Facultat d'Economia i Empresa **Facultat de Matemàtiques i Estadística**

TRABAJO FINAL DE GRADO

Comparabilidad de elecciones

Grado de Estadística

Autor: Gibran Roye Muiños

Director: Josep M. Oller Sala

Convocatoria: 2019

Resumen

Este trabajo ha tenido como objetivo establecer y estudiar una metodología de comparación de conjuntos de datos que representan a los mismos individuos pero en diferentes tiempos. Se han estudiado datos electorales donde se ha realizado un análisis exploratorio inicial para conocer los datos y su contexto utilizando algunos métodos estadísticos como ACP y clustering. Se ha determinado una metodología de comparación basada en reducir la dimensión de los datos a dos y el traslado de los conjuntos de datos al mismo espacio, observando entonces su desplazamiento mediante la visualización gráfica, y su correlación con las variables, visualizadas estas también en el mapa junto a los conjuntos de datos. Los resultados de este estudio muestran las tendencias de los individuos en el tiempo según las variables, y la metodología puede ser ampliable tanto a más datos electorales como a conjuntos de datos de otra naturaleza.

Palabras clave: *Datos electorales*, *Profiling,álgebra lineal*, *análisis de componentes principales*, *Matrices.*

Abstract

This project's aim was to define and study a methodology that allows to compare different datasets which represent the same individuals but on different spots in time. We have chosen to study electoral data. A first research analysis has been done in order to get to know the features of the data and its context. In this initial analysis, some statistical tools have been used such as PCA and clustering. The comparison methodology stablished in this project is based on, first, the dimensionality reduction of the datasets and the shift of these datasets to the same space. After, the individuals' changes from the pass of the time could be seen. If this movement is compared with the placement of variables on the same space, we can know the relationship with the groups of individuals and the variables, in order to get some ideas on why they move. The methodology stablished can be expanded not only to more electoral data, but also to other datasets with a different nature.

Keywords: *Electoral data*, *Profiling*,*Linear algebra*, *Principal components analysis*, *Matrixes*.

Índice general

Índice de figuras v

1. Introducción 1

2. Bases del estudio 3

2.1. **Determinación de los datos** 3

2.2. **Contexto de la situación** 5

3. Análisis individual 8

3.1. **Descripción del conjunto de datos** 8

3.2. **Reducción de la dimensión y visualización** 10

3.3. **Profiling** 20

4. Comparación de resultados 27

4.1. **Centralización** 28

4.2. **Rotación** 29

4.3. **Visualización y comparación** 31

5. Generalización 37

6. Conclusiones 43

Referencias 44

Índice de figuras

3.1. Variabilidad acumulada explicada 2015 14
3.2. Valores propios 2015 15
3.3. Variabilidad acumulada explicada 2017 16
3.4. Valores propios 2017 17
3.5. Mapa de individuos 2015 18
3.6. Mapa de individuos 2017 19
3.7. Dendogramas 21
3.8. Evolución varianza entre grupos 22
3.9. Profiling 1 2015 23
3.10. Profiling 2 2015 23
3.11. Profiling 3 2015 24
3.12. Profiling 1 2017 24
3.13. Profiling 2 2017 25
3.14. Profiling 3 2017 25

4.1. Centralización 29
4.2. Comparación de resultados 32
4.3. Comparación de resultados con partidos 35

Clasificación AMS

MSC clasificación principal 05C50; clasificaciones secundarias: 91F10 62H25

Lista de acrónimos

ACP	Análisis de componentes principales
CERA	Censo Electoral de los Residentes ausentes
IDESCAT	Institut d'Estadística de Catalunya
CDC	Convergencia Democrática de Catalunña
ERC	Esquerra Republicana de Catalunya
JXSí	Junts Pel Sí
UCD	Unión Democrática de Cataluña
CUP	Candidatura de Unidad Popular
PSOE	Partido Socialista Obrero Español
PP	Partido Popular
Cs	Ciudadanos
PDeCAT	Partido Democráta de Cataluña
UE	Unión Europea
PC	Principal component
SS	Sum of Squared distances

Agradecimientos

Quería agradecer a mí tutor en este trabajo, Josep M. Oller Sala, por su apoyo, guía y aportación de ideas al trabajo, ofreciendo tanto la idea inicial del proyecto como gran parte del material necesario para el desarrollo de esta.

También agradecer al profesor Esteban Vegas por las herramientas de soporte de LaTeX ofrecidas.

Capítulo 1

INTRODUCCIÓN

El mundo político, como la economía y las ciencias sociales en general, se ve influenciado y determinado por un número casi infinito de variables. Si nos centralizamos solamente en el caso de las elecciones que se realizan cada determinado tiempo en un territorio determinado, la probabilidad de tener dos elecciones completamente iguales es prácticamente nula, lo cual no obstante da sentido a realizar diferentes elecciones en diferentes períodos de tiempo. Esto es debido a que tanto los agentes involucrados en estas elecciones como la infinidad de factores que las influyen están en constante movimiento e interacción entre ellos. Cuando nos referimos a los agentes involucrados, nos referimos por una parte a los votados, que serían los partidos políticos o personas que se presentan a las elecciones, dependiendo de la votación de referencia, y por otro lado a los votantes, ciudadanos, aunque también nos podríamos referir a grupos de votantes y proceder a la agrupación de resultados por territorios, por ejemplo. Tanto los votantes como los votados varían a lo largo del tiempo, aparecen nuevos partidos, se fusionan algunos territorios o algunas personas cambian de ideología y rumbo aún manteniendo el mismo nombre o etiqueta sobre sí mismos.

Este trabajo realizará un análisis de la comparabilidad de diferentes elecciones en el mismo territorio pero en diferentes períodos. Cuando se realizan elecciones generales con cuatro años de diferencia entre ellas, por ejemplo, resulta interesante realizar una comparación entre ellas de las cuales se podría desarrollar un estudio para encontrar el origen de estas diferencias o, al menos, ser capaz de compararlas correctamente. Este trabajo hará uso de la aplicación de conocimientos matemáticos (álgebra lineal y cálculo sobre matrices) y de lógica, con soporte de la estadística, sobre un caso real, encontrando una metodología correcta desde donde iniciar la comparación. Se busca obtener una base teórica, puramente matemática, con capacidad de ser aplicable a más elecciones, en diferentes tiempos y territorios.

Se obtendrán metodologías y procesos de cómo se pueden comparar diversas elecciones en el tiempo, estudiando las características de cada una. Cabe destacar que existen muchísimos más métodos y aproximaciones que se podrían utilizar más allá de los usados en este estudio, por lo que se priorizará conseguir una metodología correcta y sólida.

Durante el proyecto, se irán definiendo los criterios necesarios en cada parte del proceso, como por ejemplo, las distancias escogidas para calcular la distancia entre matrices o qué tratamiento se le va a dar a las observaciones missings en cada situación, y se acompañará el desarrollo de ideas con argumentos convenientes y demostraciones.

Se realizará el estudio con soporte de datos reales, elecciones realizadas en España en dos años cercanos. Se elegirán votaciones pertenecientes al territorio español con el fin de tener la mayor cantidad de información posible sobre el caso. Dentro del territorio, se buscarán datos electorales de elecciones realizadas a nivel autonómico con el fin de mantener la cantidad y la variabilidad de los datos a un nivel ajustado de sencillez inicial para poder desarrollar correctamente el análisis.

Se hará uso de la herramienta R (lenguaje de programación) para realizar los cálculos y análisis pertinentes. Se ha elegido esta herramienta debido a su capacidad y facilidad para el tratamiento y análisis de datos, desde la lectura de bases hasta la obtención de conclusiones. Para la redacción se combinará esta herramienta con Látex, a través de un archivo Sweave. Látex permite la generación de una muy buena redacción de documentos con elementos matemáticos como formulación y tablas, además, permite la replicación sencilla del análisis con otro set de datos de forma automática y veloz. Sweave permitirá juntar la codificación de R con la codificación de Látex en un solo archivo, y podrá ser compilable para generar archivos PDF.

Finalmente, se obtendrán conclusiones y se dejará determinada una base matemática que podrá ser ampliada hacia varios casos según la necesidad existente.

Capítulo 2

BASES DEL ESTUDIO

2.1. Determinación de los datos

En este capítulo inicial se presentará la idea de desarrollo del trabajo a nivel general, la cual se irá desarrollando en cada apartado. Se determinarán los datos inciales necesarios para llevar a cabo la investigación y se definirán las hipótesis y suposiciones inciales. Los siguientes apartados abarcan el análisis individual de cada período, donde se desarrolla el análisis estadístico de los datos; la metodología de comparación, donde se llevará a cabo el desarrollo matemático y computacional que va a dar respuesta a la pregunta del proyecto; generalizaciones, que estudiará algunas posibilidades de generalizar los resultados; y finalmente se obtendrán las conclusiones del estudio.

La pregunta que sirve de base en este proyecto, trata de comparar dos elecciones con las mismas bases y en dos o más períodos distintos. Por lo tanto, el primer paso será definir los datos necesarios para el estudio. Estos datos deberán representar los resultados de un territorio para unas elecciones. En este estudio, se elegirán **dos elecciones al Parlament de Catalunya** que sean consecutivas. De esta forma se contaría con resultados cercanos y familiares, lo que permite un mayor conocimiento del contexto de los datos. Para los períodos se establecerán las elecciones del 2015 y del 2017 con el fin de utilizar resultados más recientes.

El resultado de las elecciones de cualquier tipo se caracterizará por el porcentaje de votos que ha tenido cada participante presentado en las elecciones en un territorio determinado. Este porcentaje de votos se expresa como aquella cantidad de votos dirigidos a un representante sobre el total de votos realizados. Los votos que un votante realiza, pueden ir a uno de los representantes presentados en las elecciones, sin embargo, también existen los votos nulos y votos en blanco. Los votos nulos son aquellos invalidados por no cumplir con las reglas de votación, como puede ser el caso de votar por dos candidatos en vota-

ciones donde solo se podría votar a uno. Los votos en blanco muestran la desconformidad del votante con los agentes presentados debido a que, aunque queriendo ejercer su derecho a voto, ninguno de los agentes presentados defiende suficientemente sus ideologías, por ejemplo. Finalmente, también existe la abstención, conocida como todo el conjunto de votantes que tienen el poder de votar por cumplir con las condiciones necesarias, pero que simplemente no han ejercido su derecho a voto, teniendo igualmente un efecto sobre los resultados al cambiar la cantidad de votos absolutos necesarios para llegar a las diferentes modalidades de mayorías.

El objetivo, es obtener una base de datos compuesta en sus filas por menores territorios votantes. Como se analizarán elecciones a nivel autonómico, se estudiarán los resultados electorales a nivel de **municipio**. No se revisarán los resultados a nivel de provincias debido a que se contaría con un nivel muy bajo de participantes, además, obtener mayor nivel de granularidad permite representar mejor al terriorio y obtener resultados más fiables al convertirse en muestras suficientemente representativas. Entonces, para representar los resultados electorales se le otorgaría a cada municipio un vector con el porcentaje destinado a cada partido:

$$W_i = (P_{i1}, P_{i2}, ..., P_{ik});\ i = 1, 2, .., n(municipios)\ e\ i = 1, 2, .., k(representantes)$$

Como resultado, al tener a todos los municipios, en diferentes filas, se obtendría una matriz con todos los municipios en las filas, y los partidos presentados en las columnas. Para un análisis que resulte a la vez más representativo y sencillo, se tomarán como columnas los partidos representados en el Parlament de Catalunya ese año, conjuntamente con tres columnas más, la de abstención, formada por el porcentaje de votos no presentados a la votación; una columna con el resto de partidos no representados sumado con los votos nulos; y finalmente los votos en blanco.

Para la obtención de estos datos, se procederá a extraerlos del organismo público IDESCAT [5]. Se buscará obtener un archivo con formato Excel para una lectura sencilla desde el programa R. Cabe destacar que para mayor eficiencia se modificará previamente esta base de datos para ajustar su estructura al programa. Entre estos ajustes entran los ajustes de estructura que eliminarían las filas iniciales de información que indican el origen de los datos y período y territorio al que hacen referencia. Entre otros ajustes realizados entraría el tratamiento a los datos CERA [6]. En el conjunto de datos sobre el que se trabajará existen un número de observaciones equivalente al número de provincias de la comunidad que hacen referencia a grupo de votantes del conjunto CERA, Censo Electoral de Residentes Ausentes, este conjunto de votantes son ciudadanos situados en el exterior que ejercen su voto mediante correo. Todo este conjunto de ciudadanos quedan agrupados por provincias, por lo que no cumple con el split determinado en este proyecto. Además,

representan un porcentaje muy bajo de los votantes, no teniendo un gran peso sobre los resultados. Como resultado, se procederá a no tener en cuenta estas observaciones para la realziación del estudio.

Importados los datos se observa la existencia de **Missings**. Para estos conjuntos de datos se llevará a cabo un tratamiento más específico de los missings. En el caso de estas bases de datos en específico, sí que se pueden encontrar missings, sin embargo, se observa que en ambos años todos los missings provienen del mismo municipio al no tener este datos. Este municipio es Mediñá, y al afectar a las dos bases de datos por igual, se procederá a su elminación en ambas.

2.2. Contexto de la situación

Resulta interesante conocer la situación social que rodea a los datos por lo que se hará un breve recorrido por los acontecimientos de esos años. La siguiente información ha sido extraída de la página web del medio de comunicación pública RTVE.

Primero sería necesario recordar los acontecimientos ocurridos en el año 2014. Comenzando por la consulta 'simbólica' del 9N, donde participaron 2,5 millones de catalanes (algo más del 33 por ciento de los llamados a votar) y donde el 81 por ciento de ellos se mostró a favor de la independencia. La Generalitat de Catalunya considera la jornada como un éxito total, mientras que el Gobierno recurre ante el Tribunal Constitucional para comenzar a poner en marcha los mecanismos que tiene a su disposición para frenar la ruta independentista. El 21 de noviembre de 2014, la Fiscalía presenta una querella en contra del President de Catalunya y otros miembros de su gobierno por desobediencia, prevaricación y malversación. Además, la consulta se realizó desoyendo las órdenes del Tribunal Constitucional, el cual la había suspendido, comenzando así una serie de choques más intensos entre independetistas y Gobierno.

Llegando ya al 2015, el President de Catalunya anuncia un adelanto electoral convocando elecciones para el 27 de septiembre. Seguidamente, los partidos independentistas con más votos, CDC (Convergencia Democratia de Catalunya) y ERC (Esquerra Republicana de Catalunya) deciden presentarse conjuntamente a las elecciones para representar al bloque independentista con el nombre de JXSí (Junts Pel Sí). El 21 de julio, el hasta entonces partido que hacía coalición con CDC, UCD (Unión Democrática de Cataluña), abandona el gobierno de la Generalitat por discrepancias entre los partidos.

El 27 de septiembre de 2015 se realizán las elecciones (de las cuales se utilizan los datos en este estudio). El bloque independentista compuesto por JXSí y la CUP (Candidatura de Unidad Popular), alcanzan la mayoría absoluta, mientras que Cs (Ciudadanos) se posiciona como líder de la oposición en el Parlament de Catalunya. El PSC (Partido Socialista de Cataluña) se fija como tercera fuerza política y PP (Partido Popular) se desploma en votos con el quinto lugar. Además, Catalunya Sí que es Pot (marca de Podemos en estos comicios) entra en el Parlament de Catalunya como cuarta fuerza. El President de Catalunya considera un éxito los resultados para continuar con la hoja de ruta independentista. El 9 de noviembre, un año después de la consulta realizada, el Parlament de Catalunyaaprueba una resolución independentista para inciar el proceso de desconexión. Los choques entre Gobierno y Govern van aumentando en intensidad y frecuencia.

En el 2016, se consigue formar un gobierno para Cataluña después de un largo período de negociaciones. Más adelante, en junio de 2016, por un fuerte asedio de varios casos de corrupción y con el objetivo de una limpieza de imagen de cara a lograr el estado catalán dentro de la UE, el partido conocido como CDC cambia de nombre a PDeCat (Partit Demócrata de Catalunya). En noviembre de 2016, el nuevo President de Catalunya, comienza el camino hacia el Referéndum creando una partida dentro de los presupuestos dirigida a este. Esta decisión queda recurrida por el Gobierno quedando cautelarmente suspendidas por la decisión del constitucional.

En el 2017, la situación se va desarrollando buscando ser desencallada tanto por Govern como por Gobierno. El 13 de marzo de 2017, la situación alcanza un punto caliente crítico con la condena de inhabilitación al expresident Artur Mas por un delito de desobediencia del 9N. El 9 de junio de 2017, el President de Catalunya pone fecha al Referéndum, siendo esta el 1 de octubre de este mismo año. Sin el apoyo del Gobierno y con una única pregunta: "Quiere que Cataluña sea un estado independiente en forma de República?". Este Referéndum contará además con una ley que prevé la independencia en el caso de la victoria del sí. Ambos gobiernos ya sólo interactúan, para anunciar decisiones que avanzan hacia una Cataluña independiente, en el caso del Govern; y anunciar recursos y actuaciones judiciales para impedir que eso ocurra, en el caso del Gobierno. En el tramo final de este proceso, se inicia una salida escalonada - por cese o dimisión- de aquellos que prefieren no aparecer vinculados con la deriva independentista o que no están convencidos del desarrollo de la consulta del 1-O. El 6 y 7 de septiembre, el Govern culmina su desafío al Estado con la aprobación de las principales leyes de desconexión que acompañarán al Referéndum, la ley de transitoriedad jurídica y fundacional de la República. Al día siguiente, el Estado de derecho responde con la suspensión de la consulta y se ordena a los cuerpos de seguridad a evitar la ejecución de este. A su vez, el TSJC admite querellas de desobediencia, prevaricación y malversación a líderes catalanes. Los siguientes días

previos al Referéndum aumentan de intensidad con algunas detenciones y movimientos desde el Estado para impedir la realización del Referéndum.

Finalmente, el 1 de octubre de 2017 quedará enmarcado como una jornada histórica en España y Cataluña. Bajo una tensión sin precedentes en el territorio en las últimas décadas, se celebró el Referéndum con una Generalitat que trató de materializar, con un sistema electoral insólito y sin garantía jurídica alguna, su desafío al Estado y que apuntó a una declaración unilateral de independencia en base al resultado: el 90 por ciento de los 2,2 millones de votantes dijo 'sí' a la secesión, frente al 7,8 por ciento que votó 'no'. El Gobierno hizo uso de la fuerza para tratar de impedir una consulta y desde el Ejecutivo de Mariano Rajoy acusaron al Govern de haber "liquidado cualquier vestigio de respetabilidad democrática çon una consulta calificada de 'bochorno'. La Generalitat cifró en casi 900 el número de heridos y el Ministerio de Interior, en 431 los agentes heridos. Tras la consulta, no reconocida internacionalmente, la UE, la ONU y la OSCE, entre otros organismos, lamentaron el uso de la violencia en Cataluña y condenaron las cargas policiales. A la crisis social y política, se sumó la brecha entre los Mossos d'Esquadra y los cuerpos y fuerzas de Seguridad del Estado con reproches y descalificaciones mútuas. Miles de catalanes salieron a la calle el 3 de octubre, durante la huelga general, para denunciar la actuación de los agentes del Estado, mientras que éstos denuncian acoso y persecución. A las 21:00 horas del 3 de octubre el Rey Felipe VI se dirige a los españoles para acusar a la Generalitat de 'deslealtad inadmisible' y llamar a asegurar el orden constitucional.

El 21 de diciembre se realizan las elecciones al Parlament de Catalunya. Donde como se muestra, quedan afectadas por los acontecimientos ocurridos a lo largo de todo el año 2017, tras los choques entre Gobierno y Govern, la violencia del 1-O, el polarismo producido en la sociedad entre independentistas y no independentistas. El segundo período elegido para el estudio ha sido este, elecciones al Parlament de Catalunya del 2017, con una gran influencia social y política [7].

Capítulo 3

ANÁLISIS INDIVIDUAL

En este capítulo, se realizará un análisis de ambos conjuntos de datos por separado de modo descriptivo, se proseguirá con la aplicación de la reducción de dimensiones con el fin de iniciar el proceso de comparación y de poder ser visualizados gráficamente, y desde el cual se llevará a cabo seguidamente un proceso de clustering y profiling para definir las características comunes de los datos.

3.1. Descripción del conjunto de datos

En el 2015, hay 6 partidos representados en el Parlament:

```
[1] "JxSí"            "C's"             "PSC"             "CatSíqueesPot"
[5] "PP"              "CUP"             "Resto"           "Votos en blanco"
[9] "Abstención"
```

Distribuidos de la siguiente forma en el conjunto de todo el territorio:

```
      JxSí              C's               PSC          CatSíqueesPot
 Min.   :0.06879   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.40341   1st Qu.:0.02882   1st Qu.:0.02589   1st Qu.:0.01909
 Median :0.51316   Median :0.05579   Median :0.04834   Median :0.02949
 Mean   :0.48989   Mean   :0.07174   Mean   :0.05561   Mean   :0.03572
 3rd Qu.:0.58987   3rd Qu.:0.10294   3rd Qu.:0.07539   3rd Qu.:0.04603
 Max.   :0.88421   Max.   :0.34615   Max.   :0.21978   Max.   :0.14065
       PP                CUP               Resto         Votos en blanco
 Min.   :0.00000   Min.   :0.01176   Min.   :0.00000   Min.   :0.000000
 1st Qu.:0.02621   1st Qu.:0.05485   1st Qu.:0.02608   1st Qu.:0.002833
```

```
Median :0.04087   Median :0.07097   Median :0.03151   Median :0.004551
Mean   :0.04439   Mean   :0.08058   Mean   :0.03425   Mean   :0.005517
3rd Qu.:0.05807   3rd Qu.:0.09397   3rd Qu.:0.03946   3rd Qu.:0.007143
Max.   :0.17088   Max.   :0.41414   Max.   :0.13265   Max.   :0.032967
  Abstención
Min.   :0.03158
1st Qu.:0.14807
Median :0.17986
Mean   :0.18229
3rd Qu.:0.21629
Max.   :0.38173
```

Y en **el 2017**, están representados en el Parlament 7 partidos:

```
[1] "C's"             "JUNTSxCAT"       "ERC-Cat Sí"      "PSC"
[5] "Cat Comú-Podem"  "CUP"             "PP"              "Resto"
[9] "Votos en blanco" "Abstención"
```

Distribuidos de la siguiente forma en el connjunto de todo el territorio:

```
     C's            JUNTSxCAT         ERC-Cat Sí           PSC
Min.   :0.00000   Min.   :0.02945   Min.   :0.03846   Min.   :0.00000
1st Qu.:0.06224   1st Qu.:0.23445   1st Qu.:0.18525   1st Qu.:0.03501
Median :0.10158   Median :0.32000   Median :0.21922   Median :0.05517
Mean   :0.11944   Mean   :0.31543   Mean   :0.22312   Mean   :0.06266
3rd Qu.:0.16242   3rd Qu.:0.39914   3rd Qu.:0.25666   3rd Qu.:0.08094
Max.   :0.42308   Max.   :0.68085   Max.   :0.44547   Max.   :0.20705
Cat Comú-Podem         CUP                PP              Resto
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
1st Qu.:0.02023   1st Qu.:0.03343   1st Qu.:0.01465   1st Qu.:0.004847
Median :0.03030   Median :0.04647   Median :0.02252   Median :0.008981
Mean   :0.03354   Mean   :0.05397   Mean   :0.02504   Mean   :0.009225
3rd Qu.:0.04396   3rd Qu.:0.06429   3rd Qu.:0.03191   3rd Qu.:0.012768
Max.   :0.10714   Max.   :0.23711   Max.   :0.14941   Max.   :0.046512
Votos en blanco      Abstención
Min.   :0.000000   Min.   :0.03896
1st Qu.:0.002082   1st Qu.:0.12437
Median :0.003703   Median :0.15085
```

```
Mean   :0.004357   Mean   :0.15322
3rd Qu.:0.005651   3rd Qu.:0.17860
Max.   :0.043011   Max.   :0.35630
```

Los conjuntos de datos cuentan con una primera columna de datos de tipo caracter con la función de identificador (nombre de municipio). Las siguientes columnas con datos de tipo numérico situados entre 0 y 1 que hacen referencia a la proporción de votos dirigidos desde cada municipio a cada partido, votos nulos, votos en blanco o porcentaje de abstención. El 2015 cuenta con 947 municipios donde los votos irán distribuidos entre 9 columnas. El 2017 cuenta con 947 municipios donde los votos irán distribuidos entre 10 columnas

3.2. Reducción de la dimensión y visualización

Análisis de componentes principales: teoría

El objetivo ahora es poder visualizar los resultados electorales. Para la representación gráfica, cada variable supone una dimensión, por lo tanto, la primera variable se mostraría en el eje X, la segunda en el eje Y y una tercera variable podría ser observada en el eje Z. No obstante, no es posible visualizar más variables al mismo tiempo con este método por lo que es necesario reducir las dimensiones del conjunto de datos. Además, no solo tendrá otras ventajas la reducción de las dimensiones debido a que algunas métricas utilizadas en métodos y modelos estadísticos se pueden ver afectados cuando se aplican sobre varias dimensiones, sino que también reducirá el número de variables de distintos conjuntos de datos a dos o tres sin importar el número inicial de variables con el que contaban, es decir, contarán con el mismo número de variables.

Para ello, se procederá a realizar el ACP, **Análisis de Componentes Principales**. El proceso se inicia con la centralización de las variables. Esto se realiza mediante el cálculo del valor medio de las variables, obteniéndose así un valor medio en cada eje. Luego, se obtiene el vector formado por el valor medio de todos los ejes:

$$V_i = (P_1, P_2, ..., P_n);\ i = 1, 2, .., n(variables)$$

Este vector, que representa el valor medio de todas las variables, es llevado al punto 0 (donde todas las coordenadas de los ejes son 0) quedando así todas las variables centradas al tener el punto medio entre el origen de todas las dimensiones. No obstante, la posición de los datos entre ellos no ha sido cambiada. Esto permitirá que una vez centrados todos los datos, se cree una recta de regresión que pase por el punto 0. Para ello, se crea una

recta ya desde el punto 0 y se rota sobre este punto; esta recta rotará hasta conseguir el máximo ajuste de los datos que pueda.

Para definir el máximo ajuste de la recta, se proyectan los puntos de las observaciones en la recta a partir de aquel punto de la recta que más cerca está de la observación. Luego, se pueden mesurar las distancias entre el punto original y el punto proyectado; entonces, se buscaría rotar la recta que hiciera que la suma de todas las distancias fuera mínima, es decir, existe menos error entre el punto proyectado y el punto original en todo el conjunto de datos. De forma alternativa y para un cálculo más sencillo, también se puede intentar maximizar la distandia de la observación proyectada frente al punto de origen. Entonces, se extraerían todas estas distancias obtenidas desde cada observación y se elevarían al cuadrado para conseguir siempre valores positivos y que las distancias no se compensen entre sí. Sumadas estas distancias, se obtiene la suma de las distacias al cuadrado, o en inglés, sum of squared distances (SS), la cual se quiere maximizar.

$$SS = \sum_{i=1}^{n}(x_i - xp_i)^2$$

El resultado obtenido, la recta, será el componente principal 1. Entonces, esta línea busca resumir la máximo la variabilidad de las variables dentro de sí misma, ya que resulta de ser una proporción de cada variable original. Luego cuando esta proporción es elevada para una variable original con respecto a esta recta, quiere decir que la variable original a la que se hace referencia tiene mucho peso a la hora de explicar ese resultado; si se tiene en cuenta que la recta resume el comportamiento de todas las variables en ella misma, esto quiere decir que la variable original a la que se hace referencia tiene mucha capacidad de resumir todo el conjunto de datos. Es decir, **las componentes principales creadas son nuevas variables producidas a partir de la combinación lineal de variables ya existentes**. Los nuevos vectores unitarios (denominados así a los vectores de norma 1) pertenecientes a las PC (componente principal) son denominados vectores singulares o eigenvector. La raíz del eigenvector del PC1, se llama valor singular del PC1.

Una vez obtenida la PC1, se puede obtener la PC2 mediante la creación de un eje ortogonal al PC1, obteniendo así, mediante la rotación de los nuevos ejes, otra vez los ejes x e y formados por las nuevas variables capaces de explicar más informacón utilizando solo dos de ellas. Luego, las observaciones definidas por las variables previas quedarán visualizadas mediante nuevas coordenadas en un nuevo conjunto de datos. Se pueden conseguir la tercera dimensión siguiendo el mismo proceso y creando un eje perpendicular a los dos ejes previos. Cabe destacar que el número máximo de componenetes que se podrían crear equivaldría al mínimo entre el número total de variables y el tamaño de la muestra en el conjunto de datos. No obstante, en este caso solo interesaría conseguir entre dos y tres

ejes para poder ser representados gráficamente.

Se puede conocer también la capacidad que ha tenido cada componente de resumir la variabilidad. Para ello sería necesario conseguir la SS calculada previamente para cada componente y dividirla por n-1, lo cual permite convertirla en variación respecto al origen donde n es el tamaño de la muestra. El porcentaje obtenido muestra la variación total que rodea a la componente principal. Es decir, este porcentaje mostraría la cantidad de información de las variables originales que se expresa en las nuevas y permitiría establecer el número mínimo de nuevas dimensiones necesarias para no perder demasiada información del conjunto de datos original.

Finalmente, cabe destacar que como todas las variables quedarían reflejadas en un subconjunto mucho más pequeño, si el resultado muestra una agrupación de observaciones podría indicar la posible presencia de grupos muy parecidos entre sí, dando lugar indirectamente a un proceso de clustering (clasificación).

Un último concepto que cabe definir antes de proceder a la realización del ACP, es la distancia que se utilizará como métrica. La distancia típica utilizada más cercana al concepto más común de distancia es la distancia euclídea. Esta distancia, además, tiene un cálculo sencillo, siendo este el siguiente:

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

No obstante, se procederá a utilizar la **Distancia de Hellinger**. Con el objetivo de obtener resultados más interpretables y que mejoren los resultados de los procesos futuros. Esta distancia se expresa de la forma siguiente:

$$h(P,Q) = 2 \cdot ||\sqrt{P} - \sqrt{Q}||$$

Sin embargo, esta distancia se obtiene de forma automática al calcular la distancia euclídea si se transforman los datos iniciales. La primera transformación necesaria de estos datos será la transformación de cada dato en su raíz y multiplicación de toda la matriz de datos por 2. Esta transformación, dará lugar a un nuevo conjunto de datos manteniendo la misma estructura de la base, no obstante, cuando un proceso como el ACP realize el cálculo de distancias, estará calculando la distancia de Hellinger en vez de la distancia Euclídea.

La razón del uso de esta distancia es su equivalencia a la distanciade Rao. La distancia de Rao cuenta con una característica interesante y es que, cuando se calcula la distancia sobre dos puntos próximos entre sí, la mayoría de distancias estadísticas tienden a asemejarse a la distancia de Rao. Si supusiéramos que los puntos sobre los cuales queremos

calcular las distancias se encuentran dentro de una esfera, la mayoría de distancias calcularían una recta entre pares de puntos, mientras que la distancia de Rao calcularía la longitud del arco de la esfera entre un punto y otro. Como resultado, entre puntos muy cercanos, ambas formas de calcular la separación entre puntos se aproximan mucho. Es debido a esto que se ha elegido una distancia equivalente a la distancia de Rao.

Análisis de componentes principales: aplicación

Se dispone a hacer el análisis de componentes principales (ACP) de la base de datos. Este análisis se basa en crear nuevas variables a partir de la combinación de las variables continuas, con la peculiaridad de que estas nuevas son las que recogen mayor inercia al proyectar sobre estas la nube de puntos multidimensional, formada por los individuos de la muestra (en este caso los municipios de la comunidad autónoma) que tienen como coordenadas los valores obtenidos en las variables cuantitativas (todas en este estudio). El objetivo de este análisis es tener una visión de la base de datos con una dimensionalidad reducida, lo suficientemente reducida como para poder representarla gráficamente (2D O 3D). Con este resultado, además, dará a conocer características de las observaciones que vengan explicadas a través de la variabilidad de los datos, por lo tanto, servirá como una primera fotografía de la situación en la que se encuentran. En el caso actual, se llevará a cabo el análisis sobre los conjuntos de datos del 2015 y 2017. El objetivo en este apartado será solo obtener estas fotografías anteriormente mencionadas, por lo que, aunque se pueda realizar alguna comparación entre ambos años, todavía no sería completamente correcto compararlos.

Para llevar a cabo el análisis, se utilizará la función PCA de R incluida en el paquete FactoMineR, [1]. La documentación de este paquete define que esta función realiza el análisis de componentes principales con elementos suplementarios, los cuales serán introducidos en este caso al ser necesario tener en cuenta la introducción de futuras variables en el análisis de cada conjunto de datos. Con respecto al tratamiento de missings, la función realiza el tratamiento de variables missing a través de su sustitución por el valor medio de la columna, aunque no será necesario que se realice este proceso al haber realizado el tratamiento de missings anteriormente. Como resultado, se requeriá de la siguiente función para la realización del análisis:

$$acp_15 < -PCA(data_15, graph = T, ncp = 3)$$

Donde data15 es la matriz de datos (posteriormente se aplicará a data17), graph = F bloquea la realización de gráficos, los cuales se generarán posteriormente; ncp = 3 indica las dimensiones que se obtendrán, donde se limitará a 3 para poder ser visualizadas. Posteriormente se analizan los componentes principales obtenidos; para ello, se comenzará

por conocer el porcentaje de variación incluida en cada componente a través de gráficos scree que permiten su visualización (en verde los valores por encima de la media):

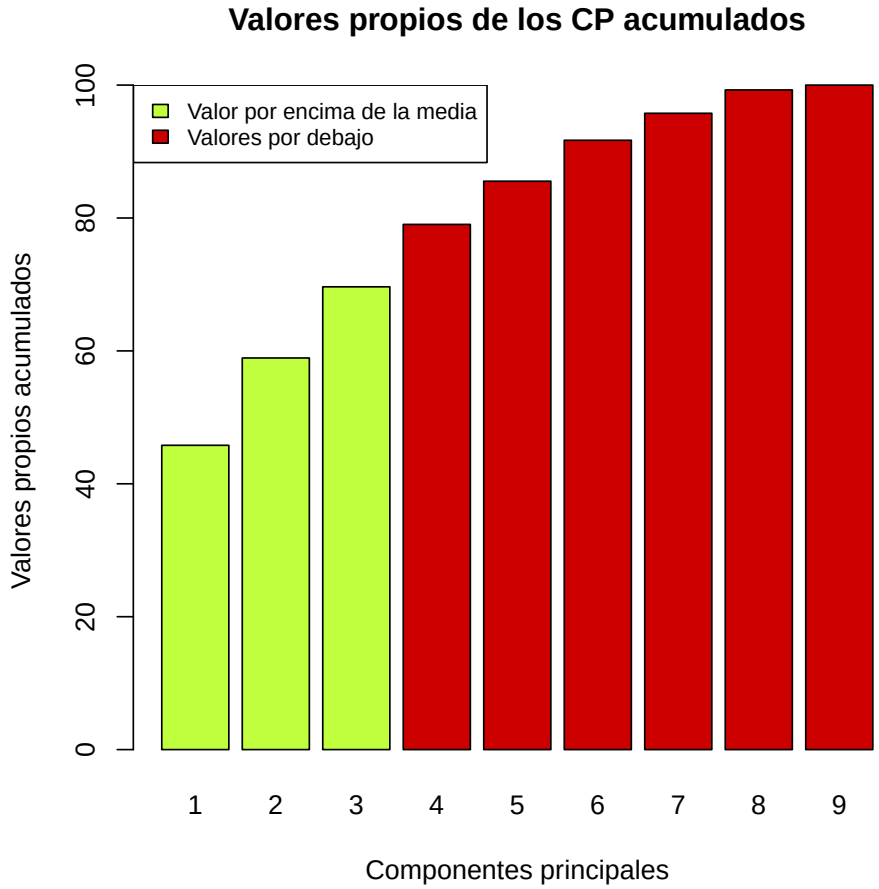

Figura 3.1: Variabilidad acumulada explicada 2015

Para el 2015, se puede observar que dos componentes principales son capaces de explicar de forma conjunta un 60 por ciento de la variabilidad, mientras que para 3 componentes se explicaría alrededor del 70 por ciento. Aún y no estar más cerca del 100 por ciento como sería deseable, debería ser suficiente información comprendida entre las dos y tres dimensinoes para resumir el conjunto de variables correctamente. También para el 2015 será interesente conocer los valores propios de forma individual ya que resulta también beneficioso para los resultados contar con valores propios que se encuentren por encima de la media. En el siguiente gráfico, se puede observar que solo los dos primeros componentes principales se sitúan por encima de la media.

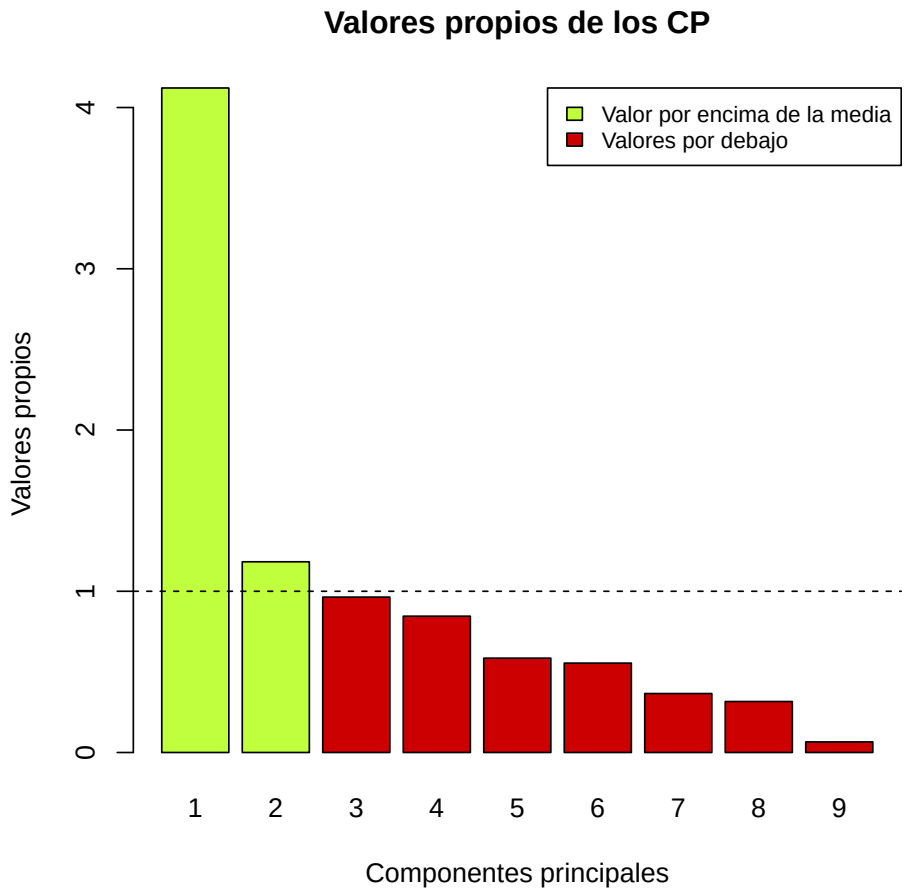

Figura 3.2: Valores propios 2015

Y a continuación se analiza el caso del 2017:

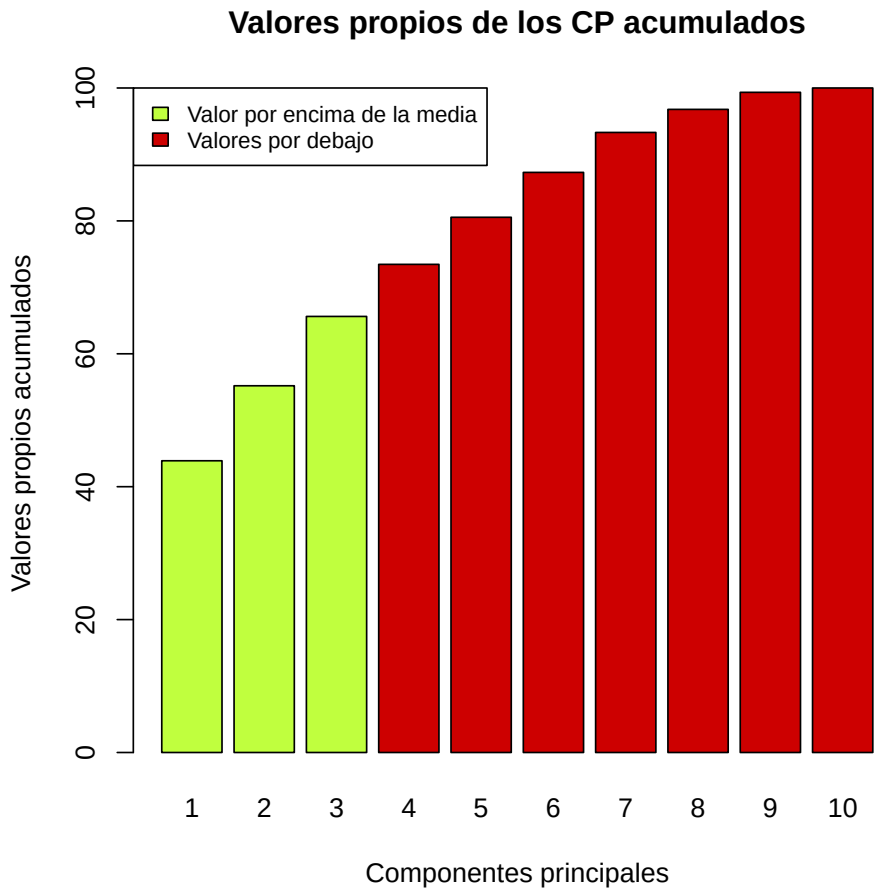

Figura 3.3: Variabilidad acumulada explicada 2017

A diferencia del caso anterior, ahora con dos dimensiones se contiene menos del 60 por ciento de la variabilidad y con tres dimensiones solo se superaría el 60 por ciento sin llegar al 70. Es decir, estos resultados tienen más dificultades para explicar el conjunto de variables originales utilizando solo dos o tres dimensiones. No obstante, se observa en el siguiente gráfico que se cuenta con un valor propio muchísimo por encima de la media, mientras que además de él solo el segundo y tercer valor propio también superan la media.

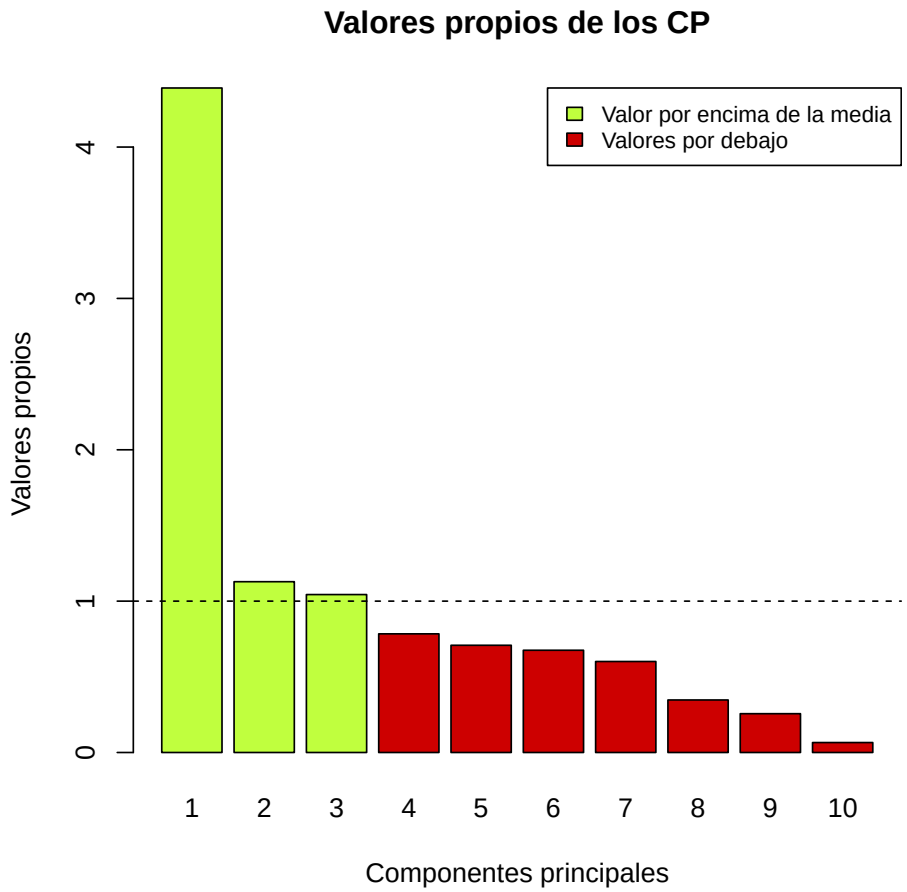

Figura 3.4: Valores propios 2017

Finalmente, se muestran los resultados obtenidos utilizando dos dimensiones. En el eje de X se encuentra el primer componente principal que explica el 46 por ciento de la variabilidad y en el eje de Y se encuentra el segundo componente, el cual abarca el 13 por ciento de la variabilidad. El mapa de puntos muestra las coordenadas de todos los individuos (municipios) en estos nuevos ejes:

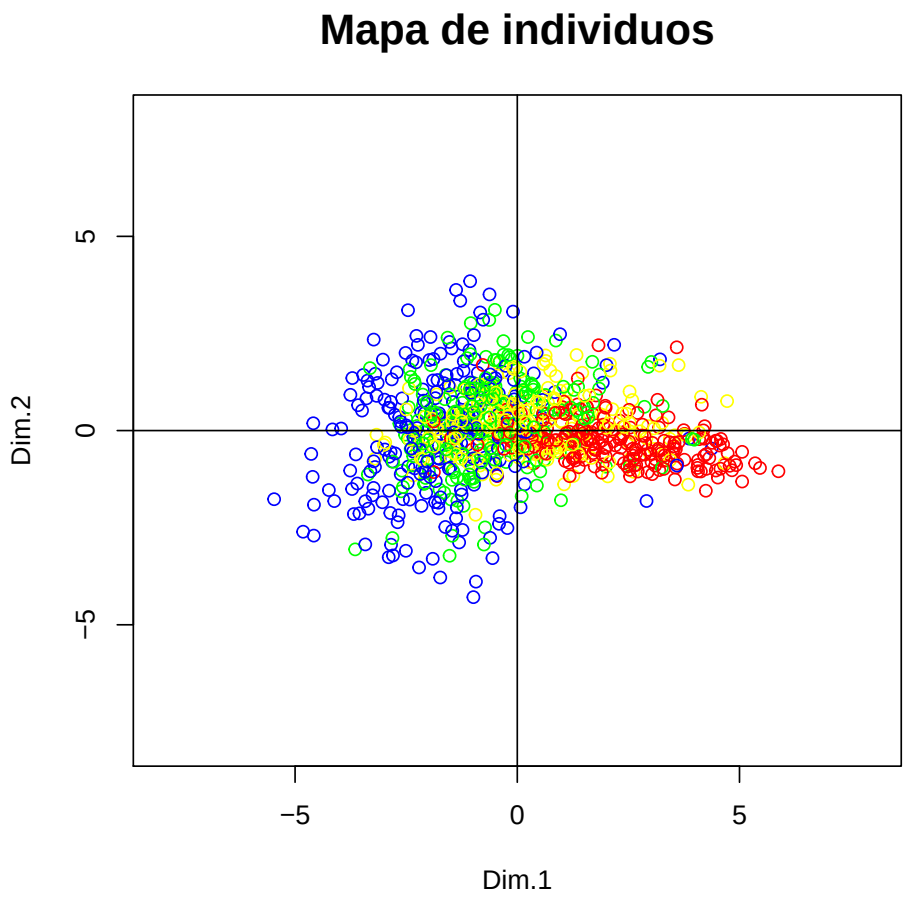

Figura 3.5: Mapa de individuos 2015

Como se puede observar, se han marcado a los individuos con cuatro colores según el tamaño del municipio. El tamaño del municipio ha sido obtenido desde la misma fuente que los datos (Instituto de Estadística de Cataluña). Para designar el color a los individuos se ha definido la partición según el cuantil en el que se encuentran:

- **Azul**, individuos situados en el primer cuartil, es decir, con menos de 313 habitantes.
- **Verde**, individuos situados en el segundo cuartil, es decir, con entre 314 y 946 habitantes.
- **Amarillo**, individuos situados en el tercer cuartil, es decir, con entre 947 y 3747 habitantes.
- **Rojo**, individuos situados en el cuarto cuartil, es decir, con más de 3748 habitantes.

Realizando el mismo mapeo para el 2017, se obtiene una nube de puntos muy diferente a la anterior. No obstante, cabe recordar que todavía no es correcto comparar estos dos

grupos de puntos:

```
integer(0)
```

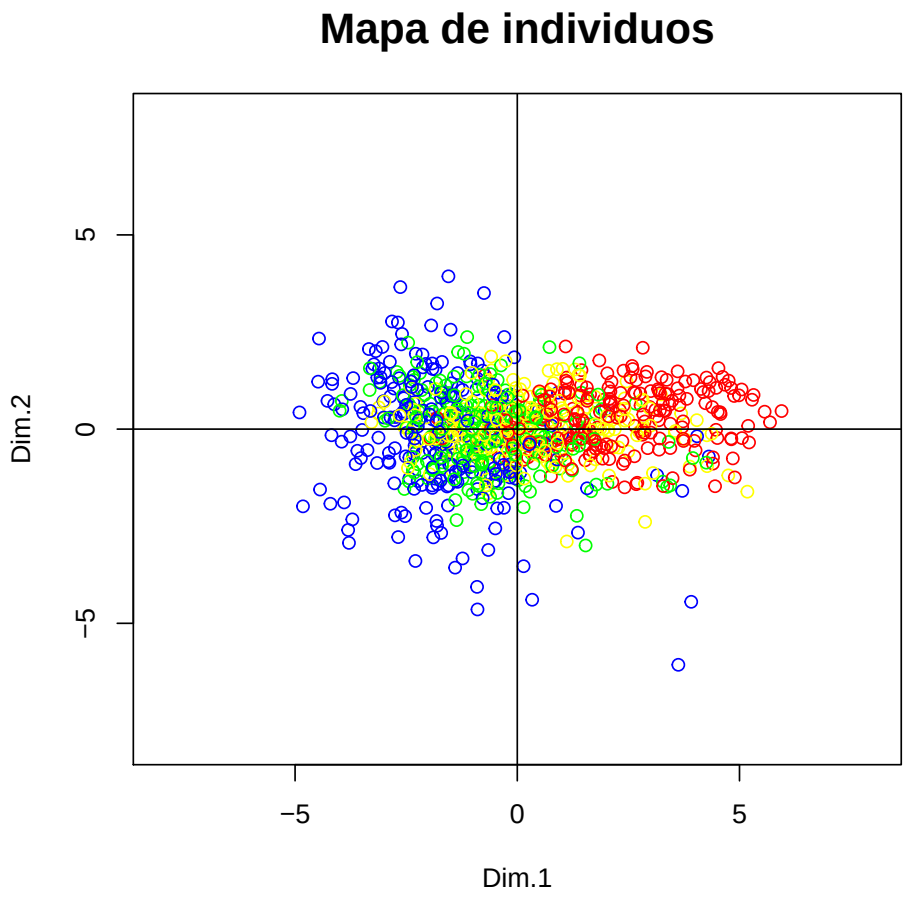

Figura 3.6: Mapa de individuos 2017

Como se puede observar en los dos casos obtenidos, no se diferencian fácilmente grupos de individuos con características comunes, siendo entonces un grupo muy homogéneo entre sí. Por otro lado, aplicando los diferentes colores para los individuos según el tamaño del municipio, se puede ver que los individuos pequeños (en azul) sí que se diferencian claramente de los individuos mayores (en rojo) al estar el conjunto pequeño en el lado negativo del eje de coordenadas x y el conjunto grande en el lado positivo. Los municipios de un tamaño intermedio se encuentran más mezclados y situados en el origen del espacio. El resultado muestra los individuos de los conjuntos de datos representados en sus nuevas dimensiones a partir de las coordenadas conseguidas y donde el eje x resume la mayor cantidad de información de los datos. Como se puede observar, para los datos del conjunto del 2015, la primera dimensión agrupa el 46 por ciento de la variabilidad, mientras que la segunda dimensión cuenta con el 13 por ciento. Luego, para el 2017, la primera dimensión

se reduce al 44 por ciento de variabilidad mientras que la primera se reduce al 11 por ciento.

Ambos casos han sido centralizados automáticamente por el método de ACP utilizado, y cuentan con la misma escala para una correcta visualización. Antes de proseguir a la metodología escogida para la comparación, se realizará un proceso de caracterización para conocer los últimos detalles de los datos.

3.3. Profiling

Finalmente, como último paso de la descripción de los datos, se buscará una agrupación de estos (clasificación) del cual se intentará determinar un perfil para los distintos grupos. Para el proceso de calsificación se llevará a cabo un clustering, el cual consiste en la agrupación de los datos en el espacio según la cercanía de estos en el espacio; para la distancia de los puntos **se utilizarán los resultados obtenidos desde el análisis de componentes principales realizado previamente**.Este proceso es complementario y puede ayudar a un estudio más profundo de los resultados conseguidos posteriormente.

Los algoritmos de agrupamiento o de clustering son un procedimiento de agrupación con el objetivo de juntar un grupo de individuos en varias clases homogéneas y distintas entre ellas. Este tipo de algoritmos tienen aplicaciones en múltiples ámbitos, como el marketing (descubrir grupos de compradores y utilizarlos por targeted marketing), astronomía (encontrar grupos de galaxias y estrellas similares), genómica (encontrar grupos de genes con expresiones similares), etc. Existen dos técnicas principales para el agrupamiento de casos: Agrupamiento o cluster jerárquico y agrupamiento o cluster no jerárquico. A partir de aquí, se realizará un cluster jerárquico, para decidir en cuantos subgrupos se dividirá la muestra.

Los métodos jerárquicos construyen una estructura en la que los elementos se agrupan en subconjuntos cada vez mayores hasta que todos pertenecen al mismo conjunto, de esta forma, no se muestra un agrupamiento sino las relaciones de proximidad que existen entre los elementos [9]. El procedimiento puede ser aglomerativo, comenzando inicialmente con clusters individuales con un único elemento y donde en cada iteración se unen los dos más próximos; o divisivos, partiendo de un úncio cluster que contiene todos los elementos y a partir del cual se define una división. A partir de aquí se pueden encontrar diferentes metodologías para determinar las proximidad de los clusters y así ir formando clusters más grandes a partir de la introducción de más individuos en un cluster. Entre estas medidas, existen por ejemplo el enlace simple, enlace completo, promedio del grupo, etc.

Para esta división se utilizará el método de Ward [2], este método permite la reducción de la pérdida en la inercia explicada. Además, este método ha permitido una mejor interpretación de las clases. El método de Ward define la proximidad como la suma de los errores al cuadrado, como resultado, en cada etapa se unen los dos clusters que dan lugar al cluster con menor suma de errores al cuadrado.

Se ha hecho uso de la función WARD.D2, la cual sigue el procedimiento de Ward al igual que las otras versiones, pero sin la necesidad de introducir los valores elevados al cuadrado. A partir de este método, se ha utilizado la distancia de Hellinger para medir la distancia entre las diferentes observaciones. El clustering jerárquico suele representarse a través de un dendograma, que muestra en qué orden se han unido los cluster y cual es el grado de proximidad que tienen los clusters que se unen. Los nodos hojas del dendograma se corresponden con los elementos individuales. El resto de nodos se corresponde con los clusters que se van formando. Este es el dendograma resultante del agrupamiento:

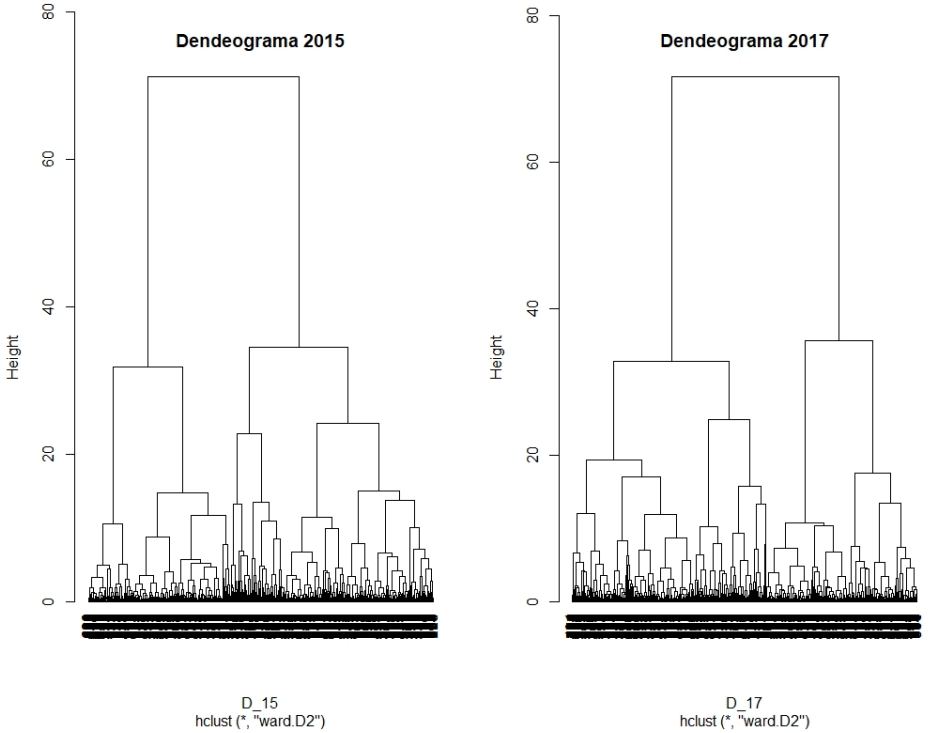

Figura 3.7: Dendogramas

Cada cluster contiene una cierta cantidad de varianza entre grupos y esta se va reduciendo a medida que se va aumentando el número de clusters, entonces, al elegir la cantidad de clusters se elegirá aquella cantidad a partir de la cual el hecho de crear un nuevo cluster no reduzca de forma significativa la variabilidad. Según los resultados obte-

nidos, visualizados a continuación, la partición ópitma estará en 3 clusters. A continuación se puede ver la evolución de la variabilidad comprendida en cada partición de los clusters:

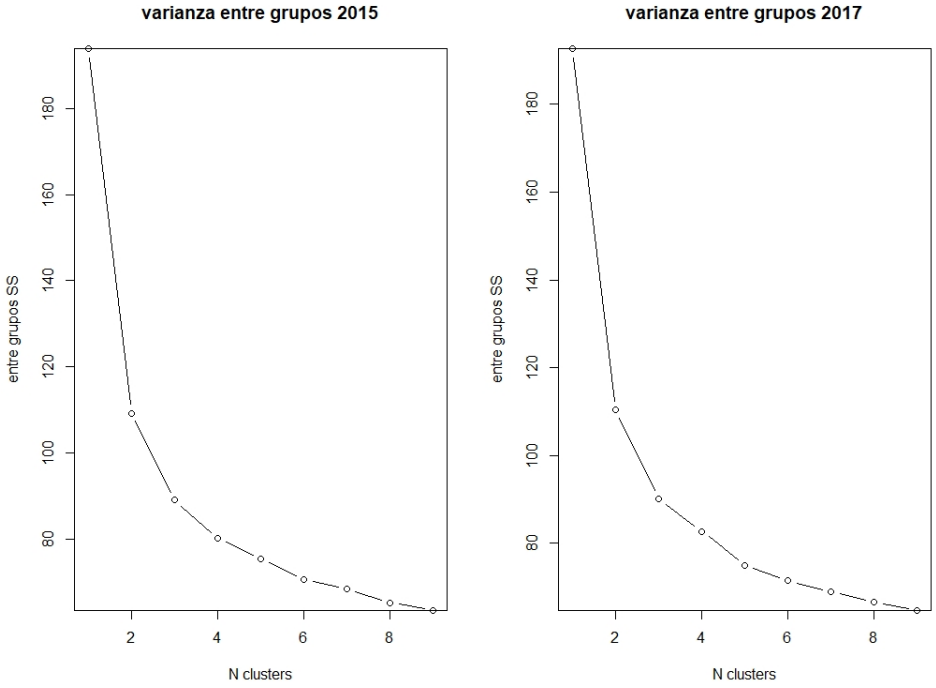

Figura 3.8: Evolución varianza entre grupos

Una vez finalizado el proceso de clustering, se pueden separar las observaciones entre los grupos determinados, en este caso, tres para cada uno de los dos conjuntos. Con estos resultados se puede proceder a realizar el profiling. Esta última fase del proceso consiste en representar gráficamente la distribución de los grupos creados con respecto a cada una de las variables. El objetivo, es conocer las tendencias de los individuos que componen cada grupo hacia cada una de las variables, o también, agrupar las variables por grupos.

Finalmente, se caracterizarán cada uno de los tres grupos según aquellas variables en las que destaquen más. Se realizará la representación gráfica para ambos conjuntos de datos y será interesante observar si la composición de cada grupo queda caracterizada por las mismas variables (partidos) en ambos años. La representación gráfica se realizará mediante gráficos de barra sin diferencia de color donde la mayor altura de la barra de un grupo con respecto a otro mostrará la tendencia que tiene ese grupo de individuos hacia esa variable. Primeramente se mostrarán los gráficos del conjunto de datos de 2015:

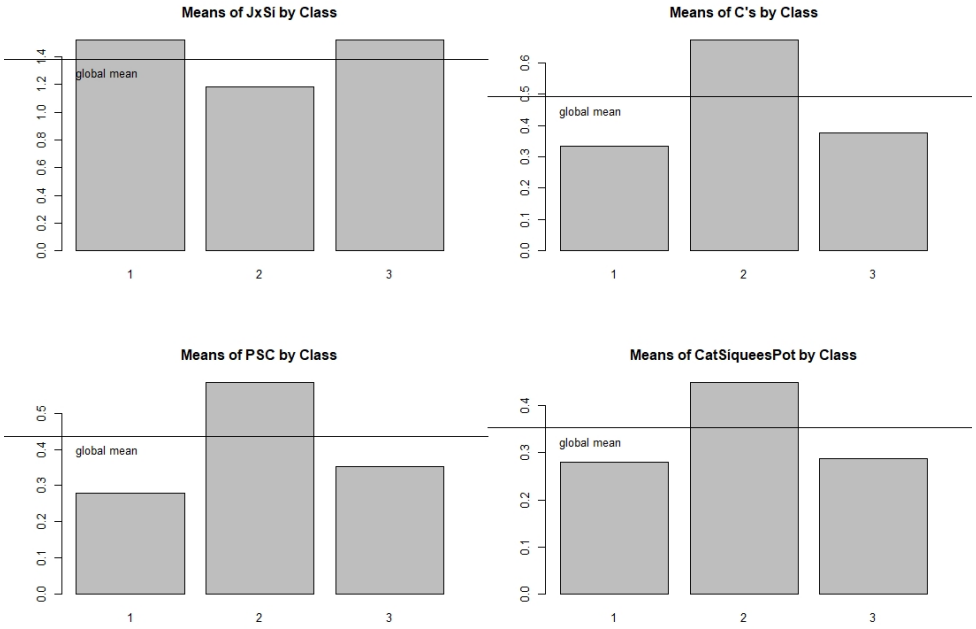

Figura 3.9: Profiling 1 2015

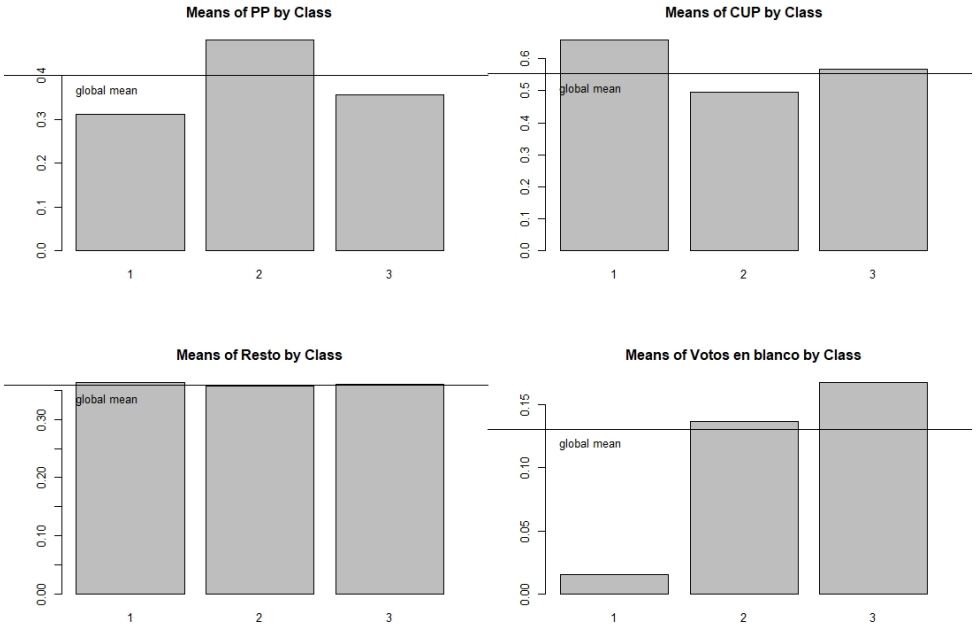

Figura 3.10: Profiling 2 2015

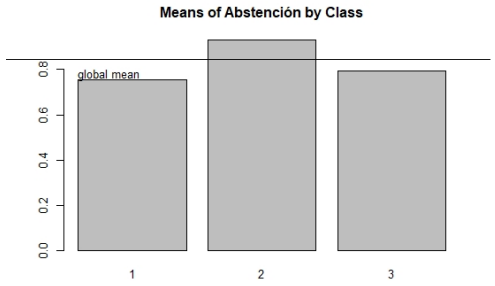

Figura 3.11: Profiling 3 2015

Seguidamente, se repetirá la visualización gráfica para los datos del 2017:

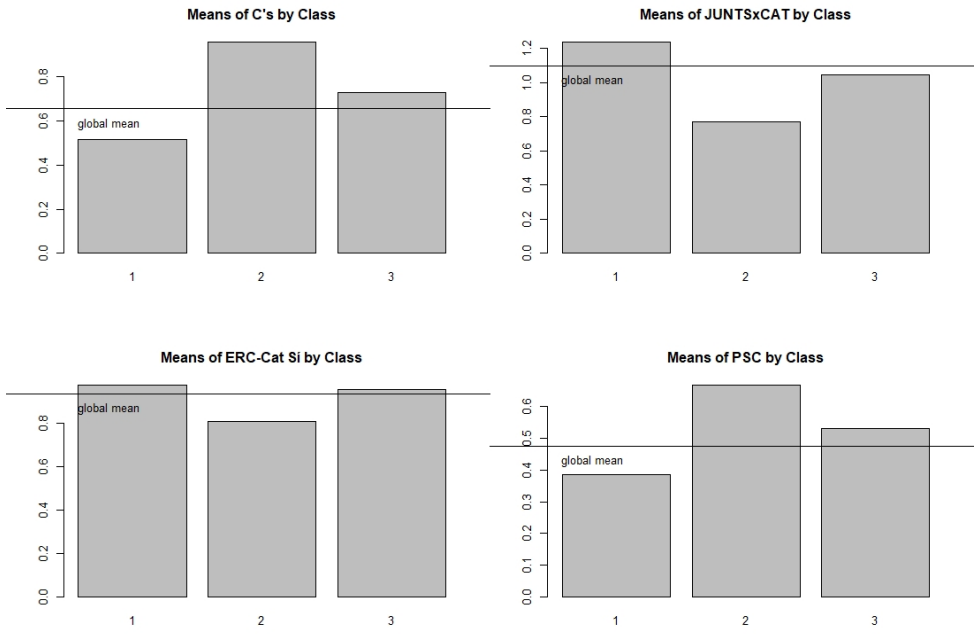

Figura 3.12: Profiling 1 2017

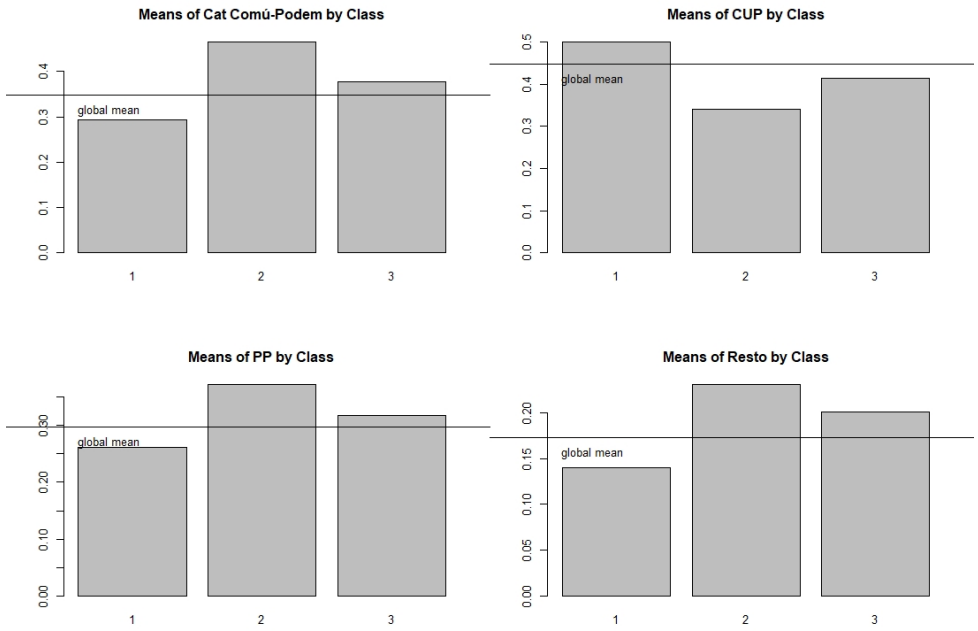

Figura 3.13: Profiling 2 2017

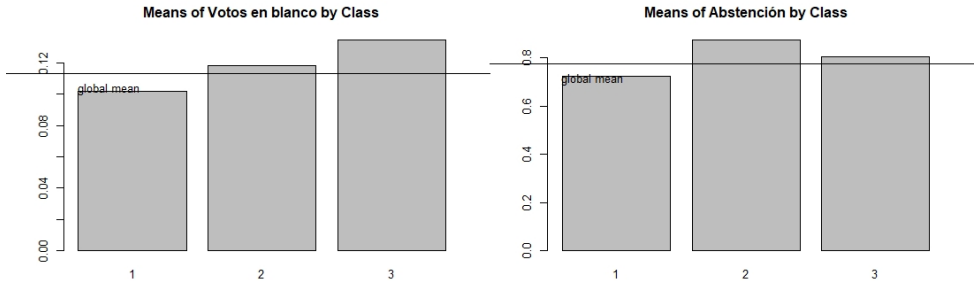

Figura 3.14: Profiling 3 2017

Finalmente se pueden extraer resultados interesantes de las características comunes entre subconjuntos de datos en ambos períodos estudiados:

- En ambos períodos, siempre se encuentran claramente separados por sus características entre los tres grupos los votos en blanco, la CUP y el conjunto Cs, PSC PP, Comuns y abstención. Es decir, frecuentemente se podrían formar tres grupos de individuos bien diferenciados por sus características los votos en blanco (en desacuerdo con los partidos presentados), el conjunto de votos a partidos que se encuentran en el resto del territotio español además de Cataluña (Cs, PSC, PP y Comuns) además de la abstención; y finalmente otro grupo más para la CUP, partido independentista que se diferencia más de otros partidos con el mismo objetivo por estar más cerca de un extremo y que no ha querido formar parte del bloque independentista en el 2015 para formar Junts pel Sí.

- Luego, en referencia a los partidos independentistas, Esquerra, JxCat y Junts pel Sí también se pueden encontrar resultados interesantes. En el 2015, las características de JxSí lo situaban entre el mismo grupo de la CUP y el grupo de los votos en blanco. Entonces, en el 2017 con la separación de los partidos, se puede ver que se han separado de siguiendo las mismas características que antes, ahora JxCat se sitúa solamente en el mismo grupo que el partido CUP, mientras que ERC se sitúa todavía entre ambos grupos igual que antes.

- Finalmente, destacar que los resultados muestran la separación de los municipios claramente entre independentistas y unionistas, manteniéndose en ambos años y cuyo resultado está claramente influenciado por los eventos sociales y políticos del entorno que rodea a los datos.

Capítulo 4

COMPARACIÓN DE RESULTADOS

Ahora que ya se cuenta con los dos conjuntos de elecciones de forma separada y se conoce su contexto y la descripción estadística de los datos, se procederá a comparar estos resultados.

Si se intentara comparar directamente el movimiento de los individuos de las elecciones de un año a otro, esta fórmula sería correcta únicamente en el caso de ni partidos ni individuos hayan cambiado. El cambio en los partidos, al cual se hace referencia, es que un partido cambie su ideología o rumbo, lo cual pasa frecuentemente debido a que siempre intentarán adapatarse a los cambios de la sociedad con tal de llamar la atención de un número mayor de votantes y hacer crecer su porcentaje de votos. Por lo tanto, el cambio en los conjuntos de datos de ambos períodos viene porque si un individuo quiere votar a un partido con ciertas características, por ejemplo podría ser un partido de derechas, un partido preocupado por cierto problema social, un partido conservador, etc. el partido que cumple estas características en el año x, posiblemente en la siguiente votación aún y aunque se busquen las mismas características, será otro partido el que las cumpla. Como resultado, los partidos vienen caracterizados más que por sus nombres y etiquetas, por características ocultas generadas a partir de la actividad histórica del partido y de sus promesas de futuro.

Como resultado, el objetivo será colocar ambos conjuntos de datos en la misma situación. Como los partidos representados pueden haber cambiado de un período a otro y además para permitir su visualización y realizar la comparación a partir de allí, los conjuntos de datos sobre los que se trabajará serán los obtenidos a partir de las herramientas de ACP. Es decir, para cada conjunto de datos electorales se tendrá una matriz de datos X, de dimensiones nxr, n hará referencia a los municipios, por lo que se tendrán tantas filas como número de municipios, mientras que r hará referencia a las coordenadas del ACP, es decir, dos columnas, una para la dimensión 1 y otra para la dimensión 2 que contendrán

la coordenada de cada municipio en el mapa de individuos de las componentes principales.

4.1. Centralización

Ahora, una vez se cuenta con ambos conjuntos de datos nuevos. Estos datos se encuentran separados en el espacio debido al movimiento en el tiempo que ha afectado a todo el conjunto de datos. Para facilitar la lectura y a modo de generalizar, al conjunto de datos de 2015 se le llamará por la matriz X, mientras que al conjunto de datos de 2017 se le llamará por la matriz Y. Lo primero, será poder poner un conjunto sobre otro para poder compararlos, por lo que se trasladarán tanto X como Y al mismo punto, el origen.

Para centralizar la matriz de datos X o Y, se hará uso del vector 1, compuesto por tantas filas como individuos tenga la matriz que se quiere centrar y solo una columna, todos sus elementos son 1:

$$\begin{pmatrix} 1_{11} = 1 \\ 1_{21} = 1 \\ \cdots \\ 1_{N1} = 1 \end{pmatrix}$$

se comienza por la multiplicación del vector 1 transpuesto por la matrix X o Y. El resultado será una nueva matriz con una sola fila y tantas columnas como coordenadas. Los valores de cada coordenada de la única fila es el sumatorio de todos los valores anteriores de aquella fila. El siguiente paso será la multiplicación del vector 1 por la matriz resultante anteriormente con el fin de obtener el mismo resultado de antes pero en vez de tener una fila se tendrían tantas filas como individuos, todas iguales entre ellas. Si entonces se divide esta matriz por N (número de municipios), entonces se conseguiría para cada coordenada el valor medio.

Por otro lado, se debe realizar la multiplicación de la matrix identidad por la matrix X o Y, como resultado, se conseguiría la misma matriz X o Y solo con valores en la diagonal. Finalmente, se restaría esta matriz con la obtenida anteriormente. El resultado final sería la matrix inicial con la que se contaba habiéndola restado por el valor medio de cada coordenada. Resumiendo este procedimeinto, sería el siguiente:

$$(I - \tfrac{1}{N} 1 1^t) \cdot X = \tilde{X}$$

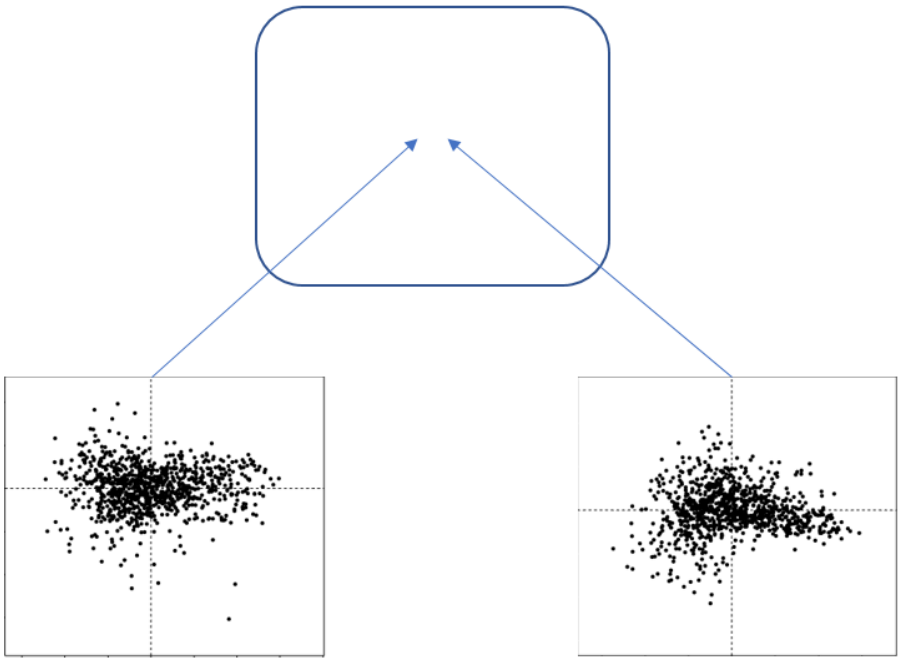

Figura 4.1: Centralización

Procedimiento realizado igual para X e Y. El siguiente paso, constará en corregir el ajuste de los partidos mediante la rotación de la matriz.

4.2. Rotación

Este apartado buscará rotar $\tilde{Y}$ de tal forma que encaje al máximo con $\tilde{X}$. Este procedimiento es equivalente a rotar a $\tilde{X}$ para encajar con $\tilde{Y}$. Cabe destacar además que $\tilde{X}$ e $\tilde{Y}$ son conjuntos fijos, constantes.

Como se requiere minimizar la separación de las matrices, y teniendo en cuenta que se quiere rotar $\tilde{Y}$, se utilizará la norma de matrices por simplicidad:

$$\Phi(S) = ||\tilde{X} - \tilde{Y}S||^2$$

Teniendo en cuenta que la norma de matrices es:

$$||A||^2 = tr(AA^2) = \sum_{i\alpha}(\alpha_{i\alpha})^2$$

Se desarrollará la fórmula:
$\Phi(S) = tr[(\tilde{X} - \tilde{Y}S)(\tilde{X} - \tilde{Y}S)^t] = tr(\tilde{X}\tilde{X}^t + \tilde{Y}\tilde{Y}^t - \tilde{Y}S\tilde{X}^t - \tilde{X}S^t\tilde{Y}^t) = tr(\tilde{X}\tilde{X}^t) + tr(\tilde{Y}\tilde{Y}^t) - tr(\tilde{Y}S\tilde{X}^t) - tr(\tilde{X}S^t\tilde{Y}^t)$

Con $tr(\tilde{Y}S\tilde{X}^t) = tr(\tilde{Y}S\tilde{X}^t)^t$; entonces:

$$\Phi(S) = tr(\tilde{X}\tilde{X}^t) + tr(\tilde{Y}\tilde{Y}^t) - 2tr(\tilde{Y}S\tilde{X}^t)$$

Para reducir al máximo el espacio entre $\tilde{X}$ e $\tilde{Y}$ se debe minimizar $\Phi(S)$. No obstante, minimizar $\Phi(S)$ equivale a maximizar $\Psi(S)$:

$$\Psi(S) = tr(\tilde{Y}S\tilde{X}^t) = tr(\tilde{X}^t\tilde{Y}S)$$

Al ser $\tilde{X}$ e $\tilde{Y}$ fijos, $\tilde{X}^t\tilde{Y}$ es constante (C) de dimensión rxr. Esta matriz C se puede descomponer en sus valores singulares: $C = ULV^t$, con L siendo una amtriz diagonal $(l_1, l_2, \cdot\cdot, l_r)$ y V y U siendo matrices de vectores propios de $\tilde{Y}^t\tilde{X}\tilde{X}^t\tilde{Y}$ y $\tilde{X}t\tilde{Y}\tilde{Y}^t\tilde{X}$.

Antes de proseguir, se definirá la descomposición en valores singulares. Teniendo presente inicialmente a $A \in nM_{mxn}(R)$, $k = min(m, n)$ y $r = rango(A)$ con $(0 \leq r \leq k)$, entonces puede demostrarse que las matrices cuadradas y simétricas AA^t y A^tA poseen todos los valores propios positivos y con la misma multiplicidad. Si A es simétrica entonces $AA^t = A^tA = A^2$. Si A es ortogonal, entonces sus valores singulares serán todos ellos iguales a 1 ya que $AA^t = A^tA = I_n$

Luego, si $A \in M_{mxn}(R)$ y λ un valor propio de la matriz AA^t o de A^tA, entonces el valor $l = \sqrt{\lambda}$ diremos es un valor singular de la matriz. A y A^t tienen los mismo valores singulares.

Entonces, sea $A_{mxn}(R)$ con $0 \leq r = rango(A)$. Luego, existen unas matrices $U \in M_{mxr}(R)$, $L \in M_{rxr}(R)$ y $V \in M_{nxr}(R)$ tales que:

$$A = ULV^t$$

Donde; **L** es una matriz diagonal, $L = diag(l_1, \cdot\cdot\cdot, l_r)$, con $l_1 \geq \cdot\cdot \geq l_r \geq 0$ donde $l_i = \sqrt{\lambda_i}$ para unos valores $\lambda_i \geq 0$ que son los valores propios no nulos correspondientes a la matriz AA^t o A^tA, considerados tantas veces como indique su multiplicidad; **U**, escrito a partir de sus r vectores columna, $U = (u_1, \cdot\cdot, u_r)$, resulta que estos son vectores unitarios y ortogonales entre sí, es decir, son vectores propios de AA^t correspondientes a los valores propios de λ_i; y V también escrita a partir de sus vectores columna, $V = (v_1, \cdot\cdot, v_r)$, resulta que estos son vectores unitarios y ortoganles entre sí, es decir, son vectores propios de A^tA correspondientes a sus valores propios λ_i. Finalmente, se le denominará a $A = ULV^t$ como la descomposición de la matriz A en valores singulares.

La expresión conseguida puede escribirse en términos de los vectores columna de U y V así como los vectores singulares de l_i como:

$A = \sum_{i=1}^{r} l_i u_i v_i^t$ Descomposición espectral generalizada de la matriz A

Entonces, siguiendo con el ejercicio, se obtiene:

$$\Phi(S) = tr(ULV^tS) = tr(LV^tSU)$$

Si denominamos $V^tSU = T$, entonces $\Phi(S) = tr(LT)$ equivalente a $\Phi(S) = l_{ij}t_{ji} = \sum_{i=1}^{r} l_i ti$ debido a que $\sum_{j=1}^{r}(t_{ij})^2 = 1$

Además, se puede conseguir que $\Psi(S) = \sum_{i=1}^{r} l_i t_{ii} \leq \sum_{i=1}^{n} l_i$ en el caso en que $T = I$. Pero, si T=I, entonces $I = V^tSU$; por lo tanto, $VU^t = S$.

Finalmente, para conseguir reducir al máximo el espacio entre $\tilde{\text{X}}$ e $\tilde{\text{Y}}$ y así conseguir la rotación que mejor encaje ambos conjuntos de datos, **se deberá conseguir la matriz S equivalente a** VU^t siendo V y U matrices de vectores propios. Esta matriz se deberá aplicar sobre $\tilde{\text{Y}}$ si se sigue el procedimiento descrito.

4.3. Visualización y comparación

Finalmente, se contará con dos conjuntos de datos: $\tilde{\text{X}}$ compuesto por los datos del 2015 transformados a coordenadas de dos dimensiones según los resultados del análisis de componentes principales realizado y habiendo sido centrado al punto de origen tras restar el valor medio de las coordenadas a sus observaciones; e $\tilde{\text{Y}}$S, la cual representa los datos del 2017 y que, además de haber pasado por el mismo procedimiento que $\tilde{\text{X}}$, también se ha multiplicado por la matriz S con tal de rotar la matriz y conseguir la máxima cercanía entre ambos resultados

El siguiente paso será la representación gráfica de ambos conjuntos de datos adapatados en un mismo plano, con la indicación del movimiento de los individuos en el espacio.

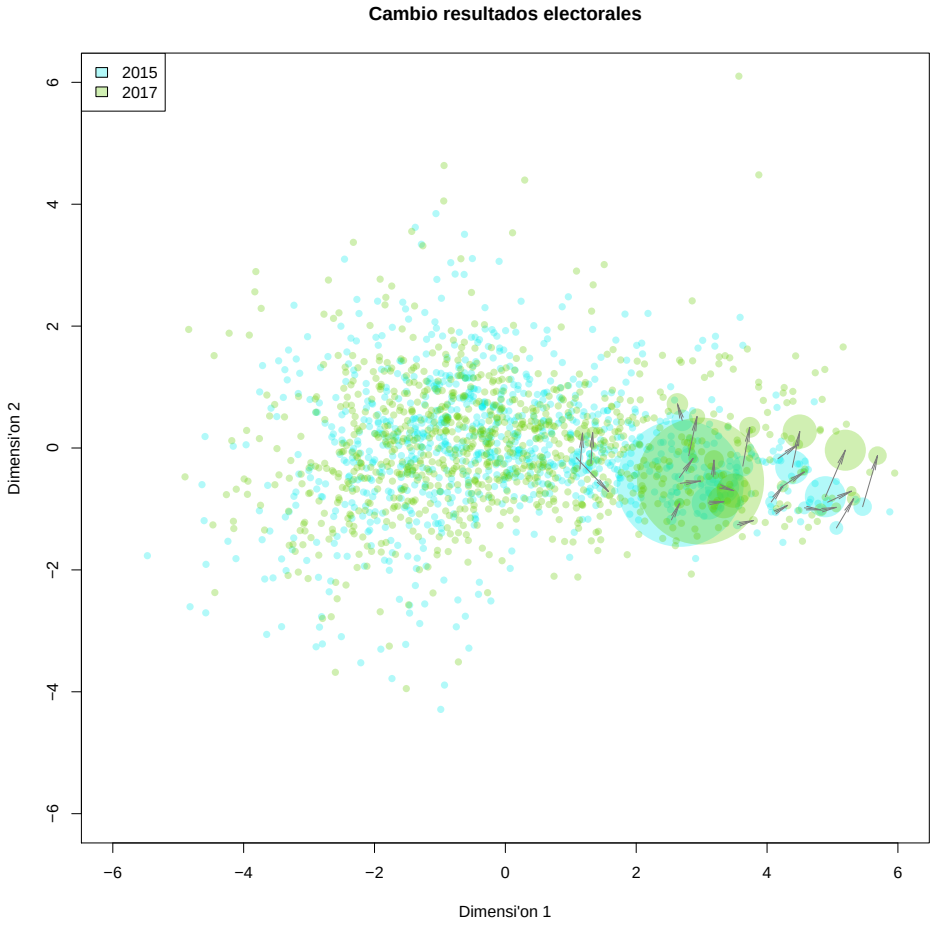

Figura 4.2: Comparación de resultados

Esta figura muestra a todos los municipios representados en el espacio de dos dimensiones. Ambos conjuntos de datos fueron representados inicialmente en las dos dimensiones gracias al procedimiento de análisis de componentes principales. El procedimiento representa a cada individuo en las dos dimensiones a través de coordenadas obtenidas sobre nuevos ejes X e Y. Luego se centralizaron ambos conjuntos de datos en el punto de origen. Finalmente se procedió a rotar el conjunto de 2017 para conseguir el máximo encaje con el conjunto de 2015. El resultado final ha sido este mapa de puntos donde se muestra en azul claro la nueva localización de los municipios del 2015, y en verde, la nueva localización de los datos del 2017. El tamaño de los puntos se ha ajustado al tamaño de los municipios (valor obtenido también desde el IDESCAT [9]), siendo este proporcional al tamaño del mayor municipio, que en este caso es Barcelona. No obstante, debido a la gran diferencia entre el municipio de mayor tamaño y lo de menos, se ha reducido el tamaño representado de Barcelona a la mitad y se ha aumentado el de los menores municipios hasta ser visibles en la nube de puntos. A modo de comprobación de que el procedimiento sea correcto, se ha realizado otra vez todo el procedimiento sustituyendo los datos del 2017 por los del 2015 también y se ha obtenido una nuve puntos de un solo color, indicando que todos

los municipios se encuentran situados igual e indicando que se cumple este check de errores.

Como se puede observar también en la figura, se ha representado con flechas el movimiento de los mayores municipios de un año a otro. Para ampliar esta información, a continuación se muestra el cambio de coordenadas de los 10 municipios más grandes y más pequeños:

Cambio mayores municipios			
Rank	Municipio	2015	2017
1	Barcelona	(2.67,-0.58)	(2.98,-0.54)
2	Hospitalet de Llobregat	(4.89,-0.80)	(5.20,-0.04)
3	Terrassa	(3.31,-0.64)	(3.49,-0.70)
4	Badalona	(4.39,-0.32)	(4.50,0.27)
5	Sabadell	(3.10,-0.90)	(3.34,-0.88)
6	Lleida	(2.70,0.45)	(2.63,0.72)
7	Tarragona	(3.63,-0.30)	(3.74,0.34)
8	Mataró	(3.18,-0.47)	(3.19,-0.20)
9	Santa Coloma de Gramenet	(5.46,-0.96)	(5.69,-0.12)
10	Reus	(2.80,-0.13)	(2.93,0.52)

Cambio menores municipios			
Rank	Municipio	2015	2017
938	Bausen	(3.21,1.83)	(4.30,0.74)
939	Savallá del Comtat	(-1.43,-0.05)	(-1.32,-1.08)
940	Quar	(-1.26,-0.32)	(-4.22,1.88)
941	Senan	(-2.42,-1.78)	(-2.60,-1.31)
942	Cava	(-3.58,0.65)	(-4.84,1.94)
943	Forés	(-1.31,0.50)	(0.11,3.53)
944	Fígols	(-1.28,3.34)	(-0.32,2.02)
945	Febró	(-1.73,-3.78)	(-2.77,-1.07)
946	Sant Jaume de Frontayá	(-3.42,-2.93)	(-3.00,-0.33)
947	Gisclareny	(-4.58,-2.70)	(-4.26,-0.77)

De estos cuadros y de la visualización gráfica se puede destacar que los municipios grandes tienden a moverse hacia un valor mayor de la componente 1, mientras que el comportamiento de los municipios más pequeños es más inestable. No obstante, estos resultados no muestran más información que la cercanía entre los votos de los municipios y si se han desplazado una gran distancia entre un período y otro (donde se ve que los

municipios de mayor tamaño tienden a desplazarse menos distancia). Por eso, para conseguir realmente resultados de esta visualización, deberán ser representados los partidos.

La nube de puntos conseguida se encuentra dentro de una circumferencia, donde en algún punto deberían estar situados los partidos, los cuales también se pueden mover de un año a otro indicando su cambio de rumbo de acuerdo a la percepción de los ciudadanos. Si por ejemplo un municipio dedicase en una votación de un año concreto el 100 por ciento a un partido, ese municipio se encontraría situado en las mismas coordenadas que el el partido al que vota. El objetivo de tener los partidos representados es ver si los municipios que ya han sido representado tienden a desplazarse hacia ellos o alejarse. Un resultado que es fácil esperar es que los partidos se encuentren alejados de la nube de puntos, al ser estos el caso extremo de la situación.

Para representar a los partidos políticos, se generarán nuevos municipios ficticios que dedidacrán el 100 por ciento de sus votos al partido que quieren representar. Estos municipios ficticios no se pueden haber introducido junto al resto para generar los anteriores procesos de ACP o clustering porque podrían influenciar en los resultados, por lo que se aplican posteriormente los mismo procedimientos sobre ellos para luego ser visualizados en el mismo plano. Los municipios ficticios tendrán el mismo nombre del partido que representan.

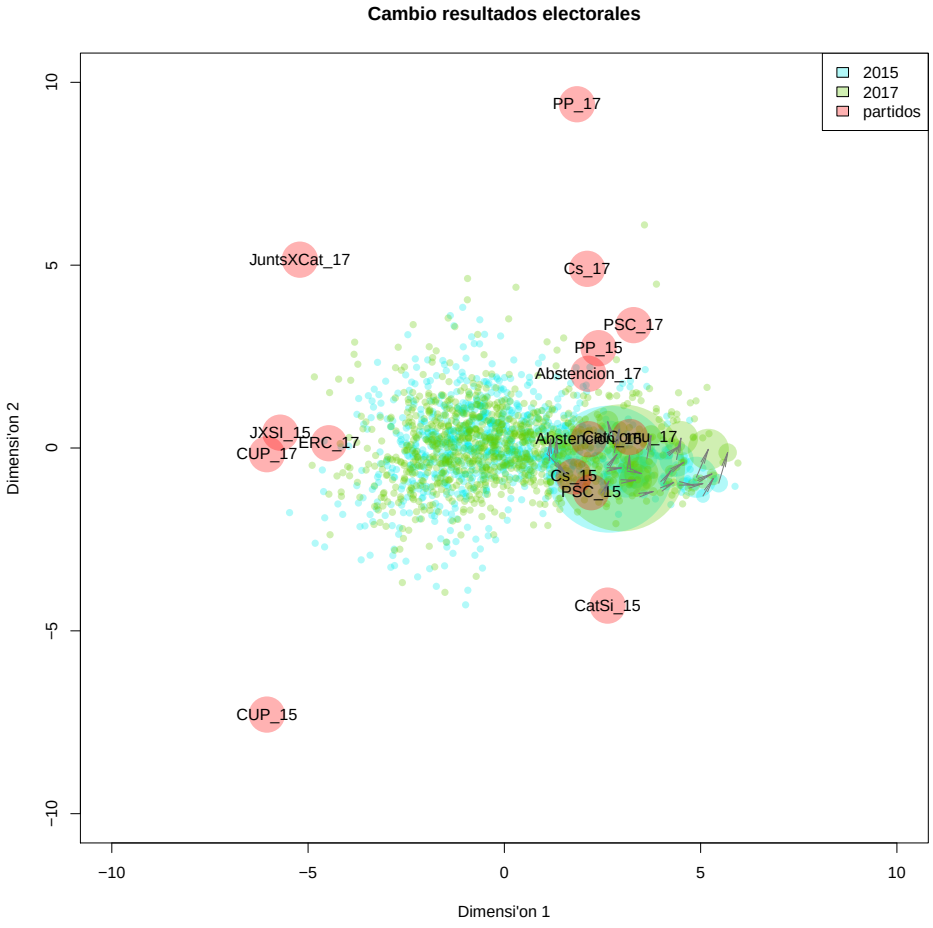

Figura 4.3: Comparación de resultados con partidos

Este es el resultado final obtenido. Este mapa está compuesto, igual que el anterior mapa, con los dos conjuntos de datos pero incluyenndo ahora la situación de los partidos. Es interesante observar que todos los partidos han cambiado de un año a otro. Esto ya permite la interpretación de los resultados. Utilizando los municipios grandes, que han sido los marcados con flechas para señalar su desplazamiento, se puede ver que tienden la mayría a moverse en dirección positiva tanto del eje x como del eje y, es decir, para el caso de este subconjunto de municipios, los votos se acercarían más a la abstención y CatComú. Si nos fijamos en el movimiento de los Comuns de un año a otro, ellos se han movido exactamente en esta dirección, por lo que, o bien, el partido se ha adaptado perfectamente a la evolución de la sociedad y ha pasado a ofrecer lo que esta busca, o bien, la sociedad ya apoyaba a este partido y solo se ha visto arrastrada por el movimiento del partido, o bien ambas cosas. Este efecto encaja con la realidad, ya que Barcelona está gobernada por este partido político.

Otro efecto a observar es la clara separación entre los partidos independentistas y el resto de partidos, tal y como se encontró en el análisis individual. En este caso se puede ver

que se sitúan en la zona más negativa del eje x, más cercanos a los municipios pequeños. Estos partidos han experimentado una tendencia hacia el eje positivo, igual que la tendencia general de los municipios. El resto de partidos como la abstenci<U+00BA>'on se han mantenido en el lado positivo del eje x, donde al igual que los partidos independentistas, han experimentado un cambio a una zona mayor del eje y. También resulta importante observar que dentro del bloque constitucionalista, hay un mayor grado de polarización entre los partidos al haber más distancia entre ambos extremos en el 2017 que en el 2015.

Realizando el mismo estudio para los municipios pequeños, se puede ver que su comportamiento és más inestable, y además tienden a acercarse mucho a un partido, indicando que el municipio vota en masa a ese partido. Si nos fijamos en el caso anterior, Barcelona se desplaze hacia CatComú, situandose muy cerca de este.

Este análisis, además de poner a disposición una herramienta de comparación de conjuntos de individuos en el tiempo, ha mostrado para este caso la situación que separa a los ciudadanos en Catalunya en los tiempos recientes, la polarización entre independentistas y unionistas. Esta situación, aún y ser de procedencia histórica, se ha incrementado de forma muy elevada en la última década, teniendo efectos en varios aspectos de la sociedad como la identidad nacionalista, o la predominación de una lengua sobre otra [11]. Como resultado, esta herramienta creada puede ayudar a un estudio más exhaustivo y específico de la región, utilizando datos de tiempos más lejanos con el fin de ir construyendo la evolución en el entorno social desde los datos políticos, los cuales cuentan en general con un buen nivel de sencillez y facilidad de conseguir.

Finalmente, mencionar que la abstención se sitúa dentro de la nube de puntos. Esto puede er debido a dos razones, por una parte, nos hace falta una dimensión más para poder separar la bastención ya que esta no se encontraría ahora e la nube de puntos, sino que se situaría por encima en el plano z. La otro razón podría ser el carácter de neutralidad que representa este voto, por consiguiente, situandolo en vez de en un extremo, en el centro.

Capítulo 5

GENERALIZACIÓN

Tras conseguir resultados con el capítulo anterior, este capítulo buscará generalizar la base teórica de la metodología utilizada. Los procedimientos anteriormente utilizados se pueden ver limitados en varios casos, como podría ser el uso de más años (más conjuntos de datos) o mucha mayor vairación de columnas (partidos políticos representados muy diferentes entre conjuntos de datos). Existen varios elementos del procedimiento de los cuales se podría estudiar una forma de generalización, en este espacio se determinará una metodología para llevar distintos conjuntos de datos con distintas columnas a un mismo espacio, para ser representados de la misma forma que los resultados anteriores.

Supongamos un espacio que represente a un conjunto de datos, R^p , $p \in N \setminus \{0\}$, con un producto escalar $< \cdot \, , \, \cdot >_1$ dado, sin pérdida de generalidad, por la matriz identidad I_p de orden $p \times p$, respecto la base canónica. Representando a otro conjunto de datos, sea $\mathbb{R}^k \, k, \in N \setminus \{0\}$ con un producto escalar $< \cdot \, , \, \cdot >$ dado, sin pérdida de generalidad, por la matriz identidad I_k de orden $k \times k$, respecto la base canónica.

Sea $f : R^p \longrightarrow R^k$ una aplicación afín que *conserve* las distancias Euclídeas asociadas a los productos escalares respectivos $< \cdot \, , \, \cdot >_1$ y $< \cdot \, , \, \cdot >$. Si usamos la representación matricial de f respecto las bases canónicas respectivas de R^p y R^k, y si denotamos por $M_{k\times p}(R)$ las matrices de k filas y p columnas de coeficientes reales, entonces, tendremos: $f(x) = Px + c$ donde $x = (x_i)_{p\times 1} \in M_{p\times 1}(R)$, $P = (p_{ij})_{k\times p} \in M_{k\times p}(R)$ y $c = (c_i)_{k\times 1} \in M_{k\times 1}(R)$. Dados $x, y \in R^p$ si hacemos $w = y - x$ y $w = (w_i)$, la conservación la distancia se puede formular, matricialmente, como

$$w^t w = w^t P^t P w \ , \quad \forall w \in M_{p\times 1}(R) \tag{5.1}$$

Además, si la distancia se debe conservar, f ha de ser inyectiva, lo que, por el Teorema de la dimensión exige $\dim f(R^p) = p$, $P = p$ y por tanto $p \leq k$. Si descomponemos P en valores singulares, tendremos

$$P = ULV^t \tag{5.2}$$

Donde U es una matriz de orden $(k \times p)$ cuyas columnas $u_1, \ldots, u_p$ son p vectores propios de PP^t, ortonormales entre sí y R^k, asociados a los p valores singulares $l_1, \cdots, l_p$, definidos como $l_i = \sqrt{\lambda_i}$, $i = 1, \ldots, p$ siendo $\lambda_i > 0$ los valores propios comunes de PP^t y P^tP, mientras que V es una matriz de orden $(p \times p)$ cuyas columnas son p vectores propios de P^tP, ortonormales en R^p, asociados a los mismos valores singulares $l_i = 1, \ldots, p$. Combinando (5.1) y (5.2) obtendremos:

$$w^tw = w^tVL^2V^tw \ , \quad \forall w \in M_{p\times 1}(R) \tag{5.3}$$

Si llamamos

$$y = V^tw \tag{5.4}$$

Como V es ortogonal conserva la norma Euclídea ordinaria, $|w| = |y|$, y podemos escribir

$$y^ty = y^tL^2y \ , \quad \forall y \in M_{p\times 1}(R) \tag{5.5}$$

Pero como además y es arbitrario, puesto que w también lo es, resulta que (5.4), solo podrá darse si $l_i^2 = 1$, $i = 1, \ldots, p$. Además, si introducimos la matriz $Q \in M_{k\times p}(R)$ definida por $Q = UV^t$ al ser $Q^tQ = I_p$ estará claro que en las columnas de Q se identifican p vectores ortonormales de R^k. Por tanto, las matrices Q forman la llamada *variedad de Stiefel* $V_p(R^k) = \{Q \in M_{k\times p}(R) \mid Q^tQ = I_p\}$. Por tanto, las aplicaciones afines f que conserven la distancia serán de la forma

$$f(x) = Qx + c \ , \quad \forall x \in M_{p\times 1}(R) \tag{5.6}$$

donde $Q \in V_p(R^k)$ y $c \in M_{k\times 1}(R)$.

Espacio Común

Consideremos observaciones sobre los mismos n individuos (municipios) $\omega_1, \ldots, \omega_n$ realizadas sobre dos juegos de variables $X_1, \ldots, X_p$ e $Y_1, \ldots Y_q$, distintos grupos de partidos políticos, que miden aspectos de una misma realidad pero no son idénticas entre sí, por ejemplo unas mismas o parecidas variables medidas sobre los mismos individuos pero realizadas en dos tiempos distintos, en el caso que podamos suponer que las variables hayan podido cambiar en el tiempo. Los datos obtenidos los podremos disponer en dos matrices $X \in M_{n\times p}(R)$ e $Y \in M_{n\times q}(R)$, matrices cuyas columnas son los valores de las

diversas variables sobre los n individuos y cuyas filas x_i^t e y_i^t con $i = 1, \ldots, n$ son los valores de las todas las variables X_i y todas las variables Y_i sobre el individuo ω_i. Consideremos R^p y R^q con el producto escalar Euclídeo ordinario. Sean f_p y f_q aplicaciones afines que conservan la distancia dadas por (5.6), i.e.,

$$f_p(x) = Q_p x + c_p\,, \quad \text{y} \qquad f_q(y) = Q_q y + c_q \tag{5.7}$$

Donde $Q_p \in V_p(R^k)$ $, Q_q \in V_q(R^k)$ $, c_p \in M_{k\times 1}(R)$ y $c_q \in M_{k\times 1}(R)$ donde $k \geq \text{máx}\,\{p, q\}$, representando *de facto* a cada individuo ω_i como dos puntos en R^k, obtenidos por $f_p(x_i)$ y $f_q(x_i)$.

El criterio que usaremos para determinar estas transformaciones afines es el que podemos formular de la siguiente manera (criterio de optimización):*Determinaremos las aplicaciones afines f_p y f_q de forma que la suma de cuadrados de las interdistancias, entre los puntos que representan un mismo individuo ω_i con los dos juegos de variables $X_1, \ldots, X_p$ y $Y_1, \ldots Y_q$, i.e., $f_p(x_i)$ y $f_q(y_i)$, sea mínima.*

La distancia al cuadrado entre $f_p(x_i)$ y $f_q(x_i)$ matricialmente expresada será, si introducimos como notación, $r_i = Q_p x_i$, $z_i = Q_q y_i$, y $\delta = c_p - c_q$ igual a

$$\begin{aligned} d_i^2 &= ((Q_p x_i + c_p) - (Q_q y_i + c_q))^t\,((Q_p x_i + c_p) - (Q_q y_i + c_q)) \\ &= ((r_i - z_i) + \delta)^t\,((r_i - z_i) + \delta) = |r_i - z_i|^2 + |\delta|^2 - 2 < r_i - z_i, \delta > \end{aligned} \tag{5.8}$$

El criterio de optimización (1) implica minimizar

$$\begin{aligned} \sum_{i=1}^{n} d_i^2 &= \sum_{i=1}^{n} \left(|r_i - z_i|^2 + |\delta|^2 - 2 < r_i - z_i, \delta >\right) \\ &= \sum_{i=1}^{n} |r_i - z_i|^2 + n|\delta|^2 - 2 < \sum_{i=1}^{n} (r_i - z_i), \delta > \end{aligned} \tag{5.9}$$

Pero por la desigualdad de Schwarz, tendremos que se alcanzará un mínimo si hacemos $\delta = \lambda \sum_{i=1}^{n} (r_i - z_i)$, por tanto

$$\begin{aligned}\sum_{i=1}^{n} d_i^2 &\geq \sum_{i=1}^{n} |r_i - z_i|^2 + n\lambda^2 \left|\sum_{i=1}^{n}(r_i - z_i)\right|^2 - 2\lambda \left|\sum_{i=1}^{n}(r_i - z_i)\right|^2 \\ &\geq \sum_{i=1}^{n} |r_i - z_i|^2 - \frac{1}{n} \left|\sum_{i=1}^{n}(r_i - z_i)\right)|^2 \\ &= n\left(\frac{1}{n}\sum_{i=1}^{n} |r_i - z_i|^2 - \left(\frac{1}{n} \left|\sum_{i=1}^{n}(r_i - z_i)\right|\right)^2\right) \end{aligned} \tag{5.10}$$

Polinomio de segundo grado en λ cuyo valor mínimo se alcanza cuando escogemos $\lambda = 1/n$, siendo pues entonces (5,10) el ínfimo de (5.8). Por tanto,

$$\delta = \frac{1}{n}\sum_{i=1}^{n}(r_i - z_i) \tag{5.11}$$

observar que (5.10) puede ser escrita como

$$\sum_{i=1}^{n} d_i^2 \geq n s^2_{|f_p(x) - f_q(y)|} \tag{5.12}$$

donde $s^2_{|f_p(x)-f_q(y)|}$ es la varianza muestral, basada en una muestra de tamaño n, de la variable $|f_p(x) - f_q(y)|$.

Una vez hemos determinado δ, tendremos que determinar cómo hemos de minimizar las diferencias en la variabilidad entre la *estructura de datos* determinada por $X_1, \ldots, X_p$ y la *estructura de datos* determinada por $Y_1, \ldots, Y_q$, punto a punto, cuantificada por la varianza $s^2_{|f_p(x)-f_q(y)|}$. Observemos:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} |r_i - z_i|^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(Q_p x_i - Q_q y_i\right)^t \left(Q_p x_i - Q_q y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(x_i^t x_i + y_i^t y_i - 2 y_i^t {Q_q}^t Q_p x_i\right) \\ &= \frac{1}{n}\left(XX^t\right) + \frac{1}{n}\left(YY^t\right) - \frac{2}{n}\left(Y {Q_q}^t Q_p X^t\right) \end{aligned} \tag{5.13}$$

Puesto que $Q_p^t Q_p = I_p$, $Q_q^t Q_q = I_q$ y además, si introducimos el vector $1_n = (1)_n \in M_{n\times 1}(R)$, i.e., un vector columna n-dimensional cuyas componentes sean todas ellas iguales a 1, resultará

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}(r_i - z_i)\right|^2 &= \left|\frac{1}{n}\ 1_n^t\left(XQ_p^t - YQ_q^t\right)\right| \\ &= \frac{1}{n^2}\ 1_n^t\left(XQ_p^t - YQ_q^t\right)\left(Q_pX^t - Y_qY^t\right)1_n \\ &= \frac{1}{n^2}\Big(XX^t1_n1_n^t\Big) + \frac{1}{n^2}\Big(YY^t1_n1_n^t\Big) - \frac{2}{n^2}\Big(YQ_q{}^tQ_pX^t1_n1_n^t\Big) \end{aligned} \tag{5.14}$$

Si llamamos $H_n = I_n - \frac{1}{n}1_n1_n^t$, combinando (5.13) y (5.14) obtenemos

$$\begin{aligned} s^2_{|f_p(x)-f_q(y)|} &= \frac{1}{n}\left(XX^tH_n\right) + \frac{1}{n}\left(YY^tH_n\right) - \frac{2}{n}\left(YQ_q{}^tQ_pX^tH_n\right) \\ &= \frac{1}{n}\left(X^tH_nX\right) + \frac{1}{n}\left(Y^tH_nY\right) - \frac{2}{n}\left(Q_q{}^tQ_pX^tH_nY\right) \\ &= (S_{XX}) + (S_{YY}) - 2\left(Q_pS_{XY}Q_q{}^t\right) \end{aligned} \tag{5.15}$$

Donde

$$\begin{aligned} S_{XX} &= \frac{1}{n}\left(X^tH_nX\right) \\ S_{YY} &= \frac{1}{n}\left(Y^tH_nY\right) \end{aligned} \tag{5.16}$$

Es la estimación por el método de los momentos de la matriz de varianzas y covarianzas de $X_1, \ldots, X_p$ y $Y_1, \ldots, Y_q$ respectivamente, mientras que S_{XY} es una estimación de la matriz de covarianzas entre ambos grupos de variables, $X_1, \ldots, X_p$ e $Y_1, \ldots, Y_q$

$$S_{XY} = \frac{1}{n}\left(Y^tH_nY\right) \tag{5.17}$$

Por tanto minimizar (5.15) va a ser equivalente a maximizar la expresión

$$\left(Q_pS_{XY}Q_q{}^t\right) \tag{5.18}$$

Para ello efectuemos la descomposición de $S_{XY} \in M_{p\times q}(R)$, en valores singulares:

$$S_{XY} = ULV^t \tag{5.19}$$

Donde, sin pérdida de generalidad, podemos suponer $l_1, \geq l_2, \geq \ldots \geq l_m > 0$, $m \leq$ mín$\{p, q\}$, U es $M_{p\times m}(R)$, V es $M_{q\times m}(R)$ y tales que $U^tU = I_m$ y $V^tV = I_m$, y maximizar (5.18) puede expresarse como

$$\begin{aligned}\left(Q_p S_{XY} {Q_q}^t\right) &= \left(Q_p U L V^t {Q_q}^t\right) \\ &= \sum_{i=1}^{m} l_i < w_i, z_i > \\ &= \sum_{i=1}^{m} l_i \end{aligned} \tag{5.20}$$

Donde $w_i = Q_p u_i$ y $z_i = Q_q v_i$, $i = 1, \ldots, m$. Además, al ser $Q_p \in V_p(R^k)$ = transforma vectores los vectores unitarios y perpendiculares $u_i \in M_{p\times 1}(R)$ en vectores unitarios y perpendiculares $w_i \in M_{k\times 1}(R)$. De forma análoga, al ser $Q_q \in V_q(R^k)$ = transforma vectores los vectores unitarios y perpendiculares $v_i \in M_{q\times 1}(R)$ en vectores unitarios y perpendiculares $w_i \in M_{k\times 1}(R)$.

La igualdad en la desigualdad (5.20) se alcanzará si elegimos Q_p y Q_q de forma que

$$Q_p u_i = \pm Q_q v_i \quad i = 1, \ldots, m \tag{5.21}$$

Hay infinitas posibilidades, una de ellas es escoger Q_q de forma que

$$Q_q = \frac{1}{u_i^t Q_p^t v_i} Q_p u_i u_i^t Q_p^t \quad i = 1, \ldots, m \tag{5.22}$$

mientras que basta escoger Q_p de manera que $u_i^t Q_p^t v_i \neq 0$.

Futura investigación

Esta solución puede ser introducida en la práctica para el caso estudiado en este proyecto rompiendo con algunas limitaciones de la metedología anterior al poderse escoger cualquier otro conjunto de datos con características muy diferentes. Es por esto que resultaría interesante realizar de nuevo el proceso práctico utilizando métodos más generales como el recién explicado y ponerlos a prueba utilizando varios conjuntos de datos provenientes de espacios temporales muy diferentes para contar con mayores diferencias posibles.

Finalmente, también se podrían ampliar los resultados conseguidos, además de generalizando los procedimientos, aumentando la capacidad de los métodos de visualización utilizados introduciendo una tercera dimensión, consiguiendo así representar mucha más información en el mapeo de datos realizado en el presente.

Capítulo 6

CONCLUSIONES

Este proyecto ha buscado una metodología que permitiera la comparación de diferentes conjuntos de datos electorales de un mismo territorio en tiempos distintos. Para ello, se han escogido dos conjuntos de datos pertenecientes a las elecciones al Parlament de Catalunya de los años 2015 y 2017. Estos conjuntos de datos servirían como visualización y comprobación a medida que se iba desarrollando la metodología.

Previamente a establecer el procedimiento de comparación entre ambos conjuntos se ha realizado un estudio por separado de ambos conjuntos de datos. Los resultados han mostrado que la cantidad de partidos representados en el Parlament de Catalunya varia alrededor de 7, a su vez, se ha realizado un proceso de clasificación (clustering) y profiling de los datos para tratar de identificar características comunes que, no solo unieran grupos de municipios, sino también grupos de variables (partidos). El resultado de los procesos de clasificación ha mostrado en ambos años una brecha clara entre dos conjuntos de partidos, los independentistas y los constitucionalistas. En el primer grupo se puede encontrar a Junts pel Sí, Esquerra, Junts per Catalunya y la CUP, mientras que en el segundo grupo se pueden encontrar a los otros partidos que además también se encuentran en el resto del territorio español, PP, PSC, Cs, CatComú y Cat sí que es Pot. Este resultado cumple con el contexto que rodea a los datos, ya que no solo ha sido una separación que se puede encontrar a lo largo de la historia, sino que en los últimos años se ha visto reforzada en la sociedad debido a la convocatoria de consultas y Referéndums y a la hoja de ruta independentista.

Finalizando el análisis individual y ya preparando la comparación de ambos conjuntos de datos, se ha realizado el análisis de componentes principales, lo que permitiría la comparación de dos conjuntos de datos con diferente número de columnas, y la visualización de estos en un plano 2D. Esto ha creado dos nubes de puntos que representan a los municipios en sus nuevas coordenadas. A pesar de no dar mucha información y de no

poder ser comparadas las dos nubes de puntos, se han marcado los municipios con cuatro colores distintos según en que cuartil del tamaño de los municipios según población se encuentran. Los resultados de la coloración de los municipios han mostrado la separación de ideología política clara entre los cuatro grupos formados. Esta separación ha mostrado a los cuatro grupos situarse de forma ordenada, al estar el grupo de los municipios más pequeños en un lado, el de los más grandes en el otro y los grupos intermedios en el medio, más entremezclados entre sí.

La metodología de comparación utilizada como base en este estudio, consta de acercar ambos conjuntos de datos lo máximo posible dentro del espacio euclídeo. Para ellos, primero se han centrado los dos conjuntos en el origen, restando a cada una de sus variables, el valor medio. Posteriormente se ha realizado una rotación de un conjunto sobre otro para conseguir el encaje máximo entre estos. Estos dos movimientos buscarían eliminar al máximo los cambios de los partidos y los municipios provenientes de los otros factores del ecosistema que influencia y afecta a los datos, aislando finalmente solo los cambios que ha generado el partido o el municipio en sí. El resultado, ha sido la representación gráfica de las dos nubes de puntos anteriores pero ahora situadas en la misma situación una sobre otra, mostrando algunos desplazamientos en el espacio de los municipios y distinguiendoles a partir de la población de cada uno. Se han introducido en este mapa también los partidos como nuevos puntos de la nube de puntos. Tanto municipios como partidos han sufrido cambios de un año a otro, al haber sido todos desplazados. Los municipios grandes han mostrado un desplazamiento corto, y en general los más grandes tienden a ir en la misma dirección, mientras que el desplazamiento de los pequeños es más acentuado y diverso. El desplazamiento de los grandes a veces va en concordancia con el desplazamiento de un partido, indicando que, o bien el municipio sigue la hoja de ruta del partido, o bien el partido realiza un buen trabajo de adaptación a los cambios de tendencia de la sociedad. Con respecto a los resultados encontrados en el análisis individual que separaban entre grupos independentistas y constitucionalistas, se puede ver en el nuevo mapa una cierta cercanía entre los partidos que pertenecen al mismo grupo y una clara separación entre grupos. Otra cercanía con la realidad es el acercamiento de Barcelona hacia las coordenadas del partido CatComú, el cual también en esa fecha, era el responsable de la gobernanza de Barcelona.

Finalmente, se han introducido algunas ideas que podrían ayudar a mejorar y generalizar el proceso, el cual puede verse limitado en ciertos aspectos o dejar muchos criterios libres en algunas situaciones. Los nuevos conceptos introducidos se basan en la aplicación de una función sobre los distintos conjuntos de datos con distintos números de columnas que situaría directamente a todos los conjuntos en un mismo espacio, sin tener que ser este el origen.

Bibliografía

[1] Lê, Sébastien and Josse, Julie and Husson, François and others, *FactoMineR: an R package for multivariate analysis*, journal: Journal of statistical software, vol.1, número 1, págs. 1-18, 2018.

[2] Murtagh, Fionn and Legendre, Pierre *Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm,* journal: arXiv preprint arXiv:1111.6285, 2011.

[3] Cristina Gil Martínez *ANÁLISIS DE COMPONENTES PRINCIPALES (PCA),* `https://rpubs.com/Cristina_Gil/PCa`, 2018.

[4] Kenneth Roy Cabrera Torres, *Distancias y similitudes,* `https://labscn-unalmed.github.io/ecologia-numerica/guiones/distancias_disimilitudes_matriz_discrepancia.html#la-distancia-de-hellinger-d_17`, 2019.

[5] IDESCAT, *Eleccions al Parlament de Catalunya. Dades generals,* `http://www.idescat.cat/pub/?id=elepc&n=389&by=mun&t=201500&lang=es`.

[6] Gobierno de España, *Censo Electoral de los Residentes Ausentes - CERA,* `http://www.exteriores.gob.es/Consulados/HaMBURGO/es/ServiciosConsulares/EnHamburgo/electorales/Paginas/CERA.aspx`.

[7] María Menéndez, *Qué ha pasado en Cataluña desde la consulta del 9-N,* `http://www.rtve.es/noticias/20171004/pasado-cataluna-desde-consulta-del-9-hasta-del-1/1583644.shtml`, 2017.

[8] IDESCAT, *Altitud, superficie y población. Municipios* `https://www.idescat.cat/pub/?id=aec&n=925&lang=es`, journal: Idescat. Anuario estadístico de Cataluña. Altitud, superficie y población. Municipios.

[9] Universidad La Laguna, *Clustering jerárquico.*

[10] Josep M. Oller Sala, material vario sobre álgebra lineal e ideas base del trabajo.

[11] Oller, J.M., Satorra, A. Tobeña, A. *Unveiling pathways for the fissure among secessionists and unionists in Catalonia: identities, family language, and media influence.* Palgrave Commun 5, 148 (2019). https://doi.org/10.1057/s41599-019-0357-z