando valores superiores a los 150000€. Y por último, el valor mínimo se encuentra en Villaverde, con un valor catastral medio de 53729.1888€.

Finalmente, vamos a realizar un análisis descriptivo de los resultados de las elecciones locales del año 2015.

- **RESULTADOS ELECCIONES LOCALES**: para analizar estos resultados, vamos a utilizar cuatro variables que hacen referencia a la proporción de personas que votaron en cada uno de los distritos a los siguientes partidos políticos: PP, PSOE, Ahora Madrid y Ciudadanos. Para ello, vamos a representar un gráfico circular con estas cuatro variables para cada distrito *(figura 2.16)*:

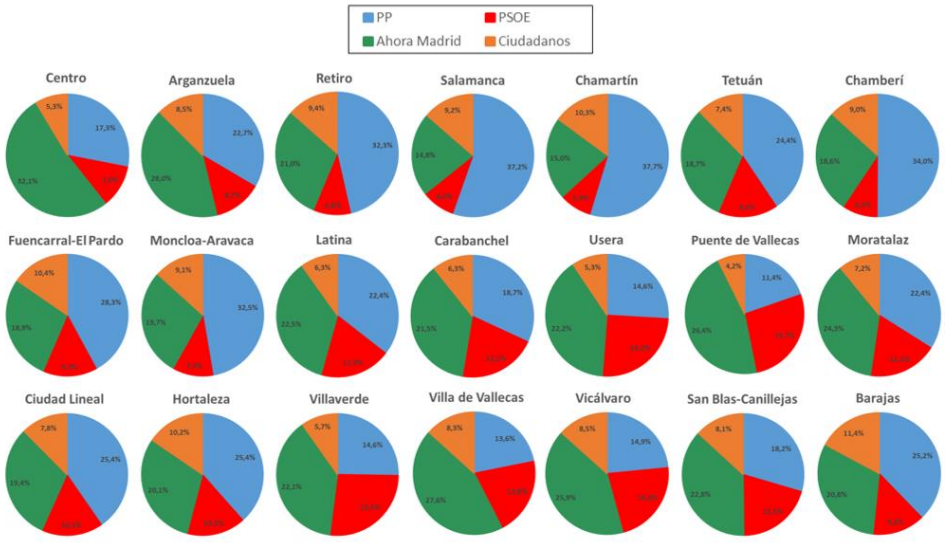

Figura 2.16

En primer lugar, podemos ver que Chamartín y Salamanca son los distritos con mayor proporción de voto al PP. Sin embargo, en el caso del PSOE la mayor proporción de voto se dio en Puente de Vallecas y Villaverde. Por otra parte, en los distritos de Barajas, Fuencarral-El Pardo y Chamartín fue donde se obtuvo la mayor proporción de voto a Ciudadanos. Y, por último, el partido Ahora Madrid fue votado mayoritariamente en Centro, Arganzuela y Villa de Vallecas.

3. Descripción de las técnicas estadísticas a utilizar y su aplicación

Para llevar a cabo el análisis de nuestra base de datos, vamos a utilizar diversas técnicas estadísticas. En primer lugar, nos encontramos con un problema debido a que existen variables que son combinación lineal directa de otras. Estas combinaciones lineales se producen porque la mayor parte de los datos están medidos en porcentaje, y como consecuencia, esto ocasiona que los grupos de ciertas variables que están relacionadas sumen el 100%. Por lo tanto, para solucionar este problema, debemos eliminar una variable de cada uno de estos grupos, para que así de este modo, dejen de existir estas combinaciones lineales.

Una vez solucionado el problema anterior, seguimos teniendo un gran número de variables, siendo este número muy superior al número de observaciones (21 distritos). Por este motivo, a la hora de realizar los diferentes análisis nos vamos a encontrar con algunos problemas. Para solucionarlos realizaremos los diferentes análisis por grupos de variables relacionadas, en lugar de hacerlo con todas las variables juntas. Estos grupos de variables los iremos analizando a lo largo del trabajo.

Finalmente, una vez solucionados todos los problemas anteriores, procederemos a realizar los diferentes análisis en los distintos grupos de variables. En primer lugar, vamos a hacer una breve descripción de estas técnicas para conocer cuáles son sus objetivos. Las técnicas que vamos a utilizar son el Análisis Factorial, el Análisis de Componentes Principales y el Análisis Cluster.

3.1 Descripción de las técnicas estadísticas

3.1.1 Análisis Factorial

El objetivo del Análisis Factorial es agrupar las variables de nuestro estudio mediante sus correlaciones, de manera que todas las variables dentro de un mismo grupo estén altamente correlacionadas entre sí, y a su vez tengan correlaciones relativamente bajas con otras variables que se encuentran en grupos diferentes. Cada grupo de variables representará una variable no observable directamente, lo que se conoce como Factor, que es el responsable de las correlaciones observadas. Además, cabe destacar que estos Factores son incorrelados entre sí.

Además, una vez realizado el Análisis Factorial, cada variable original es combinación lineal de los Factores obtenidos. Los coeficientes de estas combinaciones lineales reciben el nombre de cargas, y todas ellas se recogen en la conocida como matriz de cargas. La interpretación de estos coeficientes se corresponde con la correlación entre los Factores y las variables.

El primer paso para llevar a cabo el Análisis Factorial es calcular la matriz de correlaciones de las variables (R) para comprobar si estas variables están altamente correlacionadas. Para comprobar el grado de asociación entre ellas, se pueden utilizar varios métodos:

- Evaluación del determinante de la matriz de correlaciones. Un determinante bajo indicará correlaciones altas entre las variables, aunque hay que tener en cuenta que no puede ser cero, porque eso indicaría dependencia lineal entre ellas. Por lo tanto, es conveniente que el determinante sea bajo.
- Test de esfericidad de Bartlett. Comprueba si la matriz de correlaciones se ajusta a la matriz identidad, lo que indicaría ausencia de correlación significativa entre las variables. Por tanto, la hipótesis que se contrasta es la siguiente:

$$H_0 : R = I \qquad H_1 : R \neq I$$

- Índice KMO de Kaiser-Meyer-Olkin. Este índice viene dado por:

$$KMO = \frac{\sum_{i \neq j} \sum_{j=1}^{p} r_{i,j}^2}{\sum_{i \neq j} \sum_{j=1}^{p} r_{i,j}^2 + \sum_{i \neq j} \sum_{j=1}^{p} rp_{i,j}^2} \qquad \textit{Ecuación 3.1}$$

  Según este criterio:

$$\textit{si } KMO \leq 0.5 \textit{ valor inaceptable, se desaconseja el } A.F.$$
$$\textit{si } 0.5 \leq KMO \leq 0.6 \textit{ valor demasiado bajo}$$
$$\textit{si } 0.6 \leq KMO \leq 0.8 \textit{ valor mediocre}$$
$$\textit{si } KMO \succ 0.8 \textit{ valor excelente}$$

- Medida de adecuación de la muestra MSA. Está evaluado para cada variable.

$$MSA_j = \frac{\sum_{j' \neq j} r_{j,j'}^2}{\sum_{j' \neq j} r_{j,j'}^2 + \sum_{j' \neq j} rp_{j,j'}^2} \qquad \textit{Ecuación 3.2}$$

En nuestro caso, utilizaremos principalmente el índice KMO de Kaiser-Meyer-Olkin y la medida de adecuación de la muestra MSA para comprobar el grado de asociación entre las variables. Evaluaremos el índice KMO y si este toma valores por debajo de 0.6, eliminaremos aquella variable X_j que tengan un menor valor MSA_j, siempre que este sea menor que 0.5.

Además, como las unidades de las variables de nuestro estudio no son del mismo tipo, vamos a utilizar la matriz de correlaciones en lugar de la matriz de covarianzas.

Por otra parte, existen diferentes métodos para determinar el número de Factores que vamos a retener. En nuestro caso, vamos a elegir el número de Factores en función de la proporción de variabilidad explicada, de manera que dicha proporción se considere suficiente. Y además, también vamos a tener en cuenta otro método que únicamente es válido si utilizamos la matriz de Correlaciones, y que consiste en retener las Factores cuyos autovalores sean mayores que la unidad.

Finalmente, para mejorar la interpretación de los factores obtenidos, podríamos rotar los factores hasta que se consiga una estructura más sencilla de interpretar. Para así, conseguir que cada variable este altamente correlacionada con un único Factor, y poco correlacionada con el resto de los Factores. Además, al rotar los Factores la proporción de variabilidad explicada no varía, al igual que tampoco lo hacen las comunalidades, sin embargo lo único que varía son las cargas Factoriales. Existen varios tipos de rotaciones: Varimax, recomendable principalmente cuando el número de Factores es reducido, y Quartimax, recomendable cuando el número de Factores es elevado.

Y por último, para medir la bondad del ajuste realizado podemos utilizar las comunalidades estimadas, que expresan la parte de cada variable que puede ser explicada por los Factores retenidos.

3.1.2 Análisis de Componentes Principales

El Análisis de Componentes Principales describe la variación de un conjunto de datos multivariante en términos de un conjunto incorrelado de variables, cada una de las cuales es una combinación lineal de las variables originales. Estas nuevas variables reciben el nombre de Componentes Principales y se obtienen en orden decreciente a su importancia. Esto quiere decir que la primera Componente Principal recoge la máxima información de los datos originales, y como consecuencia, la segunda Componente recoge la mayor cantidad de información que no haya sido recogida por la primera Componente, es decir, la información de las variables iniciales que sean incorreladas con la primera Componente Principal, y así sucesivamente con el resto de Componentes Principales.

Debido a que las Componentes Principales se utilizan para resumir los datos originales con la mínima pérdida de información, esto dará lugar a importantes simplificaciones en los análisis posteriores que vamos a realizar.

En nuestra base de datos tenemos variables que no son directamente comparables, ya que no están medidas en las mismas unidades. Por este motivo, se deben tipificar de tal forma que tengan la misma dispersión, es decir, que sean centradas y reducidas, además, la matriz a diagonalizar para obtener las Componentes Principales debe ser la matriz de Correlaciones de las variables originales.

Por otro lado, la primera Componente Principal será una variable cuyo vector dirección coincide con el autovector correspondiente al mayor autovalor de la matriz de Correlaciones, y la segunda Componente Principal será una variable cuyo vector dirección coincide con el autovector correspondiente al segundo mayor autovalor de la matriz de Correlaciones, y así sucesivamente con el resto de Componentes.

Y por último, al igual que ocurría en el Análisis Factorial, existen diferentes métodos para determinar el número adecuado de Componentes Principales. En nuestro caso, elegiremos el número de Componentes en función de la proporción de variabilidad explicada, de manera que dicha proporción se considere suficiente. Además, también vamos a tener en cuenta otro método

que únicamente es válido si utilizamos la matriz de Correlaciones, y que consiste en retener las Componentes cuyos autovalores sean mayores que la unidad.

3.1.3 Análisis Cluster

El objetivo del análisis Cluster es formar grupos de distritos con características similares en función de las variables que tenemos en nuestra base de datos. Además, los grupos que se van a formar deben tener las siguientes características:

- Los distritos de cada grupo deben ser lo más parecidos posible, es decir, debe haber homogeneidad interna.
- Los grupos deben ser lo más diferentes que sea posible, es decir, debe haber heterogeneidad entre grupos.
- Los grupos obtenidos son mutuamente excluyentes, o lo que es lo mismo, cada distrito debe pertenecer a un solo grupo.

Por otro lado, existen dos tipos de Análisis Cluster: Jerárquico y No Jerárquico. En el Análisis Cluster Jerárquico no se conoce el número de grupos que queremos formar, sin embargo, en el Análisis Cluster No Jerárquico si se conoce previamente. En nuestro caso, como no conocemos el número de grupos que vamos a formar con los diferentes distritos, vamos a llevar a cabo el Análisis Cluster Jerárquico.

Los pasos que se deben seguir para realizar un Análisis Cluster Jerárquico son los siguientes:

1) En primer lugar, se parte de tantos grupos como observaciones tengamos. En nuestro caso, partiremos de 21 grupos correspondientes a los 21 distritos.
2) Posteriormente, se genera una matriz simétrica de dimensión n x n, donde n es el número de observaciones que tengamos, que indique las similitudes o distancias entre todos los pares de observaciones. Respecto a nuestra base de datos, se generará una matriz de dimensión 21 x 21, donde se indiquen las similitudes entre todos los pares de distritos.
3) A continuación, se agrupan las dos observaciones más próximas entre sí, es decir, los dos distritos más similares entre sí. Con esto se consigue que el número de Clusters existentes sea uno menos que en el paso anterior.
4) Y por último, se sigue así sucesivamente hasta que únicamente tengamos un Cluster formado por todas las observaciones existentes, es decir, hasta que todos los distritos pertenezcan al mismo Cluster.

El resultado del proceso anterior se puede representar en un diagrama de árbol que se denomina dendrograma. Esta representación permite ver el proceso de aglomeración y formación de los grupos con la distancia entre cada dos grupos unidos. Gracias a este diagrama, podemos decidir con cuantos Clusters finalmente debemos quedarnos.

Además del dendrograma, existen otros procedimientos que nos pueden ayudar a decidir el número de Clusters adecuado. En nuestro caso, los estadísticos que vamos a utilizar para determinar este número adecuado son los siguientes:

- R^2: mide la proporción de variabilidad explicada por los Clusters, por lo que interesa que su valor sea alto. En nuestro caso, vamos a considerar valores altos del R^2 hasta el 70%.
- R^2 semiparcial: mide la pérdida de homogeneidad al formar un nuevo Cluster, por lo tanto, interesan valores pequeños. Además, vamos a buscar los pasos en los que haya un salto, en cuyo caso nos quedaremos con el paso anterior (un Cluster más), debido a que sino perderíamos mucha homogeneidad.
- Pseudo F: Este criterio compara la dispersión entre Clusters (numerador) con la dispersión dentro de los Clusters (denominador). Como necesitamos que este cociente sea el máximo posible, habrá que buscar máximos relativos del valor de este estadístico.
- Pseudo test de la T^2: este test dice que si las medias de dos Clusters distintos no son significativamente diferentes, esos dos agrupamientos podrían combinarse, sin embargo si la diferencia entre esas medias es significativa, entonces los agrupamientos no deben combinarse. En la práctica, observaremos cuando se produce un máximo relativo, rechazaremos el agrupamiento en esa fase, y recomendaremos la clasificación con un Cluster más.

3.2 Aplicación de las técnicas estadísticas

Tal y como habíamos indicado anteriormente, vamos a llevar a cabo los diferentes análisis por grupos de variables similares, debido al gran número de ellas. Estos análisis se muestran a continuación.

3.2.1 Características generales y población del distrito

Dentro de este grupo de variables, hemos incluido las características generales del distrito, que hacen referencia a la superficie y a la densidad de estos, y además, la información referida a la población del distrito, en la que está incluida la estructura de la población (sexo, edad y nacionalidad), la estructura de los hogares y la dinámica demográfica. Por lo tanto, en este grupo, una vez eliminadas las variables que eran combinación lineal directa de otras, contamos con 20 variables y 21 observaciones.

En primer lugar, debido a que tenemos bastantes variables, ya que el número de variables es prácticamente igual al número de observaciones, vamos a proceder a reducir la dimensión. Para ello, utilizaremos el Análisis Factorial. Seguidamente, comprobaremos si las variables están altamente correlacionadas. Para ello, evaluamos el índice KMO y la medida de adecuación MSA.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.17701631																			
Superficie	**Densidad**	**PropPoblacion**	**PropMujeres**	**EdadMedia**	**De0a14anos**	**De15a29**	**De30a44**	**De45a64**	**De65a79**	**PersonasNacionalidadExtranjera**	**PropHogares**	**TamanoMedioHogar**	**HogaresMujerSolaMayor65anos**	**HogaresHombreSoloMayor65anos**	**HogaresMonoparentalesMujer**	**HogaresMonoparentalesHombre**	**TasaBrutaNatalidad**	**EsperanzaVidaNacerMUJERES**	**EsperanzaVidaNacerHOMBRES**
0.1126	0.2138	0.0816	0.1980	0.2707	0.2692	0.0963	0.1137	0.0761	0.1866	0.1061	0.1048	0.1667	0.2599	0.2579	0.2645	0.2292	0.1950	0.1178	0.1072

Tabla 3.1

En este caso, tal y como podemos ver en la *tabla 3.1*, el índice KMO toma un valor inferior a 0.6, por este motivo buscamos aquella variable que tenga menor valor MSA, siempre que este sea menor que 0.5. En este caso, la variable que tiene menor valor de MSA, es la proporción de personas de 45 a 64 años. Por este motivo, eliminamos esta variable y volvemos a llevar a cabo el Análisis Factorial.

Posteriormente llevaremos a cabo el mismo proceso, eliminando las variables correspondientes hasta que consigamos un valor del índice KMO superior a 0.6, este proceso se muestra en el ANEXO II *(tablas II.1)*. Aunque, cabe destacar que en el caso de que hayamos conseguido un índice KMO superior a este valor, si observamos que alguna de las variables tiene un MSA muy pequeño, la eliminaremos para comprobar si este índice ha mejorado. Finalmente, después de todo el proceso, la tabla obtenida sería la que se muestra a continuación *(tabla 3.2)*:

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.68363502												
Superficie	**Densidad**	**PropMujeres**	**EdadMedia**	**De0a14anos**	**De65a79**	**TamanoMedioHogar**	**HogaresMujerSolaMayor65anos**	**HogaresHombreSoloMayor65anos**	**HogaresMonoparentalesMujer**	**HogaresMonoparentalesHombre**	**EsperanzaVidaNacerMUJERES**	**EsperanzaVidaNacerHOMBRES**
0.5483	0.7517	0.7179	0.6724	0.8284	0.7736	0.5731	0.5943	0.6692	0.7117	0.6308	0.8898	0.5743

Tabla 3.2

Una vez eliminadas todas las variables correspondientes, podemos observar que el índice KMO toma un valor próximo a 0.7. Por otro lado, si observamos los valores MSA de las variables, podemos ver que todos ellos son superiores a 0.5. Por lo tanto, estas serán las variables definitivas con las que vamos a llevar a cabo el Análisis Factorial.

A continuación, obtenemos los autovalores de la matriz de Correlaciones, con su correspondiente proporción de varianza explicada y su proporción acumulada para ayudar a decidir con cuantos Factores nos quedamos *(tabla 3.3)*.

Eigenvalues of the Correlation Matrix: Total = 13 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	7.63637479	5.19446428	0.5874	0.5874
2	2.44191051	1.37570870	0.1878	0.7753
3	1.06620181	0.28615538	0.0820	0.8573
4	0.78004643	0.33210919	0.0600	0.9173
5	0.44793724	0.18573998	0.0345	0.9517
6	0.26219726	0.12616350	0.0202	0.9719
7	0.13603376	0.05406721	0.0105	0.9824
8	0.08196655	0.00508205	0.0063	0.9887
9	0.07688449	0.03662642	0.0059	0.9946
10	0.04025807	0.01754679	0.0031	0.9977
11	0.02271128	0.01759431	0.0017	0.9994
12	0.00511696	0.00275610	0.0004	0.9998
13	0.00236086		0.0002	1.0000

Tabla 3.3

Para decidir el número de Factores adecuado disponemos de dos métodos, ya citados anteriormente. Si nos centramos en la proporción de variabilidad explicada, podemos observar que con 3 Factores explicaríamos el 85% de la variabilidad y con 4 el 91%. Para decidir cuál es el más adecuado, vamos a tener en cuenta el método de los autovalores mayores que la unidad. Si seguimos este método, deberíamos quedarnos con 3 Factores, ya que son aquellos cuyos autovalores son mayores que la unidad. Por lo tanto, nos quedamos con 3 Factores que

explican el 85.73% de la variabilidad, y como consecuencia, resumen los datos originales con una mínima pérdida de información.

Seguidamente, se muestra la matriz de saturaciones, en la que se reflejan las correlaciones entre las variables y los 3 Factores que hemos retenido. En ella, vamos a señalar que Factor es el que está más correlacionado con cada una de las variables, para así de este modo obtener una interpretación de estos Factores.

Factor Pattern			
	Factor1	**Factor2**	**Factor3**
Superficie (Ha.)	-0.45827	0.51573	-0.32158
Densidad (hab./Ha.)	0.82445	-0.36983	-0.07595
Proporcion de mujeres	0.77215	0.42588	0.24642
Edad media de la población	0.94791	0.19523	0.00272
Proporcion poblacion de 0 a 14 anos	-0.91942	0.29107	0.21540
Proporcion poblacion de 65 a 79	0.64617	0.66255	-0.05192
Tamano medio del hogar	-0.68920	0.37496	0.55434
Proporcion de Hogares con una mujer sola >65 anos	0.90008	0.16413	0.32793
Proporcion de Hogares con un hombre solo >65 anos	0.91790	-0.07101	-0.03768
Proporcion de Hogares monoparentales: una mujer adulta con uno o mas menores	-0.96364	0.09069	0.15117
Proporcion de Hogares monoparentales: un hombre adulto con uno o mas menores	-0.78791	0.47712	-0.27576
Esperanza de vida al nacer MUJERES	0.46893	0.62044	0.31880
Esperanza de vida al nacer HOMBRES	0.30340	0.70893	-0.48002

Tabla 3.4

Tal y como vemos en la *tabla 3.4*, el Factor 3 no está correlacionado con ninguna variable. Sin embargo, el Factor 1 si lo está con la mayoría de las variables, algo esperable ya que siempre el Factor 1 está relacionado con la mayor parte de ellas. Y por último, el Factor 2 esta correlacionado con las variables superficie, proporción de personas de 65 a 79 años, y esperanza de vida al nacer en las mujeres y en los hombres.

A partir de estas correlaciones, podemos interpretar los Factores anteriores:

- Factor 1: este primer factor hace referencia a toda la información referida a la población del distrito y a sus hogares. El valor de este factor aumenta, cuando aumenta la densidad de población, la proporción de mujeres, y la edad media, y como es lógico, disminuye cuando aumenta la proporción de personas en edades tempranas. Atendiendo a la información relacionada con los hogares, este factor disminuye al aumentar el tamaño medio del hogar. Por otro lado, aumenta cuando crece la proporción de hogares con un hombre o una mujer sola mayor de 65 años, y sin embargo, disminuye cuando crece la proporción de hogares monoparentales.
- Factor 2: este factor hace referencia la superficie del distrito, a la población de mayor edad y a la esperanza de vida al nacer tanto de hombres como de mujeres. Tal y como vemos, este factor aumenta su valor cuando aumenta la superficie, la proporción de personas con edad avanzada, y también, cuando aumenta la esperanza de vida al nacer tanto de mujeres como de hombres, aunque este aumento es mayor en el caso de los hombres.

Las correlaciones que existen entre las variables y los Factores se pueden observar gráficamente en la figura 3.1, que representa las variables en el espacio de los dos primeros Factores:

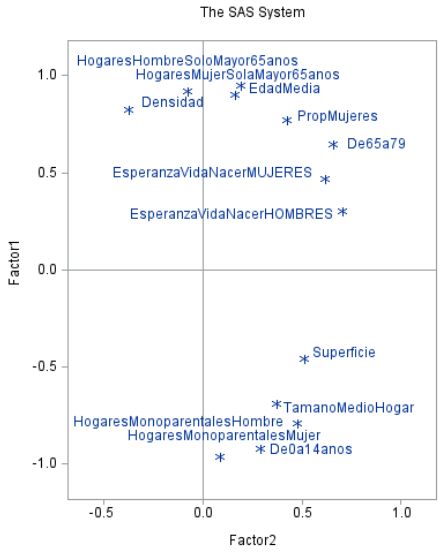

Figura 3.1

En el gráfico *(figura 3.1)*, podemos ver que el Factor 1 se representa en el eje vertical y el Factor 2 en el eje horizontal. Por lo tanto, las variables que se encuentren en la parte superior o derecha tendrán una correlación positiva con el Factor correspondiente, y las que se sitúen en la parte inferior o izquierda tendrán una correlación negativa. Finalmente, las relaciones que se pueden observar son las mismas que habíamos comentado anteriormente.

Por último, para medir la bondad del ajuste realizado, es decir, para medir el grado en que cada variable viene explicada por los factores vamos a mirar las comunalidades, ya que estas expresan la parte de cada variable que puede ser explicada por estos Factores.

Final Communality Estimates: Total = 11.144487												
Superficie	**Densidad**	**PropMujeres**	**EdadMedia**	**De0a14anos**	**De65a79**	**TamanoMedioHogar**	**HogaresMujerSolaMayor65anos**	**HogaresHombreSoloMayor65anos**	**HogaresMonoparentalesMujer**	**HogaresMonoparentalesHombre**	**EsperanzaVidaNacerMUJERES**	**EsperanzaVidaNacerHOMBRES**
0.5794	0.8222	0.8383	0.9366	0.9764	0.8592	0.9228	0.9446	0.8490	0.9596	0.9244	0.7064	0.8250

Tabla 3.5

En la *tabla 3.5*, vemos en primer lugar la suma de los autovalores, y seguidamente las comunalidades de cada variable. Todas ellas toman valores muy grandes excepto superficie, lo que quiere decir que la mayor parte de la variabilidad de cada variable puede ser explicada por los 3 Factores. En el caso de la superficie, tenemos una comunalidad de 0.5794, es decir, únicamente el 57.94% de dicha variable es explicada por los Factores, aunque este valor se puede considerar aceptable. Por lo tanto, estas variables están bien explicadas por los Factores y el análisis factorial se considera correcto.

A continuación, con los Factores obtenidos realizaremos un análisis Cluster, para formar grupos de distritos con características similares. Vamos a utilizar un análisis Cluster Jerárquico, debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Centroid Distance**	**Tie**
20	**CHAMARTIN**	**LATINA**	2	0.0003	1.00	.	.	181	.	0.1866	
19	**RETIRO**	**CL20**	3	0.0016	.998	.	.	58.7	5.5	0.3793	
18	**USERA**	**PUENTE DE VALLECAS**	2	0.0024	.996	.	.	41.2	.	0.534	
17	**SALAMANCA**	**CHAMBERI**	2	0.0025	.993	.	.	36.7	.	0.5486	
16	**HORTALEZA**	**BARAJAS**	2	0.0031	.990	.	.	33.3	.	0.6127	
15	**TETUAN**	**MORATALAZ**	2	0.0033	.987	.	.	32.1	.	0.6252	
14	**VILLA DE VALLECAS**	**VICALVARO**	2	0.0034	.983	.	.	32.1	.	0.6354	
13	**VILLAVERDE**	**SAN BLAS-CANILLEJAS**	2	0.0042	.979	.	.	31.5	.	0.7113	
12	**CARABANCHEL**	**CL13**	3	0.0073	.972	.	.	28.4	1.7	0.8107	
11	**ARGANZUELA**	**CL15**	3	0.0075	.964	.	.	27.1	2.3	0.823	
10	**CL19**	**CIUDAD LINEAL**	4	0.0089	.956	.	.	26.3	9.4	0.8415	
9	**CL10**	**MONCLOA-ARAVACA**	5	0.0111	.945	.	.	25.5	3.1	0.911	
8	**CL12**	**CL18**	5	0.0189	.926	.	.	23.1	4.1	0.9725	
7	**CL11**	**CL9**	8	0.0301	.896	.	.	20.0	5.5	0.9808	
6	**CL7**	**CL17**	10	0.0289	.867	.	.	19.5	3.5	1.0405	
5	**FUENCARRAL-EL PARDO**	**CL16**	3	0.0245	.842	.	.	21.3	7.8	1.4858	
4	**CL8**	**CL14**	7	0.0720	.770	.712	1.89	19.0	10.0	1.7388	
3	**CL6**	**CL4**	17	0.2741	.496	.551	-1.1	8.9	20.3	1.9984	
2	**CL3**	**CL5**	20	0.2159	.280	.333	-.92	7.4	7.7	2.254	
1	**CENTRO**	**CL2**	21	0.2802	.000	.000	0.00	.	7.4	4.2013	

Tabla 3.6

En la *tabla 3.6* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

En primer lugar, analizaremos el R^2 considerando valores altos hasta el 70%, por lo que podríamos elegir como mínimo 4 Clusters, ya que si elegimos menos, la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por otro lado, vamos a analizar el R^2 semiparcial, en el cual podemos observar que con 7 Clusters tenemos un salto, por lo que deberíamos quedarnos con 8. Otro de los saltos, se da con 4 Clusters, el cual es más grande que el anterior, por este motivo nos deberíamos quedar con 5 para no perder tanta homogeneidad al pasar a 4. Finalmente, teniendo en cuenta estos dos criterios, deberíamos quedarnos con 5 Clusters, ya que es el menor número de Clusters en el que coinciden ambos métodos.

También vamos a analizar la Pseudo F. Tal y como podemos ver en la *tabla 3.6*, tenemos un máximo relativo del valor de este estadístico con 5 Clusters, y como coincide con lo que habían determinado los métodos anteriores, nos quedamos con 5 Clusters.

Y por último, miraremos el Pseudo test de la T^2. En nuestro caso, tenemos dos máximos relativos, uno de ellos se da con 10 Clusters, por lo que deberíamos quedarnos con 11. Y el otro máximo relativo se observa con 3 Clusters, por lo que deberíamos quedarnos con 4, sin embargo cabe destacar que con 4 Clusters este estadístico sigue tomando un valor muy elevado, por lo que sería preferible quedarnos con 5, como habíamos determinado con el resto de métodos.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de clusters adecuado para este estudio deben ser 5. Este proceso de aglomeración y formación de los Clusters también se puede representar en una gráfica denominada dendrograma *(figura 3.2)*:

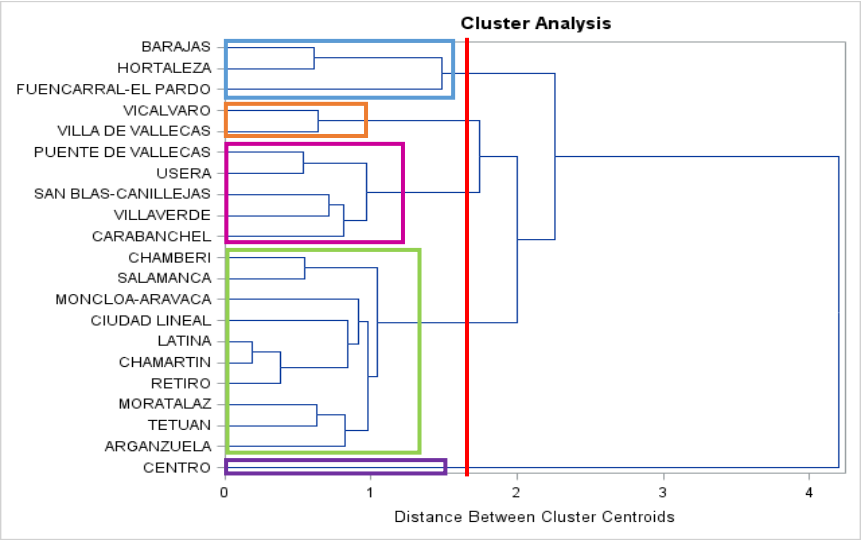

Figura 3.2

A partir de esta representación, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Como el número adecuado de Clusters es 5, deberíamos detener el proceso de agrupación en ese momento y no realizar las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos parado. Además, a partir del dendrograma *(figura 3.2)* podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, analizaremos estos 5 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en los 3 Factores que hemos utilizado.

CLUSTER=1

Obs	Distritos	Factor1	Factor2	Factor3
1	CHAMARTIN	0.76517	0.74862	0.08271
2	LATINA	0.81306	0.86151	0.22341
3	RETIRO	1.03742	0.87240	-0.12561
4	SALAMANCA	1.33248	0.41457	-0.71589
5	CHAMBERI	1.43215	-0.01474	-0.38927
6	TETUAN	0.79392	-0.55803	0.28671
7	MORATALAZ	0.86807	0.03565	0.46820
8	ARGANZUELA	0.45609	-0.20193	-0.35274
9	CIUDAD LINEAL	0.86876	0.56634	0.86013
10	MONCLOA-ARAVACA	-0.02655	0.60701	0.25635

Tabla 3.7

Como podemos observar en la *tabla 3.7*, el Cluster 1 está formado por 10 distritos, siendo este el más numeroso. Estos distritos pertenecen al mismo Cluster, debido a que tienen valores altos del Factor 1, excepto Moncloa-Aravaca que toma un valor negativo aunque mayor que el resto de distritos, por eso pertenece a este grupo. Por lo tanto, esto quiere decir que todos estos distritos se caracterizan por tener una densidad de población elevada, exceptuando el distrito de Moncloa-Aravaca, como era de esperar ya que su valor del Factor 1 no es tan elevado. Además, todos ellos también se caracterizan por tener una edad media elevada. Por otra parte, la proporción de mujeres es mayor que en el resto de distritos, y por último, la proporción de hogares con una persona sola mayor de 65 años supera a la del resto de distritos.

CLUSTER=2

Obs	Distritos	Factor1	Factor2	Factor3
11	USERA	-0.43106	-1.02687	1.11591
12	PUENTE DE VALLECAS	-0.23333	-1.48404	1.30831
13	VILLAVERDE	-0.66120	-0.24545	1.15373
14	SAN BLAS-CANILLEJAS	-0.63420	-0.33382	0.44848
15	CARABANCHEL	0.05037	-0.40320	1.19735

Tabla 3.8

El Cluster 2 está compuesto por 5 distritos *(tabla 3.8)*, y se caracterizan por tomar valores intermedios tanto del Factor 1 como del Factor 2. Todos ellos tienen una superficie media. Además, las proporciones de población en edades tempranas y avanzadas también están en la media. Por otro lado, son de los distritos con mayor tamaño medio del hogar. Y por último, la proporción de hogares monoparentales también es intermedia.

CLUSTER=3

Obs	Distritos	Factor1	Factor2	Factor3
16	HORTALEZA	-0.88646	0.86185	-0.58515
17	BARAJAS	-1.36378	0.70373	-0.93526
18	FUENCARRAL-EL PARDO	-1.08433	2.07593	-1.49076

Tabla 3.9

El Cluster 3 está formado por 3 distritos *(tabla 3.9)*. Se caracterizan por tener un valor negativo del Factor 1 y positivo del Factor 2, aunque de este último es mucho mayor el valor que toma el distrito Fuencarral-El Pardo que los otros dos. En primer lugar, estos distritos se caracterizan por tener una gran superficie, destacando Fuencarral-El Pardo cuya superficie es muy superior a la del resto. Además, también destacan por ser de los distritos donde hay mayor proporción de población en edades tempranas y mayor tamaño medio del hogar. Por último, son los distritos con mayor proporción de hogares monoparentales, en el caso de un hombre adulto con dos o más menores. No obstante, en el caso de una mujer adulta con dos o más menores, aunque no sean los distritos que tienen mayor proporción, están dentro de aquellos con mayor valor.

CLUSTER=4

Obs	Distritos	Factor1	Factor2	Factor3
19	VILLA DE VALLECAS	-1.94984	-0.40165	-0.21137
20	VICALVARO	-1.60785	-0.38031	0.32371

Tabla 3.10

Por otro lado, el Cluster 4 está formado por 2 distritos, Villa de Vallecas y Vicálvaro *(tabla 3.10)*. Se caracterizan por tener un valor muy negativo del Factor 1. Esto quiere decir que tienen una proporción de población muy elevada en edades tempranas. Además, al contrario que en el Cluster anterior, son los dos distritos con mayor proporción de hogares monoparentales, en el caso de una mujer adulta con dos o más menores. Sin embargo, en el caso de un hombre adulto con dos o más menores, aunque no sean los distritos que tienen mayor proporción, están dentro de aquellos con mayor valor.

CLUSTER=5

Obs	Distritos	Factor1	Factor2	Factor3
21	CENTRO	0.46110	-2.69759	-2.91895

Tabla 3.11

Y por último, tenemos el Cluster 5 que está compuesto únicamente por el distrito Centro *(tabla 3.11)*, esto es así porque tiene un valor negativo del Factor 2 y un valor positivo aunque no muy

alto del Factor 1. Por lo tanto, Centro está caracterizado por tener muy poca superficie y un tamaño medio del hogar muy por debajo del resto de distritos. Además, también destaca por ser de los distritos con menor proporción de población en edades avanzadas.

A continuación, a partir del Análisis Factorial que hicimos anteriormente, podemos llevar a cabo una representación gráfica de los distritos en el espacio de los dos primeros Factores, en donde podemos ver gráficamente los Clusters creados *(figura 3.3)*:

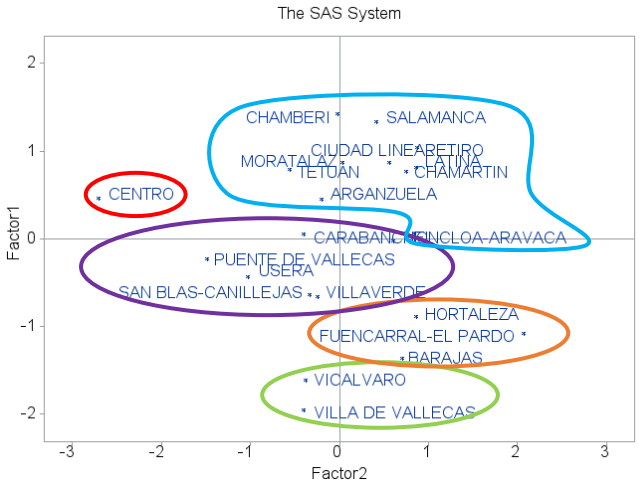

Figura 3.3

Como hemos señalado en el gráfico, podemos ver los 5 Clusters creados anteriormente. Se puede apreciar como los distritos que pertenecen a un mismo Cluster se encuentran próximos entre sí, como era de esperar.

3.2.2 Indicadores económicos e Indicadores de desempleo

En este grupo de variables están incluidos tanto los indicadores económicos como los de desempleo. Dentro de los primeros, vamos a hacer referencia a la renta neta media anual de los hogares y a la pensión media mensual tanto de los hombres como de las mujeres. Por otro lado, dentro de los indicadores de desempleo, vamos a incluir la tasa absoluta de paro registrado de la población total, así como diferenciada por sexo y por edad. Y por último, los parados de larga duración y aquellos que perciben prestaciones. Por lo tanto, una vez eliminadas las variables que eran combinación lineal directa de otras, tenemos 14 variables y 21 observaciones.

En primer lugar, debido al elevado número de variables respecto al número de observaciones, vamos a proceder a reducir la dimensión mediante el análisis Factorial. Seguidamente, comprobaremos si las variables que hemos incluido están altamente correlacionadas. Para ello, evaluamos el índice KMO y la medida de adecuación MSA *(tabla 3.12)*.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.71778682													
RentaNetaMediaAnualHogares	**PensionMediaMensualDistritoHOMBR**	**PensionMediaMensualDistritoMUJER**	**TasaAbsolutaParoRegistrado**	**TasaAbsolutaParoRegistradoMUJERE**	**ParoMujeresDe16a24anos**	**ParoMujeresDe25a44anos**	**ParoMujeresDe45a64anos**	**TasaAbsolutaParoRegistradoHOMBRE**	**ParoHombresDe16a24anos**	**ParoHombresDe25a44anos**	**ParoHombresDe45a64anos**	**ParadosLargaDuracion**	**ParadosPercibenPrestaciones**
0.7904	0.7706	0.7112	0.8118	0.6866	0.6543	0.6530	0.6521	0.9019	0.7162	0.6891	0.6359	0.7918	0.7340

Tabla 3.12

En este caso, vemos que el índice KMO toma un valor superior a 0.7. Además, si observamos los valores MSA de las variables, podemos ver que todos ellos son superiores a 0.6. Por lo tanto, no debemos eliminar ninguna variable, y como consecuencia, estas serán las variables definitivas con las que vamos a llevar a cabo el Análisis Factorial.

A continuación, obtenemos los autovalores de la matriz de Correlaciones, con su correspondiente proporción de varianza explicada y su proporción acumulada para ayudar a decidir con cuantos Factores nos quedamos *(tabla 3.13)*.

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	12.2575885	11.3424907	0.8755	0.8755
2	0.9150978	0.5849899	0.0654	0.9409
3	0.3301079	0.1705354	0.0236	0.9645
4	0.1595725	0.0330504	0.0114	0.9759
5	0.1265221	0.0342340	0.0090	0.9849
6	0.0922881	0.0342116	0.0066	0.9915
7	0.0580765	0.0220960	0.0041	0.9957
8	0.0359805	0.0217833	0.0026	0.9982
9	0.0141972	0.0089930	0.0010	0.9992
10	0.0052041	0.0005748	0.0004	0.9996
11	0.0046294	0.0039326	0.0003	0.9999
12	0.0006968	0.0006636	0.0000	1.0000
13	0.0000332	0.0000276	0.0000	1.0000
14	0.0000056		0.0000	1.0000

Tabla 3.13

Tal y como vemos en la *tabla 3.13*, solo existe un autovalor mayor que la unidad que explica el 87.55% de la variabilidad, y además es un valor bastante alto, por lo que resume los datos originales con una mínima pérdida de información. Por lo tanto, nos quedamos únicamente con un Factor.

Posteriormente, se muestra la matriz de saturaciones, en la que se reflejan las correlaciones entre las variables y el único Factor que hemos retenido *(tabla 3.14)*. A partir de ella, vamos a obtener una interpretación del Factor.

Factor Pattern	
	Factor1
Renta neta media anual de los hogares 2014	-0.89658
Pension media mensual del Distrito HOMBRES (2015)	-0.93698
Pension media mensual del Distrito MUJERES (2015)	-0.95000
Tasa absoluta de paro registrado (agosto 2017)	0.99764
Tasa absoluta de paro registrado MUJERES	0.99220
Proporcion paro registrado mujeres de 16 a 24 anos	0.98060
Proporcion paro registrado mujeres de 25 a 44 anos	0.98038
Proporcion paro registrado mujeres de 45 a 64 anos	0.98249
Tasa absoluta de paro registrado HOMBRES	0.98246
Proporcion paro registrado hombres de 16 a 24 anos	0.98574
Proporcion paro registrado hombres de 25 a 44 anos	0.97370
Proporcion paro registrado hombres de 45 a 64 anos	0.92107
Parados de larga duracion (agosto 2017)	0.53020
Parados que Si perciben prestaciones (agosto 2017)	-0.88943

Tabla 3.14

Debido a que únicamente tenemos un Factor, se podría interpretar como un índice que resume tanto los indicadores económicos como los de desempleo. Pero en realidad nos interesaría un valor del índice para cada uno de los distritos para poder interpretarlo correctamente. Como ya sabemos en el Análisis Factorial las variables son combinación lineal de los Factores, sin embargo, en el Análisis de Componentes Principales las Componentes son combinación lineal de las variables originales. En este caso, sería preferible el Análisis de Componentes Principales, para poder obtener un valor del índice para cada distrito en función de los valores que tomen las variables correspondientes.

Por este motivo, seguidamente vamos a llevar a cabo un Análisis de Componentes Principales, en el que vamos a quedarnos únicamente con una Componente Principal, que explica el 87.55% de la variabilidad *(tabla 3.15).*

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	12.2575885		0.8755	0.8755

Tabla 3.15

Como ya sabemos, las Componentes Principales son combinación lineal de las variables originales. Además, los coeficientes de estas combinaciones lineales se corresponden con los autovectores. A continuación, vamos a mostrar el único autovector correspondiente a la Componente Principal que hemos retenido.

Eigenvectors	
	Prin1
Renta neta media anual de los hogares 2014	-.256086
Pension media mensual del Distrito HOMBRES (2015)	-.267624
Pension media mensual del Distrito MUJERES (2015)	-.271345
Tasa absoluta de paro registrado (agosto 2017)	0.284951
Tasa absoluta de paro registrado MUJERES	0.283397
Proporcion paro registrado mujeres de 16 a 24 anos	0.280084
Proporcion paro registrado mujeres de 25 a 44 anos	0.280023
Proporcion paro registrado mujeres de 45 a 64 anos	0.280623
Tasa absoluta de paro registrado HOMBRES	0.280615
Proporcion paro registrado hombres de 16 a 24 anos	0.281554
Proporcion paro registrado hombres de 25 a 44 anos	0.278115
Proporcion paro registrado hombres de 45 a 64 anos	0.263081
Parados de larga duracion (agosto 2017)	0.151440
Parados que Si perciben prestaciones (agosto 2017)	-.254044

Tabla 3.16

A partir de la *tabla 3.16*, podemos interpretar la Componente que va a resumir los indicadores originales. Como podemos observar, el índice tomara valores menores cuando tanto la renta neta media anual como las pensiones medias, tanto de hombres como de mujeres, tomen valores altos, es decir, existe una relación inversa. Por otro lado, cuanto más aumenta la tasa de paro, independientemente del sexo o de la edad, y además aumenten los parados de larga duración, este índice tomara valores mayores. Y por último, si aumentan los parados que perciben prestaciones, este índice disminuirá. Por lo tanto, se puede decir que este índice tomara valores pequeños cuando se reciban ingresos, tanto de la renta como de pensiones o de prestaciones, y por otro lado, tomara valores grandes cuando no se reciban ingresos, es decir, cuando aumente la tasa de parados, independientemente del sexo y la edad, o aumente la proporción de parados

de larga duración. Esta relación que acabamos de analizar, se puede observar en el gráfico siguiente *(figura 3.4)*:

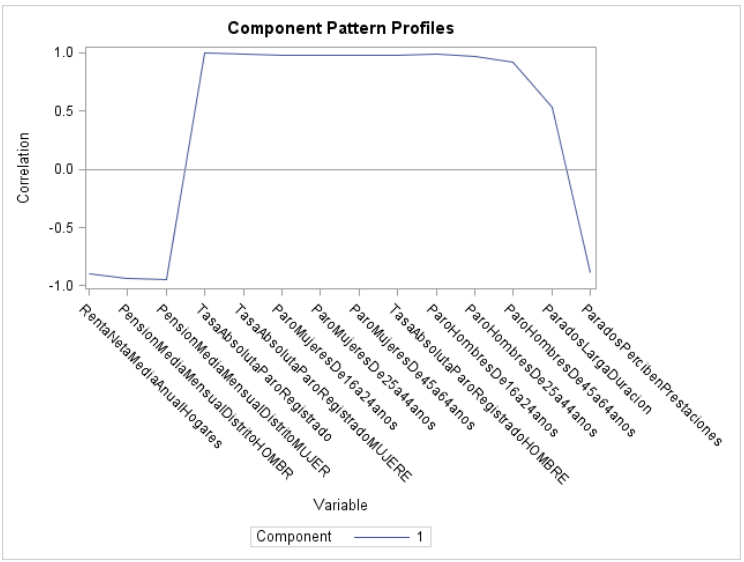

Figura 3.4

En el gráfico *(figura 3.4)* podemos ver claramente las variables que intervienen de forma positiva y aquellas que intervienen de forma negativa, tal y como ya habíamos analizado anteriormente.

A continuación, vamos a calcular el valor de esta Componente Principal para cada uno de los distritos, a partir del valor del autovector y de los valores que toman cada uno de ellos en las variables utilizadas *(tabla 3.17)*. De esta manera, podremos analizar cómo es cada uno de los distritos respecto a este índice.

Distritos	**Prin1**
CENTRO	0.12146
ARGANZUELA	-1.83252
RETIRO	-4.04205
SALAMANCA	-4.70517
CHAMARTIN	-4.87206
TETUAN	0.41562
CHAMBERI	-4.24231
FUENCARRAL-EL PARDO	-2.78969
MONCLOA-ARAVACA	-3.85043
LATINA	1.95859
CARABANCHEL	3.20467
USERA	4.07437
PUENTE DE VALLECAS	6.14223
MORATALAZ	1.75032
CIUDAD LINEAL	-0.46909
HORTALEZA	-1.81438
VILLAVERDE	5.20788
VILLA DE VALLECAS	3.43846
VICALVARO	3.72517
SAN BLAS-CANILLEJAS	1.51954
BARAJAS	-2.94058

Tabla 3.17

Para poder interpretarlo más claramente, vamos a realizar una representación gráfica de los datos anteriores.

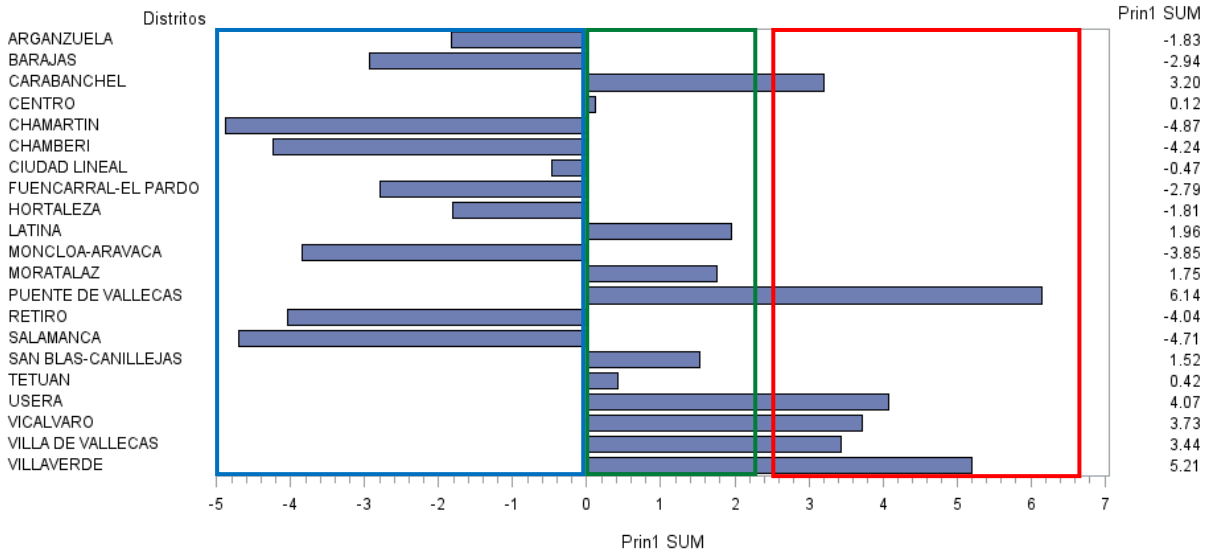

Figura 3.5

Si observamos el gráfico *(figura 3.5)*, podemos ver que hay una serie de distritos que toman valores positivos y otros que toman valores negativos, por lo que podemos hacer varios grupos.

En primer lugar, vemos que hay 6 distritos que toman valores positivos y grandes del índice. Estos distritos son Carabanchel, Puente de Vallecas, Usera, Vicálvaro, Villa de Vallecas y Villaverde. Estos 6 distritos se caracterizan por tener altas tasas de paro, independientemente del sexo y de la edad, así como elevadas proporciones de parados de larga duración.

Por otro lado, tenemos a los 5 distritos que quedan en la parte positiva, que son aquellos que toman valores positivos, pero no excesivamente grandes. Estos 5 distritos son Centro, Latina, Moratalaz, San Blas-Canillejas y Tetuán. Se caracterizan por tener tasas de paro elevadas, aunque no tan altas como las del grupo anterior, pero si mayores que las del resto de distritos. Y además, proporciones de parados de larga duración similares a las del grupo anterior.

Y por último, están los distritos con valores negativos, que son Arganzuela, Barajas, Chamartín, Chamberí, Ciudad Lineal, Fuencarral-El Pardo, Hortaleza, Moncloa-Aravaca, Retiro y Salamanca. Estos distritos se caracterizan por tener altos niveles de renta neta media anual, altas pensiones medias mensuales, tanto en hombres como en mujeres, y además una proporción elevada de parados que perciben prestaciones.

3.2.3 Educación

En este grupo tenemos las variables referidas a la educación. Las variables que vamos a incluir son aquellas que indican la proporción de población en las diferentes etapas educativas, y aquellas que indican la escolarización de alumnos por tipo de centro, además esto último también medido en los alumnos extranjeros y en los alumnos con necesidades de apoyo educativo. Y por último, también vamos a incluir las variables que miden el nivel de estudios

de la población mayor de 25 años. Por lo tanto, una vez eliminadas las variables que son combinación lineal directa de otras, contamos con 17 variables y 21 observaciones.

En primer lugar, debido a que tenemos un número bastante elevado de variables con respecto al número de observaciones, vamos a proceder a reducir la dimensión mediante el análisis Factorial. Seguidamente, vamos a comprobar si las variables que hemos incluido están altamente correlacionadas. Para ello, vamos a evaluar el índice KMO y la medida de adecuación MSA.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.40575639																
EtapasEducativasInfantil3a5anos	**EtapasEducativasPrimaria6a11anos**	**EtapasEducativasSecundaria12a15a**	**CentrosPRIVADOSCONCERTADOS**	**CentrosPUBLICOS**	**PropAlumnosExtranjeros**	**ExtranjerosCentrosPRIVADOSCONCER**	**ExtranjerosCentrosPUBLICOS**	**PropAlumnosNecesidadesApoyoEduca**	**AlumnosApoyoEducativoCentrosPRIV**	**AlumnosApoyoEducativoCentrosPUBL**	**NoSabeLeerNiEscribirOSinEstudios**	**PrimariaIncompleta**	**BachillerElementalGraduadoEscola**	**FP2gradoBachillerSuperiorOBUP**	**TituladosMediosDiplomadosArquite**	**EstudiosSuperioresLicenciadoDoct**
0.2218	0.3714	0.1718	0.3284	0.4199	0.3796	0.5093	0.3568	0.5699	0.4621	0.6578	0.4472	0.4559	0.4703	0.1548	0.4177	0.4576

Tabla 3.18

Tal y como podemos ver en la *tabla 3.18*, el índice KMO toma un valor inferior a 0.6, por este motivo buscamos aquella variable que tenga menor valor MSA, siempre que este sea menor que 0.5. En este caso, la variable que menor valor de MSA tiene es la que indica el nivel de estudios de la población mayor de 25 años: Formación profesional 2º grado, Bachiller Superior o BUP. Por este motivo, eliminamos esta variable y volvemos a llevar a cabo el Análisis Factorial.

Posteriormente, seguimos el mismo proceso, eliminando las variables correspondientes, hasta que consigamos un valor del índice KMO superior a 0.6, este proceso se muestra en el ANEXO II *(tablas II.2)*. Finalmente, una vez realizado todo el proceso, la tabla obtenida es la siguiente:

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.72280644													
EtapasEducativasInfantil3a5anos	**EtapasEducativasPrimaria6a11anos**	**CentrosPUBLICOS**	**PropAlumnosExtranjeros**	**ExtranjerosCentrosPRIVADOSCONCER**	**ExtranjerosCentrosPUBLICOS**	**PropAlumnosNecesidadesApoyoEduca**	**AlumnosApoyoEducativoCentrosPRIV**	**AlumnosApoyoEducativoCentrosPUBL**	**NoSabeLeerNiEscribirOSinEstudios**	**PrimariaIncompleta**	**BachillerElementalGraduadoEscola**	**TituladosMediosDiplomadosArquite**	**EstudiosSuperioresLicenciadoDoct**
0.5681	0.7296	0.7020	0.6381	0.5436	0.7270	0.8030	0.7557	0.8461	0.7736	0.7667	0.7084	0.6662	0.6804

Tabla 3.19

En la tabla final obtenida *(tabla 3.19)*, vemos que el índice KMO toma un valor superior a 0.7. Por otro lado, si observamos los valores MSA de las variables, podemos ver que todos ellos son superiores a 0.5. Por lo tanto, estas serán las variables definitivas con las que vamos a llevar a cabo el Análisis Factorial.

A continuación, obtenemos los autovalores de la matriz de Correlaciones, con su correspondiente proporción de varianza explicada y su proporción acumulada para ayudar a decidir con cuantos Factores nos quedamos *(tabla 3.20)*.

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	7.91926863	5.37437461	0.5657	0.5657
2	2.54489402	1.49252411	0.1818	0.7474
3	1.05236991	0.39951964	0.0752	0.8226
4	0.65285028	0.18710879	0.0466	0.8692
5	0.46574149	0.00599231	0.0333	0.9025
6	0.45974918	0.14946585	0.0328	0.9353
7	0.31028334	0.05969200	0.0222	0.9575
8	0.25059133	0.10719453	0.0179	0.9754
9	0.14339681	0.01942049	0.0102	0.9857
10	0.12397631	0.08131084	0.0089	0.9945
11	0.04266548	0.01936017	0.0030	0.9976
12	0.02330531	0.01354102	0.0017	0.9992
13	0.00976428	0.00862066	0.0007	0.9999
14	0.00114363		0.0001	1.0000

Tabla 3.20

En este caso, a partir de 3 Factores conseguimos explicar más del 80% de la variabilidad, y con 5 Factores llegamos al 90.25%. Pero para determinar con cuantos Factores nos quedamos, vamos a observar cuantos autovalores son mayores que la unidad. Deberíamos quedarnos con 3 Factores, ya que son aquellos cuyos autovalores son mayores que la unidad. Además, con estos 3 Factores explicamos el 82.26% de la variabilidad, y por lo tanto resumen los datos originales con una mínima pérdida de información.

A continuación, se muestra la matriz de saturaciones en la cual se reflejan las correlaciones entre las variables y los 3 Factores que hemos retenido. En ella, vamos a señalar que Factor es el que está más correlacionado cada una de las variables, para así de este modo obtener una interpretación de estos Factores.

Factor Pattern			
	Factor1	**Factor2**	**Factor3**
Proporcion de poblacion en etapas educativas: Infantil (3 a 5 anos)	-0.39408	-0.15477	0.83852
Proporcion de poblacion en etapas educativas: Primaria (6 a 11 anos)	0.67015	-0.35824	-0.40217
Proporcion de escolarizacion de alumnos por tipo de centro: centros PUBLICOS	0.73912	-0.32850	0.00663
Proporcion de alumnos extranjeros	0.57252	0.65024	0.09562
Proporcion de alumnos extranjeros en centros PRIVADOS CONCERTADOS	-0.03094	0.87974	0.10611
Proporcion de alumnos extranjeros en centros PUBLICOS	0.47920	-0.69972	0.00624
Proporcion de alumnos con necesidades de apoyo educativo	0.76946	0.27756	0.12805
Proporcion de alumnos con necesidades de apoyo educativo en centros PRIVADOS CONCERTADOS	-0.74639	0.41496	-0.23217
Proporcion de alumnos con necesidades de apoyo educativo en centros PUBLICOS	0.81946	-0.29985	0.29846
Nivel de estudios de la poblacion >= 25 anos: No sabe leer ni escribir o sin estudios	0.95635	0.09331	0.03002
Nivel de estudios de la poblacion >= 25 anos: Primaria incompleta	0.95742	0.16388	-0.04923
Nivel de estudios de la poblacion >= 25 anos: Bachiller Elemental, Graduado Escolar, ESO,Formacion profesional 1º grado	0.97006	0.12363	0.05497
Nivel de estudios de la poblacion >= 25 anos: Titulados medios, diplomados, arquitecto o ingeniero tecnico	-0.83074	-0.45634	0.03214
Nivel de estudios de la poblacion >= 25 anos: Estudios superiores, licenciado, arquitecto o ingeniero superior, estudios superiores no universitarios, doctorado, estudios postgraduados	-0.96288	-0.00916	-0.01661

Tabla 3.21

Tal y como vemos en la *tabla 3.21*, el Factor 1 está correlacionado con la mayoría de las variables, algo esperable ya que siempre está relacionado con la mayor parte de ellas. El Factor 2 esta correlacionado con las variables relacionadas con los extranjeros. Y por último, el Factor 3 esta correlacionado únicamente con la variable que mide la proporción de población en etapas educativas: Infantil (3 a 5 años). A partir de estas correlaciones, podemos interpretar los Factores anteriores:

- Factor 1: este primer factor hace referencia tanto a la escolarización de alumnos por tipo de centro, como a la escolarización de los alumnos con necesidades de apoyo educativo. Por otro lado, también hace referencia al nivel de estudios de la población mayor de 25 años, y por último, a la proporción de población en etapas educativas: primaria (6 a 11 años). Hay que destacar que esta última proporción influye de forma positiva, al igual que lo hace la proporción de alumnos con necesidades de apoyo educativo. Sin embargo, los centros públicos intervienen de manera directa en el Factor, y por el contrario los centros privados concertados lo hacen de forma negativa. Por último, atendiendo al nivel de estudios de la población mayor de 25 años, vemos que los niveles bajos intervienen de manera positiva, sin embargo los niveles más altos lo hacen de forma negativa.
- Factor 2: este factor hace referencia a toda la información relativa a los alumnos extranjeros. Tal y como podemos observar, la proporción de alumnos extranjeros y su escolarización en centros privados concertados interviene de forma positiva en el segundo Factor. Sin embargo, su escolarización en centros públicos influye de forma negativa.
- Factor 3: el tercer factor únicamente hace referencia a la proporción de población en etapas educativas: Infantil (3 a 5 años), y además esta variable influye de forma directa en el Factor.

Las correlaciones que existen entre las variables y todos los Factores se pueden observar en los siguientes gráficos *(figuras 3.6 y 3.7)*, que representan las variables en el espacio de los dos primeros Factores, y por otro lado del primer y tercer Factor.

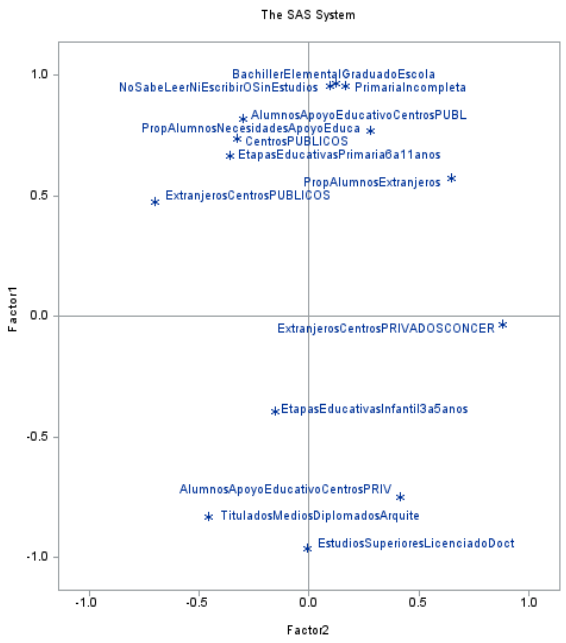

Figura 3.6

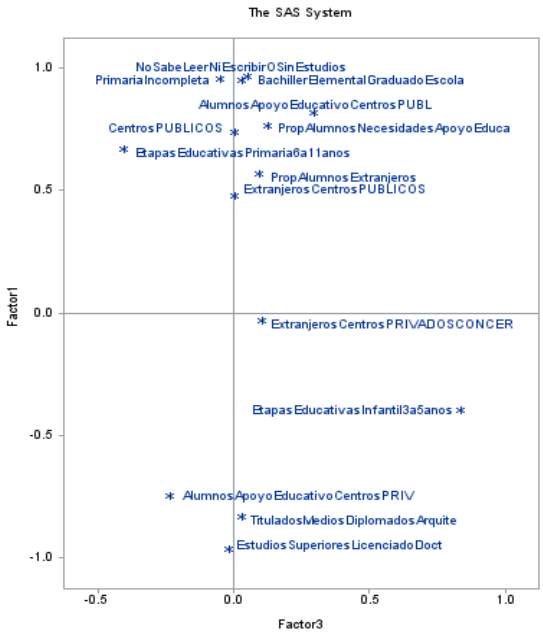

Figura 3.7

En los dos gráficos anteriores podemos ver que el Factor 1 se representa en el eje vertical *(figuras 3.6 y 3.7)*. Por lo tanto, las variables que se encuentren en la parte superior del gráfico tendrán una correlación positiva con el Factor 1, y por el contrario aquellas que se encuentren en la parte inferior tendrán una correlación negativa. Por otro lado, el Factor 2 y el Factor 3 se

representan en el eje horizontal en los dos gráficos, por lo tanto las variables situadas en la parte derecha tendrán una correlación positiva con el Factor correspondiente, y las que se encuentren en la parte izquierda tendrán una correlación negativa. Finalmente, las relaciones que se pueden observar son las mismas que habíamos comentado anteriormente.

Por último, para medir la bondad del ajuste realizado, es decir, para medir el grado en que cada variable viene explicada por los factores vamos a mirar las comunalidades, ya que estas expresan la parte de cada variable que puede ser explicada por estos 3 Factores.

Final Communality Estimates: Total = 11.516533													
EtapasEducativasInfantil3a5anos	EtapasEducativasPrimaria6a11anos	CentrosPUBLICOS	PropAlumnosExtranjeros	ExtranjerosCentrosPRIVADOSCONCER	ExtranjerosCentrosPUBLICOS	PropAlumnosNecesidadesApoyoEduca	AlumnosApoyoEducativoCentrosPRIV	AlumnosApoyoEducativoCentrosPUBL	NoSabeLeerNiEscribirOSinEstudios	PrimariaIncompleta	BachillerElementalGraduadoEscola	TituladosMediosDiplomadosArquite	EstudiosSuperioresLicenciadoDoct
0.8823	0.7391	0.6542	0.7597	0.7861	0.7192	0.6855	0.7831	0.8505	0.9242	0.9459	0.9593	0.8994	0.9275

Tabla 3.22

En la *tabla 3.22*, nos aparece en primer lugar la suma de los autovalores, y seguidamente las comunalidades de cada variable. Como podemos observar todas ellas toman valores superiores a 0.6, lo cual quiere decir que más del 60% de cada variable está explicada por los 3 Factores. Por lo tanto, estas variables están bien explicadas por los Factores y en consecuencia el análisis Factorial se considera correcto.

A continuación, a partir de los Factores obtenidos realizaremos un análisis Cluster, con el objetivo de formar grupos de distritos con características similares. Vamos a utilizar un análisis Cluster Jerárquico debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
20	LATINA	CARABANCHEL	2	0.0010	.999	.	.	52.5	.	0.3468	
19	CENTRO	TETUAN	2	0.0024	.997	.	.	32.1	.	0.5421	
18	ARGANZUELA	HORTALEZA	2	0.0033	.993	.	.	25.9	.	0.6304	
17	USERA	PUENTE DE VALLECAS	2	0.0034	.990	.	.	24.3	.	0.6419	
16	MORATALAZ	VICALVARO	2	0.0036	.986	.	.	23.9	.	0.6532	
15	MONCLOA-ARAVACA	BARAJAS	2	0.0049	.981	.	.	22.5	.	0.7685	
14	CL16	SAN BLAS-CANILLEJAS	3	0.0076	.974	.	.	20.0	2.1	0.8245	
13	SALAMANCA	CHAMBERI	2	0.0065	.967	.	.	19.7	.	0.8857	
12	CL20	VILLAVERDE	3	0.0090	.958	.	.	18.8	9.0	0.8993	
11	CL13	CHAMARTIN	3	0.0102	.948	.	.	18.3	1.6	0.9563	
10	CL12	CL17	5	0.0223	.926	.	.	15.2	5.0	1.0561	
9	RETIRO	FUENCARRAL-EL PARDO	2	0.0099	.916	.	.	16.3	.	1.0883	
8	CL18	CL9	4	0.0244	.892	.	.	15.3	3.7	1.2101	
7	CL19	CIUDAD LINEAL	3	0.0164	.875	.	.	16.3	6.7	1.215	
6	CL7	CL11	6	0.0418	.833	.	.	15.0	4.7	1.2936	
5	CL8	CL15	6	0.0660	.767	.	.	13.2	6.2	1.7239	
4	CL5	CL14	9	0.1060	.661	.712	-1.4	11.1	6.2	1.7829	
3	CL4	CL10	14	0.2200	.441	.551	-2.0	7.1	10.1	2.0267	
2	CL6	CL3	20	0.2335	.208	.333	-2.1	5.0	7.5	1.8265	
1	CL2	VILLA DE VALLECAS	21	0.2077	.000	.000	0.00	.	5.0	3.6175	

Tabla 3.23

En la *tabla 3.23* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

En primer lugar, vamos a analizar el R^2 y el R^2 semiparcial. En el caso del R^2, vamos a considerar valores altos del R^2 hasta el 70%, por lo que podríamos elegir como mínimo 5 Clusters, ya que si elegimos menos, la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por otro lado, atendiendo al R^2 semiparcial, podemos observar que con 8 Clusters tenemos un salto, por lo que deberíamos quedarnos con 9 Clusters. Otro de los saltos se observa con 4 Clusters, por este motivo nos deberíamos quedar con 5 Clusters, para no perder tanta homogeneidad al pasar a 4. Finalmente, teniendo en cuenta estos dos criterios, deberíamos quedarnos con 5 Clusters, ya que es el menor número de Clusters óptimo en ambos métodos.

Además, si analizamos la Pseudo F, tal y como podemos ver en la tabla anterior *(tabla 3.23)*, tenemos un máximo relativo del valor de este estadístico con 7 Clusters, por lo que según este criterio nos deberíamos quedar con 7 Clusters.

Y por último, miramos el Pseudo test de la T^2. En nuestro caso, tenemos dos máximos relativos, uno de ellos con 7 Clusters, por lo que deberíamos quedarnos con 8; y el otro máximo relativo se observa con 3 Clusters, por lo que deberíamos quedarnos con 4. Sin embargo, hay que destacar que con 4 Clusters este estadístico toma el mismo valor que con 5 Clusters, por lo que podríamos quedarnos con 5, al igual que habían determinado los dos primeros criterios.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de clusters adecuado para este estudio debe ser 5.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma *(figura 3.8)*, que se muestra a continuación:

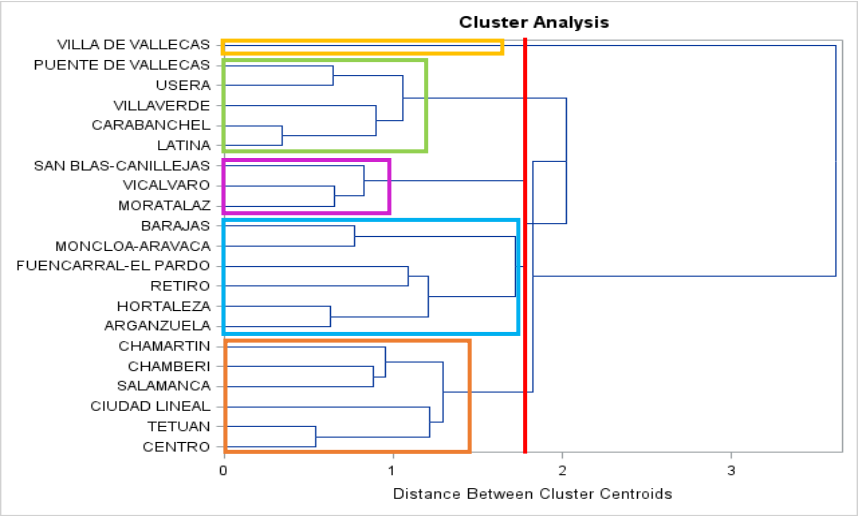

Figura 3.8

A partir de esta representación, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Pero como ya habíamos determinado anteriormente, el número adecuado de Clusters es 5, por lo que deberíamos detener el proceso de agrupación cuando tengamos 5

Clusters, y por consiguiente, no llevaremos a cabo las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos parado. Además, a partir del dendrograma anterior *(figura 3.8)*, podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, vamos a analizar estos 5 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en los 3 Factores que hemos utilizado.

CLUSTER=1

Obs	Distritos	Factor1	Factor2	Factor3
1	LATINA	0.61927	0.40211	0.01681
2	CARABANCHEL	0.94411	0.29402	0.07226
3	USERA	1.52940	0.84761	-0.02004
4	PUENTE DE VALLECAS	1.33876	1.46009	0.00419
5	VILLAVERDE	1.49899	-0.14955	-0.17132

Tabla 3.24

Como podemos observar en la *tabla 3.24*, el Cluster 1 está formado por los distritos Latina, Carabanchel, Usera, Puente de Vallecas y Villaverde. Pertenecen al mismo Cluster, por tener valores altos del Factor 1 y del Factor 2. Por lo tanto, todos ellos se caracterizan por tener un nivel bajo de estudios de la población mayor de 25 años, y también una proporción elevada de alumnos con necesidades de apoyo educativo, sobre todo en Usera y Puente de Vallecas. Por otro lado, también se caracterizan por tener una alta proporción de alumnos extranjeros.

CLUSTER=2

Obs	Distritos	Factor1	Factor2	Factor3
6	CENTRO	-0.55336	1.32763	0.27527
7	TETUAN	-0.02733	1.26573	0.15977
8	SALAMANCA	-1.60836	0.94808	-0.31916
9	CHAMBERI	-1.13617	1.07491	0.41930
10	CHAMARTIN	-1.47492	0.06904	-0.07520
11	CIUDAD LINEAL	-0.26698	1.32824	-0.99685

Tabla 3.25

El Cluster 2 está compuesto por 6 distritos *(tabla 3.25)*, que se caracterizan por tomar valores negativos del Factor 1 y positivos del Factor 2. Todos tienen una proporción elevada de alumnos extranjeros y de alumnos con necesidades de apoyo educativo en centros privados concertados. Además, la población mayor de 25 años se caracteriza por tener un nivel de estudios muy alto.

CLUSTER=3

Obs	Distritos	Factor1	Factor2	Factor3
12	ARGANZUELA	-0.51332	-0.64481	0.31173
13	HORTALEZA	-0.59585	-0.36607	-0.24762
14	MONCLOA-ARAVACA	-0.79002	-0.46119	-1.09082
15	BARAJAS	-0.74723	-1.19247	-1.32303
16	RETIRO	-0.90123	-1.61899	0.66236
17	FUENCARRAL-EL PARDO	-0.55724	-0.82404	1.32122

Tabla 3.26

El Cluster 3 está formado por los 6 distritos que se indican en la *tabla 3.26*. Estos 6 distritos pertenecen al mismo Cluster por tener un valor negativo tanto del Factor 1 como del Factor 2. En primer lugar, se caracterizan por tener una proporción elevada de alumnos extranjeros en centros públicos, exceptuando Hortaleza. Por otro lado, se caracterizan por tener un alto nivel de estudios en la población mayor de 25 años.

CLUSTER=4

Obs	Distritos	Factor1	Factor2	Factor3
18	MORATALAZ	0.65461	-1.35947	-0.83568
19	VICALVARO	1.27929	-1.33970	-0.64575
20	SAN BLAS-CANILLEJAS	0.84328	-0.54871	-0.89254

Tabla 3.27

Por otro lado, el Cluster 4 está formado por 3 distritos, Moratalaz, Vicálvaro y San Blas-Canillejas *(tabla 3.27)*. Se caracterizan por tener un valor positivo del Factor 1 y negativo del Factor 2. Esto quiere decir que tienen una proporción elevada de población en etapas educativas: Primaria (6 y 11 años). Por otro lado, también tienen una proporción elevada de alumnos en centros públicos, incluyendo tanto los alumnos extranjeros como los alumnos con necesidades de apoyo educativo. Además, son de los distritos que menos nivel de estudios tienen en la población mayor de 25 años, aunque sin llegar a ser los que menos nivel tienen.

CLUSTER=5

Obs	Distritos	Factor1	Factor2	Factor3
21	VILLA DE VALLECAS	0.46429	-0.51244	3.37510

Tabla 3.28

Y por último, tenemos al Cluster 5 compuesto únicamente por el distrito Villa de Vallecas *(tabla 3.28)*, esto es así porque tiene un valor muy alto del Factor 3, y también un valor positivo aunque no muy alto del Factor 1. Por lo tanto, Villa de Vallecas está caracterizado por tener una proporción muy elevada de población en etapas educativas: Infantil (3 a 5 años), así como una proporción elevada de alumnos con necesidades de apoyo educativo en centros públicos.

A continuación, a partir del Análisis Factorial que hicimos anteriormente, podemos llevar a cabo una representación gráfica de los distritos en el espacio de los dos primeros Factores, y por otro lado del primer y tercer Factor, donde se pueden observar los Clusters creados:

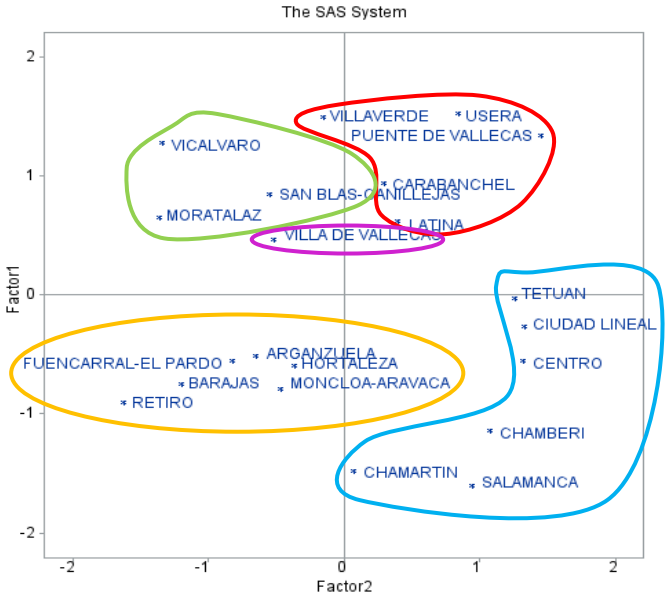

Figura 3.9

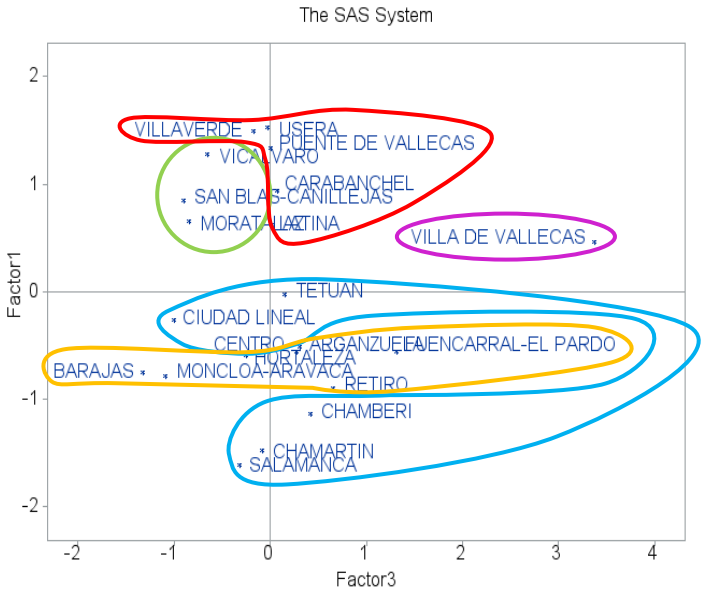

Figura 3.10

Como hemos señalado en los dos gráficos *(figuras 3.9 y 3.10)*, podemos ver los 5 Clusters que hemos creado anteriormente. Se puede apreciar como los distritos que pertenecen a un mismo Cluster se encuentran próximos entre sí como era de esperar.

3.2.4 Salud y Servicios Sociales

En este grupo vamos a introducir las variables relacionadas con la salud, donde incluimos los hábitos y estilos de vida (práctica de ejercicio físico diario, consumo de tabaco diario y consumo de medicamentos), y la proporción de personas con grado de discapacidad reconocido. En relación con los servicios sociales incluiremos las personas atendidas en la Unidad de Primera Atención, los perceptores de prestación de la Renta Mínima de Inserción y los beneficiarios de prestaciones sociales de carácter económico, y en cuanto a las personas mayores incluiremos aquellas con Servicio de Ayuda a Domicilio y las socias de los Centros Municipales de Mayores. Por otra parte, también vamos a incluir los servicios sociales municipales, dentro de ellos se incluyen centros de servicios sociales, centros municipales de mayores, centros de día de Alzheimer y Físicos, apartamentos municipales para mayores, residencias de mayores y centros de atención a la infancia. Por lo tanto, una vez eliminadas las variables que son combinación lineal directa de otras, contamos con 15 variables y 21 observaciones.

En primer lugar, debido al elevado número de variables respecto al número de observaciones, vamos a proceder a reducir la dimensión mediante el análisis Factorial. Seguidamente, comprobaremos si las variables que hemos incluido están altamente correlacionadas. Para ello, evaluamos el índice KMO y la medida de adecuación MSA.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.70645735														
Practicadeejerciciofisicodiario	**Consumodetabacodiario**	**Consumodemedicamentos**	**PropPersonasDiscapacidadReconoci**	**PropPersonasAtendidasUnidaddePri**	**PerceptoresPrestacionRentaMinima**	**BeneficiariosPrestacionesSociale**	**PropPersonasMayoresServicioAyuda**	**PropPersonasMayoresSociasCentros**	**CentrosdeServiciosSociales**	**CentrosMunicipalesMayores**	**CentrosDiaAlzheimerYFisicos**	**ApartamentosMunicipalesMayores**	**ResidenciasMayores**	**CentrosAtencionInfancia**
0.7487	0.1874	0.7070	0.6129	0.6886	0.6666	0.8119	0.8051	0.7317	0.9064	0.7360	0.9044	0.2040	0.5521	0.7236

Tabla 3.29

Como vemos en la *tabla 3.29* el índice KMO toma un valor superior a 0.7, que es un valor muy bueno, por lo que no eliminaremos ninguna variable, aunque tengamos dos valores MSA pequeños. Por lo tanto, estas serán las variables con las que realizaremos el Análisis Factorial.

A continuación, obtenemos los autovalores de la matriz de Correlaciones, con su correspondiente proporción de varianza explicada y su proporción acumulada para ayudar a decidir con cuantos Factores nos quedamos *(tabla 3.30).*

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	7.48801921	5.61807652	0.5349	0.5349
2	1.86994270	0.57484630	0.1336	0.6684
3	1.29509639	0.45415964	0.0925	0.7609
4	0.84093676	0.17081383	0.0601	0.8210
5	0.67012293	0.13174884	0.0479	0.8689
6	0.53837409	0.06443562	0.0385	0.9073
7	0.47393846	0.15152226	0.0339	0.9412
8	0.32241620	0.13717021	0.0230	0.9642
9	0.18524600	0.06470388	0.0132	0.9774
10	0.12054211	0.03168533	0.0086	0.9860
11	0.08885678	0.03231460	0.0063	0.9924
12	0.05654218	0.02659444	0.0040	0.9964
13	0.02994774	0.00992928	0.0021	0.9986
14	0.02001846		0.0014	1.0000

Tabla 3.30

En primer lugar, vemos que a partir de 3 Factores se explica más del 75% de la variabilidad, y con 6 superamos el 90%. Por otro lado, nos deberíamos de quedar con 3 Factores, ya que son aquellos cuyos autovalores son mayores que la unidad. Además, estos 3 Factores explican el 76.09% de la variabilidad, que es un valor aceptable, por lo tanto, nos quedamos con 3 Factores.

A continuación, se muestra la matriz de saturaciones en la cual se reflejan las correlaciones entre las variables y los 3 Factores que hemos retenido *(tabla 3.31)*. En ella, vamos a señalar que Factor es el que está más correlacionado con cada una de las variables, para así de este modo obtener una interpretación de estos Factores.

Factor Pattern			
	Factor1	**Factor2**	**Factor3**
Habitos y estilos de vida: Practica de ejercicio fisico diario	-0.65099	0.52962	0.45226
Habitos y estilos de vida: Consumo de tabaco diario	0.02113	-0.59019	-0.60638
Habitos y estilos de vida: Consumo de medicamentos	0.86907	-0.30702	-0.00913
Proporcion de personas con grado de discapacidad reconocido	0.60352	-0.33971	0.50957
Proporcion de personas atendidas en la Unidad de Primera Atencion en Centros de Servicios Sociales	0.83466	0.23735	0.09274
Proporcion de perceptores de prestacion de la Renta Minima de Insercion	0.85459	-0.01858	0.31622
Proporcion de beneficiarios de prestaciones sociales de caracter economico	0.92411	0.12331	-0.01219
Proporcion de personas mayores con Servicio de Ayuda a Domicilio (modalidad auxiliar de hogar)	0.90264	0.28800	-0.10217
Proporcion de personas mayores socias de los Centros Municipales de Mayores	0.72035	0.38652	-0.31070
Proporcion de Centros de Servicios Sociales	0.87016	0.11526	0.09029
Proporcion de Centros Municipales de Mayores	0.70999	0.19286	-0.15592
Proporcion de Centros de Dia de Alzheimer y Fisicos	0.67901	0.30276	0.09307
Proporcion de Apartamentos Municipales para Mayores	-0.16724	-0.27068	0.66115
Proporcion de Residencias de mayores	-0.27623	0.68559	-0.15697
Proporcion de Centros de Atencion a la Infancia (CAI)	0.70954	-0.42868	0.03302

Tabla 3.31

Debido a que estos factores no son fácilmente interpretables, una solución es rotarlos para conseguir una estructura más sencilla de interpretar. En nuestro caso, debido a que el número de factores es reducido, se recomienda usar la rotación varimax, que es la que se muestra a continuación *(tabla 3.32)*:

Rotated Factor Pattern			
	Factor1	**Factor2**	**Factor3**
Habitos y estilos de vida: Practica de ejercicio fisico diario	-0.41198	-0.85469	-0.09277
Habitos y estilos de vida: Consumo de tabaco diario	-0.19403	0.81160	-0.14185
Habitos y estilos de vida: Consumo de medicamentos	0.69319	0.48287	0.36871
Proporcion de personas con grado de discapacidad reconocido	0.42993	0.11344	0.73593
Proporcion de personas atendidas en la Unidad de Primera Atencion en Centros de Servicios Sociales	0.86231	0.00940	0.13393
Proporcion de perceptores de prestacion de la Renta Minima de Insercion	0.78391	0.06731	0.46002
Proporcion de beneficiarios de prestaciones sociales de caracter economico	0.90398	0.18340	0.13610
Proporcion de personas mayores con Servicio de Ayuda a Domicilio (modalidad auxiliar de hogar)	0.94592	0.11076	-0.03314
Proporcion de personas mayores socias de los Centros Municipales de Mayores	0.81503	0.11287	-0.29636
Proporcion de Centros de Servicios Sociales	0.84999	0.11115	0.20921
Proporcion de Centros Municipales de Mayores	0.73220	0.15810	-0.06686
Proporcion de Centros de Dia de Alzheimer y Fisicos	0.74203	-0.08397	0.06097
Proporcion de Apartamentos Municipales para Mayores	-0.26151	-0.25262	0.63730
Proporcion de Residencias de mayores	-0.00078	-0.48929	-0.57582
Proporcion de Centros de Atencion a la Infancia (CAI)	0.49955	0.50074	0.43360

Tabla 3.32

Una vez realizada la rotación obtenemos la matriz de saturaciones anterior, en la que se puede observar que el Factor 1 está correlacionado con la mayoría de las variables, algo esperable ya que siempre el Factor 1 está relacionado con la mayor parte de ellas. El Factor 2 esta correlacionado con la práctica de ejercicio físico diario, el consumo de tabaco diario y los Centros de Atención a la Infancia. Y por último, el Factor 3 esta correlacionado con la proporción de personas con grado de discapacidad reconocido y con la proporción de apartamentos municipales y de residencias para mayores. A partir de estas correlaciones, podemos interpretar los Factores anteriores:

- Factor 1: este primer factor nos da información relativa a las personas con necesidades de asistencia social, y a los mayores con servicio de ayuda a domicilio y aquellos socios de los Centros Municipales de Mayores. También hace referencia a la proporción de centros de servicios sociales, centros municipales de mayores y centros de día de Alzheimer y Físicos. Y por último, atendiendo a los hábitos y estilos de vida, hace referencia al consumo de medicamentos. Hay destacar que todas estas variables influyen de forma positiva en el Factor.
- Factor 2: este factor hace referencia a la práctica de ejercicio físico diario que influye de forma negativa, y al consumo de tabaco diario que influye de forma positiva. Por otro lado, también hace referencia a la proporción de Centros de Atención a la Infancia, que influye de forma positiva, aunque en menor medida.
- Factor 3: el tercer Factor hace referencia a la proporción de personas con grado de discapacidad reconocido influyendo positivamente. Por otro lado, también hace referencia a los apartamentos municipales para mayores, que influyen de forma positiva, y a las residencias de mayores, influyendo de forma negativa.

Las correlaciones que existen entre las variables y todos los Factores se pueden observar mediante los siguientes gráficos, que representan las variables en el espacio de los dos primeros Factores, y por otro lado del primer y tercer Factor:

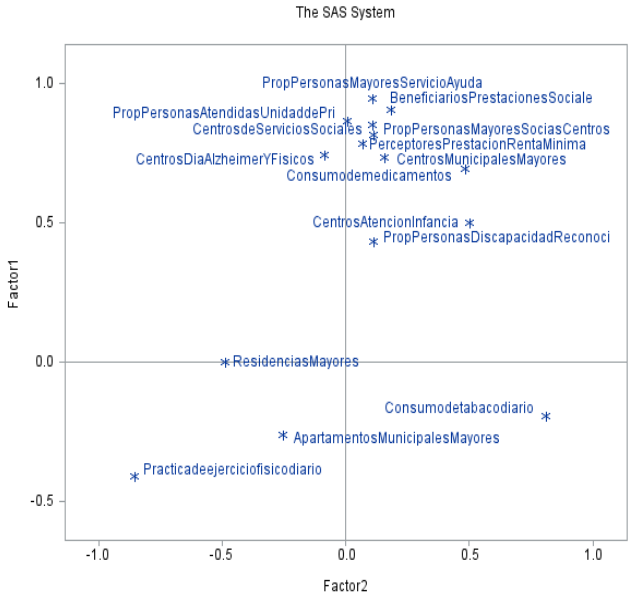

Figura 3.11

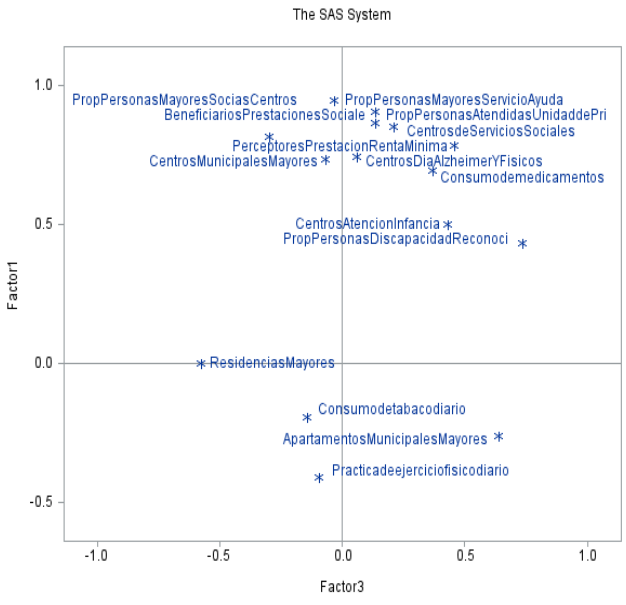

Figura 3.12

En los gráficos anteriores podemos ver que el Factor 1 se representa en el eje vertical y los Factores 2 y 3 se representan en el eje horizontal *(figuras 3.11 y 3.12)*. Por lo tanto, las variables que se encuentren en la parte superior o derecha de los gráficos tendrán una correlación positiva con el Factor correspondiente, y las que se encuentren en la parte inferior o izquierda tendrán una correlación negativa. Finalmente, las relaciones que se pueden observar son las mismas que habíamos comentado anteriormente.

Por último, para medir la bondad del ajuste realizado, es decir, para medir el grado en que cada variable viene explicada por los factores vamos a analizar las comunalidades, ya que estas expresan la parte de cada variable que puede ser explicada por estos 3 Factores.

Final Communality Estimates: Total = 11.051986														
Practicadeejerciciofisicodiario	**Consumodetabacodiario**	**Consumodemedicamentos**	**PropPersonasDiscapacidadReconoci**	**PropPersonasAtendidasUnidaddePri**	**PerceptoresPrestacionRentaMinima**	**BeneficiariosPrestacionesSociale**	**PropPersonasMayoresServicioAyuda**	**PropPersonasMayoresSociasCentros**	**CentrosdeServiciosSociales**	**CentrosMunicipalesMayores**	**CentrosDiaAlzheimerYFisicos**	**ApartamentosMunicipalesMayores**	**ResidenciasMayores**	**CentrosAtencionInfancia**
0.9088	0.7164	0.8496	0.7393	0.7616	0.8306	0.8693	0.9081	0.7648	0.7786	0.5655	0.5613	0.5383	0.5709	0.6883

Tabla 3.33

En la *tabla 3.33*, nos aparece en primer lugar la suma de los autovalores, y seguidamente las comunalidades de cada variable. Como podemos observar todas ellas toman valores superiores a 0.5, lo que quiere decir que más del 50% de cada variable está explicada por los 3 Factores, lo cual se considera un valor aceptable. Por lo tanto, estas variables están bien explicadas por los Factores y el análisis Factorial se considera correcto.

A continuación, a partir de los Factores obtenidos realizaremos un análisis Cluster, para formar grupos de distritos con características similares. Vamos a utilizar un análisis Cluster Jerárquico, ya que no conocemos el número de grupos que queremos formar. Por lo tanto determinaremos el número adecuado de grupos para obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para ello, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Centroid Distance**	**Tie**
20	**VILLA DE VALLECAS**	**VICALVARO**	2	0.0000	1.00	.	.	5973	.	0.0325	
19	**ARGANZUELA**	**MORATALAZ**	2	0.0004	1.00	.	.	250	.	0.2288	
18	**USERA**	**VILLAVERDE**	2	0.0005	.999	.	.	183	.	0.2496	
17	**SALAMANCA**	**CHAMBERI**	2	0.0005	.999	.	.	168	.	0.2504	
16	**TETUAN**	**SAN BLAS-CANILLEJAS**	2	0.0008	.998	.	.	146	.	0.3091	
15	**CL19**	**MONCLOA-ARAVACA**	3	0.0021	.996	.	.	97.2	4.8	0.4355	
14	**CHAMARTIN**	**BARAJAS**	2	0.0032	.992	.	.	70.3	.	0.6204	
13	**FUENCARRAL-EL PARDO**	**HORTALEZA**	2	0.0035	.989	.	.	59.5	.	0.6462	
12	**CARABANCHEL**	**PUENTE DE VALLECAS**	2	0.0044	.985	.	.	52.1	.	0.7256	
11	**CL15**	**CL20**	5	0.0148	.970	.	.	32.1	17.3	0.859	
10	**CL16**	**CL13**	4	0.0135	.956	.	.	26.7	6.3	0.9007	
9	**CL17**	**CL14**	4	0.0184	.938	.	.	22.6	9.9	1.05	
8	**CL11**	**CL10**	9	0.0431	.895	.	.	15.8	8.6	1.0791	
7	**LATINA**	**CL18**	3	0.0174	.877	.	.	16.7	33.6	1.2522	
6	**CL7**	**CL12**	5	0.0411	.836	.	.	15.3	5.5	1.4343	
5	**RETIRO**	**CL9**	5	0.0457	.790	.	.	15.1	6.2	1.8515	
4	**CL8**	**CL6**	14	0.2564	.534	.712	-4.1	6.5	21.7	2.1875	
3	**CL4**	**CL5**	19	0.3128	.221	.551	-5.1	2.6	11.4	2.2569	
2	**CL3**	**CIUDAD LINEAL**	20	0.0892	.132	.333	-3.2	2.9	2.1	2.3729	
1	**CENTRO**	**CL2**	21	0.1322	.000	.000	0.00	.	2.9	2.8861	

Tabla 3.34

En la *tabla 3.34* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

Comenzaremos analizando el R^2 considerando valores altos hasta el 70%, por lo que podríamos elegir como mínimo 5 Clusters, ya que si elegimos menos la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por lo tanto, teniendo en cuenta este criterio deberíamos quedarnos con 5 Clusters.

En segundo lugar, analizaremos el R^2 semiparcial, en el cual observamos que con 8 y 6 Clusters tenemos un salto, por lo que deberíamos quedarnos con 9 y 7 Clusters respectivamente. Otro de los saltos, mayor que los anteriores, se observa con 4 Clusters, por este motivo nos deberíamos quedar con 5, para no perder tanta homogeneidad al pasar a 4. Finalmente, teniendo en cuenta este criterio y comparándolo con el anterior, deberíamos quedarnos con 5 Clusters, ya que es el menor número de Clusters óptimo en ambos métodos.

A continuación, si analizamos la Pseudo F podemos ver en la *tabla 3.34* que tenemos un máximo relativo del valor de este estadístico con 7 Clusters, por lo que según este criterio, nos deberíamos quedar con 7 Clusters.

Y por último, vemos que el Pseudo test de la T2 tiene tres máximos relativos. El más pequeño se da con 9 Clusters, por lo que deberíamos quedarnos con 10. Otro de ellos se da con 8 Clusters, y por consiguiente, deberíamos quedarnos con 9. Y por último, el otro máximo relativo se observa con 4 Clusters, por lo que deberíamos quedarnos con 5. Por lo tanto, al igual que concluían los dos primeros métodos, deberíamos quedarnos con 5 Clusters.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número adecuado de Clusters para este estudio debe ser 5.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma *(figura 3.13)*, que se muestra a continuación:

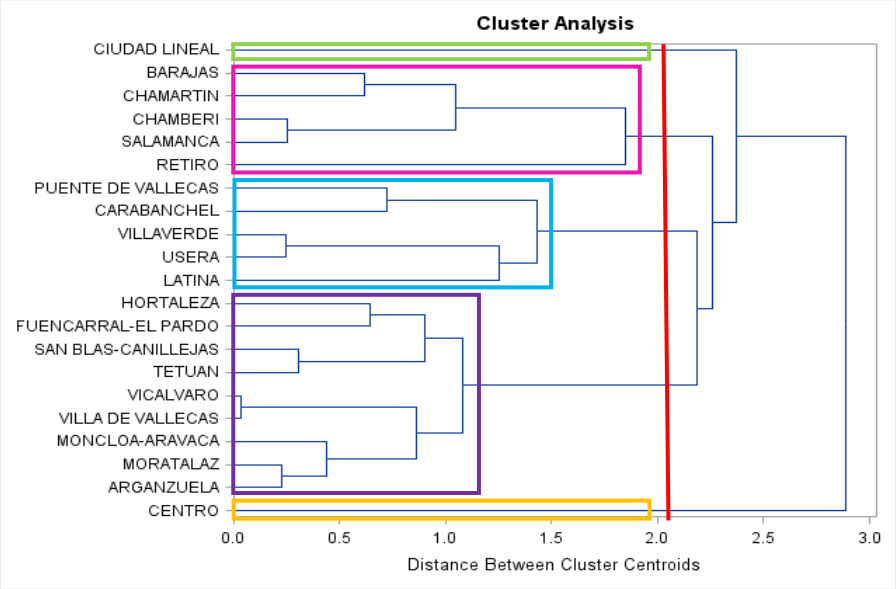

Figura 3.13

A partir de esta representación, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Pero como ya habíamos determinado anteriormente, el número adecuado de Clusters es 5, por lo que deberíamos detener el proceso en ese momento, y por consiguiente no realizaremos las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos parado. Además, a partir del dendrograma *(figura 3.13)* podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, vamos a analizar estos 5 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en los 3 Factores que hemos utilizado.

CLUSTER=1

Obs	Distritos	Factor1	Factor2	Factor3
1	VILLA DE VALLECAS	-0.96797	1.17129	0.40608
2	VICALVARO	-0.96241	1.20333	0.40636
3	ARGANZUELA	-0.76691	0.71033	-0.22748
4	MORATALAZ	-0.67162	0.72526	-0.02005
5	TETUAN	0.08976	0.75935	0.07764
6	SAN BLAS-CANILLEJAS	0.27439	0.74705	-0.16991
7	MONCLOA-ARAVACA	-0.59807	0.75647	-0.54032
8	FUENCARRAL-EL PARDO	0.41735	0.95685	-1.06650
9	HORTALEZA	-0.07516	1.18036	-0.71292

Tabla 3.35

Como podemos observar en la *tabla 3.35*, el Cluster 1 está formado por 9 distritos. Pertenecen al mismo Cluster por tener valores positivos del Factor 2. Por lo tanto, esto quiere decir que todos ellos se caracterizan por consumir mucho tabaco diario, y la mayoría de ellos por tener la mayor proporción de Centros de Atención a la Infancia.

CLUSTER=2

Obs	Distritos	Factor1	Factor2	Factor3
10	USERA	0.70916	-0.16908	0.86907
11	VILLAVERDE	0.69531	-0.05673	0.64663
12	CARABANCHEL	1.90201	-0.59368	0.39898
13	PUENTE DE VALLECAS	2.20546	-0.78117	1.03081
14	LATINA	1.23807	0.49962	-0.19383

Tabla 3.36

El Cluster 2 está compuesto por 5 distritos, que se caracterizan por tomar valores positivos tanto del Factor 1 como del Factor 3 *(tabla 3.36)*. Todos ellos se caracterizan por tener un alto consumo de medicamentos y una proporción elevada de personas con necesidades de asistencia social. Por otro lado, también se caracterizan por tener una proporción elevada de personas mayores tanto con servicio de ayuda a domicilio como socias de los centros municipales de mayores. Además, son de los distritos con mayor proporción de centros de servicios sociales, centros municipales de mayores y centros de día de Alzheimer y Físicos. Y por último, se caracterizan por ser de los distritos con mayor proporción de personas con grado de discapacidad reconocido.

CLUSTER=3

Obs	Distritos	Factor1	Factor2	Factor3
15	SALAMANCA	-0.48378	-1.77069	-1.07089
16	CHAMBERI	-0.67840	-1.64457	-0.97659
17	CHAMARTIN	-0.66593	-1.15114	-0.31862
18	BARAJAS	-1.24467	-0.92948	-0.29044
19	RETIRO	-1.11352	-1.68101	1.12875

Tabla 3.37

El Cluster 3 está formado por los 5 distritos que se indican en la *tabla 3.37*. Pertenecen al mismo Cluster por tener un valor negativo tanto del Factor 1 como del Factor 2. Estos distritos se caracterizan por tener una proporción muy pequeña de personas con necesidades de asistencia social y de personas mayores con servicio de ayuda a domicilio. Y por último, son los distritos en donde se práctica más ejercicio físico diario.

CLUSTER=4

Obs	Distritos	Factor1	Factor2	Factor3
20	CIUDAD LINEAL	1.15659	-0.093637	-2.08191

Tabla 3.38

Por otro lado, tenemos al Cluster 4 formado únicamente por Ciudad Lineal, esto es así porque tiene un valor positivo del Factor 1 y un valor negativo del Factor 3 *(tabla 3.38)*. Por lo tanto, Ciudad Lineal está caracterizado por tener una proporción elevada de personas mayores socias de los centros municipales de mayores, así como una proporción elevada de centros municipales de mayores. Y por último, se caracteriza por ser uno de los 3 distritos que tienen mayor proporción de residencias de mayores.

CLUSTER=5

Obs	Distritos	Factor1	Factor2	Factor3
21	CENTRO	-0.45965	0.16127	2.70513

Tabla 3.39

Y por último tenemos el Cluster 5 compuesto únicamente por el distrito Centro, esto es así porque tiene un valor negativo del Factor 1 y un valor positivo y muy grande del Factor 3 *(tabla 3.39)*. Por lo tanto, Centro está caracterizado por tener la mayor proporción de personas con grado de discapacidad reconocido, y por ser uno de los dos distritos con mayor proporción de apartamentos municipales de mayores. Por otro lado, también destaca por ser de los distritos con menor proporción de personas mayores socias de los centros municipales de mayores.

A continuación, a partir del Análisis Factorial que hicimos anteriormente, podemos llevar a cabo una representación gráfica de los distritos en el espacio de los dos primeros Factores, y por otro lado de los Factores 2 y 3, donde podemos ver gráficamente los Clusters creados.

Como hemos señalado en los gráficos *(figuras 3.14 y 3.15)*, vemos los 5 Clusters que hemos creado anteriormente. Se puede apreciar como los distritos que pertenecen a un mismo Cluster se encuentran próximos entre sí como era de esperar.

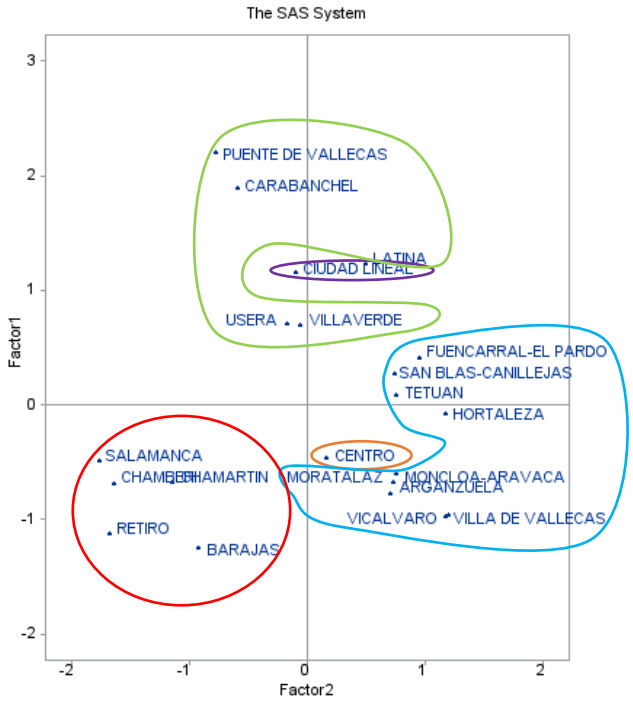

Figura 3.14

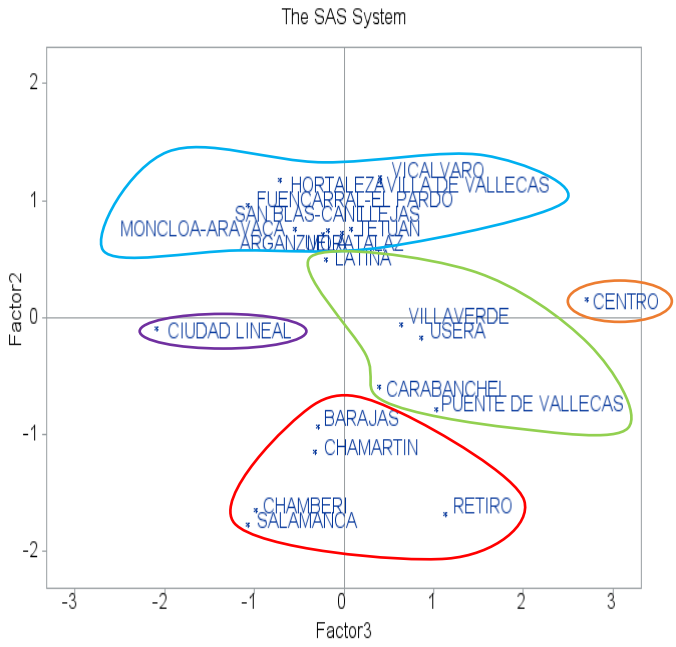

Figura 3.15

3.2.5 Vivienda

En este grupo vamos a incluir las variables relacionadas con las viviendas. En primer lugar, vamos a meter aquellas relacionadas con el estado de las viviendas. Además, incluiremos el valor catastral de los bienes inmuebles, tanto de personas físicas como de personas jurídicas, y la superficie media de la vivienda en transacción. Y por último, vamos a incluir las variables relacionadas con la tipología de las viviendas. Por lo tanto, una vez eliminadas las variables que son combinación lineal directa de otras, contamos con 11 variables y 21 observaciones.

En primer lugar, debido a que tenemos un número elevado de variables respecto al número de observaciones, vamos a proceder a reducir la dimensión mediante el análisis Factorial. Seguidamente, vamos a comprobar si las variables que hemos incluido están altamente correlacionadas. Para ello, vamos a evaluar el índice KMO y la medida de adecuación MSA.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.64687094										
Viviendas anteriores a1980	**Estadoruinoso**	**Estadomalo**	**Estadodeficiente**	**Estadobueno**	**ValorCatastralMedioBienesInmuebl**	**ValorCatastralMedioBienesInmueb1**	**SuperficieMediaVivienda**	**Principal**	**Secundaria**	**Desocupada**
0.4637	0.7643	0.7750	0.7804	0.8196	0.6360	0.8451	0.5888	0.5376	0.5900	0.4519

Tabla 3.40

Como vemos en la *tabla 3.40* el índice KMO toma un valor superior a 0.6, que es un valor bueno, por este motivo no debemos eliminar ninguna variable. Por lo tanto, estas van a ser las variables definitivas con las que vamos a llevar a cabo el Análisis Factorial.

A continuación, obtenemos los autovalores de la matriz de Correlaciones, con su correspondiente proporción de varianza explicada y su proporción acumulada para ayudar a decidir con cuantos Factores nos quedamos *(tabla 3.41)*.

Eigenvalues of the Correlation Matrix: Total = 11 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.57576109	0.84548048	0.4160	0.4160
2	3.73028061	2.75322650	0.3391	0.7551
3	0.97705411	0.18396652	0.0888	0.8439
4	0.79308758	0.34735256	0.0721	0.9160
5	0.44573503	0.22687602	0.0405	0.9565
6	0.21885901	0.11360114	0.0199	0.9764
7	0.10525786	0.01988910	0.0096	0.9860
8	0.08536876	0.04961101	0.0078	0.9938
9	0.03575775	0.00668733	0.0033	0.9970
10	0.02907043	0.02530265	0.0026	0.9997
11	0.00376778		0.0003	1.0000

Tabla 3.41

En primer lugar, vemos que a partir de 2 Factores se explica más del 75% de la variabilidad y con 4 superamos el 90%. Por otro lado, nos deberíamos quedar con 2 Factores, ya que son aquellos cuyos autovalores son mayores que la unidad. Además, explican el 75.51% de la variabilidad, cuyo valor es aceptable, por lo tanto nos quedamos con 2 Factores.

A continuación, se muestra la matriz de saturaciones en la que se reflejan las correlaciones entre las variables y los 2 Factores que hemos retenido *(tabla 3.42)*. En ella, vamos a señalar que Factor es el que está más correlacionado con cada una de las variables, para así de este modo obtener una interpretación de estos Factores.

Factor Pattern		
	Factor1	**Factor2**
Estado de las viviendas: Viviendas anteriores a 1980	0.39078	0.51739
Estado de las viviendas: Estado ruinoso	-0.65352	0.33885
Estado de las viviendas: Estado malo	-0.66806	0.62696
Estado de las viviendas: Estado deficiente	-0.71864	0.64739
Estado de las viviendas: Estado bueno	0.79000	-0.55809
Valor catastral medio de los bienes inmuebles: personas fisicas (2016)	0.84240	0.44166
Valor catastral medio de los bienes inmuebles: personas juridicas (2016)	0.66959	0.33328
Superficie media de la vivienda (m2) en transaccion (2016)	0.85365	0.01896
Tipologia de las viviendas: Principal	-0.35565	-0.89075
Tipologia de las viviendas: Secundaria	0.62712	0.72662
Tipologia de las viviendas: Desocupada	-0.05131	0.77214

Tabla 3.42

En la matriz de saturaciones anterior vemos que el Factor 1 está correlacionado con las variables que indican el estado de las viviendas, el valor catastral medio de los bienes inmuebles y la superficie media de la vivienda en transacción. Por otro lado, el Factor 2 esta correlacionado con la proporción de viviendas anteriores a 1980 y con las variables que indican la tipología de las viviendas. A partir de estas correlaciones, podemos interpretar los Factores anteriores:

- Factor 1: este primer Factor nos da información relativa al estado de las viviendas, y cómo podemos observar, este Factor toma valores negativos cuando el estado es ruinoso, malo y deficiente, y por el contrario toma valores positivos cuando el estado es bueno. Por otro lado, cuanto mayor sea el valor catastral medio de los bienes inmuebles, tanto de personas físicas como de personas jurídicas, así como cuanto mayor sea la superficie media de la vivienda en transacción, mayor es el valor de este primer Factor.

- Factor 2: este Factor hace referencia a la proporción de viviendas anteriores a 1980, y además esta variable influye de forma positiva en el Factor. Por otro lado, atendiendo a la tipología de las viviendas, cuando la vivienda es principal el factor toma un valor negativo, sin embargo cuando es secundaria o está desocupada toma valores positivos.

Las correlaciones que existen entre las variables y los dos primeros Factores se pueden observar en el siguiente gráfico, que representa las variables en el espacio de los dos primeros Factores:

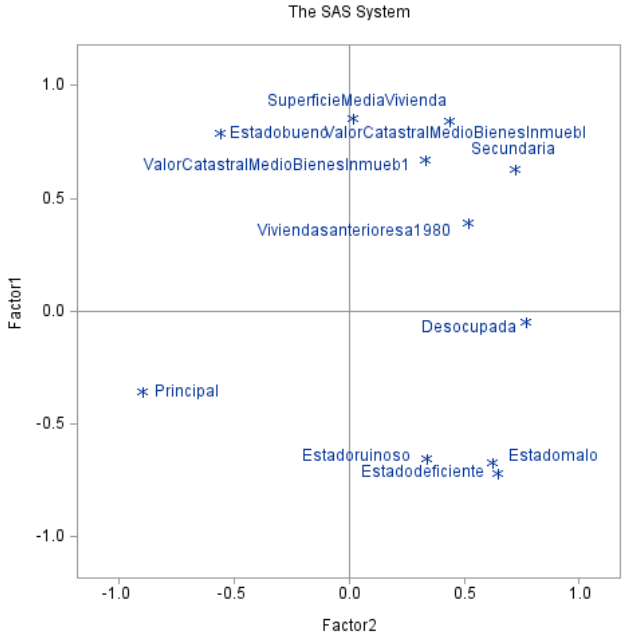

Figura 3.16

En el gráfico anterior *(figura 3.16)* podemos ver que el Factor 1 se representa en el eje vertical y el Factor 2 en el eje horizontal. Por lo tanto, las variables que se encuentren en la parte superior del gráfico tendrán una correlación positiva con el Factor 1, y por el contrario aquellas que se encuentren en la parte inferior tendrán una correlación negativa. De igual manera, las variables situadas en la parte derecha tendrán una correlación positiva con el Factor 2, y las que se encuentren en la parte izquierda tendrán una correlación negativa. Finalmente, las relaciones que se pueden observar son las mismas que habíamos comentado anteriormente.

Por último, para medir la bondad del ajuste realizado, es decir, para medir el grado en que cada variable viene explicada por los factores vamos a mirar las comunalidades, ya que estas expresan la parte de cada variable que puede ser explicada por estos Factores.

Final Communality Estimates: Total = 8.306042										
Viviendasanterioresa1980	**Estadoruinoso**	**Estadomalo**	**Estadodeficiente**	**Estadobueno**	**ValorCatastralMedioBienesInmuebl**	**ValorCatastralMedioBienesInmueb1**	**SuperficieMediaVivienda**	**Principal**	**Secundaria**	**Desocupada**
0.4203	0.5419	0.8393	0.9355	0.9355	0.9047	0.5594	0.7290	0.9199	0.9212	0.5988

Tabla 3.43

En la *tabla 3.43*, nos aparece en primer lugar la suma de los autovalores, y seguidamente las comunalidades de cada variable. Como podemos observar la mayoría de ellas toman valores

muy grandes, lo que quiere decir que la mayor parte de la variabilidad de cada variable puede ser explicada por los 2 Factores. Sin embargo, existen algunas variables que toman valores en torno a 0.4 o 0.5, lo que quiere decir que únicamente el 40-50% de la variabilidad de dichas variables puede ser explicada por los Factores, pero este valor se puede considerar aceptable. Por lo tanto, estas variables están bien explicadas por los Factores y el análisis Factorial se considera correcto.

A continuación, con los Factores obtenidos realizaremos un análisis Cluster, con el objetivo de formar grupos de distritos con características similares.

En este caso, vamos a utilizar un análisis Cluster Jerárquico, debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
20	LATINA	VICALVARO	2	0.0001	1.00	.	.	565	.	0.0863	
19	CHAMARTIN	MONCLOA-ARAVACA	2	0.0001	1.00	.	.	588	.	0.0875	
18	CIUDAD LINEAL	HORTALEZA	2	0.0001	1.00	.	.	619	.	0.0877	
17	CL20	SAN BLAS-CANILLEJAS	3	0.0014	.998	.	.	151	14.6	0.2861	
16	CL19	CHAMBERI	3	0.0023	.996	.	.	83.1	24.5	0.375	
15	RETIRO	FUENCARRAL-EL PARDO	2	0.0024	.994	.	.	67.0	.	0.4344	
14	CL15	BARAJAS	3	0.0034	.990	.	.	54.8	1.4	0.4497	
13	SALAMANCA	CL16	4	0.0046	.986	.	.	46.0	3.8	0.4941	
12	CARABANCHEL	USERA	2	0.0034	.982	.	.	45.4	.	0.521	
11	MORATALAZ	VILLA DE VALLECAS	2	0.0034	.979	.	.	46.4	.	0.523	
10	CL17	CL12	5	0.0084	.970	.	.	40.2	5.2	0.5296	
9	CL14	CL18	5	0.0095	.961	.	.	36.9	4.9	0.5633	
8	PUENTE DE VALLECAS	VILLAVERDE	2	0.0041	.957	.	.	41.2	.	0.5734	
7	ARGANZUELA	TETUAN	2	0.0113	.946	.	.	40.5	.	0.9518	
6	CL9	CL11	7	0.0387	.907	.	.	29.2	10.3	1.0415	
5	CL6	CL10	12	0.0706	.836	.	.	20.4	10.0	0.9838	
4	CL7	CL8	4	0.0311	.805	.822	-.52	23.4	4.0	1.1157	
3	CL4	CL5	16	0.2516	.554	.731	-3.4	11.2	18.7	1.8314	
2	CL3	CL13	20	0.3515	.202	.444	-3.1	4.8	14.2	2.0963	
1	CENTRO	CL2	21	0.2020	.000	.000	0.00	.	4.8	2.9126	

Tabla 3.44

En la *tabla 3.44* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

En primer lugar, vamos a analizar el R^2. En nuestro caso consideraremos valores altos del R^2 hasta el 70%, por lo que podríamos elegir como mínimo 4 Clusters, ya que si elegimos menos, la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por lo tanto, teniendo en cuenta este criterio deberíamos quedarnos con 4 Clusters.

En segundo lugar, si analizamos el R^2 semiparcial, podemos observar que con 7 Clusters tenemos un salto, por lo que deberíamos quedarnos con 8 Clusters. Otro de los saltos que se observa, mayor que el anterior, se da con 3 Clusters por este motivo nos deberíamos quedar con 4, para no perder tanta homogeneidad al pasar a 3. Finalmente, teniendo en cuenta este criterio y comparándolo con el anterior, deberíamos quedarnos con 4 Clusters, ya que es el menor número de Clusters óptimo en ambos métodos.

Seguidamente, también vamos a analizar la Pseudo F. Tal y como podemos ver en la tabla anterior, tenemos tres máximos relativos del valor de este estadístico. Los dos primeros, se dan con 11 y 8 Clusters, por lo que sería preferible un número menor de Clusters. Por consiguiente, el otro máximo relativo del valor de este estadístico con el que nos encontramos se da con 4 Clusters, por lo que según este criterio nos deberíamos quedar con 4, al igual que ocurría con los otros métodos.

Y por último, miramos el Pseudo test de la T^2, en el que podemos ver que únicamente tenemos un máximo relativo, que se da con 3 Clusters, por lo que deberíamos quedarnos con 4. Por lo tanto, al igual que concluían los otros métodos, deberíamos quedarnos con 4 Clusters.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de Clusters adecuado para este estudio debe ser 4.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma, que es la que se muestra a continuación:

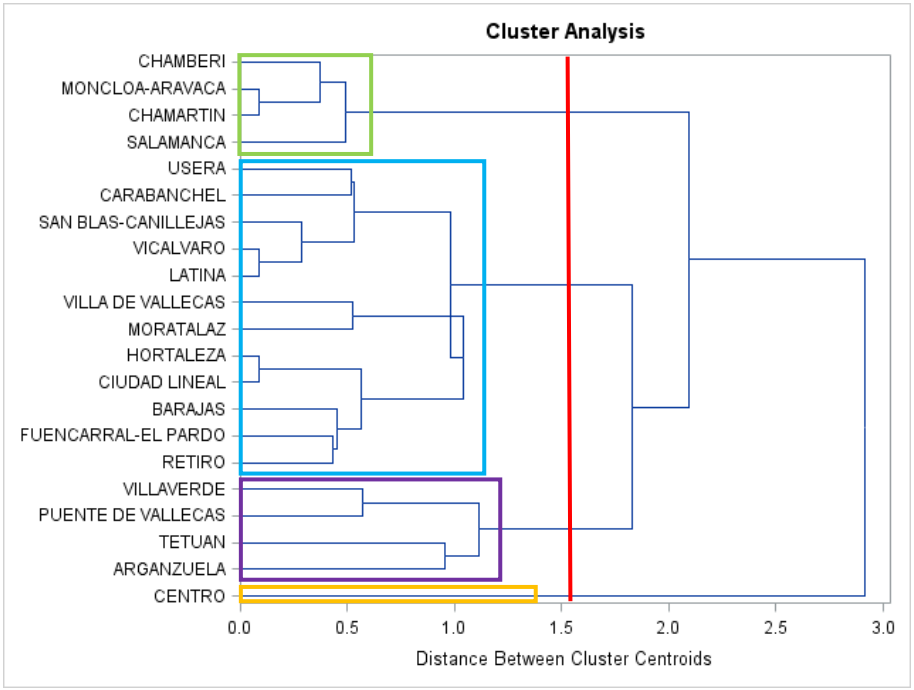

Figura 3.17

A partir de esta representación *(figura 3.17)*, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Pero como ya habíamos determinado anteriormente, el número

adecuado de Clusters es 4, por lo que deberíamos detener el proceso de agrupación cuando tengamos 4 Clusters, y por consiguiente no llevaremos a cabo las agrupaciones posteriores a la línea roja que indica el límite donde nos hemos parado. Además, a partir del dendrograma anterior, podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, vamos a analizar estos 4 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en los 2 Factores que hemos utilizado.

CLUSTER=1

Obs	Distritos	Factor1	Factor2
1	LATINA	-0.47699	-0.48490
2	VICALVARO	-0.45278	-0.56778
3	CIUDAD LINEAL	0.20701	-0.60912
4	HORTALEZA	0.13841	-0.55444
5	SAN BLAS-CANILLEJAS	-0.25997	-0.72603
6	RETIRO	0.97871	-0.66525
7	FUENCARRAL-EL PARDO	0.58613	-0.85111
8	BARAJAS	0.63995	-0.33161
9	CARABANCHEL	-0.65749	-0.04585
10	USERA	-0.95746	-0.47184
11	MORATALAZ	0.18574	-1.45992
12	VILLA DE VALLECAS	-0.33723	-1.46701

Tabla 3.45

El Cluster 1 está formado por 12 distritos, siendo el más numeroso *(tabla 3.45)*. Estos distritos pertenecen al mismo Cluster por tener valores intermedios del Factor 1 y negativos del Factor 2. Por lo tanto, esto quiere decir que todos estos distritos se caracterizan por tener un estado bueno de las viviendas, y además, la mayor parte de ellas son principales. Por otro lado, la mayoría de estos distritos tienen un valor catastral medio de los bienes inmuebles inferior a 100000€ en el caso de personas físicas, exceptuando Hortaleza y Retiro, e inferior a 300000€ en el caso de personas jurídicas, exceptuando Retiro, Fuencarral-El Pardo y Barajas. Y por último, podemos destacar que la mayoría tienen una superficie media de vivienda en transacción similar, exceptuando los distritos Hortaleza, Fuencarral-El Pardo y Retiro, cuya superficie media es mayor.

CLUSTER=2

Obs	Distritos	Factor1	Factor2
13	CHAMARTIN	1.43280	0.57394
14	MONCLOA-ARAVACA	1.51774	0.59506
15	CHAMBERI	1.14001	0.75252
16	SALAMANCA	1.77632	0.91199

Tabla 3.46

El Cluster 2 está compuesto por 4 distritos *(tabla 3.46)*, que se caracterizan por tomar valores positivos tanto del Factor 1 como del Factor 2. Todos estos distritos se caracterizan por tener un estado bueno de las viviendas y un alto valor catastral medio de los bienes inmuebles, tanto de personas físicas como de personas jurídicas. Además, también destacan por ser de los distritos que tienen una mayor superficie media de la vivienda en transacción. Por otro lado, tienen una proporción elevada de viviendas anteriores a 1980 y son los distritos que tienen mayor proporción de viviendas secundarias.

CLUSTER=3

Obs	Distritos	Factor1	Factor2
17	PUENTE DE VALLECAS	-1.77856	0.19951
18	VILLAVERDE	-1.21302	0.10469
19	ARGANZUELA	-1.52093	1.23028
20	TETUAN	-0.57579	1.11809

Tabla 3.47

El Cluster 3 está compuesto por 4 distritos *(tabla 3.47)* y se caracterizan por tomar valores negativos del Factor 1. Todos estos distritos destacan por tener un estado de las viviendas malo o deficiente, y en el caso de Arganzuela y Puente de Vallecas también un estado ruinoso.

CLUSTER=4

Obs	Distritos	Factor1	Factor2
21	CENTRO	-0.37259	2.74879

Tabla 3.48

Y por último tenemos el Cluster 4, que está compuesto únicamente por el distrito Centro *(tabla 3.48)*, esto es así porque tiene un valor negativo del Factor 1 y un valor positivo y muy grande del Factor 2. Por lo tanto, Centro está caracterizado por tener un estado de las viviendas malo o deficiente. Además, es el distrito que menor porcentaje de viviendas principales tiene, y a su vez mayor porcentaje de viviendas desocupadas.

A continuación, a partir del Análisis Factorial que hicimos anteriormente, podemos llevar a cabo una representación gráfica de los distritos en el espacio de los dos primeros Factores, en donde podemos ver gráficamente los Clusters creados:

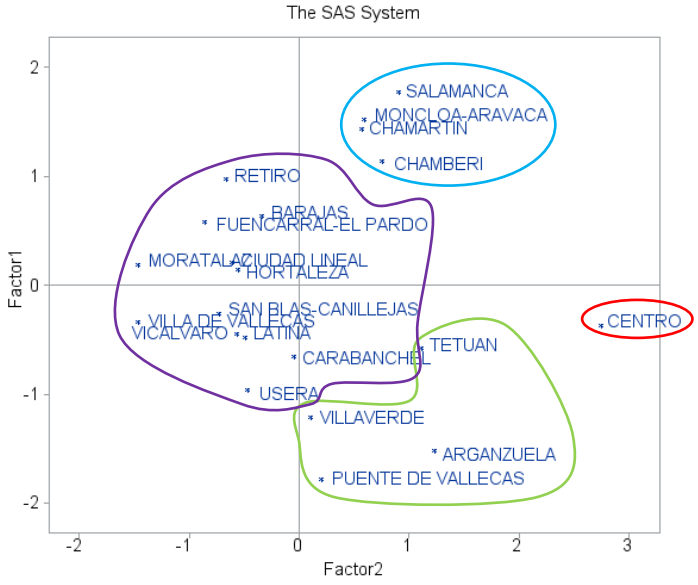

Figura 3.18

En el gráfico anterior se puede ver la distribución de los diferentes distritos en el espacio de los dos primeros Factores *(figura 3.18)*. Además, tal y como hemos señalado en el gráfico podemos ver los 4 Clusters que hemos creado anteriormente. Se puede apreciar como los distritos que pertenecen a un mismo Cluster se encuentran próximos entre sí, como era de esperar.

3.2.6 Calidad de vida: Satisfacción con los servicios públicos

A continuación, vamos a analizar la satisfacción con una serie de servicios públicos: espacios verdes, parques infantiles, centros culturales, organización de fiestas y eventos populares, instalaciones deportivas y servicios sociales municipales. Además, estas variables están medidas en una escala de 0 a 10. Por lo tanto, vamos a contar con 6 variables y 21 observaciones.

Debido a que tenemos muy pocas variables no vamos a llevar a cabo un análisis Factorial, sino que vamos a realizar directamente un análisis Cluster a partir de estas 6 variables, con el objetivo de formar grupos de distritos con características similares.

En este caso, vamos a utilizar un análisis Cluster Jerárquico, debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Centroid Distance**	**Tie**
20	**CARABANCHEL**	**VILLA DE VALLECAS**	2	0.0038	.996	.	.	13.9	.	0.951	
19	**LATINA**	**PUENTE DE VALLECAS**	2	0.0043	.992	.	.	13.7	.	1.0113	
18	**CIUDAD LINEAL**	**HORTALEZA**	2	0.0044	.988	.	.	14.1	.	1.0237	
17	**RETIRO**	**USERA**	2	0.0062	.981	.	.	13.2	.	1.218	
16	**FUENCARRAL-EL PARDO**	**BARAJAS**	2	0.0067	.975	.	.	12.9	.	1.2683	
15	**TETUAN**	**CHAMBERI**	2	0.0078	.967	.	.	12.5	.	1.3644	
14	**CL18**	**SAN BLAS-CANILLEJAS**	3	0.0142	.953	.	.	10.9	3.3	1.5985	
13	**CL15**	**CL14**	5	0.0243	.928	.	.	8.7	2.8	1.5586	
12	**CHAMARTIN**	**CL16**	3	0.0155	.913	.	.	8.6	2.3	1.6711	
11	**SALAMANCA**	**MORATALAZ**	2	0.0119	.901	.	.	9.1	.	1.6895	
10	**CL17**	**CL19**	4	0.0252	.876	.	.	8.6	4.8	1.74	
9	**CL13**	**CL20**	7	0.0399	.836	.	.	7.6	3.7	1.8303	
8	**CL12**	**VILLAVERDE**	4	0.0243	.812	.	.	8.0	2.2	1.9708	
7	**CL11**	**CL9**	9	0.0548	.757	.	.	7.3	3.6	2.0569	
6	**CL7**	**CL8**	13	0.1002	.657	.	.	5.7	5.3	2.0841	
5	**CL10**	**VICALVARO**	5	0.0351	.622	.	.	6.6	3.0	2.2942	
4	**CENTRO**	**CL6**	14	0.0573	.564	.493	1.97	7.3	2.2	2.7208	
3	**ARGANZUELA**	**CL5**	6	0.0536	.511	.372	3.75	9.4	3.0	2.7794	
2	**CL3**	**MONCLOA-ARAVACA**	7	0.0593	.451	.221	6.88	15.6	2.4	2.8823	
1	**CL4**	**CL2**	21	0.4512	.000	.000	0.00	.	15.6	3.4064	

Tabla 3.49

En la *tabla 3.49* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

En primer lugar analizaremos el R^2 considerando valores altos del R^2 hasta el 70%, por lo que podríamos elegir como mínimo 7 Clusters, ya que si elegimos menos la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por lo tanto, teniendo en cuenta este criterio deberíamos quedarnos con 7 Clusters.

En segundo lugar, analizaremos el R^2 semiparcial, en el cual podemos observar que con 14 Clusters tenemos un salto, por lo que deberíamos quedarnos con 15 Clusters. Otro de los saltos que se observa, se da con 6 Clusters, por lo que nos deberíamos quedar con 7 Clusters, para no perder tanta homogeneidad al pasar a 6. Finalmente, teniendo en cuenta este criterio y comparándolo con el anterior deberíamos quedarnos con 7 Clusters, ya que es el menor número de Clusters óptimo en ambos métodos.

Además, también vamos a analizar la Pseudo F. Tal y como podemos ver en la tabla anterior *(tabla 3.49)*, tenemos dos máximos relativos del valor de este estadístico. El primero de ellos, se da con 11 Clusters, por lo que sería preferible un número menor de Clusters. Por consiguiente, el otro máximo relativo del valor de este estadístico con el que nos encontramos se da con 8 Clusters, por lo que según este criterio, nos deberíamos quedar con 8 Clusters.

Y por último, si miramos el Pseudo test de la T^2, podemos observar que tenemos un máximo relativo con 6 Clusters, por lo que deberíamos quedarnos con 7. Por lo tanto, al igual que concluían los dos primeros métodos, deberíamos quedarnos con 7 Clusters.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de Clusters adecuado para este estudio debe ser 7.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma, que es la que se muestra a continuación:

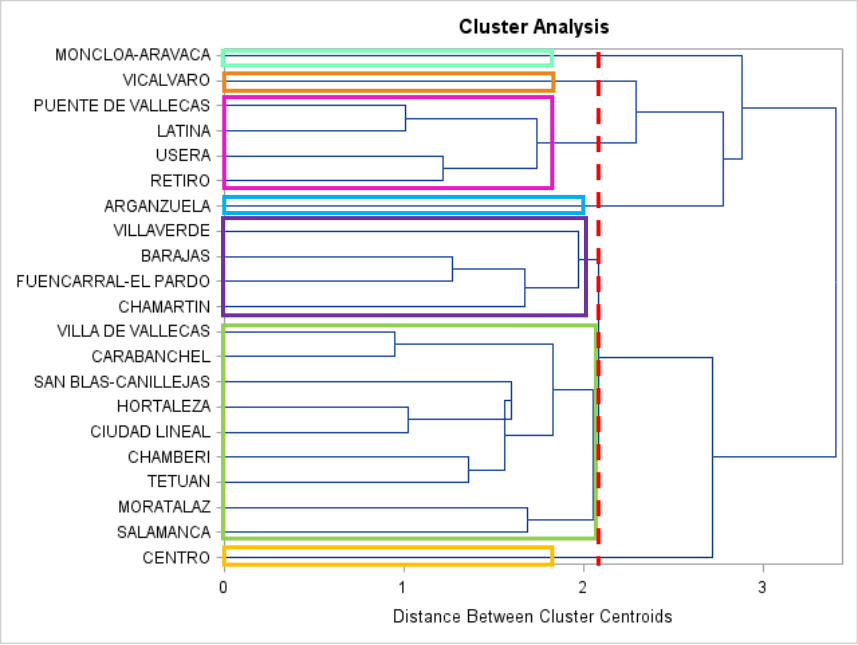

Figura 3.19

A partir de esta representación *(figura 3.19)*, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Pero como ya habíamos determinado anteriormente el número adecuado de Clusters es 7, por lo que deberíamos detener el proceso en ese momento y no realizaremos las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos

parado. Además, a partir del dendrograma anterior, podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, vamos a analizar estos 7 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en las 6 variables que hemos incluido en este grupo.

CLUSTER=1

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
1	CARABANCHEL	6.0	5.4	6.4	6.4	6.1	6.2
2	VILLA DE VALLECAS	6.3	5.5	6.4	6.6	6.2	6.0
3	CIUDAD LINEAL	5.8	5.6	6.9	6.3	6.6	6.4
4	HORTALEZA	5.8	5.9	6.6	6.2	6.5	6.3
5	TETUAN	6.0	5.5	6.4	5.8	6.5	6.1
6	CHAMBERI	6.3	5.7	6.5	5.7	6.1	6.4
7	SAN BLAS-CANILLEJAS	5.7	5.6	6.7	5.8	6.5	6.8
8	SALAMANCA	6.7	6.4	6.4	5.7	6.0	6.0
9	MORATALAZ	6.4	6.3	6.5	6.0	6.6	6.2

Tabla 3.50

El Cluster 1 está formado por los distritos que se muestran en la *tabla 3.50*. Pertenecen al mismo Cluster por tener una satisfacción con los servicios públicos anteriores ni muy buena ni muy mala en comparación con el resto de distritos, es decir, se encuentra en la media. Además, las puntuaciones obtenidas en estos distritos oscilan entre 6-7 sobre 10.

CLUSTER=2

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
10	LATINA	6.5	6.2	7.1	6.6	7.0	6.6
11	PUENTE DE VALLECAS	6.7	6.3	6.8	6.8	7.0	6.7
12	RETIRO	7.1	6.2	7.2	6.1	6.9	6.9
13	USERA	6.9	6.4	7.1	6.4	7.0	7.2

Tabla 3.51

El Cluster 2 está formado por los 4 distritos que aparecen en la *tabla 3.51*. Se caracterizan por tener una satisfacción muy alta con los centros culturales, organización de fiestas y eventos populares, instalaciones deportivas y servicios sociales municipales, tomando una puntuación en torno a 7 sobre 10.

CLUSTER=3

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
14	FUENCARRAL-EL PARDO	6.1	5.5	5.7	5.8	5.9	5.8
15	BARAJAS	6.1	5.6	6.2	5.9	6.0	5.8
16	CHAMARTIN	6.3	5.9	6.1	5.3	6.1	5.9
17	VILLAVERDE	5.5	5.1	6.3	5.5	5.8	5.9

Tabla 3.52

El Cluster 3 está formado por los 4 distritos que se muestran en la *tabla 3.52*. Estos distritos se caracterizan por tener una satisfacción bastante mala con los servicios públicos anteriores, ya que sus puntuaciones oscilan entre 5.5-6 sobre 10.

CLUSTER=4

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
18	VICALVARO	6.8	6.4	7.2	5.6	7.2	7.0

Tabla 3.53

El Cluster 4 únicamente está formado por Vicálvaro *(tabla 3.53)*. Este distrito se encuentra solo en un grupo debido a que tiene una satisfacción muy buena con las instalaciones deportivas, los servicios sociales municipales y los centros culturales, y una satisfacción muy mala con la organización de fiestas y eventos populares.

CLUSTER=5

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
19	CENTRO	5.6	5.1	6.4	6.5	5.4	6.2

Tabla 3.54

Por otro lado, el Cluster 5 está formado únicamente por Centro *(tabla 3.54)*. Este distrito está caracterizado por tener una satisfacción bastante mala con los espacios verdes, parques infantiles e instalaciones deportivas, y sin embargo una satisfacción buena con la organización de fiestas y eventos populares.

CLUSTER=6

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
20	ARGANZUELA	6.9	6.4	7.3	6.1	6.7	5.8

Tabla 3.55

En el caso del Cluster 6 tenemos solamente al distrito Arganzuela *(tabla 3.55)*. Este distrito está caracterizado por tener una satisfacción bastante buena con los espacios verdes y con los centros culturales, pero una satisfacción mala con los servicios sociales municipales.

CLUSTER=7

Obs	Distritos	EspaciosVerdes	ParquesInfantiles	CentrosCulturales	OrganizacionFiestasYEventosPopul	InstalacionesDeportivas	ServiciosSocialesMunicipales
21	MONCLOA-ARAVACA	7.3	7.0	6.4	6.2	6.3	6.4

Tabla 3.56

Y por último, tenemos el Cluster 7 compuesto por Moncloa-Aravaca *(tabla 3.56)*. Se caracteriza por tener una satisfacción muy buena con los espacios verdes y con los parques infantiles, tomando unas puntuaciones en torno a 7 sobre 10.

3.2.7 Seguridad

A continuación vamos a analizar las variables relacionadas con la seguridad, incluiremos las intervenciones de la Policía Municipal en materia de seguridad relacionadas con las personas, la tenencia de armas, el patrimonio, y la tenencia y consumo de drogas. Además, también vamos a incluir la variable que mide la proporción de detenidos e investigados de la Policía Municipal en materia de seguridad. Por lo tanto, vamos a contar con 5 variables y 21 observaciones.

Debido a que tenemos muy pocas variables, no vamos a llevar a cabo un análisis Factorial, sino que vamos a realizar directamente un análisis Cluster a partir de estas 5 variables, con el objetivo de formar grupos de distritos con características similares.

En este caso vamos a utilizar un análisis Cluster Jerárquico, debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la siguiente tabla:

Cluster History											
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Centroid Distance**	**Tie**
20	**VICALVARO**	**BARAJAS**	2	0.0002	1.00	.	.	309	.	0.1845	
19	**MORATALAZ**	**CL20**	3	0.0003	.999	.	.	216	2.0	0.2268	
18	**FUENCARRAL-EL PARDO**	**HORTALEZA**	2	0.0003	.999	.	.	225	.	0.2326	
17	**MONCLOA-ARAVACA**	**LATINA**	2	0.0004	.999	.	.	210	.	0.2844	
16	**RETIRO**	**CL19**	4	0.0007	.998	.	.	179	2.6	0.2986	
15	**CL16**	**CL18**	6	0.0016	.997	.	.	124	4.4	0.3463	
14	**CHAMARTIN**	**CHAMBERI**	2	0.0008	.996	.	.	126	.	0.3958	
13	**CL17**	**VILLAVERDE**	3	0.0011	.995	.	.	125	2.7	0.4025	
12	**CL13**	**SAN BLAS-CANILLEJAS**	4	0.0015	.993	.	.	119	2.0	0.4499	
11	**CL14**	**CL12**	6	0.0030	.990	.	.	100	3.2	0.4771	
10	**CL11**	**TETUAN**	7	0.0026	.988	.	.	96.6	1.9	0.5532	
9	**CL15**	**CL10**	13	0.0164	.971	.	.	50.4	14.5	0.713	
8	**SALAMANCA**	**USERA**	2	0.0034	.968	.	.	55.6	.	0.8245	
7	**CL9**	**CL8**	15	0.0289	.939	.	.	35.8	11.6	1.2917	
6	**PUENTE DE VALLECAS**	**CIUDAD LINEAL**	2	0.0110	.928	.	.	38.5	.	1.4861	
5	**CL7**	**VILLA DE VALLECAS**	16	0.0210	.907	.	.	38.9	4.8	1.4964	
4	**CL5**	**CARABANCHEL**	17	0.0478	.859	.537	14.0	34.5	8.7	2.2544	
3	**CL4**	**CL6**	19	0.1487	.710	.407	9.59	22.1	17.9	2.8831	
2	**ARGANZUELA**	**CL3**	20	0.1198	.590	.244	10.8	27.4	7.4	3.5519	
1	**CENTRO**	**CL2**	21	0.5903	.000	.000	0.00	.	27.4	7.8728	

Tabla 3.57

En la *tabla 3.57* se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

En primer lugar analizaremos el R^2. Vamos a considerar valores altos del R^2 hasta el 70%, por lo que podríamos elegir como mínimo 3 Clusters, ya que si elegimos menos la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por lo tanto, teniendo en cuenta este criterio deberíamos quedarnos con 3 Clusters.

En segundo lugar, si analizamos el R^2 semiparcial podemos observar que con 7 Clusters tenemos un salto, por lo que deberíamos quedarnos con 8. Otro de los saltos que se observa, se da con 3 Clusters, por lo que nos deberíamos quedar con 4, para no perder tanta homogeneidad al pasar a 3. Finalmente, teniendo en cuenta este criterio y comparándolo con el anterior, deberíamos quedarnos con 4 Clusters, ya que es el menor número de Clusters óptimo que coincide en ambos métodos.

A continuación, también vamos a analizar la Pseudo F. Como podemos ver en la tabla anterior *(tabla 3.57)*, tenemos dos máximos relativos del valor de este estadístico. El primero de ellos, se da con 8 Clusters, por lo que sería preferible un número menor de Clusters. Por consiguiente,

el otro máximo relativo del valor de este estadístico con el que nos encontramos se da con 5 Clusters, por lo que según este criterio, nos deberíamos quedar con 5 Clusters.

Y por último, miramos el Pseudo test de la T^2. En este caso, tenemos un máximo relativo que se da con 3 Clusters, por lo que deberíamos quedarnos con 4. Por lo tanto, al igual que concluían los dos primeros métodos, deberíamos quedarnos con 4 Clusters.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de Clusters adecuado para este estudio debe ser 4.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma, que es la que se muestra a continuación:

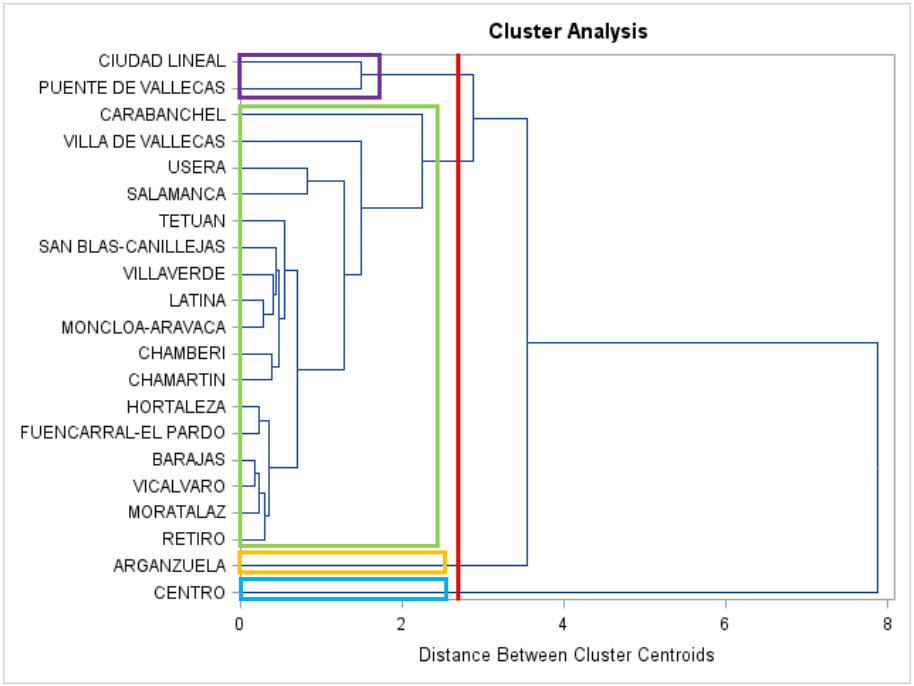

Figura 3.20

A partir de esta representación *(figura 3.20)*, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Como ya habíamos determinado anteriormente, el número adecuado de Clusters es 4, por lo que deberíamos detener el proceso de agrupación cuando tengamos 4 Clusters, y por consiguiente no llevaremos a cabo las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos parado. Además, a partir del dendrograma anterior, podemos ver los grupos que se han formado y el orden en el que se han ido formando.

A continuación, vamos a analizar estos 4 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también, del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en las 5 variables que hemos incluido en este grupo.

CLUSTER=1

Obs	Distritos	IntervenPoliciaMunicipalPersonas	IntervenPoliciaMunicipalArmas	IntervenPoliciaMunicipalPatrimon	IntervenPoliciaMunicipalDrogas	PropDetenidosEInvestigados
1	VICALVARO	1.38317	1.38427	2.38607	1.5248	1.8476
2	BARAJAS	1.48617	0.38760	1.76022	1.7021	1.7011
3	MORATALAZ	1.01530	2.21484	1.91668	1.4894	1.1476
4	FUENCARRAL-EL PARDO	2.70747	1.66113	1.58420	2.2813	2.3441
5	HORTALEZA	2.33961	2.87929	1.82867	3.3688	1.9697
6	MONCLOA-ARAVACA	4.20836	1.05205	3.62801	1.6785	2.7348
7	LATINA	4.00235	2.60244	3.98983	1.1584	3.5895
8	RETIRO	1.69217	1.49502	1.94602	1.6785	2.6941
9	CHAMARTIN	2.26604	1.16279	3.53022	2.0095	4.6557
10	CHAMBERI	2.97234	0.88594	2.24917	2.5059	4.0127
11	VILLAVERDE	3.70806	3.04540	4.49834	3.2151	4.0941
12	SAN BLAS-CANILLEJAS	2.70747	2.10410	3.65734	1.9976	2.2058
13	TETUAN	4.97351	3.37763	2.87502	4.2199	3.9720
14	SALAMANCA	2.07475	0.88594	5.10464	2.3168	7.0324
15	USERA	5.09123	2.65781	6.52259	2.4704	6.9754
16	VILLA DE VALLECAS	2.56033	5.53710	3.53022	13.0615	3.2964
17	CARABANCHEL	6.91583	4.92802	6.96264	7.3995	10.1416

Tabla 3.58

Como podemos observar en la *tabla 3.58*, el Cluster 1 está formado por 17 distritos, siendo el más numeroso. Pertenecen al mismo Cluster por tener una proporción muy baja de intervenciones de la Policía Municipal en materia de seguridad relacionadas con las personas, exceptuando el distrito de Carabanchel. Por otra parte, también destacan por tener una pequeña proporción de intervenciones relacionadas con la tenencia de armas y con el patrimonio. Además, atendiendo a las intervenciones relacionadas con la tenencia y consumo de drogas, también tienen un bajo porcentaje, exceptuando Villa de Vallecas. Y por último, exceptuando Carabanchel, el resto de distritos tienen una proporción de detenidos e investigados de la Policía Municipal en materia de seguridad igual o inferior al 7%.

CLUSTER=2

Obs	Distritos	IntervenPoliciaMunicipalPersonas	IntervenPoliciaMunicipalArmas	IntervenPoliciaMunicipalPatrimon	IntervenPoliciaMunicipalDrogas	PropDetenidosEInvestigados
18	PUENTE DE VALLECAS	17.0541	7.03212	5.98474	3.49882	6.66612
19	CIUDAD LINEAL	12.4044	1.88261	3.75513	3.87707	4.06967

Tabla 3.59

El Cluster 2 está formado por Puente de Vallecas y Ciudad Lineal *(tabla 3.59)*. Se caracterizan por tener una proporción muy elevada de intervenciones de la Policía Municipal en materia de seguridad relacionadas con las personas, y por el contrario un bajo porcentaje de intervenciones relacionadas con la tenencia de armas, el patrimonio, y la tenencia y consumo de drogas. En relación a la proporción de detenidos e investigados de la Policía Municipal en materia de seguridad, es mayor en Puente de Vallecas que en Ciudad Lineal, aunque en ambos casos se puede considerar baja.

CLUSTER=3

Obs	Distritos	IntervenPoliciaMunicipalPersonas	IntervenPoliciaMunicipalArmas	IntervenPoliciaMunicipalPatrimon	IntervenPoliciaMunicipalDrogas	PropDetenidosEInvestigados
20	ARGANZUELA	7.93114	1.27353	17.6511	3.38061	5.46964

Tabla 3.60

El Cluster 3 está formado únicamente por Arganzuela *(tabla 3.60)*. Esto es debido a que este distrito se caracteriza por tener una proporción muy elevada de intervenciones de la Policía Municipal en materia de seguridad relacionadas con las personas, y sobre todo, con el patrimonio. Sin embargo, no tiene apenas intervenciones relacionadas con la tenencia de armas, y la tenencia y consumo de drogas. Y por último, este distrito tiene un 5.5% de detenidos e investigados de la Policía Municipal en materia de seguridad.

CLUSTER=4

Obs	Distritos	IntervenPoliciaMunicipalPersonas	IntervenPoliciaMunicipalArmas	IntervenPoliciaMunicipalPatrimon	IntervenPoliciaMunicipalDrogas	PropDetenidosEInvestigados
21	CENTRO	10.5062	51.5504	14.6392	35.1655	19.3798

Tabla 3.61

Por último, tenemos el Cluster 4 formado únicamente por Centro *(tabla 3.61)*. Esto es debido a que Centro se caracteriza por tener muy poca seguridad, ya que tiene unas proporciones muy elevadas de intervenciones de la Policía Municipal en materia de seguridad relacionadas con las personas, la tenencia de armas, el patrimonio, y la tenencia y consumo de drogas. Además, también tiene una proporción de detenidos e investigados de la Policía Municipal en materia de seguridad en torno al 20%, siendo el distrito donde este valor es mayor.

3.2.8 Resultados elecciones locales

Por último, tenemos este grupo en el que vamos a incluir todas las variables relacionadas con los resultados de las elecciones locales en 2015. Dentro de ellas tenemos las 3 variables que representan: el porcentaje de abstención, de votos blancos y de votos a candidaturas, y dentro de este último la proporción de votos al PP, PSOE, Ahora Madrid y Ciudadanos. Por lo tanto, finalmente tendremos 7 variables y 21 observaciones.

Debido a que tenemos muy pocas variables, no vamos a llevar a cabo un Análisis Factorial, sino que vamos a realizar directamente un análisis Cluster a partir de estas 7 variables, con el objetivo de formar grupos de distritos con características similares.

Vamos a utilizar un análisis Cluster Jerárquico, debido a que no conocemos el número de grupos que queremos formar. Por este motivo, el primer paso es determinar el número adecuado de grupos, con el objetivo de obtener homogeneidad dentro de ellos y heterogeneidad entre ellos. Para esta elección, vamos a utilizar algunos estadísticos que aparecen en la *tabla 3.62*, en la que se muestra cómo se van formando los diferentes Clusters, junto con los estadísticos asociados a cada número de Clusters.

Cluster History											
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Centroid Distance**	**Tie**
20	**CHAMBERI**	**MONCLOA-ARAVACA**	2	0.0017	.998	.	.	31.8	.	0.6807	
19	**FUENCARRAL-EL PARDO**	**HORTALEZA**	2	0.0029	.995	.	.	24.2	.	0.9034	
18	**CL19**	**BARAJAS**	3	0.0042	.991	.	.	19.9	1.4	0.939	
17	**RETIRO**	**CL20**	3	0.0045	.987	.	.	18.6	2.7	0.9683	
16	**SALAMANCA**	**CHAMARTIN**	2	0.0036	.983	.	.	19.5	.	1.0019	
15	**TETUAN**	**CIUDAD LINEAL**	2	0.0046	.979	.	.	19.6	.	1.1335	
14	**LATINA**	**SAN BLAS-CANILLEJAS**	2	0.0056	.973	.	.	19.4	.	1.2573	
13	**CARABANCHEL**	**VILLAVERDE**	2	0.0057	.967	.	.	19.7	.	1.2644	
12	**CL13**	**USERA**	3	0.0083	.959	.	.	19.1	1.4	1.3168	
11	**CL14**	**MORATALAZ**	3	0.0085	.951	.	.	19.2	1.5	1.3339	
10	**CL17**	**CL16**	5	0.0168	.934	.	.	17.2	5.2	1.4009	
9	**CL12**	**PUENTE DE VALLECAS**	4	0.0130	.921	.	.	17.4	1.9	1.5583	
8	**VILLA DE VALLECAS**	**VICALVARO**	2	0.0087	.912	.	.	19.2	.	1.5588	
7	**CL11**	**CL8**	5	0.0182	.894	.	.	19.6	2.4	1.4574	
6	**CL10**	**CL18**	8	0.0439	.850	.	.	17.0	7.8	1.8108	
5	**CL15**	**CL7**	7	0.0404	.810	.	.	17.0	4.4	1.989	
4	**ARGANZUELA**	**CL5**	8	0.0351	.774	.462	12.3	19.5	2.4	2.3686	
3	**CL4**	**CL9**	12	0.1595	.615	.346	8.65	14.4	10.8	2.8937	
2	**CENTRO**	**CL3**	13	0.0799	.535	.206	11.5	21.9	2.9	3.4807	
1	**CL2**	**CL6**	21	0.5351	.000	.000	0.00	.	21.9	3.8892	

Tabla 3.62

Primero analizaremos el R^2 considerando valores altos del R^2 hasta el 70%, por lo que podríamos elegir como mínimo 4 Clusters, ya que si elegimos menos, la proporción de variabilidad explicada por los Clusters sería muy pequeña. Por lo tanto, teniendo en cuenta este criterio deberíamos quedarnos con 4 Clusters.

En segundo lugar, si analizamos el R^2 semiparcial podemos observar que con 10 Clusters tenemos un salto, por lo que deberíamos quedarnos con 11 Clusters. Otro de los saltos que se observa, se da con 3 Clusters, por lo que nos deberíamos quedar con 4, para no perder tanta homogeneidad al pasar a 3. Finalmente, teniendo en cuenta este criterio y comparándolo con el anterior, deberíamos quedarnos con 4 Clusters, ya que es el menor número de Clusters óptimo en ambos métodos.

Además, también vamos a analizar la Pseudo F. Como podemos ver en la tabla anterior *(tabla 3.62)*, tenemos dos máximos relativos del valor de este estadístico. El primero de ellos, se da con 7 Clusters, por lo que sería preferible un número menor de Clusters. Por consiguiente, el otro máximo relativo del valor de este estadístico con el que nos encontramos se da con 4 Clusters, por lo que según este criterio, nos deberíamos quedar con 4 Clusters.

Y por último, miraremos el Pseudo test de la T^2. En este caso, tenemos un máximo relativo que se da con 3 Clusters, por lo que deberíamos quedarnos con 4. Por lo tanto, al igual que concluían el resto de métodos, deberíamos quedarnos con 4 Clusters.

Finalmente, a partir de los estadísticos anteriores hemos decidido que el número de Clusters adecuado para este estudio debe ser 4.

Este proceso de aglomeración y formación de los Clusters también se pueden representar en una gráfica denominada dendrograma *(figura 3.21)*, que es la que se muestra a continuación:

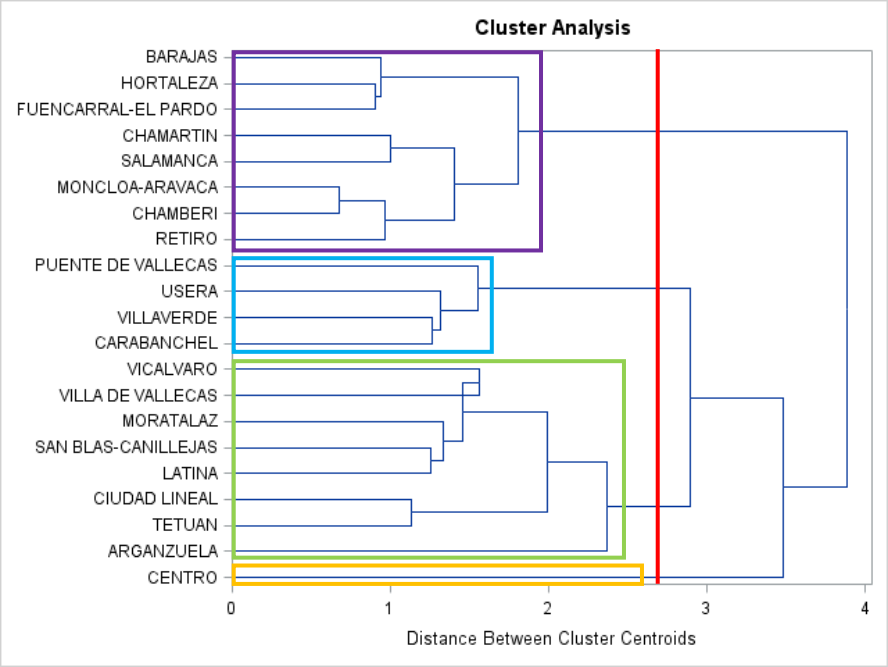

Figura 3.21

A partir de esta representación, se puede determinar en qué momento del proceso de agrupación nos debemos detener, es decir, cuál es el número de Clusters adecuado con el que nos tenemos que quedar. Pero como ya habíamos determinado anteriormente, el número adecuado de Clusters es 4, por lo que deberíamos detener el proceso de agrupación cuando tengamos 4 Clusters, y por consiguiente, no llevaremos a cabo las agrupaciones posteriores a la línea roja, que indica el límite donde nos hemos parado. Además, a partir del dendrograma anterior *(figura 3.21)*, podemos ver los grupos que se han formado y el orden en el que estos se han ido formando.

A continuación, vamos a analizar estos 4 Clusters en función de los distritos que se encuentran en cada uno de ellos, y también del valor que toman los distintos distritos pertenecientes a cada uno de los Clusters en las 7 variables que hemos incluido en este grupo.

CLUSTER=1

Obs	Distritos	Abstencion	VotosBlancos	VotosCandidaturas	PP	PSOE	AhoraMadrid	Ciudadanos
1	CHAMBERI	27.5119	0.70479	71.3203	33.9890	6.3089	18.6005	8.9537
2	MONCLOA-ARAVACA	26.4205	0.73200	72.4209	32.5314	7.4065	19.6506	9.1455
3	FUENCARRAL-EL PARDO	27.5120	0.82418	71.1733	28.2698	9.6747	18.8813	10.3561
4	HORTALEZA	28.8067	0.76761	69.8966	25.3649	10.2974	20.0709	10.2053
5	BARAJAS	27.4125	0.74544	71.2655	25.1791	9.2278	20.7850	11.4117
6	RETIRO	25.2851	0.77438	73.5106	32.3255	6.8481	21.0368	9.3635
7	SALAMANCA	28.4461	0.70743	70.4280	37.1944	5.9641	14.7904	9.1677
8	CHAMARTIN	26.5737	0.75029	72.2792	37.6910	5.8918	15.0013	10.3043

Tabla 3.63

Como podemos observar en la *tabla 3.63*, el Cluster 1 está formado por 8 distritos. Todos ellos pertenecen al mismo Cluster por tener un alto porcentaje de votos a candidaturas, y por consiguiente un porcentaje bajo de abstención. Además, son de los distritos con mayor porcentaje de votos blancos. Por otro lado, podemos destacar que estos distritos votan en su mayoría al partido político PP, seguido de Ahora Madrid, sin embargo el partido menos votado en estos distritos es el PSOE. Por otro lado, si comparamos los votos con el resto de distritos estos son los que más votan a los partidos PP y Ciudadanos, y en consecuencia son los que menos votan a los partidos PSOE y Ahora Madrid.

CLUSTER=2

Obs	Distritos	Abstencion	VotosBlancos	VotosCandidaturas	PP	PSOE	AhoraMadrid	Ciudadanos
9	TETUAN	34.9314	0.65431	63.9651	24.3772	9.5483	18.7450	7.44775
10	CIUDAD LINEAL	31.9573	0.67059	66.8572	25.4066	10.2604	19.4275	7.81070
11	LATINA	31.4722	0.62684	67.2697	22.3774	11.8999	22.5165	6.25108
12	SAN BLAS-CANILLEJAS	32.4822	0.67003	66.2669	18.1782	12.4651	22.7715	8.11640
13	MORATALAZ	28.7108	0.68319	70.0965	22.4420	12.0531	24.2877	7.19763
14	VILLA DE VALLECAS	31.7368	0.62881	67.0429	13.6454	12.8089	27.6080	8.32133
15	VICALVARO	30.3459	0.75916	68.3594	14.8919	14.1758	25.9366	8.45747
16	ARGANZUELA	27.1005	0.63954	71.7780	22.7084	8.6694	28.0494	8.50097

Tabla 3.64

El Cluster 2 está formado por los 8 distritos que se muestran en la *tabla 3.64*. Estos distritos se caracterizan por tener un porcentaje de abstención mayor que en el grupo anterior, y por consiguiente una proporción menor de votos a candidaturas. Además, la proporción de votos blancos también es menor, aunque sigue siendo alta. Por otro lado, el partido político que recibe mayor porcentaje de votos en estos distritos es Ahora Madrid seguido del PP, exceptuando en los distritos Tetuán y Ciudad Lineal, donde el orden es al contrario. Sin embargo, el partido político menos votado es Ciudadanos. Y por último, en comparación con los votos del resto de distritos, podemos destacar que se encuentran entre los que más votan al PP, PSOE y Ahora Madrid, sin embargo en el caso de Ciudadanos ocurre lo contrario, son de los distritos que menos votan a este partido.

CLUSTER=3

Obs	Distritos	Abstencion	VotosBlancos	VotosCandidaturas	PP	PSOE	AhoraMadrid	Ciudadanos
17	CARABANCHEL	35.9670	0.58179	62.8890	18.7360	12.1328	21.5192	6.29299
18	VILLAVERDE	35.8964	0.60093	62.8753	14.6497	15.5186	22.1491	5.66214
19	USERA	38.4391	0.50890	60.5257	14.6078	14.2406	22.2206	5.26230
20	PUENTE DE VALLECAS	36.7647	0.53372	62.2084	11.3711	15.6582	26.4348	4.24470

Tabla 3.65

En relación al Cluster 3, podemos ver que está formado por 4 distritos *(tabla 3.65)*. Se caracterizan por tener un alto porcentaje de abstenciones, y por consiguiente un porcentaje bajo de votos a candidaturas. Sin embargo, el porcentaje de votos en blanco es muy pequeño. Por otro lado, estos distritos votan en su mayoría a Ahora Madrid, y en su minoría a Ciudadanos. En comparación con los votos del resto de distritos, son los que más votan al PSOE, y de los que más votan a Ahora Madrid. Por el contrario, en estos distritos es donde menos se vota a Ciudadanos, y a su vez, son de los que menos votan al PP.

CLUSTER=4

Obs	Distritos	Abstencion	VotosBlancos	VotosCandidaturas	PP	PSOE	AhoraMadrid	Ciudadanos
21	CENTRO	34.2480	0.43634	64.9791	17.3407	6.95924	32.1459	5.31566

Tabla 3.66

Por último, vemos que el Cluster 4 está formado únicamente por Centro *(tabla 3.66)*. Esto es debido a que este distrito se caracteriza por tener un porcentaje elevado de abstenciones, y en consecuencia un porcentaje no demasiado alto de votos a candidaturas. Además, también destaca por tener un bajo porcentaje de votos blancos, siendo el distrito que menos tiene. Por otra parte, en este distrito el partido que más votos recibe es Ahora Madrid, seguido del PP. Sin embargo, el menos votado es Ciudadanos, aunque tiene un porcentaje de votos similar al PSOE. Por último, en comparación con los votos del resto de distritos, podemos destacar que Centro es el distrito que más vota al partido Ahora Madrid. No obstante, es de los distritos que menos vota al PP, PSOE y Ciudadanos.

4. Regresión PLS

Una vez realizados los análisis anteriores, procederemos a realizar una regresión lineal para predecir algunas de las variables de nuestro fichero. Debido a la gran cantidad de variables de las que disponemos, vamos a elegir como variables respuesta aquellas que pueden ser más interesantes. Por este motivo, realizaremos una predicción del porcentaje de votos de los cuatro partidos políticos más importantes, por lo tanto, vamos a tener cuatro variables respuesta.

Por otro lado, debido al elevado número de variables que podemos utilizar como explicativas, vamos a llevar a cabo dos regresiones, una de ellas respecto a las variables económicas y otra respecto a las variables demográficas.

En nuestro caso, debido a que nuestras variables tienen problemas de multicolinealidad, y además tenemos muchas variables y pocas observaciones, vamos a utilizar la regresión PLS. Esta regresión transforma las variables explicativas en componentes ortogonales, las cuales representan la solución al problema de multicolinealidad y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras. Además, en este caso, debido a que tenemos cuatro variables respuesta, vamos a utilizar la regresión PLS2, que se usa en el caso de que tengamos más de una variable dependiente.

El objetivo de esta regresión es encontrar un modelo que se ajuste lo mejor posible a observaciones futuras. Para ello, vamos a seleccionar el mejor modelo mediante validación cruzada, la cual se puede llevar a cabo con diferentes métodos. En nuestro caso, debido al bajo número de observaciones, utilizaremos el método CV=ONE, que consiste en ir eliminando del fichero cada observación para estimar el modelo de regresión y realizar la predicción para dicha observación con el modelo estimado sin ella. A partir de este método, decidiremos el número de factores con los que nos debemos quedar para hacer la regresión.

Para decidir el número de factores adecuado, podemos hacerlo eligiendo aquel número que tenga menor PRESS, es decir, menor error cuadrático medio, y por lo tanto este sería el mejor modelo. Por otro lado, también podemos aplicar el test de Van der Voet para decidir si el

descenso en el error es significativo, cuya construcción se basa en la diferencia en el error respecto del mínimo PRESS obtenido. En este caso, el número de factores óptimo es el menor número para el que el aumento del error respecto al error mínimo no es significativo.

4.1 Regresión PLS con variables económicas

Vamos a realizar realizaremos una regresión PLS para explicar el porcentaje de votos de los partidos políticos mediante una serie de variables económicas. Estas variables explicativas son aquellas que hacen referencia a la renta neta media anual, las pensiones medias mensuales, todos los indicadores relacionados con el desempleo, la información relativa a las personas con necesidades de asistencia social, el valor catastral medio de los bienes inmuebles y la superficie media de la vivienda en transacción.

En primer lugar, vamos a decidir el número de factores adecuado. Para ello, vamos a obtener la tabla en la que se muestran los diferentes números de factores posibles, con su correspondiente valor Root Mean PRESS, el estadístico t y su correspondiente p-valor *(tabla 4.1)*.

Cross Validation for the Number of Extracted Factors			
Number of Extracted Factors	**Root Mean PRESS**	**T**2**	**Prob > T**2**
0	1.05	11.03423	0.0070
1	0.62976	6.477928	0.1480
2	0.655517	2.881909	0.6520
3	0.609369	0	1.0000
4	0.613447	6.232551	0.1530
5	0.713994	6.526399	0.0820
6	0.71095	5.477439	0.1980
7	0.734528	4.997625	0.2540
8	0.765633	8.890846	0.0130
9	0.8468	9.558634	0.0110
10	0.927781	9.843616	0.0050
11	0.993042	9.754636	0.0060
12	1.132748	10.25922	<.0001
13	1.114361	11.03002	<.0001
14	1.08016	10.5258	<.0001
15	1.119515	10.84641	<.0001

Tabla 4.1

Minimum root mean PRESS	0.6094
Minimizing number of factors	3
Smallest number of factors with p > 0.1	1

Tabla 4.2

Como podemos observar *(tabla 4.2)*, el mínimo PRESS se obtiene con 3 factores, sin embargo si tenemos en cuenta el estadístico t deberíamos quedarnos únicamente con 1 factor, ya que es el menor número para el que el aumento del error respecto al error mínimo no es significativo. Pero finalmente para decidir el número de factores con el que nos debemos quedar, vamos a observar el porcentaje de variabilidad explicada por estos factores, el cual aparece en la siguiente tabla:

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	78.7588	78.7588	66.7332	66.7332
2	6.5072	85.2660	9.5512	76.2844
3	4.7085	89.9745	8.4074	84.6918
4	3.2190	93.1935	2.0710	86.7628
5	1.0309	94.2244	4.1189	90.8818
6	1.6753	95.8997	0.8889	91.7707
7	1.3474	97.2470	1.0711	92.8418
8	1.0368	98.2839	0.9546	93.7964
9	0.5610	98.8449	0.6660	94.4624
10	0.2065	99.0515	1.2829	95.7454
11	0.4269	99.4784	0.4813	96.2267
12	0.1205	99.5988	0.8875	97.1142
13	0.1058	99.7046	0.3215	97.4357
14	0.1199	99.8245	0.0984	97.5341
15	0.1218	99.9463	0.0756	97.6097

Tabla 4.3

En la *tabla 4.3*, podemos observar que con 1 factor el porcentaje de variabilidad explicada de las variables explicativas es 78.75% y de las variables respuesta es 66.73%. Sin embargo, con 3 factores estos porcentajes aumentan a 89.97% y 84.69%, para las variables explicativas y respuesta respectivamente. Por lo tanto, aunque con 1 factor el aumento del error respecto al mínimo PRESS obtenido no es significativo los porcentajes de variabilidad explicada son muy pequeños, por lo que sería preferible quedarnos con 3 factores, que es el número con el que obteníamos el mínimo PRESS.

Estas proporciones de variabilidad explicada también se pueden ver gráficamente:

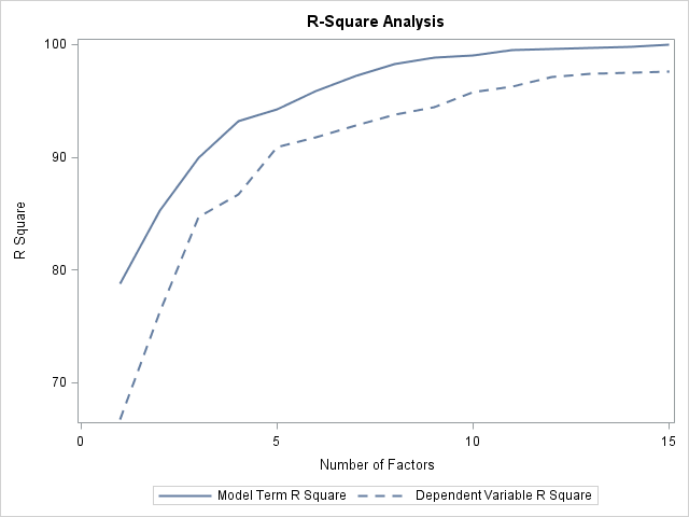

Figura 4.1

En el gráfico *(figura 4.1)*, podemos observar como al pasar de 1 factor a 3 factores aumenta notablemente el porcentaje de varianza explicada, por lo que sería preferible quedarnos con 3 factores.

Una vez que nos hemos quedado con 3 factores, podemos obtener el porcentaje de varianza explicada con cada factor para el conjunto de variables explicativas y variables respuesta, y además desglosado por variables. Debido al gran número de ellas, únicamente vamos a mostrar las variables respuesta *(tabla 4.4)*, ya que nuestro objetivo es conseguir explicar estas. El porcentaje de varianza explicada con cada Factor para las variables explicativas se encuentra en el ANEXO II *(tabla II.3)*.

Percent Variation Accounted for by Partial Least Squares Factors						
Number of Extracted Factors	**Dependent Variables**					
	PP	**PSOE**	**AhoraMadrid**	**Ciudadanos**	**Current**	**Total**
1	85.9644	85.8721	30.4381	64.6581	66.7332	66.7332
2	86.0034	95.9746	34.2928	88.8668	9.5512	76.2844
3	92.5071	96.1035	61.2873	88.8694	8.4074	84.6918

Tabla 4.4

Atendiendo a las variables respuesta, vemos que el porcentaje de votos de todos los partidos políticos, excepto Ahora Madrid, tienen un porcentaje de variabilidad explicada muy alto. En el caso de Ahora Madrid el porcentaje de varianza explicada es de 61.3%, lo que se puede deber a que su comportamiento no se ajusta totalmente a un patrón, y como consecuencia es más difícil de explicar, pero lo vamos a considerar como un valor aceptable. Por lo tanto, podemos decir que las variables respuesta quedan bien explicadas por los 3 factores.

Por otro lado, si analizamos los porcentajes de variabilidad explicada con 3 factores para cada una de las variables explicativas, podemos observar que todas ellas tienen una proporción de varianza explicada superior al 75%, excepto la variable que mide el valor catastral medio de los bienes inmuebles de las personas jurídicas, cuyo porcentaje es del 65.5%, pero este valor se puede considerar aceptable. Por lo tanto, podemos decir que las variables explicativas están muy bien explicadas por estos 3 factores (ANEXO II – *tabla II.3*).

A continuación, obtenemos los gráficos de correlaciones de los dos primeros Factores, y por otro lado del primer y tercer Factor, en los que podemos explicar el significado de la posición de las variables en dichos gráficos *(figuras 4.2 y 4.3)*:

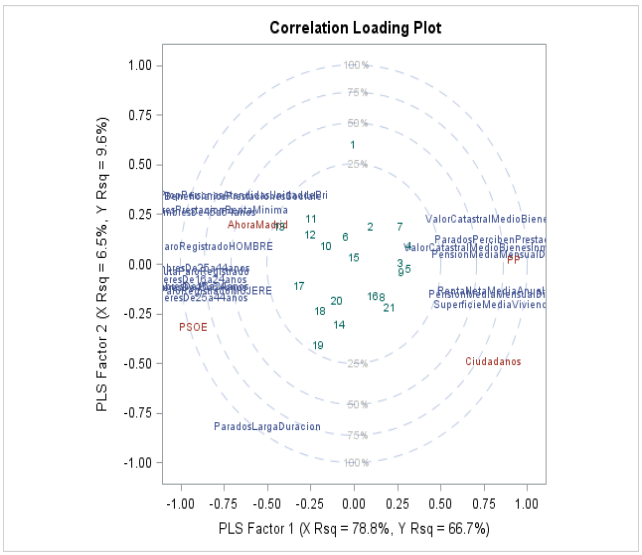

Figura 4.2

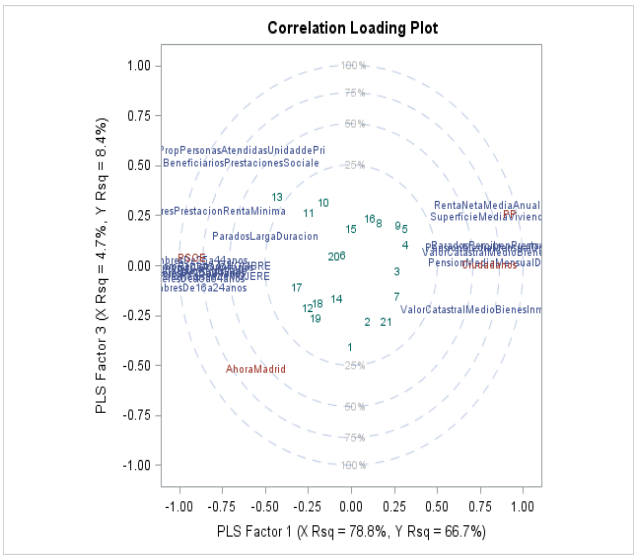

Figura 4.3

Podemos decir que el Factor 1 se representa en el eje de abscisas, y los Factores 2 y 3 en el eje de ordenadas. Por este motivo, las variables que se encuentren en la parte derecha o superior del gráfico tendrán una correlación alta y positiva con el Factor correspondiente. Sin embargo, si se encuentran en la parte izquierda o inferior del gráfico tendrán una correlación negativa y grande. Por otro lado, para conocer la proporción de varianza explicada por los dos Factores correspondientes para cada variable, podemos observar los círculos que se encuentran en el interior de los gráficos.

En primer lugar, podemos observar que las variables respuesta que miden el porcentaje de votos del PP y de Ciudadanos se encuentra en el lado derecho de los dos gráficos, por lo tanto esto quiere decir que tienen correlación positiva y grande con el Factor 1. Sin embargo con el Factor 2, Ciudadanos tiene una correlación negativa pero no muy grande, y el PP no tiene ninguna correlación con este Factor. Y por último, ambos partidos no tienen ninguna correlación con el Factor 3. Además, la proporción de varianza explicada por los dos primeros factores para estas dos variables respuesta es aproximadamente del 86-89%.

Por otro lado, la variable respuesta que mide el porcentaje de votos del PSOE tiene una correlación negativa y grande con el Factor 1, por encontrarse en la parte izquierda de los dos gráficos. En el caso del Factor 2 vemos que tiene una correlación negativa pero muy pequeña, y en relación con el Factor 3 la correlación es nula. Además, la proporción de varianza explicada por los dos primeros factores para esta variable respuesta es aproximadamente del 95%.

Y por último, la variable respuesta que mide el porcentaje de votos de Ahora Madrid tiene una correlación negativa y no excesivamente grande con el Factor 1 y con el 3, y una correlación positiva pero muy pequeña con el Factor 2. Además, la proporción de varianza explicada por los dos primeros Factores para esta variable respuesta es aproximadamente del 34%, el cual es un valor muy bajo, por este motivo con dos Factores no es suficiente para explicar esta variable y finalmente nos hemos quedado con 3 factores.

Si atendemos a las variables explicativas, podemos observar que tienen una correlación grande con el Factor 1, en algunos casos positiva y en otros negativa. Sin embargo, con el Factor 2 la correlación no es muy grande, excepto la variable que mide la proporción de parados de larga duración cuya correlación con el Factor 2 es grande y negativa. Y por último, en relación al Factor 3, podemos decir que la mayoría de las variables no tienen una correlación muy grande con este Factor, exceptuando las variables que nos aportan información relativa a las personas con necesidades de asistencia social, las cuales tienen una correlación positiva y grande con este Factor. Por otro lado, la proporción de varianza explicada por los dos primeros Factores para estas variables varía desde el 52% hasta el 99%. Por este motivo, cogemos un Factor más para que estas proporciones aumenten.

Finalmente, nuestro objetivo es encontrar el modelo adecuado. Para ello, podemos obtener los coeficientes para las variables estandarizadas y para las variables originales, los cuales se muestran en el ANEXO II (*tabla II.4 y tabla II.5*). Pero para poder analizar más claramente la aportación de las variables en dicho modelo, vamos a representar el gráfico de los parámetros del modelo.

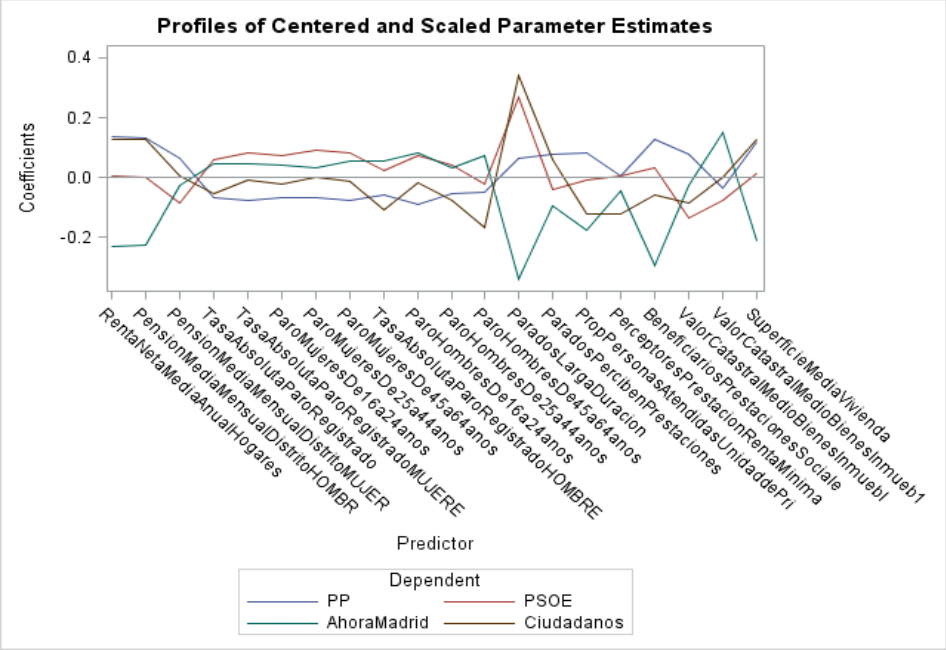

Figura 4.4

Como podemos observar en el gráfico *(figura 4.4)*, atendiendo a los indicadores económicos podemos decir, que si aumenta la renta neta media anual y las pensiones medias mensuales de los hombres aumentarán en proporción los votos al PP y a Ciudadanos, y sin embargo disminuirán los votos a Ahora Madrid. En el caso del PSOE estos factores no influyen en el porcentaje de votos. En cuanto a las pensiones medias de las mujeres, si estas aumentan aumentarán en proporción los votos al PP, y disminuirán los del PSOE, aunque el peso de esta variable en el modelo tampoco es muy grande.

Si analizamos los indicadores de desempleo, podemos observar que a medida que aumenta la tasa de paro independientemente del sexo y de la edad, aumenta la proporción de votos tanto del PSOE como de Ahora Madrid. Por el contrario, disminuye la proporción de votos al PP y a Ciudadanos. Si nos fijamos en los parados de larga duración, vemos claramente que aumenta la proporción de votos a Ciudadanos y al PSOE, y en menor medida al PP, por otro lado disminuyen los votos a Ahora Madrid. Esta variable es la que más peso tiene en el modelo. Y por último, si aumenta la proporción de parados que perciben prestaciones, se vota más al PP y Ciudadanos, y menos a Ahora Madrid seguido del PSOE.

En relación a la información relativa a las personas con necesidades de asistencia social, vemos que si aumenta la proporción de personas atendidas en la Unidad de Primera Atención en Centros de Servicios Sociales se vota más al PP, y menos a Ahora Madrid y Ciudadanos. Lo mismo ocurre cuando aumenta la proporción de beneficiarios de prestaciones sociales de carácter económico, con la diferencia de que el aumento de los votos del PP y la disminución de los votos de Ahora Madrid son más grandes que en el caso anterior. Por otra parte, teniendo en cuenta la proporción de perceptores de prestación de la Renta Mínima de Inserción, observamos que si esta aumenta, disminuyen los votos a Ciudadanos seguido de Ahora Madrid.

Y por último, con relación al valor catastral y a la superficie media, podemos destacar que si el valor catastral medio de los bienes inmuebles de las personas físicas es alto, aumentan los votos

del PP, pero disminuyen los del PSOE y Ciudadanos. Sin embargo, en el caso de las personas jurídicas aumentan los votos de Ahora Madrid y disminuyen los del PSOE. Finalmente, cuando aumenta la superficie media de la vivienda en transacción, aumentan los votos de Ciudadanos y del PP, pero disminuyen los de Ahora Madrid.

A continuación, con este modelo podemos obtener las predicciones de los porcentajes de votos de los cuatro partidos políticos para cada distrito, y también los residuos para estas observaciones (ANEXO II - *tabla II.6*). Pero lo más relevante son los residuos obtenidos, debido a que a partir de ellos podemos ver si el modelo explica bien las variables respuesta. Por este motivo, a continuación vamos a representar los diagramas de caja y bigotes de los residuos de los cuatro partidos políticos:

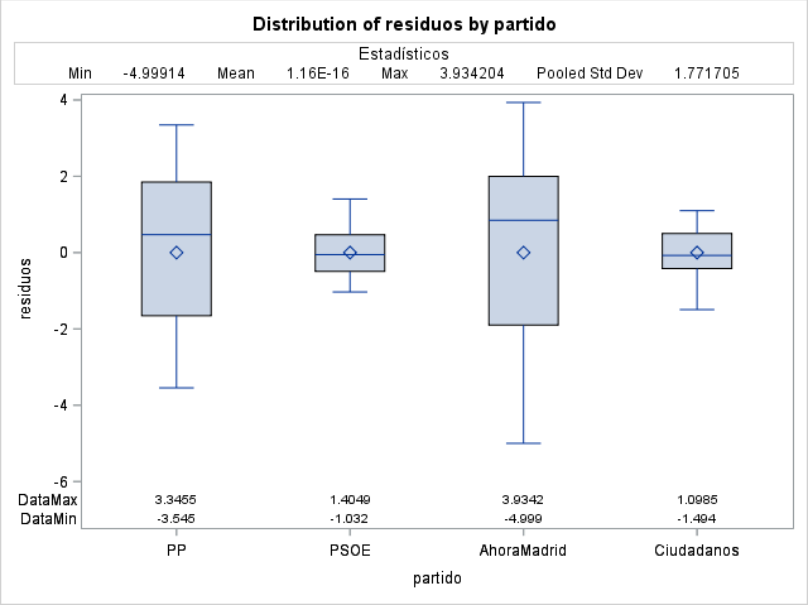

Figura 4.5

Como podemos observar en el gráfico anterior *(figura 4.5)*, los residuos obtenidos en los partidos PSOE y Ciudadanos son muy pequeños, ya que como vemos se encuentran en torno al 0 y tienen muy poca dispersión. Esto quiere decir que este modelo explica muy bien estos dos porcentajes de votos. Por otro lado, los residuos obtenidos en el porcentaje de votos del PP son bastante mayores ya que tienen mayor dispersión, pero debido a que el máximo error es -3.5, podemos decir que estos residuos no son muy grandes, y en consecuencia afirmar que este modelo explica bien esta variable. Y por último, atendiendo al porcentaje de votos de Ahora Madrid, podemos decir que es la variable con mayores residuos ya que son los que tienen mayor dispersión, y esto es debido a que era la que tenía menor porcentaje de variabilidad explicada. Pero aun así, estos residuos no son excesivamente grandes, y los podemos considerar aceptables. Por lo tanto, concluimos que esta variable está bien explicada por el modelo.

4.2 Regresión PLS con variables demográficas

En segundo lugar, vamos a realizar otra regresión PLS para explicar el porcentaje de votos de los partidos políticos, pero en este caso mediante un conjunto de variables demográficas. Dentro de ellas, incluiremos aquellas que indican las edades, la nacionalidad y todas las referentes a la

educación, dentro de las cuales está la población en etapas educativas, la escolarización de alumnos por tipo de centro y el nivel de estudios de la población mayor de 25 años.

A continuación, para decidir el número de factores adecuado, vamos a obtener la tabla en la que se muestran los diferentes números de factores posibles, con su correspondiente valor Root Mean PRESS, el estadístico t y su correspondiente p-valor *(tabla 4.5)*.

Cross Validation for the Number of Extracted Factors			
Number of Extracted Factors	**Root Mean PRESS**	**T**2**	**Prob > T**2**
0	1.05	11.50963	0.0050
1	0.677153	7.490602	0.0680
2	0.62589	3.273651	0.6140
3	0.605471	3.629848	0.5030
4	0.568284	4.55928	0.3710
5	0.537601	0	1.0000
6	0.577832	5.747431	0.2030
7	0.590738	6.396894	0.1170
8	0.577154	7.197648	0.0930
9	0.630855	9.985004	0.0070
10	0.624244	9.927602	0.0060
11	0.667639	8.94255	0.0130
12	0.696007	10.00545	0.0060
13	0.67872	10.77849	0.0040
14	0.634394	10.88562	0.0020
15	0.591381	9.163203	0.0090

Tabla 4.5

Minimum root mean PRESS	0.5376
Minimizing number of factors	5
Smallest number of factors with p > 0.1	2

Tabla 4.6

Como podemos observar *(tabla 4.6)*, el mínimo PRESS se obtiene con 5 factores, sin embargo si tenemos en cuenta el estadístico t, deberíamos quedarnos con 2 factores, ya que es el menor número para el que el aumento del error respecto al error mínimo no es significativo. Pero finalmente para decidir el número de factores con el que nos debemos quedar, vamos a observar el porcentaje de variabilidad explicada por estos factores, el cual aparece en la siguiente tabla:

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	**Model Effects**		**Dependent Variables**	
	Current	**Total**	**Current**	**Total**
1	39.5269	39.5269	64.2149	64.2149
2	19.3720	58.8988	9.5098	73.7247
3	11.8842	70.7830	12.1056	85.8303
4	9.9559	80.7390	4.3637	90.1940
5	5.0033	85.7422	2.3938	92.5878
6	2.3341	88.0764	2.0504	94.6382
7	3.0038	91.0802	1.1281	95.7663
8	2.3890	93.4691	0.8118	96.5781
9	0.8048	94.2739	0.9110	97.4891
10	1.1132	95.3871	0.4214	97.9104
11	1.1839	96.5710	0.5627	98.4731
12	0.9304	97.5014	0.6109	99.0840
13	0.9069	98.4083	0.2180	99.3020
14	0.9793	99.3876	0.1441	99.4461
15	0.4168	99.8044	0.0866	99.5327

Tabla 4.7

En la *tabla 4.7* podemos observar que con 2 factores el porcentaje de variabilidad explicada de las variables explicativas es de 58.89% y de las variables respuestas es de 73.72%. Sin embargo, con 5 factores estos porcentajes aumentan a 85.74% y 92.58% respectivamente. Por lo tanto, aunque con 2 factores el aumento del error respecto al mínimo PRESS obtenido no es significativo, los porcentajes de variabilidad explicada son pequeños, sobre todo el de las variables explicativas. Por este motivo, para que estos porcentajes sean mayores, y en consecuencia mejore el modelo, sería preferible quedarnos con 5 factores, ya que es el número en el que obteníamos el mínimo PRESS, y además nos son demasiados factores. Estas proporciones de variabilidad explicada también se pueden ver gráficamente:

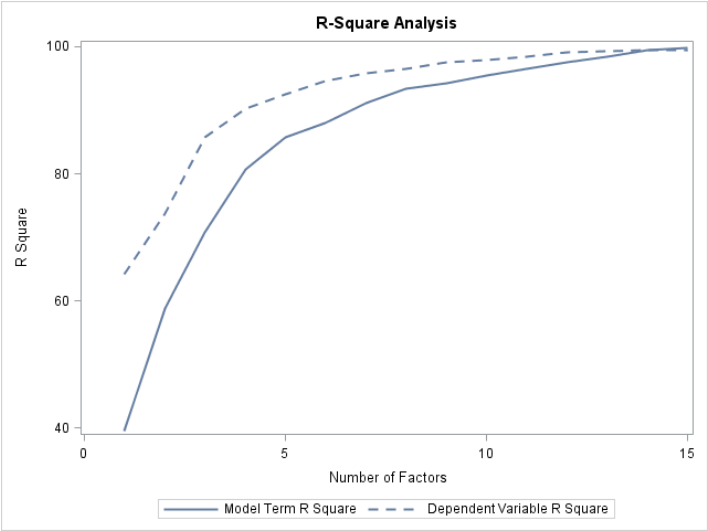

Figura 4.6

En el gráfico anterior *(figura 4.6)*, podemos observar claramente que al pasar de 2 factores a 5 aumenta notablemente el porcentaje de varianza explicada, por lo que como ya hemos dicho anteriormente, sería preferible quedarnos con 5 factores.

Una vez que hemos decidido quedarnos con 5 factores, podemos obtener el porcentaje de varianza explicada con cada factor para el conjunto de variables explicativas y para las variables respuestas y además, desglosado por variables. Debido al gran número de ellas, únicamente vamos a mostrar las variables respuestas *(tabla 4.8)*, ya que nuestro objetivo es conseguir explicar estas. El porcentaje de varianza explicada con cada Factor para las variables explicativas se encuentra en el ANEXO II *(tabla II.7)*.

Percent Variation Accounted for by Partial Least Squares Factors						
Number of Extracted Factors	Dependent Variables					
	PP	PSOE	AhoraMadrid	Ciudadanos	Current	Total
1	84.0232	86.5220	27.1050	59.2092	64.2149	64.2149
2	86.8166	90.2466	31.5271	86.3086	9.5098	73.7247
3	94.7362	93.7183	66.2887	88.5781	12.1056	85.8303
4	94.7793	95.0271	80.4311	90.5385	4.3637	90.1940
5	95.2167	95.6344	84.0698	95.4303	2.3938	92.5878

Tabla 4.8

Atendiendo a las variables respuesta, vemos que el porcentaje de votos de todos los partidos políticos, excepto Ahora Madrid, tienen un porcentaje de variabilidad explicada en torno al

95%, que es un valor muy alto. En el caso de Ahora Madrid es del 84%, cuyo valor también se puede considerar un valor grande y muy bueno. El motivo por el que Ahora Madrid tiene un porcentaje menor que el resto de partidos puede ser porque su comportamiento no se ajusta totalmente a un patrón, y como consecuencia es más difícil de explicar, pero aun así su valor es bastante alto. Por lo tanto, podemos decir que las variables respuesta quedan bien explicadas por los 5 factores.

Por otro lado, si analizamos los porcentajes de variabilidad explicada con 5 factores para cada una de las variables explicativas, vemos que casi todas tienen una proporción de varianza explicada bastante alta, superior al 70%, sin embargo hay algunas de ellas que aunque no tienen un proporción tan alta, esta es superior al 60%, por lo que se pueden considerar aceptables. Por lo tanto, podemos decir que las variables explicativas están muy bien explicadas por estos 5 factores (ANEXO II – *tabla II.7*).

A continuación, obtenemos los gráficos de correlaciones del primer factor con respecto a los otros 4 factores *(figuras 4.7, 4.8, 4.9 y 4.10)*, en los que podemos explicar el significado de la posición de las variables:

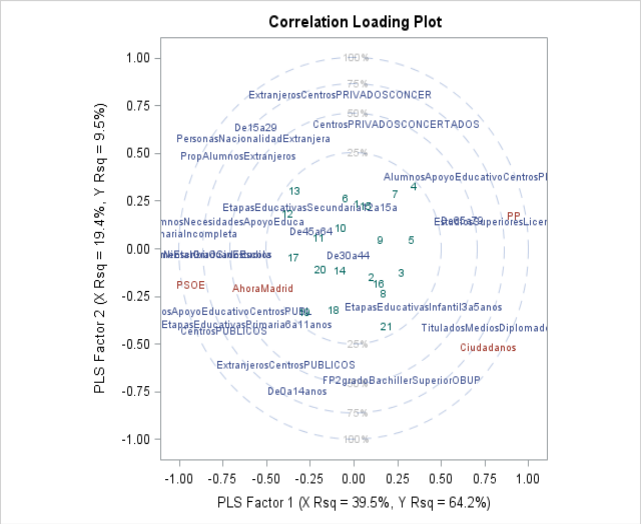

Figura 4.7

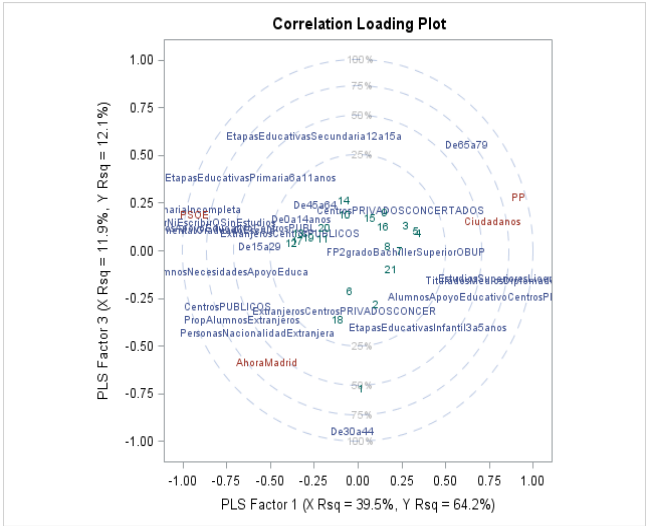

Figura 4.8

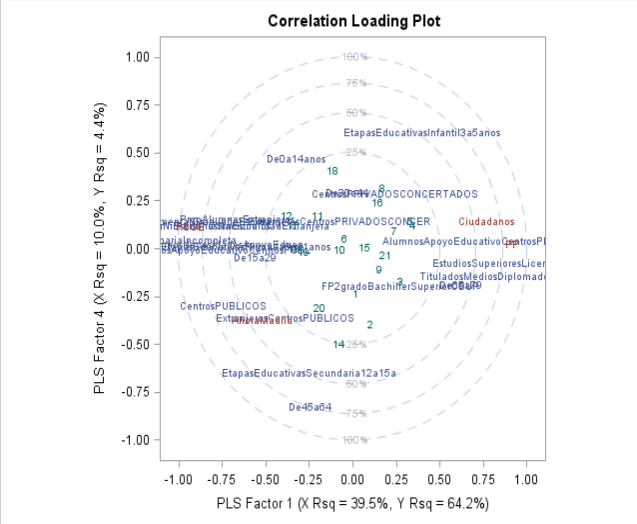

Figura 4.9

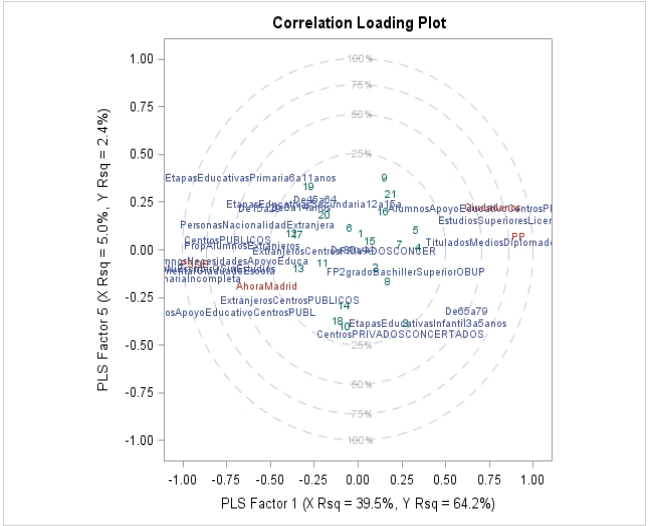

Figura 4.10

Podemos decir que el Factor 1 se encuentra en el eje de abscisas, y por el contrario el resto de Factores se encuentran en el eje de ordenadas. Por este motivo, las variables que se encuentren en la parte derecha o superior del gráfico tendrán una correlación alta y positiva con el Factor correspondiente, sin embargo las que se encuentren en la parte izquierda o inferior tendrán una correlación negativa y grande. Por otro lado, para conocer la proporción de varianza explicada por los dos Factores correspondientes para cada variable, podemos observar los círculos que se encuentran en el interior de los gráficos.

En primer lugar, vemos las variables respuesta que miden el porcentaje de votos del PP y de Ciudadanos que se encuentran en el lado derecho de los gráficos, por lo tanto tienen correlación positiva y grande con el Factor 1. Sin embargo, Ciudadanos tiene una correlación negativa pero no muy grande con el Factor 2, y casi nula con el resto de Factores. En el caso del PP, la correlación con los 4 últimos Factores es mínima. Además, la proporción de varianza explicada por los dos primeros Factores para estas dos variables respuesta es aproximadamente del 86%.

Por otro lado, la variable respuesta que mide el porcentaje de votos del PSOE tiene una correlación negativa y grande con el Factor 1, por encontrarse en la parte izquierda de los gráficos. En relación con el resto de Factores, tiene una correlación muy pequeña con ellos. Además, la proporción de varianza explicada por los dos primeros Factores para esta variable respuesta es aproximadamente del 90%.

Y por último, la variable respuesta que mide el porcentaje de votos de Ahora Madrid tiene una correlación negativa y no excesivamente grande con el Factor 1, el 3 y el 4, y una correlación negativa pero muy pequeña con el Factor 2 y el 5. Además, la proporción de varianza explicada por los dos primeros Factores para esta variable respuesta es aproximadamente del 31%, el cual es un valor muy bajo. Por este motivo con dos Factores no es suficiente para explicar esta variable, y nos hemos quedado finalmente con 5.

Si atendemos a las variables explicativas, podemos observar que la mayoría de ellas tienen una correlación grande con el Factor 1, en algunos casos positiva y en otros negativa. Por otro lado, podemos destacar la que representa a los extranjeros en centros privados concertados que tiene una correlación positiva y grande con el Factor 2. Sin embargo, las variables que representan a las personas de 0 a 14 años y a las personas mayores de 25 años con un nivel de estudios bastante alto, tienen una correlación negativa y grande con el Factor 2. En relación con el Factor 3, podemos destacar la variable que representa a las personas de 30 a 44 años que tiene una correlación negativa y alta con este Factor. Por otra parte, la variable que representa la población de 45 a 64 años tiene una correlación negativa y muy alta con el Factor 4. Y por último, no hay ninguna variable que tenga una correlación muy alta con el Factor 5. Por otro lado, la proporción de varianza explicada por los 2 primeros Factores para estas variables en algunos casos es muy pequeña, por lo que finalmente cogemos 5 Factores para que estas proporciones aumenten.

Finalmente, nuestro objetivo es encontrar el modelo adecuado. Para ello, podemos obtener los coeficientes para las variables estandarizadas y para las variables originales (ANEXO II - *tabla II.8 y tabla II.9)*. Pero para poder analizar más claramente la aportación de las variables en dicho modelo, vamos a representar el gráfico de los parámetros del modelo *(tabla 4.11)*.

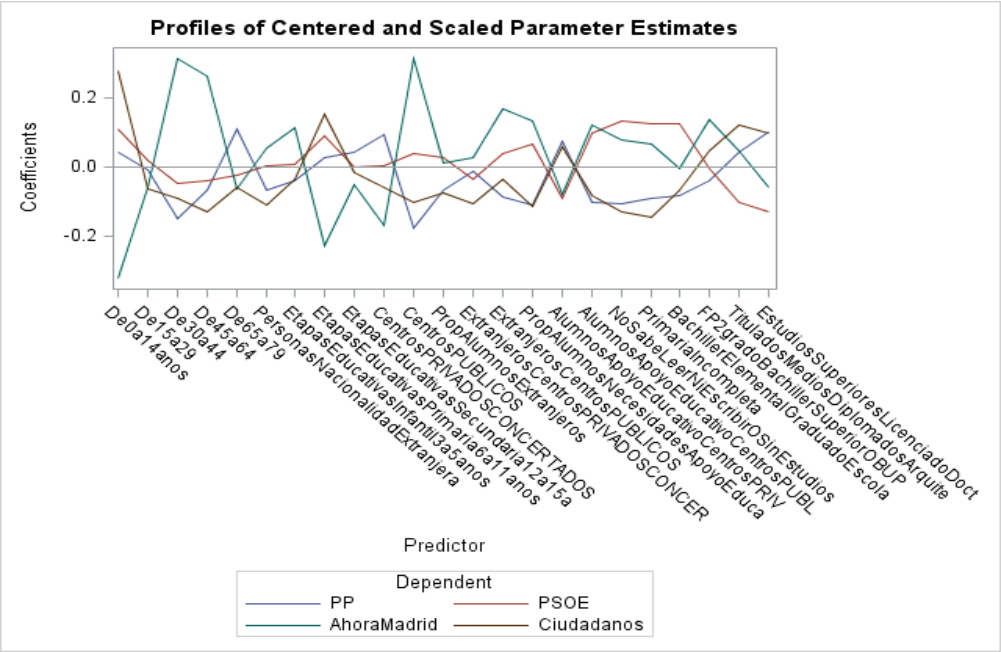

Figura 4.11

En primer lugar atendiendo a la edad de los individuos, podemos ver que cuando aumenta la proporción de personas entre 0 y 14 años, aumentan los votos a Ciudadanos, y en menor medida al PSOE y al PP, por el contrario disminuyen los de Ahora Madrid. En cuanto a la población joven de 15 a 29 años, tanto el incremento como la disminución en el porcentaje de votos son pequeños, siendo el PSOE el único partido en el que aumentan. Sin embargo, en las edades centrales de 30 a 64 años vemos como aumenta de forma clara el porcentaje de votos de Ahora Madrid, y por el contrario disminuye el porcentaje del PP y Ciudadanos, y en menor medida el del PSOE. Y por último, atendiendo a la gente de mayor edad, podemos ver como aumenta el porcentaje de votos del PP, pero disminuyen los porcentajes de los otros 3 partidos políticos.

Cuando el porcentaje de población extranjera aumenta, vemos que también lo hace el porcentaje de votos de Ahora Madrid, sin embargo disminuyen los de Ciudadanos y del PP. En el caso de los votos del PSOE, apenas hay variación.

A continuación, en relación con la población en etapas educativas, observamos que cuando aumenta la proporción de población en la etapa educativa infantil de 3 a 5 años, aumenta el porcentaje de votos de Ahora Madrid, sin embargo disminuyen los del PP y Ciudadanos. En el caso del aumento del porcentaje de la etapa de primaria de 6 a 11 años, vemos como claramente disminuye el porcentaje de votos de Ahora Madrid, pero aumentan los de los otros 3 partidos, dándose el mayor aumento en Ciudadanos. Y finalmente, si aumenta la proporción de la etapa educativa secundaria de 12 a 15 años, podemos observar que aumenta el porcentaje de votos del PP, pero disminuye el de Ahora Madrid.

En relación a la escolarización de alumnos por tipo de centro, vemos como el porcentaje de votos de Ahora Madrid aumenta notablemente en los centros públicos y disminuye en los privados concertados, sin embargo en el caso del PP ocurre lo contrario. En ambos centros disminuye el porcentaje de votos de Ciudadanos, siendo mayor este descenso en los centros

públicos. Y por último, en los centros públicos aumentan los votos al PSOE, mientras que en los privados concertados su aportación es mínima.

Si aumenta la proporción de alumnos extranjeros, cabe destacar que disminuyen los votos del PP y Ciudadanos, y aumenta los de los otros dos partidos mínimamente. Dentro de este tipo de alumnos, si pertenecen a centros privados concertados aumentan los votos de Ahora Madrid, pero disminuyen el resto destacando Ciudadanos. Y en el caso de que pertenezcan a centros públicos, aumentan los votos de Ahora Madrid seguido del PSOE, y disminuyen tanto los del PP como los de Ciudadanos, aunque estos últimos en menor medida.

En relación a los alumnos con necesidades de apoyo educativo, si estos aumentan vemos que aumentan los votos de Ahora Madrid y PSOE, y disminuyen los del PP y Ciudadanos. Dentro de este tipo de alumnos, si pertenecen a los centros públicos, el comportamiento es el mismo que acabamos de comentar. Y sin embargo, si pertenecen a centros privados concertados, ocurre lo contrario.

Y por último, en relación al nivel de estudios de la población mayor de 25 años, podemos decir que cuando aumenta el porcentaje de personas con un nivel de estudios bajo, hace que aumente el porcentaje de votos del PSOE y Ahora Madrid, y que disminuya el de Ciudadanos y PP. Pero cuando aumenta un poco el nivel de estudios, es decir, Bachiller Elemental, Graduado Escolar, ESO o Formación profesional 1º grado, ocurre lo mismo que anteriormente con la diferencia de Ahora Madrid que tiene una aportación mínima. Por otro lado, si aumenta el porcentaje de personas con un nivel de estudios de Formación profesional 2º grado, Bachiller Superior o BUP, esto provoca que aumenten los votos de Ahora Madrid seguido de Ciudadanos, pero disminuyan los del PP. Y por último, en el nivel de estudios más alto aumentan los votos de Ciudadanos y PP, y disminuyen los del PSOE y Ahora Madrid.

A continuación, con este modelo vamos a obtener las predicciones de los porcentajes de votos de los cuatro partidos políticos para cada uno de los distritos, y también los residuos para estas observaciones (ANEXO II – *tabla II.10*). Para ver si nuestro modelo es bueno, vamos a representar los diagramas de caja y bigotes de los residuos de los cuatro partidos políticos.

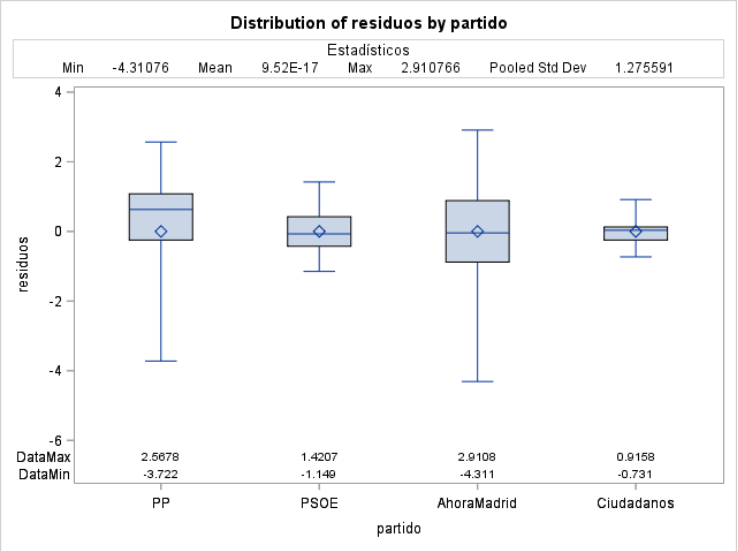

Figura 4.12

Como podemos observar en el gráfico anterior *(figura 4.12)*, Ciudadanos es el partido que menos residuos tiene, ya que la dispersión es mínima. En el caso del PSOE, aunque su dispersión es mayor, también sigue siendo muy pequeña. Por lo tanto, en ambos casos los residuos son muy pequeños, lo que quiere decir que este modelo explica muy bien estos dos porcentajes de votos. Por otro lado, los residuos obtenidos en el porcentaje de votos del PP son mayores que en los dos partidos anteriores, pero aun así no son muy grandes y se pueden considerar aceptables, por lo que podemos afirmar que esta variable está bien explicada por el modelo. Y por último, los residuos del porcentaje de votos de Ahora Madrid son los mayores debido a que es el partido que tiene menor porcentaje de variabilidad explicada, aunque se pueden considerar aceptables, y por consiguiente podemos decir que este modelo explica bien esta variable.

5. Conclusión

El objetivo principal de este trabajo es el estudio sociodemográfico de los distritos de Madrid. Para ello, en primer lugar hemos llevado a cabo varios análisis Factoriales, con sus posteriores análisis Clusters, para diferentes grupos de variables creados. Esto tiene el objetivo de agrupar los distritos de Madrid en grupos similares, y de esta forma poder caracterizar estos grupos en base a las variables utilizadas. En segundo lugar, hemos realizado una regresión PLS para explicar el porcentaje de votos de los partidos políticos mediante una serie de variables divididas en dos grupos (económicas y demográficas), con el objetivo de poder analizar cuáles son las que más influyen en el porcentaje de votos de cada partido.

A partir de los resultados obtenidos y tras el análisis de los mismos, se obtienen las siguientes conclusiones:

- Según los Clusters obtenidos en relación a las características generales (superficie y densidad) y a la población del distrito, podemos destacar a los distritos Hortaleza, Barajas y Fuencarral-El Pardo por tener una gran superficie. Por el contrario, Centro se caracteriza por tener muy poca superficie y un tamaño medio del hogar muy por debajo del resto de distritos.
- En cuanto a los Clusters realizados para los indicadores económicos e indicadores de desempleo, destacan los distritos Carabanchel, Puente de Vallecas, Usera, Vicálvaro, Villa de Vallecas y Villaverde, por tener altas tasas de paro, así como elevadas proporciones de parados de larga duración.
- En los Clusters realizados para la educación, los distritos Moratalaz, Vicálvaro y San Blas-Canillejas tienen una proporción elevada de población en etapas educativas de primaria, entre 6 y 11 años, y una proporción elevada de alumnos en centros públicos, incluyendo tanto los alumnos extranjeros como los alumnos con necesidades de apoyo educativo.
- Atendiendo a los Clusters de salud y de servicios sociales, los distritos Salamanca, Chamberí, Chamartín, Barajas y Retiro tienen una proporción muy pequeña de personas con necesidades de asistencia social, de personas mayores con servicio de ayuda a domicilio, además de ser los distritos en donde se práctica más ejercicio físico diario.

- En los Clusters realizados en base a las viviendas, destaca Centro por tener un estado de las viviendas malo o deficiente, y por ser el distrito que menos viviendas principales y más viviendas desocupadas tiene.
- En relación a los Clustes sobre la satisfacción con los servicios públicos, destacamos el distrito de Arganzuela por tener una satisfacción bastante buena con los espacios verdes y con los centros culturales. Sin embargo, tiene una satisfacción mala con los servicios sociales municipales.
- Según los Clusters en base a la seguridad de los distritos, Centro destaca por ser el distrito con menor seguridad.
- En los últimos Clusters creados en función de los resultados de las elecciones locales, los distritos Carabanchel, Villaverde, Usera y Puente de Vallecas tienen un alto porcentaje de abstenciones, un porcentaje bajo de votos a candidaturas y un porcentaje de votos en blanco muy pequeño. Además, votan en su mayoría a Ahora Madrid y en su minoría a Ciudadanos.
- Atendiendo a la regresión PLS realizada con las variables económicas, podemos destacar que las variables que más influyen de forma positiva en el porcentaje de votos del PP son la renta media de los hogares y la pensión media de los hombres. Y las que más influyen de forma negativa son las tasas de paro. Por otro lado, los parados de larga duración influyen de forma positiva en los partidos PSOE y Ciudadanos, y negativamente en Ahora Madrid. Y por último, el valor catastral medio de los bienes inmuebles influye de forma negativa en el PSOE y positiva en Ahora Madrid.
- Y por último en la regresión PLS realizada con las variables demográficas, podemos ver que la proporción de personas entre 65 y 79 años influye positivamente y los centros públicos lo hacen de forma negativa en la proporción de votos del PP. En relación a los votos del PSOE, el nivel de estudios más bajo de la población mayor de 25 años influye de forma positiva y por el contrario el nivel más alto de estudios lo hace de forma negativa. Por otra parte, las variables que indican la proporción de personas de 30 a 64 años y los centros públicos influyen positivamente en los votos de Ahora Madrid, mientras que la proporción de personas de 0 a 14 años lo hace de forma negativa. Finalmente, la proporción de votos de Ciudadanos está influida positivamente por la proporción de personas de 0 a 14 años, y negativamente por los dos niveles de estudios más bajos de la población mayor de 25 años.

6. Bibliografía

- Apuntes de la asignatura Estudio y Depuración de Datos de la profesora Juana Mª Alonso Revenga.
- Apuntes de la asignatura Técnicas Estadísticas Multivariantes I de la profesora Mª Lina Vicente Hernanz.
- Apuntes de la asignatura Técnicas Avanzadas de Predicción de la profesora Juana Mª Alonso Revenga.
- Ayuntamiento de Madrid, 2017a. Panel de indicadores de distritos y barrios de Madrid. Estudio sociodemográfico: Estudio de indicadores de distritos y barrios 2017, Portal de Datos Abiertos del Ayuntamiento de Madrid. Disponible en: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=71359583a773a510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default [Última consulta: 26/02/2018].
- Ayuntamiento de Madrid, 2017b. Anexo gráfico del panel de indicadores 2017, Portal de Datos Abiertos del Ayuntamiento de Madrid. Disponible en: https://datos.madrid.es/FWProjects/egob/Catalogo/SectorPublico/Ficheros/ANEXO%20GR%C3%81FICO%20PANEL%20DE%20INDICADORES%202017.pdf [Última consulta: 4/02/2018].

7. ANEXOS

ANEXO I: Sintaxis de SAS

- Análisis Factorial (apartado 3.2.1: Características generales y población del distrito)

```
proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie -- EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie -- De30a44 De65a79 -- EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie Densidad PropMujeres -- De30a44 De65a79 -- EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie Densidad PropMujeres -- De30a44 De65a79 PersonasNacionalidadExtranjera
        TamanoMedioHogar -- EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie Densidad PropMujeres -- De15a29 De65a79 PersonasNacionalidadExtranjera
        TamanoMedioHogar -- EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie Densidad PropMujeres -- De15a29 De65a79 PersonasNacionalidadExtranjera
        TamanoMedioHogar -- HogaresMonoparentalesHombre EsperanzaVidaNacerMUJERES
        EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion2;
        var Superficie Densidad PropMujeres -- De0a14anos De65a79 PersonasNacionalidadExtranjera
        TamanoMedioHogar -- HogaresMonoparentalesHombre EsperanzaVidaNacerMUJERES
        EsperanzaVidaNacerHOMBRES;
run;

proc factor data=tfg.datos_eliminvar msa scree out=solucion1 outstat=solucion2 n=3;
        var Superficie Densidad PropMujeres -- De0a14anos De65a79 TamanoMedioHogar --
        HogaresMonoparentalesHombre EsperanzaVidaNacerMUJERES EsperanzaVidaNacerHOMBRES;
run;

proc transpose data=solucion2 out=represen;
        where _type_ = 'PATTERN';
run;
%plotit(data=represen, plotvars=Factor1 Factor2, labelvar=_name_, href=0, vref=0);
%plotit(data=solucion1, plotvars=Factor1 Factor2, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
```

- Análisis Cluster (apartado 3.2.1: Características generales y población del distrito)

```
proc cluster data=solucion1 noeigen std method=centroid nonorm ccc pseudo out=salcluster;
        var Factor1 -- Factor3;
        id distritos;
run;
```

```
proc tree data=salcluster out=saltree nclusters=5;
        id distritos;
        copy Factor1 -- Factor3;
run;

proc sort data=saltree;
        by cluster;
run;
proc print data=saltree;
        by cluster;
        var distritos Factor1 -- Factor3;
run;
```

- Análisis Factorial (apartado 3.2.2: Indicadores económicos e Indicadores de desempleo)

```
proc factor data=tfg.datos_eliminvar msa scree out=solucion3 outstat=solucion4 n=1;
        var RentaNetaMediaAnualHogares -- ParadosPercibenPrestaciones;
run;
```

- Análisis de Componentes Principales (apartado 3.2.2: Indicadores económicos e Indicadores de desempleo)

```
proc princomp data=tfg.datos_eliminvar plot=all outstat=statdat out=datscores n=1;
        var RentaNetaMediaAnualHogares -- ParadosPercibenPrestaciones;
run;

proc print data=datscores noobs;
        var distritos PRIN1;
run;

proc gchart data=datscores;
        hbar distritos / sumvar=PRIN1;
run;
```

- Análisis Factorial (apartado 3.2.3: Educación)

```
proc factor data=tfg.datos_eliminvar msa scree outstat=solucion6;
        var EtapasEducativasInfantil3a5anos -- EstudiosSuperioresLicenciadoDoct;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion6;
        var EtapasEducativasInfantil3a5anos -- BachillerElementalGraduadoEscola
        TituladosMediosDiplomadosArquite EstudiosSuperioresLicenciadoDoct;
run;

proc factor data=tfg.datos_eliminvar msa scree outstat=solucion6;
        var EtapasEducativasInfantil3a5anos EtapasEducativasPrimaria6a11anos
        CentrosPRIVADOSCONCERTADOS -- BachillerElementalGraduadoEscola
        TituladosMediosDiplomadosArquite EstudiosSuperioresLicenciadoDoct;
run;

proc factor data=tfg.datos_eliminvar msa scree out=solucion5 outstat=solucion6 n=3;
        var EtapasEducativasInfantil3a5anos EtapasEducativasPrimaria6a11anos CentrosPUBLICOS --
        BachillerElementalGraduadoEscola TituladosMediosDiplomadosArquite
        EstudiosSuperioresLicenciadoDoct;
run;
```

```
proc transpose data=solucion6 out=represen2;
        where _type_ = 'PATTERN';
run;

%plotit(data=represen2, plotvars=Factor1 Factor2, labelvar=_name_, href=0, vref=0);
%plotit(data=represen2, plotvars=Factor1 Factor3, labelvar=_name_, href=0, vref=0);

%plotit(data=solucion5, plotvars=Factor1 Factor2, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
%plotit(data=solucion5, plotvars=Factor1 Factor3, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
```

- Análisis Cluster (apartado 3.2.3: Educación)

```
proc cluster data=solucion5 noeigen std method=centroid nonorm ccc pseudo out=salcluster2;
        var Factor1 -- Factor3;
        id distritos;
run;

proc tree data=salcluster2 out=saltree2 nclusters=5;
        id distritos;
        copy Factor1 -- Factor3;
run;

proc sort data=saltree2;
        by cluster;
run;
proc print data=saltree2;
        by cluster;
        var distritos Factor1 -- Factor3;
run;
```

- Análisis Factorial (apartado 3.2.4: Salud y Servicios Sociales)

```
proc factor data=tfg.datos_eliminvar msa scree out=solucion7 outstat=solucion8 n=3;
        var Practicadeejerciciofisicodiario -- Consumodemedicamentos PropPersonasDiscapacidadReconoci
        PropPersonasAtendidasUnidaddePri -- PropPersonasMayoresSociasCentros CentrosdeServiciosSociales
        -- CentrosAtencionInfancia;
run;

proc factor data=tfg.datos_eliminvar msa scree out=solucion7 outstat=solucion8 n=3 rotate=varimax plot=all;
        var Practicadeejerciciofisicodiario -- Consumodemedicamentos PropPersonasDiscapacidadReconoci
        PropPersonasAtendidasUnidaddePri -- PropPersonasMayoresSociasCentros CentrosdeServiciosSociales
        -- CentrosAtencionInfancia;
run;

proc transpose data=solucion8 out=represen3;
        where _type_ = 'PATTERN';
run;

%plotit(data=represen3, plotvars=Factor1 Factor2, labelvar=_name_, href=0, vref=0);
%plotit(data=represen3, plotvars=Factor1 Factor3, labelvar=_name_, href=0, vref=0);

%plotit(data=solucion7, plotvars=Factor1 Factor2, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
%plotit(data=solucion7, plotvars=Factor2 Factor3, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
```

- Análisis Cluster (apartado 3.2.4: Salud y Servicios Sociales)

```
proc cluster data=solucion7 noeigen std method=centroid nonorm ccc pseudo out=salcluster3;
        var Factor1 -- Factor3;
        id distritos;
run;

proc tree data=salcluster3 out=saltree3 nclusters=5;
        id distritos;
        copy Factor1 -- Factor3;
run;

proc sort data=saltree3;
        by cluster;
run;
proc print data=saltree3;
        by cluster;
        var distritos Factor1 -- Factor3;
run;
```

- Análisis Factorial (apartado 3.2.5: Vivienda)

```
proc factor data=tfg.datos_eliminvar msa scree out=solucion9 outstat=solucion10 n=2;
        var Viviendasanterioresa1980 -- Desocupada;
run;

proc transpose data=solucion10 out=represen4;
        where _type_ = 'PATTERN';
run;

%plotit(data=represen4, plotvars=Factor1 Factor2, labelvar=_name_, href=0, vref=0);

%plotit(data=solucion9, plotvars=Factor1 Factor2, labelvar=distritos, tsize=1.5, symsize=0.5, ls=125, href=0,
vref=0);
```

- Análisis Cluster (apartado 3.2.5: Vivienda)

```
proc cluster data=solucion9 noeigen std method=centroid nonorm ccc pseudo out=salcluster4;
        var Factor1 -- Factor2;
        id distritos;
run;

proc tree data=salcluster4 out=saltree4 nclusters=4;
        id distritos;
        copy Factor1 -- Factor2;
run;

proc sort data=saltree4;
        by cluster;
run;
proc print data=saltree4;
        by cluster;
        var distritos Factor1 -- Factor2;
run;
```

- Análisis Cluster (apartado 3.2.6: Calidad de vida: Satisfacción con los servicios públicos)

```
proc cluster data=TFG.datos_eliminvar noeigen std method=centroid nonorm ccc pseudo out=salcluster5;
        var EspaciosVerdes -- ServiciosSocialesMunicipales;
        id distritos;
run;

proc tree data=salcluster5 out=saltree5 nclusters=7;
        id distritos;
        copy EspaciosVerdes -- ServiciosSocialesMunicipales;
run;

proc sort data=saltree5;
        by cluster;
run;
proc print data=saltree5;
        by cluster;
        var distritos EspaciosVerdes -- ServiciosSocialesMunicipales;
run;
```

- Análisis Cluster (apartado 3.2.7: Seguridad)

```
proc cluster data=TFG.datos_eliminvar noeigen std method=centroid nonorm ccc pseudo out=salcluster6;
        var IntervenPoliciaMunicipalPersonas -- PropDetenidosEInvestigados;
        id distritos;
run;

proc tree data=salcluster6 out=saltree6 nclusters=4;
        id distritos;
        copy IntervenPoliciaMunicipalPersonas -- PropDetenidosEInvestigados;
run;
proc sort data=saltree6;
        by cluster;
run;
proc print data=saltree6;
        by cluster;
        var distritos IntervenPoliciaMunicipalPersonas -- PropDetenidosEInvestigados;
run;
```

- Análisis Cluster (apartado 3.2.8: Resultados elecciones locales)

```
proc cluster data=TFG.datos_eliminvar noeigen std method=centroid nonorm ccc pseudo out=salcluster7;
        var Abstencion -- Ciudadanos;
        id distritos;
run;

proc tree data=salcluster7 out=saltree7 nclusters=4;
        id distritos;
        copy Abstencion -- Ciudadanos;
run;
proc sort data=saltree7;
        by cluster;
run;
proc print data=saltree7;
        by cluster;
        var distritos Abstencion -- Ciudadanos;
run;
```

- Regresión PLS con variables económicas (apartado 4.1)

```
proc pls data=tfg.datos_eliminvar cv=one cvtest;
        model PP PSOE AhoraMadrid Ciudadanos = RentaNetaMediaAnualHogares --
        ParadosPercibenPrestaciones PropPersonasAtendidasUnidaddePri -- BeneficiariosPrestacionesSociale
        ValorCatastralMedioBienesInmuebl -- SuperficieMediaVivienda/ solution;
        output out=out_pred1 p=p_PP p_PSOE p_AhoraMadrid p_Ciudadanos;
run;

proc pls data=tfg.datos_eliminvar method=pls nfac=3 varss censcale details plots=(parmprofiles
corrload(nfac=3 unpack));
        model PP PSOE AhoraMadrid Ciudadanos = RentaNetaMediaAnualHogares --
        ParadosPercibenPrestaciones PropPersonasAtendidasUnidaddePri -- BeneficiariosPrestacionesSociale
        ValorCatastralMedioBienesInmuebl -- SuperficieMediaVivienda/ solution;
        output out=out_pred_pls1 p=p_PP p_PSOE p_AhoraMadrid p_Ciudadanos yresidual=PP_res PSOE_res
        AhoraMadrid_res Ciudadanos_res;
run;

proc print data=out_pred_pls1;
        var distritos PP PSOE AhoraMadrid Ciudadanos p_PP p_PSOE p_AhoraMadrid p_Ciudadanos PP_res
        PSOE_res AhoraMadrid_res Ciudadanos_res;
run;

data residuos_PP;
        set out_pred_pls1 (keep=PP_res);
        residuos = PP_res;
        partido='PP        ';
        drop PP_res;
run;
data residuos_PSOE;
        set out_pred_pls1 (keep=PSOE_res);
        residuos = PSOE_res;
        partido='PSOE      ';
        drop PSOE_res;
run;
data residuos_AhoraMadrid;
        set out_pred_pls1 (keep=AhoraMadrid_res);
        residuos = AhoraMadrid_res;
        partido='AhoraMadrid';
        drop AhoraMadrid_res;
run;
data residuos_Ciudadanos;
        set out_pred_pls1 (keep=Ciudadanos_res);
        residuos = Ciudadanos_res;
        partido='Ciudadanos ';
        drop Ciudadanos_res;
run;

data residuos_partido;
        set residuos_PP residuos_PSOE residuos_AhoraMadrid residuos_Ciudadanos;
run;
proc print data=residuos_partido;
run;

proc boxplot data= residuos_partido;
        plot residuos*partido;
        inset min mean max stddev /header ='Estadísticos' pos = tm;
        insetgroup min max / header = 'Errores mínimo y máximo para cada fichero';
run;
```

- Regresión PLS con variables demográficas (apartado 4.2)

```
proc pls data=tfg.datos_eliminvar cv=one cvtest;
	model PP PSOE AhoraMadrid Ciudadanos = De0a14anos -- PersonasNacionalidadExtranjera
EtapasEducativasInfantil3a5anos -- EstudiosSuperioresLicenciadoDoct/ solution;
	output out=out_pred1 p=p_PP p_PSOE p_AhoraMadrid p_Ciudadanos;
run;

proc pls data=tfg.datos_eliminvar method=pls nfac=5 varss censcale details plots=(parmprofiles
corrload(nfac=5 unpack trace=OFF));
	model PP PSOE AhoraMadrid Ciudadanos = De0a14anos -- PersonasNacionalidadExtranjera
	EtapasEducativasInfantil3a5anos -- EstudiosSuperioresLicenciadoDoct/ solution;
	output out=out_pred_pls2 p=p_PP p_PSOE p_AhoraMadrid p_Ciudadanos yresidual=PP_res PSOE_res
	AhoraMadrid_res Ciudadanos_res;
run;

proc print data=out_pred_pls2;
	var distritos PP PSOE AhoraMadrid Ciudadanos p_PP p_PSOE p_AhoraMadrid p_Ciudadanos PP_res
PSOE_res AhoraMadrid_res Ciudadanos_res;
run;

data residuos_PP2;
	set out_pred_pls2 (keep=PP_res);
	residuos = PP_res;
	partido='PP        ';
	drop PP_res;
run;
data residuos_PSOE2;
	set out_pred_pls2 (keep=PSOE_res);
	residuos = PSOE_res;
	partido='PSOE      ';
	drop PSOE_res;
run;
data residuos_AhoraMadrid2;
	set out_pred_pls2 (keep=AhoraMadrid_res);
	residuos = AhoraMadrid_res;
	partido='AhoraMadrid';
	drop AhoraMadrid_res;
run;
data residuos_Ciudadanos2;
	set out_pred_pls2 (keep=Ciudadanos_res);
	residuos = Ciudadanos_res;
	partido='Ciudadanos ';
	drop Ciudadanos_res;
run;

data residuos_partido2;
	set residuos_PP2 residuos_PSOE2 residuos_AhoraMadrid2 residuos_Ciudadanos2;
run;
proc print data=residuos_partido2;
run;

proc boxplot data= residuos_partido2;
	plot residuos*partido;
	inset min mean max stddev /header ='Estadísticos' pos = tm;
	insetgroup min max / header = 'Errores mínimo y máximo para cada fichero';
run;
```

ANEXO II: Salidas de SAS

- *Tablas II.1*: Índice KMO y medida de adecuación MSA (apartado 3.2.1)

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.23801118																		
Superficie	Densidad	PropPoblacion	PropMujeres	EdadMedia	De0a14anos	De15a29	De30a44	De65a79	PersonasNacionalidadExtranjera	PropHogares	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	TasaBrutaNatalidad	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.1937	0.2514	0.0998	0.3345	0.3153	0.3078	0.1208	0.3098	0.2224	0.1385	0.1277	0.3010	0.3027	0.2997	0.3073	0.2671	0.2119	0.2141	0.1360

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.48581971																		
Superficie	Densidad	PropMujeres	EdadMedia	De0a14anos	De15a29	De30a44	De65a79	PersonasNacionalidadExtranjera	PropHogares	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	TasaBrutaNatalidad	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.7355	0.4407	0.6381	0.5022	0.5700	0.4966	0.3697	0.3773	0.3695	0.3265	0.4183	0.4967	0.4684	0.5018	0.4852	0.4587	0.6757	0.7461

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.52287585																	
Superficie	Densidad	PropMujeres	EdadMedia	De0a14anos	De15a29	De30a44	De65a79	PersonasNacionalidadExtranjera	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	TasaBrutaNatalidad	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.8280	0.4939	0.7187	0.5281	0.5978	0.4755	0.3822	0.4088	0.4027	0.4453	0.5126	0.4996	0.5412	0.5135	0.5255	0.6676	0.7456

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.53834984																
Superficie	Densidad	PropMujeres	EdadMedia	De0a14anos	De15a29	De65a79	PersonasNacionalidadExtranjera	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	TasaBrutaNatalidad	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.9020	0.5264	0.6698	0.5623	0.6602	0.4644	0.4211	0.5163	0.4893	0.5198	0.5178	0.5659	0.5132	0.4083	0.5315	0.6709

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.59173134														
Superficie	Densidad	PropMujeres	EdadMedia	De0a14anos	De15a29	De65a79	PersonasNacionalidadExtranjera	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.7184	0.7417	0.5434	0.7195	0.6568	0.3481	0.6933	0.3891	0.4454	0.6226	0.6119	0.6207	0.5847	0.5560	0.5236

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.62639695														
Superficie	Densidad	PropMujeres	EdadMedia	De0a14anos	De65a79	PersonasNacionalidadExtranjera	TamanoMedioHogar	HogaresMujerSolaMayor65anos	HogaresHombreSoloMayor65anos	HogaresMonoparentalesMujer	HogaresMonoparentalesHombre	EsperanzaVidaNacerMUJERES	EsperanzaVidaNacerHOMBRES
0.6421	0.6411	0.6571	0.6577	0.6831	0.7819	0.2784	0.5033	0.7147	0.6053	0.7108	0.6963	0.7637	0.3688

- *Tablas II.2*: Índice KMO y medida de adecuación MSA (apartado 3.2.3)

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.54527889															
EtapasEducativasInfantil3a5anos	EtapasEducativasPrimaria6a11anos	EtapasEducativasSecundaria12a15a	CentrosPRIVADOSCONCERTADOS	CentrosPUBLICOS	PropAlumnosExtranjeros	ExtranjerosCentrosPRIVADOSCONCER	ExtranjerosCentrosPUBLICOS	PropAlumnosNecesidadesApoyoEduca	AlumnosApoyoEducativoCentrosPRIV	AlumnosApoyoEducativoCentrosPUBL	NoSabeLeerNiEscribirOSinEstudios	PrimariaIncompleta	BachillerElementalGraduadoEscola	TitulosMediosDiplomadosArquite	EstudiosSuperioresLicenciadoDoct
0.3066	0.6728	0.2371	0.2593	0.4506	0.6355	0.3903	0.3408	0.5050	0.6616	0.5900	0.7861	0.7934	0.5973	0.5956	0.6422

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.58332213														
EtapasEducativasInfantil3a5anos	EtapasEducativasPrimaria6a11anos	CentrosPRIVADOSCONCERTADOS	CentrosPUBLICOS	PropAlumnosExtranjeros	ExtranjerosCentrosPRIVADOSCONCER	ExtranjerosCentrosPUBLICOS	PropAlumnosNecesidadesApoyoEduca	AlumnosApoyoEducativoCentrosPRIV	AlumnosApoyoEducativoCentrosPUBL	NoSabeLeerNiEscribirOSinEstudios	PrimariaIncompleta	BachillerElementalGraduadoEscola	TitulosMediosDiplomadosArquite	EstudiosSuperioresLicenciadoDoct
0.3667	0.8001	0.2361	0.4967	0.7685	0.3880	0.3636	0.6976	0.5577	0.5564	0.7355	0.7102	0.6839	0.5328	0.6609

- *Tabla II.3*: Porcentaje de varianza explicada con cada factor para el conjunto de variables explicativas, y desglosado por variables (apartado 4.1)

Percent Variation Accounted for by Partial Least Squares Factors												
Number of Extracted Factors	Model Effects											
	RentaNetaMediaAnualHogares	PensionMediaMensualDistritoHOMBR	PensionMediaMensualDistritoMUJER	TasaAbsolutaParoRegistrado	TasaAbsolutaParoRegistradoMUJERE	ParoMujeresDe16a24anos	ParoMujeresDe25a44anos	ParoMujeresDe45a64anos	TasaAbsolutaParoRegistradoHOMBRE	ParoHombresDe16a24anos	ParoHombresDe25a44anos	ParoHombresDe45a64anos
1	81.2503	88.5884	90.2340	98.9877	96.7382	94.7444	94.0363	94.7871	97.3929	95.4638	94.1882	87.5135
2	83.0451	90.8553	90.4544	99.1181	98.5189	95.3427	96.9273	96.1414	98.1922	96.6777	94.2262	94.4384
3	92.1949	91.6780	90.4778	99.1992	98.7850	95.4731	97.0814	96.6068	98.1936	97.8670	94.2802	94.4449

Percent Variation Accounted for by Partial Least Squares Factors										
Number of Extracted Factors	Model Effects									
	ParadosLargaDuracion	ParadosPercibenPrestaciones	PropPersonasAtendidasUnidaddePri	PerceptoresPrestacionRentaMinima	BeneficiariosPrestacionesSociale	ValorCatastralMedioBienesInmuebl	ValorCatastralMedioBienesInmueb1	SuperficieMediaVivienda	Current	Total
1	24.7173	75.2465	40.4628	74.6016	41.1007	79.5777	60.0369	65.5079	78.7588	78.7588
2	91.8311	76.8331	52.7257	82.1400	52.6712	84.8119	60.7583	69.6116	6.5072	85.2660
3	93.9253	77.8603	86.2197	89.4609	79.2836	85.2341	65.5391	75.6857	4.7085	89.9745

- *Tabla II.4*: Coeficientes para las variables estandarizadas (apartado 4.1)

Parameter Estimates for Centered and Scaled Data				
	PP	PSOE	AhoraMadrid	Ciudadanos
Intercept	0.0000000000	0.0000000000	0.0000000000	0.0000000000
RentaNetaMediaAnualHogares	0.1349337815	0.0059724114	-.2318444985	0.1273357771
PensionMediaMensualDistritoHOMBR	0.1335097068	0.0021622726	-.2246720008	0.1287149315
PensionMediaMensualDistritoMUJER	0.0624042294	-.0850607675	-.0238501237	0.0043259767
TasaAbsolutaParoRegistrado	-.0650397282	0.0583157273	0.0472700595	-.0509724697
TasaAbsolutaParoRegistradoMUJERE	-.0737116378	0.0847068918	0.0449761859	-.0069649671
ParoMujeresDe16a24anos	-.0686333949	0.0732132197	0.0441343508	-.0234651893
ParoMujeresDe25a44anos	-.0687859263	0.0906972478	0.0314668421	0.0031457566
ParoMujeresDe45a64anos	-.0771261789	0.0809402857	0.0547725250	-.0102842771
TasaAbsolutaParoRegistradoHOMBRE	-.0558428515	0.0220517415	0.0564617741	-.1075277689
ParoHombresDe16a24anos	-.0903923851	0.0759103288	0.0838673454	-.0165584772
ParoHombresDe25a44anos	-.0516195118	0.0426914908	0.0336256436	-.0742466428
ParoHombresDe45a64anos	-.0481373862	-.0208570258	0.0751008320	-.1668139587
ParadosLargaDuracion	0.0634942129	0.2702927421	-.3364841777	0.3408550868
ParadosPercibenPrestaciones	0.0795165015	-.0397481387	-.0929814313	0.0592631462
PropPersonasAtendidasUnidaddePri	0.0841411870	-.0098621191	-.1757758920	-.1196981726
PerceptoresPrestacionRentaMinima	0.0041612451	0.0067065816	-.0429337866	-.1199524012
BeneficiariosPrestacionesSociale	0.1285943456	0.0344303066	-.2935914540	-.0568783272
ValorCatastralMedioBienesInmuebl	0.0798113614	-.1334459650	-.0242785270	-.0840795375
ValorCatastralMedioBienesInmueb1	-.0368376024	-.0749114973	0.1516242241	0.0008156549
SuperficieMediaVivienda	0.1182806461	0.0141474613	-.2094948617	0.1266541604

- *Tabla II.5*: Coeficientes para las variables originales (apartado 4.1)

Parameter Estimates				
	PP	**PSOE**	**AhoraMadrid**	**Ciudadanos**
Intercept	-0.48101555	-6.88744144	65.21794161	-5.31464529
RentaNetaMediaAnualHogares	0.00007451	0.00000130	-0.00006879	0.00001711
PensionMediaMensualDistritoHOMBR	0.00605782	0.00003877	-0.00547744	0.00142142
PensionMediaMensualDistritoMUJER	0.00407044	-0.00219260	-0.00083588	0.00006868
TasaAbsolutaParoRegistrado	-0.23776026	0.08424607	0.09284782	-0.04535105
TasaAbsolutaParoRegistradoMUJERE	-0.24269234	0.11021538	0.07956602	-0.00558123
ParoMujeresDe16a24anos	-0.22951143	0.09675241	0.07929953	-0.01909783
ParoMujeresDe25a44anos	-0.20274405	0.10564430	0.04983411	0.00225665
ParoMujeresDe45a64anos	-0.26576970	0.11022290	0.10141249	-0.00862519
TasaAbsolutaParoRegistradoHOMBRE	-0.22403670	0.03496213	0.12171136	-0.10499373
ParoHombresDe16a24anos	-0.32065545	0.10641702	0.15985444	-0.01429612
ParoHombresDe25a44anos	-0.22130806	0.07233160	0.07746026	-0.07747317
ParoHombresDe45a64anos	-0.17227506	-0.02949821	0.14441412	-0.14529927
ParadosLargaDuracion	0.23232259	0.39083649	-0.66152607	0.30354160
ParadosPercibenPrestaciones	0.15778339	-0.03116910	-0.09913460	0.02862067
PropPersonasAtendidasUnidaddePri	0.20542005	-0.00951498	-0.23057835	-0.07112346
PerceptoresPrestacionRentaMinima	0.00731036	0.00465607	-0.04052653	-0.05128795
BeneficiariosPrestacionesSociale	0.30281869	0.03204093	-0.37147469	-0.03259859
ValorCatastralMedioBienesInmuebl	0.00001876	-0.00001240	-0.00000307	-0.00000481
ValorCatastralMedioBienesInmueb1	-0.00000135	-0.00000109	0.00000299	0.00000001
SuperficieMediaVivienda	0.07169842	0.00338905	-0.06823298	0.01868558

- *Tabla II.6*: Predicciones y residuos de los cuatro partidos políticos para cada distrito (apartado 4.1)

Obs	Distritos	PP	PSOE	AhoraMadrid	Ciudadanos	p_PP	p_PSOE	p_AhoraMadrid	p_Ciudadanos	PP_res	PSOE_res	AhoraMadrid_res	Ciudadanos_res
1	CENTRO	17.3407	6.9592	32.1459	5.3157	20.1924	7.6012	28.2117	5.4544	-2.85176	-0.64199	3.93420	-0.13875
2	ARGANZUELA	22.7084	8.6694	28.0494	8.5010	24.1976	8.2112	24.4621	7.8438	-1.48927	0.45827	3.58729	0.65720
3	RETIRO	32.3255	6.8481	21.0368	9.3635	31.8552	6.9539	19.5642	9.8065	0.47026	-0.10578	1.47260	-0.44292
4	SALAMANCA	37.1944	5.9641	14.7904	9.1677	34.6135	6.0445	18.1040	9.7802	2.58098	-0.08036	-3.31362	-0.61246
5	CHAMARTIN	37.6910	5.8918	15.0013	10.3043	35.2764	6.5677	16.9012	10.2645	2.41457	-0.67591	-1.89990	0.03977
6	TETUAN	24.3772	9.5483	18.7450	7.4478	22.5285	10.3852	22.5506	7.1115	1.84874	-0.83695	-3.80556	0.33620
7	CHAMBERI	33.9890	6.3089	18.6005	8.9537	30.9060	6.0643	21.4515	9.0276	3.08296	0.24464	-2.85101	-0.07391
8	FUENCARRAL-EL PARDO	28.2698	9.6747	18.8813	10.3561	30.6812	9.0806	17.6199	9.8559	-2.41141	0.59416	1.26140	0.50020
9	MONCLOA-ARAVACA	32.5314	7.4065	19.6506	9.1455	34.1392	7.1240	17.1236	10.0596	-1.60790	0.28247	2.52706	-0.91412
10	LATINA	22.3774	11.8999	22.5165	6.2511	21.1101	12.2232	20.9673	6.5423	1.26729	-0.32327	1.54918	-0.29120
11	CARABANCHEL	18.7360	12.1328	21.5192	6.2930	18.0988	12.6261	22.8348	5.3893	0.63719	-0.49333	-1.31553	0.90367
12	USERA	14.6078	14.2406	22.2206	5.2623	13.6155	12.8357	27.2198	5.6795	0.99231	1.40489	-4.99914	-0.41716
13	PUENTE DE VALLECAS	11.3711	15.6582	26.4348	4.2447	12.8732	15.1672	23.8185	4.3297	-1.50205	0.49098	2.61621	-0.08499
14	MORATALAZ	22.4420	12.0531	24.2877	7.1976	19.0965	12.6942	23.4077	8.6915	3.34548	-0.64108	0.88003	-1.49384
15	CIUDAD LINEAL	25.4066	10.2604	19.4275	7.8107	25.1653	10.3133	20.3775	7.8853	0.24126	-0.05289	-0.95004	-0.07458
16	HORTALEZA	25.3649	10.2974	20.0709	10.2053	28.9096	9.8312	18.0718	9.4056	-3.54467	0.46621	1.99915	0.79965
17	VILLAVERDE	14.6497	15.5186	22.1491	5.6621	12.2249	14.8545	25.9500	6.3054	2.42483	0.66410	-3.80084	-0.64329
18	VILLA DE VALLECAS	13.6454	12.8089	27.6080	8.3213	15.2999	13.8404	25.0186	7.6393	-1.65444	-1.03157	2.58943	0.68199
19	VICALVARO	14.8919	14.1758	25.9366	8.4575	14.2584	14.6292	25.2307	8.2811	0.63345	-0.45342	0.70598	0.17632
20	SAN BLAS-CANILLEJAS	18.1782	12.4651	22.7715	8.1164	20.6765	12.4184	21.9241	8.1227	-2.49825	0.04662	0.84738	-0.00630
21	BARAJAS	25.1791	9.2278	20.7850	11.4117	27.5586	8.5436	21.8193	10.3132	-2.37956	0.68418	-1.03427	1.09853

- *Tabla II.7*: Porcentaje de varianza explicada con cada factor para el conjunto de variables explicativas, y desglosado por variables (apartado 4.2)

Percent Variation Accounted for by Partial Least Squares Factors													
Number of Extracted Factors	Model Effects												
	De0a14anos	De15a29	De30a44	De45a64	De65a79	PersonasNacionalidadExtranjera	EtapasEducativasInfantil3a5anos	EtapasEducativasPrimaria6a11anos	EtapasEducativasSecundaria12a15a	CentrosPRIVADOSCONCERTADOS	CentrosPUBLICOS	PropAlumnosExtranjeros	ExtranjerosCentrosPRIVADOSCONCER
1	10.5399	31.1287	0.0851	5.9453	38.7329	32.8442	15.9475	38.0076	6.2549	5.8843	55.2987	43.8757	0.6376
2	66.5291	71.2226	0.2204	6.7560	40.9651	66.3931	25.2821	53.9822	10.9248	48.6376	74.1345	67.6135	65.7575
3	69.1296	71.2720	90.4153	12.4776	71.6849	85.0160	41.7358	68.5649	46.9799	52.9426	82.8210	80.8779	75.7923
4	91.2057	71.4817	98.6861	81.6830	75.3469	86.3502	78.4267	68.5720	89.0895	61.1930	92.0512	83.1590	77.7726
5	96.1432	76.0854	98.6875	88.4896	86.0440	87.9654	93.5663	82.6118	94.8440	80.9817	92.2784	83.1963	77.7828

Percent Variation Accounted for by Partial Least Squares Factors												
Number of Extracted Factors	Model Effects											
	ExtranjerosCentrosPUBLICOS	PropAlumnosNecesidadesApoyoEduca	AlumnosApoyoEducativoCentrosPRIV	AlumnosApoyoEducativoCentrosPUBL	NoSabeLeerNiEscribirOSinEstudios	PrimariaIncompleta	BachillerElementalGraduadoEscola	FP2gradoBachillerSuperiorOBUP	TituladosMediosDiplomadosArquite	EstudiosSuperioresLicenciadoDoct	Current	Total
1	15.0599	65.6114	46.5636	57.6424	89.7448	88.0145	92.0784	7.5372	74.7986	86.8847	39.5269	39.5269
2	52.4430	67.5898	60.5294	68.4380	89.8573	88.6330	92.1622	55.7274	92.0069	88.8676	19.3720	58.8988
3	53.2551	68.8453	66.5325	69.8238	92.0820	93.1931	93.3412	55.7307	94.4426	91.0535	11.8842	70.7830
4	66.5512	68.8814	66.6801	69.8364	93.4297	93.3596	95.3303	59.7371	96.5355	91.6371	9.9559	80.7390
5	73.5972	69.2387	71.3304	80.7635	94.3931	95.7239	96.6221	61.1457	96.6867	93.8932	5.0033	85.7422

- *Tabla II.8*: Coeficientes para las variables estandarizadas (apartado 4.2)

Parameter Estimates for Centered and Scaled Data				
	PP	PSOE	AhoraMadrid	Ciudadanos
Intercept	0.0000000000	0.0000000000	0.0000000000	0.0000000000
De0a14anos	0.0415422822	0.1083025895	-.3207384167	0.2782520225
De15a29	-.0076580441	0.0186560571	-.0595196921	-.0649120797
De30a44	-.1498743957	-.0471400000	0.3136433068	-.0911670329
De45a64	-.0683700962	-.0390233786	0.2610564468	-.1317426821
De65a79	0.1081816986	-.0239467523	-.0627165976	-.0586504955
PersonasNacionalidadExtranjera	-.0690137472	0.0041204770	0.0556363199	-.1120169902
EtapasEducativasInfantil3a5anos	-.0412220860	0.0071032172	0.1147862590	-.0372883056
EtapasEducativasPrimaria6a11anos	0.0256470556	0.0908043414	-.2262451108	0.1510456047
EtapasEducativasSecundaria12a15a	0.0418514330	-.0002305635	-.0499652370	-.0153917251
CentrosPRIVADOSCONCERTADOS	0.0931664343	0.0039366574	-.1696932222	-.0598640571
CentrosPUBLICOS	-.1760755894	0.0379858690	0.3129381387	-.1021877271
PropAlumnosExtranjeros	-.0663656736	0.0287203758	0.0096375207	-.0767013045
ExtranjerosCentrosPRIVADOSCONCER	-.0117597248	-.0359927054	0.0268695227	-.1064797404
ExtranjerosCentrosPUBLICOS	-.0858075651	0.0393153700	0.1677416088	-.0340606606
PropAlumnosNecesidadesApoyoEduca	-.1124135858	0.0649911715	0.1331305203	-.1136195115
AlumnosApoyoEducativoCentrosPRIV	0.0732527189	-.0913046722	-.0791594605	0.0594993701
AlumnosApoyoEducativoCentrosPUBL	-.1031122962	0.0988307809	0.1213608326	-.0845744237
NoSabeLeerNiEscribirOSinEstudios	-.1066967500	0.1338862626	0.0788245216	-.1305570138
PrimariaIncompleta	-.0920881872	0.1248934054	0.0656740127	-.1469902296
BachillerElementalGraduadoEscola	-.0833419307	0.1267844971	-.0047792400	-.0687442202
FP2gradoBachillerSuperiorOBUP	-.0405019456	-.0045112684	0.1356016617	0.0451999902
TituladosMediosDiplomadosArquite	0.0445204609	-.1008944853	0.0447030731	0.1217227626
EstudiosSuperioresLicenciadoDoct	0.1016860529	-.1312025159	-.0587403188	0.0957046655

- *Tabla II.9*: Coeficientes para las variables originales (apartado 4.2)

Parameter Estimates				
	PP	**PSOE**	**AhoraMadrid**	**Ciudadanos**
Intercept	50.51562392	3.95971461	1.38574682	12.43028240
De0a14anos	0.11659136	0.12012082	-0.48367320	0.19006647
De15a29	-0.06448060	0.06207754	-0.26927555	-0.13302336
De30a44	-0.38226771	-0.04751527	0.42983445	-0.05659384
De45a64	-0.37581321	-0.08476843	0.77101939	-0.17624772
De65a79	0.37237025	-0.03257401	-0.11599217	-0.04913421
PersonasNacionalidadExtranjera	-0.13338217	0.00314712	0.05777575	-0.05269109
EtapasEducativasInfantil3a5anos	-0.24622298	0.01676707	0.36839484	-0.05420788
EtapasEducativasPrimaria6a11anos	0.10474425	0.14655571	-0.49647447	0.15013845
EtapasEducativasSecundaria12a15a	0.14348505	-0.00031239	-0.09204269	-0.01284325
CentrosPRIVADOSCONCERTADOS	0.07232790	0.00120775	-0.07078413	-0.01131107
CentrosPUBLICOS	-0.10136428	0.00864194	0.09679865	-0.01431778
PropAlumnosExtranjeros	-0.11705385	0.02001869	0.00913340	-0.03292578
ExtranjerosCentrosPRIVADOSCONCER	-0.00811225	-0.00981211	0.00995931	-0.01787733
ExtranjerosCentrosPUBLICOS	-0.04636068	0.00839441	0.04869563	-0.00447887
PropAlumnosNecesidadesApoyoEduca	-0.63793014	0.14575119	0.40593544	-0.15692726
AlumnosApoyoEducativoCentrosPRIV	0.04041462	-0.01990725	-0.02346619	0.00798948
AlumnosApoyoEducativoCentrosPUBL	-0.04969354	0.01882284	0.03142628	-0.00992018
NoSabeLeerNiEscribirOSinEstudios	-0.28793847	0.14278676	0.11429701	-0.08575113
PrimariaIncompleta	-0.15284374	0.08191938	0.05856826	-0.05937770
BachillerElementalGraduadoEscola	-0.08206383	0.04933531	-0.00252855	-0.01647463
FP2gradoBachillerSuperiorOBUP	-0.20752654	-0.00913482	0.37332562	0.05636730
TituladosMediosDiplomadosArquite	0.15999309	-0.14328886	0.08631852	0.10646443
EstudiosSuperioresLicenciadoDoct	0.05951784	-0.03034809	-0.01847341	0.01363358

- *Tabla II.10*: Predicciones y residuos de los cuatro partidos políticos para cada distrito (apartado 4.2)

Obs	Distritos	PP	PSOE	AhoraMadrid	Ciudadanos	p_PP	p_PSOE	p_AhoraMadrid	p_Ciudadanos	PP_res	PSOE_res	AhoraMadrid_res	Ciudadanos_res
1	CENTRO	17.3407	6.9592	32.1459	5.3157	18.3514	7.2197	30.3119	6.0462	-1.01070	-0.26044	1.83398	-0.73051
2	ARGANZUELA	22.7084	8.6694	28.0494	8.5010	22.6257	8.2013	27.8857	8.3142	0.08265	0.46818	0.16361	0.18679
3	RETIRO	32.3255	6.8481	21.0368	9.3635	31.6949	7.6933	21.0792	9.5843	0.63057	-0.84519	-0.04234	-0.22073
4	SALAMANCA	37.1944	5.9641	14.7904	9.1677	37.2860	5.5423	15.5678	9.0609	-0.09155	0.42185	-0.77738	0.10687
5	CHAMARTIN	37.6910	5.8918	15.0013	10.3043	35.6692	6.3681	16.1891	10.4394	2.02177	-0.47628	-1.18784	-0.13514
6	TETUAN	24.3772	9.5483	18.7450	7.4478	21.8094	9.6578	23.0558	6.5320	2.56783	-0.10953	-4.31076	0.91575
7	CHAMBERI	33.9890	6.3089	18.6005	8.9537	32.9106	6.6498	17.8989	8.4689	1.07833	-0.34088	0.70156	0.48482
8	FUENCARRAL-EL PARDO	28.2698	9.6747	18.8813	10.3561	27.5259	9.5657	19.5158	10.2279	0.74385	0.10905	-0.63447	0.12814
9	MONCLOA-ARAVACA	32.5314	7.4065	19.6506	9.1455	31.4217	8.1957	17.6320	9.6277	1.10965	-0.78924	2.01864	-0.48217
10	LATINA	22.3774	11.8999	22.5165	6.2511	22.5746	11.9707	21.7983	6.5012	-0.19724	-0.07074	0.71826	-0.25014
11	CARABANCHEL	18.7360	12.1328	21.5192	6.2930	17.9809	13.2819	22.1744	6.5900	0.75512	-1.14908	-0.65516	-0.29697
12	USERA	14.6078	14.2406	22.2206	5.2623	13.2129	14.9820	23.1025	5.1371	1.39486	-0.74141	-0.88181	0.12518
13	PUENTE DE VALLECAS	11.3711	15.6582	26.4348	4.2447	15.0934	14.2375	23.5240	4.3811	-3.72226	1.42072	2.91077	-0.13636
14	MORATALAZ	22.4420	12.0531	24.2877	7.1976	21.8070	11.8909	24.9757	7.1763	0.63497	0.16220	-0.68799	0.02132
15	CIUDAD LINEAL	25.4066	10.2604	19.4275	7.8107	28.5931	9.3379	18.5416	7.7428	-3.18646	0.92247	0.88592	0.06788
16	HORTALEZA	25.3649	10.2974	20.0709	10.2053	28.8582	9.4879	17.6290	10.5755	-3.49329	0.80952	2.44199	-0.37018
17	VILLAVERDE	14.6497	15.5186	22.1491	5.6621	12.9866	15.1458	23.9324	6.3076	1.66304	0.37284	-1.78328	-0.64546
18	VILLA DE VALLECAS	13.6454	12.8089	27.6080	8.3213	13.8943	12.7989	26.8755	8.0076	-0.24887	0.00999	0.73251	0.31371
19	VICALVARO	14.8919	14.1758	25.9366	8.4575	14.1172	14.6018	24.2567	8.3369	0.77467	-0.42598	1.67995	0.12062
20	SAN BLAS-CANILLEJAS	18.1782	12.4651	22.7715	8.1164	18.0799	12.7289	24.5846	7.3535	0.09834	-0.26380	-1.81304	0.76290
21	BARAJAS	25.1791	9.2278	20.7850	11.4117	26.7844	8.4520	22.0982	11.3780	-1.60528	0.77577	-1.31312	0.03369