



---

**Universidad de Valladolid**

Facultad de Ciencias

## TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA  
2019/2020

Caracterización de los municipios de Castilla y  
León atendiendo a factores demográficos

Autor: Álvaro Cardona Arévalo

Tutores:

Jesús Alberto Tapia García

Jesús María Rodríguez Rodríguez



### **Resumen**

En este trabajo de fin de grado se va a realizar una clasificación de los municipios de Castilla y León en función de factores demográficos. Se emplearán datos obtenidos a partir del módulo de datos básicos del Sistema de Información Estadística de la Junta de Castilla y León del año 2018. La clasificación obtenida va a permitir describir características comunes de los municipios. Para ello, se aplicarán diversas técnicas multivariantes gracias al lenguaje de programación R.



### **Abstract**

The present dissertation includes a classification of the municipalities of “Castilla y León” according to demographic factors. It will be used the data obtained from the basic data module of the Statistical Information System of the ”Junta de Castilla y León” for the year 2018. The classification obtained will make it possible to describe common characteristics of the municipalities. For this purpose, it will be applied several multivariate techniques thanks to the programming language R.

## Índice

<b>1. Introducción</b>	<b>10</b>
<b>2. Obtención de la base de datos</b>	<b>11</b>
<b>3. Análisis estadístico descriptivo</b>	<b>13</b>
<b>4. Análisis estadístico multivariante</b>	<b>17</b>
4.1. Descripción de las técnicas estadísticas . . . . .	18
4.1.1. Análisis factorial . . . . .	18
4.1.2. Análisis de Componentes Principales . . . . .	19
4.1.3. Análisis Cluster . . . . .	20
4.2. Aplicación de las técnicas estadísticas . . . . .	22
4.2.1. Análisis Factorial . . . . .	22
4.2.2. Análisis de Componentes Principales . . . . .	28
4.2.3. Análisis Cluster . . . . .	44
<b>5. Conclusiones</b>	<b>66</b>

## Índice de figuras

1.	Consulta de los datos . . . . .	11
2.	Consulta de los datos . . . . .	11
3.	Diagrama de barras de la población total . . . . .	13
4.	Diagrama de barras del número de nacimientos . . . . .	14
5.	Diagrama de barras de emigraciones a municipios de la misma provincia . . . . .	15
6.	Ecuación de la suma total del cuadrado dentro del cluster . . . . .	21
7.	Matriz de correlaciones de las variables de estadística de población . . . . .	22
8.	Matriz de correlaciones de las variables de estadística de movimiento natural de la población . . . . .	23
9.	Matriz de correlaciones de las variables de estadística de migraciones . . . . .	24
10.	Gráfico de dispersión de los Scores de las variables de estadística de migraciones . . . . .	25
11.	Gráfico de dispersión de los municipios según los factores para el grupo de variables de estadística de migraciones . . . . .	26
12.	Matriz de correlaciones de las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	27
13.	Biplot para dos componentes del ACP de las variables de estadística de población . . . . .	29
14.	Biplot para dos componentes del ACP de las variables de estadística de población . . . . .	29
15.	Biplot para dos componentes del ACP de las variables de estadística de población (sin capitales de provincia) . . . . .	31
16.	Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población . . . . .	32
17.	Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población . . . . .	33
18.	Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia) . . . . .	34
19.	Biplot para dos componentes del ACP de las variables de estadística de migraciones . . . . .	36
20.	Biplot para dos componentes del ACP de las variables de estadística de migraciones . . . . .	36
21.	Biplot para dos componentes del ACP de las variables de estadística de migraciones (sin capitales de provincia) . . . . .	38
22.	Biplot para dos componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	40
23.	Biplot para dos componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	40
24.	Biplot para dos componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia) . . . . .	43
25.	Grupos principales para las variables de estadística de población . . . . .	46
26.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de población . . . . .	47
27.	Grupos principales para las variables de estadística de población (no jerárquico) . . . . .	48
28.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de población (no jerárquico) . . . . .	49
29.	Grupos principales para las variables de estadística de movimiento natural de población . . . . .	51
30.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de movimiento natural de población . . . . .	52

31.	Grupos principales para las variables de estadística de movimiento natural de población (No Jerárquico) . . . . .	53
32.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de movimiento natural de población (No Jerárquico) . . .	54
33.	Grupos principales para las variables de estadística de migraciones (10 clusters) . .	56
34.	Grupos principales para las variables de estadística de migraciones (11 clusters) . .	56
35.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de migraciones . . . . .	57
36.	Grupos principales para las variables de estadística de migraciones (No Jerárquico) .	58
37.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de migraciones (No Jerárquico) . . . . .	59
38.	Grupos principales para las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	62
39.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el conjunto de variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	63
40.	Grupos principales para las variables del Grupo 1, Grupo 2 y Grupo 3 (No Jerárquico)	64
41.	Mapa de CyL que muestra el grupo al que pertenece cada municipio para el conjunto de variables del Grupo 1, Grupo 2 y Grupo 3 (No Jerárquico) . . . . .	65

## Índice de tablas

1.	Resumen estadístico de la población total . . . . .	13
2.	Resumen estadístico del número de nacimientos . . . . .	14
3.	Resumen estadístico de las emigraciones a municipios de la misma provincia . . . . .	15
4.	KMO test de las variables de estadística de movimiento natural de la población . . . . .	23
5.	Análisis factorial de las variables de estadística de movimiento natural de la población . . . . .	23
6.	KMO test de las variables de estadística de migraciones . . . . .	24
7.	Análisis factorial de las variables de estadística de migraciones . . . . .	25
8.	Análisis factorial de las variables de estadística de migraciones . . . . .	26
9.	ACP de las variables de estadística de población . . . . .	28
10.	Variabilidad explicada por las componentes del ACP de las variables de estadística de población . . . . .	30
11.	ACP de las variables de estadística de población (sin capitales de provincia) . . . . .	30
12.	Variabilidad explicada por las componentes del ACP de las variables de estadística de población (sin capitales de provincia) . . . . .	31
13.	ACP de las variables de estadística de movimiento natural de población . . . . .	32
14.	Variabilidad explicada por las componentes del ACP de las variables de estadística de movimiento natural de población . . . . .	33
15.	ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia) . . . . .	34
16.	Variabilidad explicada por las componentes del ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia) . . . . .	35
17.	ACP de las variables de estadística de migraciones . . . . .	35
18.	Variabilidad explicada por las componentes del ACP de las variables de estadística de migraciones . . . . .	37
19.	ACP de las variables de estadística de migraciones (sin capitales de provincia) . . . . .	37
20.	Variabilidad explicada por las ocho componentes del ACP de las variables de estadística de migraciones (sin capitales de provincia) . . . . .	38
21.	ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	39
22.	Variabilidad explicada por cada una de las catorce componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	41
23.	ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia) . . . . .	42
24.	Variabilidad explicada por las componentes del ACP de las variables de estadística de Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia) . . . . .	43
25.	Coefficientes de los métodos de agregación para las variables de estadística de población . . . . .	45
26.	Coefficientes de los métodos de agregación para las variables de estadística de movimiento natural de la población . . . . .	50
27.	Coefficientes de los métodos de agregación para las variables de estadística de migraciones . . . . .	55
28.	Coefficientes de los métodos de agregación para las variables del Grupo 1, Grupo 2 y Grupo 3 . . . . .	61



## 1. Introducción

El objetivo principal de este trabajo de fin de grado es la clasificación de los municipios de Castilla y León utilizando como información diversos factores demográficos del año 2018. El análisis estadístico se va a realizar a nivel de desagregación municipal.

Para llevar a cabo este análisis se utilizan los datos que ofrece el Sistema de Información Estadística de la Junta de Castilla y León (SIE), que se pueden encontrar en la página web de la Junta de Castilla y León. El SIE está estructurado en función de módulos temáticos, en los que se agrupa información con características comunes. En este caso se ha decidido utilizar el módulo de “Datos Básicos”, que contiene información acerca de agricultura, turismo, etc., a nivel provincial y municipal. Para este estudio se utilizará la información de la sección “Indicadores Demográficos”, donde están disponibles diferentes características demográficas como población, nacimientos y defunciones, emigraciones e inmigraciones, de los diferentes municipios de Castilla y León. [1]

Una vez obtenida la base de datos conformada por 16 variables, correspondientes a los diferentes indicadores y 2248 registros, que hacen referencia a cada uno de los municipios de Castilla y León, el estudio se dividirá en dos fases. En la primera se realizará el análisis estadístico utilizando diversas técnicas, con el objetivo de agrupar el conjunto de observaciones en diferentes subconjuntos según su similitud. En la segunda fase, se realizará una representación gráfica de los grupos obtenidos en la fase anterior, a través de un mapa de Castilla y León, en el que se mostrará cada municipio con un color, que representa el grupo al que ese municipio ha sido asignado. De esta forma, el resultado del estudio será muy visual y fácilmente interpretable.

Para este estudio se utilizarán diversas herramientas, como Excel y el entorno de desarrollo R-Studio.

## 2. Obtención de la base de datos

Una vez en el módulo de “Datos Básicos” del SIE, para obtener los datos necesarios para este estudio se debe rellenar una consulta personalizada de la información requerida para el análisis estadístico. Mediante esta consulta se seleccionarán los años, las provincias o municipios y los indicadores de los que se quiere obtener la información (*Figura 1*).

Figura 1: Consulta de los datos

La consulta realizada en este caso corresponde al año 2018, engloba todos los municipios de Castilla y León, y la familia de indicadores “Indicadores Demográficos”.

FECHA	2018							
INDICADOR	POBLACIÓN DE DERECHO (TOTAL)	POBLACIÓN DE DERECHO (MUJERES)	POBLACIÓN DE DERECHO (VARONES)	NACIMIENTOS	DEFUNCIONES	MATRIMONIOS DE DIFERENTE SEXO	EMIGRACIONES CON DESTINO A LA MISMA PROVINCIA	INMIGRACIONES PROCEDENTES DE LA MISMA PROVINCIA
MUNICIPIO								
05001 ADANERO	202	89	113	1	2	.	2	.
05002 ADRADA (LA)	2.517	1.219	1.298	19	26	5	35	17
05005 ALBORNOS	190	90	100	.	5	.	2	1
05007 ALDEANUEVA DE SANTA CRUZ	119	56	63	.	2	.	1	.
05008 ALDEASECA	233	118	115	.	.	1	1	6
05010 ALDEHUELA (LA)	168	83	85	.	7	.	2	2
05012 AMAVIDA	135	66	69	2	2	1	1	2
05013 ARENAL (EL)	972	477	495	3	9	3	4	13
05014 ARENAS DE SAN PEDRO	6.454	3.319	3.135	36	86	25	47	71
05015 AREVALILLO	75	40	35	.	2	.	1	3

Figura 2: Consulta de los datos

En la *Figura 2* se encuentra parte de la información obtenida a través de nuestra consulta. Para

que la lectura de los datos sea posible mediante R-Studio, estos datos deben tener un formato diferente al original, por lo que antes de comenzar con el análisis estadístico, es de gran importancia realizar un acondicionamiento de estos datos.

Se puede observar en los datos, bastantes registros, correspondientes generalmente a municipios pequeños, de los cuales no se encuentra información para todos los indicadores. Estos campos vienen señalados con un ".". Si se eliminan las observaciones con ausencia de datos, se perdería una gran cantidad de información correspondiente a municipios pequeños, por lo que el estudio no sería representativo. Lo que se hará, por tanto, será rellenar aquellos campos sin datos con 0's.

También se puede apreciar que la variable indicadora del municipio cuenta con dos formas de identificar este municipio, un número formado por 5 dígitos y el nombre del municipio. Se separará esta información en dos variables por cuestión de comodidad y mayor claridad en los gráficos.

Por último, para completar la preparación de los datos, con R-Studio se elimina la última observación, que corresponde a la suma de todas las observaciones de una variable. Esto es, debido a que esta observación es combinación lineal directa de todas las anteriores, lo que llevaría a un problema de colinealidad.

### 3. Análisis estadístico descriptivo

Para conocer con más detalle la información de las variables de las que se dispone, se realizará un análisis descriptivo.

- Variables identificadoras:
  - *INDICADOR*: Variable numérica de 5 cifras. Identificador único para cada municipio.
  - *MUNICIPIO*: Variable alfanumérica que hace referencia al nombre del municipio.
- Variables a estudiar (numéricas):
  - *POBLACIÓN TOTAL (P\_TOT)*: Hace referencia al número total de habitantes de un municipio.

Media	Moda	Desviación	Varianza	Min.	Max.	1st cuartil	3rd cuartil
1071.7	34/61	90062.732	82133103	6.0	298866.0	84.0	409.2

Tabla 1: Resumen estadístico de la población total

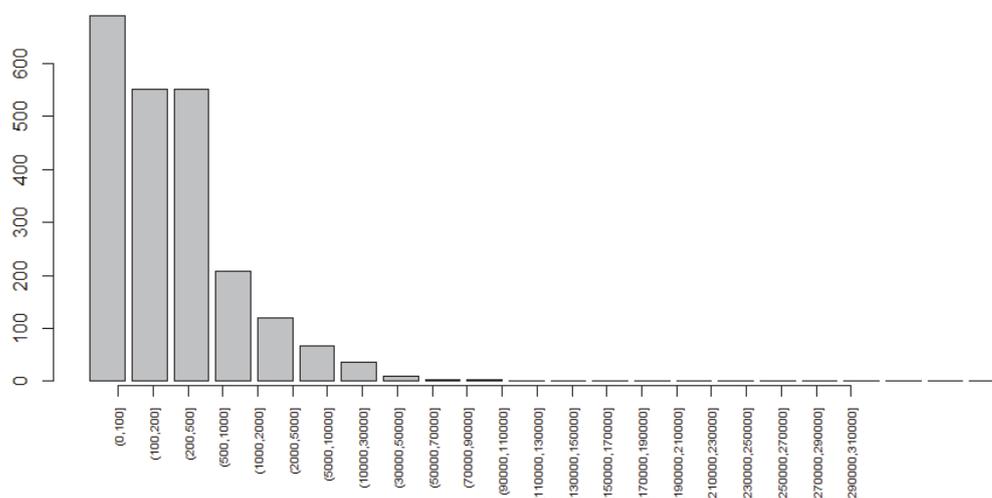


Figura 3: Diagrama de barras de la población total

A partir de varios estadísticos (*Tabla 1*) se puede intuir que en Castilla y León existen un gran número de municipios pequeños, ya que el 75% de ellos no tienen más de 409 habitantes. La cantidad de población más repetida (moda) corresponde a municipios de 34 y 61 habitantes, 15 veces repetidos. La *Figura 3*, muestra la cantidad de municipios de Castilla y León con una población determinada por intervalos. Se puede ver que predominan los municipios de 0 a 100 habitantes, de los que se pueden encontrar más de 700.

- *POBLACIÓN HOMBRES (P\_HOM)*: Hace referencia al número de hombres que hay en cada municipio de Castilla y León.
- *POBLACIÓN MUJERES (P\_MUJ)*: Hace referencia al número de mujeres que hay en cada municipio en Castilla y León.
- *NACIMIENTOS (NAC)*: Indicador del número de nacimientos que tienen lugar en un municipio a lo largo de un año.

Media	Moda	Desviación	Varianza	Min.	Max.	1st cuartil	3rd cuartil
6.692	0	61.91302	3833.223	0.0	1927.0	0.000	2.000

Tabla 2: Resumen estadístico del número de nacimientos

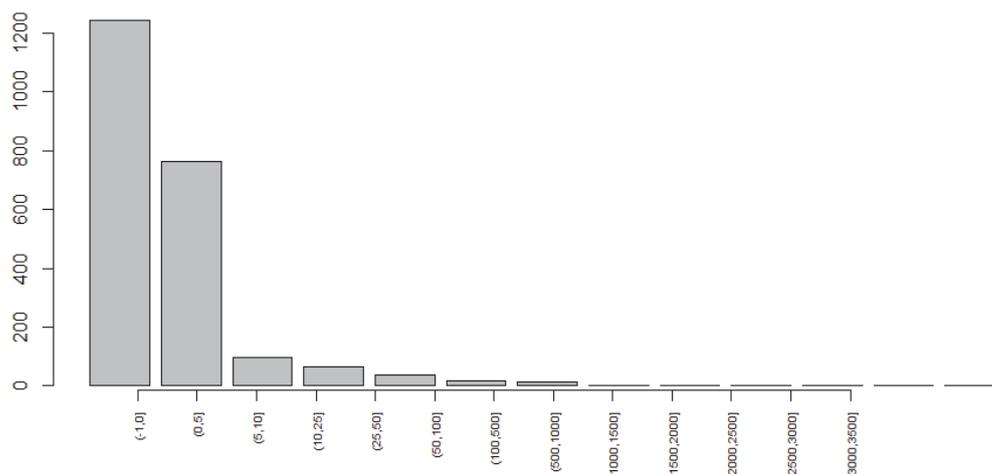


Figura 4: Diagrama de barras del número de nacimientos

Si se observan los estadísticos de la *Tabla 2*, llama la atención que la media de nacimientos es claramente pequeña, y la moda 0, lo que indica que es el número de nacimientos más repetido. También se observa que el 75% de los municipios cuenta con 2 nacimientos como máximo, de nuevo indicador de la existencia de una gran cantidad de municipios pequeños.

En la *Figura 4* se muestra que hay más de 1200 municipios de Castilla y León que en 2018 no contaron con ningún nacimiento.

- *DEFUNCIONES (DEF)*: Indicador del número de defunciones que tienen lugar en un municipio a lo largo de un año.
- *MATRIMONIOS DE DISTINTO SEXO (MAT)*: Indicador del número de matrimonios de diferente sexo que tienen lugar en un municipio a lo largo de un año.

- *EMIGRACIONES A LA MISMA PROVINCIA (E\_A\_M\_P)*: Hace referencia al número de personas que emigran a un municipio de la misma provincia a lo largo de un año.

Media	Moda	Desviación	Varianza	Min.	Max.	1st cuartil	3rd cuartil
18.33	0	117.3112	13761.92	0.0	3334.00	1.00	10.00

Tabla 3: Resumen estadístico de las emigraciones a municipios de la misma provincia

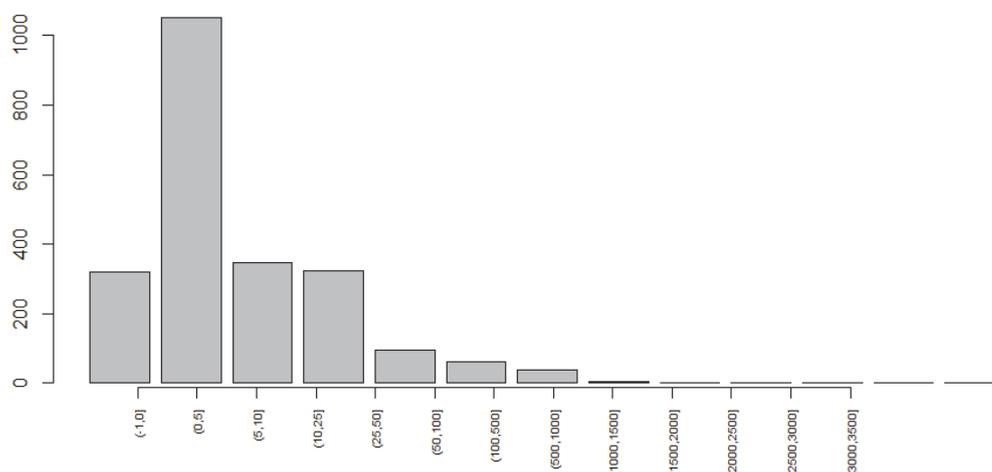


Figura 5: Diagrama de barras de emigraciones a municipios de la misma provincia

Los estadísticos de la *Tabla 3* muestran un valor de la moda igual a 0. El 75% de los municipios cuentan como máximo con 10 emigraciones a otro municipio de la misma provincia. En el gráfico de barras (*Figura 5*) se puede ver que el número de emigraciones en 2018 fue 0 en más de 300 municipios, no obstante, el pico lo encontramos entre 1 y 5 emigraciones, en más de 1000 municipios. De nuevo, estos indicadores ponen en manifiesto que en Castilla y León existe una cantidad de pueblos pequeños muy notable.

- *INMIGRACIONES DE LA MISMA PROVINCIA (I\_D\_M\_P)*: Hace referencia al número de personas que inmigran desde un municipio de la misma provincia a lo largo de un año.
- *EMIGRACIONES A OTRA PROVINCIA (E\_A\_O\_P)*: Hace referencia al número de personas que emigran a un municipio de otra provincia a lo largo de un año.
- *INMIGRACIONES DE OTRA PROVINCIA (I\_D\_O\_P)*: Hace referencia al número de personas que inmigran desde un municipio de otra provincia a lo largo de un año.

- *EMIGRACIONES A OTRA COMUNIDAD (E\_A\_O\_C)*: Hace referencia al número de personas que emigran a un municipio de otra comunidad a lo largo de un año.
- *INMIGRACIONES DE OTRA COMUNIDAD (I\_D\_O\_C)*: Hace referencia al número de personas que inmigran desde un municipio de otra comunidad a lo largo de un año.
- *EMIGRACIONES A OTRO PAÍS (E\_A\_O\_PAI)*: Hace referencia al número de personas que emigran a un municipio de otro país a lo largo de un año.
- *INMIGRACIONES DE OTRO PAÍS (I\_D\_O\_PAI)*: Hace referencia al número de personas que inmigran desde un municipio de otro país a lo largo de un año.

Tras esta breve descripción de las variables empleadas para el estudio, y al observar el análisis descriptivo realizado a alguna de ellas, se puede llegar a la conclusión de que son variables altamente correlacionadas, ya que todas ellas dependen de la población total.

## 4. Análisis estadístico multivariante

Para realizar este análisis, se utilizarán diferentes técnicas estadísticas. En primer lugar, se cuenta con un gran número de variables, por lo que se probará a realizar varios análisis agrupando las variables relacionadas. Posteriormente se realizará un análisis con todas las variables para comparar con los análisis por grupos.

Los grupos de variables que se van a utilizar son:

### 1. Variables de estadística de población

- Población Total (P\_TOT)
- Población Hombres (P\_HOM)
- Población Mujeres (P\_MUJ)

### 2. Variables de estadística de movimiento natural de la población

- Nacimientos (NAC)
- Defunciones (DEF)
- Matrimonios del mismo sexo (MAT)

### 3. Variables de estadística de migraciones

- Emigraciones a municipios de la misma provincia (E\_A\_M\_P)
- Inmigraciones desde municipios de la misma provincia (I\_D\_M\_P)
- Emigraciones a municipios de otra provincia (E\_A\_O\_P)
- Inmigraciones desde municipios de otra provincia (I\_D\_O\_P)
- Emigraciones a municipios de otra comunidad (E\_A\_O\_C)
- Inmigraciones desde municipios de otra comunidad (I\_D\_O\_C)
- Emigraciones a municipios de otro país (E\_A\_O\_PAÍ)
- Inmigraciones desde municipios de otro país (I\_D\_O\_PAÍ)

### 4. Variables de Grupo 1, Grupo 2 y Grupo 3

## 4.1. Descripción de las técnicas estadísticas

### 4.1.1. Análisis factorial

El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Estos grupos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros. Fundamentalmente lo que se pretende con el Análisis Factorial es simplificar la información que nos da una matriz de correlaciones para hacerla más fácilmente interpretable. [2]

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. [3]

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. [2]

En primer lugar, para realizar el Análisis Factorial, se debe calcular la matriz de correlaciones, para comprobar el grado de asociación de las variables. Esto se puede conseguir a través de varios métodos:

- Evaluación del determinante de la matriz de correlaciones: Un determinante nulo indicaría dependencia lineal entre las variables, por lo que no sería válido. No obstante, un determinante bajo sería lo óptimo, ya que es indicador de correlaciones altas entre las variables. [4]
- Test de Bartlett: La prueba de Bartlett permite contrastar si más de 2 muestras presentan igualdad de varianzas. Se utiliza para probar la hipótesis nula,  $H_0$ : todas las varianzas de una población  $k$  son iguales, frente a la hipótesis alternativa de que al menos dos son diferentes. [5]
- Índice de Kaiser-Meyer-Olkin (KMO): se utiliza para comparar las magnitudes de los coeficientes de correlación parcial, de forma que cuánto más pequeño sea su valor, mayor será el valor de los coeficientes de correlación parciales  $r_{ij}(p)$  y, en consecuencia, menos apropiado es realizar un Análisis Factorial. [2]

$$KMO < 0,5 \Rightarrow \text{Inaceptable} \quad (1)$$

$$KMO \geq 0,5 \Rightarrow \text{Aceptable} \quad (2)$$

$$KMO \geq 0,75 \Rightarrow \text{Bien} \quad (3)$$

Por último, se pueden rotar los factores para mejorar su interpretación. Así, cada variable estará altamente correlacionada con un único factor, y poco correlacionada con el resto. Tanto las communalidades como la proporción explicada no varían, al contrario que las cargas factoriales. La rotación *Varimax* es recomendable ante un número de factores reducido; la *Quartimax*, cuando el número de factores es elevado. [4]

#### 4.1.2. Análisis de Componentes Principales

El método de componentes principales tiene por objeto transformar un conjunto de variables, a las que se denomina originales, en un nuevo conjunto de variables denominadas componentes principales. Estas últimas se caracterizan por estar incorrelacionadas entre sí y, además, pueden ordenarse de acuerdo con la información que llevan incorporada. Por esto, cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes.[6]

Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la cantidad de información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquella que tenga mayor varianza, mientras que la última componente es la de menor varianza. [6]

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.

La extracción de componentes principales se efectúa sobre variables tipificadas, para evitar problemas de escala. En este estudio, todas las variables tienen la misma unidad de medida, por lo que no sería necesario realizar un proceso de tipificación. [6]

Por último, el nuevo conjunto de variables que se obtiene por el método de componentes principales es igual en número al de las variables originales, ya que la suma de sus varianzas es igual a la suma de las varianzas de las variables originales. Por tanto, quedarse con todas ellas no simplificaría el problema, por lo que deberemos seleccionar aquellas componentes que, siendo pocas e interpretables, expliquen una proporción aceptable de la varianza global. [6]

### 4.1.3. Análisis Cluster

El análisis Cluster es una técnica estadística multivariante que busca agrupar variables tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Los grupos obtenidos son mutuamente excluyentes, es decir, un municipio pertenecerá a un único grupo. [7]

La variedad de formas de algoritmos de clasificación y de medir diferencias multivariadas o distancias entre individuos proporciona diversas posibilidades de análisis. Puesto que la utilización del análisis Cluster ya implica un desconocimiento o conocimiento incompleto de la clasificación de los datos, emplearemos varios métodos, con el fin de contrastar los resultados. [7]

Existen dos grandes tipos de análisis de clusters: no jerárquicos y jerárquicos.

Los métodos o agrupamientos jerárquicos van generando grupos en cada una de las fases del proceso buscando el número de clúster que hacer una agrupación óptima. [8]

El agrupamiento jerárquico es capaz de fijar por sí solo el número de clusters, por ello se puede utilizar de forma exploratoria y posteriormente aplicar un análisis no jerárquico con el número de clúster ya fijado. [8]

Para realizar un Análisis Cluster Jerárquico se deben seguir varios pasos:

1. Partimos de número de grupos igual al de observaciones, de forma que cada municipio representará un grupo.
2. Se genera una matriz simétrica que contiene las similitudes o distancias entre todos los pares de observaciones.
3. A partir de esta matriz, se agrupan las dos observaciones más próximas entre sí. Este proceso es recursivo, de forma que el número de clusters será uno menos en cada paso, hasta llegar a un único Cluster, que contenga todas las observaciones. [9]

Este proceso de aglomeración de los grupos puede verse gráficamente a través de un dendograma, a partir del cual podemos decidir con cuántos clusters nos quedaremos finalmente.

En el caso del agrupamiento no jerárquico, el número de clusters debe conocerse previamente. El algoritmo más utilizado es el de las K-medias, el cual clasifica los objetos en varios grupos, de modo que los objetos dentro del mismo grupo son lo más similares posible, mientras que los objetos de diferentes grupos son lo más diferentes posible. En la agrupación de k-medias, cada grupo está representado por su centro (centroide) que corresponde a la media de los puntos asignados al grupo. [10]

Para realizar un Análisis Cluster No Jerárquico con el algoritmo de las K-medias se deben seguir varios pasos:

1. Especificar el número de grupos (K) que se crearán.

2. Seleccionar al azar  $K$  objetos del conjunto de datos como los centros o medios iniciales del cluster.
3. Asignar cada observación a su centroide más cercano, en función de la distancia euclidiana entre el objeto y el centroide.
4. Para cada uno de los  $K$  grupos, actualizar el centroide del grupo calculando los nuevos valores medios de todos los puntos de datos en el grupo. El centroide de un grupo  $K$ th es un vector de longitud  $p$  (siendo  $p$  el número de variables) que contiene las medias de todas las variables para las observaciones en el grupo  $K$ th.
5. Minimizar iterativamente el total dentro de la suma del cuadrado *Figura 6*. Es decir, repetir los pasos 3 y 4 hasta que las asignaciones del cluster dejen de cambiar o se alcance el número máximo de iteraciones. [10]

$$\sum_{k=1}^k W(C_k) = \sum_{k=1}^k \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Figura 6: Ecuación de la suma total del cuadrado dentro del cluster

Tras el análisis cluster, se realizará un mapa representativo del resultado de la agrupación. Esto es, un mapa que muestra todos los municipios de Castilla y León, y les asigna un color en función del grupo obtenido por el análisis.

Para la realización de este mapa se ha empleado la librería de R-Studio “sp”. Esta librería proporciona el mapa de Castilla y León con todos los municipios. Para asignar el grupo a cada municipio del mapa, se realizará un merge, es decir, una combinación de los datos en una sola tabla, de la tabla que contiene cada municipio representado por el mapa y la tabla que contiene el resultado del agrupamiento. Para este paso es necesario realizar un tratamiento de los datos. En función de las fuentes de las que se obtienen los datos, es posible que se encuentren municipios escritos de forma diferente, problema que se debe solucionar para posteriormente encontrar las coincidencias al combinar las tablas.

Para el análisis cluster se prescindirá de las capitales de provincia, buscando una mayor claridad de los resultados. Estos municipios aparecerán en el mapa, aunque no pertenecerán a ningún grupo.

Por último, se añadirá una paleta de colores al mapa, para colorear los municipios en función del número al que pertenezcan. Para una mayor claridad, se utilizará una escala de degradados, en la que los colores más oscuros representan municipios con datos más elevados y los colores más claros, municipios con estadísticas menores.

## 4.2. Aplicación de las técnicas estadísticas

### 4.2.1. Análisis Factorial

Tal y como se ha indicado anteriormente, aparte de realizar un análisis global sobre todas las variables, se llevarán a cabo diferentes análisis en función de subgrupos.

#### 1. Variables de estadística de población

Este subgrupo hace referencia a la población total de cada uno de los municipios, así como de hombres y mujeres.

En la *Figura 7* se puede apreciar la correlación que existe entre las variables. En este caso, la población total corresponde a la suma de las otras dos variables, por lo que no tendría sentido aplicar análisis factorial, ya que los datos se encuentran en un espacio de dimensión 2.

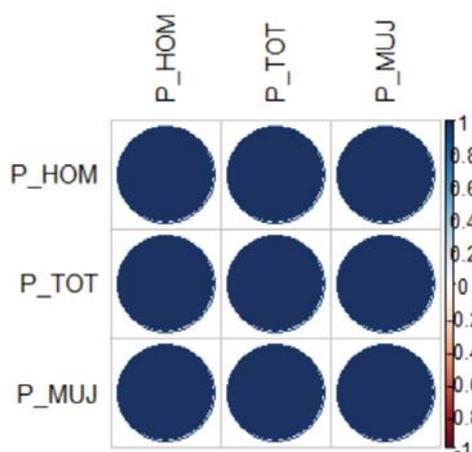


Figura 7: Matriz de correlaciones de las variables de estadística de población

#### 2. Variables de estadística de movimiento natural de la población

Este subgrupo engloba diversas características como el número de nacimientos, de defunciones y de matrimonios del mismo sexo que tienen lugar en un municipio.

En este caso también se puede apreciar una correlación muy alta entre las variables “Nacimientos”, “Defunciones” y “Matrimonios de diferente sexo” (*Figura 8*), no obstante, las correlaciones no son tan próximas a uno como en el subgrupo anterior, por lo que se calculará el índice de KMO y la medida de adecuación MSA para comprobar si se puede realizar el análisis factorial.

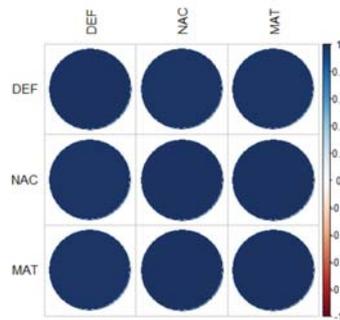


Figura 8: Matriz de correlaciones de las variables de estadística de movimiento natural de la población

Se puede observar un KMO de 0.76 (Tabla 4), que, junto a los MSA de las variables, que se pueden considerar aceptables, son indicadores de que sería adecuado realizar un análisis factorial.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = datagenerales)
Overall MSA = 0.76
MSA for each item =
  NAC DEF MAT
0.68 0.96 0.70
```

Tabla 4: KMO test de las variables de estadística de movimiento natural de la población

```
Call:
factanal(x = datagenerales, factors = 1, score
s = c("regression"), rotation = "varimax")

Uniquenesses:
  NAC DEF MAT
0.005 0.028 0.005

Loadings:
  Factor1
NAC 0.998
DEF 0.986
MAT 0.998

SS loadings      Factor1
Proportion Var  2.965
                0.988

The degrees of freedom for the model is 0 and
the fit was 0.1959
```

Tabla 5: Análisis factorial de las variables de estadística de movimiento natural de la población

En este grupo, al contar únicamente con tres variables, con un solo factor se explica prácticamente toda la variabilidad, por lo que parece evidente que quedarse con este factor sería suficiente.

### 3. Variables de estadística de migraciones

Para este último subgrupo también se aprecia una alta correlación entre todas las variables, entre 0.85 y 0.99 (*Figura 9*). En primer lugar, se obtiene el índice KMO para comprobar que se puede realizar el análisis factorial.

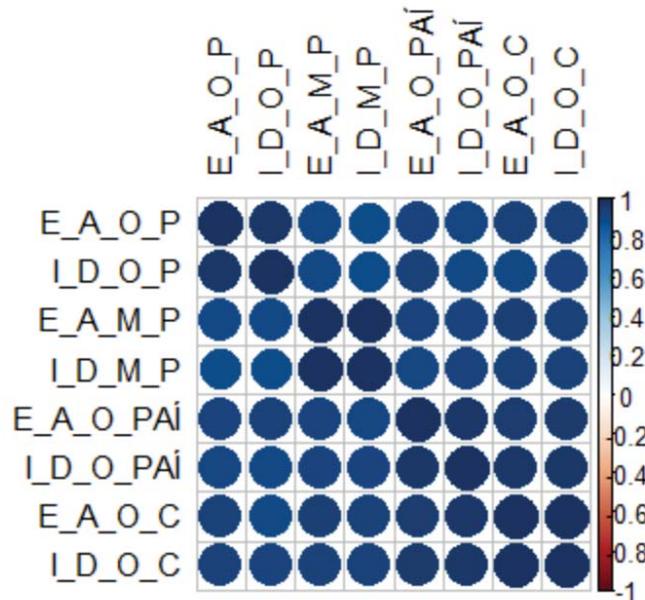


Figura 9: Matriz de correlaciones de las variables de estadística de migraciones

En este grupo, se aprecia un KMO bastante elevado (*Tabla 6*), así como los MSA de las variables, por tanto, a partir de estos resultados se puede concluir en que es posible realizar un análisis factorial.

```

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = datamovimientos)
Overall MSA = 0.83
MSA for each item =
  E_A_M_P  I_D_M_P  E_A_O_P  I_D_O_P  E_A_O_C
  0.83    0.83    0.81    0.78    0.81
  I_D_O_C  E_A_O_PAÍ  I_D_O_PAÍ
  0.83    0.85    0.89
    
```

Tabla 6: KMO test de las variables de estadística de migraciones

En este caso se ha aplicado el método de rotación ortogonal *varimax*, que minimiza el número de variables que tienen cargas altas en cada factor.

Se puede observar en la *Tabla 7* que con 2 factores se obtiene un 95.2% de variabilidad explicada. El primer factor tiene peso sobre las variables que refieren a la emigración hacia otras

provincias, comunidades y países, y a la inmigración procedente de otras provincias, comunidades y países, mientras que el segundo factor tiene más peso en las variables que refieren a la emigración a un municipio de la misma provincia y la inmigración procedente de un municipio de la misma provincia.

Loadings:

	Factor1	Factor2
E_A_M_P	0.601	0.797
I_D_M_P	0.587	0.807
E_A_O_P	0.744	0.571
I_D_O_P	0.713	0.594
E_A_O_C	0.824	0.560
I_D_O_C	0.832	0.550
E_A_O_PAÍ	0.781	0.570
I_D_O_PAÍ	0.812	0.549

	Factor1	Factor2
SS loadings	4.409	3.208
Proportion Var	0.551	0.401
Cumulative var	0.551	0.952

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 5935.41 on 13 degrees of freedom.  
The p-value is 0

Tabla 7: Análisis factorial de las variables de estadística de migraciones

En los siguientes gráficos se muestran los scores de ambos factores (*Figura 10*) y la agrupación de los municipios según los dos factores (*Figura 11*). Observando el primer gráfico se puede observar la poca utilidad del segundo factor, por lo que para comprobar si se puede prescindir de este factor, se probará este análisis factorial con un factor único.

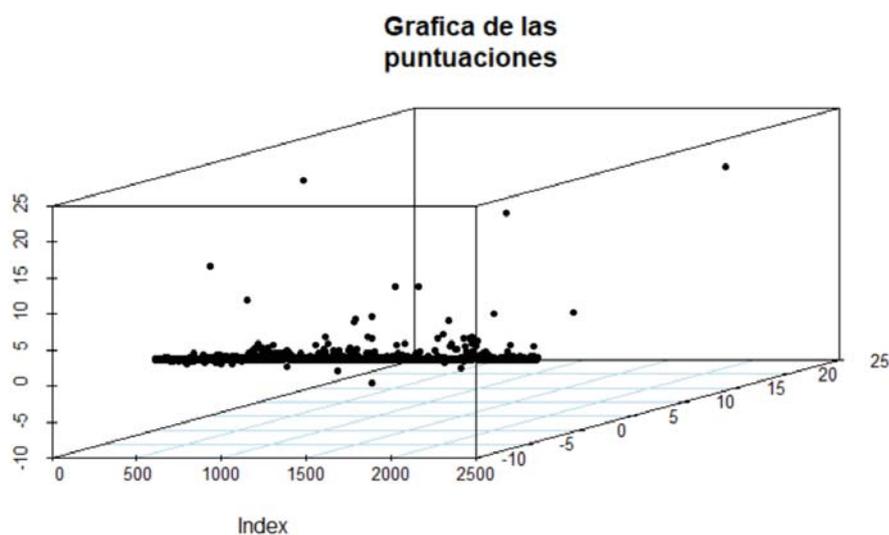


Figura 10: Gráfico de dispersión de los Scores de las variables de estadística de migraciones

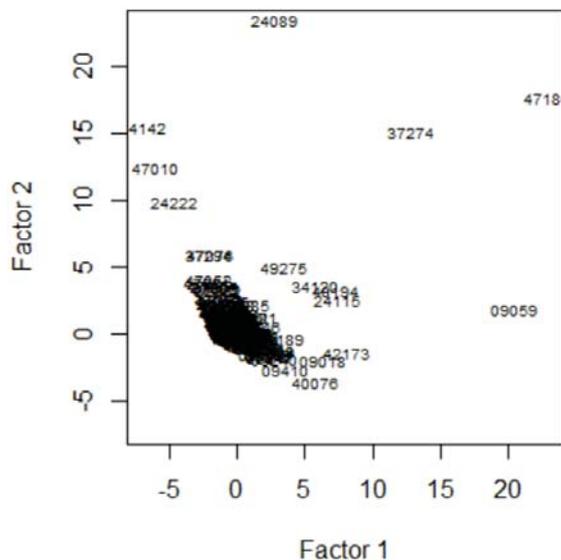


Figura 11: Gráfico de dispersión de los municipios según los factores para el grupo de variables de estadística de migraciones

Como se puede apreciar en la tabla de resultados (*Tabla 8*), un único factor explicaría un 92.5% de la variabilidad, y asigna un peso positivo elevado para todas las variables.

```

Loadings:
          Factor1
E_A_M_P  0.946
I_D_M_P  0.941
E_A_O_P  0.938
I_D_O_P  0.926
E_A_O_C  0.997
I_D_O_C  0.997
E_A_O_PAÍ 0.967
I_D_O_PAÍ 0.981

          Factor1
SS loadings    7.402
Proportion Var 0.925

Test of the hypothesis that 1 factor is
sufficient.
The chi square statistic is 11701.48 on
20 degrees of freedom.
The p-value is 0
    
```

Tabla 8: Análisis factorial de las variables de estadística de migraciones

Finalmente, se puede concluir en que lo ideal sería quedarse con un único factor.

#### 4. Variables del Grupo 1, Grupo 2 y Grupo 3

Finalmente se realiza un análisis factorial con el conjunto formado por todas las variables.

Se puede observar que todas las variables están altamente correlacionadas (*Figura 12*).

De nuevo se aprecia un problema ocasionado porque algunas de las variables son linealmente dependientes, lo que hace que el determinante de la matriz de correlaciones sea prácticamente 0, y por tanto el sistema sería computacionalmente singular. Es por esto, que no es posible realizar un análisis factorial sobre el conjunto completo de las variables.

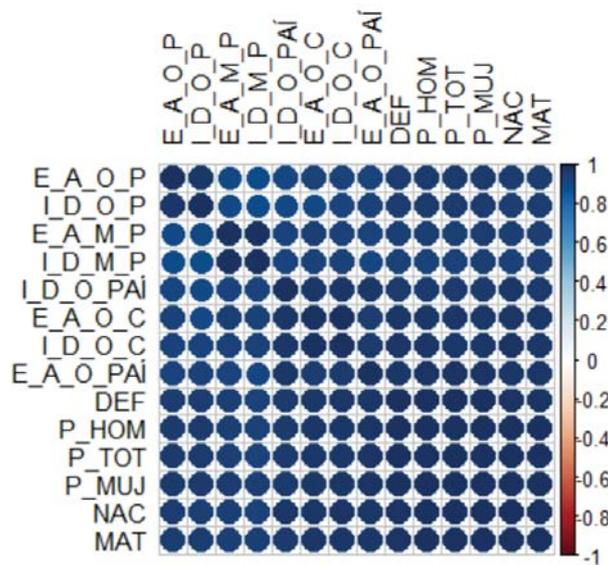


Figura 12: Matriz de correlaciones de las variables del Grupo 1, Grupo 2 y Grupo 3

#### Conclusión

Al realizar un análisis factorial por grupos de variables, se obtiene que para los grupos de estadística de movimiento natural de la población y de estadística de migraciones, lo idóneo es quedarse con un factor en ambos casos. No obstante, puesto que para el grupo de variables de estadística de población y de variables del Grupo 1, Grupo 2 y Grupo 3 (todas las variables) no es posible realizar análisis factorial, se tratará de realizar un análisis de componentes principales para realizar una reducción del número de variables.

#### 4.2.2. Análisis de Componentes Principales

El objetivo del análisis de componentes principales (ACP) es describir el conjunto de datos en términos de nuevas variables no correlacionadas, que permitan reducir la dimensión del número de variables. Para ello se realizará el ACP sobre la agrupación de variables definidas al principio de la sección 4. [11]

Para mejorar el estudio, se realizará el análisis también para el conjunto de municipios sin las capitales de provincia, que recogen las cantidades de población más elevadas.

##### 1.1. Variables de estadística de población

En primer lugar, se obtiene la cantidad de información que aporta cada variable (“Población Total” (P\_TOT), “Población Hombres” (P\_HOM) y “Población Mujeres” (P\_MUJ)) a cada componente. En este caso se observa (*Tabla 9*), que la primera componente representa una medida de tamaño, al asignar pesos positivos y distintos de 0 a todas las variables, por lo que la población de un municipio será mayor según crezca la componente.

También se puede observar que la segunda componente está representada principalmente por el número de hombres, con peso positivo, y por el número de mujeres, con peso negativo. Por tanto, se puede deducir que el número de hombres es directamente proporcional al valor de la componente, al contrario que el número de mujeres, que es inversamente proporcional, y será mayor cuanto más baja sea esta componente.

Rotation (n x k) = (3 x 3):			
	PC1	PC2	PC3
P_TOT	0.8159580	0.02965232	-0.5773503
P_MUJ	0.4336586	-0.69181417	0.5773503
P_HOM	0.3822993	0.72146649	0.5773503

Tabla 9: ACP de las variables de estadística de población

Las *Figuras 13 y 14* corresponden a un biplot de las dos primeras componentes. Se observa como las observaciones se extienden a lo largo de la primera componente, mientras que varían mínimamente con la segunda componente. Esto pone de manifiesto que la cantidad de variabilidad explicada por la primera componente es muy elevada.

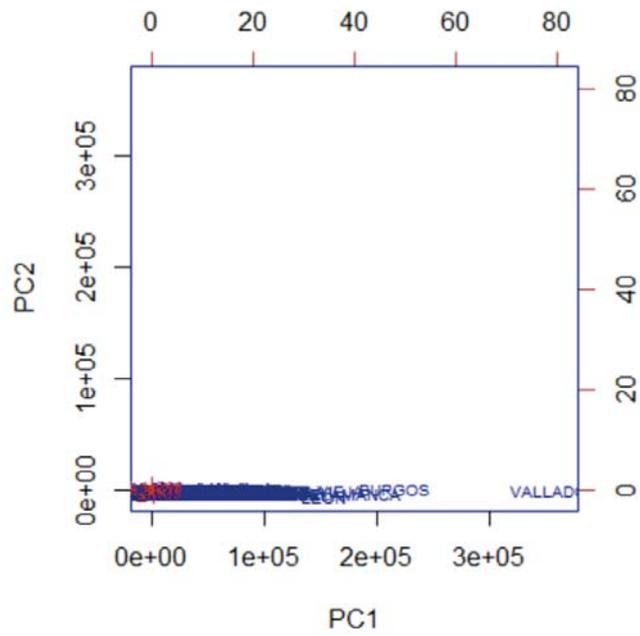


Figura 13: Biplot para dos componentes del ACP de las variables de estadística de población

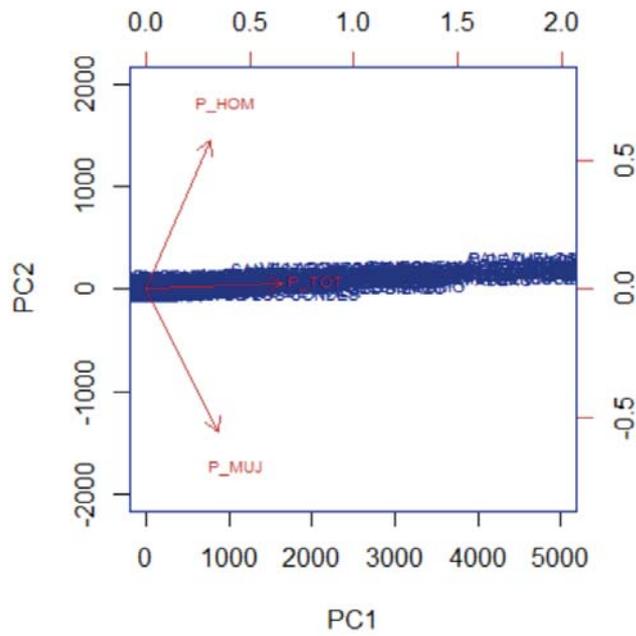


Figura 14: Biplot para dos componentes del ACP de las variables de estadística de población

La *Tabla 10* muestra la proporción de varianza explicada por cada una de las tres componentes. En este caso, es evidente que la primera componente es suficiente, ya que representa casi el 100 % de la información.

**[1] 0.99988 0.00012 0.00000**

Tabla 10: Variabilidad explicada por las componentes del ACP de las variables de estadística de población

### 1.2. Variables de estadística de población sin capitales de provincia

En la *Tabla 11* se puede apreciar que apenas varían los pesos, siendo la población total la mayor influencia para la primera componente. También se observa un peso nulo de esta variable para la segunda componente. En este caso, son las variables de población en función del sexo las variables que aportan toda la información a la componente. La población de hombres influye de forma positiva en la componente, mientras que la población de mujeres lo hace de forma negativa. Estos resultados pueden observarse también en la *Figura 15*.

Rotation (n x k) = (3 x 3):			
	PC1	PC2	PC3
P_TOT	0.8163843	0.01354319	0.5773503
P_MUJ	0.4199209	-0.70023791	-0.5773503
P_HOM	0.3964634	0.71378110	-0.5773503

Tabla 11: ACP de las variables de estadística de población (sin capitales de provincia)

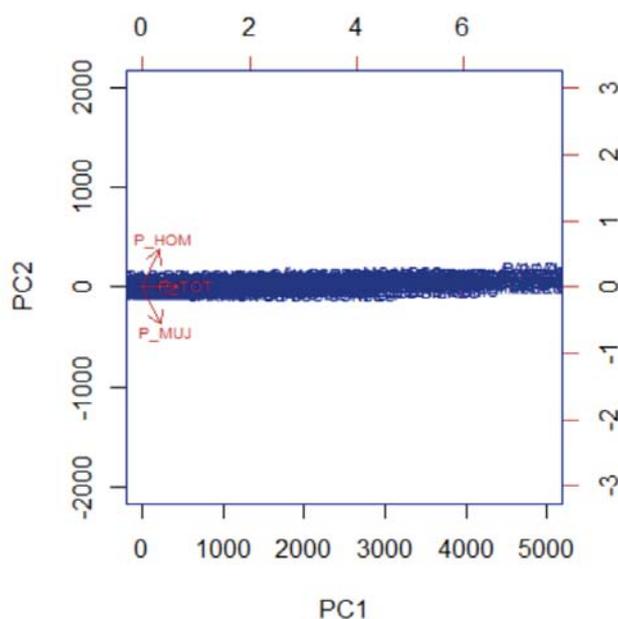


Figura 15: Biplot para dos componentes del ACP de las variables de estadística de población (sin capitales de provincia)

La *Tabla 12* muestra el porcentaje de población aportado por cada componente. En este caso se puede ver que el resultado es el mismo omitiendo las capitales de provincia.

$$[1] \quad 0.9998 \quad 0.0002 \quad 0.0000$$

Tabla 12: Variabilidad explicada por las componentes del ACP de las variables de estadística de población (sin capitales de provincia)

### 2.1. Variables de estadística de movimiento natural de población

Las variables utilizadas son “Nacimientos” (NAC), “Defunciones” (DEF) y “Matrimonios de distinto sexo” (MAT).

La primera componente, de nuevo representa una medida de tamaño, ya que todos los pesos son positivos (*Tabla 13*). Por este motivo, el crecimiento de la respuesta supone el crecimiento del valor de estas características. El número de defunciones supone una elevada aportación a esta componente.

Con respecto a la segunda componente, se encuentra una asociación negativa con el número de defunciones, y positiva en el caso del número de nacimientos y de matrimonios, este último en

menor medida. Se puede interpretar esto de forma que, si el valor de la segunda componente de un municipio es elevado, mayor es el número de nacimientos y de matrimonios, y menor el de defunciones. Por el contrario, si se encuentra ante un municipio con un pequeño valor de la componente, mayor será el número de defunciones y, por tanto, menor el de nacimientos y de matrimonios. Normalmente, un municipio con mayor cantidad de defunciones está relacionado con poblaciones con una media de edad elevada, por lo que parece evidente que cuanto mayor sea esta media de edad, menor será el número de nacimientos y de matrimonios, explicando así las relaciones entre estos tres indicadores.

Rotation (n x k) = (3 x 3):

	PC1	PC2	PC3
NAC	0.5206958	0.7442703	-0.418255355
DEF	0.8181440	-0.5749946	-0.004655619
MAT	0.2439596	0.3397689	0.908317556

Tabla 13: ACP de las variables de estadística de movimiento natural de población

Visualmente (*Figuras 16 y 17*) se puede notar la gran influencia de la primera componente a este estudio, confirmando de nuevo la gran dependencia de estas características con respecto al número de habitantes, siendo los grandes municipios, los que cuentan con más nacimientos, defunciones y matrimonios.

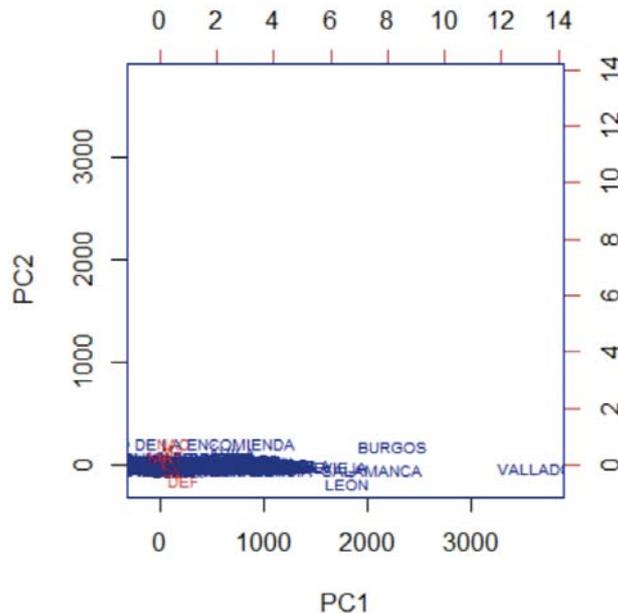


Figura 16: Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población

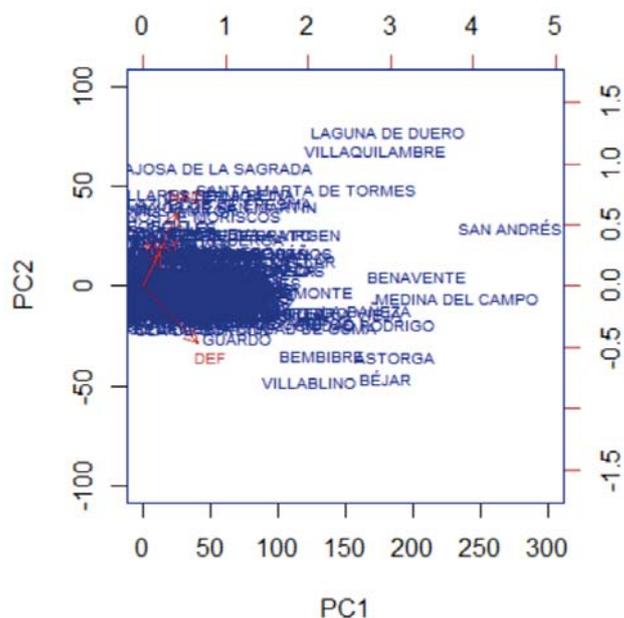


Figura 17: Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población

Gracias a la siguiente tabla (*Tabla 14*) se puede ver que la primera componente explica prácticamente toda la variabilidad, por lo que la se considerará como suficiente para continuar nuestro estudio.

$$[1] \quad 0.99305 \quad 0.00670 \quad 0.00025$$

Tabla 14: Variabilidad explicada por las componentes del ACP de las variables de estadística de movimiento natural de población

## 2.2. Variables de estadística de movimiento natural de la población sin capitales de provincia

Con ayuda de la *Tabla 15* y la *Figura 18*, se puede apreciar una primera componente de tamaño, al contar con pesos positivos y distintos de 0, y con una elevada influencia por parte de la variable defunciones y una influencia aceptable por parte de la variable nacimientos. La segunda componente cuenta con una aportación positiva por parte de nacimientos y matrimonios, y negativa con las defunciones.

Rotation (n x k) = (3 x 3):

	PC1	PC2	PC3
NAC	0.5405201	0.7125324	-0.44736521
DEF	0.8010509	-0.5984147	0.01474024
MAT	0.2572070	0.3663297	0.89422989

Tabla 15: ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia)

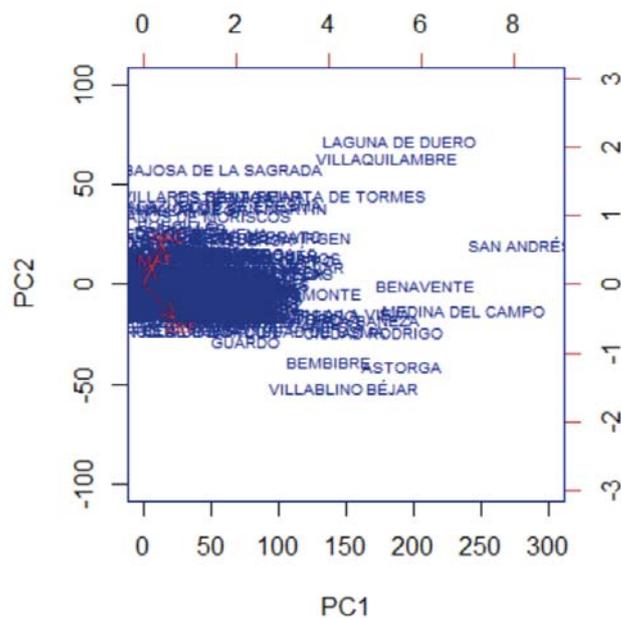


Figura 18: Biplot para dos componentes del ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia)

En la *Tabla 16* se puede apreciar una disminución en cuanto a proporción explicada por la primera componente. No obstante, esta componente sigue siendo la principal aportación de información.

[1] 0.94194 0.05536 0.00270

Tabla 16: Variabilidad explicada por las componentes del ACP de las variables de estadística de movimiento natural de población (sin capitales de provincia)

### 3.1. Variables de estadística de migraciones

Las variables utilizadas son “Emigraciones a municipios de la misma provincia” (E\_A\_M\_P), “Inmigraciones desde municipios de la misma provincia” (I\_D\_M\_P), “Emigraciones a municipios de otras provincias” (E\_A\_O\_P), “Inmigraciones desde municipios de otras provincias” (I\_D\_O\_P), “Emigraciones a municipios de otras comunidades” (E\_A\_O\_C), “Inmigraciones desde municipios de otras comunidades” (I\_D\_O\_C), “Emigraciones a municipios de otros países” (E\_A\_O\_PAÍ) e “Inmigraciones desde municipios de otros países” (I\_D\_O\_PAÍ).

La primera componente es claramente una medida de tamaño (*Tabla 17*), ya que todos los pesos son positivos y no nulos. No obstante, se encuentran algunas variables que destacan sobre las demás con una influencia aceptable, como son las correspondientes a emigraciones a municipios de la misma provincia y de otra comunidad, y a inmigraciones desde municipios de la misma provincia.

Para la segunda componente, se observa una asociación positiva con las variables “Emigraciones a municipios de la misma provincia” e “Inmigraciones desde municipios de la misma provincia”, y una asociación negativa y distinta de 0 en “Emigraciones a municipios de otras comunidades”, “Inmigraciones desde municipios de otras comunidades”, “Inmigraciones desde municipios de otros países” y “Emigraciones a municipios de otros países”, esta última en menor medida. Esto quiere decir, que municipios con la segunda componente elevada, tienen una respuesta mayor en emigraciones a municipios de la misma provincia e inmigraciones desde municipios de la misma provincia, mientras que, si tienen valores bajos para esta componente, tendrán una respuesta elevada en emigraciones a municipios de otras comunidades y países, e inmigraciones desde municipios de otras comunidades y países.

Rotation (n x k) = (8 x 8):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
E_A_M_P	0.4723155	0.49880343	0.009837493	-0.001630066	0.38755599	-0.51423449	0.33011013	0.06636490
I_D_M_P	0.4722069	0.54686554	-0.097179686	-0.056261965	-0.36390342	0.49277878	-0.28642952	-0.08972672
E_A_O_P	0.1290285	-0.07760221	0.460634801	0.199164262	-0.43476303	-0.37803591	-0.03261994	-0.62648592
I_D_O_P	0.1615511	-0.05965501	0.755498922	0.174915695	-0.05298357	0.20458684	0.10575147	0.55957304
E_A_O_C	0.4570280	-0.39947949	-0.330310658	0.349402546	-0.16565252	-0.31388935	-0.40096644	0.33706518
I_D_O_C	0.3855240	-0.35939524	-0.097032525	0.351961100	0.30638251	0.45420087	0.41302671	-0.34373632
E_A_O_PAÍ	0.1812139	-0.14734355	0.297814324	-0.311005029	0.57625359	0.04120415	-0.61576479	-0.21702913
I_D_O_PAÍ	0.3476345	-0.36344665	-0.016807089	-0.764125087	-0.26864344	-0.03797763	0.29360957	0.05539838

Tabla 17: ACP de las variables de estadística de migraciones



La variabilidad explicada por la primera componente alcanza un 95'7 (*Tabla 18*). Se realizará un análisis Cluster con dos componentes, para comprobar si la segunda componente carece o no de utilidad.

[1] 0.95753 0.02862 0.00609 0.00437 0.00171 0.00094 0.00052 0.00022

Tabla 18: Variabilidad explicada por las componentes del ACP de las variables de estadística de migraciones

### 3.2. Variables de estadística de migraciones sin capitales de provincia

La *Tabla 19* muestra los pesos de cada variable en cada una de las componentes. Se observa una primera componente de tamaño, con pesos positivos y no nulos. Esta componente recoge principalmente, información de las variables “Emigraciones a municipios de la misma provincia”, “Inmigraciones desde municipios de la misma provincia”, “Emigraciones a municipios de otra comunidad” e “Inmigraciones desde municipios de otra comunidad”. La segunda componente también recoge información de estas variables, no obstante, en este caso las variables “Emigraciones a municipios de otra comunidad” e “Inmigraciones desde municipios de otra comunidad” tienen una influencia negativa. Se pueden contemplar estos resultados de forma gráfica mediante la *Figura 22*.

Rotation (n x k) = (8 x 8):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
E_A_M_P	0.55809266	0.38122782	-0.080714914	0.709142811	-0.09882274	0.11897855	-0.04071637	-0.090663444
I_D_M_P	0.56744489	0.46480604	-0.002979974	-0.657603221	0.11944686	-0.03568544	0.04467397	0.109425803
E_A_O_P	0.09623236	-0.04794099	0.496166027	0.125017785	-0.29115156	-0.46501833	-0.03839026	0.651264715
I_D_O_P	0.10104573	-0.02957069	0.476231885	-0.121833264	-0.40400331	-0.24494602	-0.09505921	-0.717650128
E_A_O_C	0.42812169	-0.55096255	-0.153586036	0.076518653	0.41803630	-0.47487037	0.25338159	-0.138743878
I_D_O_C	0.34422895	-0.49738144	-0.331658921	-0.167841848	-0.58415734	0.29595580	-0.22140624	0.134533451
E_A_O_PAÍ	0.09363181	-0.11825342	0.184343417	0.009222748	0.41033686	0.05428260	-0.87853665	0.005662572
I_D_O_PAÍ	0.19110026	-0.26554856	0.594079020	0.009520392	0.20882558	0.62633198	0.31746928	0.057074698

Tabla 19: ACP de las variables de estadística de migraciones (sin capitales de provincia)

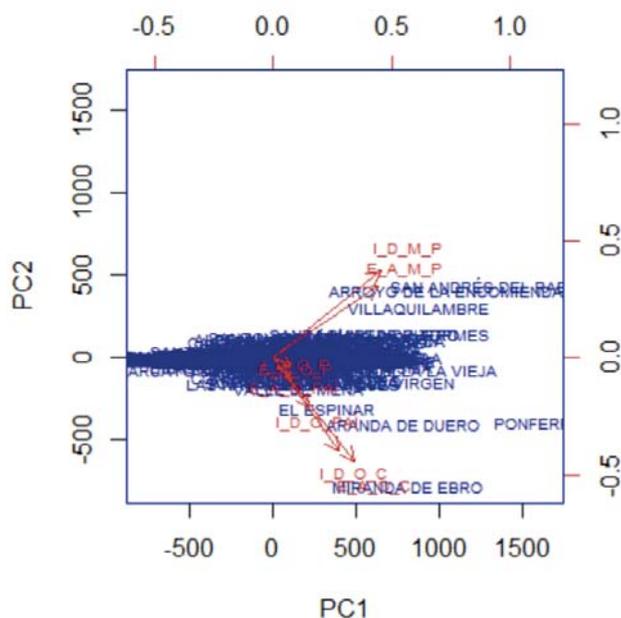


Figura 21: Biplot para dos componentes del ACP de las variables de estadística de migraciones (sin capitales de provincia)

Si se observa la *Tabla 20*, se puede apreciar una importante disminución de la proporción de información que recoge la primera componente, en este caso un 82'8%. A pesar de ser un resultado elevado, la segunda componente gana algo de fuerza, por lo que se considera este cambio una mejora para el posterior análisis cluster.

[1] 0.82810 0.13914 0.01402 0.00790 0.00402 0.00358 0.00230 0.00094

Tabla 20: Variabilidad explicada por las ocho componentes del ACP de las variables de estadística de migraciones (sin capitales de provincia)

#### 4.1. Variables del Grupo 1, Grupo 2 y Grupo 3

La primera componente principal, en el caso de utilizar todas las variables, se puede considerar una medida de tamaño (*Tabla 21*), al asignar peso positivo a todas las variables. No obstante, únicamente cuenta con valores no nulos para los indicadores de población. Esto quiere decir que, un valor elevado de la primera componente indica, mayor población del municipio.

La segunda componente, principalmente se ve asociada con la población en función del sexo, con una gran influencia positiva por parte del número de hombres, y negativa por parte del número de mujeres. Se encuentran también otras dos variables que se consideran no nulas para esta segunda

componente, referentes a emigraciones a municipios de la misma provincia, y a inmigraciones desde municipios de la misma provincia, con peso asignado negativo. Por tanto, cuanto mayor resulte el valor de la segunda componente, más elevado será el número de hombres, así como cuanto más bajo sea, mayor será el número de mujeres, y en menor medida, el número de inmigraciones y emigraciones de/a municipios de la misma provincia. Con esto se podría deducir que, el número de mujeres que tiende a realizar migraciones es ligeramente superior al de hombres.

Rotation (n x k) = (14 x 14):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
P_TOT	0.815722376	0.032761212	0.005792428	0.00845501	0.001051996	-0.003369537	-0.0004932163
P_MUJ	0.433533793	-0.669920305	0.150191576	0.05087457	-0.023919252	-0.011108022	0.0116767546
P_HOM	0.382188583	0.702681517	-0.144399148	-0.04241956	0.024971248	0.007738485	-0.0121699709
NAC	0.005548973	0.007092554	-0.027216840	-0.08353333	0.160113030	0.086428751	-0.0619960300
DEF	0.008651230	-0.061207904	0.037572550	-0.01951066	-0.170325514	-0.234235703	-0.3006998966
MAT	0.002601909	0.005898713	-0.019312418	-0.01163787	0.051222231	0.020322308	-0.0328573356
E_A_M_P	0.010030972	-0.134794632	-0.647718741	0.12473440	0.014847321	-0.217647471	0.1445248149
I_D_M_P	0.009945403	-0.155374445	-0.714661224	0.06373001	0.075320247	0.238916738	-0.1200451719
E_A_O_P	0.002912314	0.000800562	0.028554912	0.05717453	-0.195023017	0.514696977	-0.3830842162
I_D_O_P	0.003663361	-0.002805217	0.040826542	0.16577460	-0.161196892	0.700046581	0.3189914122
E_A_O_C	0.009931502	-0.054039613	-0.119365872	-0.66951338	-0.349896383	-0.035931860	-0.2603474727
I_D_O_C	0.008420132	-0.033333300	-0.067320736	-0.51731650	-0.324184654	0.077064526	0.5519222244
E_A_O_PAÍ	0.004050996	-0.016428130	0.044615019	-0.04274066	0.326415383	-0.086546047	0.4700255571
I_D_O_PAÍ	0.007586916	-0.076942016	0.022342484	-0.46902594	0.733178253	0.250810692	-0.1650499503

	PC8	PC9	PC10	PC11	PC12	PC13	PC14
P_TOT	-0.004889636	0.001984528	-8.617169e-05	-0.004501127	0.0049485504	-0.0001721231	-5.773503e-01
P_MUJ	-0.056584480	0.026633544	8.552613e-04	0.004055676	0.0039598648	-0.0005435815	5.773503e-01
P_HOM	0.051694844	-0.024649016	-9.414330e-04	-0.008556804	0.0009886857	0.0003714584	5.773503e-01
NAC	-0.299516302	0.004739437	-1.779480e-01	0.483413439	-0.6132134694	0.4734273408	-2.955705e-16
DEF	0.662727721	-0.535525000	-1.642444e-01	0.078705927	-0.2455563233	0.0250000441	-1.099478e-14
MAT	-0.150812985	-0.016429401	-4.844568e-02	0.260315013	-0.3628149419	-0.8778232481	-1.970069e-14
E_A_M_P	0.384919623	0.548033759	-9.023789e-02	0.074112261	-0.1264473529	0.0049708027	1.987343e-16
I_D_M_P	-0.269078926	-0.529101818	8.061093e-02	-0.057480768	0.1470836769	0.0021282257	-1.162692e-15
E_A_O_P	0.206002234	0.204145351	4.470532e-02	0.565790214	0.3738375701	-0.0149060281	1.671332e-15
I_D_O_P	0.261300410	-0.007496973	1.087369e-01	-0.326800976	-0.4073253453	0.0124445493	7.883707e-15
E_A_O_C	-0.046329713	0.151762547	5.098955e-01	-0.139093949	-0.1954601081	0.0161818358	3.491646e-15
I_D_O_C	0.008184749	-0.137079829	-4.602496e-01	0.195220487	0.2033770008	-0.0292572196	-1.217283e-15
E_A_O_PAÍ	0.228515080	-0.204776859	6.292457e-01	0.406837982	0.0699601902	0.0205133116	-8.523393e-15
I_D_O_PAÍ	0.238170828	0.076303033	-2.068169e-01	-0.183176342	0.0651399367	-0.0521776330	-2.282196e-15

Tabla 21: ACP de las variables del Grupo 1, Grupo 2 y Grupo 3

Observando en las Figuras 22 y 23, se puede confirmar que la primera componente explica un porcentaje de variabilidad muy elevado, siendo la población total la variable predominante, lo que indica, como se ha comentado anteriormente, que los municipios que se encuentran situados a la derecha del gráfico son los que cuentan con una población mayor. En cuanto a la segunda componente, parece evidente su altísima influencia de la cantidad de población en función del sexo.



Analizando los resultados de esta tabla (*Tabla 22*) se puede confirmar que la primera componente explica un gran porcentaje de variabilidad y, por tanto, representa casi toda la información. No obstante, se realiza un análisis Cluster con 2/3 componentes, para valorar si la influencia de las variables del grupo de migraciones resulta relevante.

```
[1] 0.99984 0.00012 0.00002 0.00001 0.00000 0.00000 0.00000  
[8] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
```

Tabla 22: Variabilidad explicada por cada una de las catorce componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3

#### 4.2. Variables del Grupo 1, Grupo 2 y Grupo 3 sin capitales de provincia

En la *Tabla 23*, se muestra una primera componente que se obtiene en su totalidad a partir de las variables de estadística de población, siendo los pesos del resto de variables correspondientes a variables de estadística de movimiento natural de población y a variables de estadística de migraciones prácticamente nulos. Se observa una segunda componente que recoge principalmente información de las variables de estadística de población en función del sexo, siendo el peso de población de hombres positivo, mientras que el de mujeres es negativo. Esta segunda componente también se apoya, aunque en menor medida, en las variables que refieren a las emigraciones a municipios de la misma provincia e inmigraciones desde municipios de la misma provincia. Estos resultados pueden apreciarse de forma gráfica en la *Figura 24*.

Rotation (n x k) = (14 x 14):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
P_TOT	0.816087935	0.004858766	-0.01337304	0.01734698	0.004450267	-0.006907943	-0.007874778
P_MUJ	0.419767735	-0.596375619	0.35008186	-0.08710235	-0.048760458	0.012937519	-0.014866666
P_HOM	0.396320201	0.601234385	-0.36345491	0.10444933	0.053210724	-0.019845462	0.006991888
NAC	0.005707806	0.035139157	0.01501237	-0.02458352	-0.221832158	-0.099259015	-0.126418065
DEF	0.007940384	-0.094286247	-0.08121012	0.04102482	0.312558294	0.128036127	0.866401476
MAT	0.002723797	0.020995539	0.02413509	-0.02001804	-0.080033107	-0.058457872	-0.034295667
E_A_M_P	0.013002175	0.329577793	0.52509275	-0.14604011	0.320099757	0.665526805	-0.161730285
I_D_M_P	0.012811819	0.395213740	0.58134495	-0.26868995	-0.314098742	-0.452892034	0.311725358
E_A_O_P	0.002698385	0.013014194	-0.01004028	0.13574176	-0.313798627	0.351280910	0.280227390
I_D_O_P	0.002689855	0.048011325	-0.01298210	0.09929705	-0.429070078	0.159856689	0.130622921
E_A_O_C	0.012052077	-0.060627444	-0.20550238	-0.61755217	0.002347407	0.032277425	0.083367808
I_D_O_C	0.009555886	0.002560478	-0.25211707	-0.68557976	0.026608244	0.071497859	-0.033260738
E_A_O_PAÍ	0.002691153	-0.015692553	-0.04655726	-0.05428888	-0.179448942	0.098905379	0.069495746
I_D_O_PAÍ	0.005590652	-0.032340092	-0.12588056	-0.04867436	-0.569961160	0.395982982	-0.007321064

	PC8	PC9	PC10	PC11	PC12	PC13	PC14
P_TOT	0.0004420781	0.000163827	-7.385554e-05	-0.0056553908	0.001308612	-0.0005694047	5.773503e-01
P_MUJ	0.0223196190	-0.024679224	-2.581079e-02	-0.0048538957	0.009110482	0.0027822425	-5.773503e-01
P_HOM	-0.0218775409	0.024843051	2.573693e-02	-0.0008014951	-0.007801869	-0.0033516472	-5.773503e-01
NAC	0.0162528532	-0.085634889	-1.233659e-01	0.7269908866	-0.401348123	0.4587448733	1.993831e-17
DEF	-0.1883516586	-0.139367521	-2.661940e-02	0.2485117045	0.014103334	0.0215931181	-9.574420e-16
MAT	0.0774887414	-0.013889591	-6.037094e-02	0.4012812865	-0.185847215	-0.8843991159	9.516614e-15
E_A_M_P	-0.0101640769	0.027030196	-1.311230e-02	0.1413410662	0.037630125	0.0146168130	-5.690486e-16
I_D_M_P	-0.0814746019	-0.034794104	4.074276e-02	-0.1507644501	-0.017123491	0.0034580391	2.426597e-16
E_A_O_P	0.5192003286	0.090517751	-1.661222e-02	-0.3112412058	-0.553046244	0.0114814633	-8.945761e-16
I_D_O_P	0.4311161035	-0.044768600	-1.400492e-01	0.2433410041	0.702270472	0.0326164582	-1.628978e-15
E_A_O_C	0.1632770502	0.496001741	5.151561e-01	0.1528436704	0.045849544	0.0324549183	1.143867e-15
I_D_O_C	0.1007676243	-0.460441853	-4.616915e-01	-0.1513713117	-0.032066856	-0.0114835710	-1.802604e-15
E_A_O_PAÍ	-0.3478978601	0.663834481	-6.201131e-01	-0.0372113345	0.008410651	-0.0106280644	1.143150e-15
I_D_O_PAÍ	-0.5807409744	-0.246950550	3.070704e-01	-0.0468084677	-0.009033453	-0.0647021423	-3.160111e-16

Tabla 23: ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia)

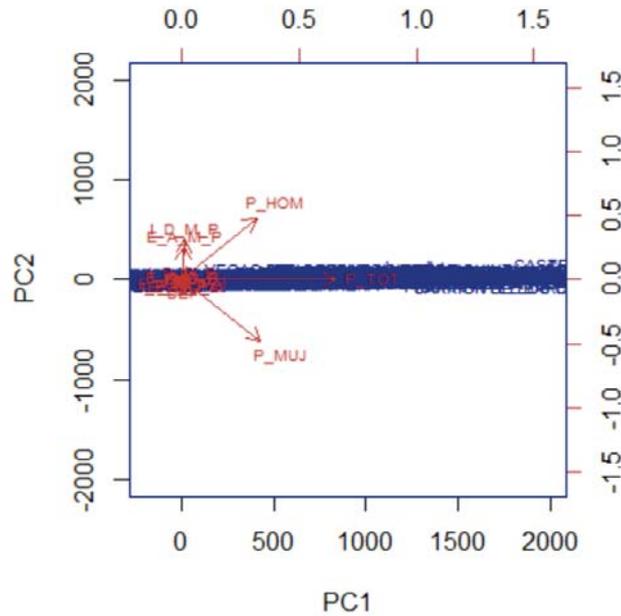


Figura 24: Biplot para dos componentes del ACP de las variables del Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia)

En la *Tabla 24* se muestran los pesos de las componentes obtenidas mediante el análisis de componentes principales. Se puede observar como la primera componente aporta prácticamente toda la información, aunque para continuar con el análisis cluster, se emplearán 3 componentes, para que las variables referentes a estadística de migraciones tengan algo de peso en la información representada.

[1]	0.99958	0.00024	0.00013	0.00003	0.00001	0.00001	0.00000
[8]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Tabla 24: Variabilidad explicada por las componentes del ACP de las variables de estadística de Grupo 1, Grupo 2 y Grupo 3 (sin capitales de provincia)

Se puede observar, tras realizar este análisis, que la diferencia entre añadir las capitales de provincia al estudio no es muy grande, principalmente en el análisis de componentes principales con indicadores de estadística de población. No obstante, esta diferencia, sí que será importante al realizar el análisis cluster, ya que, al eliminar observaciones con diferencias significativas con respecto al resto, permite que la diferenciación por grupos sea más estable, y por tanto se obtenga un análisis más claro e interpretable. Por tanto, se continuará este estudio omitiendo las capitales de provincia.

A continuación, se realizará el análisis Cluster a partir de las componentes recogidas en cada uno de los grupos.

### 4.2.3. Análisis Cluster

#### Jerárquico

En primer lugar, se realizará el Análisis Cluster Jerárquico, que busca construir grupos de municipios similares a partir de los datos recogidos en las variables de estudio.

El principio del algoritmo utilizado para realizar el Análisis Cluster Jerárquico consiste en crear, en cada etapa, una partición obtenida mediante agregación dos a dos de los municipios más próximos. Existen diferentes maneras de considerar la nueva pareja de elementos agregados. Dentro de las medidas de agregación de individuos que hemos considerado se encuentran: [12]

- Promedio: Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. El valor promedio de todas ellas se selecciona como la distancia entre los dos clusters.
- Simple: Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. La menor de todas ellas se selecciona como la distancia entre los dos clusters.
- Completo: Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. La mayor de todas ellas se selecciona como la distancia entre los dos clusters.
- Ward: Hacer variar lo menos posible la inercia intra-clases en cada etapa de agregación es buscar el mínimo de pérdida de inercia inter-clases resultante de la agregación de dos elementos. Así en cada etapa la inercia intra-clases aumenta en la cantidad  $\Delta$ . [12]

#### No Jerárquico

Posteriormente se va a realizar el Análisis Cluster no Jerárquico. Este busca agrupar individuos en una clasificación de K clusters, donde K se especifica a priori o bien se determina como una parte del proceso. La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición mejor. [13]

En este caso, se utilizará el método de las K – medias, que agrupa objetos en K grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática. [14]

Se realizará este análisis sobre los grupos definidos al principio de la sección 4, utilizando las componentes principales que hayamos considerado necesarias en el análisis de componentes principales.

## 1. Variables de estadística de población

### Jerárquico

En primer lugar, se va a realizar un análisis cluster sobre los tres indicadores “Población Total”, “Población Hombres” y “Población Mujeres”. Para empezar, se determinará la medida de agregación que proporcionará una solución óptima.

En la *Tabla 25* se pueden observar los coeficientes obtenidos por cada medida de agregación. Se considera el óptimo el más elevado, por lo que en este caso el método de Ward realizaría la mejor agrupación de los datos.

Promedio	Simple	Completo	Ward
0.9993815	0.9992367	0.9995782	0.9997022

Tabla 25: Coeficientes de los métodos de agregación para las variables de estadística de población

El número de clusters se determinará teniendo en cuenta la gran cantidad de observaciones. En este caso, se han seleccionado 13 clusters, ya se considera que la distribución de los municipios en grupos es bastante adecuada.

Se observa en la *Figura 25* la agrupación de los municipios (no capitales de provincia) con tamaño de población más elevado, obtenida del análisis cluster. Ponferrada representa un grupo debido a que, cuenta con una población que dobla la del segundo municipio más grande. El siguiente grupo destacado es el correspondiente a los municipios de Miranda de Ebro, Aranda de Duero y San Andrés del Rabanedo, muy distanciados en el gráfico del cluster de Ponferrada, con unas poblaciones significativas de aproximadamente 30000 habitantes. Los tres grupos restantes, cuentan con municipios con una población entre 8000 y 22000 habitantes, con diferencias significativas entre los municipios pertenecientes a distintos grupos. Para un valor más pequeño de la primera componente, menor es la población del municipio.

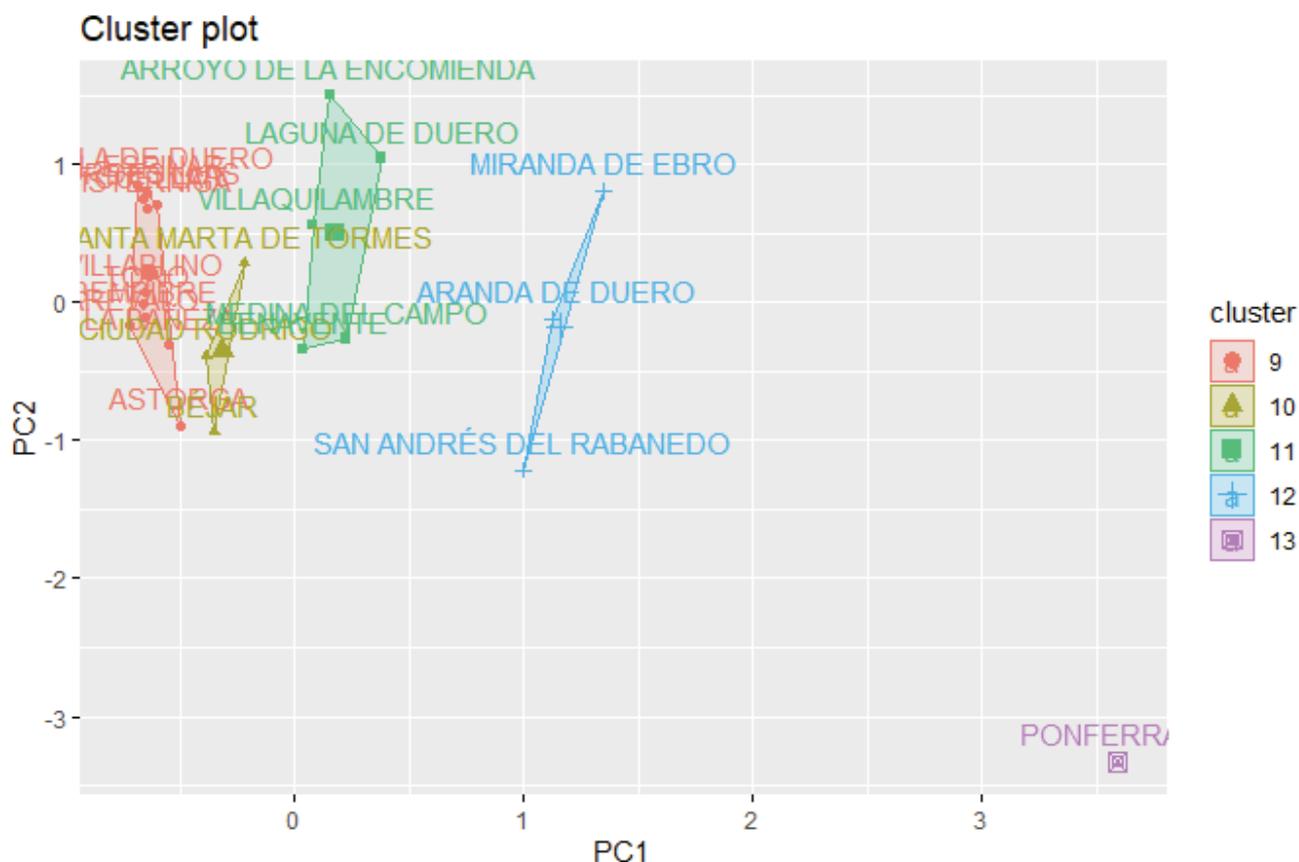


Figura 25: Grupos principales para las variables de estadística de población

Para mostrar los datos con una mayor claridad se utiliza la *Figura 26*. Este representa de una forma más visual, los municipios de Castilla y León en función de la agrupación resultante del Análisis Cluster. Se puede observar algún municipio perteneciente al grupo 0, el cual realmente no es un cluster, sino que contiene las capitales de provincia, de las que se ha prescindido para el estudio. Se aprecia una enorme cantidad de municipios pertenecientes al grupo 1, que en este caso se corresponden con municipios con una población muy pequeña, entre 0 y 250 habitantes aproximadamente. Los grupos representados están ordenados, de forma que el grupo 1 cuenta con los municipios de menor población, y el 13 con los municipios de mayor población, de forma que el mapa representa una mayor población con colores más fuertes. Esto es indicador de que Castilla y León tiene un alto porcentaje de municipios pequeños, por lo que predomina el medio rural.

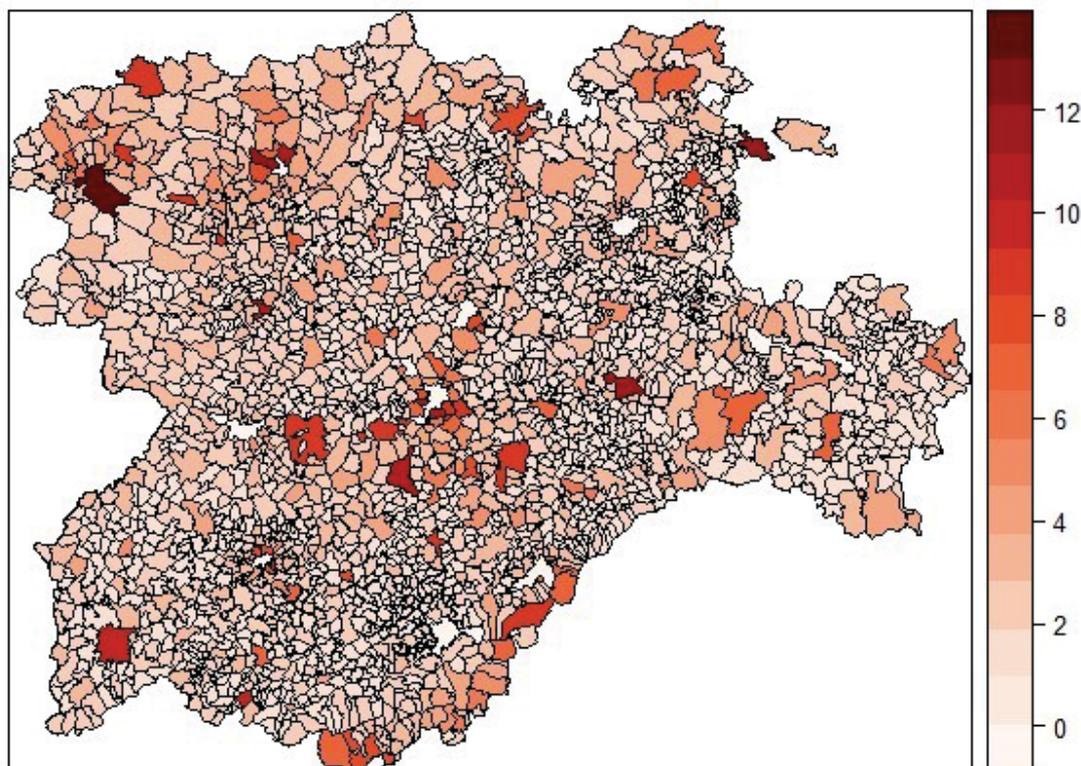


Figura 26: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de población

### No Jerárquico

Se ha realizado un Análisis Cluster no Jerárquico a partir del mismo número de grupos que los seleccionados en el Jerárquico. Para el caso de las variables de estadística de población, por tanto, 13 grupos.

Se observa en la *Figura 27* una gran diferencia con respecto a los grupos obtenidos mediante el análisis jerárquico. En este último, los municipios con una población más elevada (mostrados en la *Figura 25*), se clasifican en 5 grupos diferentes, mientras que, en el análisis no jerárquico, esta clasificación se reduce a 3 grupos. Esto implica, que existe una mayor diferenciación entre los municipios con una población más reducida. El cluster 12, que recoge municipios como Miranda de Ebro, Aranda de Duero y San Andrés del Rabanedo, de más de 30000 habitantes, los cuales antes representaban un único grupo, también cuenta ahora con otros municipios Arroyo de la Encomienda o Villaquilambre, que no superan los 20000 habitantes. Es evidente que hay diferencias importantes entre estos municipios, por lo que supone un problema, al no encontrarlas reflejadas en los resultados. A su vez, el cluster 11 recoge municipios con diferencias significativas, como Santa Marta de Tormes, con 15000 habitantes, y Bembibre, con 9000. Ponferrada representa un

único grupo.

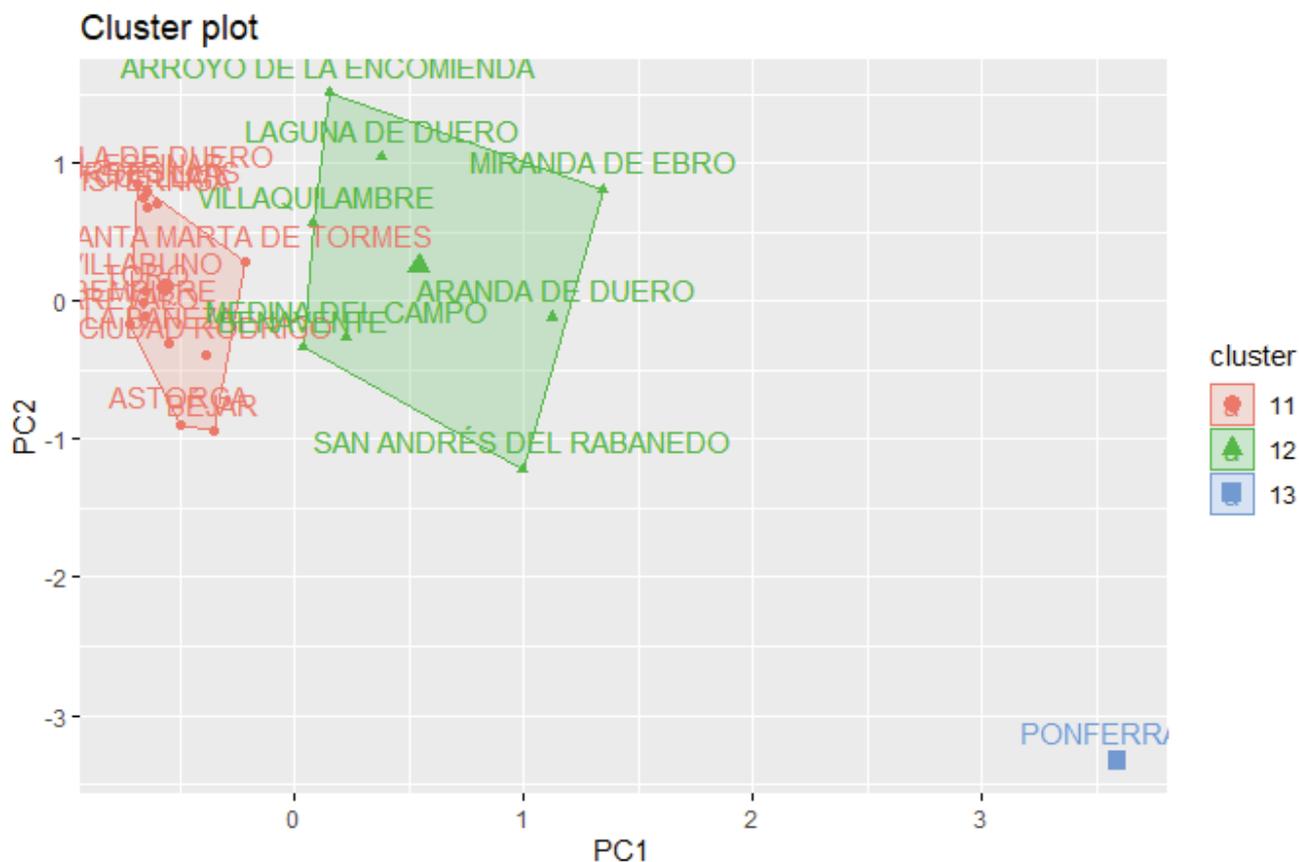


Figura 27: Grupos principales para las variables de estadística de población (no jerárquico)

La *Figura 28* muestra a qué grupo se asigna cada municipio de Castilla y León. Se aprecian grupos muy numerosos, que recogen en su mayoría los municipios con menor cantidad de población. No obstante, existe una mayor diversidad entre grupos, estando los municipios más repartidos entre los distintos grupos.

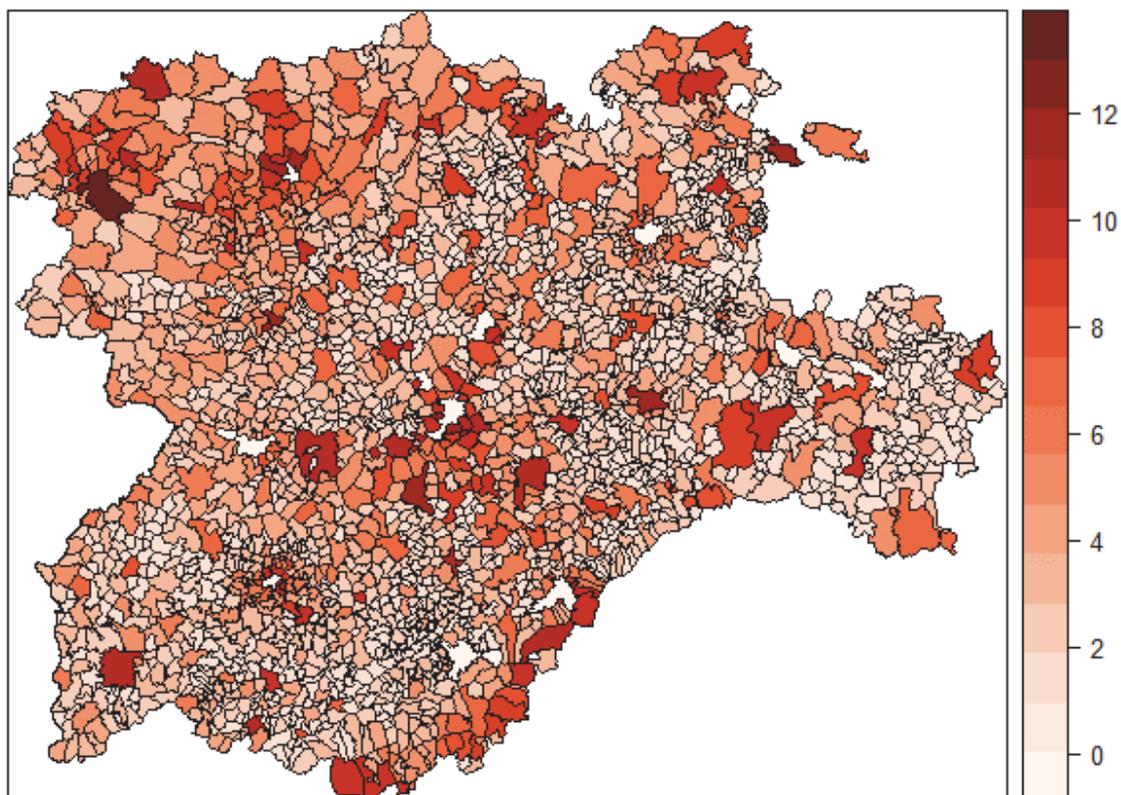


Figura 28: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de población (no jerárquico)

### Jerárquico vs No Jerárquico

- El Análisis Cluster Jerárquico prioriza la diferenciación de municipios con poblaciones elevadas. De esta forma, se obtienen más grupos con pocos municipios (los más destacados en cuanto a población), y por consecuencia, grupos muy numerosos que recogen municipios pequeños.
- El Análisis Cluster No Jerárquico, por el contrario, hace mayor diferenciación entre municipios de menor población, lo que da lugar a grupos menos heterogéneos, es decir, que recogen una cantidad de municipios más equilibrada.

A partir de los resultados, se puede deducir que el Análisis Cluster Jerárquico da lugar a mejores resultados, ya que clasifica mejor los municipios más poblados, que cuentan con diferencias significativas y necesarias de plasmar en los resultados.

2. Variables de estadística de movimiento natural de la población

**Jerárquico**

En la *Tabla 26* se puede observar que la medida de agrupación con la que se obtienen mejores resultados en la de Ward, por lo que será la empleada para el análisis Cluster.

Promedio	Simple	Completo	ward
0.9987976	0.9982278	0.9987700	0.9993449

Tabla 26: Coeficientes de los métodos de agregación para las variables de estadística de movimiento natural de la población

Los datos que refieren a movimiento natural de la población cuentan con menor heterogeneidad que los relacionados con la población, por lo que se selecciona un número menor de clusters, 10 en este caso. Se puede observar de nuevo, con ayuda de la *Figura 29*, un cluster que cuenta con un único municipio, Ponferrada. Parece evidente que, al contar con una gran diferencia de población con respecto al resto de municipios, también contará con mayor número de nacimientos, defunciones y matrimonios. El cluster 9, que cuenta con municipios de destacada población, está formado por 3 municipios, Miranda de Ebro, Aranda de Duero y San Andrés del Rabanedo. La primera componente muestra, un valor mayor para aquellos municipios que cuentan con más defunciones y más nacimientos, este último en menor medida. Por tanto, los municipios situados más a la derecha tienen valores mayores para estas características. Por el contrario, para un valor más alto de la segunda componente, disminuye el número de defunciones, aumentando únicamente nacimientos y matrimonios. En este aspecto destaca el municipio de Arroyo de la Encomienda, que cuenta con una gran diferencia entre nacimientos y matrimonios, y defunciones, siendo esta última muy baja. El grupo en el que se encuentra este municipio, también formado por Laguna de Duero y Villaquilambre, que también cuentan con valor de la segunda componente significativo, representa un grupo con alta natalidad y una cantidad pobre de defunciones, lo que es un claro indicador de que estos municipios cuentan con población joven. Lo contrario ocurre con el grupo 7, formado por municipios con una población semejante a los municipios del grupo 8, como Benavente, Medina del Campo, etc. Estos municipios cuentan con un valor negativo para la segunda componente, lo que asegura que el número de defunciones es más significativo que el número de nacimientos y matrimonios.

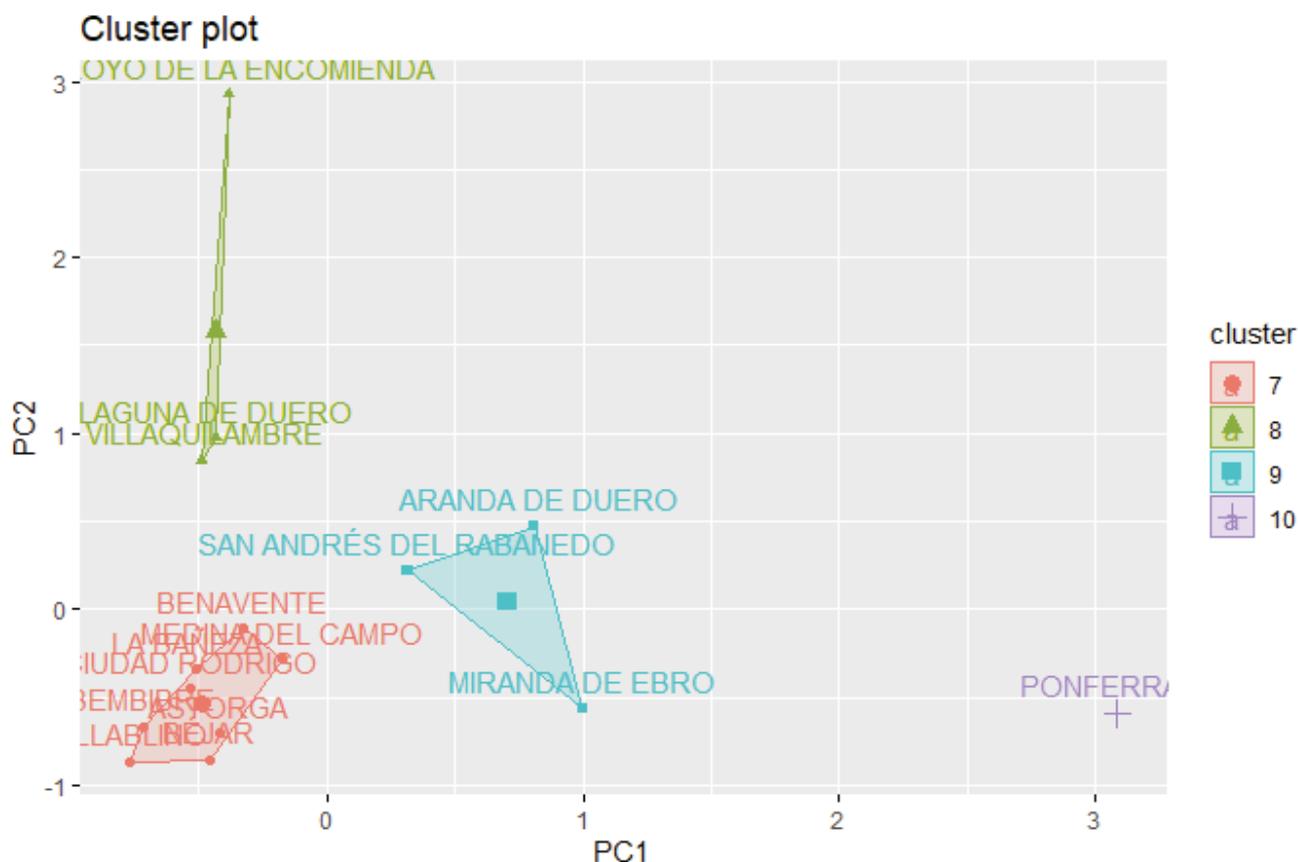


Figura 29: Grupos principales para las variables de estadística de movimiento natural de población

La *Figura 30* muestra la asignación de clusters de cada municipio de Castilla y León. Los grupos representados con colores más claros hacen referencia a municipios con menor número de nacimientos, defunciones y matrimonios. El cluster 1, representado en el mapa con el color más claro, contiene un número elevado de municipios que apenas cuentan con uno o dos nacimientos, defunciones o matrimonios como máximo. Se observa que este grupo es bastante numeroso, lo que es indicador de la gran cantidad de municipios pequeños existentes en Castilla y León. En general, una importante cantidad de los municipios cuentan con un número de defunciones superior al de nacimientos y matrimonios, con una diferencia importante con respecto a este último. Existe algún grupo, que contiene municipios que no siguen la estadística anterior. El segundo y tercer cluster, cuenta con una media de características generales semejante, no obstante, existe una gran diferencia en cuanto a la distribución de los datos. En el segundo cluster se observa un número de defunciones bastante elevado, mientras que en el tercer cluster, el número de nacimientos es significativamente superior al número de defunciones. Esto mismo ocurre con los clusters 5 y 8, que cuentan con medias similares a los grupos 6 y 7, respectivamente. En este caso, el número de nacimientos de los dos primeros son claramente superiores al número de defunciones. En el grupo 8, además, se observa un número de matrimonios superior en muchos casos al número de defun-

ciones. Estos resultados, por tanto, muestran, una diferenciación de grupos que no solo se basa en la población, sino que las tasas de natalidad y mortalidad son factores importantes. Se puede deducir que, un municipio con un número de nacimientos y de matrimonios superior es indicador de población joven, mientras que un mayor número de defunciones es indicador de población más anciana.

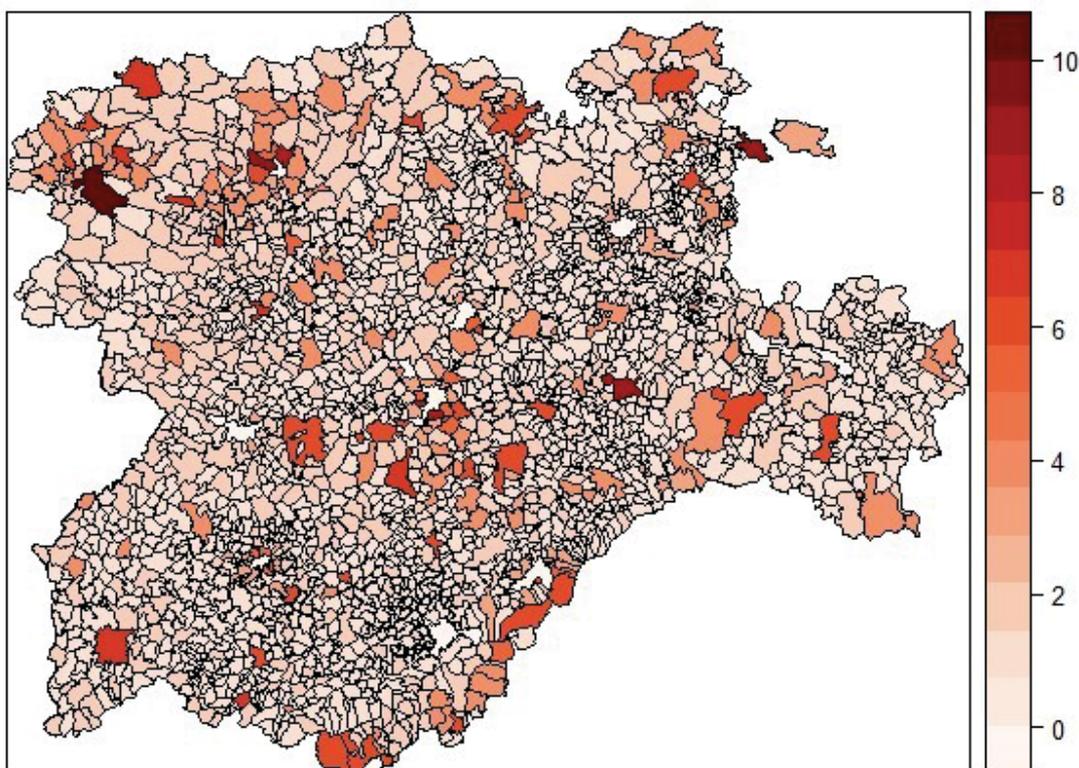


Figura 30: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de movimiento natural de población

### No Jerárquico

La *Figura 31* muestra los grupos con los municipios más significativos para las variables de estadística de movimiento natural de la población. Al igual que en el Análisis Cluster Jerárquico, se han clasificado los municipios en 10 grupos. Se observa que Ponferrada representa un grupo, al contar con diferencias muy significativas con respecto al resto de municipios. Tiene un valor elevado para la primera componente, lo que supone un mayor número de nacimientos y defunciones. No obstante, para la segunda componente cuenta con un valor negativo, que hace referencia a un menor número de nacimientos y matrimonios. Por tanto, existe una mayor cantidad de defunciones que de nacimientos y matrimonios. El siguiente grupo más significativo engloba los municipios de Aranda de Duero, Miranda de Ebro y San Andrés del Rabanedo. El grupo que recoge los municipios de Arroyo de la Encomienda, Laguna de Duero y Villaquilambre, cuenta con un valor



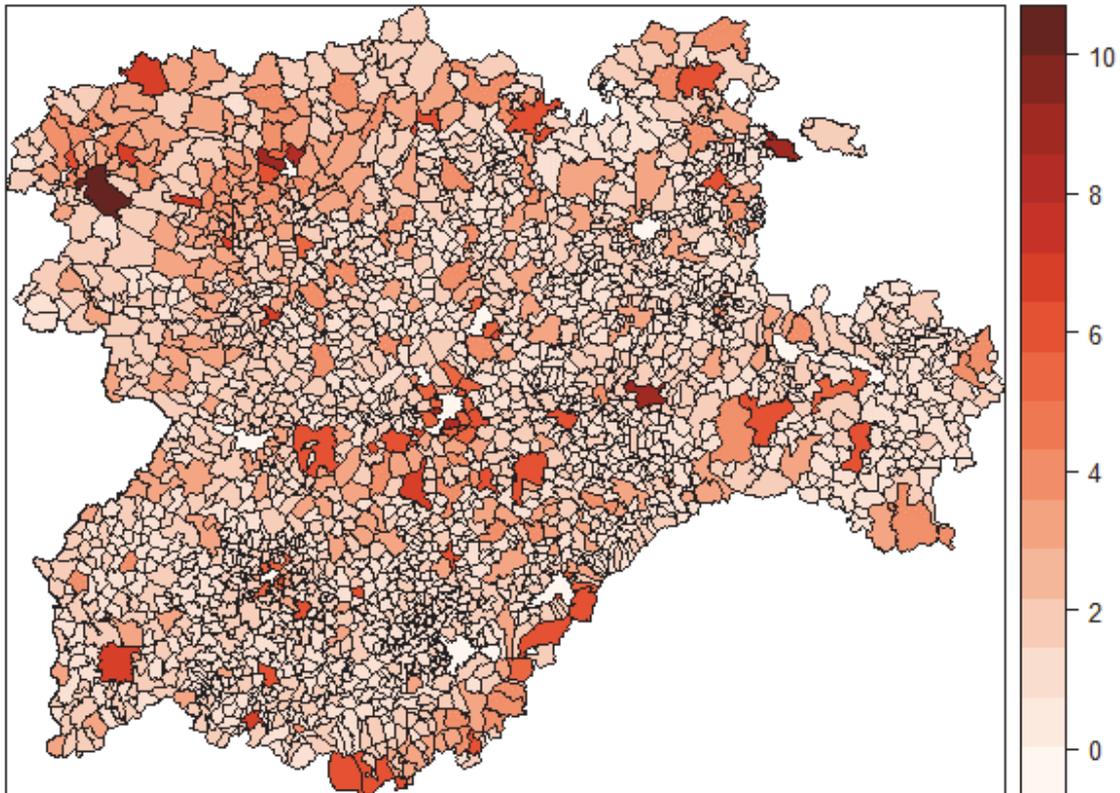


Figura 32: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de movimiento natural de población (No Jerárquico)

### Jerárquico vs No Jerárquico

En este caso se aprecia una clasificación por grupos bastante similar entre el Análisis Jerárquico y el No Jerárquico. No obstante, este segundo ofrece un mayor equilibrio entre grupos en cuanto a cantidad de municipios recogidos.

### 3. Variables de estadística de migraciones

#### Jerárquico

Se observa en la *Tabla 27* un coeficiente más elevado utilizando la medida de agrupación de Ward, por lo que será la que se utilice para este grupo de variables.

Promedio	Simple	Completo	Ward
0.9976009	0.9975437	0.9984842	0.9993228

Tabla 27: Coeficientes de los métodos de agregación para las variables de estadística de migraciones

Las *Figuras 33 y 34* representan los grupos obtenidos mediante el análisis cluster con 10 y 11 clusters, respectivamente. La diferencia entre estos dos gráficos es que en el primero, los municipios de Arroyo de la Encomienda, Villaquilambre, Santa Marta de Tormes y Laguna de Duero forman un grupo, mientras que en el segundo gráfico se divide en dos clusters. Se aprecian diferencias suficientes entre estos dos clusters, por tanto, finalmente se han seleccionado 11 clusters para el grupo de migraciones. Se observan dos grupos representados por un único municipio, Ponferrada y San Andrés del Rabanedo. Estos dos grupos, tienen un valor bastante alto para la primera componente. Esto es debido al elevado movimiento migratorio para las variables “Emigraciones a municipios de la misma provincia”, “Inmigraciones desde municipios de la misma provincia”. La principal diferencia entre estos dos grupos se encuentra en las variables “Emigraciones a municipios de otra comunidad” e “Inmigraciones desde municipios de otra comunidad”, para las que Ponferrada ofrece valores muy elevados, al contrario que San Andrés del Rabanedo, lo cual se puede apreciar gracias al valor de la segunda componente en la *Figura 34*. Llama la atención el grupo formado por los municipios de Aranda de Duero y Miranda de Ebro, que a pesar de tener una población bastante superior a los municipios del grupo 9 (*Figura 34*), este último cuenta con un movimiento migratorio dentro de la misma provincia bastante más elevado. En cambio, Aranda de Duero y Miranda de Ebro, destacan por su gran movimiento con otras comunidades y, principalmente, por su elevada cifra de inmigraciones desde municipios de otros países.

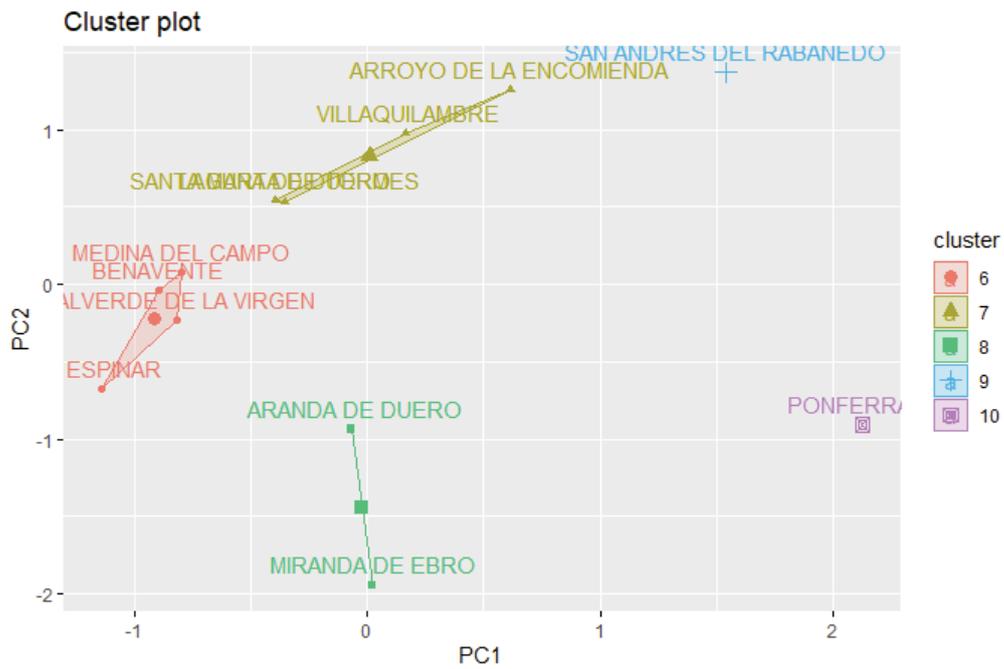


Figura 33: Grupos principales para las variables de estadística de migraciones (10 clusters)

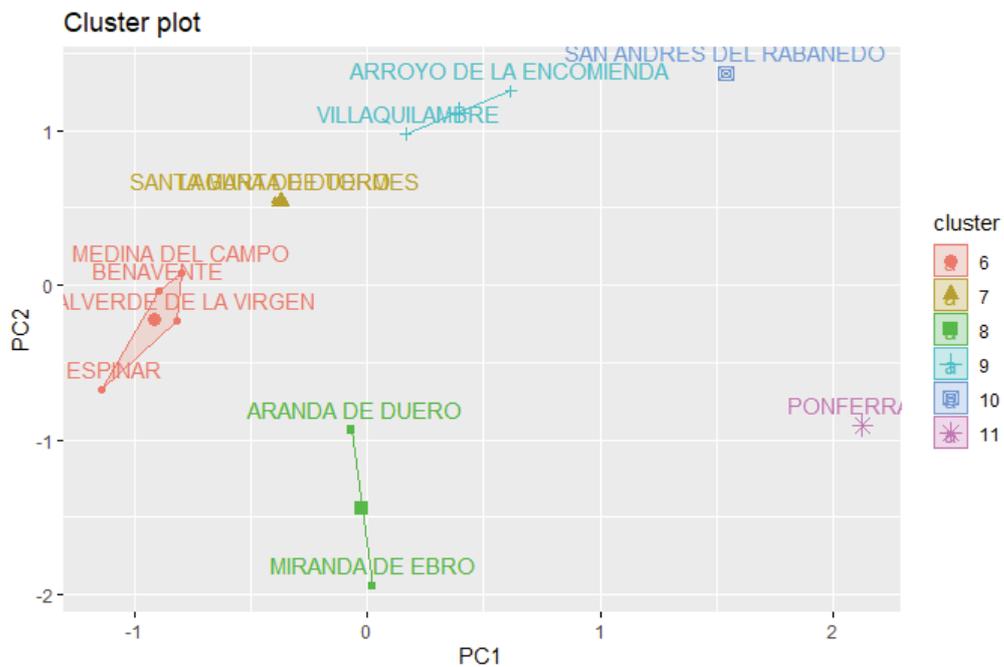


Figura 34: Grupos principales para las variables de estadística de migraciones (11 clusters)

La *Figura 35* representa los grupos formados a partir del análisis cluster jerárquico para el grupo de variables de estadística de migraciones. Visualmente se aprecia un gran número de clusters que cuentan con muy pocos municipios. Esto ocurre debido a las grandes diferencias que tienen unos pocos municipios con respecto al resto. Una vez más, Ponferrada cuenta con cifras bastante elevadas en comparación con el resto para prácticamente todas las variables. San Andrés del Rabanedo, municipio que representa el décimo cluster, destaca por su elevado movimiento migratorio dentro de la misma provincia. Por otro lado, municipios de dimensiones poblacionales semejantes como Aranda de Duero y Miranda de Ebro, que representan otro cluster, destacan por sus elevadas cifras en movimientos migratorios fuera de la comunidad, como se comentaba anteriormente, así como sobresale sobre cualquier otro municipio el número de inmigraciones desde otro país. Se pueden considerar, por tanto, estos dos municipios, los destinos preferidos por extranjeros para vivir en Castilla y León. En el mapa también se observan bastantes clusters formados por una enorme cantidad de municipios. El primer grupo se compone, de los municipios que tienen un movimiento migratorio prácticamente nulo. Este grupo es muy abundante, lo que quiere decir que Castilla y León es una comunidad con escaso movimiento migratorio.

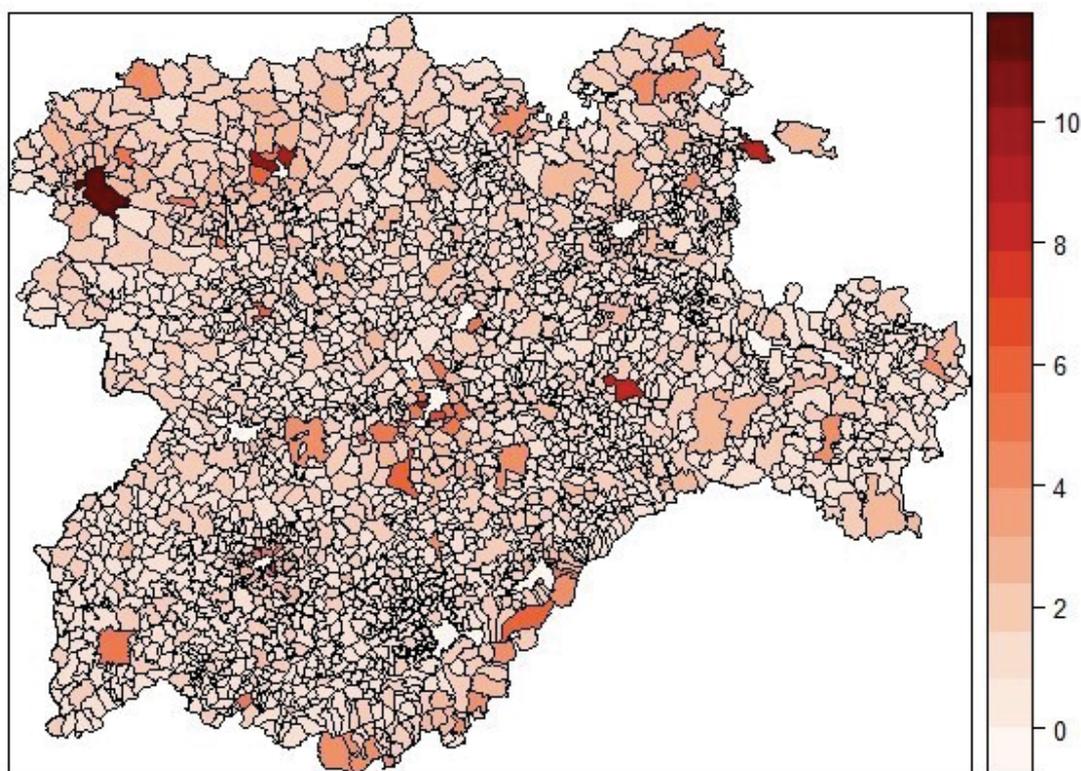


Figura 35: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de migraciones

### No Jerárquico

Se ha realizado un Análisis Cluster No Jerárquico a partir de un número de grupos igual al obtenido en el Análisis Jerárquico, en este caso 11 grupos. La *Figura 36* muestra los más relevantes. Se observa un grupo formado por un único municipio, Ponferrada, el cual toma un valor negativo de la segunda componente, lo que implica que existe menor movimiento migratorio dentro de la misma provincia y mayor movimiento migratorio principalmente con otras comunidades, y, en menor medida, con otros países. Arroyo de la Encomienda, San Andrés del Rabanedo y Villaquilambre, forman un grupo, el cual toma un valor elevado para la segunda componente. Esto quiere decir que predominan los movimientos migratorios dentro de la misma provincia. El grupo formado por Aranda de Duero y Miranda de Ebro, de la misma forma que Ponferrada, toma valores elevados para las variables “Emigraciones a municipios de otras comunidades”, “Inmigraciones desde municipios de otras comunidades” e “Inmigraciones desde municipios de otros países”.

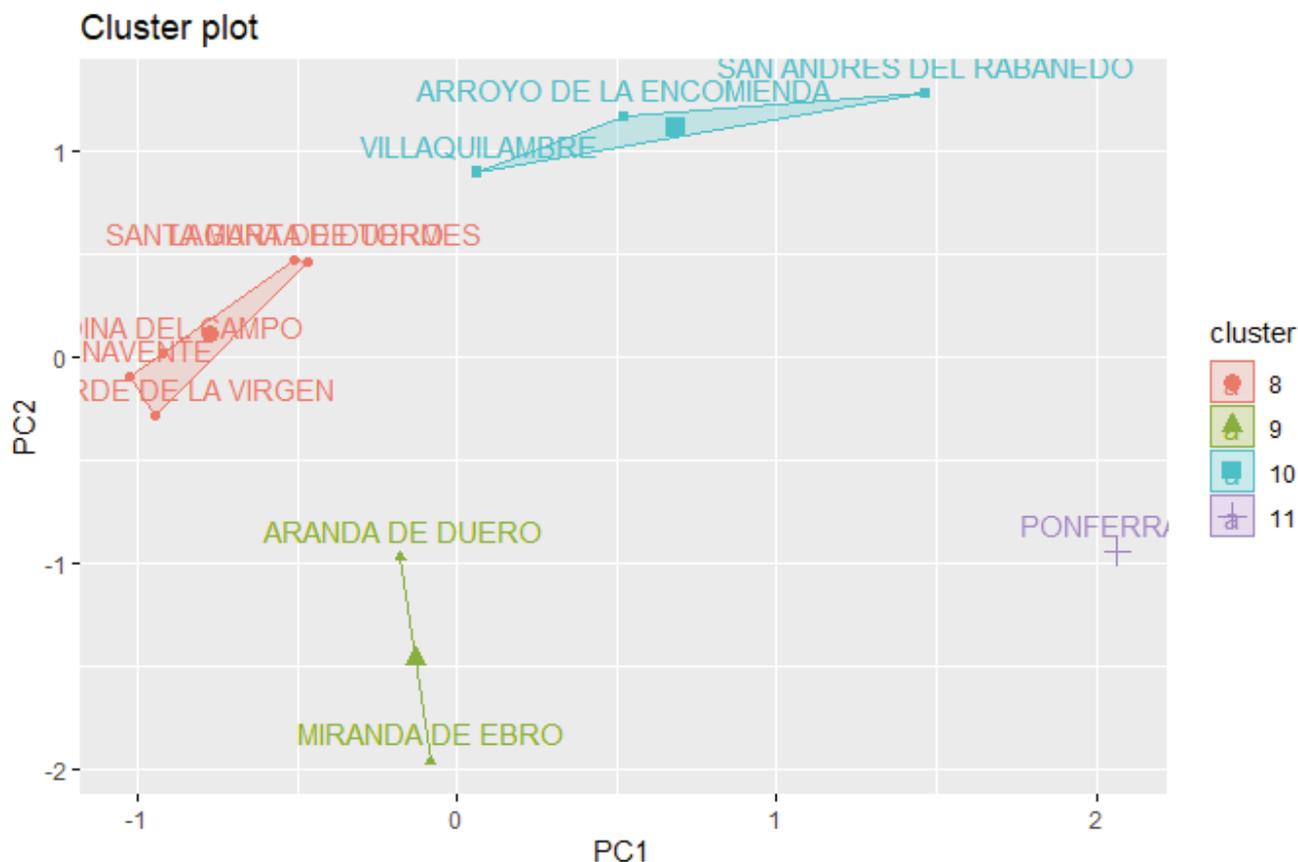


Figura 36: Grupos principales para las variables de estadística de migraciones (No Jerárquico)

En la *Figura 37* se observan los municipios de Castilla y León con sus correspondientes grupos. Se aprecia una enorme cantidad de municipios en colores claros, los cuales representan grupos con municipios con valores bajos para las variables de estadística de migración, lo que hace referencia a la importante población rural con la que cuenta Castilla y León.

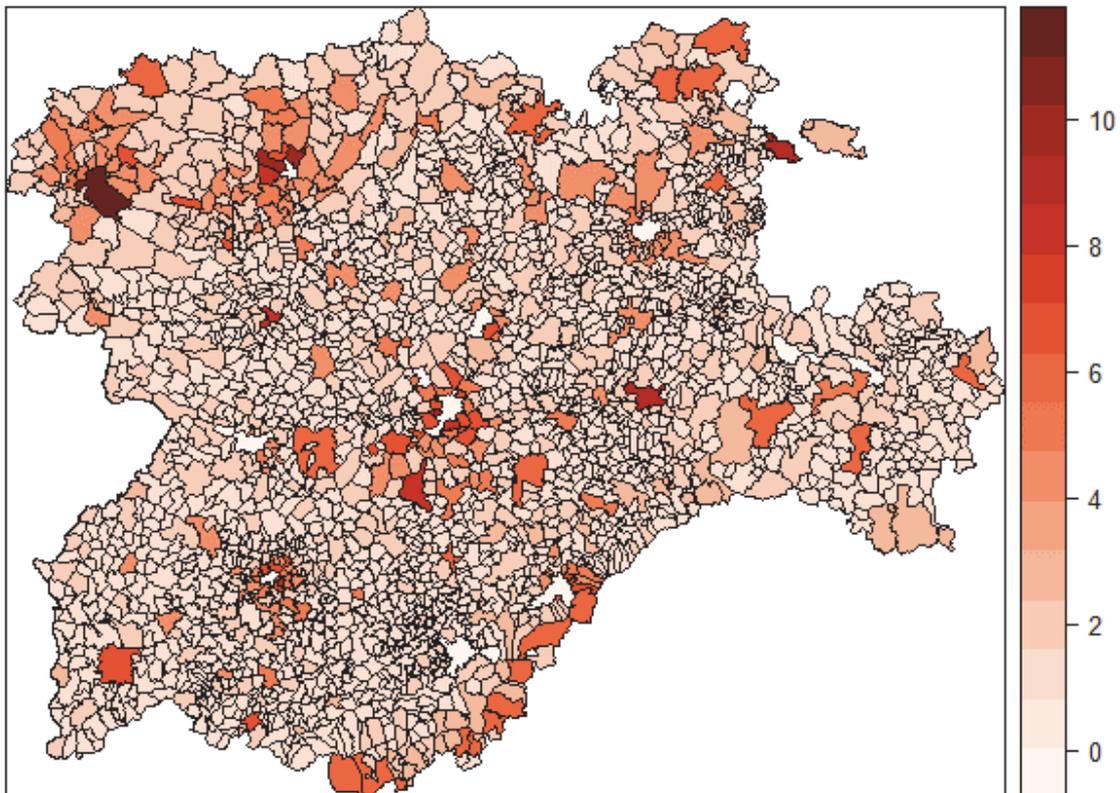


Figura 37: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el grupo de variables de estadística de migraciones (No Jerárquico)

### Jerárquico vs No Jerárquico

- San Andrés del Rabanedo, que en el Análisis Jerárquico formaba un grupo, pasa a clasificarse junto con Villaquilambre y Arroyo de la Encomienda en el Análisis No Jerárquico.
- Los municipios de Santa Marta de Tormes y Laguna de Duero también formaban un grupo tras realizar un agrupamiento de tipo jerárquico. Sin embargo, en el Análisis No Jerárquico forman grupo junto con los municipios de Benavente, Medina del Campo y Valverde de la Virgen.
- El municipio de El Espinar, que se clasifica con los municipios de Benavente, Medina del Campo y Valverde de la Virgen siguiendo un Análisis Jerárquico, no se muestra dentro de los grupos relevantes obtenidos tras el agrupamiento no jerárquico.
- El Análisis no Jerárquico ofrece un mayor equilibrio entre grupos en cuanto a cantidad de municipios recogidos.

Se puede deducir, por tanto, que el Análisis Jerárquico clasifica mejor los municipios de estadísticas más elevadas, haciendo una mayor diferenciación, mientras que el Análisis no Jerárquico cuida más dicha diferenciación entre grupos con estadísticas más bajas.

#### 4. Variables del Grupo 1, Grupo 2 y Grupo 3

##### Jerárquico

La *Tabla 28* muestra los coeficientes resultantes de un Análisis Cluster sobre el conjunto completo de variables según las diferentes medidas de agrupación. El más elevado se consigue agrupando las observaciones con la técnica de Ward, por lo que será la empleada para realizar el Análisis Cluster.

Promedio	Simple	Completo	ward
0.9992585	0.9990049	0.9994527	0.9996343

Tabla 28: Coeficientes de los métodos de agregación para las variables del Grupo 1, Grupo 2 y Grupo 3

En este caso se ha realizado el Análisis Cluster sobre 3 componentes, para recoger más información de variables que no son de población.

En la *Figura 38* se observan los grupos seleccionados tras el Análisis Cluster. Se han seleccionado 13 clusters por las grandes diferencias existentes entre los datos, principalmente con respecto a la población. Las variables más significativas son las de población, por lo que los grupos que se aprecian en el gráfico contienen observaciones semejantes en cuanto a población, teniendo menos en cuenta el resto de las variables. El municipio de Ponferrada, con 64000 habitantes, de nuevo supone un grupo, por lo que claramente se puede concluir en que cuenta con importantes diferencias con respecto al resto de municipios. También destacan los municipios de San Andrés del Rabanedo, Miranda de Ebro y Aranda de Duero, con más de 30000 habitantes, que forman otro cluster. El gráfico muestra una clara diferencia entre los municipios de San Andrés del Rabanedo y Aranda de Duero y Miranda de Ebro para la segunda componente, ocasionada por el elevado valor que toman las variables “Emigraciones a municipios de la misma provincia” e “Inmigraciones desde municipios de la misma provincia” para el primer municipio. No obstante, que pertenezcan a un mismo grupo a pesar de esta diferencia reafirma la fuerza que tienen las variables poblacionales en este estudio. Se pueden observar varios grupos que cuentan con municipios con elevada población, no obstante, estos grupos no son muy numerosos. Ocurre lo contrario con los grupos que recogen municipios pequeños, que cuentan con una gran cantidad de observaciones. Esto es un claro indicador de que Castilla y León es una comunidad rural, es decir, formada por numerosos pueblos pequeños.

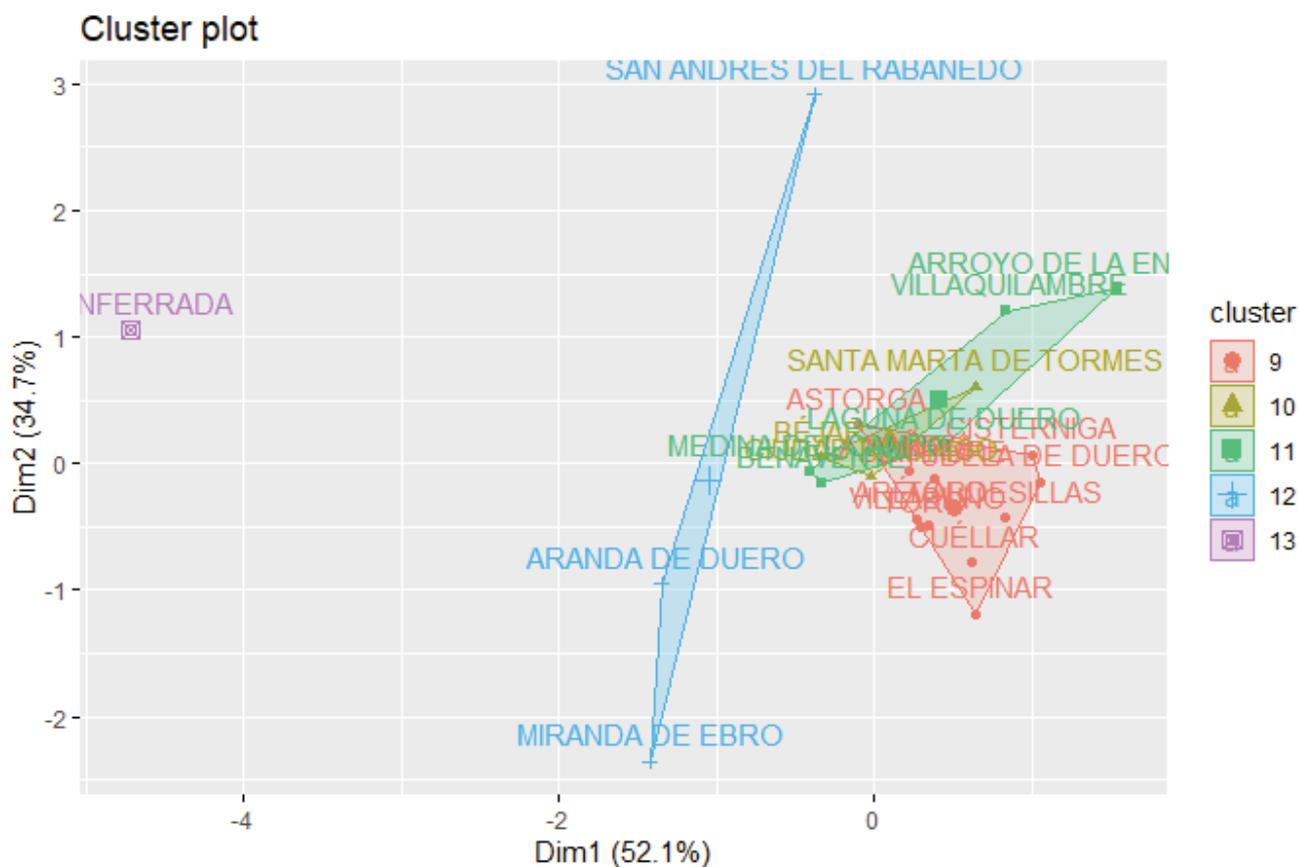


Figura 38: Grupos principales para las variables del Grupo 1, Grupo 2 y Grupo 3

La *Figura 39* muestra los municipios de Castilla y León con sus respectivos grupos obtenidos mediante el Análisis Cluster. En el mapa que representaba los movimientos naturales de la población de Castilla y León se apreciaba una importante cantidad de grupos que recogían muy pocas observaciones, y por tanto algún grupo que recogía una enorme cantidad de municipios. A diferencia, en este se puede observar un mayor equilibrio, con varios grupos que recogen numerosos municipios. Estos grupos tan numerosos contienen, en general, municipios con pequeñas poblaciones. De hecho, los 9 primeros grupos contienen municipios que no superan los 11000 habitantes. De esta forma, se puede asegurar lo que se comentaba anteriormente, que Castilla y León es una comunidad rural, ya que cuenta con una numerosa cantidad de municipios pequeños.

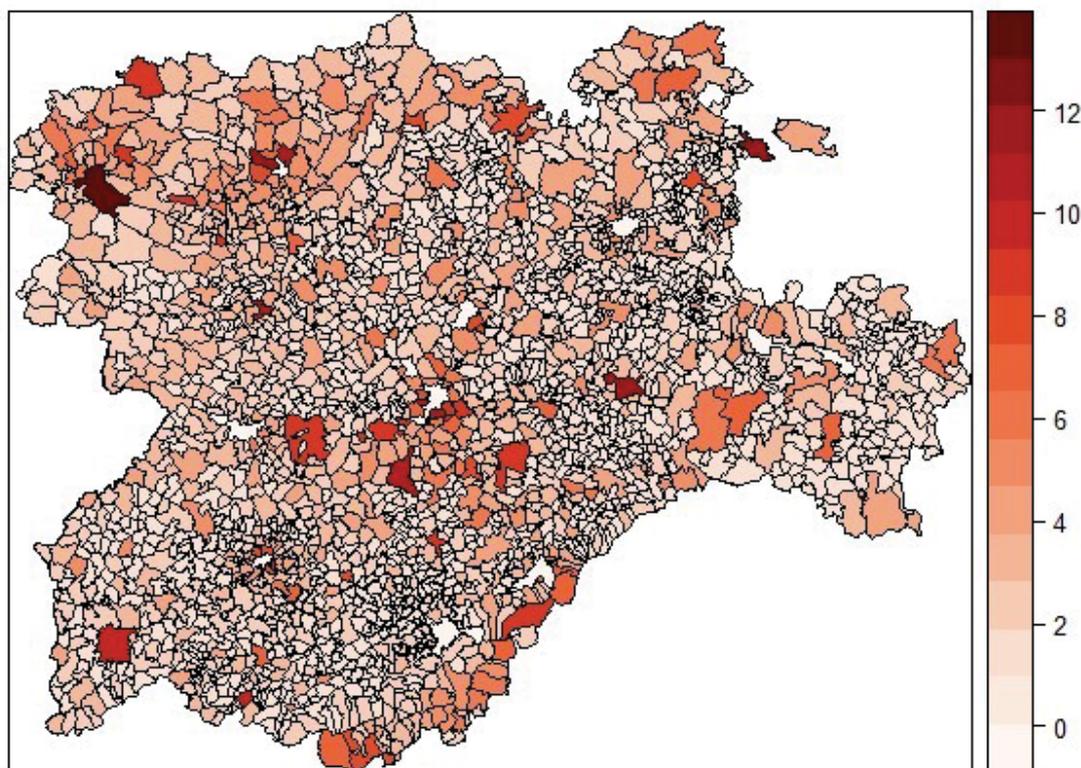


Figura 39: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el conjunto de variables del Grupo 1, Grupo 2 y Grupo 3

### No Jerárquico

Para este Análisis Cluster no Jerárquico, se han utilizado las variables de estadística de población, las variables de estadística de movimiento natural de la población y las variables de estadística de migración. El número de grupos con que se ha realizado la clasificación de municipios es el mismo que el obtenido mediante el Análisis Jerárquico, 13.

En la *Figura 40* se observan 3 grupos, considerados los más relevantes por contar con valores más elevados para todas las variables. Uno de ellos está formado por un único municipio, Ponferrada, que, cuenta con importantes diferencias con respecto al resto de municipios. En el grupo 12, formado por municipios como San Andrés del Rabanedo, Aranda de Duero o Medina del Campo, se aprecian diferencias significativas entre los municipios incluidos. En cantidad de población, se encuentran desde municipios de 18000 habitantes, como Benavente, hasta municipios de 35000 habitantes, como Miranda de Ebro, lo que supone una gran diferencia. Por otro lado, también se aprecian importantes diferencias en variables de estadística de migración. Se observa un valor de la segunda componente muy elevado para el municipio de San Andrés del Rabanedo, lo que indica valores muy significativos para las variables “Emigraciones a municipios de la misma pro-

vincia” e “Inmigraciones desde municipios de la misma provincia”. Por el contrario, municipios como Miranda de Ebro y Aranda de Duero, toman valores muy bajos para la segunda componente, lo que representa valores bajos para las variables mencionadas, y muy significativos para las variables “Emigraciones a municipios de otras comunidades”, “Inmigraciones desde municipios de otras comunidades” e “Inmigraciones desde municipios de otros países”.

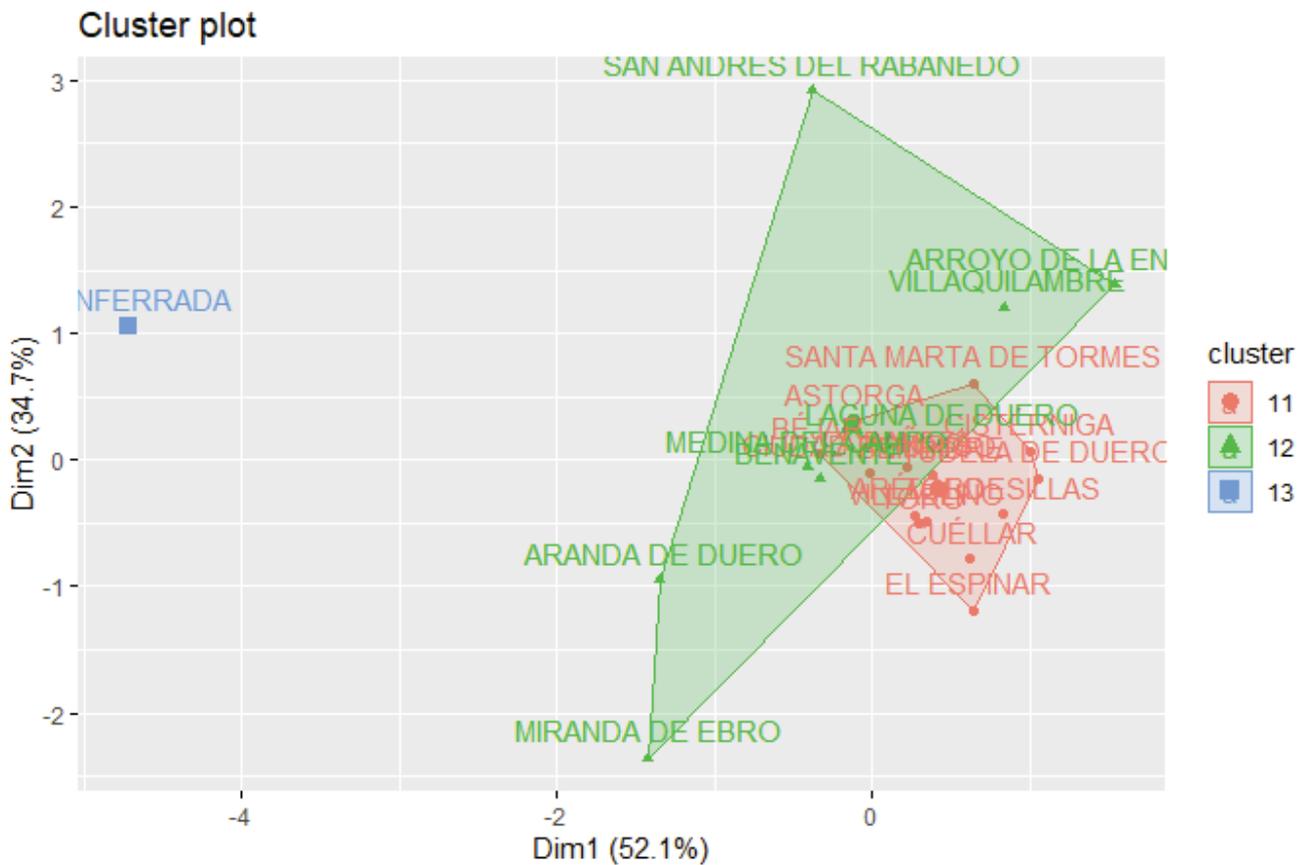


Figura 40: Grupos principales para las variables del Grupo 1, Grupo 2 y Grupo 3 (No Jerárquico)

La Figura 41 muestra todos los municipios de Castilla y León y el grupo asignado para cada uno de ellos. Se observan grupos bastante numerosos, exceptuando el formado por Ponferrada. En gran parte, estos grupos recogen municipios con valores bajos para las variables que se han empleado en el estudio.

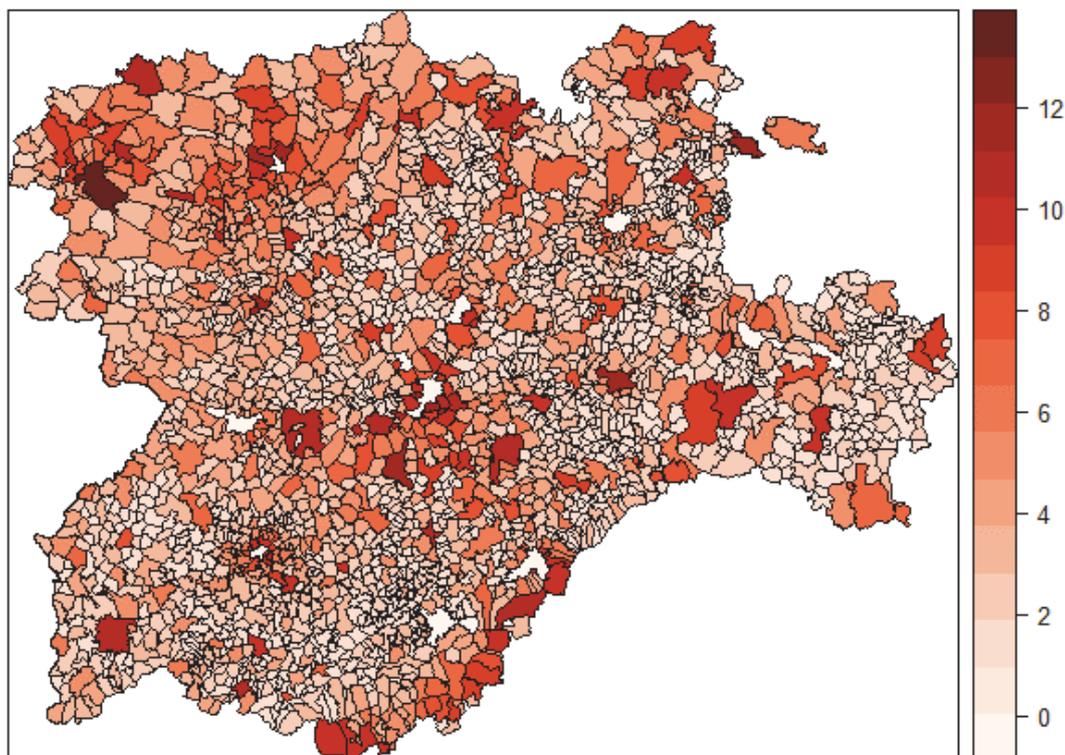


Figura 41: Mapa de CyL que muestra el grupo al que pertenece cada municipio para el conjunto de variables del Grupo 1, Grupo 2 y Grupo 3 (No Jerárquico)

### Jerárquico vs No Jerárquico

- El Clustering Jerárquico clasifica los municipios más significativos en 5 grupos.
- El Clustering no Jerárquico clasifica los municipios más significativos en 3 grupos. De esta forma, existe un grupo que cuenta con municipios con importantes diferencias, como Miranda de Ebro y Benavente, que para la variable “Población Total”, el primero de ellos dobla al segundo. A su vez, otro cluster agrupa municipios como Santa Marta de Tormes, de 14500 habitantes, con otros como El Espinar, con 9000.

Se ha observado, para todos los grupos de variables analizados, que el Análisis Cluster Jerárquico, en general hace mayor diferenciación en municipios más significativos, lo que da lugar a determinados grupos muy numerosos, y otros con escasos municipios. El Análisis Cluster no Jerárquico, por el contrario, realiza una clasificación más minuciosa para municipios con valores más bajos para las variables estudiadas, ofreciendo así grupos menos numerosos, y, por tanto, más equilibrados. Teniendo en cuenta los resultados, se considera mejor el Análisis Cluster Jerárquico, ya que ofrece grupos más homogéneos.

## 5. Conclusiones

El análisis estadístico multivariante realizado utilizando la información proporcionada por el SIE de la Junta de Castilla y León pone de manifiesto que:

- El municipio de Ponferrada hace referencia a un cluster para todos los grupos de variables, mostrando una gran diferencia con el resto de los municipios.
- En el grupo de variables de estadística de movimiento natural de la población, los municipios de Villaquilambre, Laguna de Duero y Arroyo de la Encomienda, representan población joven, debido a su significativa cantidad de nacimientos y matrimonios.
- En el grupo de variables de estadística de movimiento natural de la población, el municipio de Ponferrada, con un elevado número de defunciones en comparación con los nacimientos y los matrimonios, se considera que cuenta con población más anciana. Lo mismo ocurre con municipios como Béjar o Astorga.
- El número de defunciones aporta más información al estudio que el número de nacimientos, ya que en general, es bastante superior, lo que indica que Castilla y León cuenta una importante cantidad de población anciana.
- En el grupo de variables de estadística de migraciones, destacan por su aportación las migraciones dentro de la misma provincia y con municipios de otras comunidades.
- En el grupo de variables de estadística de migraciones, San Andrés del Rabanedo representa un cluster, siendo el municipio con más movimientos migratorios dentro de la misma provincia.
- Villaquilambre y Laguna de Duero, también destacan por su significativa cantidad de movimientos migratorios dentro de la misma provincia. Las cifras son semejantes a las de Ponferrada, a pesar de que este último municipio cuenta con tres veces la población de Villaquilambre y Laguna de Duero.
- En el grupo de variables de estadística de migraciones, los municipios de Aranda de Duero y Miranda de Ebro suponen la mayor aportación a las inmigraciones desde municipios de otros países, pudiendo considerarlas así, los destinos preferidos por extranjeros en Castilla y León.
- Se aprecia una evidente mejora del estudio con la eliminación de capitales de provincia, ofreciendo una mayor claridad de los datos y una clasificación de los municipios más adecuada.
- El cluster que recoge los municipios con valores más bajos para los indicadores demográficos utilizados, es muy numeroso para cada grupo de variables, lo que demuestra que Castilla y León tiene una cantidad muy significativa de pueblos pequeños.
- Todas las variables utilizadas para el estudio tienen una elevada dependencia de la variable que mide la población total, por tanto, para el análisis estadístico de todos los grupos de variables, el tamaño de población aporta la principal información en la clasificación de los municipios.

Sería interesante, como trabajo futuro, analizar la evolución temporal de los municipios de Castilla y León utilizando información demográfica de diversos años. Esta información está también disponible en el Sistema de Información Estadística (SIE) de la Junta de Castilla y León.

## Referencias

- [1] SIE, “Selección de datos - Datos Básicos.” <https://www.jcyl.es/sie/v2/datosbasv2irAmodulo.html>, 2019. [Online; último acceso 18 de Abril de 2020].
- [2] S. de la Fuente Fernández, “Análisis Factorial UAM.” <http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIAL/analisis-factorial.pdf>, 2011. [Online; último acceso 25 de Abril de 2020].
- [3] U. C. I. de Madrid, “Análisis Factorial.” <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema4am.pdf>. [Online; último acceso 29 de Abril de 2020].
- [4] M. Chávez, “Introducción a los métodos multivariantes.” [https://www.rpubs.com/marcelo-chavez/multivariado\\_1](https://www.rpubs.com/marcelo-chavez/multivariado_1), 2017. [Online; último acceso 5 de Mayo de 2020].
- [5] “Test de Bartlett.” <https://vivaelsoftwarelibre.com/test-de-bartlett-para-contrastar-la-homocedasticidad-de-mas-de-2-muestras-usando-r-com> [Online; último acceso 12 de Mayo de 2020].
- [6] S. de la Fuente Fernández, “Análisis de Componentes Principales UAM.” [http://www.estadistica.net/Master-Econometria/Componentes\\_Principales.pdf](http://www.estadistica.net/Master-Econometria/Componentes_Principales.pdf), 2011. [Online; último acceso 14 de Mayo de 2020].
- [7] S. de la Fuente Fernández, “Análisis Cluster UAM.” [http://www.estadistica.net/Master-Econometria/Analisis\\_Cluster.pdf](http://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf), 2011. [Online; último acceso 18 de Mayo de 2020].
- [8] D. Calvo, “Clúster Jerárquicos y No Jerárquicos.” <http://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/>, 2018. [Online; último acceso 18 de Mayo de 2020].
- [9] M. Ángel Giménez Cuadrillero, “Clustering Jerárquico en R.” <https://rpubs.com/mjimcua/clustering-jerarquico-en-r>, 2018. [Online; último acceso 20 de Mayo de 2020].
- [10] U. of Cincinnati, “K-means Cluster Analysis.” [https://uc-r.github.io/kmeans\\_clustering](https://uc-r.github.io/kmeans_clustering), 2018. [Online; último acceso 23 de Mayo de 2020].
- [11] J. A. Rodrigo, “Análisis de Componentes Principales y t-SNE.” [https://rpubs.com/Joaquin\\_AR/287787](https://rpubs.com/Joaquin_AR/287787), 2017. [Online; último acceso 29 de Mayo de 2020].
- [12] D. Garavito, “Agrupamiento (clustering).” [http://rstudio-pubs-static.s3.amazonaws.com/474357\\_50a1f0fed64d4428b99216cd460a4896.html](http://rstudio-pubs-static.s3.amazonaws.com/474357_50a1f0fed64d4428b99216cd460a4896.html). [Online; último acceso 1 de Junio de 2020].
- [13] J. Ángel Gallardo San Salvador, “Métodos No Jerárquicos de Análisis Cluster.” <https://www.ugr.es/~gallardo/pdf/cluster-4.pdf>. [Online; último acceso 1 de Junio de 2020].

- [14] U. de Oviedo, “El algoritmo k-means.” [https://www.unioviedo.es/compnum/laboratorios\\_py/kmeans/kmeans.html#:~:text=procesamiento%20de%20im%C3%A1genes.-,El%20algoritmo%20k%2Dmeans,suele%20usar%20la%20distancia%20cuadr%C3%A1tica.](https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html#:~:text=procesamiento%20de%20im%C3%A1genes.-,El%20algoritmo%20k%2Dmeans,suele%20usar%20la%20distancia%20cuadr%C3%A1tica.) [Online; último acceso 3 de Junio de 2020].