Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en estadística

Clasificación de los municipios de Castilla y León en función de su evolución demográfica

Autor: Marina Toquero Asensio

Tutor/es:

Eusebio Arenal Gutiérrez

Jesús M. Rodríguez Rodríguez

RESUMEN

El trabajo tiene como objetivo clasificar los municipios de Castilla y León en función de su evolución demográfica. Para ello se utilizan varios indicadores demográficos. Los datos se obtienen del módulo de Datos Básicos del SIE (Sistema de Información Estadística de Castilla y León), así como del módulo Padrón de este mismo sistema. La clasificación se realiza mediante diversas técnicas de análisis de datos implementadas en el lenguaje R, obteniéndose grupos de municipios con características demográficas similares.

ABSTRACT

This proyect's objective is to classify the municipalities of Castilla y León according to their demographic evolution using various demographic indicators to measure them. The data obtained from the SIE (Sistema de Información Estadística de Castilla y León), referred to as “Datos Básicos” (Basics data) and “Padron” (Census), will help provide a classification system by similarities between municipalities. Al the tasks had been carried out using different data analysis techniques applied with the R software.

INDICE

TABLA DE ILUSTRACIONES

1 INTRODUCCIÓN

En este Trabajo de Fin de Grado se utilizan datos demográficos de Castilla y León de un largo periodo de años, desde 1996 a 2019. Utilizando los indicadores demográficos de este periodo se realiza una clasificación de los municipios de Castilla y León. El número de municipios de Castilla y León es elevado, la cuarta parte de los municipios de España, y gran mayoría de ellos tienen poca población. La revolución industrial produjo que muchas personas se desplazaran a zonas urbanas, lo que provocó un aumento de estas zonas en detrimento de las zonas rurales. Por este motivo gran parte de la población en Castilla y León se concentra en un número reducido de municipios.

Para la obtención de los datos se utiliza el servicio SIE (Sistema de Información Estadística) de la Junta de Castilla y León. En el SIE se encuentra la Información Estadística más relevante de la comunidad autónoma de Castilla y León, está organizado en módulos temáticos según característica comunes de dichos módulos. La información está disponible para cualquier usuario, que podrá en cada momento elegir la información que desee dentro de las opciones ofrecidas por el sistema.

El capítulo 2 trata de la obtención y tratamiento de los datos. Al final de este proceso se dispone de 2.248 registros por variable correspondientes con la desagregación municipal. Al realizar una descripción de los indicadores demográficos se observa que es conveniente hacer una primera división de los municipios en dos grupos según su población: un grupo formado por los municipios de más de 20000 habitantes y otro por los de menos.

En el capítulo 3 se realiza una breve descripción de los métodos estadísticos que se utilizaran en los análisis de la capítulo 4.

En el capítulo 4 se obtienen diferentes clasificaciones para cada uno de estos dos grupos. En todos los casos se realiza un análisis de componentes principales para la reducción de la dimensionalidad, y diferentes análisis cluster (jerárquicos y no jerárquicos) para la clasificar los municipios. Finalmente se utiliza el multidimensional scaling para representar los grupos formados. Además de realizarse para todas las variables, cada uno de estos análisis se hace para 3 grupos de variables: movimiento natural de la población, estadísticas de población y migración.

2 OBTENCIÓN, TRATAMIENTO Y ANÁLISIS DESCRIPTIVO DE LOS DATOS

Los datos se obtienen del SIE (Sistema de Información Estadística) de la Junta de Castilla y León. Se realiza una consulta personalizada de la siguiente manera, ya que deseamos obtener para cada indicador demográfico un fichero de datos.

Para este Trabajo de Fin de Grado usaremos fundamentalmente el módulo de "Datos Básicos ", y algún dato obtenido del módulo "Padrón". En el módulo de "Datos Básicos" tenemos varias familias de indicadores, usaremos los "Indicadores demográficos" junto de la opción de a "Selección por municipios". En los datos procedentes del Padrón no se dispone del año 1997, ya que en dicho año se produjo un parón en el Padrón de habitantes.

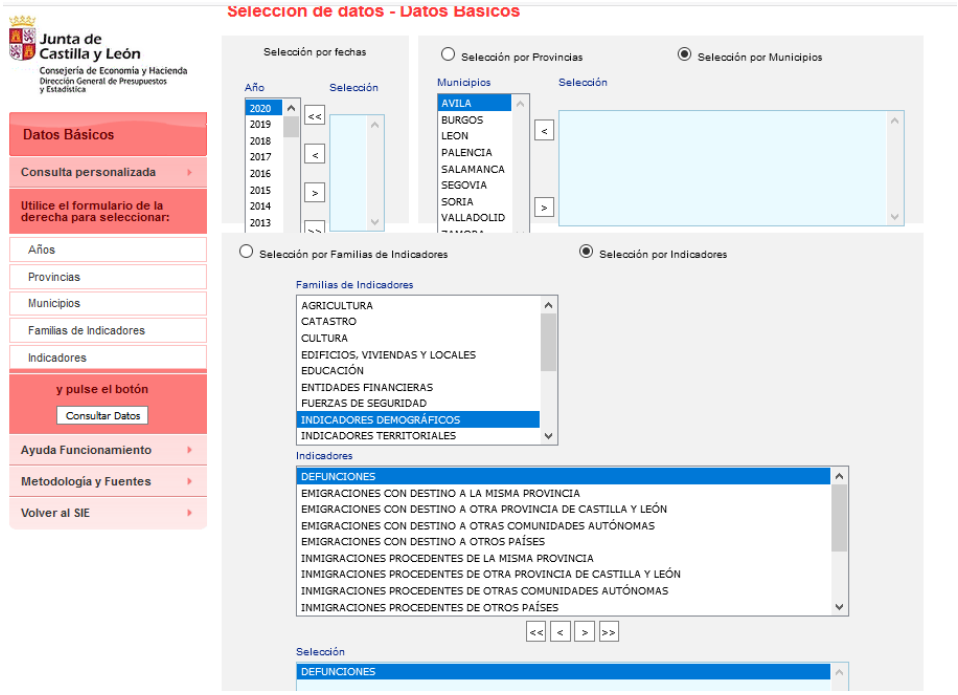

Ilustración 1 : Consulta personalizada de datos

Seleccionamos el periodo de años desde 1996 a 2019 con las opciones, como podemos observar en la *Ilustración 1*, e iremos seleccionado cada indicador demográfico. En las tablas tenemos como filas y columnas los municipios y los años, respectivamente.

Los datos para utilizarlos en R no están acondicionados. Para hacerlos "entendibles" para R debemos realizar alguna modificación, la cual realizamos con el software Excel.

Al visualizar los datos observamos que tenemos valores como ".". Estos son valores se sustituyen por cero ya que él SIE es la manera que tiene para dar la ausencia de ese municipio en lo que se mide. Se dan, sobre todo, en municipios con poca población.

Además de estos cambios debemos realizar alguna modificación más como seria la eliminación de la fila de total. Esta última fila es la suma de cada columna y, no es un municipio. La fila que pone nombre de la variable, así como los espacios en blanco que aparecen en el inicio deberán ser eliminados también.

Cuando cargamos los datos con R, los usaremos como DataFrames. Tendremos cada variable como un DataFrame, donde estarán los distintos valores para cada municipio así como el código INE del municipio. Es decir, en los ficheros tenemos por filas cada municipio y, por columnas cada año, desde 1996 a 2019. En la última columna tenemos el código INE de cada municipio. Dado que en unos ficheros tenemos el año 1997 y, en otros no, trabajaremos sin ese año.

Para describir las variables se dividen en grupos: uno formado por las variables clásicas que se usan en demografía y otro con las variables de migración.

Dada la naturaleza de los datos será necesario utilizar los datos originales del *conteo* y las tasas relativas a la población del municipio.

$$TT = \frac{N_{ij}}{P_{ij}} * 1000$$

TASA DE VARIABLE T
N_{ij} : número de variable T en periodo *i* en municipio *j*
P_{ij}: población de periodo *i* en municipio *j*

Estas tasas nos permiten una mejor comparación entre municipios.

Con los datos originales realizamos los gráficos de evolución total y de evolución por municipio de cada variable . En el gráfico de evolución por municipio podremos ver la evolución de cada municipio en todos los periodos. En estos gráficos algunas curvas se encuentran por encima de las demás; estas corresponden, en la mayoría de los casos, a municipios de más de 20.000 habitantes. Esto nos da una idea de la necesidad de realizar una división previa en dos grupos: los municipios de más de 20.000 habitantes y los de menos de 20.000 habitantes.

Los mapas se realizan con la variación relativa entre los años 1998 y 2018 (20 años) relativa al año de inicio para cada variable D:

$$D_{2018} - D_{1998} \,/\, D_{1998}$$

2.1 VARIABLES CLÁSICAS

Dentro de esta división podemos encontrar variables, como número población de derecho mujeres, población de derecho varón, número de nacimientos, etc. Además, dentro de variables clásicas podríamos dividirlo en 2 grupos: las de estadísticas de población y las relativas al movimiento natural de la población.

- Variables de movimiento natural de la población: Defunciones, Nacimientos y Mujeres en edad reproductiva.
- Variables de estadísticas de población: Población total, Población de derecho varón y Población de derecho mujer.

Defunciones

El número de defunciones que se han producido en cada municipio por cada año. El número medio de muertes que se producen por año es 27.126. Existen muy pocos municipios, en los cuales, en algún momento no exista ninguna muerte o, para los que

no tengamos datos, es decir, para la gran mayoría conocemos el número de muertes en todo el periodo de tiempo.

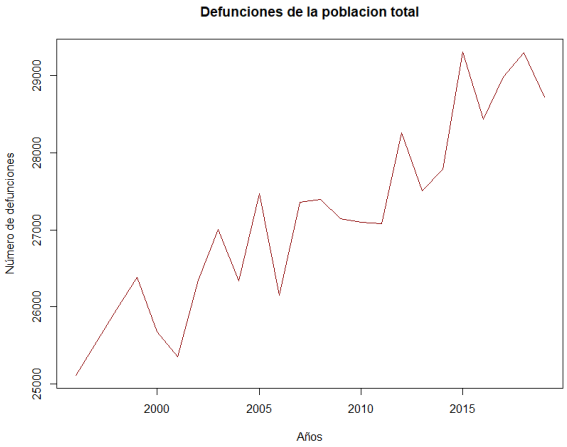

Ilustración 2: Defunciones totales por año

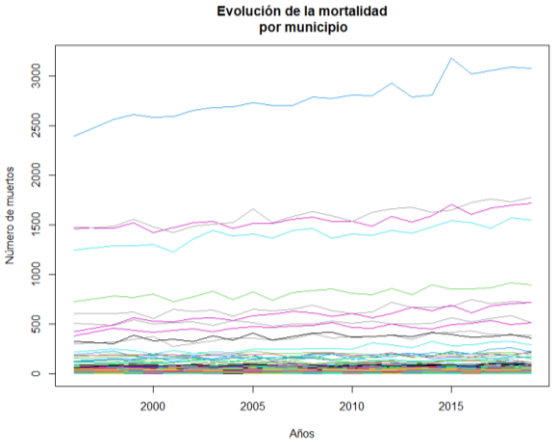

Ilustración 3: Evolución mortalidad por año

En *Ilustración 2* tenemos el número de muertes total por año. Observamos una tendencia creciente en el número de las defunciones. En *Ilustración 3* podemos observar la evolución de la mortalidad en cada municipio, siendo cada línea un municipio. Dada la cantidad de municipios que tenemos, no podemos discernir ninguna conclusión, únicamente, podemos ver que existen unos municipios con muchas muertes. Para poder observar alguna conclusión, realizamos la tasa de mortalidad bruta en cada año.

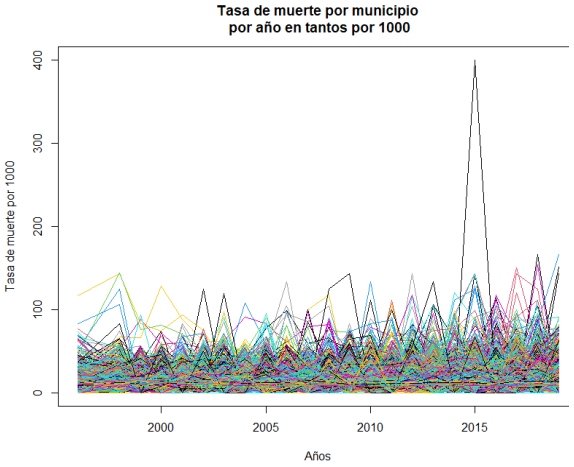

Ilustración 4:Tasa mortalidad por municipio y año

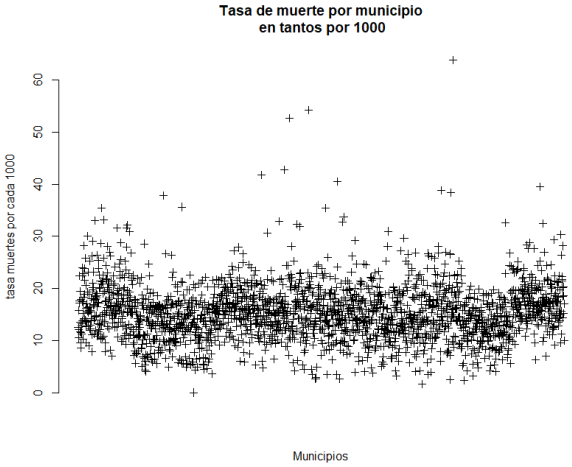

Ilustración 5: Media de tasa de muerte por municipio

Podemos ver en *Ilustración 4* que las tasas de muerte por año se encuentran entre 0 y 150 por cada 1.000 habitantes. Observamos un aumento o pico en 2015, correspondiente al municipio con código INE 09184 que corresponde con Jaramillo Quemado. En cuanto a la *Ilustración 5*, tenemos la tasa media de muerte por municipio. Se ve que la mayoría de la media de las tasas para los municipios se encuentra entre 5 y 25 por cada 1.000 habitantes.

En *Ilustración 6* vemos un mapa de la variación proporcional de defunciones en Castilla y León entre los años 1998 y 2018. Los municipios que observamos en color blanco son en los cuales no podemos calcular esta variación ya que en el año 1998 no hubo

defunciones. Los municipios que vemos en un color más oscuro corresponden con los que han tenido más defunciones en el año 1998 en comparación con 2018. Según la escala de color va aclarándose, se evidencia que tienen lugar más defunciones en 2018 que en 1998.

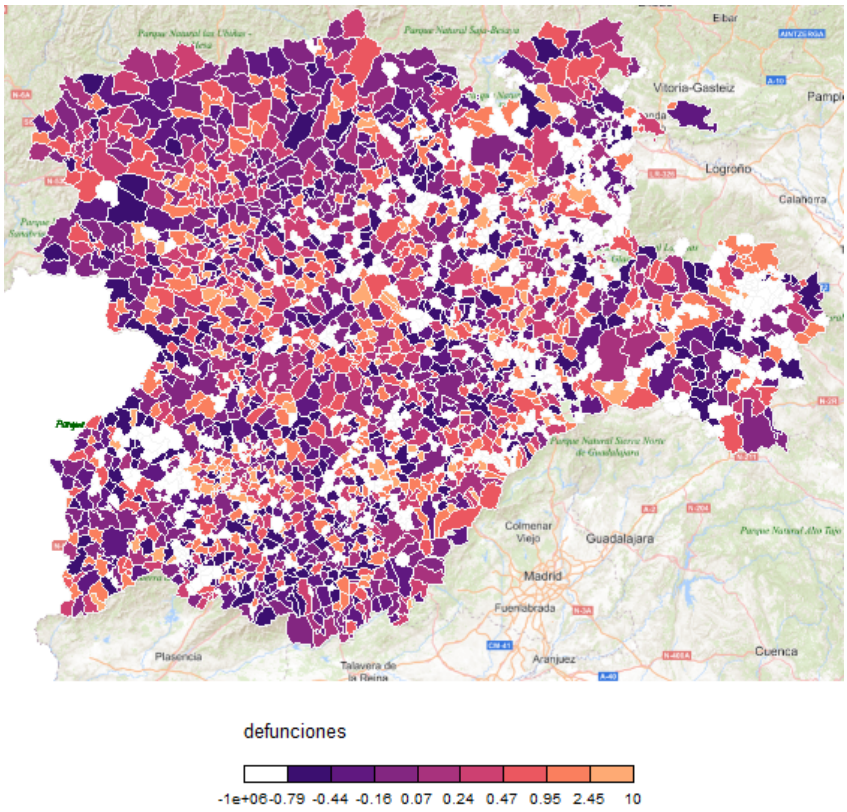

Ilustración 6: Mapa de variación de defunciones en municipios de Castilla y León

Nacimientos

En los nacimientos tenemos el número de nacimientos que se producen en cada municipio por cada año. El número medio por año de nacimientos que existen en este periodo de tiempo es 18.189,17. En este caso tenemos muchos valores que son 0.

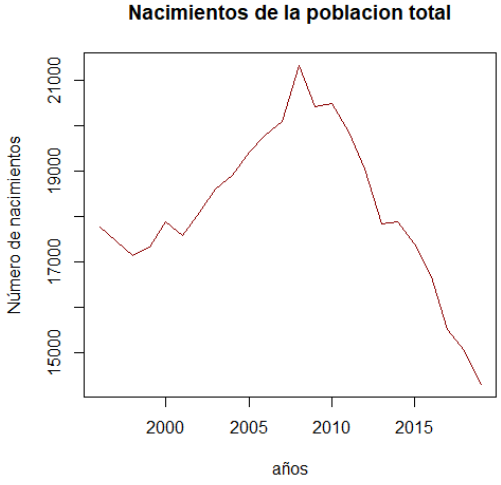

Ilustración 7: Total de nacimiento por año

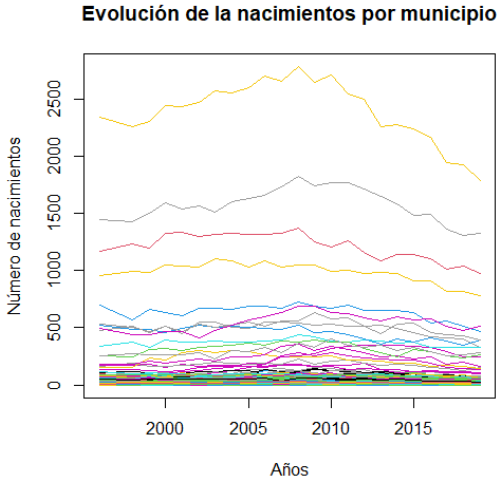

Ilustración 8: Evolución de nacimientos por cada municipio

En las ilustraciones podemos ver cómo evoluciona el total y cada municipio por los años. Podemos observar en *Ilustración 7* que, en el periodo de 2000 a 2006 podemos ver un aumento en el total de nacimientos y, de 2010 en adelante, podemos ver que decrece. En *Ilustración 8* podemos ver estas ideas, pero de un manera menos clara, ya que además, al ser tantos no podemos sacar ninguna idea clara. Dado este problema, calcularemos la tasa de nacimientos.

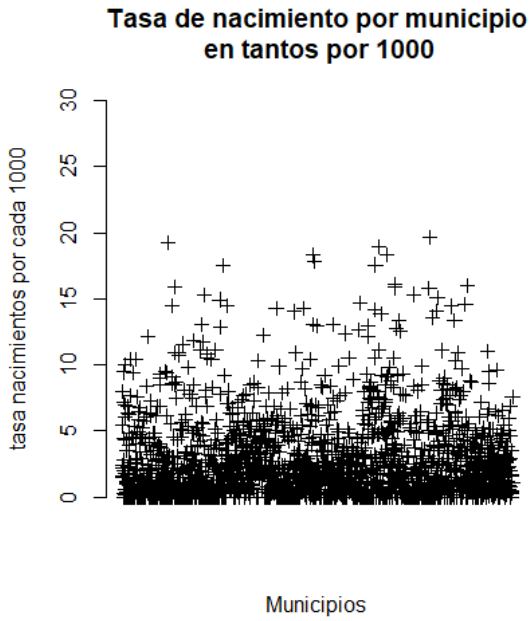

Ilustración 9: Media de la tasa de nacimiento por cada 1000

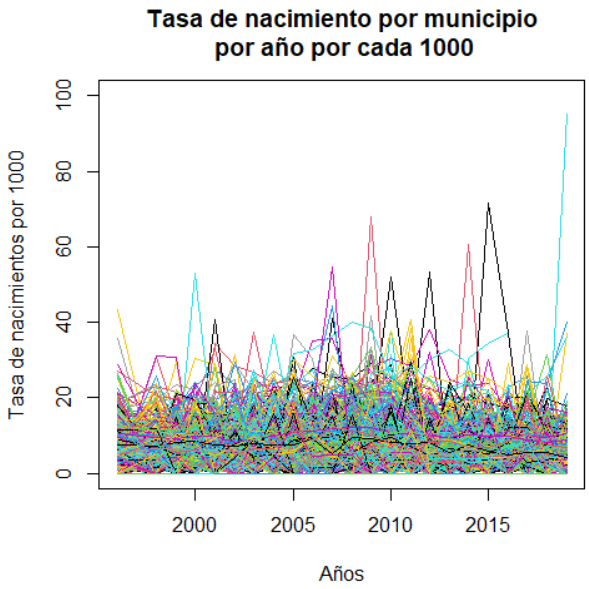

Ilustración 10: Tasa de nacimiento por cada municipio por cada 1000

Villarcayo de Merindad de Castilla la Vieja, código INE 09903 es un municipio que, en ambas ilustraciones (9 y 10), se presenta como un municipio con un valor mucho más elevado que los demás. Para una buena visualización de los demás valores, los ejes se han ajustado a los demás, quedando muy encima de los ejes este municipio.

En la *Ilustración 9* tenemos la media de la tasa de nacimientos para cada municipio. Vemos que las medias de las tasas de nacimientos son muy bajas. En la *Ilustración 10* seguimos con esta idea de tasas de nacimientos bajas, pero podemos observar algunos picos en algunos municipios, sobre todo, de 2010 en adelante.

En la *Ilustración 11* vemos la variación proporcional de los nacimientos entre los años 1998 y 2018. Los municipios que observamos en color blanco son en los cuales no podemos calcular esta variación ya que en el año 1998 no hubo nacimientos. Vemos en colores oscuros los municipios en los que se produjeron más nacimientos en 1998 que en 2018 y, según va aclarándose el color, cambia esta idea al suceso contrario, es decir, más nacimientos en 2018 que en 1998.

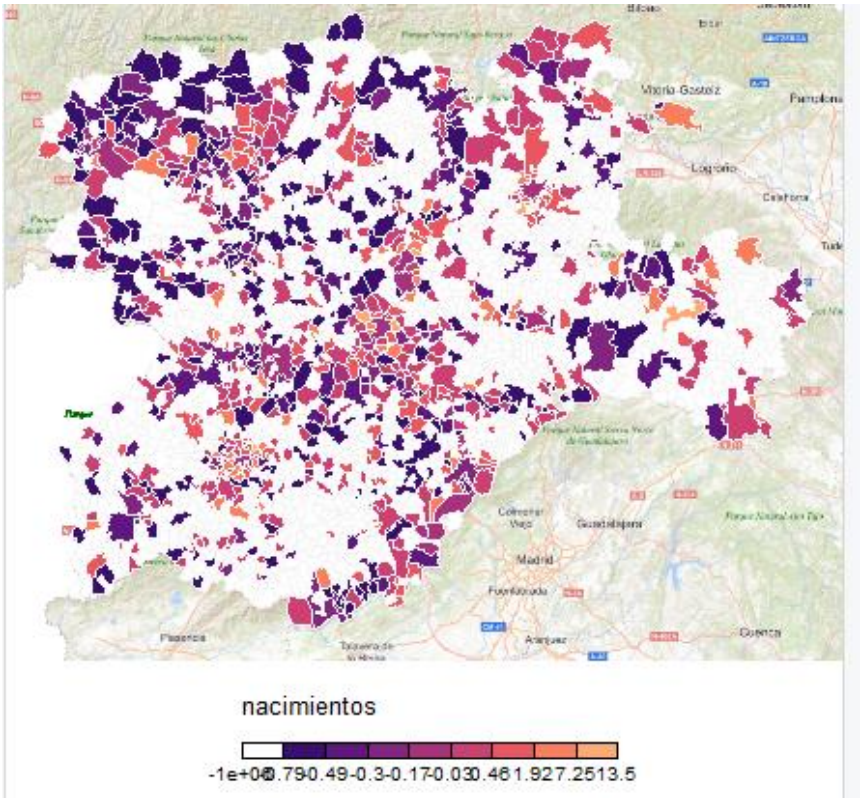

Ilustración 11: Mapa de variación de nacimientos en municipios de Castilla y León

Mujeres en edad reproductiva

Disponemos de las mujeres que se encuentran en edad reproductiva, considerando edad reproductiva desde 15 a 49 años. Tenemos el número de mujeres que se encuentran en esa franja de edad en cada municipio por cada año.

El número medio anual de mujeres en edad reproductiva es 562.408,4. En las mujeres en edad reproductiva tenemos los datos de todos los municipios y, además, casi no existen valores 0. Podemos ver que, en algunos municipios, el número de mujeres en edad reproductiva permanece casi constante varios años.

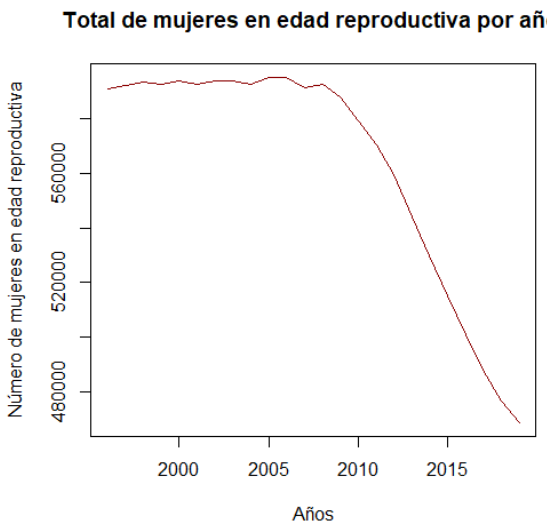

Ilustración 12:Total de mujeres en edad reproductiva por años

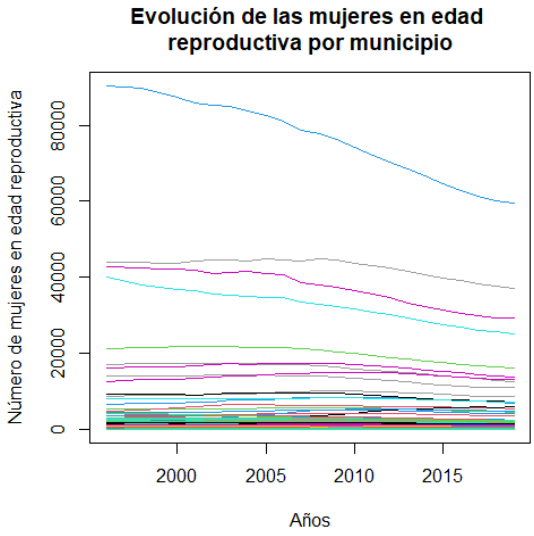

Ilustración 13: Evolución de mujeres en edad reproductiva por municipios por años

Como podemos observar en la *Ilustración 12*, la curva del total de mujeres en edad reproductiva estaba estable hasta 2010, donde podemos ver una bajada muy brusca. Observando la *Ilustración 13* podemos discernir esta misma idea para los valores que vemos. Al encontrarnos con tantos valores solo podemos ver los altos, ya que debajo hay muchas líneas, en las cuales no podemos distinguir claramente como actúa cada municipio. Realizamos una tasa de mujeres en edad reproductiva, para ver si podemos decir algo más sobre estas ideas.

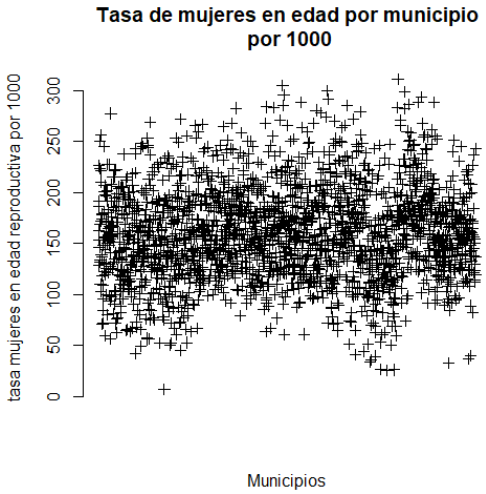

Ilustración 14: Tasa media de mujeres en edad reproductiva por municipio por cada 1000

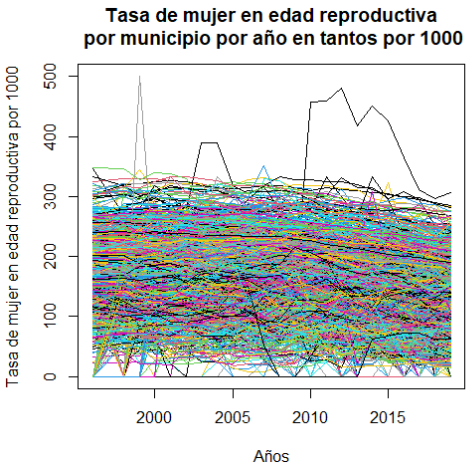

Ilustración 15: Tasa de nacimiento por cada municipio por cada 1000

En la *Ilustración 14* vemos que las tasas medias de mujeres de edad reproductiva se encuentran entre los valores de 0 y 300 por cada 1.000 habitantes, lo cual no son unas tasas muy elevadas. En cuanto a la *Ilustración 15*, no podemos observar ningún dato claro, tan solo vemos el mismo intervalo de 0 a 300, pero observamos unos picos.

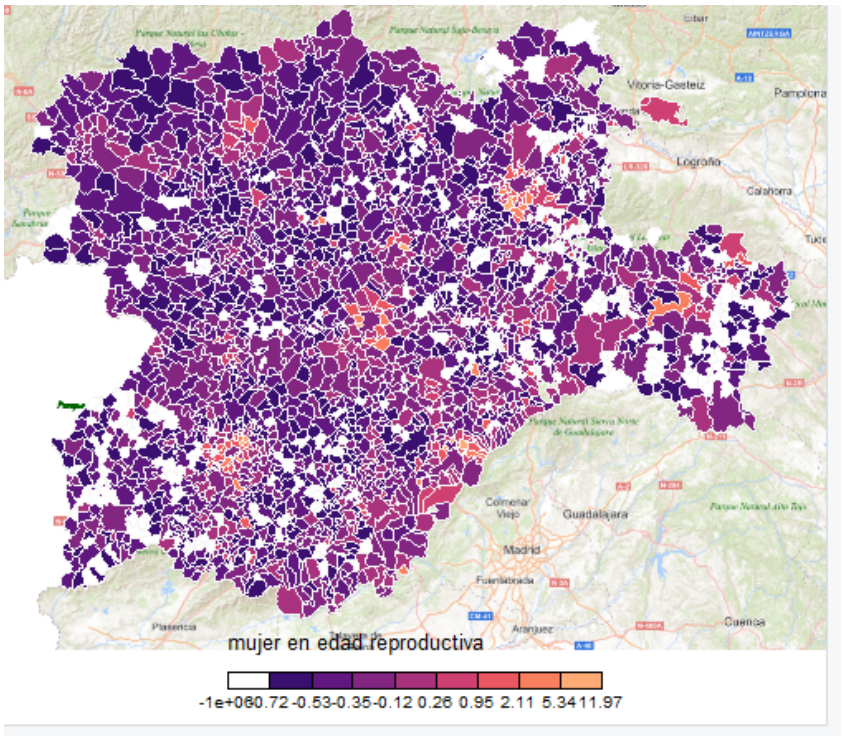

Ilustración 16:Mapa de variación de número de mujeres en edad reproductiva en municipios de Castilla y León

Para crear el mapa *Ilustración 16* calculamos la variación proporcional del número de mujeres en edad reproductiva entre los años 1998 y 2008. Los municipios que observamos en color blanco son en los cuales no podemos calcular esta variación ya que en el año 1998 no había ninguna mujer en edad reproductiva. Los colores más oscuros corresponden a los municipios en los que 1998 había más mujeres en edad reproductiva que en 2018 y, según se aclaran los colores, este suceso se vuelve el contrario.

Población total

Pertenece a las variables estadísticas de población, representa el total de población de derecho que hay en cada municipio en cada año. Sería la suma de población de derecho mujer y población de derecho hombre. Al ser unos datos provenientes del padrón podemos ver que no disponemos del año 1997. El número medio de personas por año total es 2.493.074. En estos datos tenemos todos los valores, no disponemos de valores perdidos. En algunos casos, el número de individuos en un municipio varia poco durante un número de años.

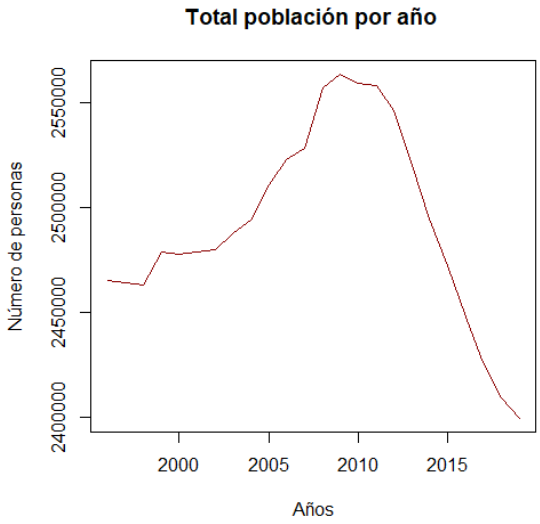

Ilustración 17: Total de la población por años

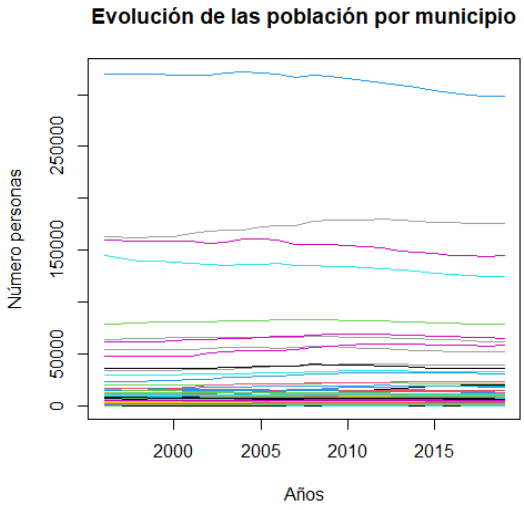

Ilustración 18: Evolución del total de la población por municipio

En las ilustraciones podemos observar que decrece el total de la población desde el año 2010. En *Ilustración 17* vemos que, hasta 2010, la población total crecía, pero, desde 2010, donde alcanzo un "pico", podemos ver un rápido decrecimiento. En *Ilustración 18* observamos que la mayoría de los municipios perdieron población, pero las diferencias no son apreciables dada la cantidad de municipios.

El mapa de la *Ilustración 19* representa la variación que hemos calculado de la variación proporcional del total de la población en los años 1998 y 2008. Los colores más claros representan una población era mayor en 2018 que en 1998 y, según se oscurecen los colores cambia esta idea a la contraria. Podemos ver que la mayoría de los municipios tiene un color oscuro, es decir, que tenían más población en 1998 que en 2018, lo que indica una disminución de la población.

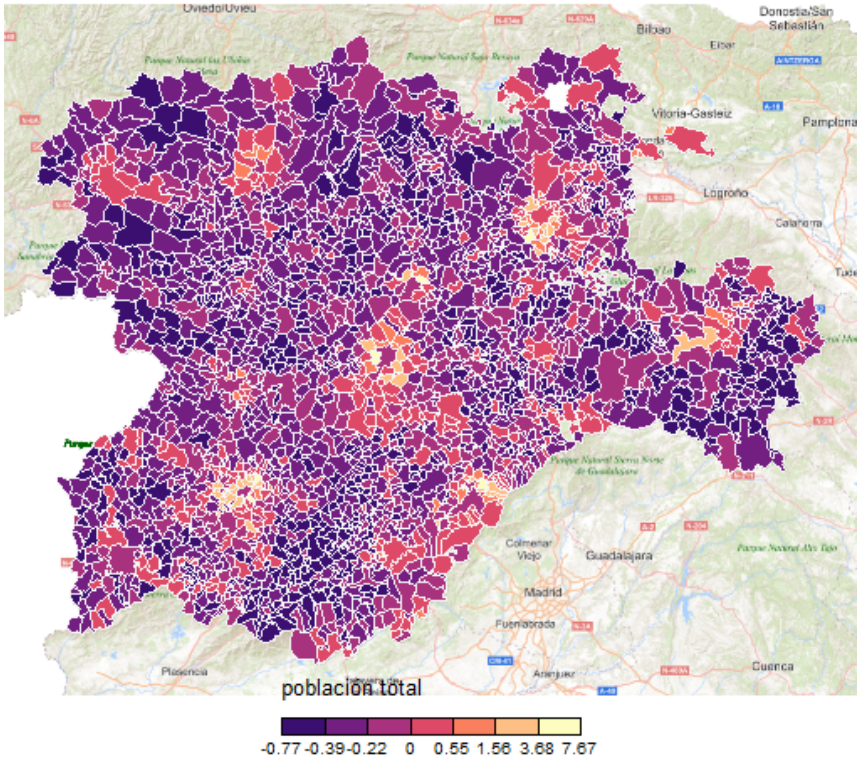

Ilustración 19: Mapa de variación la población total en municipios de Castilla y León

Población de derecho varón

En la población de derecho varón tenemos los datos del número de varones empadronados en cada municipio. [6]

La población de derecho varón media por años es 1.233.240. El número de valores 0 no es elevado, dado que en muchos municipios existen datos y algún varón.

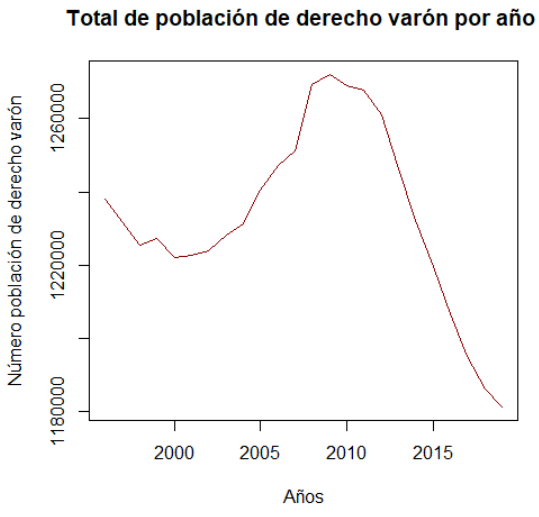

Ilustración 20: Total de población de derecho varón por año

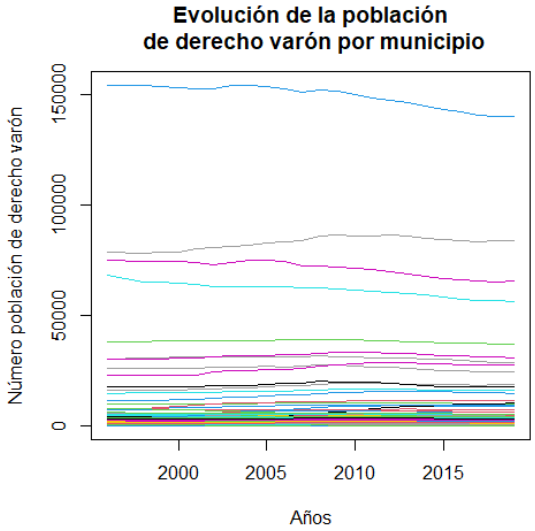

Ilustración 21: Evolución del total de la población de derecho varón por municipio

Podemos ver en *Ilustración 20* un crecimiento hasta el año 2010 y, después un decrecimiento, de la misma manera que se veía en la población total (*Ilustración 17*). En el gráfico de la Ilustración 21 al ser muchos municipios no podemos ver claramente estas ideas. Calculamos la tasa de población de derecho varón.

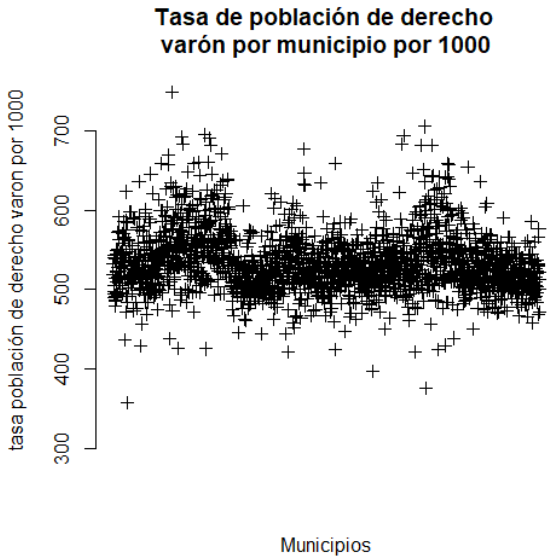

Ilustración 22: Tasa media de población de derecho varón por municipio por cada 1000

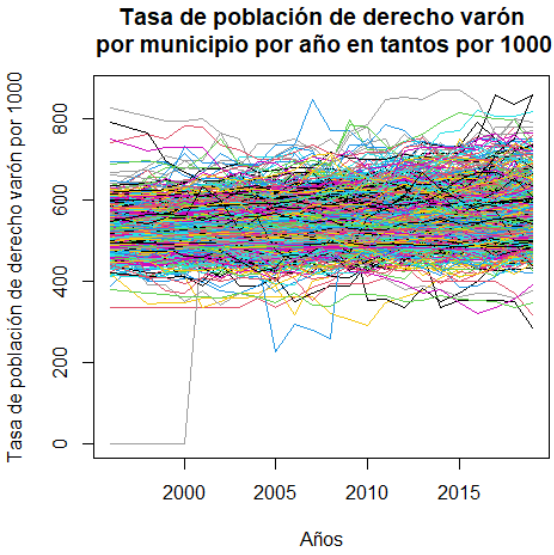

Ilustración 23: Tasa de población de derecho varón por cada municipio por cada 1000

En *Ilustración 22* vemos que la tasa de población de derecho varón se encuentra entre 300 y 700 individuos por cada 1.000 habitantes. 700 es una tasa elevada, por lo que en muchos municipios la tasa media indica que son casi todos varones. Con la *Ilustración 23* podemos ver que se confirman estas ideas, pues la variación se ha quedado en estas franjas. El municipio que vemos con valor 0 varios periodos es San Cristóbal de Segovia ya que es un municipio que se creó nuevo en el año 2001 como segregación del municipio de Palanzuelos de Eresma. Este es el primer grafico en que se observa ya que en otras variables no se puede observar por tener valores cercanos a 0.

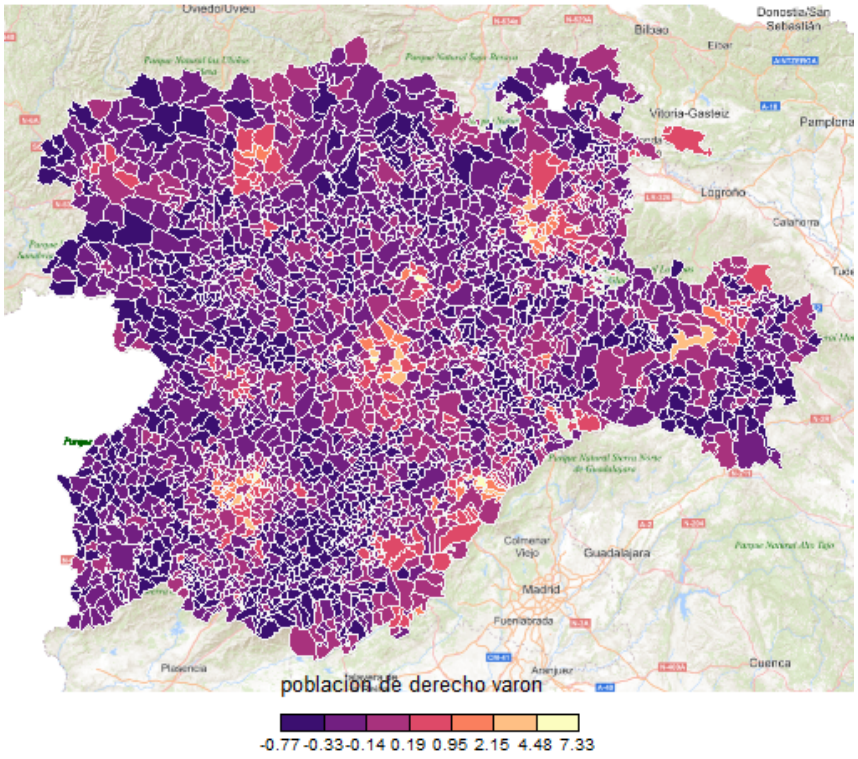

Ilustración 24: Mapa de variación la población de derecho varón en municipios de Castilla y León

En el mapa de la *Ilustración 24* vemos que la mayoría de los municipios han perdido población, ya que gran parte de los municipios están representados en color oscuro, a excepción de los alrededores de las capitales. En la *Ilustración 24* tenemos calculada la variación proporcional de derecho varón entre los años 1998 y 2018. Los colores más claros son los que tiene más población en 2018 que en 1998, según se oscurecen los colores se cambia esta idea a la contraria.

Población de derecho mujer

En la población de derecho mujer tenemos los datos del número de mujeres empadronadas en cada municipio. [6]

La media anual de la población de derecho mujer es 1.263.177. De la población de derecho mujer tenemos valor para todos los años y todos los municipios. En muchos municipios podemos observar que, durante varios años, se tiene el mismo valor, es decir, en muchos el número de población de derecho mujer que había no varía.

Podemos ver que en el de total vemos un ascenso hasta el año 2010 y, después de 2010 un decrecimiento.

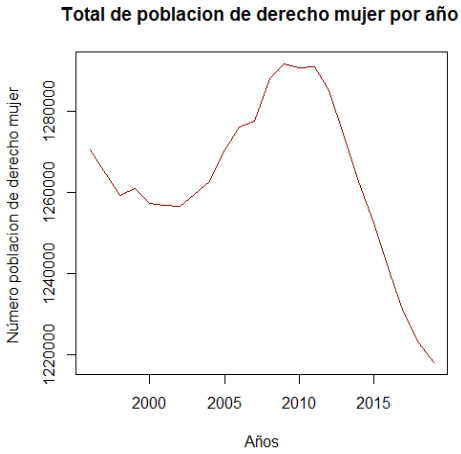

Ilustración 25: Total de población de derecho mujer por año

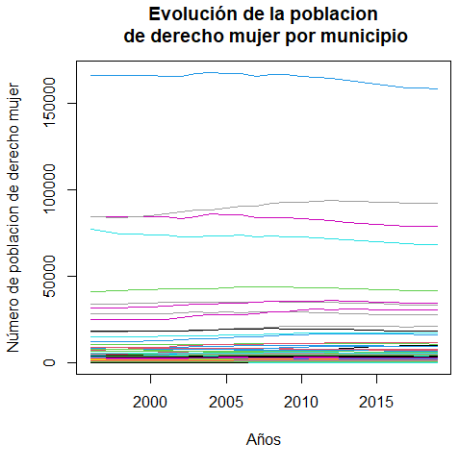

Ilustración 26: Evolución del total de la población de derecho mujer por municipio

Se observa en *Ilustración 25* un decrecimiento hasta el año 2000, seguidamente de un crecimiento hasta el año 2010 y, desde 2010 vemos un decrecimiento mucho más pronunciado. En la *Ilustración 26* no podemos ver muchas cosas, ya que tenemos muchos municipios. Lo que podemos ver es que se mantuvo, más o menos, estable la población. Realizamos la tasa de población de derecho mujer.

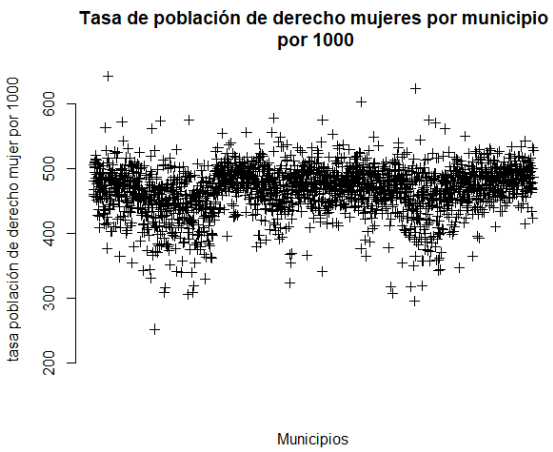

Ilustración 27: Tasa media de población de derecho mujer por municipio por cada 1000

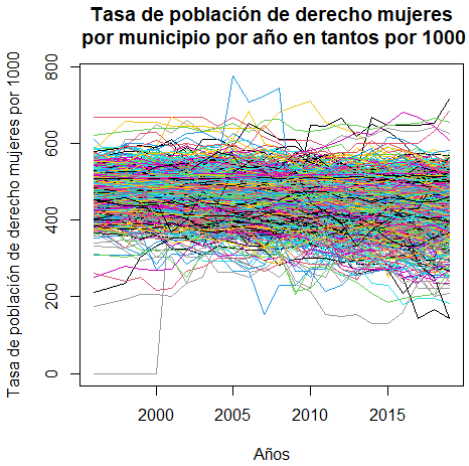

Ilustración 28: Tasa de población de derecho mujer por cada municipio por cada 1000

En la *Ilustración 27* vemos que las tasas medias de población de derecho mujer por municipio se encuentran entre 200 y 600 por cada 1.000 individuos. En la *Ilustración 28* vemos que la tasa anual por municipio se encuentra en dichos intervalos para la mayoría, pero no podemos saber, exactamente, por dónde es cada uno, ya que tenemos muchas líneas y no distinguirlo adecuadamente, tan solo vemos picos por encima y por debajo de donde está la gran masa de las líneas. De la misma manera que ocurría en la población de derecho varón en la *Ilustración 28* podemos ver que existe una línea en 0, esto corresponde con el municipio de San Cristóbal de Segovia.

En la *Ilustración 29* tenemos el mapa de variación proporcional de la población de derecho mujer entre los años 1998 y 2018. La mayoría de los municipios tienen un tono oscuro, lo que nos indica que la población de derecho mujer era mayor en 1998 que en 2018, según se aclara la tonalidad, esta idea cambia hasta la posición contraria de que la población en 1998 era menor que en 2018.

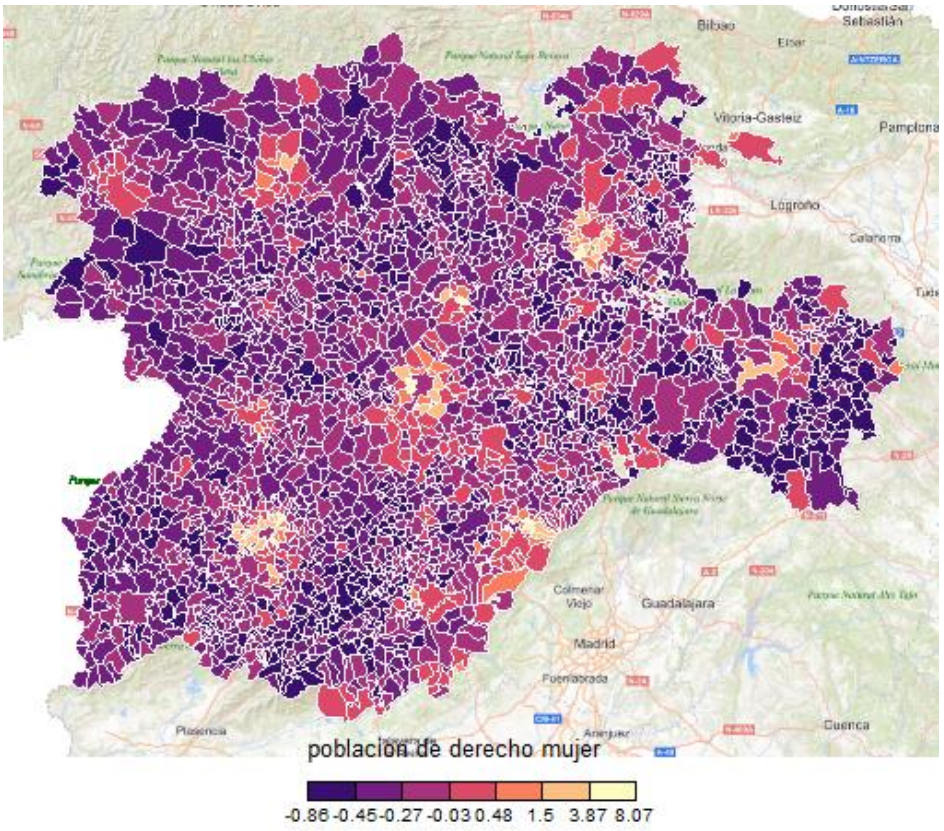

Ilustración 29: Mapa de variación la población de derecho mujer en municipios de Castilla y León

2.2 VARIABLES DE MIGRACIÓN.

Denominamos variables de migración a las variables que indican el movimiento de la población en términos de residencia temporal o definitiva, dentro de este grupo podemos dividirlas en 2 subgrupos :

- Emigraciones
  - Emigraciones dentro de la misma provincia.
  - Emigraciones con destino otra provincia de Castilla y León.
  - Emigraciones a otra comunidad autónoma.
  - Emigraciones con destino otro país.
- Inmigraciones
  - Inmigraciones dentro de la misma provincia.
  - Inmigraciones procedentes de otra provincia de Castilla y León.
  - Inmigraciones de otra comunidad autónoma.
  - Inmigraciones procedentes de otro país.

Todas estas variables proceden del Padrón, por lo que el año 1997 no está disponible. Todos los DataFrame tienen las n-1 primeras columnas como los años, en la última columna el código INE y, como filas tienen los distintos municipios. Para estos datos hay municipios para los que no disponemos de los datos en algunas variables y, además, tenemos muchos valores que son desconocidos o son 0. Dada la naturaleza de estos datos y, para mayor comodidad, se juntaran en 2 DataFrame todas estas variables: uno de emigraciones y otro de inmigraciones.

Inmigraciones

Las inmigraciones son la entrada de personas en la comunidad para establecer su residencia en la zona considerada de Castilla y León de manera temporal o permanente.[7]

Para la creación de este fichero, realizamos la suma de las distintas variables de inmigración en el mismo municipio y para el mismo periodo de tiempo. Las variables de inmigración son las siguientes: "Inmigraciones dentro de la misma provincia", "Inmigraciones procedentes de otra provincia de Castilla y León", "Inmigraciones de otra comunidad autónoma" e "Inmigraciones procedentes de otro país". En este fichero existen muchos valores 0, ya que en muchos municipios en ciertos periodos no se produjo ninguna inmigración. La media por año del número total de inmigraciones por año es 86.999,74.

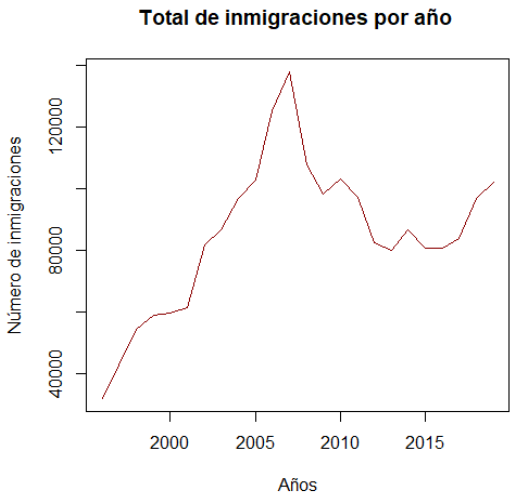

Ilustración 30: Total de inmigraciones por año

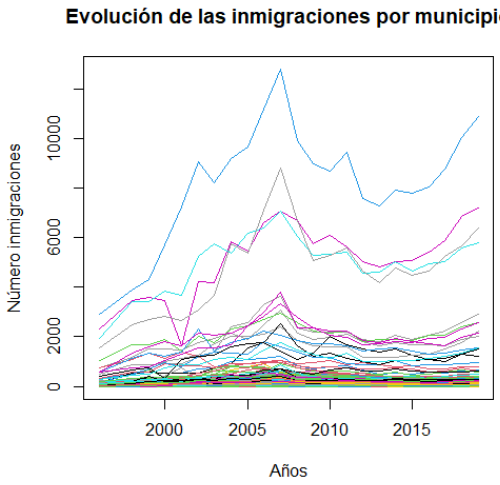

Ilustración 31: Evolución de la inmigración por municipio

En la *Ilustración 30* podemos ver que el total de las inmigraciones crecía hasta, aproximadamente, 2006; desde 2006 ha ido decreciendo, a excepción de algún momento, hasta 2015, donde vemos otro repunte de las inmigraciones. En cuanto a la evolución de las inmigraciones por municipios, en la *Ilustración 31* podemos ver que el mismo pico se produce, aproximadamente, en 2006 y, luego un comportamiento simular al total, aunque, aquí, no podemos verlo claro dado que al tener una línea por municipio se juntan las líneas. Para una mejor visualización de los datos, realizaremos una tasa de inmigración.

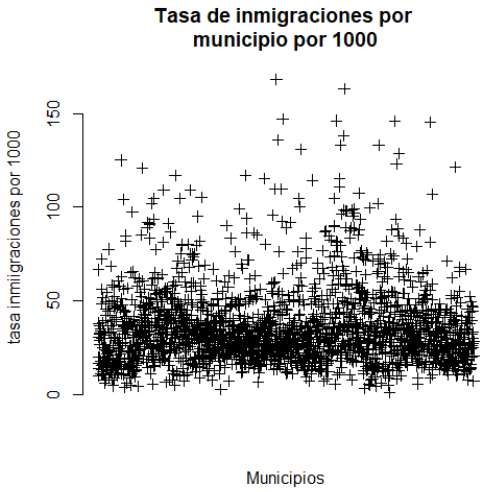

Ilustración 32: Tasa media de inmigración por municipio por cada 1000

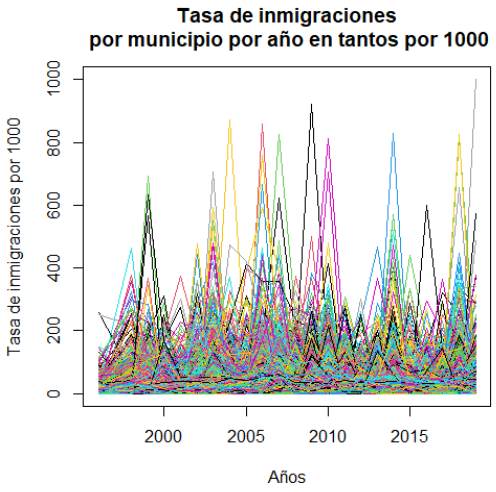

Ilustración 33: Evolución de tasa de inmigración por cada municipio por cada 1000

En la *Ilustración 32* podemos ver que las tasas medias de inmigración por municipio se encuentran entre 0 y 150 por cada 1.000 habitantes. En la *Ilustración 33* tenemos la evolución de la tasa de inmigración por cada municipio, donde podemos ver unos picos, que se producen cada 5 años, donde la inmigración aumenta.

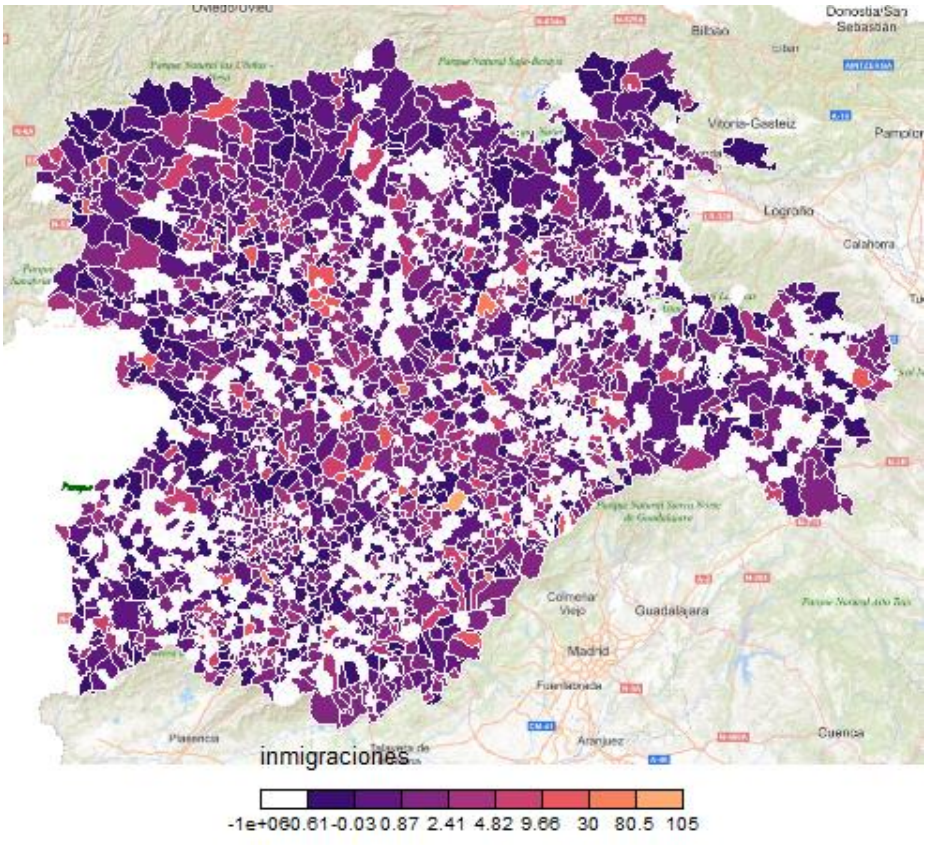

Ilustración 34: Mapa de variación inmigración en municipios de Castilla y León

En mapa de la *Ilustración 34* podemos ver la variación proporcional de la inmigración en 20 años, entre los años 1998 y 2018. Los municipios que observamos en color blanco son en los cuales no podemos calcular esta variación ya que en el año 1998 no hubo

ninguna inmigración en el municipio. En los colores claros tenemos los municipios que tenían más inmigración en 2018 que en 1998, frente a los colores oscuros, que sucede la idea contraria, más inmigración en 1998 que en 2018.

Emigraciones

Las emigraciones son la salida de personas una zona de la comunidad de Castilla y León para establecer su residencia fuera de esa misma zona de la comunidad de manera temporal o permanente. [8]

Para la creación de fichero emigraciones, lo que hacemos es la suma para cada periodo y cada municipio de las variables "Emigraciones dentro de la misma provincia", "Emigraciones con destino otra provincia de Castilla y León", "Emigraciones a otra comunidad autónoma" y "Emigraciones con destino otro país". En las emigraciones, tenemos muchos municipios con el valor 0 en algunos periodos, ya que no se produjo ninguna inmigración. La media total de emigraciones por año es 83.961,22.

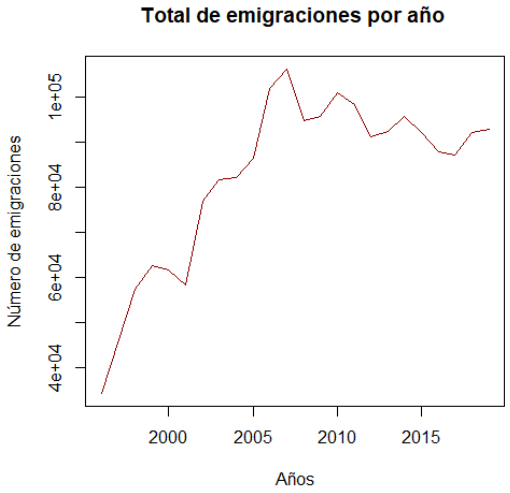

Ilustración 35: Total de emigraciones por año

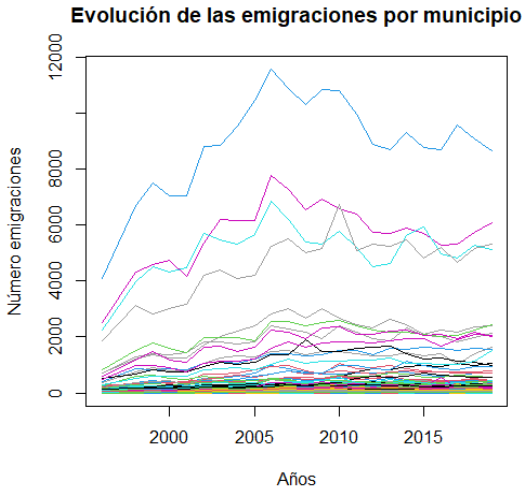

Ilustración 36: Evolución de la emigración por municipio

Podemos ver un aumento en *Ilustración 35* bastante pronunciado, hasta el año 2005, del total de emigraciones. Tras 2005 vemos una fase de picos y bajadas. En la *Ilustración 36* vemos la misma idea que en la *Ilustración 35*, pero no de forma tan clara, ya que al ser la evolución de cada municipio tenemos muchas líneas y, además, unos tienen muchas inmigraciones y otras casi ninguna, dada la naturaleza de los municipios. Realizaremos la tasa de emigración.

En *Ilustración 37* tenemos la tasa de emigración media por cada municipio, donde vemos que se encuentra entre 10 y 120 por cada 1000, aproximadamente. En la *Ilustración 38* tenemos picos que se producen cada 5 años, en especial vemos un gran pico en 2012, aproximadamente. Esto, nos dice que algunos municipios han tenido mucha emigración en esos años concretos.

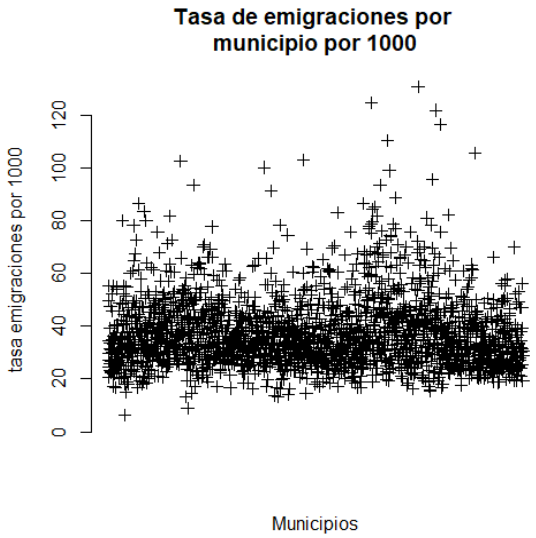

Ilustración 37: : Tasa media de emigración por municipio por cada 1000

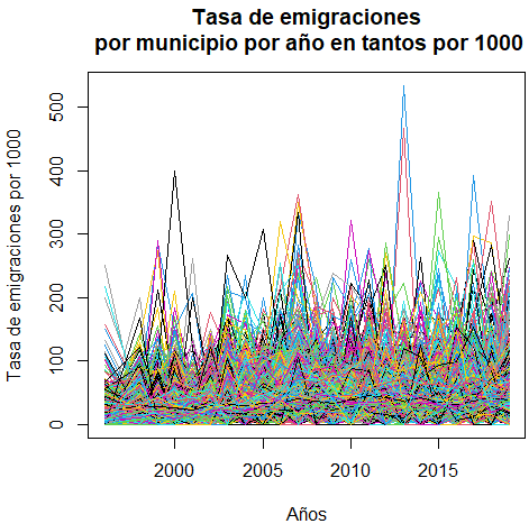

Ilustración 38: Evolución de tasa de inmigración por cada municipio por cada 1000

En mapa de Castilla y León de la *Ilustración 39* tenemos la variación proporcional de las inmigraciones entre los años 1998 y 2018. Los municipios que observamos en color blanco son en los cuales no podemos calcular esta variación ya que en el año 1998 no hubo ninguna emigración en el municipio. Los colores claros representan que, en 2018, tenía más emigración y los oscuros que tenía más emigración en 1998. Podemos ver que predominan más los colores oscuros que los claritos, indicando que había más emigración en 1998 que en 2018.

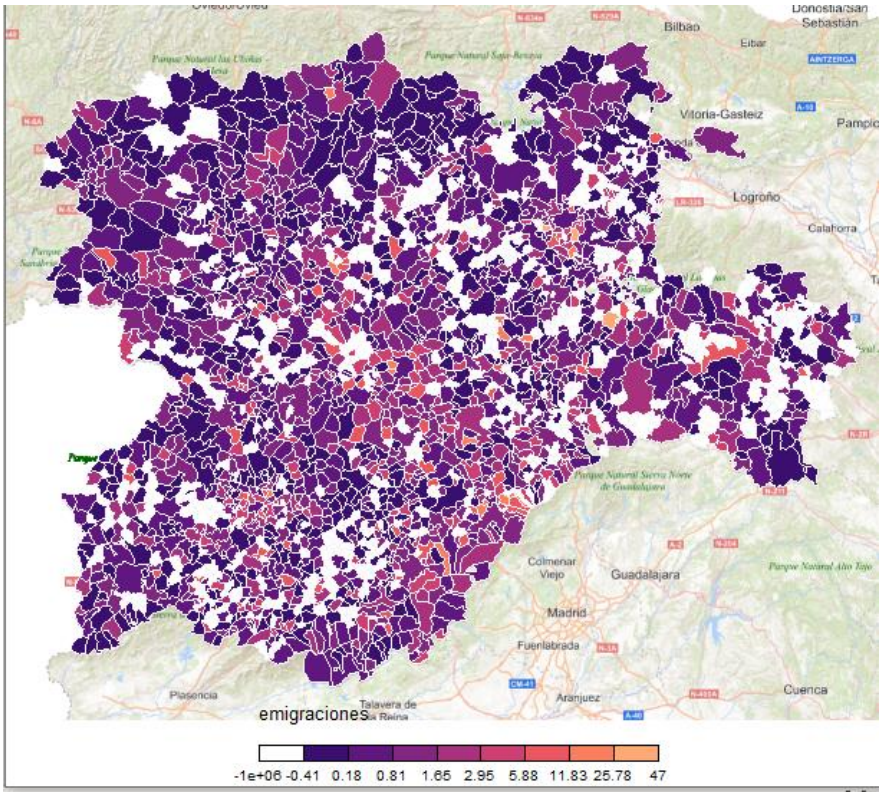

Ilustración 39: Mapa de variación emigración en municipios de Castilla y León

3 MÉTODOS ESTADÍSTICOS MULTIVARIANTES

3.1 ANÁLISIS DE COMPONENTES PRINCIPALES

El Análisis de Componentes Principales (APC) es una técnica estadística multivariante, en la cual se sintetiza la información, es decir, simplifica la complejidad de los espacios muestrales en cuales la dimensionalidad es muy elevada. Busca que la información de esta manera pueda ser interpretada, así como la búsqueda de la reducción de la perdida de información. Por esto, las ACP son un método de simplificación de la información.

En la muestra tenemos *n* individuos y *p* variables que deben estar correladas, ya que si no están correladas este análisis no tiene sentido. Partiendo de este conjunto de variables, en dimensión *p,* lo que queremos es tener un conjunto de variables en dimensión m siendo $m < p$. Estas variables en dimensión m deben ser incorreladas y, sus varianzas, deben decrecer en una progresión, de tal manera que pueden ordenarse según la información que posean. Estas nuevas variables creadas caracterizan al individuo de la misma manera que lo hacían las p variables originales.

Para el cálculo de las componentes principales es necesario una tipificación de los datos, dado que deben estar en la misma escala, así como una centralización de las variables para tener media 0. Cada componente y_i *(i= 1,2,..m)* es una combinación lineal de las x_1, x_2 ,… x_p variables originales. La primera componente es una combinación normalizada de estas variables para el primer individuo, esto es, $\sum_{k=1}^{p} \propto_{kj}^{2} = 1$

La primera componente es de la forma $y_1=\alpha_{11}x_1 + \alpha_{21}x_2+ \ldots+ \alpha_{p1}x_p$. Los α pueden interpretarse como el peso que tiene cada variable en cada componente principal. Para calcular los α se trata de encontrar los autovalores y los autovectores de la matriz de covarianzas para, de esta manera, maximizar la varianza. De la misma manera que se calcula la primera componente se calculan las demás. Debemos tener en cuenta que las componentes tienen que ser incorreladas entre sí, o lo que es lo mismo, que la combinación lineal no puede estar correlada.

Para decir cuanta información proporciona una componente, se utiliza la medida de la varianza explicada por esa componente. Cuanta mayor sea la varianza más información posee esa componente. Como los datos los tenemos tipificados y centrados, podemos ver que:

$\sum_{k=1}^{p} var(x_k) = \sum_{k=1}^{p} \frac{1}{n} x_{ik}^2$	$\frac{1}{n}\sum_{i=1}^{n} y_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{p} \alpha_{km} x_{ik})^2$	$\frac{\sum_{i=1}^{n}(\sum_{k=1}^{p} \alpha_{km} x_{ik})^2}{\sum_{i=1}^{n}(\sum_{k=1}^{p} \alpha_{km} x_{ik})^2}$
Varianza total	Varianza explicada componente m	Proporción de varianza explicada componente m

Como queremos reducir el número de variables para, así, reducir la dimensionalidad, escogemos el número de componentes principales, usamos un número que explique la variabilidad. Una manera común de escoger el número de componentes es usando la proporción de varianza explicada acumulada y, seleccionamos el mínimo número de componentes con el cual queda explicada la variabilidad. Encontramos un punto, en el cual incrementar el número de componentes no produce una mejora en la explicación

de la varianza, por lo que nos quedaremos con ese punto como el número de componentes. [9,10,11]

3.2 MULTIDIMENSIONAL SCALING

El multidimensional scaling o escalado multidimensional es una técnica de análisis multivariante, en la cual intentamos representar en un número reducido de dimensiones las disimilaridades que tiene los objetos. Para esto, partimos de una matriz de disimilaridades que, en muchas ocasiones, es una matriz de distancias e, intentamos que la nueva representación respete y represente las disimilaridades originales. Lo más común es bajar la dimensión a 2 dimensiones, de manera que se pueda representar estas disimilaridades en un gráfico.

La semejanza que tiene con análisis de componentes principales es que, si usamos multidimensional scaling con la distancia euclídea de los elementos, estamos realizando el mismo análisis componentes principales.

El procedimiento parte de una matriz *D* de disimilaridades o distancias, esta matriz es cuadrada *n x n*, siendo *n* el número de individuos u objetos. Los elementos de esta matriz *D* son de la manera δ_{ij}, siendo *i=1,2,..n, j =1,2,.. n* además, cumplen unas propiedades.

1. La diagonal es de 0, es decir, que la disimilaridad de un elemento con el mismo es 0, o lo que es lo mismo, $\delta_{ii}=0$.
2. La disimilaridad de i a j debe ser la misma que de j a i, $\delta_{ij}= \delta_{ji}$.
3. Las disimilaridades deben ser mayores o iguales que 0, $\delta_{ij} \geq 0$.

Pretendemos representar en una matriz *n x k*, siendo *k* el nuevo número de dimensiones y $k \leq n$. Cuando tenemos *un k >2, lo* que hacemos es ordenarlas por importancia y hacer un gráfico, especialmente cuando es 2 o 3. Tenemos varios métodos de scaling: métrico y no métrico.

Para valorar la calidad de la representación usamos la medida de STRESS, que se calcula de la manera:

$$STRESS = (\sum_{i<j} w_{ij} (d_{ij} - \widehat{d_{ij}})^2)^{1/2}$$

siendo w_{ij} los pesos. Los pesos más utilizados son raw stress, stress-1 y stress-2.

En cuanto a la evaluacion de este stress lo mejor es que tenga un valor lo más proximo a 0 posible. Como convenio se suele ulitizar esta escala: menor que 0,025 es excelente, entre 0,025 y 0,05 es bueno,entre 0,05 y 0,1 regular y más de 0,2 es pobre.

3.2.1 Scaling métrico

Tenemos que $d_{ij} \approx f(\delta_{ij})$. Para el scaling clásico se toma $f(x)=x$ y, para el scaling métrico $f(x) = a + bx$, donde $a, b \geq 0$. Observando las fórmulas para $f(x)$ podríamos introducir el scaling clásico dentro de los métricos, siendo $a=0$ y $b=1$.

El scaling métrico preserva la distancia original de los objetos o una representación de ellas. En el scaling métrico usamos unas variantes de los mínimos cuadrados, los mínimos cuadrados con pesos o weighted least squares (wls).

3.2.2 Scaling no métrico

Tenemos que $d_{ij} \approx f(\delta_{ij})$ tomamos $f(x)$ como una función monótona creciente. No se conserva la relación de las distancias originales, sino que el orden entre los deltas se mantiene. No tenemos por qué tener distancias, sino que podemos tener simplemente disimilaridades, es un scaling más relajado. [13,14]

3.3 ANÁLISIS CLÚSTER

El análisis clúster es una técnica de análisis multivariante, en la cual creamos grupos de individuos con características similares, es decir, creamos categorías y categorizamos a los individuos. En estos grupos, los individuos son homogéneos entre sí y distintos a los individuos que conforman otros grupos. Además, los individuos únicamente pertenecen a un único grupo. El análisis clúster es un aprendizaje no supervisado. [3]

En un análisis clúster, lo primero que debemos hacer es seleccionar la muestra sobre la cual vamos a realizar el análisis, así como la selección de las variables que vamos a usar. Una utilización excesiva de variables puede desvirtuar el análisis, de la misma manera que un número insuficiente de variables lo haría también desvirtuado. Una vez tengamos el número de variables y la muestra debemos decidir si transformamos las variables o las usamos de la manera en la que se encuentran, una manera de transformación podría ser una tipificación para que estén centradas, teniendo así media 0 y normalizadas para tener varianza 1. El realizar este tipo de cambios a las variables puede hacer que las diferencias entre grupos no estén visibles de cara al análisis. Este tipo de cambios es adecuado para algunos tipos de variables, como las que están medidas en distintas escalas. [15]

Existen distintos maneras de realizar un clúster, según sea jerárquico, no jerárquico o, una mezcla de ambos.

3.3.1 Clúster no jerárquico (partitioning clustering)

Necesitamos que el número de clúster este fijado de antes para realizarlo. Algunos de los métodos que son no jerárquicos son k-medias, k-medioides, Block-Clustering o Análisis factorial de tipo Q. De los métodos no jerárquicos, el que más destaca por su utilización es las k-medias. [16]

El método de las k-medias es un método que está basado en particiones y, estos grupos que crea son grupos no solapados, es decir, crea grupos, clasifica los individuos y los individuos, únicamente, pertenecen a un único grupo. Los grupos no tienen ninguna estructura interna. La técnica de las k-medias está basada en los centroides de los grupos. El método de las k-medias trata de maximizar las diferencias entre unos grupos y otros, a la vez que intenta minimizar las diferencias entre los individuos asignados a un mismo grupo. [17]

En las k-medias para el cálculo de las disimilitudes se usa, en muchas ocasiones, la distancia de Minkowski y después se tipifican los datos. Pero se pueden usar otras distancias como son la distancia promedio o la similitud del coseno. [17]

En el funcionamiento de las k-medias lo más importante es que la elección de los centros es al azar dentro de los del grupo. Podemos tener 2 ideas a la hora de escoger esos k elementos. La primera sería escoger k elementos a azar y situados como centroides y, la segunda sería crear unos puntos como centroides de manera artificial. Debemos seguir los siguientes pasos para realizar el método:

1. Elegir el número *k* de particiones, con lo que inicializamos *k.*
2. Calculamos la distancia de cada punto a los centroides.
3. Asignamos a cada observación al grupo más cercano, es decir, miramos el centroide más cercano y lo asignamos a ese.
4. Calculamos la media de cada grupo o partición y lo asignamos como nuevo centroide del grupo.
5. Repetir pasos 2 al 4 hasta que los centroides no se muevan, es decir, si el método converge o, si se puso un límite en el número de iteraciones.

Como las k-medias son un método heurístico, no podemos asegurar que convergen a la solución óptima global o distribución óptima, solo podemos decir que lo hacen a una solución. El método depende mucho de los **n** iniciales, por eso es importante su elección. Para asegurar que converge a una solución que no sea un óptimo local, es común que se realice k-medias desde distintos puntos de partida.

3.3.2 Clúster jerárquico (hierarchical clustering)

Los análisis jerárquico no necesitamos decidir el número de grupos que queremos, ya que el método escoge el número de clúster de manera automática. Los clúster se crean usando una jerarquía. En la manera de elegir los grupos podríamos dividirlo en dos tipos:

- Métodos divisivos: los individuos se encuentran todos en el mismo clúster y los vamos dividiendo en varios clúster, según sus diferencias, hasta llegar a un único clúster por individuo. Esta técnica recibe el nombre de DIANA. Es un buen método para buscar y encontrar clúster grandes.
- Métodos aglomerativos: se conoce como AGNES o aglomeración de anidación. Inicialmente, cada individuo tiene su propio clúster y se van juntando, según sus semejanzas. Finalmente, acabamos con único grupo con todos los individuos. Es un buen método para encontrar clúster pequeños. Estos son los métodos que más se utilizan. [18]

El algoritmo Hierarchical o HAC es el más popular y que más se utiliza de los métodos jerárquicos, este método consta de los siguientes pasos:

1. Si tenemos *M* individuos partimos con *M* clúster.
2. Construimos una matriz de distancias entre los distintos grupos o clústeres que se tengan actualmente.
3. Encontramos los clúster más cercanos a cada uno y los juntamos, creando así menos clústeres.
4. Repetimos los pasos 2 y 3 hasta que tengamos solo un clúster de tamaño *M.* [18]

Para las distancias entre grupos tenemos varios tipos métodos de cálculo de las disimilitudes:

- Complete linkage o enlace completo. Es la vinculación completa entre clústeres similares. Es la distancia más larga entre 2 los individuos de grupos.

  $L(i,j) = max(d(xi,x_j))$, siendo i y j los distintos grupos o clústeres.

- Single linkage o enlace simple. la distancia más corta entre 2 individuos de distintos grupos.

  $L(i,j) = min\ (d(xi,x_j))$, siendo i y j los distintos grupos o clústeres.

- Average linkage. Es la distancia media de un punto a los demás de otro grupo.

- Centroid method. Usa los centroides de los grupos para encontrar los más cercanos. El centroide de un grupo es la media de los que son de un mismo grupo.

- Ward´s method. El método de Ward minimiza la perdida de asociación en cada grupo, es decir, minimiza la varianza intragrupo.

La principal ventaja de usar métodos jerárquicos, en lugar de los no jerárquicos, es el no tener que definir el número de clústeres al inicio. Otras ventajas, serían que los dendrogramas (representación de los clúster en forma de árboles) son más fáciles de interpretar, aunque si tenemos muchos datos puede ser complicado ver los grupos. Estos métodos son más fáciles de implementar, pero son mucho más lentos que las k-medias, en cuanto al tiempo de cálculo. Una desventaja que tenemos es que los pasos que vamos dando son sin retorno, si juntamos 2 clúster quedan unidos, sea la unión correcta o incorrecta. Pero una ventaja sería que los clúster que hace siempre son los mismos a diferencia de las k-medias. [19]

4 APLICACIÓN A LA CLASIFICACIÓN DE LOS MUNICIPIOS DE CASTILLA Y LEÓN.

Como se indicó en el capítulo 2, dada la naturaleza de los datos es conveniente realizar los análisis con las tasas en lugar de los datos originales.

Como hemos visto en los análisis descriptivos es conveniente dividir de los municipios en 2 grupos. Una división administrativa muy común es separar los municipios de más de 20.000 habitantes y de los menos de 20.000. Se ha realizado esta agrupación según la población en el año 2019.

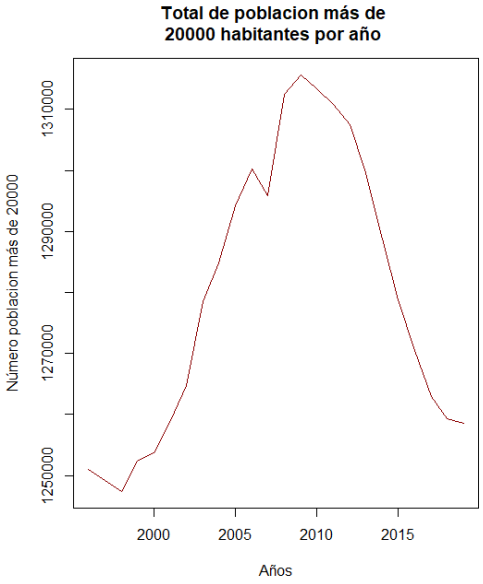

Ilustración 40:Total por año de población grupo de más de 20000 habitantes

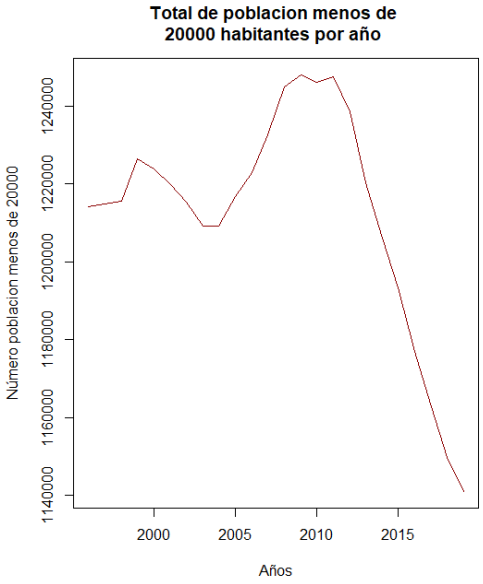

Ilustración 41: Total por año de población grupo de menos de 20000 habitantes

A la vista de las *Ilustraciones 40 y 41*, podemos observar que en ambos casos la población total alcanzo su máximo en 2010, pero desde dicho año la curva baja de forma muy pronunciada. En *Ilustración 40*, aparte de ese máximo en 2010, podemos ver que la curva tiene una pendiente muy elevada hasta 2005, es decir, que hasta 2005 creció muy rápido. En la *Ilustración 40* también se puede observar que entre los años 2007 y 2010 se produjo un gran aumento de la población en estos municipios.

La media anual del número de habitantes de los municipios de más de 20.000 habitantes es de 1.280.866, frente a los 1.212.209 de esa misma media de los municipios de menos de 20.000 habitantes. Aunque estas dos cantidades no son muy diferentes, dada el número de habitantes medio por municipio es de 543,104 para los de menos de 20.000 habitantes es y 80.054,100 para los de más de 20.000 habitantes. El municipio con menor población en el grupo de más de 20.000 habitantes es Arroyo de la Encomienda (código INE 47010) con 20.179 habitantes. Los dos municipios con mayor número de habitantes en el grupo de municipios de menos de 20.000 son Villaquilambre (código INE 24222) y Benavente (código INE 49021) con 18.638 y 17.935 respectivamente, el resto de los municipios no superan los 15.000 habitantes. Dada la evolución demográfica en ambos municipios es previsible que Benavente siga perteneciendo a este grupo, mientras que Villaquilambre podría pertenecer al grupo de más de 20.000 habitantes en unos años.

Los análisis se realizaran para cada uno de los 2 grupos de municipios y dentro de cada grupo se realizan para las variables clásicas y para las variables de migración; además de un análisis para todas las variables. Se utilizaran 22 variables por cada indicador demográfico ya que se cuenta con 22 observaciones por indicador, una por año.

4.1 ANÁLISIS DE LOS MUNICIPIOS DE MÁS DE 20.000 HABITANTES

Los municipios que se encontraban en 2019 en esta lista de más de 20.000 habitantes son: Ávila, Aranda de Duero, Burgos, Miranda de Ebro, León, Ponferrada, San Andrés del Rabanedo, Palencia, Salamanca, Segovia, Soria, Arroyo de la Encomienda, Laguna de Duero, Medina del Campo, Valladolid y Zamora. En 1996 los municipios de Arroyo de la Encomienda y Laguna de Duero no estaban entre los municipios de más de 20.000 habitantes, pero estos municipios han tenido un fuerte crecimiento a lo largo de todo el período estudiado. El municipio de Arroyo de la Encomienda supera los 20.000 habitantes en el año de 2019 aunque algunos periodos antes están en valores muy cercanos; mientras que Laguna de Duero en el año 2003 ya había superado estos 20.000 habitantes. Dada la evolución demográfica de estos municipios es previsible que sigan perteneciendo a este grupo.

4.1.1 Análisis de todas las variables

En este grupo tenemos todas las variables: nacimientos, defunciones, mujeres en edad reproductiva, población de derecho varón, población de derecho mujer, emigraciones e inmigraciones. Trataremos todas las variables como tasas por cada 1.000 habitantes, según se menciona al inicio de este capítulo. En la *Ilustración 42* podemos ver las primeras variables para los primeros municipios de este grupo.

```
> head(TODOT)
                  NAC1996  NAC1998  NAC1999  NAC2000  NAC2001  NAC2002  NAC2003  NAC2004   NAC2005
AVILA           10.553754 9.317943 9.269745 9.844700 9.798403 8.220378 9.082530 9.901368 10.796016
ARANDA DE DUERO  8.794744 9.345164 8.823034 8.769572 8.666956 9.370154 7.522960 9.878543  9.376900
BURGOS           8.850425 8.821859 9.238216 9.763832 9.281147 9.305676 8.935901 9.429403  9.488403
MIRANDA DE EBRO  7.182694 6.864989 8.805244 8.952878 8.570623 8.995585 9.049774 9.400324  9.611300
LEON             6.616543 7.102547 7.023868 7.586627 7.591859 7.570290 8.132179 8.034524  7.557875
PONFERRADA       7.762891 8.329402 7.391269 8.125539 7.432828 8.279956 7.785813 8.093870  8.289888
```

Ilustración 42: captura de la cabecera del dataframe más de tasa por 1000 de más de 20000 habitantes

4.1.1.1 Componentes principales

Para el análisis de las componentes principales se realiza en todo los casos con la función "*prcomp*" de R con la opción *scale=T*. De esta manera, se tipifican los datos, dado que esta función ya realiza la centralización de los datos. Esta función calcula automáticamente los *principal components scores*, es decir, el valor de cada componente para cada variable, esto lo podemos observar los resultados de las 16 componentes en *Ilustración 43*.

```
> head(pca2$x)
                      PC1       PC2        PC3        PC4        PC5         PC6         PC7        PC8         PC9
AVILA           -1.427762 -4.251123  3.8687436 -0.1683707 -0.7797484 -1.90615366 -1.49560792  0.7110925  1.37523947
ARANDA DE DUERO  3.053112 -5.051499 -2.3476805 -2.3013566 -1.1125890 -0.62652346  3.08360216 -0.5612857 -0.08225158
BURGOS          -3.710051 -4.751471  1.5067436  0.4717395  0.5391705 -0.06027804 -0.14784596 -0.2374137  0.16577583
MIRANDA DE EBRO  0.318427 -3.952312 -7.1674419 -4.2014013  0.7585806  2.28779058 -1.30449193  0.2382656  1.01234200
LEON            -9.211298 10.197695  1.5533387  0.7850257  0.4317878  2.06751533  0.07542339 -0.2444203 -0.79730582
PONFERRADA      -3.418424 -4.015933 -0.2515121  0.3705968 -1.8026454 -0.45509770  0.57672817  0.5338852 -0.28417452
                       PC10        PC11        PC12       PC13         PC14         PC15         PC16
AVILA           -0.86233289  0.12858332 -0.55024069 0.51529686 -0.386610962  0.194312625 1.268039e-14
ARANDA DE DUERO -1.14111310 -0.38185215 -0.35472851 0.07059097 -0.001823293 -0.002581307 1.272767e-14
BURGOS          -0.35635177 -0.01814486  1.48433001 0.67898842  0.472480262 -0.522712023 1.268950e-14
MIRANDA DE EBRO  0.07544293 -0.15433349 -0.03996685 0.05252473 -0.062726470  0.063546242 1.238246e-14
LEON            -0.39671918 -0.13132729 -0.56465148 0.73356340 -0.532333534 -0.364290382 1.155673e-14
PONFERRADA       1.93064600 -0.28484639 -0.52565975 0.71357077  0.365192145  0.050585393 1.283175e-14
>
```

Ilustración 43: cabecera de principal components scores análisis de todas las variables grupo más de 20000

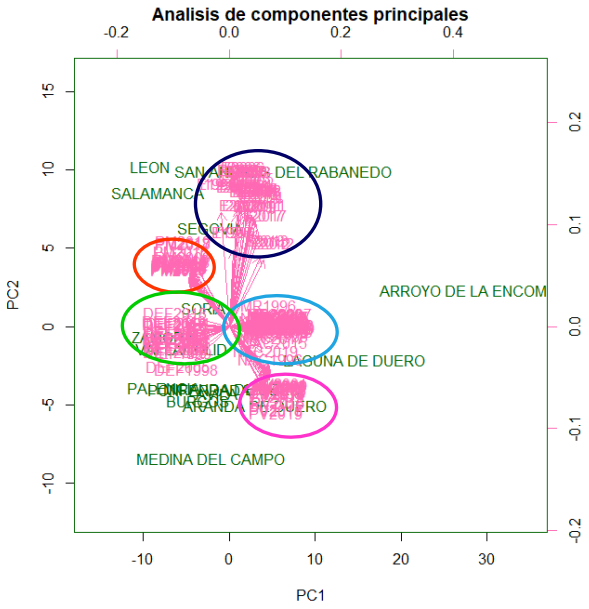

Ilustración 44:biplot de análisis de componentes principales con círculos. Todas las variables en grupo más de 20000 habitantes.

En el siguiente biplot, *Ilustración 44*, podemos ver que las 2 primeras componentes son muy influyentes, dada la configuración que observamos. Observamos que los grupos de variables están separados según los grupos de variables, con unas elipses se remarcan las variables más cercanas. Estos las elipses más se usan para marcar los grupos de variables y dónde influyen más.

- Color azul oscuro: se corresponde con las emigraciones y las inmigraciones. Podemos ver que los municipios más cercanos son San Andrés del Rabanedo, León y Salamanca.
- Color rojo: corresponde con la población de derecho mujer. Vemos que el municipio que más le influyen estas variables es Segovia.
- Color verde: son las variables de las defunciones. Los municipios más influidos por ellas son Zamora y Valladolid.
- Color rosa: son las variables de población de derecho varón. Algún municipio de los mas influenciados son Aranda de Duero, Burgos o Medina del Campo.
- Color azul claro: son las variables de las mujeres en edad reproductiva y los nacimientos. Los municipios con altas tasas en estas variables son Arroyo de la Encomienda y Laguna de Duero.

Para la selección del número óptimo de componentes principales, necesitamos saber la proporción de varianza explicada por cada componente, así como la variabilidad explicada acumulada. En *Ilustración 45* tenemos la varianza explicada por cada componente. Observamos que las primeras componentes explican la mayoría de la variabilidad. En especial, destacamos que la primera componente explica un 60% de la variabilidad. En la *Ilustración 46* tenemos la proporción de varianza explicada acumulada según vamos introduciendo componentes. Observamos que con 4 componentes tenemos más de un 90% de la varianza explicada, con lo que podemos decir que un número adecuado de componentes principales sería de 4 componentes.

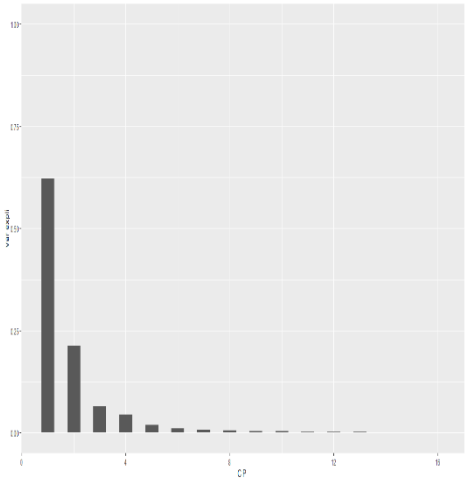

Ilustración 45: varianza explicada por cada componente. Todas las variables en grupo más de 20000 habitantes.

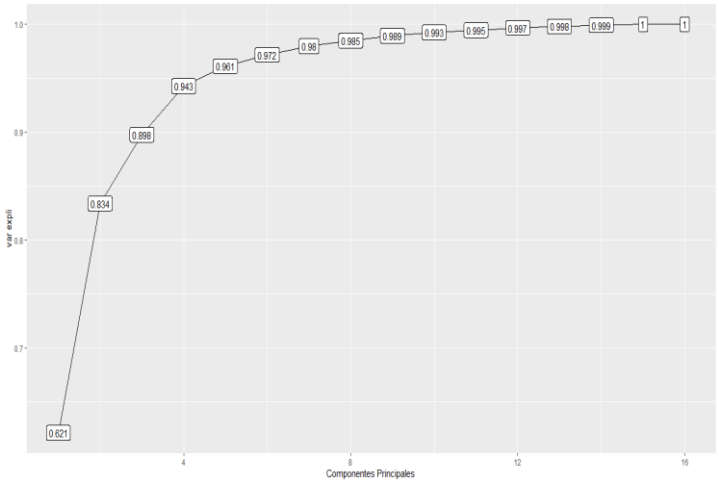

Ilustración 46: proporción de varianza explicada acumulada. Todas las variables en grupo más de 20000 habitantes.

4.1.1.2 Análisis clúster de todas las variables

Para la selección del número óptimo de clúster se utiliza la función "fviz_nbclust" de la librería "factoextra" de R. Para usar esta función, es necesario que los datos estén normalizados. Se usa la opción method="wss" indica que usamos que la suma de cuadrados intragrupo. A la vista de la *Ilustración 47* seleccionamos el número óptimo de clústeres como 3, dado que es el punto donde vemos que se produce el "codo" de la curva.

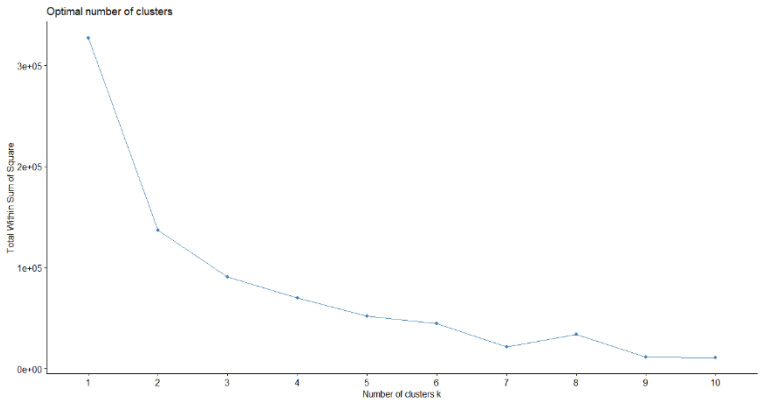

Ilustración 47: número óptimo de clústeres

4.1.1.2.1 Clasificación jerárquica

Para saber que método para calcular las disimilitudes entre los clúster es el mejor para este análisis, miramos la correlación de la matriz de distancias con cada método y, consideramos el mejor método el que tenga más correlación. Para el cálculo de estas correlaciones usaremos las funciones *hclust* y *cophenetic* de R. Para el cálculo usaremos la distancia euclídea. En Tabla 1 podemos ver que el método que más correlación tiene es Average, realizaremos el análisis clúster con este método.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,8517623	0,9038909	0,8776988	0,8958386	0,8882697	0,8978481

Tabla 1: correlaciones de métodos y clúster

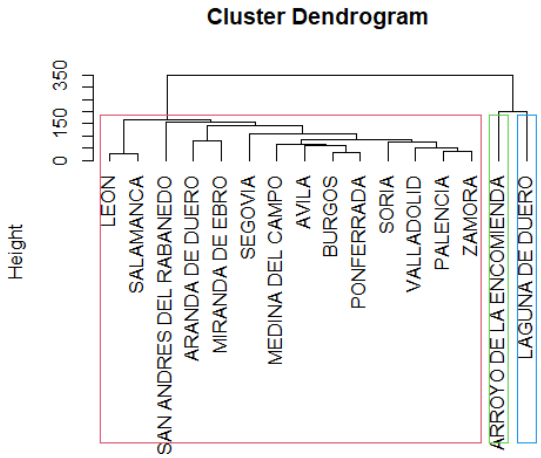

Ilustración 48: dendrograma método average linkage. Todas las variables en grupo más de 20000 habitantes.

En *Ilustración 48* podemos ver los 4 grupos que habíamos visto que eran necesarios en *Ilustración 47*. Los grupos que tenemos son según colores:

- Rojo, es el grupo más numeroso en el encontramos los municipios de León y Salamanca, San Andrés del Rabanedo, Aranda de Duero, Miranda de Ebro, Segovia, Medina del Campo, Ávila, Burgos, Ponferrada, Soria, Valladolid, Palencia y Zamora. La población media por municipio es de 89.246,9 habitantes. La población media en estos municipios fue aumentando lentamente a esta 2011, año en el que se produce la población media por municipio más alta. Tras este año 2011 la población ha ido disminuyendo a lo largo de los años de una manera más rápida a la moneda en que crecía hasta 2011. Sus tasas de natalidad se sitúan en torno a 8,3 por cada 1000 sin variar mucho en los años, lo mismo sucede con la tasa de defunciones pero aquí vemos un aumento en los últimos años. La tasa de mujeres en edad reproductiva se sitúa en los 230 por cada 1.000 habitantes. Este grupo tiene unas tasas de inmigración de 35 por cada 1.000 habitantes. En este grupo destacamos San Andrés del Rabanedo que posee una tasa de inmigración de 50 por cada 1.000 habitantes. Las tasas de emigración se sitúan en torno a 40 por cada 1.000 habitantes.

- En verde, tenemos Arroyo de la Encomienda, su código INE es 47010. La población media de este municipio es de 10.927,39. Este municipio ha tenido un aumento de población muy rápido a lo largo de los periodos, ya que en 1996 era un municipio que no alcanzaba los 2.000 habitantes y en 2019 pasaba de los 20.000 habitantes. Podemos destacar una tasa de natalidad elevada de 12 por cada 1.000 habitantes y una baja tasa de defunciones de solo 2 por cada 1.000 habitantes. Se ve una ligera bajada de las tasas mujeres en edad reproductiva, de valores de 317 a valores de 297 por cada 1.000 habitantes. Las tasas de inmigración son de 9 por cada 1.000 habitantes pero tiene un aumento rápido y en los últimos periodos se sitúa en 51 por cada 1.000 habitantes. Sus tasa de emigración son de 50 por cada 1.000 habitantes, siendo algo más bajas en los primeros periodos.

- En azul Laguna de Duero con código INE es 47076. Población media por año es de 20.481,52. El aumento en la población de este municipio no se ha producido

de manera brusca como sucedía con Arroyo de la encomienda, sino que a lo largo de los años ha ido aumentando un poco la población. Sus tasas de natalidad han ido decreciendo desde los 15.7 hasta los 7 por cada 1.000 habitantes, sin embargo, sus tasas de defunciones se sitúan en torno a 5 por cada 1.000 habitantes. La tasa de mujeres en edad reproductiva ha descendido mucho desde 1996 a 2019, desde valores de 325 a valores de 234 por cada 1.000 habitantes. Las tasas de inmigración en los primeros y los últimos periodos tienen valores de 19 por cada 1.000 habitantes; pero entre 2006 y 2014 alcanza valores de 45 por cada 1.000 habitantes. En cuanto a las tasas de emigración se sitúan en torno a 30 por cada 1.000 habitantes, con una subida entre los años 2006 y 2013.

En todos los grupos las tasas de población de derecho mujer y varón se sitúan en 500 por cada 1.000 habitantes. Esto nos indica que más o menos existe el mismo número de varones y mujeres en los municipios.

4.1.1.2.2 Clasificación no jerárquica

Para la clasificación jerárquica usaremos el método de las k-medias mencionado en capítulo 3, el cual implementaremos mediante la función *kmeans* de R. Hemos decidido que el número de clúster que tomaremos será 3, número que debemos proporcionarle a las k-medias.

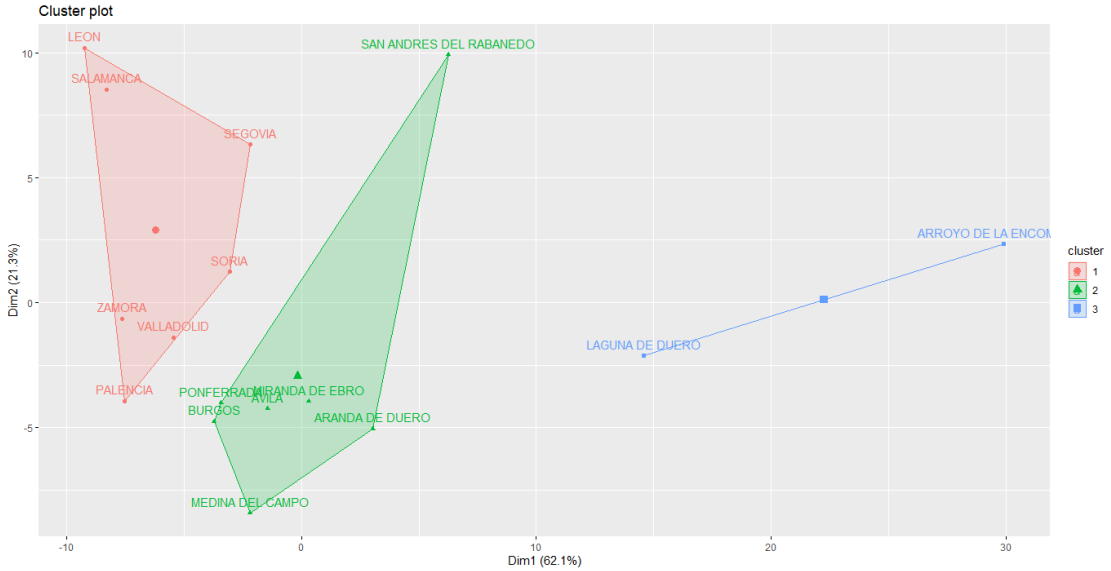

Ilustración 49: Clúster creados por k-medias. Grupo de toda las variables en más de 20000 habitantes.

En *Ilustración 49* tenemos los clúster que nos crea k-medias, donde podemos ver los grupos que creados con los distintos colores. Destacamos el grupo de Arroyo de la Encomienda y Laguna de Duero, municipios donde sabemos que el número de nacimientos, las mujeres en edad reproductiva, así como la población, ha crecido mucho en los últimos años. La población media por municipio es 15.704,46. Tenemos sus tasas explicadas en 4.1.1.2.1 .

El grupo en color rojo lo forman León, Salamanca, Segovia, Soria, Soria, Valladolid, Palencia y Zamora. La población media por municipio se encuentra en 119.766,2. Esta población media ha disminuido ligeramente a lo largo de los años.

El grupo verde está formado por San Andrés del Rabanedo, Aranda de Duero, Medina del Campo, Ávila, Miranda de Ebro, Burgos y Ponferrada. La población media por municipio se encuentra en torno a los 58.727. Esta población media aumento ligeramente hasta el año 2012 y desde ese año ha ido decreciendo.

4.1.1.3 Análisis clúster usando las componentes principales

Para realizar este análisis clúster usaremos las componentes principales seleccionadas en apartado 4.1.1.1, teníamos seleccionadas las 4 primeras componentes principales. Realizamos un análisis como el de clúster del 4.1.1.2, pero usando los componentes principal scores como valores. En este caso, normalizaremos los datos también para poder usar la función "fviz_nbclust" de la librería "factoextra" de R. Lo primero que debemos determinar es el número óptimo de clúster. Podemos ver en *Ilustración 50* que el un número adecuado de clúster es 3.

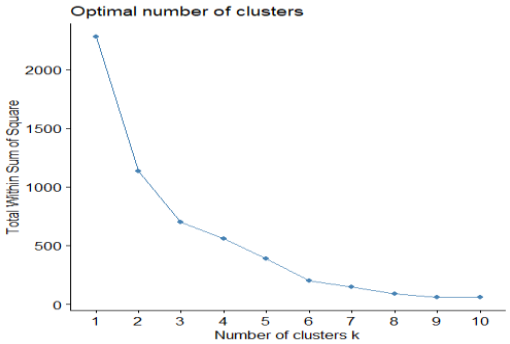

Ilustración 50: número óptimo de clústeres usando las componentes principales en el grupo de más de 20000 habitantes y usando todas las variables.

4.1.1.3.1 Clasificación jerárquica

Debemos seleccionar el método que vamos a utilizar en el clúster jerárquico. Seleccionaremos el que mayor correlación tenga en la *Tabla 2*. Podemos ver que el de mayor correlación es los centroides.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,8175448	0,940068	0,9363586	0,860068	0,9274171	0,9406769

Tabla 2:correlaciones de métodos

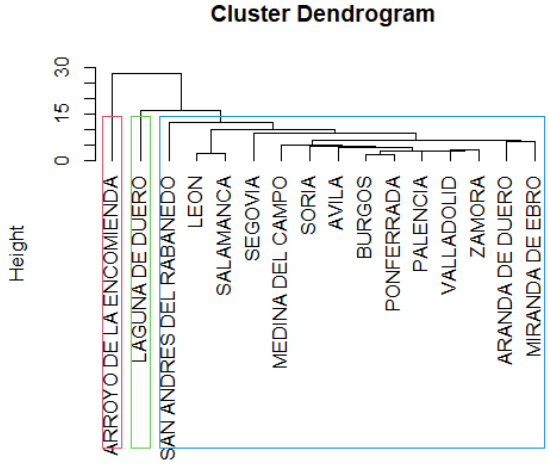

Ilustración 51: dendrograma de PCA + clúster por método de los centroides para el grupo más de 20000 habitantes.

En *Ilustración 51* tenemos el dendrograma del análisis clúster, con el método de los centroides, que realiza los mismos grupos que observábamos en la *Ilustración 50*. Como son las tasas de los grupos lo podemos ver en 4.1.1.2.1.

4.1.1.3.2 Método no jerárquico

Para el método no jerárquico se utilizaran las k-medias, junto con los datos normalizados, usando 3 grupos.

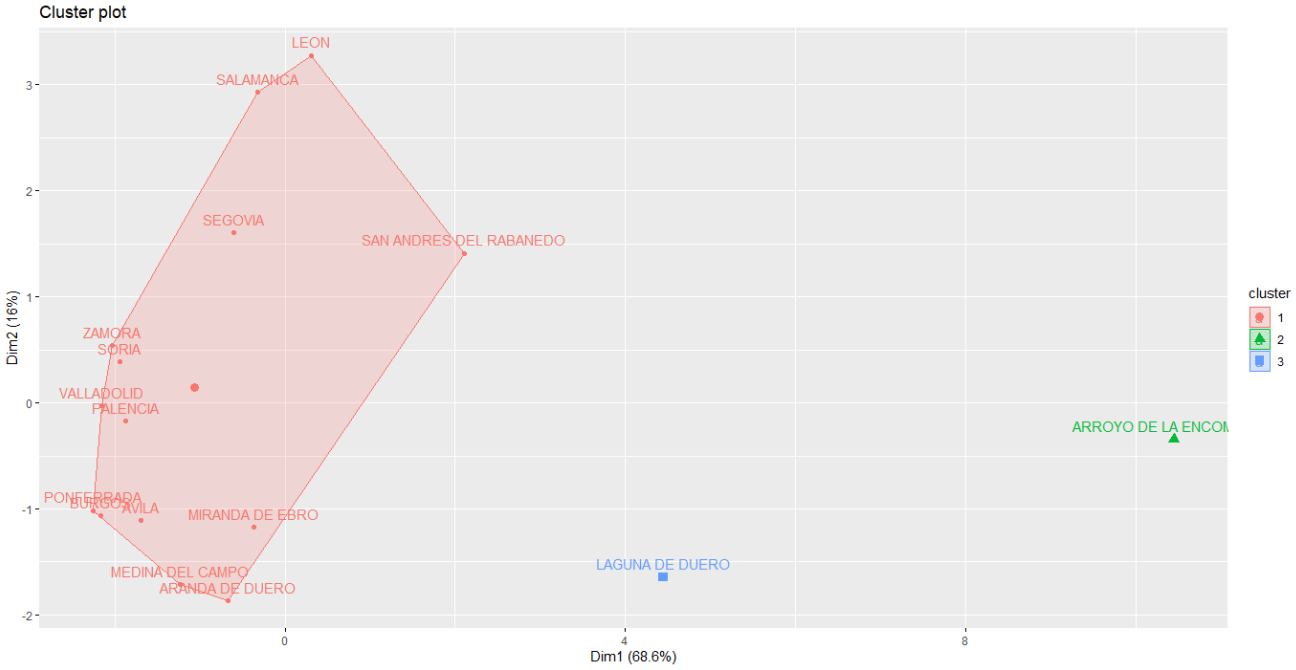

Ilustración 52: clúster de las k-medias usando las componentes principales para el grupo más de 20000 habitantes.

Podemos ver en *Ilustración 52* que se crean 3 grupos, como le habíamos indiciado. Estos 3 grupos son uno formado por Laguna de Duero, otro Arroyo de la Encomienda y, el último grupo el resto de los municipios. Como son las tasas de las variables en los distintos grupos lo podemos ver en 4.1.1.2.1.

4.1.1.4 Multidimensional scaling

Para el multidimensional scaling, como numero óptimo de grupos, usaremos el número de clústeres óptimos seleccionado en análisis clúster de todas las variables (apartado 4.1.1.2). Este número que usaremos es 3.

Dado que en este caso tenemos pocos datos, únicamente 16, los métodos gráficos se verán correctamente. Podemos realizar un método grafico junto con un multidimensional scaling no métrico.

- <u>No métrico</u>

  Para el método no métrico usaremos la función "SAMMON" de R de la librería MASS. Las opciones que pondremos serán: sammon(mdist2, cmdscale(mdist2,2), k=2, niter = 100, trace=T, magic=0.1). Dado que queremos representarlo en dos dimensiones, ponemos k=2 y, el número de iteraciones que usaremos será 100. Como distancias introducidas, usaremos las distancias de los elementos mediante las distancias de Manhattan.

  Podemos ver en *Ilustración 53* que Arroyo de la Encomienda y Laguna de Duero se encuentran muy separados de todos los demás. En este método no vemos muchas

más diferencias para hacer los grupos, a excepción del grupo León-Salamanca, en la parte superior.

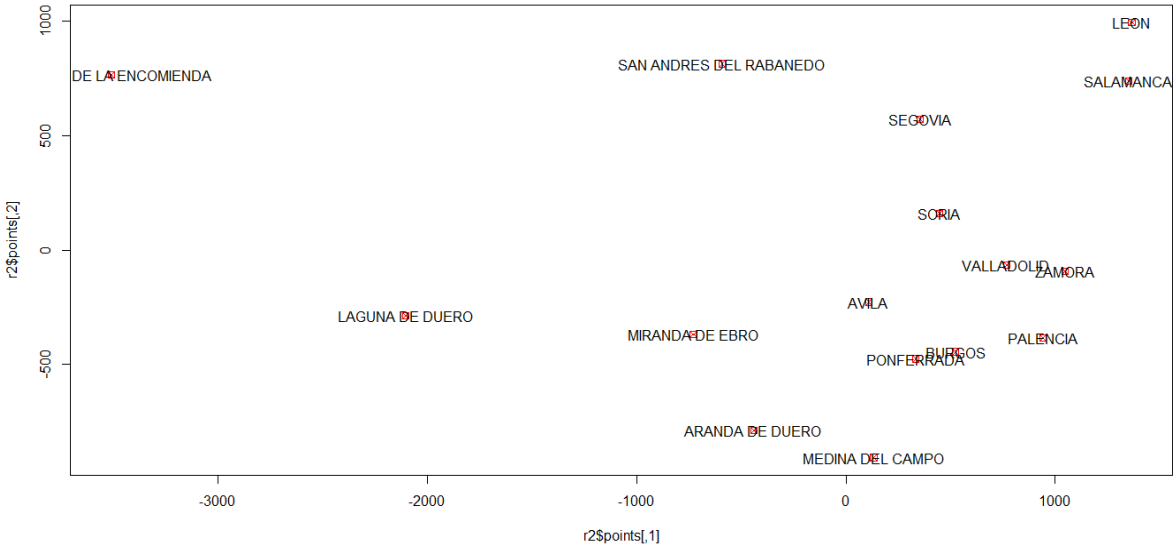

Ilustración 53: multidimensional scaling no métrico para el grupo más de 20000 habitantes.

- Métrico

  Para realizar el método usamos la distancia de Manhattan, representaremos en 2 dimensiones los grupos. Usamos la función cmdscale() de librería *stats* R, junto con unas k-medias sobre ellos, para formar los grupos.
  Podemos ver que nos crea 3 grupos como le habíamos indicado, los grupos que nos ha creado son:

  - Color verde : Arroyo de la Encomienda y Laguna de Duero. Como son las tasas de estos 2 municipios lo tenemos en apartado 4.1.1.2.1. En estas tasas podemos destacar el aumento de la natalidad en estos municipios así como el aumento de la población.

  - Color rojo: San Andrés de Rabanedo, Aranda de Duero, Miranda de Ebro, Ávila, Ponferrada, Burgos y Medina del Campo. En este grupo las tasas de población de derecho mujer se sitúan algo por encima de 500 por cada 1.000 habitantes frente a la tasa de población de derecho varón que se sitúa por debajo de 500 por cada 1.000 habitantes. Las tasas de inmigración se sitúan en 50 por cada 1.000 habitantes frente a las de emigración que lo hacen en 30 por 1.000 habitantes. La población total en estos municipios ha sido más o menos estable.

  - Color azul: León, Salamanca, Segovia, Soria, Valladolid, Palencia y Zamora. La tasa de nacimientos es de 7,3 por cada 1.000 habitantes y vemos un ligero descendimiento a lo largo de los años. Las tasas de defunciones por el contrario han ido aumentando, situándose en valores de 11 por cada 1.000 habitantes. Con las mujeres en edad reproductiva sucede similar, un descendimiento del número a lo largo de los periodos. Las tasas de población de derecho mujer son más elevadas que las de varón, superan los 520 por cada 1.000 habitantes. Las tasas de inmigración han ido aumentadas desde los 10 a los 41 por cada 1.000 habitantes. Las tasas de emigración han

aumentado ligeramente hasta los 35, pero su máximo lo alcanzan en 2006 y 2007 con valores de 46 por cada 1.000 habitantes.

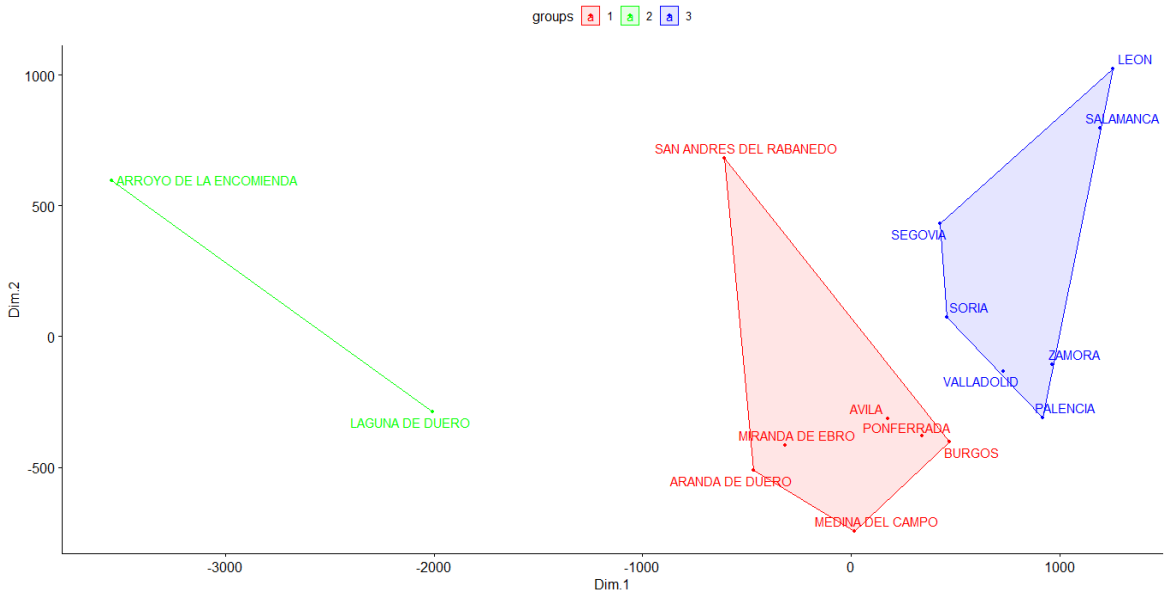

Ilustración 54: multidimensional scaling métrico para el grupo más de 20000 habitantes.

4.1.1.5 Comparación de métodos

Podemos ver que en todos los métodos el número de grupos adecuado es 3. Dependiendo de los métodos podemos ver que realiza una división de Arroyo de la Encomienda y Laguna de Duero como un único cluster o como 2 clúster individuales. Multidimensional Scaling y k-medias para todas las variables realizan esta división en 3 grupos teniendo como un único grupo a Laguna de Duero y Arroyo de la Encomienda. El resto de los métodos realiza una división de municipios como la vista en 4.1.1.2.1. Dada la similitud de las tasas de los municipios podemos decir que esta segunda división como la más adecuada aunque ambas son buenas.

La separación de estos 2 municipios de los demás se produce ya en ellos podemos ver un gran aumento de la población. Esto se ve además reflejado en el crecimiento de las tasas de natalidad y la baja tasas de defunción. Que estas tasas sean de esta manera indica que en estos municipios vive población joven, cosa que corroboramos con las tasas de mujeres en edad reproductiva.

4.1.2 Análisis de las variables de movimiento natural de la población

En este grupo se encuentran los nacimientos, las defunciones y las mujeres en edad reproductiva. Trabajaremos con las tasas para cada 1.000 individuos. De la misma manera que realizábamos un conjunto con todas las variables para el 4.1.1., lo realizamos para solo este grupo de variables.

4.1.2.1 Componentes principales

El análisis de componentes realiza componentes en el mínimo de datos-1 y número de variables, en este caso realizan 16 componentes. Mediante la función *prcomp*, ya mencionada anteriormente, obtenemos el valor de cada componente principal para cada variable, observando los valores para los primeros municipios en la *Ilustración 55*.

```
> head(pca2$x)
                    PC1        PC2        PC3        PC4        PC5        PC6        PC7        PC8        PC9       PC10        PC11
AVILA          -1.12297358  1.7840654  0.2380832 -1.1859663 -0.2484886 -1.2763364  0.7528156  0.29116863  0.4559742 -0.3372433  0.24389335
ARANDA DE DUERO -0.01895166  0.6581587 -1.4716336  1.4547065 -0.8594792 -0.8448681 -0.8006718  0.41792066 -0.5986094 -0.2444475 -0.16562069
BURGOS          1.90204924  0.9823361  0.3944855  0.2198931 -0.1894717 -0.2122898  0.2292739  0.04655690  0.1030837  0.0961212 -0.21532494
MIRANDA DE EBRO  4.75870358  1.2765809 -0.3343311  1.2448806 -0.4439057  0.9713812  1.1908331 -0.07938612 -0.3316621 -0.4419685 -0.03099278
LEON            5.85874354 -0.1106844  0.7721988  0.3934213  1.3751282  0.3371032 -0.4017926  0.54460815  0.3095817  0.1248548 -0.09228614
PONFERRADA      2.83821799 -0.7916868 -1.1308778 -0.4238401  0.1754514  0.3532599 -0.4230126 -0.09242680 -0.0739908 -0.2807271  1.01826888
                    PC12        PC13         PC14         PC15         PC16
AVILA          -0.09459832  0.15535196 -0.128365151 -0.013358674 2.976785e-15
ARANDA DE DUERO -0.15793579 -0.01216112  0.007903568 -0.021556726 3.028827e-15
BURGOS          0.32749075  0.10036733  0.409274656  0.362745578 2.900458e-15
MIRANDA DE EBRO -0.09950776  0.09931628 -0.083587825  0.001310182 3.025358e-15
LEON           -0.42029657  0.04867080 -0.181529544  0.256388877 3.053113e-15
PONFERRADA      0.14500223 -0.19281454  0.040258652  0.106220480 3.004541e-15
>
```

Ilustración 55: valores de cada componente principal en los primeros municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

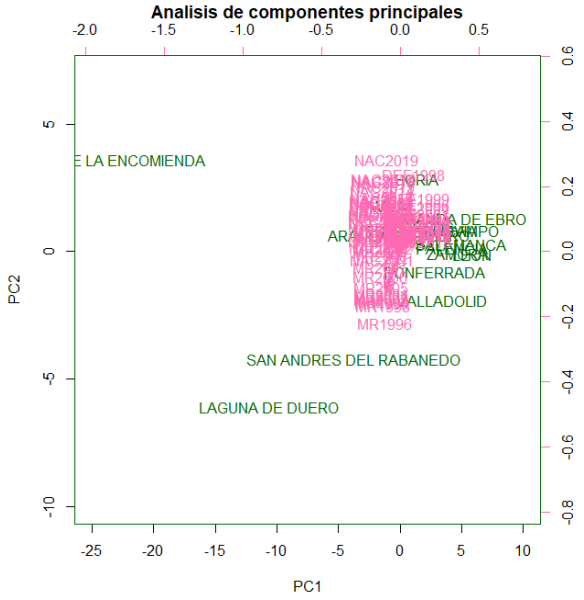

Ilustración 56:biplot de las dos primeras componentes principales municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

En el biplot, *Ilustración 56*, se representan las dos componentes principales en las 2 primeras dimensiones, así como las variables. Podemos decir acerca de esta Ilustración que, Arroyo de la Encomienda está muy influenciado por la segunda componente principal, mientras que Laguna de Duero lo está por la primera componente. En cuanto a que grupos de variables son las que más influyen en cada municipio, no podemos decir mucho, ya que se presentan todas juntas.

Para la selección del número adecuado de componentes principales, debemos calcular la proporción de varianza o variabilidad que explica cada componente. Tenemos la proporción de varianza que explica cada componente en *Ilustración 57*. En ella, podemos ver que la primera componente explica más de un 80% de la variabilidad que tenemos, siendo así la componente que más variabilidad explica en comparación con las demás. Para decidir el número de componentes principales que seleccionamos, debemos mirar la *Ilustración 58*, en ella tenemos la curva de la proporción de varianza acumulada, según vamos introduciendo componentes principales. Al igual que en *Ilustración 57*, vemos que la primera componente explica más del 80% de la variabilidad. Como queremos que, al menos un 90% de la variabilidad quede explicada, seleccionaremos 2 compontes principales, explicando así un 93,6% de la varianza.

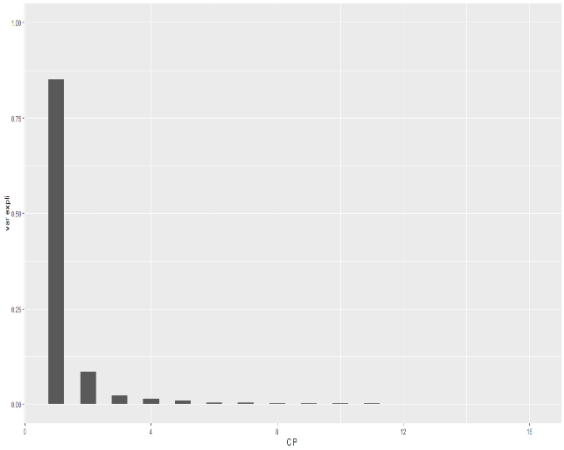

Ilustración 57: componentes principales municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

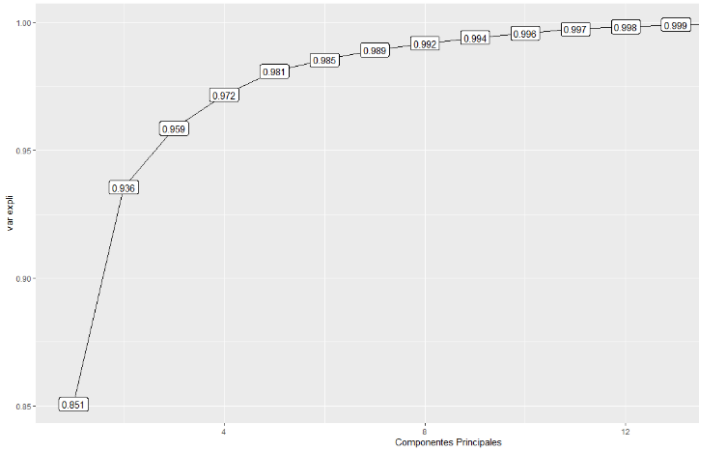

Ilustración 58: proporción de varianza explicada por cada componente municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.2 Análisis clúster de todas las variables

Para el análisis clúster lo realizamos de la misma manera que explicamos previamente. En la *Ilustración 59* vemos que con 2 clúster podemos tener un numero adecuado.

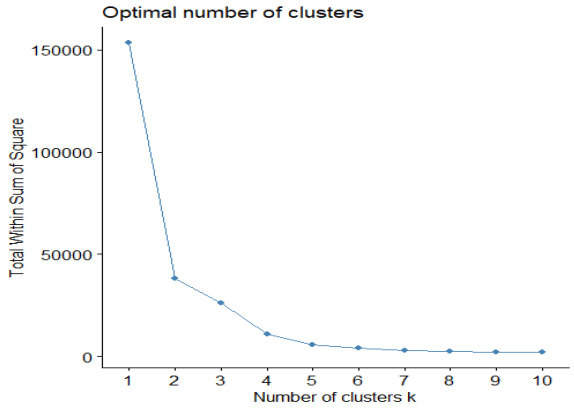

Ilustración 59: número óptimo de clúster municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.2.1 Clasificación jerárquica

Realizaremos la clasificación jerárquica mediante la función hclust. Para la elección del método más adecuado para calcular las disimilitudes entre clúster, elegiremos según la mayor correlación en la matriz de distancias de los métodos. En *Tabla 3* podemos ver que el método con más correlación es single.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,9328723	0,9381556	0,9611173	0,9324461	0,9527595	0,9344101

Tabla 3: tabla de correlaciones de métodos

En *Ilustración 60* podemos ver 2 grupos, uno formado por Arroyo de la Encomienda y, un grupo mucho más grande con el resto de los municipios. Sabemos que Arroyo de la Encomienda es un municipio con la natalidad alta en los últimos años y, además, un numero alto de mujeres en edad reproductiva. Sus tasas las podemos ver en 4.1.1.2.1.

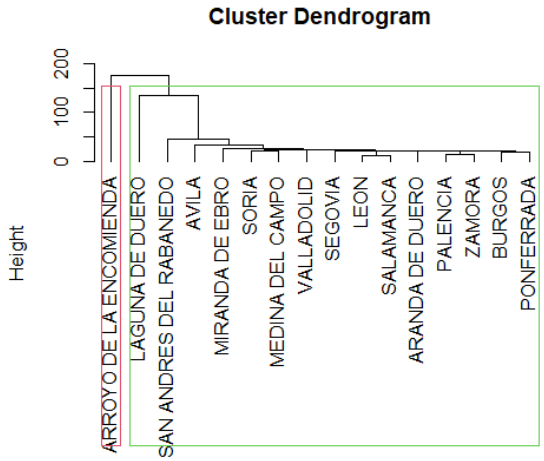

Ilustración 60: dendrograma del clúster de municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.2.2 Clasificación no jerárquica

Para la clasificación no jerárquica usaremos el método de las k- medias, usando la función de kmeans de R, seleccionado el número de grupos como 2.

En *Ilustración 61* tenemos los 2 grupos que habíamos creado, juntos tenemos los municipios de Laguna de Duero y Arroyo de la Encomienda, donde, además, sabemos que tienen valores altos en las variables del grupo1. En el otro grupo tenemos el resto de los municipios. Las tasas de estos grupos las tenemos en 4.1.1.2.1.

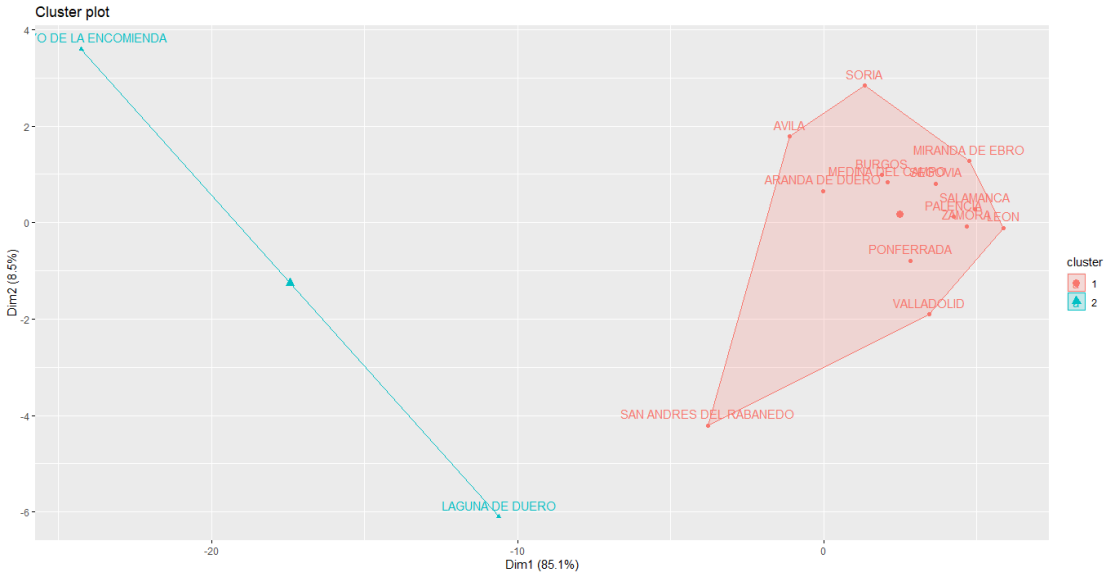

Ilustración 61: cluster de las k-medias municipios para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.3 Análisis clúster usando las componentes principales

En el análisis de componentes principales seleccionamos que, el número de componentes principales para explicar más del 90% de la variabilidad, son 2. Ahora, miraremos el número óptimo de cluster observando donde se produce el "codo" de la gráfica, la cual es la *Ilustración 64* que el número óptimo de cluster es 2.

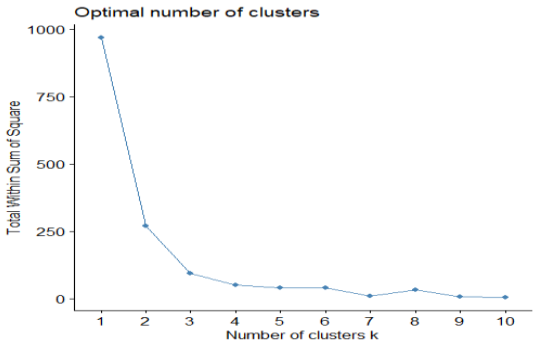

Ilustración 62: curva de número óptima de clúster usando las componentes principales para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.3.1 Método jerárquico

De la misma manera que en el análisis cluster, usando todas las variables, en este caso, comprobaremos la correlación de las matrices de distancias de los métodos de cálculo de disimilitudes de los datos. Escogeremos como mejor método el que tenga la correlación más elevada, en este caso sería centroides (*Tabla 4*).

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,9475034	0,9614434	0,9645921	0,9390812	0,9721567	0,9759956

Tabla 4: tabla de correlaciones de métodos

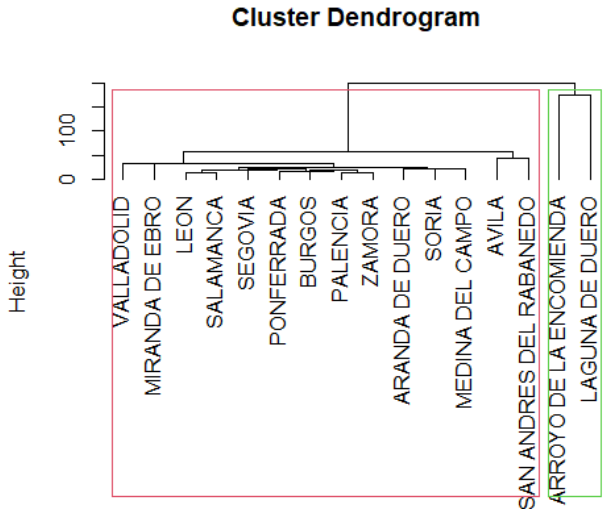

Ilustración 63: dendrograma de centroides usando las componentes principales para el grupo más de 20000 habitantes y variables de movimiento natural.

Podemos ver en la *Ilustración 63*, el dendrograma, 2 grupos claramente desde las primeras ramas del árbol. El primer grupo estaría formado por Laguna de Duero y Arroyo de la Encomienda y, el segundo el resto de los municipios. Sabemos que en los municipios que vemos en cuadrado verde tienen tasas elevadas en algunas variables. Lo explicamos en 4.1.1.2.1.

4.1.2.3.2 Método no jerárquico

Para el método de no jerárquico usaremos las k-medias, seleccionando como número de grupos 2, donde vimos que se producía el "codo" previamente. En la *Ilustración 64* vemos estos grupos. Tenemos un grupo formado por Laguna de Duero y Arroyo de la

Encomienda y, otro, por el resto de los municipios. Estos son, exactamente, los mismos grupos que teníamos en el análisis cluster jerárquico, *Ilustración 63*.

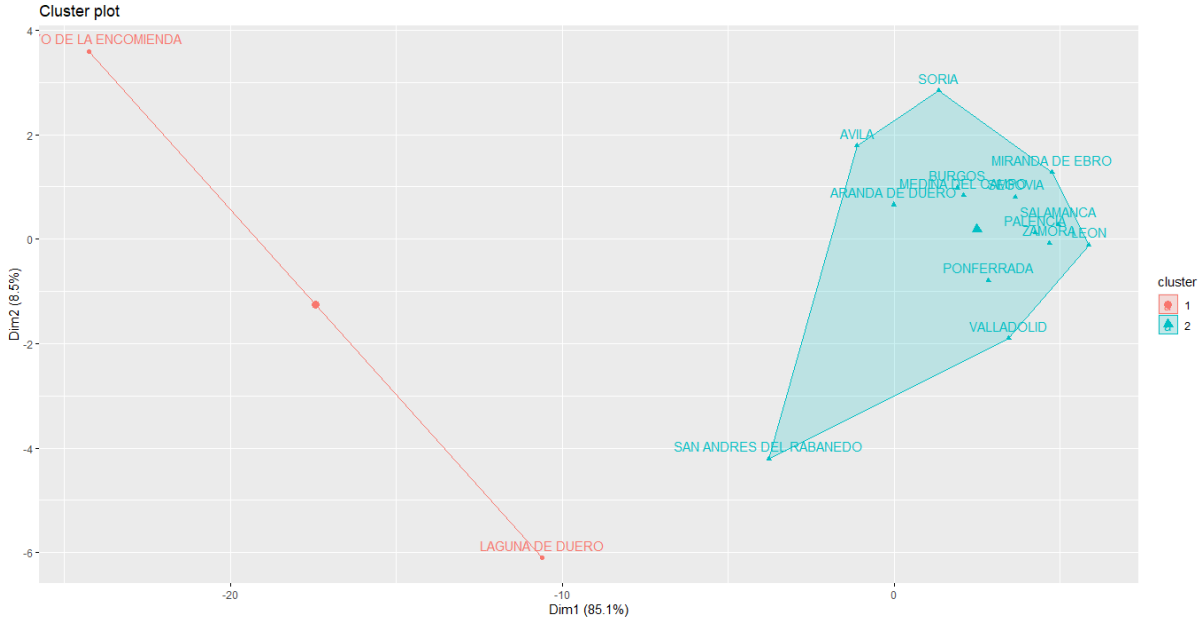

Ilustración 64: cluster de las k-medias usando las componentes principales para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.4 Multidimensional scaling

Para el multidimensional scaling como numero de grupos se usará el número de cluster seleccionado en los apartados previos, que en este caso seria 2. Dado que, en este caso, tenemos pocos datos, tan solo 16, los métodos gráficos se pueden visualizar correctamente y, aplicare un multidimensional scaling no métrico aparte del métrico, de la misma manera que hacía cuando tenía el grupo de todas las variables no solo las de movimiento natural de la población.

- <u>No métrico</u>

  Para el método no métrico usaremos la función "SAMMON" de R de la librería MASS. Las opciones que pondremos serán número de iteraciones 100 y la visualización en 2D. Como distancias introducidas, usaremos las distancias de los elementos mediante las distancias de Manhattan.

  En la *Ilustración 65* no podemos ver claramente ningún grupo, tan solo podemos decir que Laguna de Duero, Arroyo de la Encomienda y San Andrés del Rabanedo son 3 municipios que están separados de los demás. El valor del stress de este análisis es 0,003396459, lo que es un valor muy bueno.

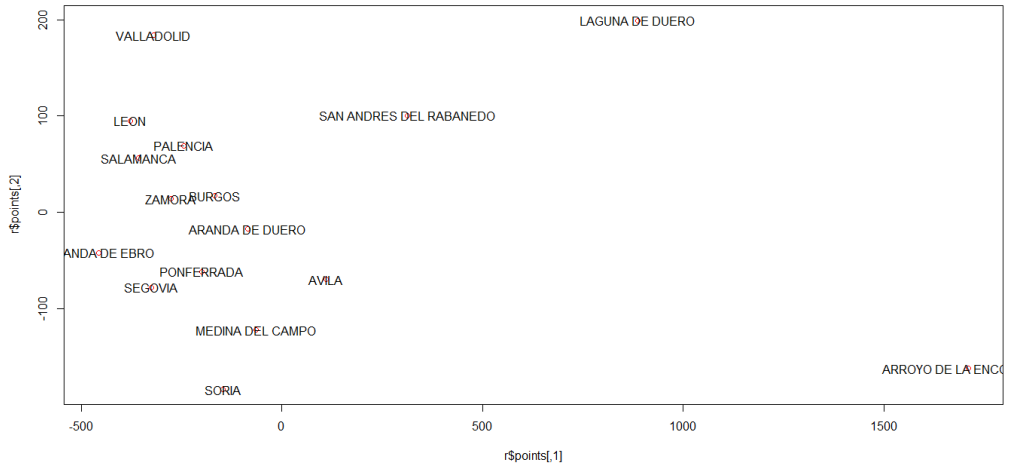

Ilustración 65: multidimensional scaling no métrico para el grupo más de 20000 habitantes y variables de movimiento natural.

- <u>Métrico</u>

  Para el análisis multidimensional scaling métrico en la *Ilustración 68*, vemos los 2 grupos que tenemos representado por colores. El análisis que obtenemos es el mismo que obteníamos por los métodos de cluster: un grupo formado por Laguna de Duero y Arroyo de la Encomienda y, otro grupo formado por los demás municipios.

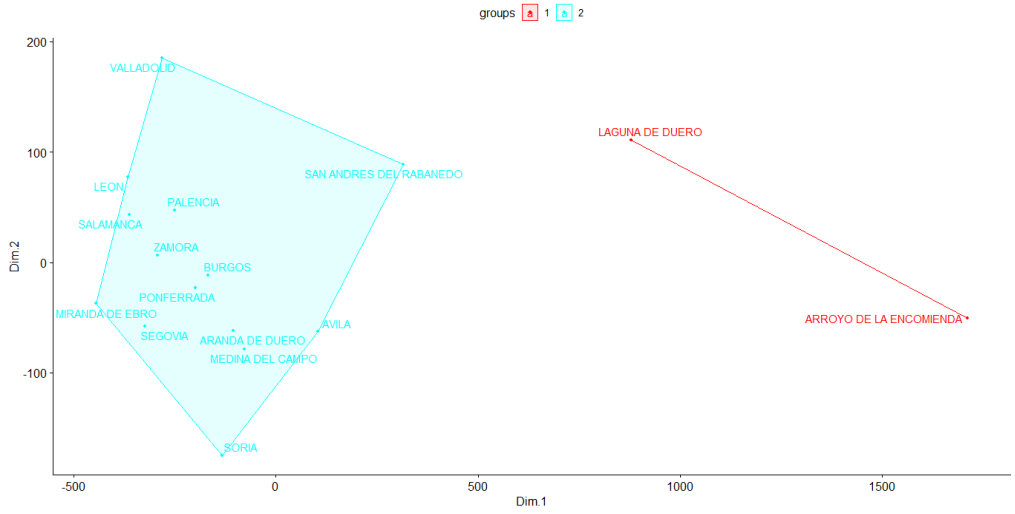

Ilustración 66: multidimensional scaling métrico para el grupo más de 20000 habitantes y variables de movimiento natural.

4.1.2.5 Comparación de métodos

En este caso se crean 2 grupos con todos los métodos. Todos los métodos menos la clasificación jerárquica con las variables realizan un grupo con Laguna de Duero y Arroyo de la Encomienda; la clasificación jerárquica realiza un cluster únicamente con Arroyo de la Encomienda. Sucede esta clasificación de los municipios, ya que al usar las tasas para las variables de movimiento natural estos municipios tienen valores más altos que los demás municipios. Además sucede que en estos municipios las tasas han ido en aumento a lo largo de los periodos. En cuanto a las tasas de defunciones son bajas en estos municipios. Con estas ideas podemos decir que estos municipios han ido creciendo en los periodos, ya que las tasas de nacimiento son más elevadas que las defunciones.

Por estas conclusiones podemos ver que todos los métodos serian igual de efectivos y realizarían el mismo cluster.

4.1.3 Análisis de las variables de estadísticas de población

En este grupo se encuentran las variables de población de derecho varón y población de derecho mujer. En este grupo de variables trataremos los datos calculados en tasa por cada 1.000 habitantes, es decir, que realizaremos los análisis con las tasas por cada 1.000 habitantes de las variables.

4.1.3.1 Componentes principales

El análisis de componentes realiza componentes en el mínimo de datos-1 y número de variables, en este caso realizan 16 componentes. Mediante la función *prcomp*, ya mencionada anteriormente, obtenemos el valor de cada componente principal para cada variable. En la *Ilustración 67* podemos ver como son los valores de las primeras componentes principales.

```
> head(pca2$rotation)
                PC1         PC2         PC3         PC4
varon1996 0.1469482 -0.1857856 0.11125205 -0.1658378 -
varon1998 0.1467397 -0.2141977 0.11197071 -0.1788681 -
varon1999 0.1469823 -0.2111936 0.05598509 -0.1899123 -
varon2000 0.1470098 -0.2145277 0.03339092 -0.1487281
varon2001 0.1474758 -0.1826980 0.03079892 -0.0702985
varon2002 0.1478522 -0.1421763 0.02322417  0.1810646
```

Ilustración 67: primeras componentes principales para el grupo más de 20000 habitantes y variables estadísticas de población.

En *Ilustración 68* podemos ver que, las mujeres, se encuentran en valores negativos de la primera componente y, los varones, en la positiva de la primera componente. Los primeros años, se encuentran en la parte negativa de la segunda componente y, los años, más actuales en la positiva. En cuanto a la influencia, observamos que la mayoría de los municipios se encuentran centrados, es decir, que afectan todas por igual.

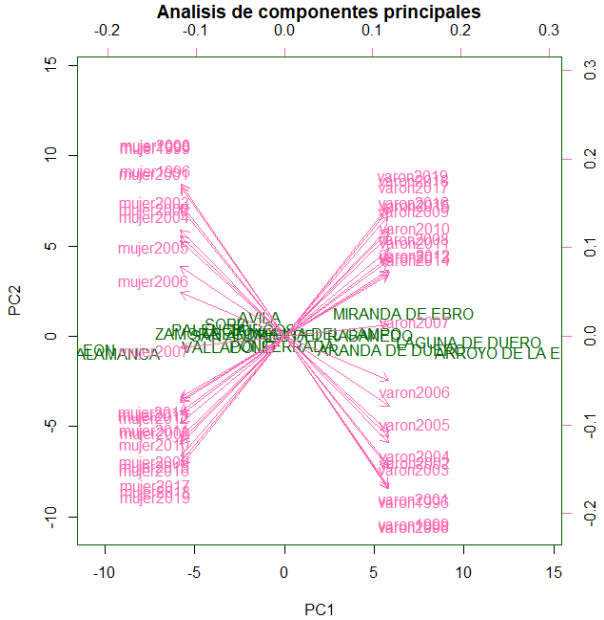

Ilustración 68: biplot de componentes principales para el grupo más de 20000 habitantes y variables estadísticas de población.

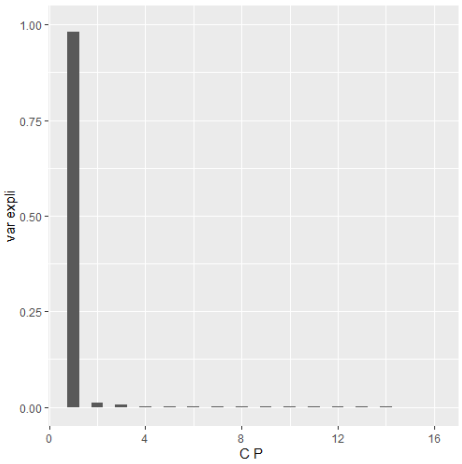

Ilustración 69: componentes principales para el grupo más de 20000 habitantes y variables estadísticas de población.

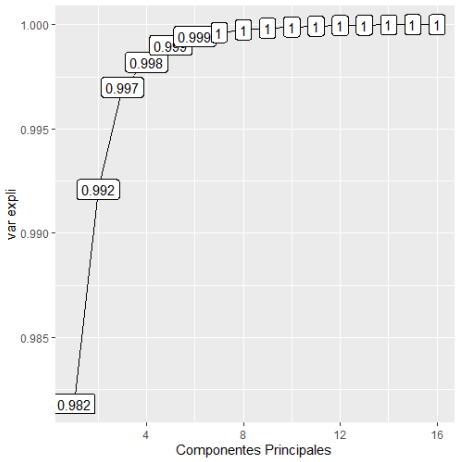

Ilustración 70: proporción de varianza acumulada para el grupo más de 20000 habitantes y variables estadísticas de población.

Para la selección del número adecuado de componentes principales, debemos ver los gráficos de las *ilustraciones 69 y 70*. En la *Ilustración 69* tenemos la variabilidad explicada por cada componente principal. Podemos ver que la primera componente principal explica más del 90% de la variabilidad. Esta idea la corroboramos con la *Ilustración 70*, donde tenemos la proporción de variabilidad explicada acumulada. Tras este razonamiento, podemos decir que el número de componentes principales, para explicar al menos el 90% de la variabilidad, es 1.

4.1.3.2 Análisis clúster de todas las variables

Para el análisis clúster calcularemos un número adecuado de grupos. En la *Ilustración 71* podemos ver la curva del número óptimo de cluster. El "codo" de esta curva se encontraría en 3, por lo que seleccionaremos el número de grupos como 3 grupos .

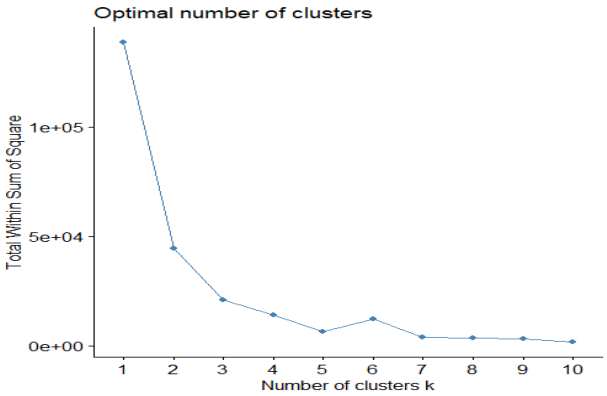

Ilustración 71: número óptimo de cluster para el grupo más de 20000 habitantes y variables estadísticas de población.

4.1.3.2.1 Cluster jerárquico

Para la elección del método más adecuado para calcular las disimilitudes entre clúster, elegiremos según la mayor correlación en la matriz de distancias de los métodos. En la *Tabla 5* tenemos estas correlaciones. Podemos observar que la mayor correlación la vemos en el método Average.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,5750416	0,7915554	0,7240553	0,778613	0,7765988	0,7811139

Tabla 5: tabla de correlaciones de los métodos

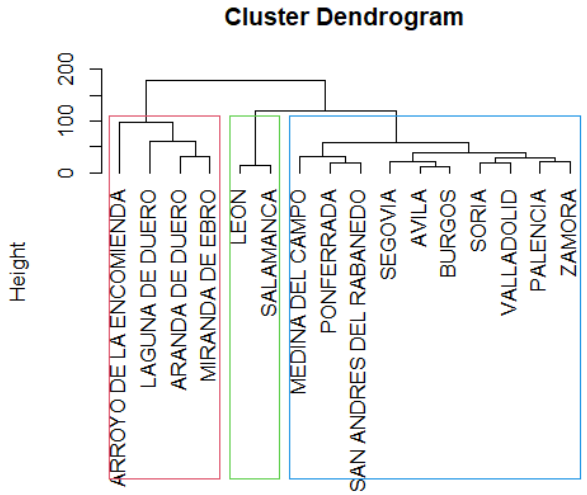

Ilustración 72: dendrograma de método average para el grupo más de 20000 habitantes y variables estadísticas de población.

En la *Ilustración 72* podemos ver que se crean 3 grupos claramente, con las ramas del árbol. Los grupos que se han creado son (según los colores de los cuadrados):

- Grupo de color rojo: lo conforman los municipios de Arroyo de la Encomienda, Laguna de Duero, Aranda de Duero y Miranda de Ebro.
  En este grupo las tasas de población de derecho varón son de más de 500 por cada 1.000 habitantes y además han ido aumentado a lo largo de los periodos. Existe algo más población varón que mujer en los municipios.

- Grupo de color verde: lo forman los municipios de León y Salamanca.
  Las tasas de población de derecho varón han disminuido a lo largo de los periodos. Además estas tasas se encuentran en valores en torno a 465 por cada 1.000 habitantes. En estos municipios la población de mujeres es más elevada que la de varones.

- Grupo de color azul: lo forman el resto de los municipios, siendo este el más numeroso.
  En este grupo las tasas de población de derecho varón se sitúan en torno a los 475 por cada 1.000 habitantes. Esto indica que existe más población mujer que hombres en los municipios pero esta diferencia es muy pequeña.

4.1.3.2.2 Cluster no jerárquico

Para el cluster no jerárquico usaremos el método de las k-medias, mediante la función *kmeans*. Como número de grupos cogeremos el número de grupos que hemos obtenido previamente, este número es 3. En *Ilustración 73* podemos ver que los grupos que realiza son los mismos que realiza el cluster jerárquico *Ilustración 72*. El análisis de los grupos lo podemos encontrar en 4.1.3.2.1 .

Ilustración 73: cluster no jerárquico con k-medias para el grupo más de 20000 habitantes y variables estadísticas de población.

4.1.3.3 Análisis clúster usando las componentes principales

Para el análisis clúster realizaremos y usaremos los mismos análisis que el cluster con las variables. El número de componentes que explicaban más del 90% de la variabilidad era 1 componente. Comenzaremos calculando un número adecuado de grupos, pero no podremos usar la función "fviz_nbclust" de la librería "factoextra" de R, dado que solo tenemos una columna y nos pide que sea una matriz. Usaremos los datos normalizados y calcularemos mediante la función *kmeans* para número de grupos de 1 a 10, apuntando, en cada caso, la suma de cuadrados de los grupos. Después, con estos valores realizamos un gráfico, como la *Ilustración 76*. En esta ilustración podemos ver que el "codo" se produce en 2, con lo que seleccionaremos el número de grupos como 2.

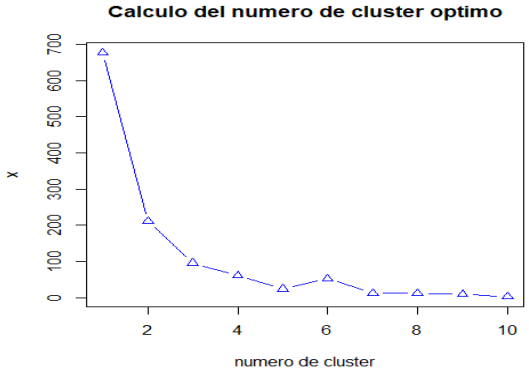

Ilustración 74: cluster óptimos usando PCA para el grupo más de 20000 habitantes y variables estadísticas de población.

4.1.3.3.1 Cluster jerárquico

Para escoger el método más adecuado usaremos las correlaciones de distancias de los cluster en los distintos métodos. El criterio de elección del mejor método será el de mayor correlación, es decir, el que en la *Tabla 6* tenga un valor más alto. Podemos observar que el método con el valor más alto es Average, por lo que será el que escojamos.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,5859623	0,7922196	0,7401351	0,7802835	0,7850046	0,7845993

Tabla 6: tabla de correlaciones de los métodos

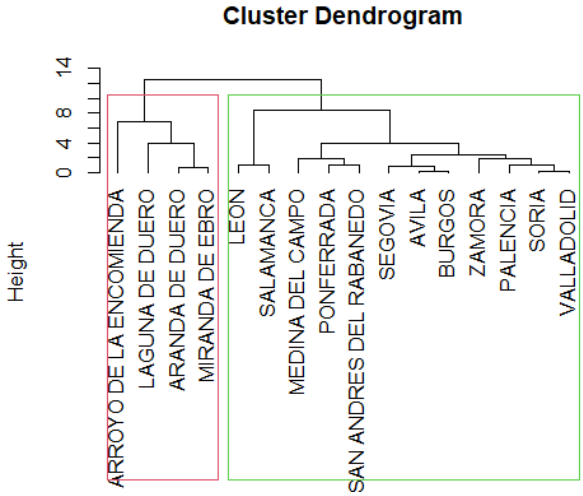

Ilustración 75: cluster jerárquico método average usando PCA para el grupo más de 20000 habitantes y variables estadísticas de población.

Podemos ver en *Ilustración 75* los grupos que nos crea: uno contiene solo 4 municipios y el otro 12. El grupo de 4 municipios lo forman Arroyo de la Encomienda, Laguna de Duero, Aranda de Duero y Miranda de Ebro, mientras que el otro lo forman el resto de los municipios. Podemos ver que el grupo del cuadrado rojo es el mismo que teníamos en 4.1.3.2.1 .En este apartado pudimos ver que en este grupo existe mayor población que son varones, ya que las tasas de población de derecho varón son más elevadas que las de mujer. En el grupo del cuadrado verde sucede el suceso contrario.

4.1.3.3.2 Cluster no jerárquico

Para el cluster no jerárquico realizaremos las k-medias con la función kmeans, seleccionando el número de grupos como 2. Como solo tenemos una componente, no podemos realizar el gráfico como en el caso anterior, como consecuencia ponemos la *Ilustración 76*, siendo los grupos los mismos que encontrabamos en el cluster jerarquico.

```
> kmedias$cluster
                AVILA         ARANDA DE DUERO                BURGOS        MIRANDA DE EBRO
                    1                       2                     1                      2
                LEON             PONFERRADA SAN ANDRES DEL RABANEDO              PALENCIA
                    1                       1                     1                      1
           SALAMANCA                SEGOVIA                  SORIA ARROYO DE LA ENCOMIENDA
                    1                       1                     1                      2
      LAGUNA DE DUERO          MEDINA DEL CAMPO             VALLADOLID                ZAMORA
                    2                       1                     1                      1
```

Ilustración 76: grupos de k medias usando PCA para el grupo más de 20000 habitantes y variables estadísticas de población.

4.1.3.4 Multidimensional scaling

Para el multidimensional scaling tomare el número de grupos que obtuve de cluster (apartado 4.1.3.2), en este caso son 3 grupos. Dado que en este caso solo contamos con 16 municipios, los métodos gráficos se verán correctamente y, realizare un método métrico y uno no métrico.

- <u>No métrico</u>
  Para el método no métrico usaremos la función "SAMMON" de R de la librería MASS. Las opciones que usaré serán número de iteraciones 100 y la visualización

en 2D, niter=100 y k=2. Como distancias introducidas usaremos las distancias de los elementos, mediante las distancias de Manhattan.

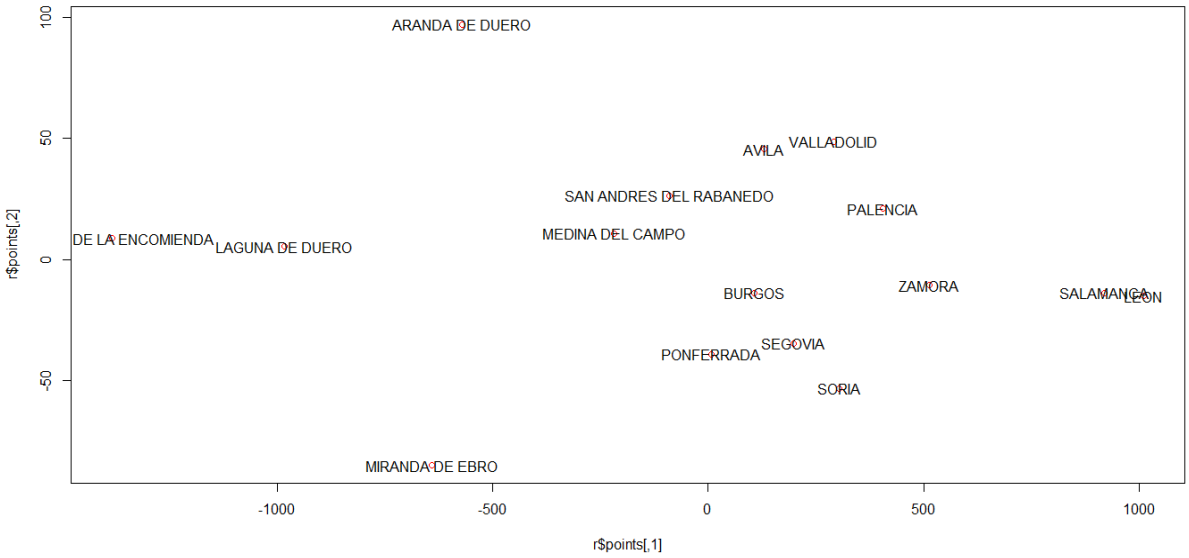

Ilustración 77:multidimensional scaling no métrico para el grupo más de 20000 habitantes y variables estadísticas de población.

Los grupos que podemos observar en la *Ilustración* 77 son los siguientes: Arroyo de la Encomienda y Laguna de Duero se encuentran juntos, Aranda de Duero y Miranda de Ebro se encuentran separados de los demás, Salamanca y León se encuentran juntos y, los demás municipios están, relativamente, cerca unos de otros. El valor del stress es 0,0002570923, con lo que es una buena representación en 2D de los datos.

- <u>Métrico</u>
  Para el análisis multidimensional scaling métrico usare el número de grupos 3. En la *Ilustración 78* podemos ver que se forman 3 grupos de la manera:

  - en color verde San Andrés del Rabanedo, Medina del Campo, Ponferrada, Ávila, Valladolid, Burgos, Segovia, Soria, Palencia.
    Este grupo las tasas de población de derecho varón se sitúan en torno a los 485 por cada 1.000 habitantes, presentando un decrecimiento desde los primeros periodos a los últimos.

  - en color azul Zamora, Salamanca y León.
    Las tasas de población de derecho varón han ido disminuyendo a lo largo de los periodos. Se sitúa en valores de 465 por cada 1.000 habitantes, lo que indica que la población de mujeres es mayor que la de varones.

  - en color rojo, Arroyo de la Encomienda, Laguna de Duero, Aranda de Duero y Miranda de Ebro.
    En estos municipios la tasa de población de derecho varón es mayor que 500 por cada 1.000 habitantes. En algunos municipios de este grupo además estas tasas han ido en aumento a lo largo de los periodos. En estos municipios existe mayor población de derecho varón que mujer.

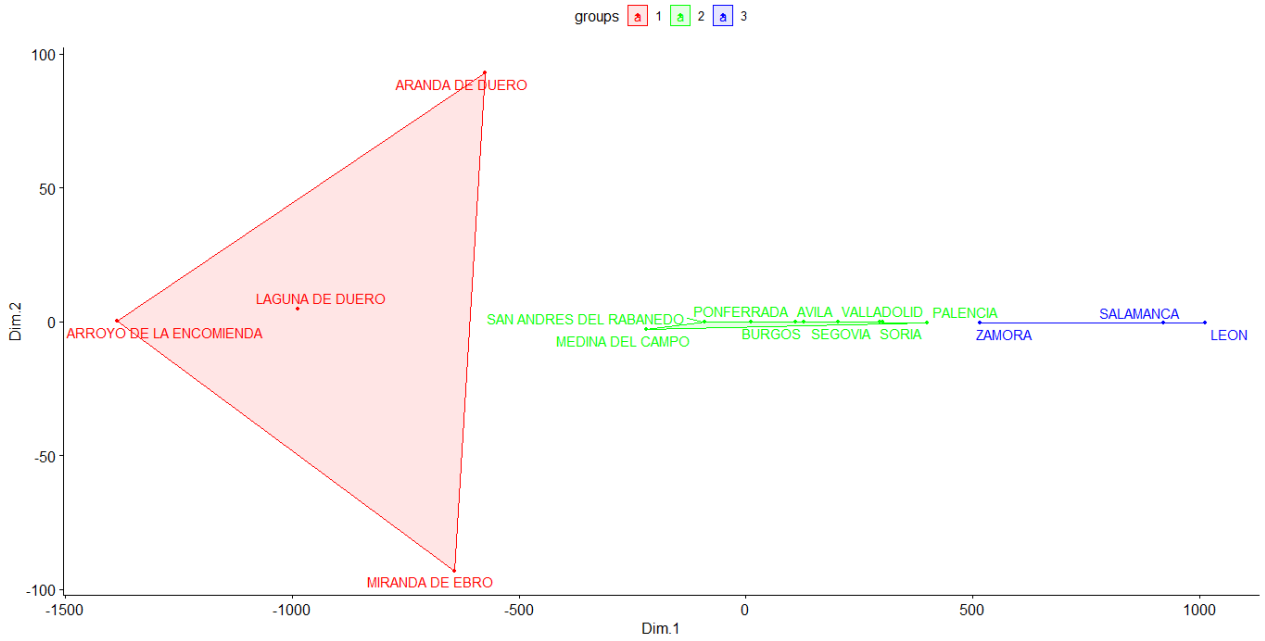

Ilustración 78: multidimensional scaling métrico para el grupo más de 20000 habitantes y variables estadísticas de población.

4.1.3.5 Comparación de métodos

En estos análisis tenemos 2 o 3 grupos. La separación que podemos ver es los municipios que tienen mayor población de derecho mujer y los que tienen mayor población de derecho varones. En los casos en los que tenemos 3 grupos, la división se hace dentro de los que tienen mayor población de derecho mujer.

4.1.4 Análisis de las variables de migración.

Este grupo de variables lo forman las variables de migración o, en otras palabras, las emigraciones y las inmigraciones. No trabajaremos con las variables tal cual, sino que trabajaremos igual que en anteriores apartados con la tasa por cada 1.000 individuos. Estas tasas se calculan como la *variable$_i$ en periodo$_j$ para municipio$_k$ / población total en periodo$_j$ para municipio$_k$* *1000.

4.1.4.1 Componentes principales

En el análisis de las componentes principales se obtiene que el número de compontes principales es 16, ya que es el mínimo de municipios y variables que tenemos. En la *Ilustración 79* podemos ver los valores de las primeras componentes principales y las primeras variables.

```
> head(pca2$rotation)
                PC1         PC2         PC3         PC4         PC5          PC6         PC7
inmi1996 0.09191962 0.360997907 -0.1788286  0.25657589 -0.11389075  0.147505989  0.10639367
inmi1998 0.10645340 0.311935934  0.1959725 -0.16282293 -0.30089512  0.082131211  0.07209993
inmi1999 0.14598139 0.132468896  0.1890248 -0.07035015 -0.07453324 -0.233079938 -0.01503987
inmi2000 0.15380757 0.084024339  0.2035029  0.06414104 -0.07960250 -0.097935725  0.11701039
inmi2001 0.14991968 0.021060625  0.2035098  0.19875209  0.04852617 -0.090504414  0.04129404
inmi2002 0.15488001 0.003046409  0.1252127  0.21929000  0.16336875 -0.006609888 -0.01503600
```

Ilustración 79: componentes principales los primeros valores para el grupo más de 20000 habitantes y variables de migración.

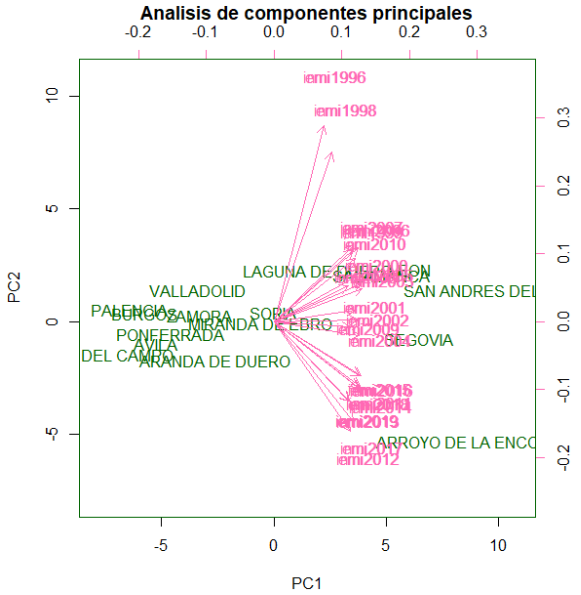

Ilustración 80: biplot de las 2 primeras componentes principales para el grupo más de 20000 habitantes y variables de migración.

En la *Ilustración 80* podemos ver el biplot de las dos primeras componentes principales. En las flechas rosas tenemos las variables, las cuales se encuentran en una línea vertical, por lo que afectan más a la segunda componente. La primera componente explica la mayoría de la variabilidad, ya que podemos ver que todos los municipios se encuentran en una línea horizontal, donde estaría una línea que trazaríamos en el valor 0 de la segunda componente principal. Podríamos decir que los municipios que más emigraciones o inmigraciones han tenido por cada 1.000 habitantes son San Andrés del Rabanedo, Arroyo de la Encomienda y Segovia.

Para la selección del número de componentes adecuado con el que se cumple que un 90% de la variabilidad queda explicada, deberemos mirar la varianza que explica cada componente. En las *Ilustraciones 81 y 82* tenemos la varianza explicada por cada componente. En la *Ilustración 81* tenemos la proporción de la variabilidad explicada por cada componente; podemos ver que la primera componente es la que más explica con mucha diferencia respecto del resto, ya que supera el 75% y las demás ninguna llega a alcanzar el 20%. En la *Ilustración 82* tenemos la curva de la variabilidad explicada acumulada, tras visualizar esta curva podemos ver que, para alcanzar al menos el 90% de variabilidad explicada, debemos seleccionar 4 componentes principales.

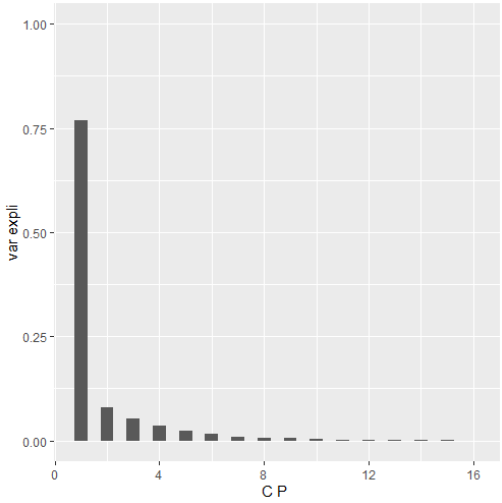

Ilustración 81: explicación de cada componente principal para el grupo más de 20000 habitantes y variables de migración.

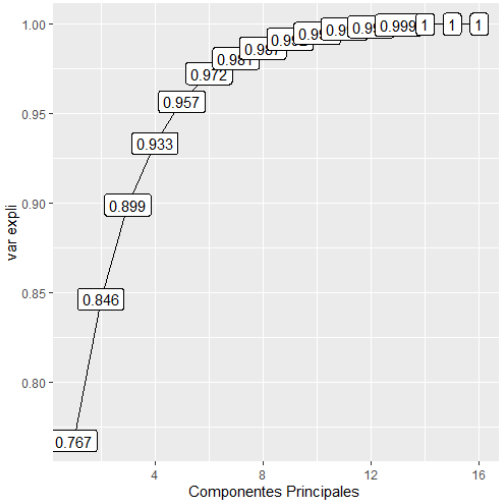

Ilustración 82: variabilidad explicada por las componentes principales acumulada para el grupo más de 20000 habitantes y variables de migración.

4.1.4.2 Análisis clúster de todas las variables

Para saber el número de grupos en el análisis cluster debemos tener las variables normalizadas, ya que la función "fviz_nbclust" de la librería "factoextra" de R necesita que estén normalizados los datos. Para la selección del número de grupos miramos la *Ilustración 85,* en ella podemos ver que el número adecuado de clúster es 3, ya que es donde se produce el "codo".

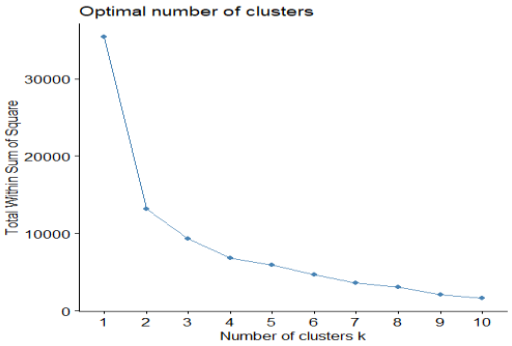

Ilustración 83: número óptimo de clúster para el grupo más de 20000 habitantes y variables de migración.

4.1.4.2.1 Cluster jerárquico

Para la realización del clúster jerárquico debemos seleccionar el método para el cálculo de las disimilitudes entre clúster. En la *Tabla 7* podemos ver la correlación de las matrices de disimilitudes, según distintos métodos. Tomaremos como el mejor método el que presente una correlación más elevada de esta matriz de disimilitudes. Podemos ver que este método con la correlación más elevada es Average, por lo que le tomaremos como el método a utilizar en el análisis cluster como opción de la función hclust.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,8037094	0,8188967	0,7270277	0,803235	0,2723285	0,7716478

Tabla 7: tabla de correlaciones de los métodos

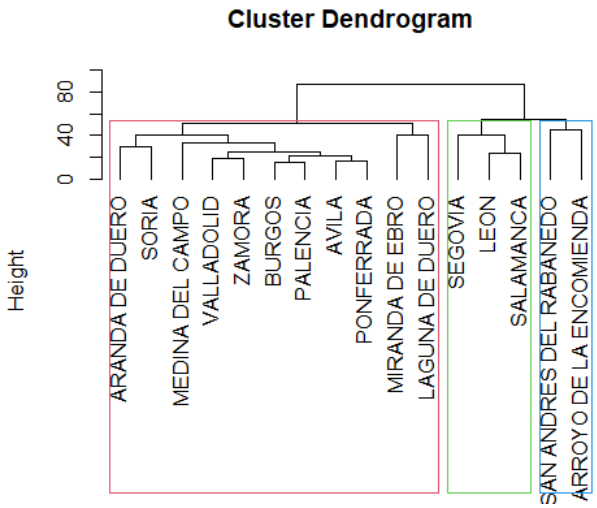

Ilustración 84: dendrograma de cluster jerárquico para el grupo más de 20000 habitantes y variables de migración.

En la *Ilustración 84* podemos ver los 3 grupos que el cluster nos ha formado. Según los rectángulos de colores que podemos ver, tenemos en el grupo azul los municipios de San Andrés del Rabanedo y Arroyo de la Encomienda. Sabemos que este grupo tiene tasas más elevadas de emigración. Sus tasas de emigración se sitúan en los 50 por cada 1.000 habitantes y observamos un aumento según los años. En cuanto a las tasas de inmigración observamos el mismo suceso. En el color verde tenemos Segovia, León y Salamanca. En cuanto a las inmigraciones vemos un aumento de las tasas en los primeros años hasta 2010 y, después, un ligero descendimiento. Sobre las tasas de emigración podemos decir que aumentan hasta los años 2005 y 2006, alcanzando valores de 49 por cada 1.000 habitantes y, después, disminuyen unos años; produciéndose un ligero aumento en los últimos años. Como grupo rojo, tenemos el resto de los municipios. En este grupo las inmigraciones se encuentran en valores de 35 por cada 1.000 habitantes y las emigraciones en valores similares.

4.1.4.2.2 Cluster no jerárquico

En el cluster no jerárquico usaremos el método de las k-medias mediante la función *kmeans* de R, como número de grupos tomaremos el que hemos seleccionado previamente, es decir, 3. En la *Ilustración 85* tenemos el gráfico de los cluster que nos forman las k-medias, observamos que son muy distintos a los que nos formaban el método jerárquico. El grupo de color rojo tiene un aumento de las tasas de emigración e inmigración a lo largo de los periodos y, además, tiene valores cercanos a 50 por cada 1.000 habitantes. En cuanto al grupo azul, las tasas de inmigración se sitúan en 30 por cada 1.000 habitantes, percibiéndose un ligero aumento de los primeros periodos a los últimos. En cuanto a las emigraciones sucede un comportamiento similar.

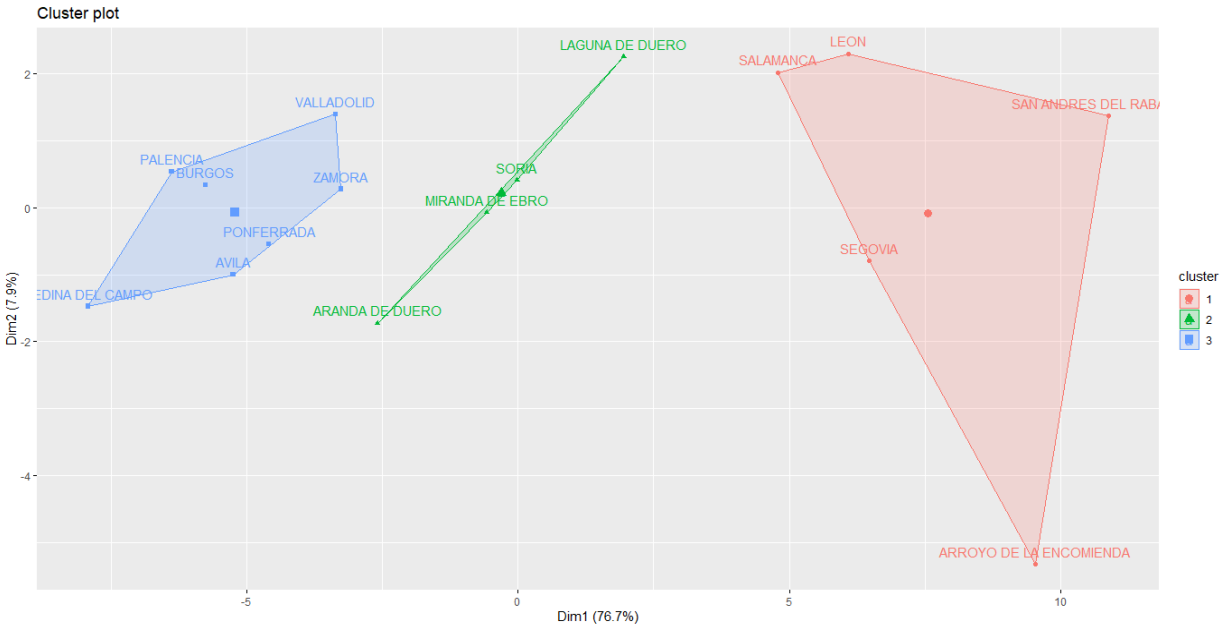

Ilustración 85: cluster de las k-medias para el grupo más de 20000 habitantes y variables de migración.

4.1.4.3 Análisis clúster usando las componentes principales

Los datos deben estar normalizados para poder usar la función "fviz_nbclust". En este caso, los datos que usaremos son los valores de las componentes principales para cada municipio. Para la selección del número de clúster miramos la *Ilustración 86,* en la cual podemos ver que el número óptimo de cluster sería 3.

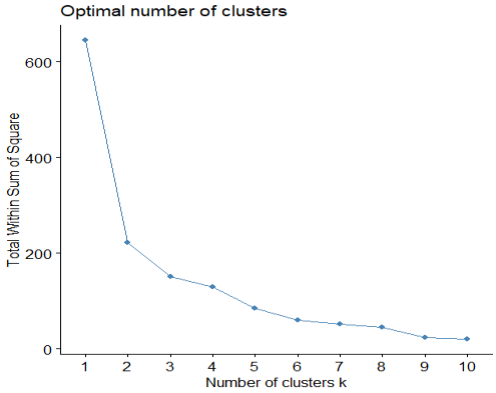

Ilustración 86: número óptimo de clúster usando PCA para el grupo más de 20000 habitantes y variables de migración.

4.1.4.3.1 Cluster jerárquico

Para la selección del mejor método para el cálculo de las disimilitudes entre clústeres, debemos mirar las correlaciones de las matrices de disimilitudes entre grupos calculado en los distintos métodos. Para esta selección del mejor método escogeremos el método con mayor correlación, en este caso es Median.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,8079052	0,8344292	0,7598696	0,8123471	0,3468242	0,8206501

Tabla 8: tabla de correlaciones en métodos

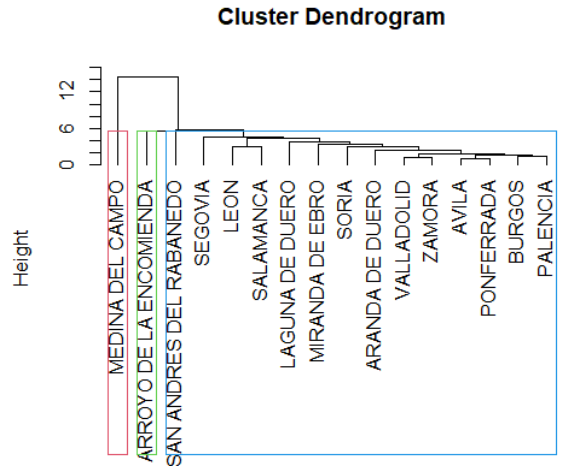

Ilustración 87: cluster del método median usando PCA para el grupo más de 20000 habitantes y variables de migración.

Podemos ver en la *Ilustración 87* los 3 grupos que se han formado. A primera vista observamos que 2 grupos son solo de un municipio, Medina del Campo y Arroyo de la Encomienda y, un tercer grupo tiene el resto de los municipios.

Del grupo formado por Medina del Campo podemos decir que sus tasas de inmigraciones aumentaron hasta 2013, año desde el cual se observa un decrecimiento. Se sitúan en valores cercanos a 27 por cada 1.000 habitantes. En cuanto a las emigraciones, vemos que va oscilando, a lo largo de los años, en valores cercanos a 26 por cada 1.000 habitantes.

En el grupo formado por Arroyo de la Encomienda las tasas de inmigración son de 9 por cada 1.000 habitantes, pero tiene un aumento rápido y, en los últimos periodos, se sitúa en 51 por cada 1.000 habitantes. Sus tasa de emigración son de 50 por cada 1.000 habitantes, siendo algo más bajas en los primeros periodos.

En el grupo formado por el resto de los municipios (cuadrado azul) las tasas de emigración se encuentran cerca de 30 por cada 1.000 habitantes, presentando un mayor aumento desde el año 2008. Las tasas de inmigración se sitúan en valores cercanos a 30 por cada 1.000 habitantes.

4.1.4.3.2 Cluster no jerárquico

Para el cluster no jerárquico usaremos las k-medias con la función kmeans de R y, el número de grupos que tomaremos es 3, dado que es el que obtuvimos como óptimo.
En este caso los 3 grupos tienen un tamaño similar, esto lo observamos en la *ilustración 88*. Los grupos, según colores, que podríamos ver son los siguientes: en color verde tenemos los municipios de Laguna de Duero, Miranda de Ebro y Soria, con color rojo tenemos Salamanca, León, San Andrés del Rabanedo, Arroyo de la Encomienda y Segovia y, con color azul están el resto de los municipios. Los grupos son los mismos que los del apartado 4.1.4.2.2, a excepción de Segovia que cambia de grupo.

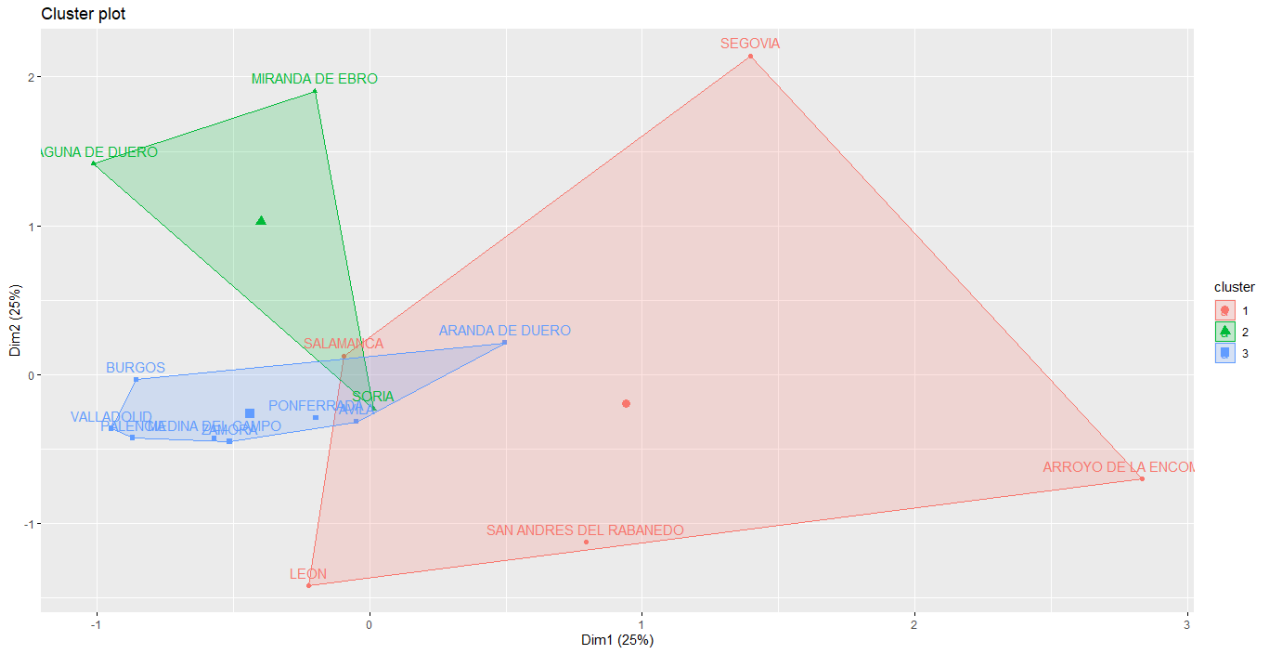

Ilustración 88: cluster de las k-medias usando PCA para el grupo más de 20000 habitantes y variables de migración.

4.1.4.4 Multidimensional scaling

Para el multidimensional scaling usaremos como número de grupos los que nos producía cluster, es decir, 3 grupos. Realizaremos 2 tipos de scaling, dado que, al tener pocos municipios, podemos ver gráficamente las conclusiones.

- <u>No métrico</u>

  Para el método no métrico usaremos la función "SAMMON" de R de la librería MASS. Lo realizaremos en 2 dimensiones la representación, por lo que pondremos como k=2 y, además, realizaremos 100 iteraciones. Usaremos como método de calcular las distancias la distancia de Manhattan.

  A primera vista en la *ILUSTRACIÓN 89* podemos ver que los municipios están dispersos y, como consecuencia, no podríamos ver grupos claramente.

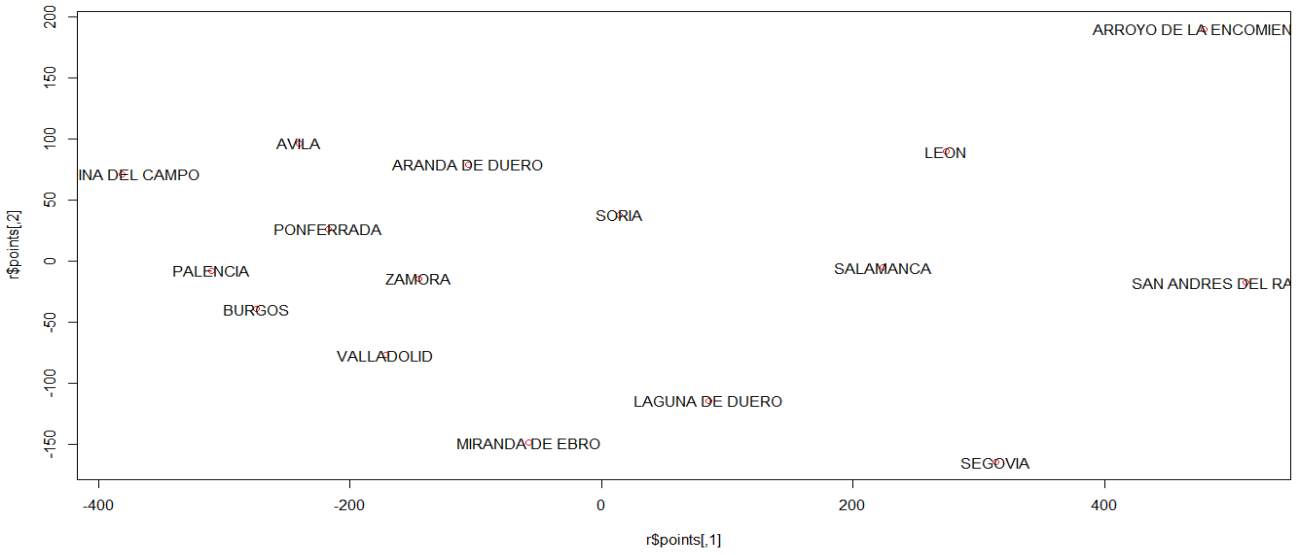

Ilustración 89: multidimensional scaling no métrico para el grupo más de 20000 habitantes y variables de migración.

- Métrico
  En el multidimensional scaling métrico podemos ver, en la *ILUSTRACIÓN 92,* que los 3 grupos que se forman se encuentran bien separados. Los grupos que vemos según sus colores:
    - Azul: lo forman Medina del Campo, Palencia, Ávila, Ponferrada, Zamora, Valladolid y Burgos. En este grupo las tasas de emigración aumentaron.
    - Rojo: formado por los municipios de Aranda de Duero, Soria, Miranda de Ebro y Laguna de Duero. Vemos un aumento de las emigraciones e inmigraciones.
    - Verde: están Arroyo de la Encomienda, León, Salamanca, San Andrés del Rabanedo y Segovia. Los valores de las inmigraciones se sitúan en torno a 50 por cada 1.000 habitantes y los de las emigraciones en valores similares.

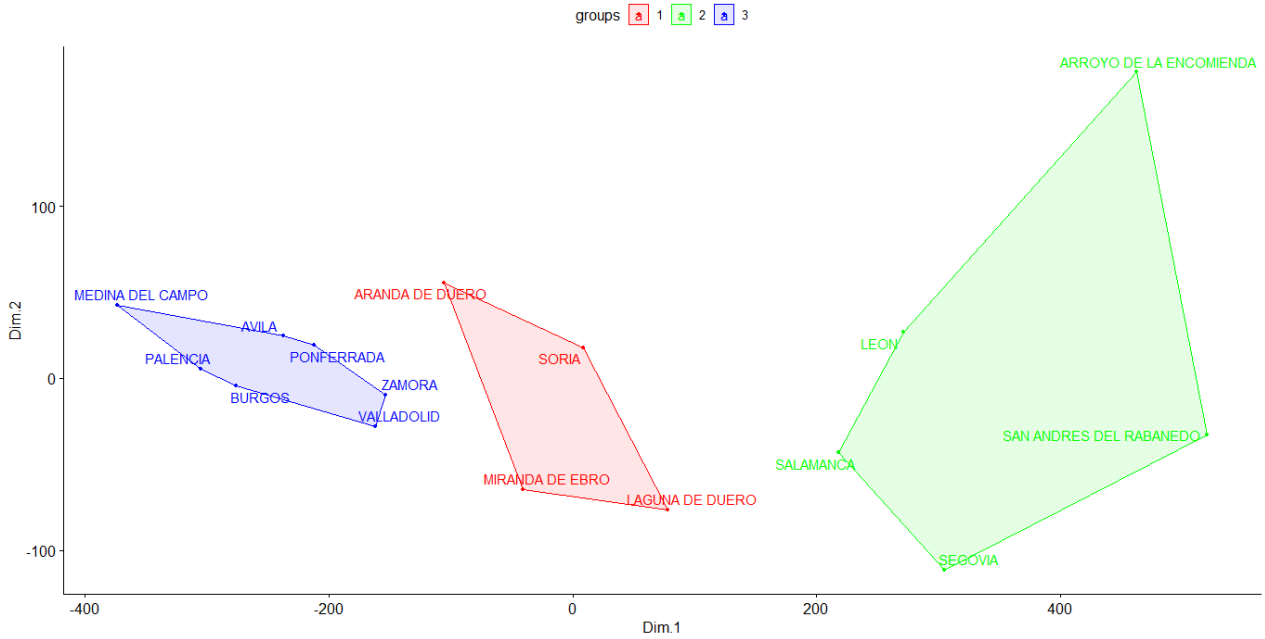

Ilustración 90: multidimensional scaling métrico para el grupo más de 20000 habitantes y variables de migración.

4.1.4.5 Comparación de métodos

En este caso, el número de grupos óptimo es 3. Según los distintos métodos obtenemos distintas clasificaciones. La mayoría de las clasificaciones producen grupos de tamaños similares. Las divisiones que tenemos son municipios que tienen más emigración o menos y, si esta emigración aumento, o no, a lo largo de los periodos.

4.2 ANÁLISIS DE LOS MUNICIPIOS DE MENOS DE 20.000 HABITANTES

De la misma manera que en el *apartado 4.1* tenemos todas las variables que son: nacimiento, defunciones, mujeres en edad reproductiva, población de derecho varón, población de derecho mujer, emigraciones e inmigraciones, pero, en este caso, para los municipios de menos de 20.000 habitantes. Trataremos todas las variables como tasas por cada 1000 habitantes, según se menciona al inicio de este capítulo. Disponemos de 2.232 municipios en este grupo, en la *ILUSTRACIÓN 91* podemos ver las primeras variables en los primeros municipios. Las variables las tenemos seguidas unas detrás de otras, de la misma manera que la *ILUSTRACIÓN 91*.

```
> head(TODO)
                              NAC1996  NAC1998   NAC1999   NAC2000   NAC2001   NAC2002   NAC2003
ADANERO                      0.000000 6.230530  5.405405  8.797654 11.799410  2.824859 21.671827
ADRADA (LA)                  4.493260 8.590197  5.456349  6.128703 10.131712  4.967710  6.419753
ALBORNOS                     0.000000 0.000000 11.538462  8.097166 12.345679 17.094017  4.504505
ALDEANUEVA DE SANTA CRUZ    10.362694 5.291005 10.695187 10.362694  0.000000  5.291005  0.000000
ALDEASECA                    5.524862 8.620690  0.000000  0.000000  3.030303  3.058104  0.000000
ALDEHUELA (LA)               3.436426 7.042254  0.000000  0.000000  3.731343  0.000000  4.132231
```

Ilustración 91: primeras líneas de los datos de menos de 20000 habitantes

4.2.1 Análisis de todas las variables

En el análisis de todas las variables tenemos las variables sin separar en los grupos de variables que hemos creado. Estas variables, de las cuales tenemos las tasas por cada 1000 habitantes, son las siguientes: nacimientos, defunciones, mujeres en edad reproductiva, población de derecho varón, población de derecho mujer, emigraciones e inmigraciones. De la misma manera que en los municipios de más de 20.000 habitantes, realizaré unos análisis estadísticos para clasificar los municipios.

4.2.1.1 Componentes principales

El análisis de las componentes principales lo realizaré de la misma manera que lo realizaba en otras ocasiones. Como el análisis se realiza el mínimo de número de variables-1 y número de datos, el número de componentes que tenemos es 161. Es un número muy elevado, por lo que tenemos que coger el número de componentes para que expliquen la mayoría de la variabilidad, pero que no sea un número excesivo de componentes, ya que queremos reducir la dimensionalidad.

```
> head(pca$rotation)
              PC1         PC2         PC3          PC4         PC5         PC6
NAC1996 0.04673712 -0.06119934 0.05997381 -0.016488670 0.06422862 -0.03500896
NAC1998 0.04844753 -0.05250999 0.03944152 -0.003934219 0.04190520 -0.07206137
NAC1999 0.04086406 -0.06216873 0.02940100 -0.005402551 0.03672034 -0.10481321
NAC2000 0.02899899 -0.07580256 0.04440942 -0.020377481 0.03798881 -0.11845587
NAC2001 0.03976997 -0.07473712 0.02398834 -0.018313071 0.08894061 -0.12489530
NAC2002 0.04062634 -0.06981246 0.01788578  0.007674733 0.07542103 -0.11625281
```

Ilustración 92: cabecera de las componentes principales análisis de todas las variables grupo menos de 20000

En la *ILUSTRACIÓN 92* podemos ver las primeras componentes principales y sus primeros valores. En la *ILUSTRACIÓN 93* podemos ver los *principal components scores*, que es el valor de cada componente principal en cada municipio.

```
> head(pca$x)
                                PC1        PC2        PC3        PC4        PC5
ADANERO                   2.2866243 -3.4260044  0.4569085 -0.9262735 -1.1067518
ADRADA (LA)               6.0484309 -7.0796790 -1.9332850 -0.3195593  0.2944100
ALBORNOS                 -0.5673947 -0.1256284  4.9405816 -2.4269209  0.3285105
ALDEANUEVA DE SANTA CRUZ -2.9642363  2.8284469  2.3339840 -0.9568014  0.1907006
ALDEASECA                 6.3038966  1.8129929  0.1243241 -1.5541137 -1.0119785
ALDEHUELA (LA)            4.2406778  4.8179379 -0.0992717 -0.9138084  1.8528780
```

Ilustración 93: valor de cada componente en cada municipio análisis de todas las variables grupo menos de 20000

En la *ILUSTRACIÓN 94* podemos ver el biplot de la representación en las 2 primeras componentes principales. Al ser tantos municipios, la verdad es que no podemos ver a que municipios le afecta más un grupo de variables u otros. Por la masa de puntos verdes que son los municipios, podemos decir que tienen valores altos en ambas componentes la mayoría de ellos, ya que se encuentran en el centro y, allí, es donde afectan ambas por igual.

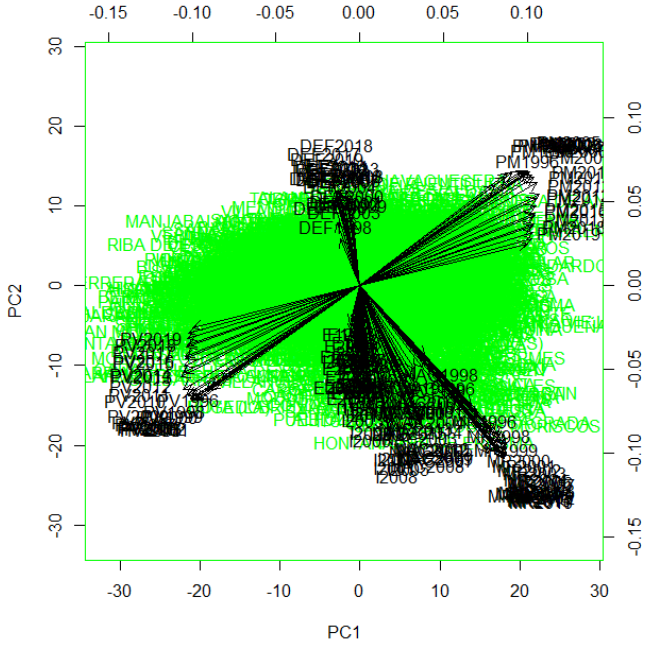

Ilustración 94: biplot de componentes principales análisis de todas las variables grupo menos de 20000

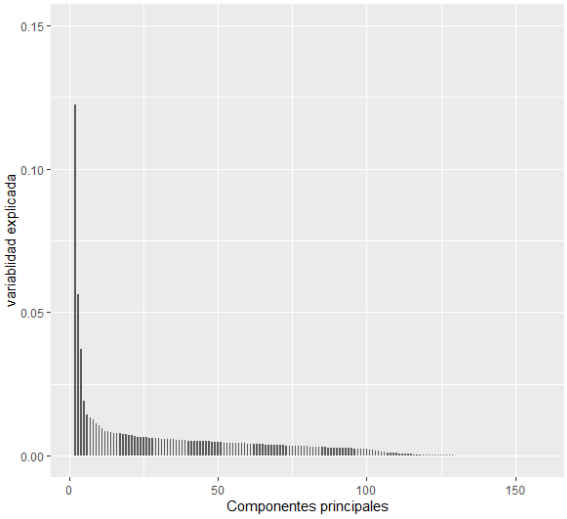

Ilustración 95: proporción de varianza explicada por cada componente principal

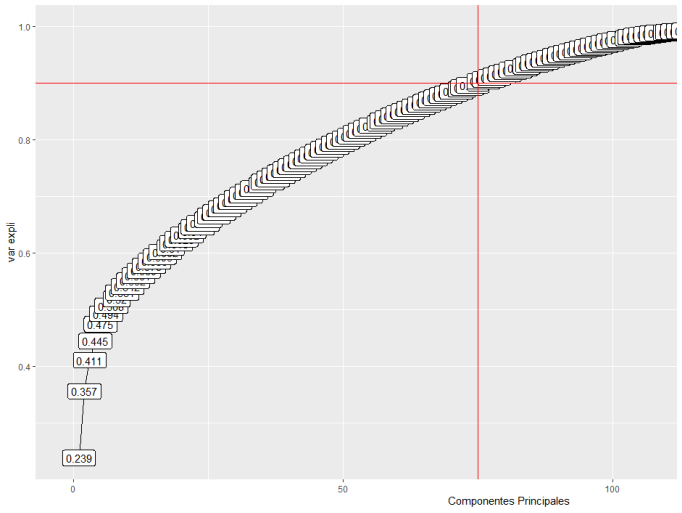

Ilustración 96 : variabilidad explicada acumulada

Para la selección del número de componentes principales debemos mirar las proporciones de varianza o variabilidad explicada por cada componente principal. En las *ILUSTRACIONES 95 Y 96* tenemos la variabilidad explicada por las componentes principales. En la ILUSTRACIÓN 95 tenemos la proporción de varianza explicada por cada componente. En esta ilustración podemos observar que ninguna componente llega a explicar el 25% de la variabilidad, siendo la primera componente la componente que más explica, comparándola con las demás. Observando la *ILUSTRACIÓN 96* podemos ver la proporción de varianza acumulada. Dado que tenemos muchas etiquetas, los valores de las proporciones no se ven correctamente. Deseamos que al menos un 90% de la variabilidad quede explicada y, en esta ilustración, en la línea roja horizontal, vemos donde correspondería este 90%. Hemos trazado una línea vertical para cruzarse con

dicha línea horizontal y, así, ver el número de componentes necesarias para este 90% explicado. Esta explicación se consigue con 75 componentes.

4.2.1.2 Análisis clúster de todas las variables

Para la selección del número óptimo de clúster usaré la función "fviz_nbclust", de la misma manera que utilizaba en otras ocasiones. En la *ILUSTRACIÓN 97* tenemos la curva del número de clúster, donde podemos ver que se produce una curva en el 4, un "codo".

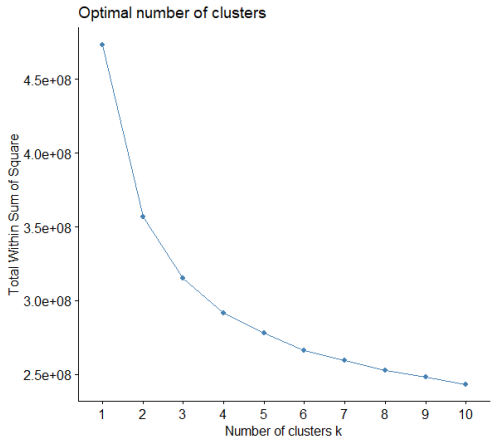

Ilustración 97: número óptimo de cluster análisis de todas las variables grupo menos de 20000.

4.2.1.2.1 Cluster jerárquico

Para un cluster jerárquico debemos mirar primero que método tiene la correlación más alta en la matriz de disimilitudes en los distintos métodos. El método que tenga la correlación más alta lo usaremos como el método más adecuado para este análisis. En la *Tabla 9* podemos ver estas correlaciones, la más alta se encuentra en el método Average.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,4778178	0,831835	0,7733857	0,3765498	0,7278835	0,7518443

Tabla 9: tabla de correlación de distintos métodos.

El método del average solo nos ha creado 4 grupos. Los grupos que tenemos son los siguientes: 3 grupos formados por un solo municipio y, después, un grupo conformado por los demás municipios. Los municipios que se encuentran formando cada uno un único cluster son Jaramillo Quemado, Fuente El Olmo De Fuentidueña y Beraton.

El municipio de Jaramillo Quemado tiene el código INE 09184 y se encuentra en la provincia de Burgos. La población media de este municipio es de 8,826087 personas. Su tasa de mortalidad es 0 en la mayoría de los años, pero en el año 2015 alcanza un valor muy alto, ya que se sitúa en 400 por cada 1.000 habitantes. En cuanto a la migración en este pueblo no se ha producido casi, ya que la mayoría de los años su tasa es 0. Podemos destacar que la tasa de nacimientos es 0. En este pueblo la tasa de varones es de 700 por cada 1000 frente a la de mujeres, que alcanza 300 por 1.000 habitantes.

El municipio de Fuente el Olmo de Fuentidueña tiene el código INE 40083 y se encuentra en la provincia de Segovia. La población media por año es de 152,4348. La tasa de población de derecho varón media está sobre los 400 por cada 1000 habitantes frente la de mujeres, que está sobre los 600. La tasa de mujeres en edad reproductiva se sitúa en 320 por cada 1000, aunque la tasa de nacimientos es de 6 por cada 1.000 habitantes, siendo en muchos casos 0. La tasa de defunción se sitúa en 17 por cada 1.000 habitantes. En cuanto a las tasas de migración se sitúan en torno a 70 por 1.000 habitantes, pero destaca las inmigraciones en 2009, con un valor de 920 por cada 1.000 habitantes.

El municipio de SAN CRISTOBAL DE SEGOVIA tiene código INE 40906 y se encuentra en la provincia de Segovia. La población media por año es de 2.240 habitantes. La población en este municipio ha ido en aumento a lo largo de los periodos ya que es un municipio de nueva creación; dada esta creación para los primeros periodos sus tasas son 0 ya que el municipio no existía. Las tasas de población de derecho varón y mujer se encuentran ambas en los 500 por cada 1.000 habitantes. Las tasas de mujeres en edad reproductiva están en torno a los 280 por cada 1.000 habitantes; observándose una disminución a lo largo de los periodos. Sus tasas de natalidad han ido decreciendo a lo largo de los periodos ya que comienza con tasas de 34 por cada 1.000 habitantes y en el año 2019 su tasa tan solo es de 6 por cada 1.000 habitantes. Las tasas de defunción son de en torno a 2 por cada 1.000 habitantes, lo que son unas tasas bajas en comparación con la tasa de nacimientos. En la tasa de inmigración varían entre 40 y 100 por cada 1.000 habitantes y las tasas de emigraciones están en torno a 50 por cada 1.000 habitantes.

El grupo de los demás municipios presentan una población media de 542,758 habitantes. Tasas de varones y mujeres se sitúan en los 500 por cada 1.000 habitantes. La tasa de mujeres ene dad reproductiva en los 165 por cada 1.000 habitantes. Las emigraciones se sitúan en 21 por cada 1.000 habitantes y las inmigraciones en 6 por cada 1.000 habitantes. En cuanto a la tasa de nacimientos en muchos casos es 0 y, la tasa de defunciones se sitúa en los 160 por cada 1.000 habitantes.

4.2.1.2.2 Cluster no jerárquico

Para el cluster no jerárquico usaremos las k-medias, mediante la función kmeans. Como número de grupos tomaremos 4, dado que es el número que determinamos antes como adecuado. En este caso, como el cluster tiene muchos individuos, el dendrograma no se ve correctamente, con lo que representamos los grupos del cluster con un mapa.

En la *ILUSTRACIÓN 98* podemos ver los grupos formados por las k-medias. Denominaremos grupo 1 al color verde, el cual lo forman 645 municipios y tiene una población media de 360,0278 habitantes. En este grupo la población media ha ido descendiendo a lo largo de los años. Este grupo 1 posee unas tasas de defunción por cada 1000 habitantes bajas, de en torno a 10 por cada 1000; en cuanto a sus tasas de nacimientos podemos decir que son muy bajas, de 5 por cada 1000, siendo en muchos municipios, para varios periodos, de 0. Si hablamos de las tasas de emigración e inmigración nos encontramos en la misma situación, dado que, en general, se sitúan en torno al 15 por cada 1000, salvo excepción de algún municipio en algún año, donde encontramos valores tan distintos como 0 o 120 por cada 1000. La tasa de mujeres en edad reproductiva se sitúa en torno a 150 por cada 1000 individuos, observándose una ligera bajada en las tasas desde los primeros periodos a los últimos. Las tasas de población de derecho varón y mujer se encuentran ambas cercanas a 500 por cada 1000,

lo que indica que, en estos municipios, no existe mayor población varón o mujer, que están equilibrados en ese sentido entre ambos.

El grupo 2 lo forman los municipios que presentan el color amarillo y, el número de municipios en este son 241, lo cual lo convierte en el grupo menos numeroso. La población media de estos municipios se situaría en los 81,21162 individuos. Las tasas de defunciones no son elevadas en este grupo, se sitúan en torno al 25 por cada 1000, siendo 0 en muchos municipios en algunos periodos. Podemos considerarla muy elevada en comparación con la tasa de nacimiento, ya que en la mayoría de los periodos en la mayoría de los municipios es 0. Las tasas de migración se sitúan en torno a 40 por cada 1000, siendo, en algunos casos, para algunos periodos 0. En estos municipios la tasa de población de derecho varón se sitúa sobre los 650 por cada 1000; es más elevada que la de mujer, con lo que podemos decir que en estos municipios existen más varones que mujeres; además, la tasa de mujeres en edad reproductiva es muy baja.

En el grupo 3 tenemos los municipios azul. Este grupo lo conforman 622 municipios. La población media de estos municipios es de 1.290,311. La tasa de mujeres en edad reproductiva se encuentra en torno a 170 por cada 1000, donde podemos apreciar una disminución en las tasas desde los primeros periodos a los últimos. Las tasas de población de derecho varón y mujer se encuentran un poco por encima de 500 por cada 1000 habitantes, con lo que, más o menos, esta equilibrada la población entre mujeres y hombres. La tasa de emigración se sitúa en torno a 45 por cada 1000, mientras que la de inmigración es de 60 por cada 1000, aproximadamente, siendo en muchas ocasiones más elevada. La tasa de defunción se sitúa en torno a 15 por cada 1000 y la de nacimientos se sitúa en torno a 6 por cada 1000 habitantes.

El grupo 4 lo encontramos en los municipios de color rosa, este grupo es el más numeroso y lo forman 724 municipios. La población media por municipio es de 218,0186 habitantes. La tasa de población de derecho varón se sitúa sobre los 570 por cada 1000. Esto indica que existe algo más de población de derecho varón que de mujer en estos municipios. La tasa de emigración se sitúa en torno a los 50 por cada 1000, sobrepasando los 100 en algunos periodos para algún municipio y, la de inmigración entre 0 y 40 por cada 1000. La tasa de mortalidad se sitúa sobre los 20 por cada 1000, siendo 0 en algunos municipios para varios periodos. Podemos destacar de este grupo que la tasa de natalidad, en general, se sitúa en torno a 0.

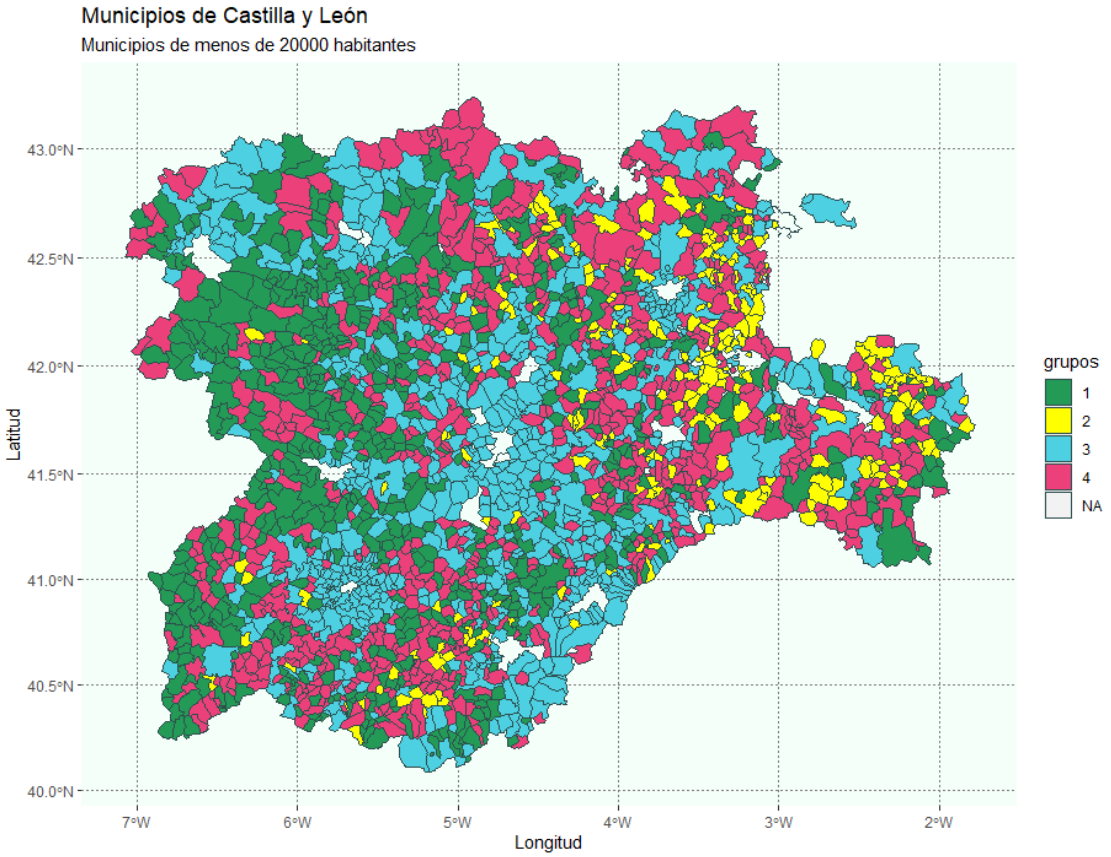

Ilustración 98: mapa de cluster de las k-medias análisis de todas las variables grupo menos de 20000.

4.2.1.3 Análisis clúster usando las componentes principales

Para el análisis cluster usando las componentes principales, realizamos el análisis de la misma manera que hemos hecho en el apartado anterior. El número de componentes principales que usaremos serían las seleccionadas en el apartado 4.2.1.2, este número es 75 componentes. Para seleccionar el número óptimo de grupos miramos la *ILUSTRACIÓN 103*. En esta ilustración podemos ver que el número óptimo de cluster sería 4.

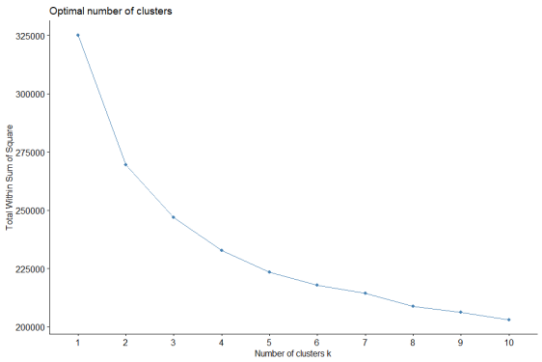

Ilustración 99: curva de numero óptimo de clúster usando PCA análisis de todas las variables grupo menos de 20000.

4.2.1.3.1 Cluster jerárquico

Para el cluster jerárquico usaremos hclust debemos saber el método que tiene mayor correlación en la matriz de disimilitudes de los grupos. Cogeremos como mejor método el de la correlación más alta. En la *Tabla 10* podemos ver que el valor más alto en la correlación lo tiene el método Average.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,6049932	0,886807	0,8443148	0,301088	0,8145584	0, 8367404

Tabla 10: tabla de correlaciones

Los grupos que crea son 4. Tres de ellos solo tiene un único municipio por grupo y, el cuarto grupo tiene el resto de los municipios. Los municipios que están solos cada uno en único cluster son Jaramillo Quemado, Villarmentero de Campos y Gormaz.

Jaramillo Quemado que tiene el código INE 09184, es un municipio que se encontraba formado un único grupo también en 4.2.1.2.1, apartado donde lo describimos.

El municipio de Villarmentero de Campos tiene un código INE 34230 y se encuentra en la provincia de Palencia. Tiene una población media de 15,6087. Su tasa de nacimientos es 0 en todos los periodos menos en el año 2015. La tasa de mujeres en edad reproductiva ha ido variando según periodos, entre 1996 y 2004 se encuentra entre 62 y 133 por cada 1.000 habitantes, para la franja de 2005 a 2009 las tasas son 0 y desde 2010 las tasas han ido oscilando con un valor máximo en el año 2015 con un valor de 238 por cada 1.000 habitantes. Las tasa de población de derecho varón se sitúa en torno a los 700 por cada 1.000 habitantes, dejando así la de mujeres sobre los 200. Con estas tasas podemos ver que en este municipio la mayoría de la población son hombres. En este municipio la tasa de defunciones se sitúa en 0 en la mayoría de los años. La tasa de inmigración son 0 en la mayoría de los periodos hasta 2009; desde ese año vemos unas tasas entre 86 y casi 300 por cada 1.000 habitantes. Las tasas de emigración son 0 la mayoría de los periodos; el año en el observamos el mayor valor es 2017 con una tasa de 391 por cada 1.000 habitantes.

El municipio de Gormaz tiene un código INE 42097 y se encuentra en provincia de Soria. Tiene una población media de 21,17391. Su tasa de nacimientos es 0 en la mayoría de los años. La tasa de mujeres en edad reproductiva se encuentra en 70 por 1.000 habitantes y la tasa de varones y mujeres se encuentran en los 500 por cada 1.000 habitantes. La tasa de defunciones se sitúa en 0 en la mayoría de los periodos. La tasa de emigración se sitúa en 270 por 1.000 habitantes y la inmigración en 200 por cada 1.000 habitantes.

El grupo del resto de municipios tiene una población media de 543,8069 habitantes. La tasa de nacimientos se sitúa en 13 por cada 1.000 habitantes, siendo 0 en muchos periodos para muchos de los municipios. La tasa de mujeres en edad reproductiva se sitúa en 150 por 1.000 habitantes y la tasa de varones y mujeres se encuentran en los 500 por cada 1.000 habitantes. La tasa de defunción se sitúa en 27 por 1.000 habitantes. En cuanto a las variables de migración se sitúan en tasas de 40 por cada 1.000 habitantes.

4.2.1.3.2 Clasificación no jerárquica

Para la clasificación no jerárquica usaremos las k-medias con la función kmeans de R. En la *ILUSTRACIÓN 100* podemos ver los grupos. Los municipios que aparecen como NA son los de más de 20.000 habitantes. Vemos claramente que el color más numeroso es el rosa. A este grupo de color rosa lo denominaremos grupo 4. Este está formado por 834

municipios y la población media por municipio es de 536,9842 habitantes. Las tasas de nacimientos son muy bajas, con valores que oscilan entre 0 y 8 por cada 1.000 habitantes; siendo mayoritariamente 0. Esto contrasta con las tasas de defunción, las cuales alcanzan valores superiores a 30 por cada 1.000 habitantes en algunas ocasiones. Las tasas de emigración han ido oscilando en crecimiento y decrecimiento a lo largo de los periodos situándose en valores de 33 por cada 1.000 habitantes. Las tasas de inmigración en varios periodos para algunos municipios son 0, pero, en general, se sitúan en torno a 20 por cada 1.000 habitantes. En estos municipios, podemos decir que la proporción de mujer y hombres está, más o menos, equilibrada. En cuanto a la tasa de mujer en edad reproductiva vemos una ligera disminución entre los primeros periodos y los más actuales, situándose actualmente en valores de 130 por cada 1.000 habitantes.

El grupo 3 lo podemos ver en color azul. Este grupo está formado por 381 municipios y su población media es de 1.514,872 habitantes. Sus tasas de natalidad son cercanas a los 3 por cada 1.000 habitantes, pero son prácticamente 0 en algunos municipios para la mayoría de los periodos. Las tasa de defunciones en muchas ocasiones son 0, pero se sitúan en torno a 15 por cada 1000. Las tasa de emigración se sitúan en torno a 30 por cada 1000. Las tasas de inmigración han ido aumentando a lo largo de los periodos produciéndose en los últimos periodos valores muy elevados, se sitúan en general en los 45 por cada 1.000 habitantes. La proporción de mujeres y hombres es similar en estos municipios. En cuanto a las tasas de las mujeres en edad reproductiva se ve una ligera disminución, situándose en valores cercanos a 150 por cada 1.000 habitantes en los últimos periodos.

El grupo 2 lo podemos identificar en color amarillo. Este grupo lo forman 792 municipios. La población media por municipio es de 213,5591 habitantes. Las tasas de natalidad se sitúan entre 0 y 10 por cada 1.000 habitantes, las cuales son unas tasas bajas en comparación con las tasas de defunciones que están situadas entre 9 y 20 por cada 1.000 habitantes. Las tasas de inmigración están en torno entre 0 y 130 por cada 1.000 habitantes y las de emigración están en 30 por cada 1.000 habitantes. En cuanto a las tasas de población de derecho varón, observamos un aumento a lo largo de los años. Estas superan los 600 por cada 1000, por lo que podemos decir que hay algo más de población varón que mujer en los municipios.

El grupo 1 lo podemos identificar en color verde. Este es un grupo es el menos numeroso y está formado por 225 municipios, con una población media de 80,26261 habitantes. Las tasas de natalidad se sitúan en torno a 3 por cada 1000, pero son 0 en muchos municipios en varios periodos. Las tasas de defunciones rondan los 30 por cada 1.000 habitantes, aunque en ocasiones son 0, son unas tasas algo elevadas comparándolas con las de los nacimientos. Observamos un aumento en las tasas de emigración, a lo largo de los periodos; estas se sitúan en torno a 50 por cada 1.000 habitantes. En las tasas de inmigración, observamos, también, un aumento, sobrepasando incluso valores de 100 por cada 1.000 habitantes en algunos periodos. En estos municipios, podemos decir que la proporción de mujer y hombres no está equilibrada ya que a lo algo de los años ha ido aumentado la población de derecho varón; además las tasas de mujeres en edad reproductiva están en valores de 130 por cada 1000 habitantes pero en algunas ocasiones son muy inferiores.

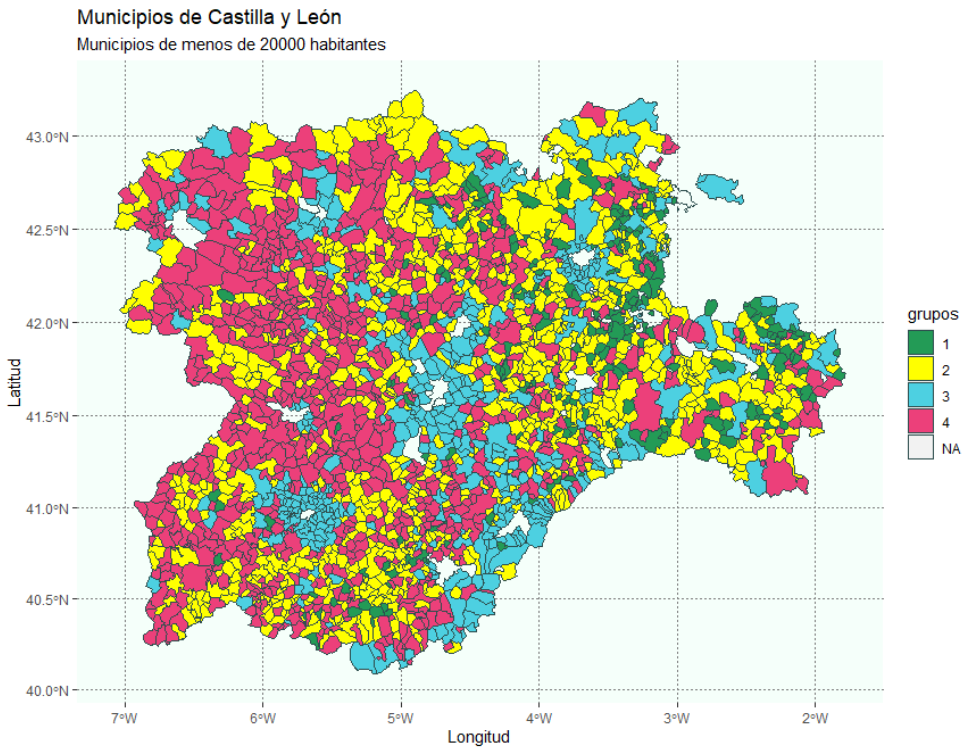

Ilustración 100: mapa de cluster de las k-medias usando PCA análisis de todas las variables grupo menos de 20000.

4.2.1.4 Multidimensional scaling

Para el multidimensional scaling como número óptimo de grupos usaremos el número de clústeres óptimos seleccionado en análisis clúster de todas las variables, este número serían 4 grupos. Para el multidimensional scaling, dada la cantidad de datos, usaremos un método métrico y representaremos en un mapa los distintos grupos que tenemos. En la *ILUSTRACIÓN 101* podemos ver una representación sobre un mapa de los grupos. Los municipios que aparecen como NA son los de más de 20.000 habitantes.

El grupo 1 es el de color verde. Es el grupo menos numeroso, únicamente lo conforman 239 municipios, su población media es de 77,71803 habitantes. Las tasas de defunciones oscilan entre 10 y 40 por cada 1000, pero en algunos periodos son 0. Estos valores contrastan con la tasa de nacimiento que, prácticamente, son 0. Las tasas de migración están entre 0 y 70 por cada 1000. En este caso la población de derecho varón es mayor que la de mujeres en los municipios, ya que la tasa de varones se sitúa sobrepasando los 700 por cada 1000.

El grupo 2 es de color amarillo. Cuenta con 619 municipios y su población media es de 1.328,443 habitantes. Las tasas de defunción se sitúan en valores algo superiores a 10 por cada 1000. Las tasas de nacimiento oscilan entre 0 y 7, lo que son unas tasas bajas. Las tasas de inmigración se sitúan entre 20 y 50 y las de emigración entre 30 y 70 por cada 1000. En este caso la población entre varones y mujeres esta equilibrada.

El grupo 3 es de color azul, lo forman 638 municipios y la población media por municipio es 331.7747. Las tasas de defunción oscilan entre 0 y 40 por cada 1000. Estos valores contrastan con las bajas tasas de nacimientos, que oscilan entre 0 y 6 por cada 1000, siendo 0 el valor más común. Las tasas de migración están en torno a 50 por cada 1000.

Existe algo más de población de derecho varón, ya que su tasa sobrepasa ligeramente los 500 por cada 1000.

El grupo 4 es de color rosa y está formado por 736 municipios. Este grupo es el más numeroso. La población media por municipio es de 216,9231 habitantes. En las tasas de defunción podemos ver un aumento a lo largo de los periodos, al final encontramos tasas sobre 25 por cada 1000. Las tasas de nacimientos se sitúan en 10 por cada 1000, siendo 0 en muchos casos. Las tasas de inmigración se sitúan sobre 10 y las de emigración sobre 20 por cada 1000. En este caso, la población de derecho varón es mayor que la de mujeres en los municipios, ya que la tasa de varones se sitúa sobrepasando los 600 por cada 1000.

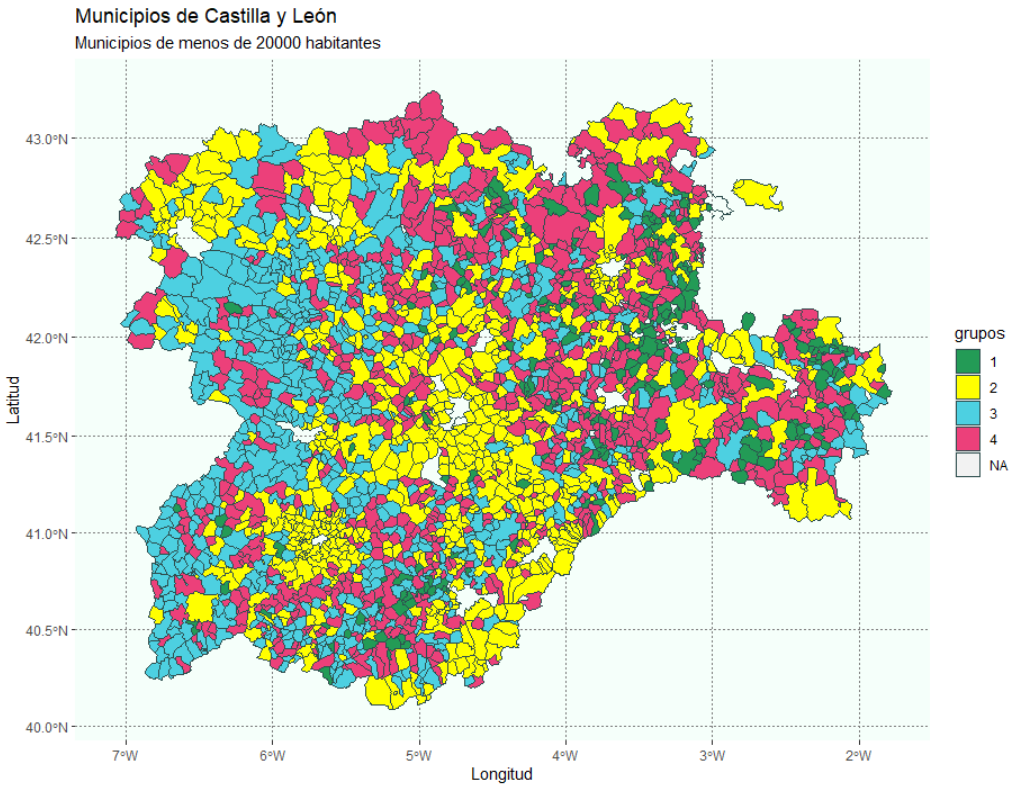

Ilustración 101: grupos de multidimensional scaling

4.2.1.5 Comparación de métodos

Los métodos crean grupos diversos. Los métodos que crean grupos separados por categorías que tiene más relación entre sí serían las k-medias, ya sea con uso de componentes o sin ellas, y el multidimensional scaling. Además, estos métodos proporcionan grupos, más o menos, del mismo número de municipios. Los métodos jerárquicos proporcionan grupos en los cuales un único municipio forma un grupo, es decir, varios grupos que contienen únicamente un municipio. Un método cluster jerárquico con todas las variables podemos ver que forman estos grupos de un único municipio los municipios de Fuente el Olmo de Fuentidueña, San Cristóbal de Segovia y Jaramillo Quemado. Mientras que si aplicamos primero las componentes principales y, después el cluster jerárquico los municipios obtenidos como únicos son Jaramillo Quemado, Villarmentero de Campos y Gormaz; podemos ver que son municipios distintos los que vemos formando un único grupo cada uno, a excepción de Jaramillo Quemado.

Podríamos escoger como mejor método el cluster no jerárquico usando las componentes principales, ya que reducimos la dimensionalidad y, además, proporcionamos unos grupos con relación entre grupos y diferencia con otros grupos.

4.2.2 Análisis de las variables de movimiento natural de la población

En este grupo se encuentran los nacimientos, las defunciones y las mujeres en edad reproductiva. Trabajaremos con las tasas para cada 1000 individuos, de la misma manera que realizábamos un conjunto con todas las variables para el 4.2.1. lo realizamos para solo este grupo de variables.

4.2.2.1 Componentes principales

En el análisis de las componentes principales se realiza de la misma manera que en otras ocasiones, el número de componentes principales que en este caso se realizan son 69.

En la *Ilustración 102* podemos ver un biplot de la representación de las 2 primeras componentes. Podemos ver que los municipios se representan en la 1 dimensión, ya que los vemos como una masa de horizontal donde esta solo distribuidos para valores de la 1 componente, para la segunda están todos en la línea del 0.

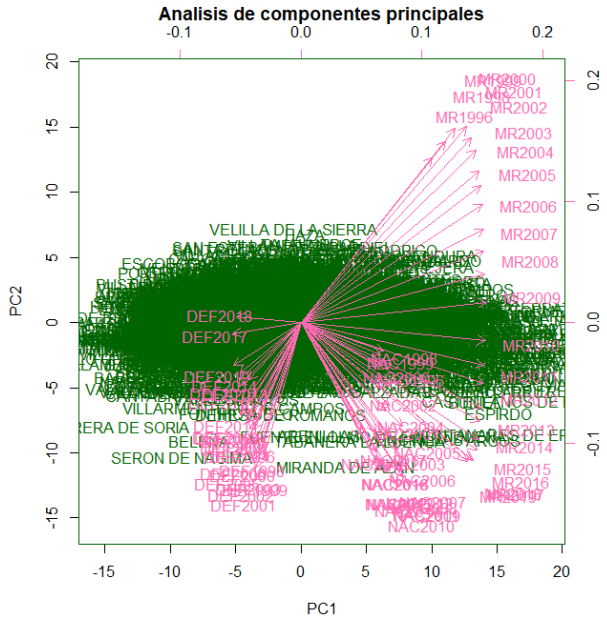

Ilustración 102: biplot de componentes principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

Para la selección del número de componentes principales para explicar al menos el 90% de la variabilidad debemos mirar la proporción de variabilidad explicada por cada componente. En la *Ilustración 103* podemos ver la proporción de variabilidad explicada por cada componente. Observamos que la primera componente es la que más explica pero que no llega a explicar el 35% de la variabilidad. En la Ilustración 104 podemos ver que al menos el 90% de la variabilidad explicada se alcanza con 40 componentes principales. La línea roja tenemos donde se encuentra el 90%, en la línea roja tenemos el valor de las componentes con el que se alcanzaría ese valor del 90%. Seleccionamos 40 componentes principales.

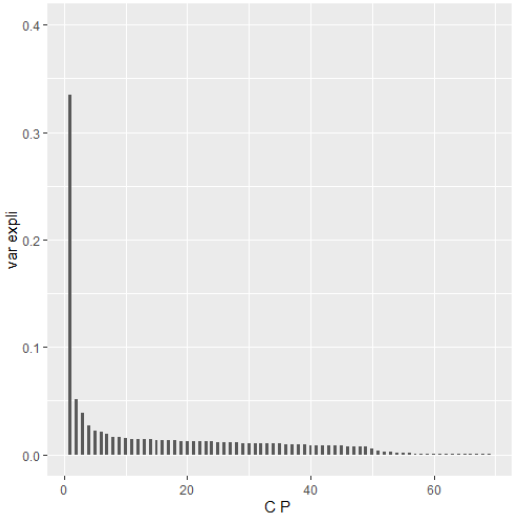

Ilustración 103: variabilidad explicada por cada componente principal principales para el grupo menos de 20000 habitantes y variables de movimiento natural

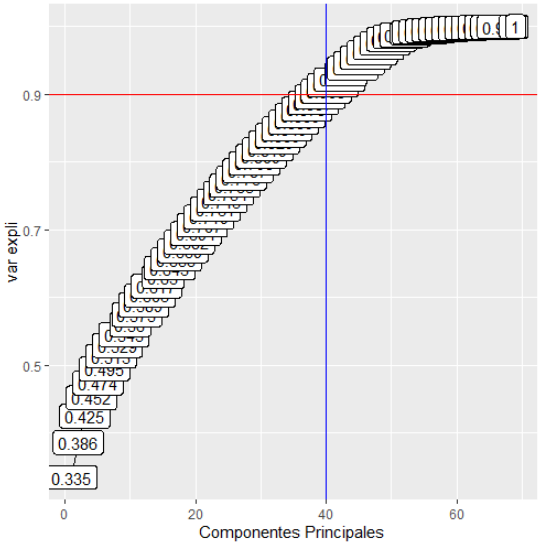

Ilustración 104: variabilidad explicada acumulada principales para el grupo menos de 20000 habitantes y variables de movimiento natural

4.2.2.2 Análisis clúster de todas las variables

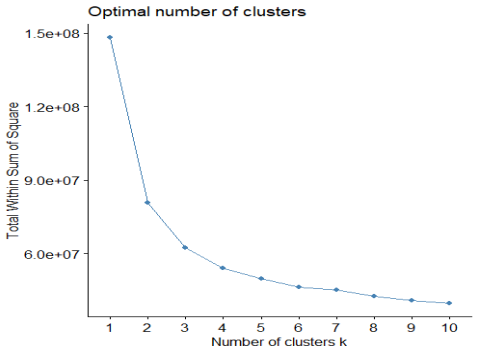

Ilustración 105: número óptimo de clúster principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

En la *Ilustración 105* podemos ver esta curva para la selección del número óptimo de clúster, que se realiza de la misma manera que en otros apartados. Observamos una especie de "codo" en el número 3 de clúster, en ese punto podemos decir que se produce el número más adecuado de clúster.

4.2.2.2.1 Cluster jerárquico

Para la elección del método más adecuado para calcular las disimilitudes entre clúster, elegiremos según la mayor correlación en la matriz de distancias de los métodos. Usaremos después la función hclust de R con el método seleccionado. En la *Tabla 11* podemos ver estas correlaciones de las matrices de disimilitudes entre grupos usando los distintos métodos. En esta tabla observamos que el método con mayor correlación es average linkage con lo que lo tomaremos como el método a usar.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0, 5684995	0,7096938	0,453051	0,5147426	0,4770978	0, 6445614

Tabla 11:tabla de correlaciones de métodos

Este método nos crea 3 grupos, 2 de ellos formados únicamente por un municipio y el otro formado por el resto de los municipios. Uno de los municipios que se encuentra solo es el municipio de Jaramillo Quemado tiene el código INE 09184 se encuentra en la provincia de Burgos. Este municipio tenemos descritas sus características en el apartado 4.2.1.2.1. El otro municipio que se encuentra en solitario es el municipio de Fuente el Olmo de Fuentidueña tiene el código INE 40083 se encuentra en la provincia de Segovia, este municipio tenemos descritas como son sus tasas en el apartado 4.2.1.2.1 . El grupo del resto de municipios contiene 2230 municipios y tiene una población media de 543,5191 habitantes. Sus tasas para este grupo de variables son: 170 por cada 1.000 habitantes para las mujeres en edad reproductiva. En cuanto a la tasa de defunciones en muchos casos es 0 pero se situaría sobre los 50 por cada 1.000 habitantes. La tasa de nacimientos se sitúa en entre 0 y 7 por cada 1.000 habitantes pero siendo 0 en muchos de los casos para varios periodos de años.

4.2.2.2.2 Cluster no jerárquico

Para el cluster no jerárquico usaremos el método de las k-medias con la función kmeans de R. Como número de grupos tomaremos 3 ya que son los que habíamos determinado como óptimo.

En la *Ilustración 106* podemos ver la representación de los grupos de las k-medias. Podemos distinguir 3 grupos con los colores diferentes. Los valores que pone como NA son los municipios de más de 20.000 habitantes.

Con color verde encontramos el grupo 1, este grupo tiene 1.059 municipios y su población media es de 340,7747 habitantes. La tasa de natalidad oscila entre 0 y 15 por cada 1.000 habitantes pero la mayoría de los casos es 0; solo se observan algunos valores con una tasa distinta de 0 para algunos municipios y periodos. Además se ve un descendimiento de los primeros periodos a los últimos; con esto podemos decir que la natalidad en estos municipios es casi nula. La tasa de mujeres en edad reproductiva se sitúa en 118 por cada 1.000 habitantes. Los de las tasas de defunción se sitúan en los 27 por cada 1.000 habitantes, lo que es una tasa elevada dada la baja natalidad que se ve.

Con el color amarillo tenemos el grupo 2, podemos ver que es un color numeroso sobre todo en la frontera con otras comunidades y Portugal. Este grupo lo conforman 666 municipios y su población media es de 156,2432 habitantes. Las tasas de nacimientos para estos municipios son 0 y excepcionalmente algún municipio tiene una tasa de 10 por cada 1.000 habitantes, lo que sigue siendo una tasa baja. Las tasas de mujeres en edad reproductiva se sitúan en 120 por cada 1.000 habitantes. Las tasas de defunciones se sitúan por encima de 25 por cada 1.000 individuos. A la vista de estas tasas podíamos ver que estos municipios tienen una población muy baja y envejecida.

Con color azul tenemos el grupo 3, es un color que podemos ver alrededor de los que son NA, es decir, de los de más 20.000 habitantes. Este grupo lo conforman 507 municipios y la población media por municipio es de 1.473,906 habitantes. La tasa de natalidad se sitúa en 9 por cada 1.000 habitantes, en algunas ocasiones la tasa es 0 pero no tenemos tantas ocasiones como en otros grupos. La tasa de mujeres en edad reproductiva es de 200 por cada 1.000 habitantes. La tasa de defunciones es de 8 por cada 1.000 habitantes, lo que no es una tasa muy elevada.

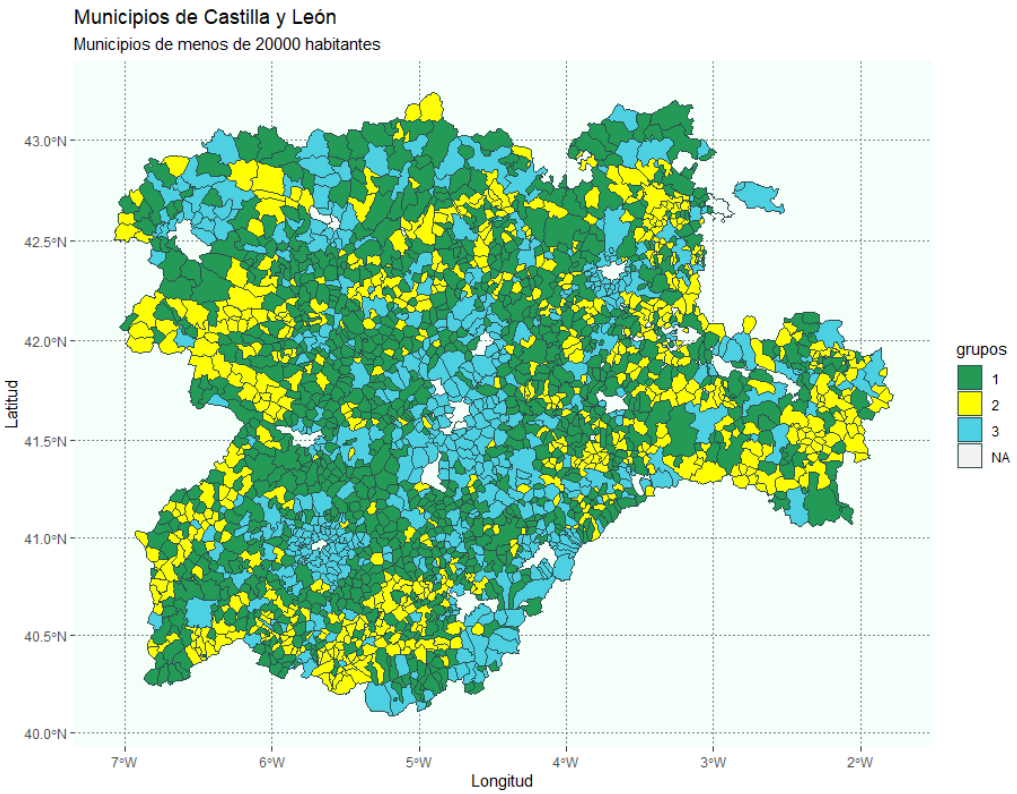

Ilustración 106: mapa de los grupos de las k-medias principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

4.2.2.3 Análisis clúster usando las componentes principales

Para este análisis realizaremos lo mismo que en el apartado anterior pero ahora usando las componentes principales. Habíamos visto que el número de componentes principales para explicar al menos un 90% de la variabilidad es de 40 componentes.

Para la selección del número de grupos debemos mirar la *Ilustración 109*. En esta ilustración podemos ver que el número optimo seria 3, ya que es donde se produce un "codo" en la curva.

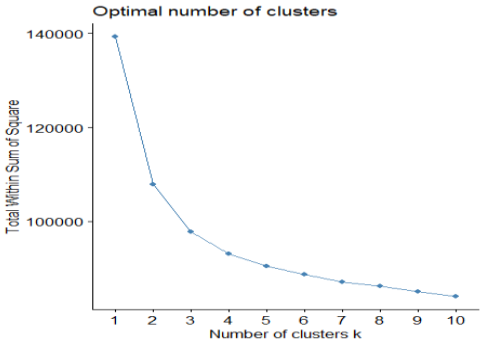

Ilustración 107: número óptimo de clúster principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

4.2.2.3.1 Cluster jerárquico

En el cluster jerárquico debemos seleccionar le método más adecuado para utilizar con la función *hclust* de R. Para la selección del método más adecuado miraremos la correlación de las matrices de disimilaridades entre clúster calculadas con los distintos métodos. En la *Tabla 12* podemos ver esta información de las correlaciones. Observamos que la correlación más alta se encuentra en el método *average linkage*.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,6625177	0,8055946	0,7180806	0,0.3302024	0,7037751	0,7664272

Tabla 12: tabla de correlaciones

Los grupos que se producen son los mismos analizados en el apartado anterior el 4.2.2.2.1 .Dos grupos formados solo por un municipio , Jaramillo Quemado y Fuente el Olmo de Fuentidueña y después un grupo con el resto de los municipios.

4.2.2.3.2 Cluster no jerárquico

Para el cluster no jerárquico usaremos las k-medias con la función kmeans de R, como número de grupos tomaremos el que habíamos seleccionado como adecuado. Este número de clúster es 3. En la *Ilustración 108* podemos ver la representación de los grupos de los cluster por el método de las k-medias. Los municipios que aparecen como NA son los de más de 20.000 habitantes.

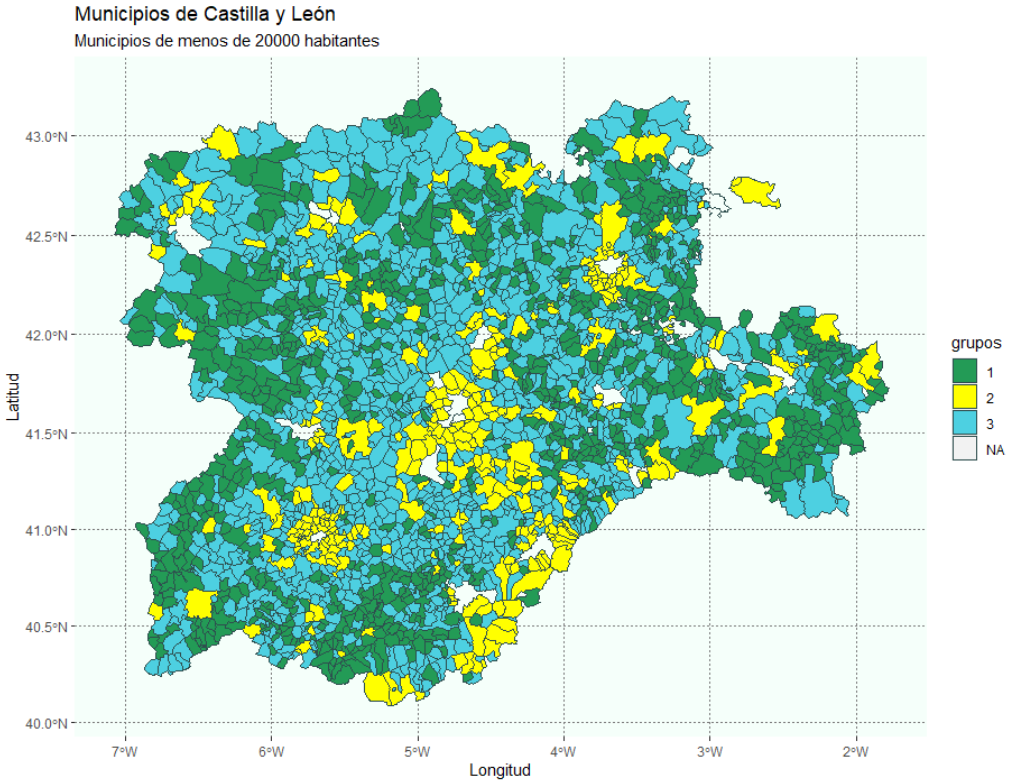

Ilustración 108: mapa de representación de grupos de k-medias principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

De los grupos que vemos en la *Ilustración 108* el grupo 1 lo forman 851 municipios. Este grupo es el que vemos en verde, podemos ver que se encuentra principalmente en zonas fronterizas con otras comunidades o Portugal. Tiene una población media por municipio de 177,9188 habitantes. Las poblaciones medias por municipio han ido disminuyendo ligeramente a lo largo de los años. Las tasas de nacimientos en este grupo son muy bajas, oscilan entre 0 y 6 por cada 1.000 habitantes; siendo 0 el valor más predominante. Las tasas de mujeres en edad productiva se sitúan en 120 por cada 1.000 habitantes; además vemos una ligera disminución de estas tasas a lo largo de los periodos. Las tasas de defunciones se sitúan en 32 por cada 1.000 habitantes, siendo más elevadas en

algunos periodos para algunos municipios. Estas tasas de defunción podemos considerarlas elevadas en comparación con las de nacimiento.

El grupo 2 lo conforman los municipios en el color amarillo y podemos ver que se encuentran alrededor de los municipios de más de 20.000 habitantes. Este grupo lo forman 312 municipios siendo el menos numeroso, su población media por municipio es de 1.969,825 habitantes. Esta población media ha ido aumentando a lo largo de los años. Su tasa de nacimientos se sitúa en 9 por cada 1.000 habitantes. Las tasas medias de nacimientos crecieron hasta 2008, año desde el cual observamos un ligero descendimiento en los valores. La tasa de mujeres en edad reproductiva observamos esta misma idea, un aumento hasta 2008 y después un decrecimiento; las tasas se sitúan en 220 por cada 1.000 habitantes. La tasa de defunciones es de media de 10 por cada 1.000 habitantes.

El grupo 3 lo conforman los municipios de azul, podemos ver que es el color más abundante. Este grupo lo forman 1.069 municipios y su población media por municipio es de 417,4129 habitantes. Sus tasas de nacimientos oscilan entre 0 y 4 por cada 1.000 habitantes, siendo la mayoría 0. Sus tasas de mujeres en edad reproductiva están en los 170 por cada 1.000 habitantes y la tasa de defunciones entre 0 y 31 por cada 1.000 habitantes, siendo esta última una tasa bastante elevada en comparación con la de nacimientos. Además la tasa media de defunciones ha ido en aumento a lo largo de los periodos.

4.2.2.4 Multidimensional scaling

Para el multidimensional scaling dada la cantidad de datos usaremos un método métrico y representaremos en un mapa los distintos grupos que tenemos así como el grafico de grupos tradicional. En el multidimensional scaling como número óptimo de grupos usaremos el número de clústeres óptimos seleccionado en análisis clúster de todas las variables, este número serian 3.

En las *Ilustraciones 109 y 110* podemos ver los 3 grupos que tenemos, en una representación más clasica y sobre un mapa . Los valores que salen como NA son los municipios de más de 20.000 habitantes, que hemos analizado anteriormente.

En color amarillo podemos ver el grupo 2 este grupo posee 1.045 individuos, él más numeroso, y su población media es de 344,772. Sus tasas de defunciones se sitúan entre 0 y 50 por cada 1.000 individuos, esto contrasta con sus tasas de natalidad, ya que son prácticamente cero en todos los periodos para todos los municipios. En cuanto a la tasa de mujeres en edad reproductiva podemos ver que se encuentra en torno a en torno a 150 por cada 1.000 habitantes; además vemos un descendimiento desde los primeros periodos a los últimos.

El grupo 2 lo podemos observar en color verde; es un grupo predominante alrededor de los municipios de más de 20.000 habitantes. Este grupo lo forman 490 municipios y su población media por municipio es de 1.516,779. En cuanto a sus tasas de nacimiento podemos observar que en los primeros períodos existen unas tasas debes por cada 1.000 habitantes, frente a los últimos periodos que presentan tasas tan solo de 7 por cada 1.000 habitantes; estas tasas son bajas para ambos periodos. En cuanto a las tasas de mujeres en edad reproductiva observamos una gran diferencia entre los años. Los primeros años presentaban tasas de 240 por cada 1.000 habitantes y los últimos períodos tan solo de tan solo de 150 por cada 1.000 habitantes; esto es una disminución

bastante importante. Tasas de defunción se encuentran en torno a 20 por cada 1.000 habitantes; existen algunos períodos para algunos municipios que estas tasas son cero.

Grupo 3 lo podemos observar en color azul este grupo lo forman 697 municipios y la población media por municipio es de 156,0338. Las tasas de natalidad de estos municipios contrastan con las tasas de defunción; ya que las primeras tienen valor de prácticamente cero y las otras alcanzan valores de 37 por cada 1.000 habitantes. Las tasas de mujeres en edad reproductiva son muy bajas ya que solo alcanzan 75 por cada 1.000 habitantes en la mayoría de los municipios. Estas características comunes en municipios en los cuales la población está envejecida.

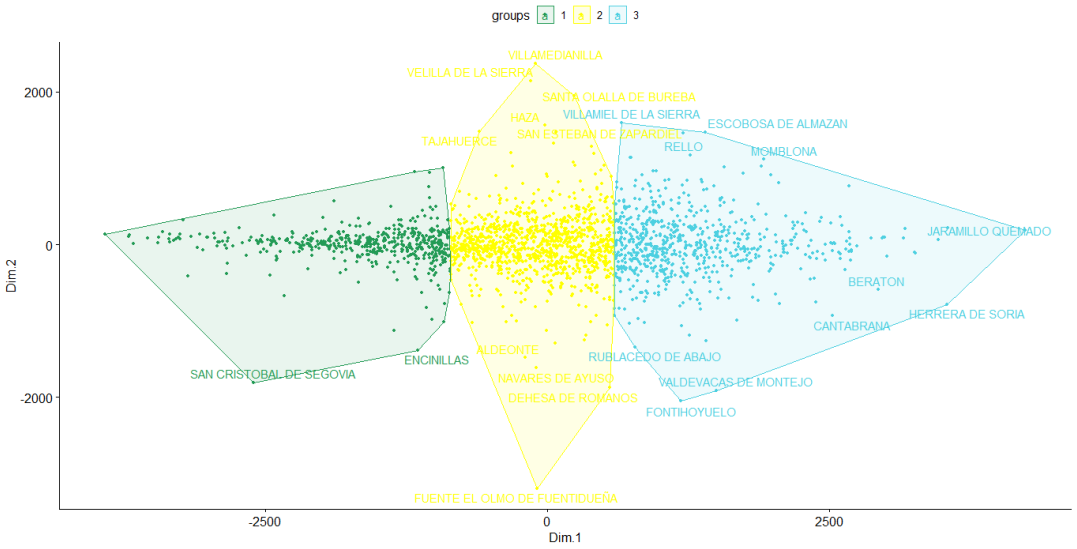

Ilustración 109: representación clasica de los grupos principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

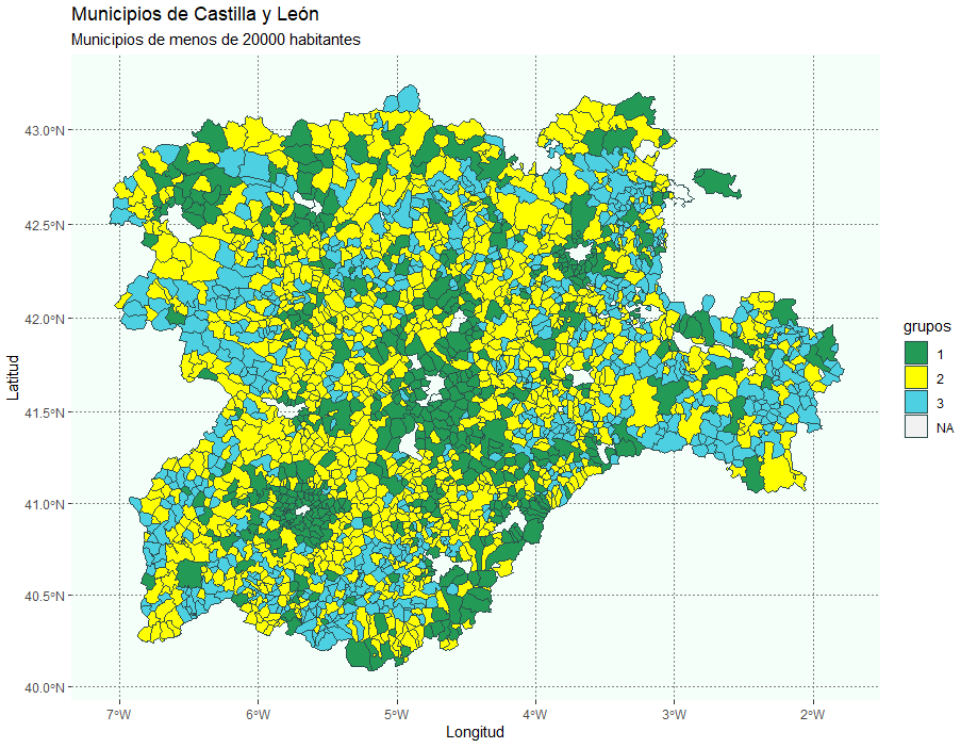

Ilustración 110: mapa de grupos de multidimensional scaling principales para el grupo menos de 20000 habitantes y variables de movimiento natural.

4.2.2.5 Comparación de métodos.

En los métodos como número óptimo de grupos es 3. En estas clasificaciones predominan un grupo de mayor tamaño y dos de un tamaño similar, eso lo podemos ver en la clasificación de 4.2.2.2.2. Grupo menos numeroso solo está reformado por municipios que se encuentran alrededor de los de más de 20.000 habitantes; además en este grupo tiene municipios con población media en torno a 1.500 habitantes. Estos municipios tienen como características que sus tasas de natalidad no son cero y su mortalidad no es tan elevada; además de una población media que ha aumentado a lo largo de los periodos. El grupo más numeroso suele estar formado por municipios de pequeño tamaño en cuanto a número de habitantes su característica principal es unas tasas de natalidad prácticamente nulas. El grupo que denominaremos intermedio en cuanto a número de municipios suele ser uno en el cual las tasas de defunciones son bastante elevadas en comparación con las de natalidad. Los métodos que proporcionan estos tipos de análisis; es decir estos grupos, son métodos como el clúster no jerárquico o el multidimensionales scaling. Dada la utilidad de las componentes principales para la reducción de la dimensionalidad podemos tomar como el mejor método el análisis cluster no jerárquico, k-medias.

4.2.3 Análisis de las variables de estadísticas de población

Este grupo lo forman las variables las variables de población de derecho varón y población de derecho mujer. En este grupo de variables trataremos los datos calculados en tasa por cada 1000 habitantes, es decir, que realizaremos los análisis con las tasas por cada 1000 habitantes de las variables.

4.2.3.1 Componentes principales

El análisis de componentes realiza de la misma manera que en otros apartados, en este caso realizan 45 componentes. En la *ilustración 111* podemos ver los *principal components scores* que es el valor de cada componente principal en cada municipio.

```
> head(pca2$x)
                              PC1        PC2          PC3          PC4         PC5         PC6
ADANERO                 0.2924082  1.2732461 -2.956631555 -0.009674917 -0.37952526 -0.34087367
ADRADA (LA)             2.2114692  0.9655640  0.001793914  0.590462880 -0.04970137 -0.09523141
ALBORNOS               -1.2981300  3.1634336 -0.668751947  0.047757812 -0.40848772 -0.40306995
ALDEANUEVA DE SANTA CRUZ -1.5320063  0.7812823  1.698888765  1.352826266  0.04181563 -0.04684596
ALDEASECA               6.0748981  2.2958494 -0.869546201 -0.527702473  0.23673382  0.19829522
ALDEHUELA (LA)          5.6019645 -0.5051874 -0.429682396  0.249596263 -0.56877411 -0.57959147
                              PC7        PC8          PC9         PC10         PC11
```

Ilustración 111: primeras componentes para primeros municipios para el grupo menos de 20000 habitantes y variables estadísticas de población.

En la *Ilustración 112* podemos ver que los municipios tienen valores más altos en la primera componente normalmente, dado que la masa de punto se sitúa en torno a la horizontal del 0. Podemos ver que las mujeres se sitúan en la primera componente en valores positivos y los años más actuales se sitúan en valores positivos de la segunda componente. En cuanto a los varones se sitúan en valores negativos de la primera componente y los años más actuales sucede al contrario que las mujeres, los más actuales se sitúan en valores negativos de la segunda componente.

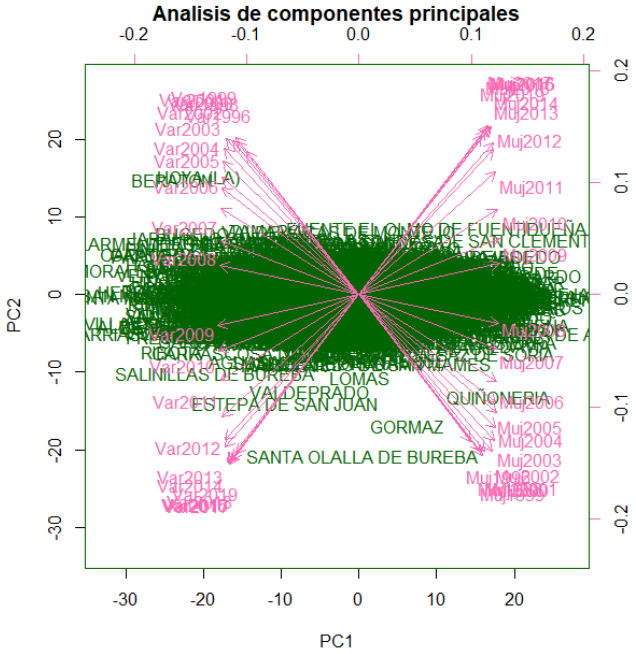

Ilustración 112:biplot de las componentes principales para el grupo menos de 20000 habitantes y variables estadísticas de población.

Para la selección del número adecuado de componentes principales miramos las *Ilustraciones 113 y 114* donde tenemos la variabilidad explicada según las componentes principales. En la *Ilustración 113* podemos ver la proporción de variabilidad explicada por cada componente. En la *Ilustración 113* destaca la primera componente principal, ya que es muy superior al resto en cuanto a la variabilidad que explica. Aunque a pesar de este hecho no llega a alcanzar el 75% explicado y las demás componentes ni si quiera alcanzan el 20% cada una de ellas. En la *Ilustración 114* podemos ver la curva de la variabilidad explicada acumulada. Observando estas dos Ilustraciones podemos ver que al menos 90% de la variabilidad explicada se produce con 3 componentes principales.

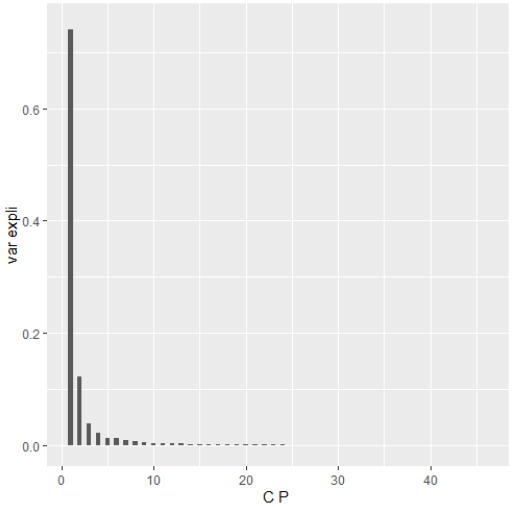

Ilustración 113: proporción de varianza explicada por cada componente para el grupo menos de 20000 habitantes y variables estadísticas de población.

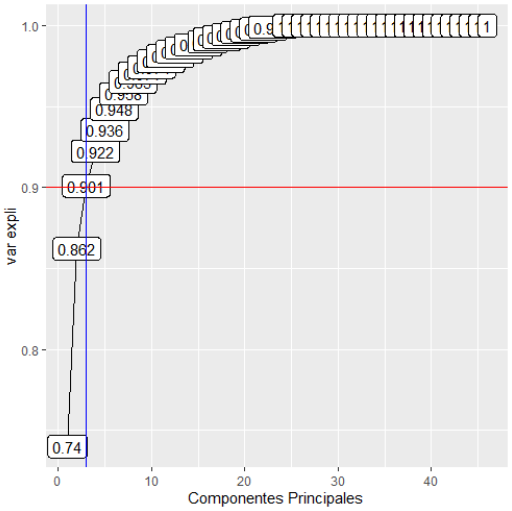

Ilustración 114: varianza explicada acumulada para el grupo menos de 20000 habitantes y variables estadísticas de población.

4.2.3.2 Análisis clúster de todas las variables

Para la selección del número de cluster debemos las variables normalizadas para poder usar la función "fviz_nbclust", de la misma manera que lo teníamos en otras ocasiones. El número optimo es 3, ya que en la *Ilustración 115* podemos ver que se produce un "codo" en la curva.

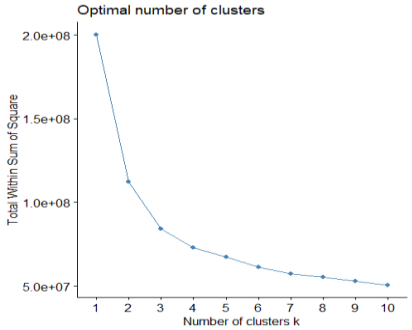

Ilustración 115: número óptimo de cluster

4.2.3.2.1 Cluster jerárquico

Para el cluster jerárquico usaremos la función hclust de R. Esta función debemos escoger el método que queremos para el cálculo de las disimilitudes entre cluster. Para escoger el método calcularemos las correlaciones de las matrices de disimilitudes según cada método; escogeremos como mejor método el que tenga la correlación más elevada. Mirar la *Tabla 13* podemos ver que el método con mas correlación es average.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0, 37899	0,7999523	0,6851932	0,4693664	0,6864599	0, 7975767

Tabla 13: tabla de correlaciones de métodos

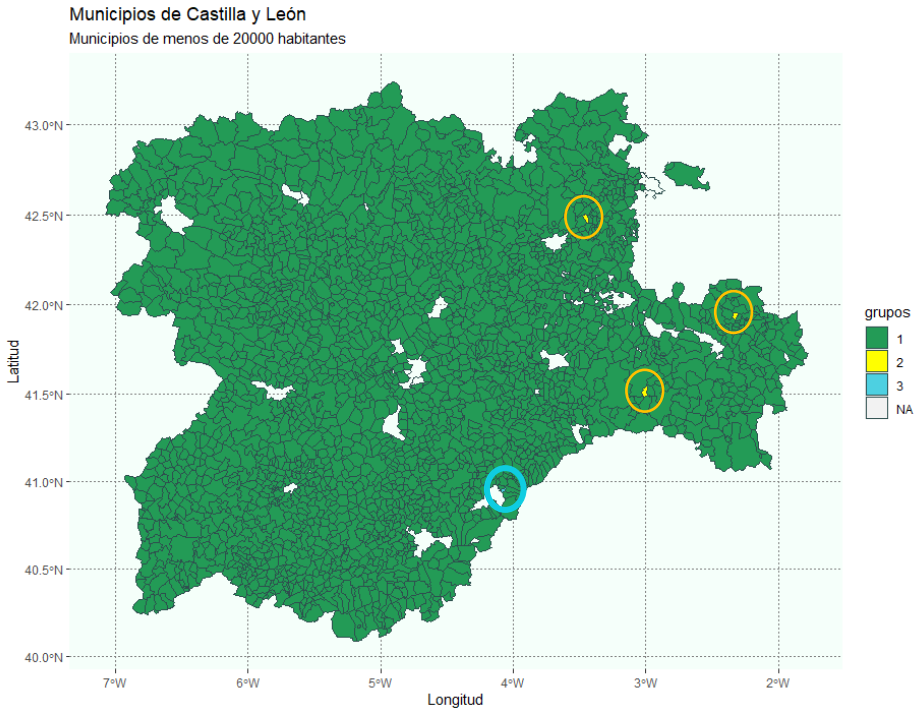

Ilustración 116: mapa de representación de grupos con el método de average para el grupo menos de 20000 habitantes y variables estadísticas de población.

En la *Ilustración 116* podemos ver los distintos grupos que nos ha creado el análisis cluster, a simple vista vemos que el grupo más numeroso es el 1, color verde. Los valores que vemos como Na son los municipios de más de 20.000 habitantes.

El grupo 1 lo conforman 2228 municipios y lo podemos ver en color verde. Su población media es de 543,0432. En estos municipios la tasa media de población de derecho varón se sitúa en los 525 por cada 1000. Esto indica que existe más población de derecho varón que mujer en los municipios pero no es una tasa elevada, es decir, podemos decir que existen más varones pero no muchos más.

El grupo 2 lo podemos ver en color amarillo, además tenemos los círculos amarillos para identificarlos mejor. Lo conforman con 3 municipios Santa Olalla de Bureba con código INE 09354, Estepa de San Juan con código INE 42082 y Gormaz con código INE 42097. La población media es de 22,84058 habitantes. En este grupo las tasas de población de derecho varón han ido aumentando a lo largo de los periodos, como son tan pocos habitantes de media un aumento de un varón supone una subida grande de la tasa. Las tasas se sitúan en los últimos periodos en valores cercanos a los 800 por cada 1.000 habitantes. Esta tasa nos indica que la población de derecho varón sobre la de mujer es muy superior.

El grupo 3 con 1 municipios y lo podemos encontrar en color azul, además de un círculo a su alrededor para hacer más fácil su identificación. Este municipio es San Cristóbal de Segovia tiene código INE 40906. Este municipio ha tenido un gran aumento de su población ya que se apareció nuevo y ha ido creciendo su población. En cuanto a su población se encuentra equilibrada entre mujeres y hombres.

4.2.3.2.2 Cluster no jerárquico

Para el cluster no jerárquico usamos las k-medias mediante los kmeans de R. Como número de grupos tomamos 3.

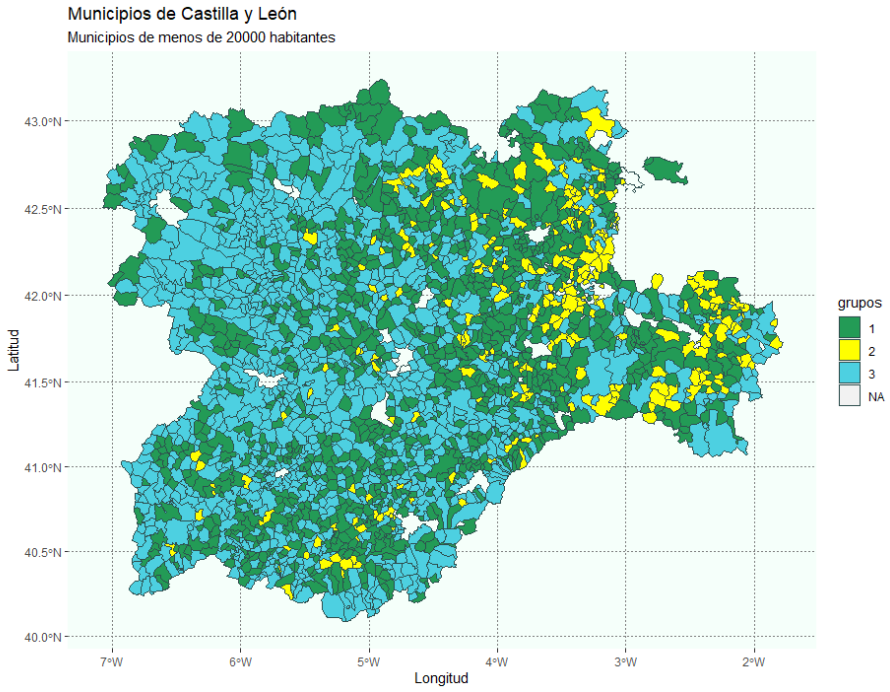

Ilustración 117: mapa de las k-medias para el grupo menos de 20000 habitantes y variables estadísticas de población.

En la *Ilustración de 117* podemos observar la clasificación de los grupos por las k-medias. Los valores que observamos como NA son los municipios de más de 20.000 habitantes. El grupo 1 lo podemos observar en color verde. Cuenta con 951 municipios y su población media es de 268,311. En este grupo la tasa de población de derecho varón se sitúa un poco por encima de 500 por cada 1.000 habitantes, habiendo aumentado un poco a lo largo de los periodos. Esto nos indica que son municipios ,más o menos, equilibrados entre mujeres y hombres.

El grupo 2 lo podemos ver en el grupo dos lo podemos ver en color amarillo. Este grupo lo conforman 259 municipios y su población media es de 90,21202. En este grupo las tasas de población de derecho varón han aumentado a lo largo de los periodos y se sitúan en torno a 650 por cada 1.000 habitantes, lo que indica que en estos municipios existe mayor población varón que mujeres.

El grupo 3 lo podemos ver en color azul, este es el grupo más numeroso. Está formado por 1.022 municipios. Su población media por municipio es de 913,5813 habitantes. En este caso la población de derecho varón se sitúa entre 470 y 540 por cada 1000 habitantes. Lo que indica que la población esta equilibrada.

4.2.3.3 Análisis clúster usando las componentes principales

Realizamos el análisis cluster de la misma manera que en el apartado anterior pero en este caso usamos las componentes principales. Debemos usar las componentes normalizadas. Usamos las 5 componentes principales que habíamos seleccionado como numero de componentes para que al menos un 90% de la variabilidad quede explicada. Para la selección del número de grupos miramos la ilustración . El número de grupos es 3, ya que es donde se produce el "codo" de la curva.

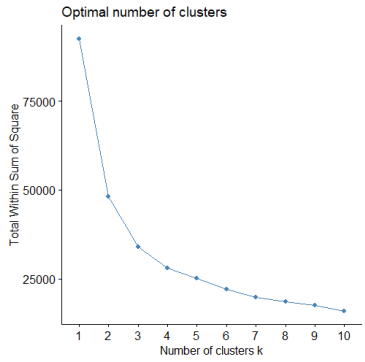

Ilustración 118: número óptimo de cluster usando PCA para el grupo menos de 20000 habitantes y variables estadísticas de población.

4.2.3.3.1 Cluster jerárquico

Para la elección del método más adecuado para calcular las disimilitudes entre clúster, elegiremos según la mayor correlación en la matriz de distancias de los métodos. En la *Tabla 14* tenemos estas correlaciones. Podemos observar que la mayor correlación la vemos en el método de los centroides.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0, 3815859	0,7837281	0,6910214	0,5925686	0,5121236	0, 7965745

Tabla 14: tabla de correlaciones de métodos

Los tamaños de los grupos son 2.209, 20 y 3. Vemos que estos tamaños son muy distintos. Las poblaciones medias son 268,311 , 90,21202 y 913,581 respectivamente. Las tasas de población de derecho varón del grupo más numeroso son en torno a 520 por cada 1.000 habitantes, lo que indica que existe algo más de población de derecho varón que de mujer en los municipios. El grupo formado por 20 municipios. Sus tasas de población de derecho varón se sitúan han ido aumentando a lo largo de los periodos situándose en los últimos años en valores incluso sobrepasando los 800 por cada 1.000 habitantes indicando que existen más varones que mujeres en los municipios. El grupo menos numeroso está formado por los municipios de Santa Olalla de Bureba con código INE 09354, Estepa de San Juan con código INE 42082 y Gormaz con código INE 42097. Las tasas de población de derecho varón han ido evolucionando a lo largo de los años con unos inicios en 1996 por debajo de 500 hasta llegar a valores de 750 o incluso 850 por cada 1.000 habitantes. Lo que indica que las tasas de población de derecho varón han aumentado de manera muy elevada, siendo estos municipios prácticamente toda la población varones.

4.2.3.3.2 Cluster no jerárquico

Para el método no jerárquico usaremos las k-medias con la función kmeans de R. El número de grupos lo tomaremos como 3, ya que seleccionamos que era el óptimo.

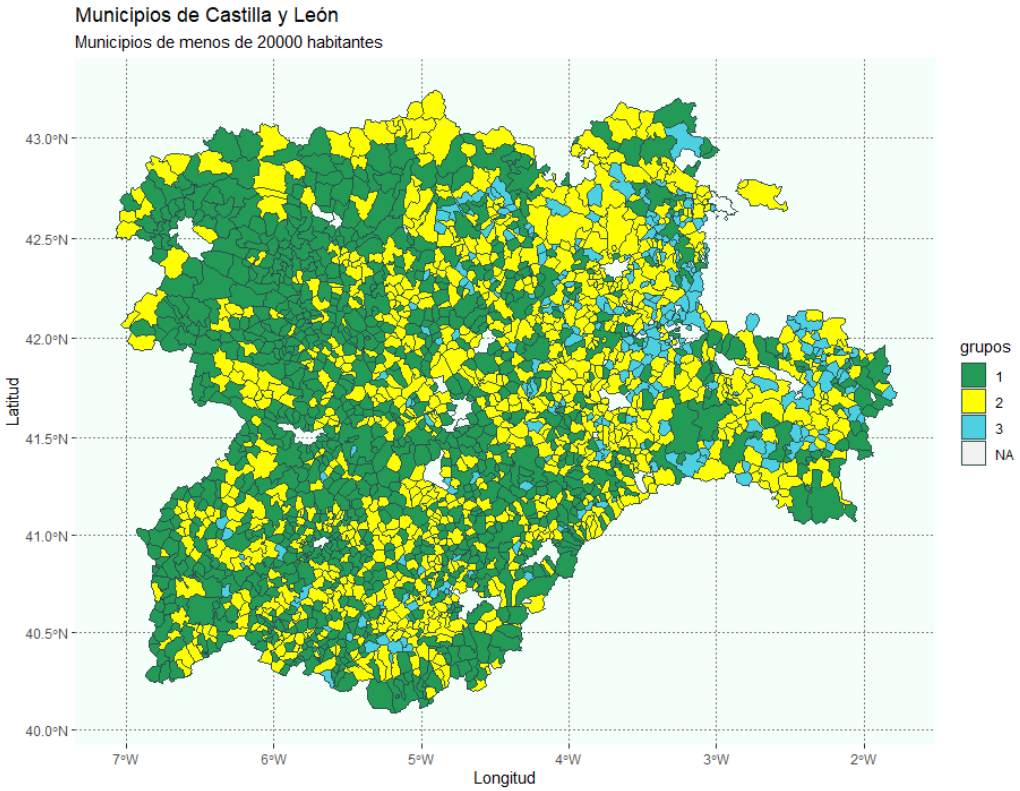

Ilustración 119: mapa de los grupos de las k-medias usando PCA para el grupo menos de 20000 habitantes y variables estadísticas de población.

En la *Ilustración 119* podemos ver los grupos representados. A primera vista parece que el grupo más numeroso es el verde. Los tamaños de los grupos son 1.013 , 970 y 249; sus poblaciones medias son 915,5341 , 269,6265 y 93,31343 respectivamente. Los municipios que vemos con color NA son los de más de 20.000 habitantes.

El grupo más numeroso lo conforman los municipios en color verde. Las tasas de población de derecho varón de este grupo se sitúan, en general, entre 470 y 520 por cada 1.000 habitantes a lo largo de todos los periodos. Podríamos decir que más o menos los municipios se encuentran equilibrados entre mujeres y hombres.

El grupo intermedio lo conforman los municipios en color amarillo. Este grupo tienen unas tasas de población de derecho varón que han ido aumentando a lo largo de los periodos. Estas tasas en los últimos periodos alcanzan valores cercanos a 580 por cada 1.000 habitantes. Existe algo más población de derecho varón en los municipios.

El grupo menos numeroso lo podemos ver en color azul. Las tasas de población de derecho varón en este grupo son de entre 560 y 700 por cada 1.000 habitantes. Estas altas tasas indican que existe mayor población varón qué mujeres en los municipios.

4.2.3.4 Multidimensional scaling

Para el multidimensional scaling dada la cantidad de datos usaremos un método métrico y representaremos en un mapa los distintos grupos que tenemos así como el grafico de grupos tradicional. En el multidimensional scaling como número óptimo de grupos usaremos el número de clústeres óptimos seleccionado en análisis clúster de todas las variables, este número es 3.

En las *ilustraciones 120 y 121* podemos ver las representaciones de los grupos. El grupo 1 lo podemos ver en color verde. Es un grupo bastante compacto en cuanto a distancia de esos municipios para la primera dimensión y para la segunda dimensión; existe algún municipio algo disperso en ambas dimensiones. Este grupo lo forman 1.009 municipios y su población media es de 918,2563. Las tasas de población de derecho varón se sitúan en las 520 por cada 1.000 habitantes, lo que indica que existen algún varón más en los municipios.

El grupo 2 lo podemos en color amarillo. Estos municipios están bastante separados, una parte de los municipios se encuentran juntos y el resto están bastante dispersos esto sucede sobre todo en la segunda dimensión. Este grupo lo forman 243 municipios y su población media es de 89,85382. Las tasas de población de derecho varón están sobre los 650 por cada 1.000 habitantes, lo que indica que la población de derecho varón es superior a la de mujer.

El grupo 3 lo podemos ver en colora azul. Este grupo se encuentra bastante compacto en la primera dimensión pero algo más disperso en la segunda dimensión. Este grupo lo forman 980 municipios y su población media por municipio es de 269,2385. Las tasas de población de derecho varón se sitúan en torno a 600 por cada 1.000 habitantes; indican que existe más población varón que mujer en los municipios.

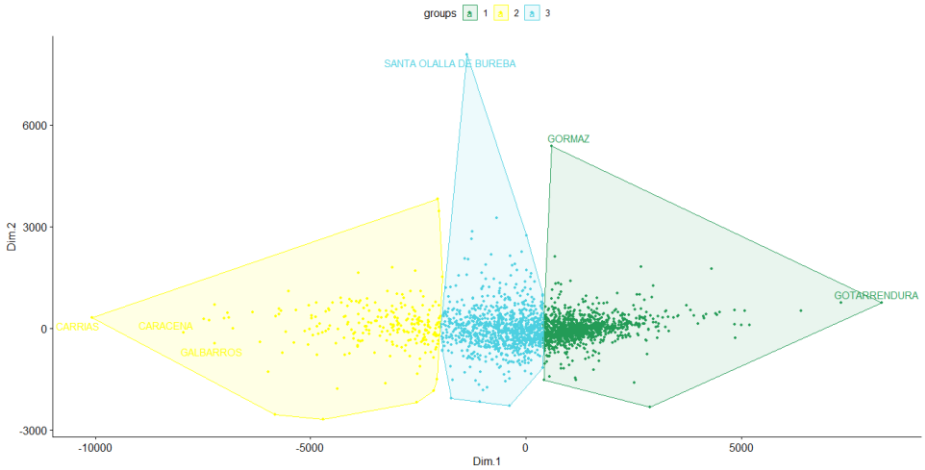

Ilustración 120: representación clasica de multidimensional scaling usando PCA para el grupo menos de 20000 habitantes y variables estadísticas de población.

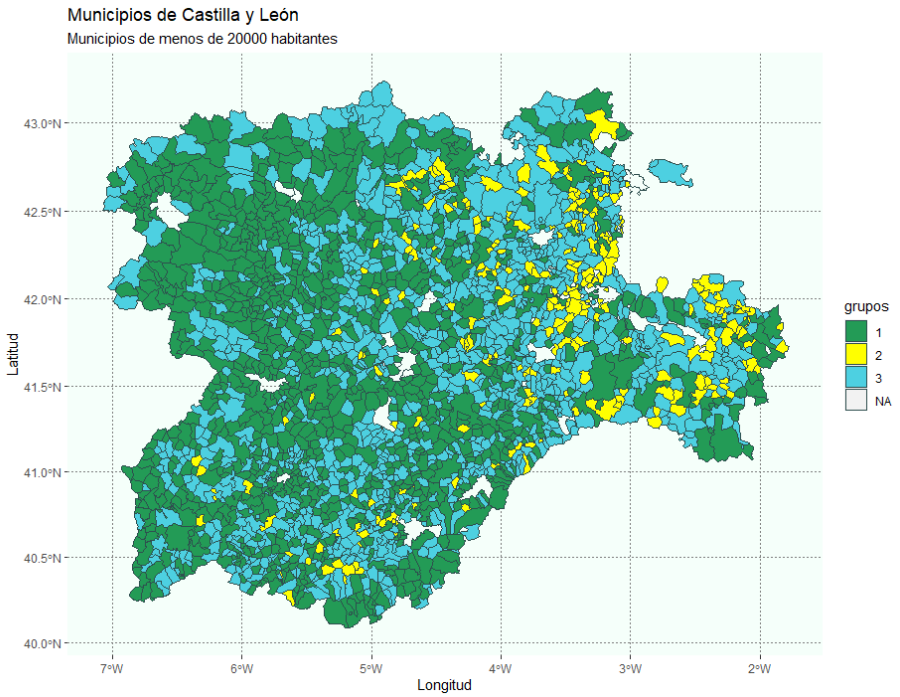

Ilustración 121: representación de los grupos de multidimensional scaling usando PCA para el grupo menos de 20000 habitantes y variables estadísticas de población.

4.2.3.5 Comparación de métodos

Los métodos proporcionan clasificación en 3. Los métodos cluster no jerárquico y el multidimensionales scaling proporcionan grupos de tamaños similares. Los grupos formados por métodos jerárquicos tienen tamaños muy distintos. Se crea un grupo muy numerario con población media de 500 habitantes y más o menos equilibrado entre población de derecho varón y población de derecho mujer. Un segundo grupo mucho menos numerario con una población media de 250 habitantes y una población de derecho varón mayor que la de mujeres. El tercer grupo lo conforman municipios con población media en torno a 50 habitantes, en este grupo la población de derecho varón es muy superior a la de mujeres. El cuarto grupo tiene aproximadamente 100 habitantes y una población de derecho equilibrada entre varones y mujeres.

Los grupos de métodos no jerárquicos son más equilibrados en cuanto al número de municipios. Los grupos se forman 2 equilibrados entre la población de derecho mujer y varón; y dos grupos en los cuales la población de derecho varón es mayor que la de mujer. En el método de las k-medias usando las componentes principales se forman 3 grupos uno que va variando entre mayor población de mujer o varón según el municipio y el periodo, pero siempre en torno al equilibrio entre ambos. Un grupo en el que la población de derecho varón es algo mayor que la de mujeres y un último grupo en el cual la población de derecho varón es muy superior a la de mujeres.
Todas estas características el mejor método sería clúster no jerárquica utilizando las componentes principales, ya que así reducimos la dimensionalidad.

4.2.4 Análisis de las variables de migración.

En este grupo están las emigraciones y las inmigraciones. En este grupo de variables trataremos los datos calculados en tasa por cada 1.000 habitantes, es decir, que realizaremos los análisis con las tasas por cada 1.000 habitantes de las variables.

4.2.4.1 Componentes principales

En el análisis de las componentes principales se utiliza la función "*prcomp*" de la misma manera que en otras ocasiones, en este caso son 46 las componentes principales que se realizan. Las primeras componentes para los primeros valores los podemos ver en la *Ilustración 122*.

```
> head(pca2$x)
                                PC1         PC2         PC3         PC4         PC5         PC6
ADANERO                   1.1273842  3.57578242  2.01145833 -1.5388060 -0.8659671  0.2829545
ADRADA (LA)               4.9398947 -0.02727701 -1.07763148 -0.2609807  0.3034572  0.3651821
ALBORNOS                 -4.2722381 -0.54824470 -0.79067754 -0.4196270 -0.2723939 -0.5009058
ALDEANUEVA DE SANTA CRUZ -3.1828927 -0.56311173 -0.09473109 -0.3488121  0.4646504 -0.5404916
ALDEASECA                -0.2277756 -1.41805487 -0.77758017 -0.4573145 -0.3780048 -0.3104535
ALDEHUELA (LA)           -2.2201326 -0.77259816 -0.27790434  0.1433950  0.3322930 -0.1695948
```

Ilustración 122:primeras componentes para primeros municipios para el grupo menos de 20000 habitantes y variables de migración.

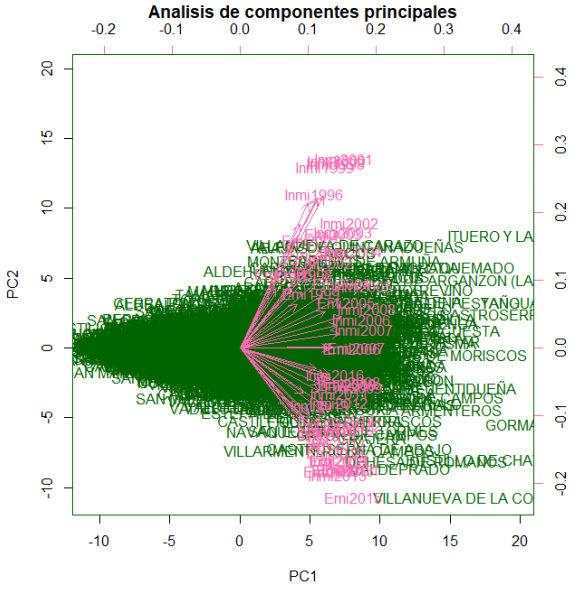

Ilustración 123: biplot de las 2 primeras componentes para el grupo menos de 20000 habitantes y variables de migración

En la *Ilustración 123* podemos ver el biplot de la representación de las 2 primeras componentes. Podemos ver que para la mayoría de los municipios la primera

componente es muy importante de aquí se distribuyen como una masa de puntos horizontal a excepción de los que se sitúan en valores positivos de la primera componente a los cuales les afecta también bastante la segunda componente. Las variables se encuentran en la segunda componente a la mayor distribución, ya que en la primera se encuentran en torno al valor cero. No podemos decir nada acerca de la influencia de una variable u otra en algún municipio, ya que al ser tantos municipios no podemos distinguirlos bien.

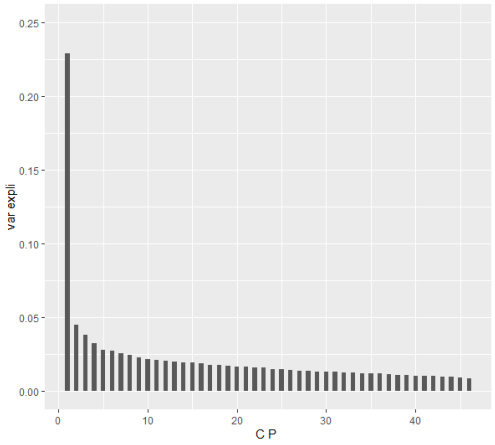

Ilustración 124: proporción de varianza explicada por cada componente principal para el grupo menos de 20000 habitantes y variables de migración.

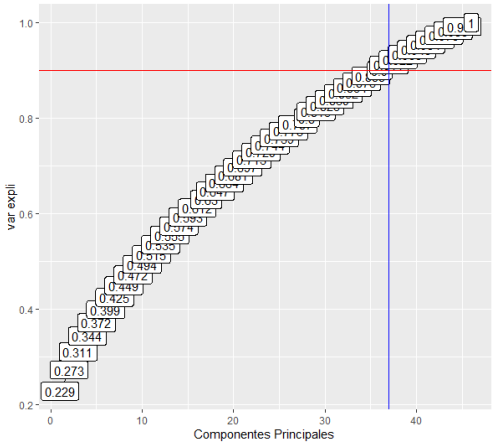

Ilustración 125: proporción de varianza explicada acumulada para el grupo menos de 20000 habitantes y variables de migración

Para la selección del número de componentes principales debemos mirar las *ilustraciones 124 y 125*. En la *Ilustración 124* podemos ver la proporción de varianza explicada por cada componente. Las componentes no llegan a explicar ni el 20% de la variabilidad a excepción de la primera que explica algo más del 20%. En la *Ilustración 125* tenemos la variabilidad explicada acumulada y podemos ver que con 37 componentes se alcanza al menos el 90%. Con estas conclusiones decidimos que el número de componentes a seleccionar es 37.

4.2.4.2 *Análisis clúster de todas las variables*

En este caso los datos que usaremos son los valores de las componentes principales para cada municipio Los datos deben estar normalizados para poder usar la función "fviz_nbclust", la cual. usaremos para calcular el número óptimo de grupos. En *Ilustración 126* podemos observar el punto en el cual se produce una bajada considerable en el valor de la inercia. Este cambio brusco es por qué se produce en el valor 3 ya que es donde podemos observar un "codo".

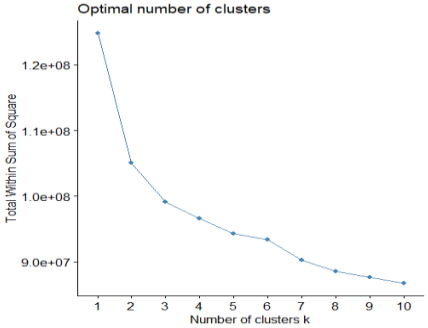

Ilustración 126: número de cluster óptimo para el grupo menos de 20000 habitantes y variables de migración.

4.2.4.2.1 Cluster jerárquico

Para escoger el método más optimo mirar la *Tabla 15*, podemos ver que el método con mayor correlación es average. En esta tabla tenemos la correlación entre la matriz de disimilitudes entre grupos calculada con cada método.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0,8357607	0,9594265	0,9317824	0,4343382	0,9111663	0,927762

Tabla 15: correlaciones de los métodos

Uno de los municipios que se encuentra solo es el municipio de Jaramillo Quemado tiene el código INE 09184 se encuentra en la provincia de Burgos. Este municipio tenemos descritas sus características en el apartado 4.2.1.2.1 . Otro municipio que aparece solo es el municipio de Arenillas tiene un código INE 42026 y se encuentra en la provincia de Soria, esto ya nos sucedía en 4.2.1.3.1.

El grupo de los demás está formado por 2.230 municipios y la población media por municipio es de 543,5725. En cuanto a la tasa de las tasas de inmigración son muy dispersos están entre 0 y 100 por cada 1000 dependiendo de los municipios y de los periodos. Cuando las tasas de inmigración el suceso es similar se encuentran muy variadas entre 0 y 100, pero habitualmente está en torno a 30 por cada 1000.

4.2.4.2.2 Cluster no jerárquico

Para este grupo usamos las k-medias con la función kmeans de R y como numero de grupos tomaremos 3.

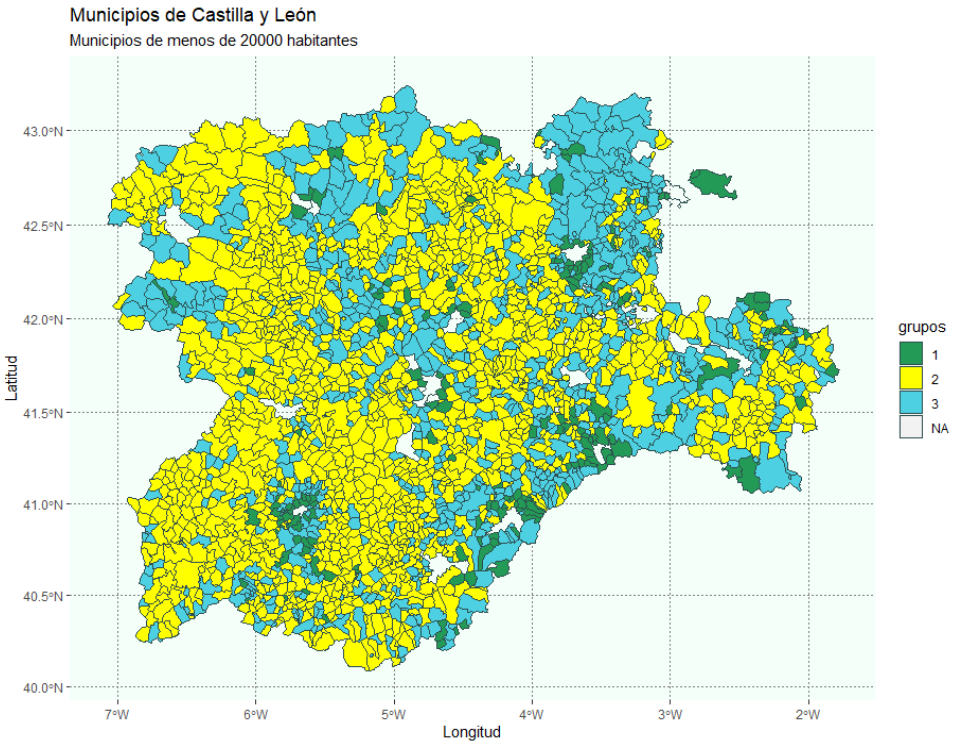

Ilustración 127: mapa de grupos de k-medias para el grupo menos de 20000 habitantes y variables de migración.

En la ilustración 127 podemos observar los distintos grupos formados por las k-medias. El primer grupo lo observamos en color verde, podemos ver que son los grupo menos

numeroso y que se encuentra principalmente alrededor de municipios de más de 20.000 habitantes. Este grupo está formado por 174 municipios y la población media por municipio es de 695,7711 habitantes. Las tasas de inmigración han ido aumentando ligeramente a lo largo de los periodos situándose actualmente en valores entre 60 y 100 por cada 1.000 habitantes. Pasas de emigración se sitúan entre 60 y 120 por cada 1.000 habitantes a lo largo de todos los periodos.

Pero todos lo podemos ver en amarillo es el grupo mayoritario. Está formado por 1.332 municipios y su población media por municipio se encuentra en torno 537,5546. Sus tasas de inmigración se sitúan entre 20 y 60 por cada 1.000 habitantes lo que no son unas tasas muy elevadas. Además existen muchos casos en los cuales para algún periodo y algún municipio estas tasas son cero. Las tasas de emigración se sitúan entre 30 y 80 por cada 1.000 habitantes.

El grupo 3 lo podemos ver en color azul y está formado por 726 municipios. La población media por municipio es de 516,6967. sus tasas de inmigración se sitúan entorno 35 por cada 1.000 habitantes. Sus tasas de emigración alcanzan valores más elevados incluso de 100 por cada 1.000 habitantes.

4.2.4.3 Análisis clúster usando las componentes principales

En este caso los datos que usaremos son los valores de las componentes principales para cada municipio. Los datos deben estar normalizados para poder usar la función "fviz_nbclust", la cual usaremos para calcular el número óptimo de grupos. En *Ilustración 128* podemos observar el punto en el cual se produce una bajada considerable en el valor de la inercia. Este cambio brusco es por qué se produce en el valor 3, ya que es donde podemos observar un "codo".

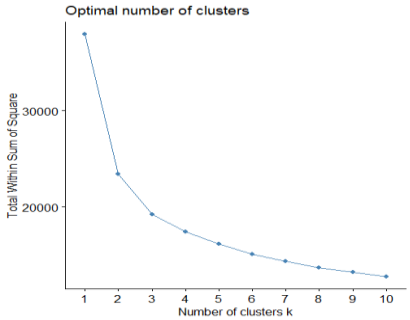

Ilustración 128: número óptimo de cluster usando PCA para el grupo menos de 20000 habitantes y variables de migración.

4.2.4.3.1 Clasificación jerárquica

Para la selección del método adecuado a utilizar con la función hclust de R. Observaremos la correlación de las matrices de disimilitudes entre grupos calculadas en los distintos métodos. Para la selección de un método adecuado mirar la tabla 16 de correlaciones. Ver que la correlación más alta se produce con el método average linkage.

Método	Complete	Average	Single	Ward	Median	Centroides
Correlación	0, 7558109	0,9578404	0,9339184	0,3557049	0,9149175	0, 9295361

Tabla 16: tabla de correlaciones

Los grupos que nos forman tienen los tamaños 2.230 , 1 y 1, es decir tenemos dos grupos únicamente con un municipio y un tercer grupo muy numerario. El primer grupo formado únicamente por el municipio está formado por el municipio de Jaramillo Quemado que tiene el código INE 09184. Es un municipio que se encontraba formado un único grupo también en varios análisis, en 4.2.1.2.1 lo describimos. Este municipio tiene como características que su tasa de inmigración es prácticamente cero salvo en algunos puntos en los cuales tiene tasas muy elevadas. Otro municipio que aparece solo es el municipio de Castroserracin tiene un código INE 40051 y se encuentra en la provincia de Segovia. Su población media es de 50,69565 habitantes. En los años 1998 y 1999 sus tasas de emigración superan los 280 por cada 1.000 habitantes, lo mismo sucede en los años 2003, 2007, 2010 y 2014. El resto de los años tiene tasas que son entre 50 y 100 por cada 1.000 habitantes y algún año esta tasas son 0. En cuanto a las inmigraciones sus tasas son superior a 100 por cada 1.000 habitantes la mayoría de los años.

Tercer grupo tiene unas tasas de emigración que se encuentran entre 20 y 40 por cada 1.000 habitantes, aunque ocasionalmente estas tasas son cero. Sus tasas de inmigración se encuentran en torno a 30 por cada 1.000 habitantes. Aunque estos valores pueden variar para algunos años y municipios y observarse valores tan dispares como cero o superar los 100 por cada 1.000 habitantes.

4.2.4.3.2 Clasificación no jerárquica

Para la no jerárquica usamos las k-medias. Como número de grupos 3, ya que detectamos que era el número óptimo de grupos.

En la Ilustración 129 vemos en un mapa representando los grupos formados por las k-medias. Como valores NA tenemos los municipios de más de 20.000 habitantes.

En color verde podemos ver el grupo 1, es el grupo más numeroso, que está formado por 1.224 municipios. La población media por municipio es de 554,1925 habitantes. Las tasas de emigración crecen en los primeros periodos y de la mitad de los años en adelante observamos que decrece lentamente. Su valores en general se encuentran entre 30 y 50 por cada 1.000 habitantes. Sus tasas de inmigración se encuentran cercanas a 35 por cada 1.000 habitantes, para algunos municipios en algunos periodos estas tasas son 0.

En color amarillo podemos ver el grupo 2. Este grupo está formado por 804 municipios y su población media por municipio es de 502,8864. Sus tasas de inmigración se sitúan en torno a 70 por cada 1.000 habitantes. En ocasiones se producen valores muy superiores a 100 por cada 1.000 habitantes o justo lo contrario 0. En cuanto a las tasas de emigración se encuentran en general entre 30 y 70 por cada 1.000 habitantes.

En color azul tenemos el grupo 3. Es el grupo menos numeroso, formado únicamente por 204 municipios. La población media por municipio es de 635,0808 habitantes. Sus tasas de inmigración han ido bajando en los últimos periodos, situándose en valores de entre 20 y 100 por cada 1.000 habitantes, salvo excepciones. Las tasas de emigración están en general entre 60 y 100 por cada 1.000 habitantes, en algunos periodos para algún municipio se producen valores muy elevados.

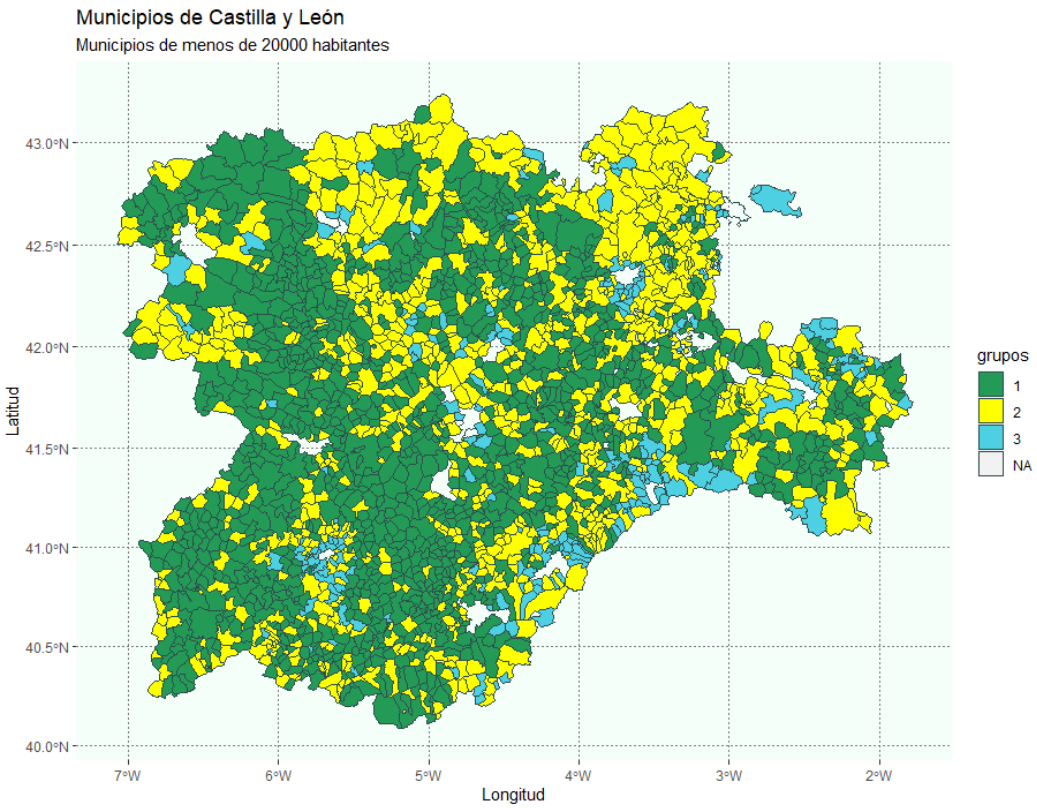

Ilustración 129: mapa de grupos de las k-medias usando PCA para el grupo menos de 20000 habitantes y variables de migración

4.2.4.4 Multidimensional scaling

Para el multidimensional scaling dada la cantidad de datos usaremos un método métrico y representaremos en un mapa los distintos grupos que tenemos así como el grafico de grupos tradicional. En el multidimensional scaling como número óptimo de grupos usaremos el número de clústeres óptimos seleccionado en análisis clúster de todas las variables, este número es 3.

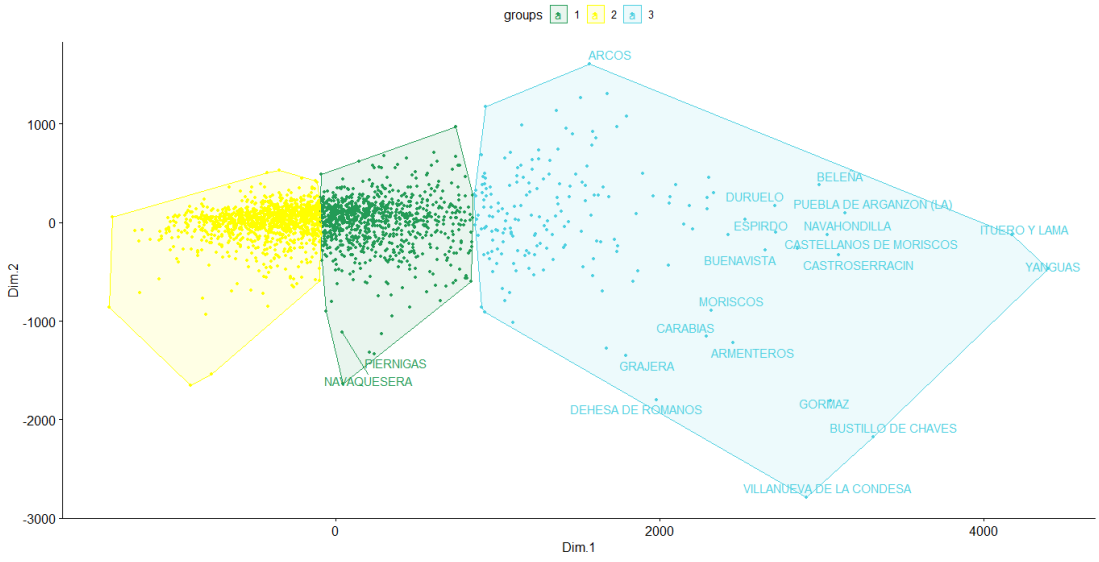

Ilustración 130: representación clasica de grupos para el grupo menos de 20000 habitantes y variables de migración.

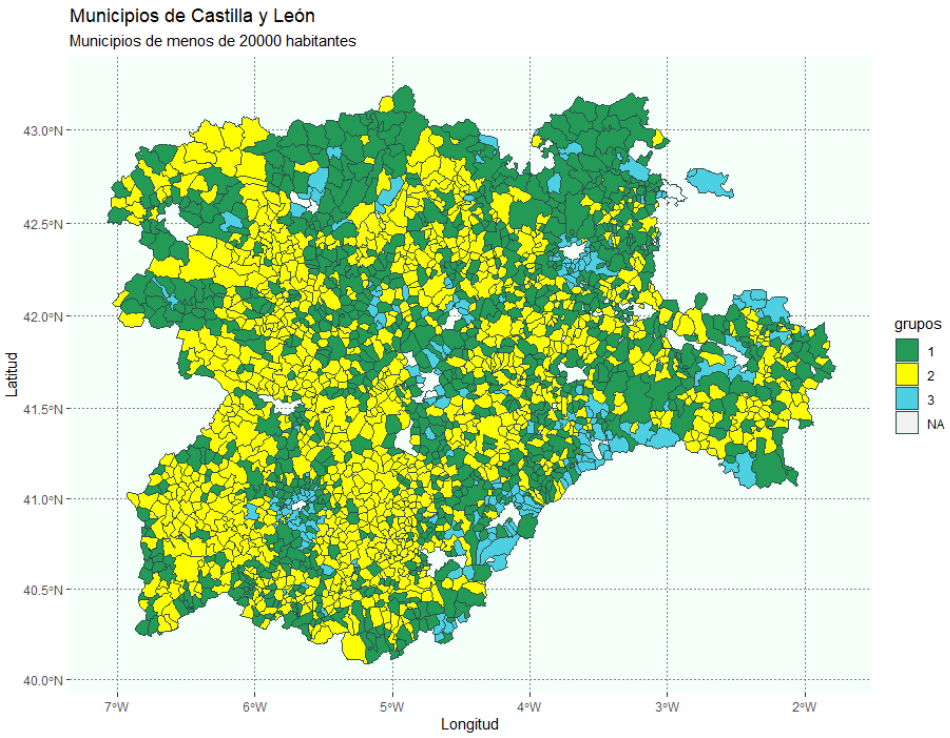

Ilustración 131: mapa de representación de grupos usando PCA para el grupo menos de 20000 habitantes y variables de migración.

En las *Ilustraciones 130 y 131* podemos ver una representación de los distintos grupos. En la *Ilustración 130* tenemos una representación clasica de las variables, esta representación es en las 2 primeras componentes. En la *Ilustración 131* podemos observar la representación de estos grupos en un mapa.

El grupo 1 lo podemos observar en color verde es un grupo bastante numeroso. En la *ilustración 130* podemos ver que es un grupo que los municipios se encuentran bastante cercanos unos de otros. Está formado por 912 municipios y la población media por municipio es de 451,0278. Población media se ha ido disminuyendo ligeramente a lo largo de los períodos. La tasa media de emigraciones por municipio y año se encuentra en 42,67187 por cada 1.000 habitantes. Las tasas de emigración han ido aumentando a lo largo de los periodos, comenzando con valores de 15 por cada 1.000 habitantes en los primeros períodos encontrándonos en valores de 50 por cada 1.000 habitantes en los últimos periodos. En torno a los años 2007 y 2008 se produjeron unas tasas más elevadas en la mayoría de los municipios. En cuanto a las tasas de inmigración suma su media se encuentra en torno a 41 por cada 1.000 habitantes. Estas medias han ido oscilando a lo largo de los años, aumentaron hasta el año 2006 y desde ese año las tasas han ido oscilando entre 40 y 60 por cada 1.000 habitantes. Algún caso para algún municipio se producen valores mucho más elevados a estas tasas que sobrepasan los 100 por cada 1.000 habitantes o valores que son cero.

El grupo 2 lo podemos ver en color amarillo. Es el grupo más numeroso está formado por 1.137 municipios. La población media por municipio se sitúa en 451,0278. La población media ha disminuido considerablemente desde los primeros periodos a los últimos; en los primeros se encontraba en los 500 habitantes y en 2019 estaba tan solo en 383. Su tasa media de emigración se encuentra en torno a 30 por cada 1.000 habitantes. Estas tasas medias crecieron rápidamente hasta 2007 después se produjo

un ligero bajón es 2017 vemos otro aumento en las tasas. En algunos municipios como Bustillo Páramo Carrión con código INE 34039 vemos que sus tasas de inmigración son cero a partir de 2012. En cuanto a las tasas de inmigración las tasas medias fueron aumentando en los primeros períodos hasta estabilizarse en valores entre 30 y 35 por cada 1.000 habitantes. En este caso existe en muchos municipios para los que en algunos períodos estas tasas son cero.

Grupo 3 lo encontramos en color azul, es el grupo menos numeroso. En la *Ilustración 130* podemos ver que es un grupo en el cual los municipios están más separados. este grupo está formado por 183 municipios. La población media por municipio es de 834,1511 habitantes. En estos grupos la población media por municipio se ha duplicado entre los primeros periodos y los últimos, en 1996 no alcanzaba los 500 sin embargo en 2019 sobrepasa los 1050 habitantes. La tasa media de migraciones se sitúa en torno a 63 por cada 1.000 habitantes. Estas tasas medias han aumentado mucho desde los primeros periodos a los últimos, ya que al principio estas se encontraban en torno a 20 por cada 1.000 habitantes y en los últimos años supera los 75 por cada 1.000 habitantes. En este grupo la tasa de inmigración media se sitúa en 84,32 por cada 1.000 habitantes, es una tasa bastante elevada. Hasta el año 2007 se produjo una subida elevada de estas tasas medias, pero desde ese año las tasas se han estabilizado entre 70 y 75 por cada 1.000 habitantes.

4.2.4.5 Comparación de métodos.

A la vista de los resultados con los distintos métodos el número óptimo de grupos sería 3. Los grupos que encontramos es uno con alta población media por municipio y unas altas tasas de inmigración. Un segundo grupo muy numeroso en el cual la población media ha descendido a lo largo de los períodos y además no es muy elevada; no supera los 500 habitantes de media. En este caso las tasas de migración se encuentran en torno a 30 por cada 1.000 habitantes. El tercer grupo que encontramos son los municipios en los cuales la población media no ha variado mucho a lo largo de los períodos y se encuentra en torno a 500 habitantes por municipio. En estos municipios las tasas medias de emigración aumentaron mucho entre 2006 y 2007.

Con estas características de grupos la mayoría de los métodos realizan unos grupos similares a los descritos pero el método de clúster no jerárquico utilizando las componentes principales o el multidimensional scaling son los que mejor realizan los grupos. Dado que el método de clúster no jerárquico junto con las componentes principales reduce la dimensionalidad lo tomaremos como el mejor método.

Los métodos jerárquicos realizan métodos en los cuales un par de municipios quedan como únicos el único clúster y los del resto de municipios quedan todos en un grupo muy grande.

5 CONCLUSIONES

A la vista de los análisis descriptivos previos a la clasificación de los 2.248 municipios Castilla y León se observa que hay un grupo de municipios que tienen un comportamiento diferente al resto. Estos municipios son en la mayoría de los casos los que tienen más de 20000 habitantes en 2019. Dado que la división de los municipios en dos grupos según tengan más o menos de 20000 habitantes es una división administrativa muy común, se considera necesaria esta división en grupos antes de la clasificación de los mismos.

Los 16 municipios que tiene más de 20.000 habitantes en 2019 son Aranda de Duero, Arroyo de la Encomienda, Ávila, Burgos, Laguna de Duero, Medina del Campo, Miranda de Ebro, León, Palencia, Ponferrada, Salamanca, San Andrés del Rabanedo, Segovia, Soria, Valladolid y Zamora. Estos municipios, en general, presentan altas tasas de nacimientos y aumento de la población. También se observa bajas tasas de defunciones y un elevado número de mujeres en edad reproductiva. En general, en estos municipios la proporción de mujeres y hombres es similar y las tasas de emigración no son elevadas.

Destacan especialmente los municipios de Laguna de Duero y Arroyo de la Encomienda, que son los municipios que más han aumentado su población total en el periodo estudiado. De hecho estos municipios se han incorporado a este grupo en los años 2003 y 2019 respectivamente.

En el grupo de 2.232 municipios menos de 20.000 habitantes en 2019, se observa que la población total, en general, ha disminuido. Este hecho junto con el aumento de la población de en el grupo de los municipios más de 20.000 habitantes refleja el movimiento de la población de las zonas rurales a las urbanas.

En este grupo destaca el municipio de Jaramillo Quemado (código INE 09184) que en todos los análisis de cluster jerárquicos se encuentra formando solo un único grupo. La población media anual de este municipio es de 8,826 personas. Tiene una tasa de mortalidad casi nula con un pico muy elevado en 2015. Lo que más destaca de este municipio es su elevada tasa de población de derecho varón. Además también merece una mención especial San Cristóbal de Segovia (código INE 40906) ya que es un municipio creado recientemente. En este municipio la población media anual ha aumentado en el período estudiado y la proporción de hombres y mujeres es similar. La tasas de mujeres en edad reproductiva y la tasa de natalidad han ido disminuyendo a lo largo de los periodos.

Del análisis de las variables de movimiento natural de la población se forman 3 grupos, uno de los cuales tiene un gran tamaño. En este grupo destaca que los municipios cuentan con muy poca población y una tasa de natalidad prácticamente nula en todos los periodos estudiados. Los otros dos grupos tienen un tamaño similar. Uno de ellos está formado por municipios que son colindantes a los de más de 20.000 habitantes, estos municipios los denominamos periurbanos. Tienen una población en torno a los 1.500 habitantes, una tasa la mortalidad que no es elevada y una natalidad distinta de 0. El último grupo lo forman municipios pequeños que tienen mayor tasa de defunciones que de nacimientos, incluso durante largos periodos de tiempo no hay

nacimientos. Con esto podemos ver que Castilla y León tiene un problema despoblación en determinadas zonas rurales.

En cuanto al estudio de las variables de estadísticas de población la principal diferencia entre los grupos formados es la proporción de mujeres y hombres en los municipios. Existe un grupo de municipios pequeños con población media anual en torno a los 50 habitantes en los cuales la población de derecho varón es muy elevada frente a la población de derecho mujer.

Al analizar las variables de migración se forma un grupo con elevada población media por municipio y altas tasas de inmigración. En el grupo con los municipios con población media anual por debajo los 500 habitantes encontramos que ha descendido mucha esta población media anual a lo largo de estos años. Para el grupo en el cual los municipios tienen una población media en torno a los 500 habitantes se produjeron más emigraciones, entre 2006 y 2007, que las que se produjeron antes y después de este periodo.

En los diferentes análisis se observó que dentro de los municipios de menos de 20.000 habitantes existe un grupo que tiene un comportamiento especial. Estos municipios podríamos denominarlos municipios periurbanos dado que son los que se encuentran alrededor de los municipios de más de 20.000 habitantes. En estos municipios periurbanos la población media anual por municipio ha ido aumentado a lo largo del periodo estudiado. Las tasa de defunción se mantienen estables en torno a 9 por cada 1.000 habitantes y la proporción de hombres y mujeres es similar. Las tasas de emigraciones media aumentaron hasta 2007; año desde el cual se han estabilizado las tasas en torno a 50 por cada 1.000 habitantes. Las tasas de inmigración media aumentaron hasta el año 2007, alcanzando valores de en torno a 86 por cada 1.000 habitantes. Después de este año las tasas han decrecido hasta valores de entre 40 y 50 por cada 1.000 habitantes.

Una futura línea de trabajo sería realizar estos análisis para más años, ya que, así, podríamos ver si los grupos han cambiado a lo largo de los periodos, además, de realizar una predicción de lo que podría ocurrir en el futuro. Otra posible línea de trabajo sería realizar los análisis de clasificación de municipios con más indicadores, como por ejemplo *Paro registrado en diciembre* o *Centros de enseñanza,* o usar estos otros indicadores para intentar explicar las características encontradas en los grupos formados.

6 BIBLIOGRAFÍA

[1] - *S I E* -. (s. f.). Sistema de Información Estadística Junta de Castilla y León. Recuperado 27 de abril de 2021, de https://www.jcyl.es/sie/sas/broker?_PROGRAM=mddbpgm.v2.indexv2.scl&_SERVICE=saswebl&_DEBUG=0&menu=index

[2] *Cálculo del crecimiento de la población*. (s. f.). Apuntes de demografía. Recuperado 8 de junio de 2021, de https://apuntesdedemografia.com/curso-de-demografia/temario/tema-3-crecimiento-y-estructura-de-la-poblacion/calculo-del-crecimiento-de-la-poblacion/

[3] Cardona Arévalo, A. (2020). *Caracterización de los municipios de Castilla y León atendiendo a factores demográficos.* (Trabajo final de grado). Universidad de Valladolid, Valladolid, España.

[4] Hernangómez, D (2021). mapSpain: Administrative Boundaries of Spain. R package version 0.2.3.9000. http://doi.org/10.5281/zenodo.4318024. Package url: https://CRAN.R-project.org/package=mapSpain

[5] Torrejón Valenzuela, A. (s. f.). *RPubs - Mapas de España en R*. Mapas de España con R. Recuperado 20 de abril de 2021, de https://rpubs.com/albtorval/595824

[6] Definición Población de derecho. (2021). Retrieved 9 May 2021, from https://www.eustat.eus/documentos/opt_0/tema_159/elem_1441/definicion.html

[7] Real Academia Española - RAE. (s. f.). *Definición de inmigración*. Diccionario panhispánico del español jurídico - Real Academia Española. Recuperado 10 de junio de 2021, de https://dpej.rae.es/lema/inmigraci%C3%B3n

[8] Real Academia Española - RAE. (s. f.-a). *Definición de emigración*. Diccionario panhispánico del español jurídico - Real Academia Española. Recuperado 10 de junio de 2021, de https://dpej.rae.es/lema/emigraci%C3%B3nb

[9] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[10] Amat Rodrigo, J. (2017, junio). *RPubs - Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. https://rpubs.com/Joaquin_AR/287787

[11] Rodrigo, J. A. (2017, junio). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. cienciadedatos. https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[13] Análisis de Multidimensional Scaling. Retrieved 11 June 2021, from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema4dm.pdf

[14] Fernández Temprano , M.A. (curso 2020/2021). Material del campus virtual de la Asignatura de Análisis Multivariante. Grado en Estadística. Universidad de Valladolid, Valladolid, España.

[15] Parra, F. (s. f.). *7 Agrupación de la información | Estadística y Machine Learning con R*. bookdown. Recuperado 10 de junio de 2021, de https://bookdown.org/content/2274/agrupacion-de-la-informacion.html#analisis-cluster

[16] Calvo, D. (2018, 1 diciembre). *Clúster Jerárquicos y No Jerárquicos*. Diego Calvo. https://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/

[17] Hojas, I. M. (2021, 17 enero). *Agrupación por K-medias*. StatDeveloper. https://www.statdeveloper.com/agrupacion-por-k-medias/

[18] Jiménez Cuadrillero, M. Á. (2018, 9 mayo). *RPubs - Clustering Jerárquico en R*. Clustering Jerárquico en R. https://rpubs.com/mjimcua/clustering-jerarquico-en-r

[19] Hojas, I. M. (2021a, enero 17). *Agrupación en clúster jerárquica*. StatDeveloper. https://www.statdeveloper.com/agrupacion-en-cluster-jerarquica/

[20] *Análisis multivariante. Clasificación*. (s. f.). mamutCola. Recuperado 29 de mayo de 2021, de http://ares.inf.um.es/00Rteam/pub/mamutCola/modulo6.html

[21] LRomero, MRamirez, JRojas, EDarghan. (2020, 1 julio). *RPubs - Análisis de Cluster en R*. RPubs. https://rpubs.com/lhromeroj/analisisdeclusterR

7 BIBLIOGRAFÍA NO CITADA

[1] Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

[2] Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

[3] D, Hernangómez (2021). mapSpain: Administrative Boundaries of Spain. R package version 0.2.3. http://doi.org/10.5281/zenodo.4318024. Package url: https://CRAN.R-project.org/package=mapSpain

[4] Giraud, T. and Lambert, N. (2016). cartography: Create and Integrate Maps in your R Workflow. JOSS, 1(4). doi: 10.21105/joss.00054.

[5] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[6] Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

[7] Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm

[8] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

[9] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.

[10] Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

[11] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[12] Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.1.

[13] Tennekes M (2018). “tmap: Thematic Maps in R.” _Journal of Statistical Software_, *84*(6), 1-39. doi:10.18637/jss.v084.i06 (URL: https://doi.org/10.18637/jss.v084.i06).

[14] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[15] Walesiak M, Dudek A (2020). “The Choice of Variable Normalization Method in Cluster Analysis.” In Soliman KS (ed.),_Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges_, 325-340. ISBN 978-0-9998551-4-1.

[16] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686