Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

ANÁLISIS DE RESULTADOS ELECTORALES EN ESPAÑA DE ELECCIONES EUROPEAS

Autor:
Hugo Mata Merino
Tutores:
M. Pilar Rodríguez del Tío
Miguel Alejandro Fernández Temprano

Resumen

En este Trabajo de Fin de Grado se analizan los resultados de las elecciones al Parlamento Europeo para las nueve convocatorias celebradas en España entre el año 1987 y 2024, teniendo como objetivos principales la identificación de patrones ideológicos a nivel provincial y el estudio de la abstención. Para ello, se emplean técnicas como el Análisis de Correspondencias Simples, Análisis de Correspondencias Múltiples, la Regresión Lineal Simple y Múltiple y el ANOVA, que permiten observar la evolución del comportamiento electoral en España a lo largo de este periodo analizado.

Palabras clave

Elecciones al Parlamento Europeo, ideología política, abstención, análisis de correspondencias, polarización, fragmentación, comportamiento electoral.

Abstract

In this work, the results of the European Parliament elections for the nine elections held in Spain between 1987 and 2024 are analyzed, with the main objectives being the identification of ideological patterns at the provincial level and the study of the abstention. To achieve this, techniques such as Simple Correspondence Analysis, Multiple Correspondence Analysis, Multiple and Simple Lineal Regression, and ANOVA are employed, which allow for the observation of the evolution of electoral behavior in Spain throughout the analyzed period.

Keywords

European Parliament elections, political ideology, abstention, correspondence analysis, polarization, fragmentation, electoral behavior.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Durante los últimos años, el panorama político español ha experimentado numerosas transformaciones, marcadas por una creciente polarización y el surgimiento de nuevos partidos políticos de ideología extrema. El comportamiento electoral ha sufrido un proceso evolutivo, partiendo de una situación estable, con la predominancia del eje izquierda-derecha, hasta momentos de alta crispación política y del auge de partidos emergentes con discursos más extremos. Esta crispación ha dado lugar a un claro desapego de la población española a la política en general, lo cual podría traducirse en un mayor número de personas que no participan en las elecciones.

En este trabajo, se van a analizar y a estudiar en profundidad los resultados de las elecciones al Parlamento Europeo a nivel municipal y provincial para las nueve convocatorias celebradas del año 1987 hasta el 2024. Estas elecciones presentan varias peculiaridades que las hacen diferentes a las elecciones generales, municipales o autonómicas. En primer lugar, la circunscripción nacional es única, es decir, todos los votos tienen el mismo peso, lo cual provoca que partidos regionales y provinciales (como los nacionalistas) tengan que agruparse en coaliciones [1]. Además, la participación es significativamente inferior al resto de elecciones [2], por ejemplo, en la convocatoria de 2024 fue del 49,2 %, frente al 66,59 % de participación que hubo en las elecciones generales de 2023 en España [3]. Este elevado nivel de abstención puede explicarse por la percepción de lejanía de las instituciones europeas, las cuales a menudo se denominan elecciones de 'segundo orden'. El voto en estas elecciones suele ser más ideológico, de tal manera que la población no suele votar a las mismas opciones que en su país, tratando de dar un mensaje de disconformidad [4].

Para abordar el análisis, se van a aplicar técnicas como el Análisis de Correspondencias Simples, Análisis de Correspondencias Múltiples, la Regresión Múltiple o el Análisis de la Varianza. Esto permitirá estudiar tanto el fenómeno de la abstención, como la evolución ideológica del voto a lo largo de las convocatorias.

Este trabajo se organiza en varios capítulos, que abordan la descripción de la metodología

usada, la obtención y el tratamiento realizado a los datos descargados de la página oficial del Ministerio del Interior, el posterior análisis de estos datos y las conclusiones.

1.1. Asignaturas relacionadas

Para la realización de este Trabajo Fin de Grado, las siguientes asignaturas han sido especialmente relevantes, ya que de ellas se ha extraído la mayoría de información y conocimientos necesarios tanto para la aplicación de las técnicas empleadas como para el uso del software empleado:

- **Computación estadística:** en esta asignatura se explican los conocimientos fundamentales y básicos sobre el uso del software estadístico `R`, que ha sido el software usado a lo largo de todo el trabajo para el tratamiento de datos y la implementación de los análisis.
- **Análisis de datos:** esta asignatura proporciona las bases de diversas técnicas estadísticas empleadas en el análisis de datos, en concreto el Análisis de Correspondencias Simples y el Análisis de Correspondencias Múltiples, los cuales han sido aplicados en este estudio.
- **Regresión y Anova:** proporciona los fundamentos teóricos de los modelos de Regresión Lineal Simple y Múltiple, además de los procedimientos para la validez del modelo. También se enseñan conceptos clave del Análisis de la Varianza, como los distintos tipos, la validez del modelo, además del análisis post-hoc.

Capítulo 2

Metodología

En este apartado se explicarán las técnicas empleadas para poder llevar a cabo el análisis de las elecciones al Parlamento Europeo a nivel de municipios y provincias en España.

2.1. Análisis de Correspondencias

El Análisis de Correspondencias es una técnica estadística usada para explorar y visualizar las relaciones existentes entre dos o más variables categóricas [5].

El objetivo principal es reducir la dimensión del conjunto de datos, de manera que se obtiene una representación gráfica (generalmente en dos dimensiones) fácilmente interpretable [6]. Es una extensión del Análisis en Componentes Principales (PCA). Existen dos tipos de Análisis de Correspondencias de manera general:

- **Análisis de Correspondencias Simples (ACS):** cuando se estudia la asociación entre dos variables categóricas.
- **Análisis de Correspondencias Múltiple (ACM):** cuando se estudia la asociación entre más de dos variables categóricas.

En este caso se van a aplicar ambos tipos de Análisis de Correspondencias.

2.1.1. Elementos básicos

En esta sección se definen los conceptos básicos para comprender el Análisis de Correspondencias.

Tabla de contingencia:

Para realizar el Análisis de Correspondencias se parte de una tabla con dos variables X e Y (en el caso del Análisis de Correspondencias Simples), de r filas y c columnas.

$X \backslash Y$	Cl. 1	$\cdots$	Cl. c	Total$_i$
Cl. 1	n_{11}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$	$\vdots$
Cl. r	n_{r1}	$\cdots$	n_{rc}	n_{r+}
Total$_j$	n_{+1}	$\cdots$	n_{+c}	n_{++}

Tabla 2.1: *Ejemplo tabla de contingencia*

- n_{ij}**:** número de individuos pertenecientes a la categoría i (variable X) y categoría j (variable Y).
- n_{1+}**:** número total de individuos pertenecientes a la categoría 1 de la variable X independientemente de la categoría de Y.
- n_{+1}**:** número total de individuos pertenecientes a la categoría 1 de la variable Y independientemente de la categoría de X.
- n_{++}**:** número total de individuos.

Perfiles fila y columna:

Son las distribuciones de frecuencias condicionadas por filas y columnas. Se expresan como:

$$\text{Perfiles fila} = \left(\frac{n_{11}}{n_{1+}}, \quad \cdots \quad , \frac{n_{1c}}{n_{1+}}\right)$$

$$\text{Perfiles columna} = \left(\frac{n_{11}}{n_{+1}}, \quad \frac{n_{21}}{n_{+1}}, \quad \cdots, \quad \frac{n_{r1}}{n_{+1}}\right)^{\top}$$

No todos los perfiles tienen el mismo peso. A diferencia del Análisis de Componentes Principales, estos perfiles suelen generarse a partir de cantidades distintas de individuos, por lo que se asignan pesos en función del número de individuos que los generan.

Distancia χ^2:

Se utiliza para medir las diferencias entre perfiles. Esta distancia posee la propiedad de equivalencia distribucional: si dos perfiles fila son iguales y se agregan, las distancias entre los perfiles columna deben mantenerse inalteradas.

Inercia total:

Se calcula como la suma del cuadrado de la distancia de cada perfil al centro de gravedad, ponderada por el peso correspondiente. Evalúa la calidad de representación de los ejes en el Análisis de Correspondencias.

2.1.2. Análisis de Correspondencias Simples

Antes de comenzar se aplica el test χ^2 de independencia con el objetivo de comprobar si son o no independientes las variables a estudiar, el interés de realizar este test radica en que si las variables son independientes, no tiene sentido aplicar esta técnica, en el caso contrario de que se rechace la hipótesis de independencia si tendría sentido aplicar el Análisis de Correspondencias ya que indica que existe asociación entre las variables.

Se parte de una tabla de contingencia con el objetivo de representar gráficamente las relaciones entre las variables disminuyendo la dimensión. Lo primero de todo es calcular la distancia χ^2 entre los perfiles fila y columna. Se obtiene una matriz de desviaciones respecto al modelo de independencia, sobre esta matriz se aplica la descomposición en valores singulares (SVD), construyéndose así un nuevo sistema de dimensiones ortogonales (ejes independientes entre sí), que explican la mayor parte de la inercia total. El número máximo de dimensiones que se pueden extraer viene dado por $\min(r-1,c-1)$ donde r y c son el número de filas y de columnas de la tabla de contingencia inicial.

El número de dimensiones extraídas depende de la inercia explicada por cada eje, de forma general se extraen dos dimensiones, ya que facilita la interpretación de los gráficos. Tras extraer las dimensiones principales, se representan sobre ellas los perfiles fila y columna.

2.1.3. Análisis de Correspondencias Múltiples

Es una extensión del Análisis de Correspondencias Simples, que permite estudiar la asociación entre tres o más variables categóricas. Permite representar gráficamente las relaciones entre las distintas categorías de las variables. Al igual que en el ACS, el objetivo es reducir la dimensión del conjunto de datos, lo cual facilita la interpretación.

Se parte de una matriz de indicadores, que es una matriz binaria donde cada fila representa un individuo y cada columna una categoría de las variables. Al ser una matriz binaria, un individuo solo puede tomar el valor de 1 en una categoría de cada variable, en el resto de categorías toma el valor de 0. Sobre dicha matriz se construye una matriz de desviaciones respecto al modelo de independencia, sobre la que se aplica la descomposición en valores singulares (SVD). Al igual que en el ACS, esto permite la construcción de un nuevo sistema de dimensiones ortogonales que explican la mayor parte de la inercia total.

Una alternativa a la matriz de indicadores, es usar la matriz de Burt, la cual contiene los posibles cruces entre las variables categóricas. El procedimiento a aplicar es el mismo que se ha descrito anteriormente, con la diferencia de que la inercia total es mayor.

Respecto al número máximo de dimensiones extraídas ocurre lo mismo que en el caso de dos variables, y es que de forma general se extraen dos dimensiones, ya que esto permite

una representación clara de las relaciones entre las categorías. Una vez extraídas las dimensiones, se representan sobre ellas los perfiles fila y columna, permitiendo identificar posibles asociaciones.

2.1.4. Gráficos del Análisis de Correspondencias

Los gráficos que se van a usar son:

- **Scree plot:** se representa el % de variabilidad explicada por cada dimensión. Se usa para ver el número de dimensiones que se extraen.
- **Biplot:** se representa de forma simultánea las proyecciones de las filas y las columnas. Se usa para ver como se relacionan las distintas categorías de las variables.

2.2. Regresión

La regresión se utiliza para describir relaciones estadísticas entre variables, en concreto el análisis de regresión es una técnica estadística usada para modelar y explorar la relación de una variable dependiente con una o más variables independientes.

2.2.1. Descripción del modelo de regresión

El modelo de Regresión Lineal Múltiple [7] se puede escribir de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

El principal objetivo es la estimación de los coeficientes β_i. Los elementos del modelo son los siguientes:

- Y_i**:** variable respuesta o dependiente. Es la variable a estudiar.
- X_i**:** variables explicativas o independientes. Estas variables aportan información sobre la variabilidad de la variable respuesta y es interesante el efecto que producen sobre ella.
- β_i**:** coeficientes de regresión.
- ε_i**:** componente aleatoria. Se asume que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d..

Se puede escribir de forma matricial el modelo [8]:

$$Y = X * \boldsymbol{\beta} + \varepsilon$$

Donde:

- Y**:** vector con las observaciones de la variable respuesta.
- X**:** matriz de diseño en la que cada fila representa una observación y cada columna una variable explicativa.
- $\boldsymbol{\beta}$**:** vector de parámetros.
- ε**:** vector de errores aleatorios.

2.2.2. Estimación por mínimos cuadrados

Partiendo del modelo escrito en forma matricial, para estimar los coeficientes de la regresión (vector de parámetros $\boldsymbol{\beta}$) se minimiza la suma de los cuadrados de los residuos, de la siguiente manera, hasta obtener las ecuaciones normales:

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) \Rightarrow X^T X\hat{\beta} = X^T Y \Rightarrow \hat{\beta} = (X^T X)^{-1}X^T Y$$

La distribución de los estimadores es:

$$\hat{\beta} \sim N\left(\beta, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]\right)$$

donde σ^2 es la varianza del error.

2.2.3. Residuos y residuos estudentizados

Residuos: son la diferencia entre el valor observado y el valor predicho por el modelo.

$$e_i = Y_i - \hat{Y}_i$$

Se usan para comprobar la validez del modelo a partir de las asunciones que se explicarán posteriormente.

Residuos estudentizados: se obtienen:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad \text{donde} \quad \hat{\sigma}^2 = \frac{SSE}{n - p}$$

Cabe destacar que h_{ii} es el elemento i-ésimo de la diagonal de la matriz: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, y que p es el número total de parámetros.

Estos residuos verifican:

$$E(r_i) = 0 \quad \text{y} \quad Var(r_i) = 1$$

Asumiendo la normalidad de los errores, y bajo la hipótesis de que la i-ésima observación sigue el modelo ajustado:

$$r_i \sim t_{n-p}$$

Estos residuos permiten identificar los valores atípicos de forma más precisa.

2.2.4. Asunciones del modelo de regresión

Para poder usar un modelo de regresión se deben validar una serie de hipótesis [9], en concreto son:

Homogeneidad de varianza: esta hipótesis asume que la varianza de los errores es constante, es decir los residuos tienen una dispersión homogénea. Para comprobarlo se realiza el llamado plot de residuales. De forma complementaria se pueden usar varios contrastes de hipótesis como el de Brown-Forsythe o el de Breusch-Pagan.

Normalidad: se asume que los residuos siguen una distribución normal de media cero y varianza constante. Para comprobarlo se representan los residuos en un gráfico cuantil-cuantil. Para muestras grandes no es crucial esta hipótesis.

Linealidad: se requiere que la media de la variable respuesta esté relacionada de forma lineal con las variables explicativas. Se puede comprobar fácilmente viendo que la media de los residuos es cercana a 0.

Independencia: se requiere que los residuos sean independientes entre sí. Para comprobarlo se puede usar el test de Durbin-Watson.

2.3. Análisis de la Varianza

El ANOVA es una técnica estadística que permite comparar las medias de tres o más grupos de un factor, evaluando si existen o no diferencias significativas entre ellas [10].
En un modelo ANOVA de un factor con k niveles:

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \delta_i + \varepsilon_{ij}, \quad i = 1, \ldots, k \text{ y } j = 1, \ldots, n_i$$

con $\varepsilon_{ij} \sim N(0,\sigma^2)$, se quiere saber si todos los niveles del factor tienen o no el mismo efecto sobre la variable respuesta.
Para ello se realiza el siguiente contraste de hipótesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \quad \text{vs} \quad H_1 : \mu_i \neq \mu \text{ para algún } i = 1, \ldots, k$$

Para realizar este contraste se usa un estadístico F basado en la descomposición de la variabilidad total de los datos.

$$\begin{aligned} \text{Variabilidad total (SST)} = &\text{ Variabilidad dentro de cada nivel (SSE)} \\ &+ \text{Variabilidad entre niveles (SSR)} \end{aligned}$$

2.3.1. Tabla ANOVA

En la tabla ANOVA se resume la descomposición de la variabilidad total de los datos en las diferentes fuentes (entre-grupos, intra-grupos y total) y permite calcular el estadístico F para contrastar la hipótesis explicada anteriormente.

Fuente	**Suma de cuadrados (SS)**	**Grados de libertad (gl o df)**	**Media cuadrática (MS)**	F_{obs}
Entre-grupos	SSR	$k-1$	MSR	$\frac{MSR}{MSE}$
Intra-grupos	SSE	$N-k$	SSE	
Total	SST	$N-1$		

Tabla 2.2: *Ejemplo de tabla ANOVA*

2.3.2. Tipos de ANOVA

- **ANOVA unifactorial:** estudia el efecto de un único factor sobre la variable respuesta Y. Es el que se va a usar en este trabajo.
- **ANOVA multifactorial:** estudia el efecto de dos o más factores y sus interacciones sobre la variable respuesta Y.
- **ANOVA de efectos fijos:** los niveles del factor están fijados de forma previa, en este tipo de modelos es interesante observar las diferencias en la respuesta entre los niveles específicos.
- **ANOVA de efectos aleatorios:** los niveles del factor son una muestra representativa de una población de posibles niveles. Es interesante concluir si tiene un efecto significativo el factor en la variable respuesta en general.
- **ANOVA de efectos mixtos:** es un modelo donde se mezclan factores fijos y aleatorios.

2.3.3. Hipótesis del ANOVA

Para determinar que un modelo ANOVA es válido es necesario comprobar que se cumplan las siguientes hipótesis:

Normalidad de los residuos: los residuos tienen que seguir una distribución normal. Para comprobar este supuesto se realiza un histograma (debe tener la apariencia de una distribución normal centrada en cero) o un Q-Q plot (los puntos deben estar próximos a la recta $y = x$). También se puede comprobar de manera numérica usando el test de Shapiro-Wilk. Cuando el tamaño muestral es suficiente, por el teorema central del límite,

la falta de normalidad tendrá poco efecto en el contraste de igualdad de medias y las estimaciones.

Homocedasticidad de la varianza: la varianza de los residuos debe ser constante y no variar en los distintos niveles del factor. Para evaluar esta hipótesis se realiza un gráfico de los residuos frente a los valores predichos, donde no se debe observar ningún patrón ni ningún aumento o disminución de la varianza a lo largo de todo el eje x. Este gráfico es usado para la detección de valores atípicos también. De forma numérica se puede comprobar con el test de Levene.

Independencia de los residuos: las observaciones de cada grupo del factor deben ser independientes entre sí. Para comprobar este supuesto se realiza el llamado plot secuencial, en el que se representan los residuos frente al orden en el que se recopilaron los datos. En este gráfico si se observan rachas o cualquier tendencia creciente o decreciente, se rechaza la hipótesis.

2.3.4. Análisis post-hoc

Una vez ajustado y validados los supuestos del modelo ANOVA, si se rechaza la hipótesis de igualdad entre todas las medias, se puede realizar un análisis post-hoc con el objetivo de comprobar los grupos que son significativamente distintos [11]. Se pueden usar distintos test:

- **Test de Tukey:** es útil cuando se quieren realizar todos los pares de comparaciones de medias $\mu_i - \mu_j$. Para cada par de medias, se estima el error estándar $E_{ii'}$ y se calcula el umbral $HSD_{ii'}$.

$$E_{ii'} = \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}, \quad HSD_{ii'} = \frac{q_{\alpha;k,N-k}}{\sqrt{2}} \cdot E_{ii'}$$

  Es un test conservador esto quiere decir que para que existan diferencias significativas estas diferencias deben ser grandes.

- **Test de Duncan:** es útil cuando el diseño es balanceado y se quiere evaluar las diferencias de todos lo niveles respecto de uno. Se ordenan las medias de todos los niveles de menor a mayor, se denota con p al número de grupos cuyas medias están entre las dos que queremos comparar y se calcula el umbral R_p.

$$R_p = q_{\alpha_p,p,N-k} \cdot \sqrt{\frac{MSE}{n}}$$

En ambos casos la hipótesis de igualdad de media se rechaza (a nivel de significancia α), si la diferencia entre las medias a comparar es mayor que el umbral.

Capítulo 3

Obtención y tratamiento de datos

En este apartado se va a explicar como se han obtenido los distintos conjuntos de datos y el tratamiento llevado a cabo hasta lograr el conjunto final.

Se consideran las 9 convocatorias a las elecciones al Parlamento Europeo que se han celebrado en España. Estas son las correspondientes a los años: 1987, 1989, 1994, 1999, 2004, 2009, 2014, 2019 y 2024.

3.1. Descarga de los datos

La descarga de los datos se realizó desde la página web del Ministerio del Interior, para ello se accedió al apartado de Elecciones Celebradas y por último al Área de descargas [12].

En el área de descargas se selecciona primero la convocatoria, que indica el tipo de elección, aparecen las elecciones al Parlamento Europeo, al congreso, al senado... En este caso se seleccionan las elecciones al Parlamento Europeo.

Posteriormente se selecciona la fecha, que indica las distintas convocatorias que ha habido, una vez seleccionada aparece un botón de consultar, que presionándole te da la opción de descargar información relativa a tres ámbitos distintos, aparece información sobre la mesa, el municipio y superior al municipio, en este caso se descarga la información relativa al municipio ya que no solo contiene información sobre los municipios, si no que contiene información sobre las provincias lo cuál será útil para los posteriores análisis.

Una vez descargados los datos se obtiene una carpeta comprimida que cuenta con 10 documentos, de estos documentos se pueden diferenciar dos tipos:

- **Archivos .dat:** de la carpeta comprimida se obtienen 8 ficheros con esta extensión. Estos ficheros contienen diversos datos, posteriormente se indicaran aquellos que se usan.

- **Archivos .doc:** de la carpeta comprimida se obtienen 2 ficheros con esta extensión. No son ficheros distintos, si no que son el mismo pero de manera duplicada. Son documentos explicativos donde se describe el contenido de cada fichero .dat de manera detallada.

3.2. Descripción de los conjuntos de datos

De los 8 archivos de datos que se tienen son interesantes los 3 siguientes, los cuales son tratados como los ficheros base a partir de los cuales se construye el conjunto de datos final.

- **Fichero 03: candidaturas.** Contiene información sobre las candidaturas presentadas en cada una de las convocatorias. En concreto contiene, tanto las siglas, como la denominación y el código asignado de las candidaturas.
- **Fichero 05: datos comunes de municipios.** Incluye datos globales para cada municipio. Contiene información sobre el comportamiento electoral en cada municipio. De este conjunto serán de utilidad variables como el código de cada provincia y de cada municipio, los votos (tanto nulos como en blanco) y la información relativa al censo.
- **Fichero 06: datos de candidaturas de municipios.** Incluye los votos obtenidos por cada candidatura que aparece en el fichero de candidaturas en cada municipio. De este conjunto serán de utilidad campos como el código de cada provincia y de cada municipio, junto con el código de la candidatura (estos 3 campos permitirán relacionarlo con los otros dos conjuntos) y los votos obtenidos.

Cabe destacar que el fichero 05 y el fichero 06 contienen una variable que indica el número del distrito, esto será explicado en la sección siguiente de tratamiento de datos.
Una vez explicado lo anterior, para cada convocatoria de elecciones al Parlamento Europeo, se tienen 3 conjuntos de datos.

3.3. Tratamiento de datos

En este apartado se va a detallar el procedimiento hasta lograr el conjunto de datos final, el objetivo es conseguir una tabla final donde aparezcan todas las elecciones con las distintas ideologías y los votos obtenidos. Esto se ha realizado a través del software estadístico R.

3.3.1. Tratamiento de cada convocatoria

Se comienza uniendo el fichero 03 (candidaturas) con el fichero 06 (datos de candidaturas por municipio). Esto se realiza a través de la función `merge` en R [13], la cuál permite combinar ambos conjuntos de datos a partir de columnas comunes. En este caso, se usa el código de la candidatura como clave de unión.

Una vez realizada la combinación, se obtiene un único conjunto de datos que contiene, para cada candidatura presentada, tanto información descriptiva como el número de votos obtenidos en cada municipio.

En este punto se dispone de dos conjuntos de datos: el conjunto resultante explicado y el fichero 05 (datos comunes de municipios). Tal como se mencionó anteriormente, ambos conjuntos contienen una columna que indica el número del distrito. Cada municipio está desglosado en varios distritos, pero dado que el objetivo de este trabajo es analizar los datos a nivel municipal, se filtran las observaciones de tal manera que se seleccionan aquellas que el valor del campo 'Distrito' sea 99. Según mencionan los documentos descriptivos descargados junto con los datos, este valor indica que la información se refiere al total de votos del municipio y no a un distrito específico.

Posteriormente, se realiza una unión de tipo `left join` [14] entre ambos conjuntos de datos, usando como claves de unión el código de la provincia y el código del municipio.

Se eliminan algunas columnas que no aportan nada en el análisis como son el mes en el que se realizó dicha elección, el número de vuelta, el número del distrito, así como el código de la candidatura a nivel nacional, autonómico y provincial.

Se realiza este procedimiento para cada una de las convocatorias, de tal manera que se obtienen 9 conjuntos de datos, con la información organizada por municipios.

3.3.2. Clasificación de candidaturas

Para cada convocatoria ha sido necesario realizar una clasificación de las candidaturas de forma detallada.

En primer lugar se identificaron las distintas coaliciones que han participado en las elecciones europeas a lo largo de todas las convocatorias. No se consideraron aquellas coaliciones que no tuvieron una continuidad en el tiempo o que se presentaron en una única elección, si no que se priorizó aquellas que han mantenido una continuidad o persistencia histórica.

Se observo que, en algunas convocatorias, un mismo partido político se presentaba mediante distintas candidaturas. Esto ocurre, por ejemplo, con el PSOE y el PSC (Partido de

los Socialistas de Cataluña), o el caso de Podemos que en una misma convocatoria existen varias candidaturas. Aunque a priori aparecen en el conjunto de datos como candidaturas diferentes (distinto código), como representan al mismo partido político, se agruparon bajo una única candidatura.

Como uno de los objetivos de este trabajo es poder observar como han ido evolucionando las diversas ideologías políticas a lo largo de los años en las elecciones al Parlamento Europeo, se agruparon las candidaturas finales en las principales corrientes ideológicas.

En un principio se optó por la siguiente división en ideologías: Izquierda, Derecha, Centro, Centro derecha, Centro Izquierda, nacionalistas de izquierda y nacionalistas de derecha. Esta primera clasificación suponía algunos problemas:

- **Nacionalistas:** la mayoría de candidaturas nacionalistas provienen de coaliciones, como pueden ser Pueblos de Europa, Por la Europa de los Pueblos, Coalición por una Europa solidaria... Estas coaliciones si bien en algunas convocatorias están bien diferenciadas, en otras son bastante ambiguas, como es el caso de la candidatura Galeusca-Pueblos de Europa en las elecciones del año 2004, ya que está integrada por partidos de ideologías nacionalistas de izquierdas y de derechas. Esto hizo muy difícil la clasificación, por tanto se optó finalmente por agrupar las ideologías nacionalistas en un solo grupo.

- **Centro:** al igual que con el caso de los nacionalistas, la clasificación de una candidatura en centro derecha o centro izquierda se hacía compleja, por tanto también se optó en agrupar las ideologías de centro en un solo grupo.

- **Extrema derecha:** se observa que desde las elecciones de 1989 hasta 2014, no aparece esta ideología, pero la decisión es mantenerla, para poder ver el claro auge de las ideologías extremas en los últimos años.

- **Extrema izquierda:** durante las elecciones de 1987 hasta 2009 (incluida), no tenía en principio mucha importancia, pero se observó que durante estos años era bastante relevante la ideología ecologista, contando con candidaturas como Los Verdes, los cuales aparecen formando parte de candidaturas de extrema izquierda en convocatorias posteriores [15], por tanto la decisión tomada fue clasificar la ideología ecologista como extrema izquierda. Por tanto en los posteriores análisis cuando se mencione esta ideología, cabe destacar que desde las elecciones de 1987 hasta las de 2009 se está refiriendo a los ecologistas.

Por tanto, una vez tomadas todas estas decisiones, se tiene una división en bloques ideológicos mucho más simple contando con: Izquierda, Derecha, Centro, Extrema Izquierda, Extrema Derecha y Nacionalistas.

Otro gran problema a la hora de realizar la clasificación es el elevado número de candidaturas presentadas, ya que en algunas convocatorias había más de 50. Se optó por tomar un umbral mínimo del 0,5 % de los votos nacionales para realizar la clasificación ideológica, de tal manera que aquellas candidaturas que en la convocatoria superaron dicho umbral fueron clasificadas según su ideología correspondiente, mientras que aquellas candidaturas que no superaron el umbral se agruparon en una categoría genérica denominada 'OTROS'. En esta categoría se encuentran partidos minoritarios de todas las ideologías posibles.

Cabe destacar que en las elecciones de 2004 no se tomó el umbral de 0,5 % ya que usándolo se obtenían muy pocas candidaturas, por tanto fue necesario reducirlo a un 0,434 %, siendo más transigente para permitir la presencia de más candidaturas.

Las coaliciones consideradas y la clasificación ideológica de las candidaturas se encuentran en el Apéndice A y en el Apéndice B respectivamente.

3.3.3. Conjunto de datos final

Se parte de los 9 conjuntos de datos descritos anteriormente, donde cada fila representa una candidatura concreta en un municipio, y se tiene información relevante como los votos y el censo.

Una vez tratados los datos por convocatoria, se unen los datos de cada convocatoria para construir un único conjunto de datos donde esté la información de cada elección. Para poder realizar esta unión, se usó la función `rbindlist` de R [16], concatenando las tablas de cada elección en una misma tabla.

Se eliminaron variables que no se van a usar posteriormente como es el voto CERA (voto de los residentes permanentes en el extranjero) ya que para la mayoría de las convocatorias dicha variable está rellenada con ceros, salvo para las convocatorias de 1999, 2004 y 2009.

Tras esto, se agruparon los datos según cada convocatoria, provincia, municipio y candidatura, obteniendo el número total de votos. Esto se realizó para evitar que existan duplicados.

Posteriormente se creó un nuevo bloque ideológico con el objetivo de agrupar los votos en blanco y los votos nulos en una única candidatura, para poder analizar la abstención. Una vez hecho esto se obtiene la abstención tanto a nivel municipal como a nivel provincial.

Para obtener la abstención a nivel municipal, simplemente se obtiene el censo y el número de votos totales emitidos, teniendo en cuenta también los votos nulos y los votos en blanco, en el municipio, en cada convocatoria y se realiza la siguiente operación:

$$\text{Porcentaje Abstención} = \left(\frac{\text{Censo} - \text{Votos\_totales}}{\text{Censo}}\right)$$

Para obtener la abstención provincial, se realiza el mismo proceso anterior, pero con los datos a nivel de cada provincia.

Se creó una nueva variable en el conjunto la cual agrupa a los municipios en tres categorías en función del tamaño del censo. Estas categorías vienen definidas por el Ministerio de Agricultura, Pesca y Alimentación [17].

- **Pequeños:** municipios con un censo inferior a 5.000.
- **Medianos:** municipios con un censo entre 5.000 y 30.000.
- **Grandes:** municipios con un censo superior a 30.000.

Para finalizar la construcción del conjunto de datos final a partir del cual se van a realizar los posteriores análisis, se comprueba lo siguiente con el fin de tener unos datos sin errores o inconsistencias:

- **Detección de valores ausentes:** se comprueba para todo el conjunto si existen datos ausentes (NA), tras realizar esta comprobación se observó que no existía ningún ausente en el conjunto de datos.
- **Valores negativos:** se comprueba si existe algún valor negativo en las columnas número de votos y censo. No existía ninguna observación con número de votos o censo negativo.
- **Inconsistencias en los votos:** se comprobó si existía alguna observación en la que el número de votos fuese mayor que el censo, y sí, existía una observación. En la convocatoria del año 1989, en la fila correspondiente al municipio de Caracena, provincia de Soria, el número de votos registrados a la candidatura del Partido Popular (11) es superior al censo de dicho municipio (10), por tanto se elimina dicha observación.
- **Inconsistencias en la abstención:** se comprobó si existía algún municipio donde el porcentaje relativo de abstención fuese superior a 1 o inferior a 0. Se encontraron 6 observaciones con un porcentaje de abstención inferior a 0. Esto se debe a que el número de personas censadas en dicho municipio es inferior al número de votos totales registrados, por tanto al hacer la operación descrita anteriormente, se obtiene un resultado negativo. Al tratarse de un número tan pequeño de observaciones, la decisión fue eliminarlos del conjunto.

Tras realizar estas comprobaciones se tiene el conjunto sobre el cual aplicar los posteriores análisis.

Capítulo 4

Análisis de los datos

En este capítulo se exponen los resultados obtenidos a partir de la aplicación de las técnicas estadísticas descritas en la Metodología.

En primer lugar, se estudia en profundidad la abstención por provincias, mediante el ajuste de diversos modelos de Regresión Lineal para analizar el efecto del censo en el porcentaje de abstención, también se observará si este efecto es igual en todas las convocatorias. Para completar el estudio de la abstención se realiza un Análisis de la Varianza con la convocatoria como factor para identificar si existen diferencias significativas en el porcentaje de abstención medio en las nueve convocatorias celebradas. Posteriormente, se exploran las asociaciones entre las ideologías políticas definidas en el tercer capítulo y las convocatorias, además de la relación entre provincias e ideologías mediante el Análisis de Correspondencias Simples, lo que permitirá observar la evolución de la política española en las elecciones al Parlamento Europeo a lo largo de los años. Finalmente, se aplicará un Análisis de Correspondencias Múltiples para estudiar de manera conjunta las asociaciones entre las provincias, las ideologías y las convocatorias.

A lo largo de todo el capítulo, se muestran los resultados obtenidos, tanto numéricos como gráficos, junto con la interpretación extraída.

4.1. Estudio de la abstención

Tal y como se comentó en la introducción, en este apartado se va a estudiar de manera general la abstención en las elecciones al Parlamento Europeo. La variable de interés es la abstención, la cual es un valor entre 0 y 1 que representa la proporción de personas censadas que no han ejercido su derecho a voto.

En primer lugar, se realizan una serie de diagramas de cajas que permiten visualizar y comparar los niveles de abstención por provincias y municipios en las distintas convocatorias. Estos gráficos facilitarán la identificación de patrones, constituyendo un análisis

exploratorio. Posteriormente, se ajustarán varios modelos de Regresión Lineal para estudiar el efecto de distintas variables explicativas, como el tamaño del censo y el año de la convocatoria, sobre la abstención. Y por último se ajustará un ANOVA con la convocatoria como factor observando si hay diferencias significativas en la media de la abstención en las distintas convocatorias.

4.1.1. Análisis exploratorio

Se realiza un diagrama de cajas con el porcentaje relativo de abstención a nivel general en función del tamaño de los municipios. Para esto se ha usado la variable categórica 'Tamaño' ya descrita en el tercer capítulo, la cual agrupa a los municipios en pequeños, medianos y grandes, en función del tamaño de su censo.

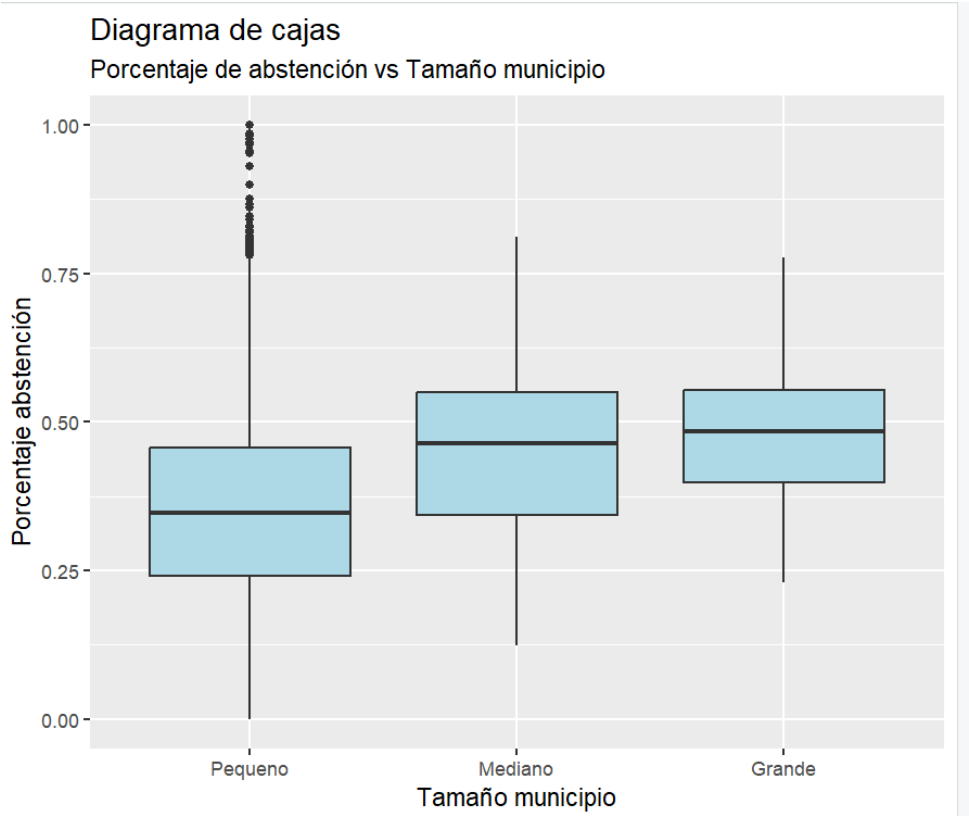
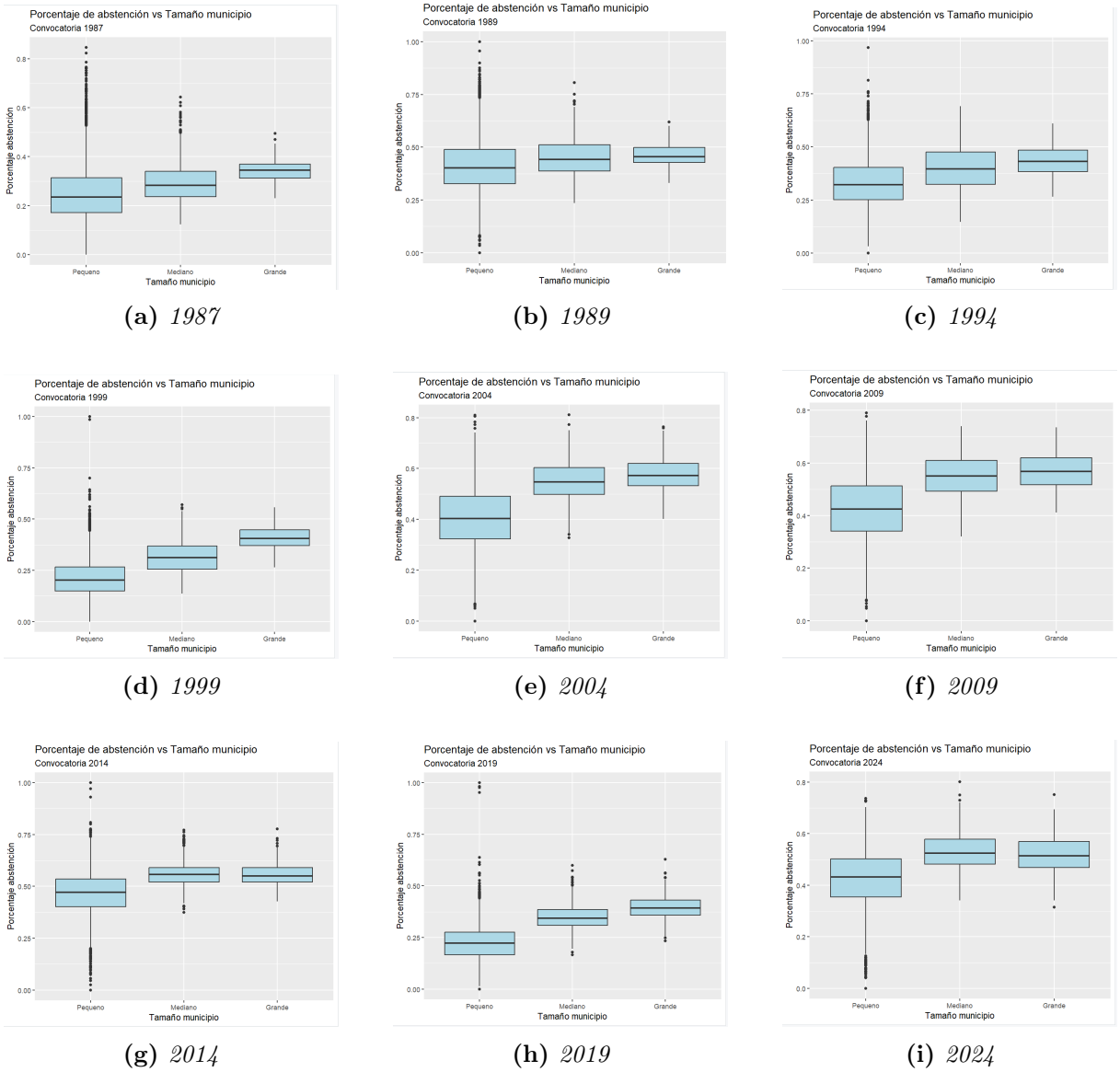


Figura 4.1: *Abstención por tamaño de municipio*

En el gráfico, se observan diferencias claras en el porcentaje de abstención en función del tamaño del censo de los municipios. Los municipios pequeños destacan por tener unos niveles de abstención inferiores a los niveles de los municipios medianos y grandes. También es reseñable, que en estos municipios pequeños se aprecia una mayor dispersión en los valores y una gran cantidad de valores atípicos. Esta variabilidad se explica porque en municipios con censo reducido, es posible que haya casos de participación total y de abstención completa.

Por el contrario, los municipios medianos y grandes muestran unos porcentajes de abstención más homogéneos y superiores (en general) al de los municipios pequeños. El rango intercuartílico es más estrecho, indicando una mayor estabilidad en los niveles de abstención.

Se va a realizar este mismo diagrama, pero para cada una de las 9 convocatorias, con el fin de observar si se sigue el mismo esquema en cada una de ellas.



(a) *1987* **(b)** *1989* **(c)** *1994*

(d) *1999* **(e)** *2004* **(f)** *2009*

(g) *2014* **(h)** *2019* **(i)** *2024*

Figura 4.2: *Porcentaje de abstención por tamaño de municipio en distintas convocatorias*

De manera general, se mantiene el esquema explicado anteriormente, y es que los municipios pequeños muestran niveles menores de abstención, además de una mayor dispersión de los datos y gran cantidad de valores atípicos, respecto de los municipios medianos y grandes. Esto se ve en la mayoría de convocatorias, sugiriendo que el tamaño del censo del municipio tiene un efecto significativo sobre la abstención.

Sin embargo, esto no se observa en las convocatorias de 2014, figura g y 2024, figura i, donde la abstención en municipios medianos es superior a la de los municipios grandes.

Se representan ahora los datos en dos diagramas de cajas múltiples, con el fin de analizar el porcentaje relativo de abstención municipal y provincial en las convocatorias.

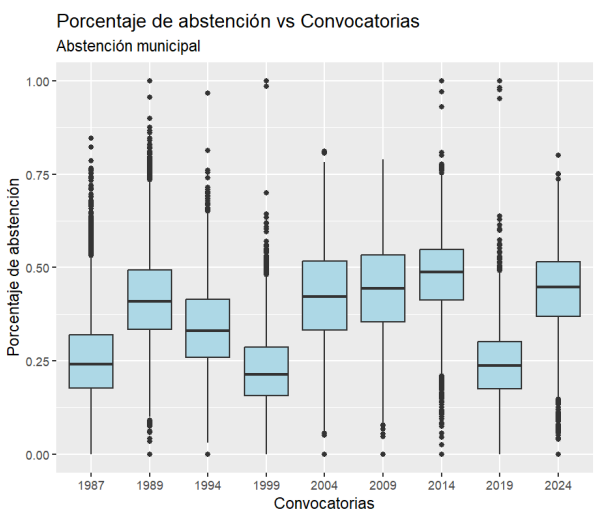


Figura 4.3: *Abstención nivel municipal.*

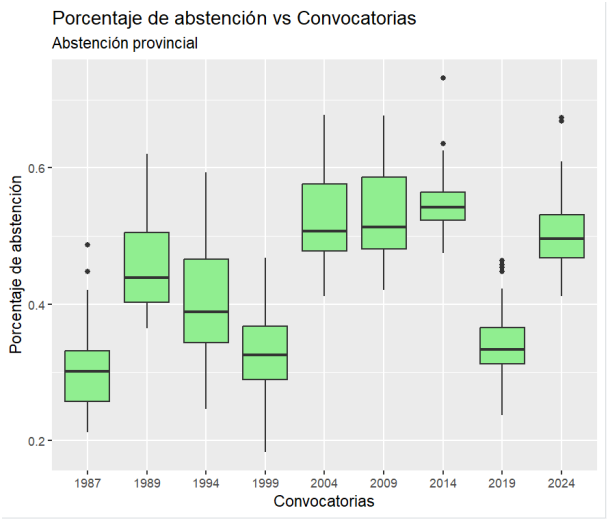


Figura 4.4: *Abstención nivel provincial.*

En la Figura 4.3, en la cual cada observación representa un municipio, se observa que las convocatorias de 1987, 1999 y 2019 presentan porcentajes de abstención menores que el resto de las convocatorias, mientras que las convocatorias de 2009 y 2014 destacan por tener un porcentaje de abstención superior al resto. El diagrama de cajas muestra la existencia de una gran cantidad de valores atípicos en todas las convocatorias, especialmente hacia valores más elevados de la abstención, apreciándose una asimetría en la cola derecha para la mayoría de las convocatorias. En algunas convocatorias también se aprecia una asimetría en la cola izquierda, hacia valores más bajos de la abstención. En general, se observa una gran variabilidad en la abstención municipal, con un rango de valores del 0 % de abstención hasta valores cercanos al 100 %.

En la Figura 4.4, donde cada observación representa una provincia, se aprecia que las convocatorias donde la abstención es más alta son las de 2004, 2009 y 2014. El rango intercuartílico de estas convocatorias es amplio, indicando que la participación fue baja y hay una gran dispersión, es decir, la diferencia en la abstención entre provincias fue mayor. Las convocatorias donde la abstención es más baja son la de 1987, 1999 y 2019, observándose también una menor dispersión, evidenciando un comportamiento más homogéneo entre provincias en esos años.

En ambos diagramas, la convocatoria del año 2019 es una de las que menores niveles de abstención tiene, esto es debido a que coincidió con las elecciones generales en España.

Tanto a nivel municipal como a nivel provincial, la abstención tiene un comportamiento muy similar y presenta esquemas parecidos. La diferencia entre ambos diagramas radica en que a nivel municipal los valores son mucho más dispersos, contando también con numerosos valores atípicos, mientras que a nivel provincial, el número de estos valores atípicos y la dispersión se reduce.

La menor dispersión de los datos observada a nivel provincial facilita la interpretación, por lo cual se ha optado por trabajar con la abstención provincial para los posteriores análisis.

4.1.2. Modelos de regresión

En este apartado se construyen varios modelos de regresión con el fin de analizar el efecto del tamaño del censo de la provincia, y el año de la convocatoria a los niveles de abstención, por tanto, la variable respuesta de los modelos será el porcentaje relativo de abstención a nivel provincial.

Antes de ajustar los modelos de regresión, se va a analizar la distribución de la variable censo a partir de un histograma.

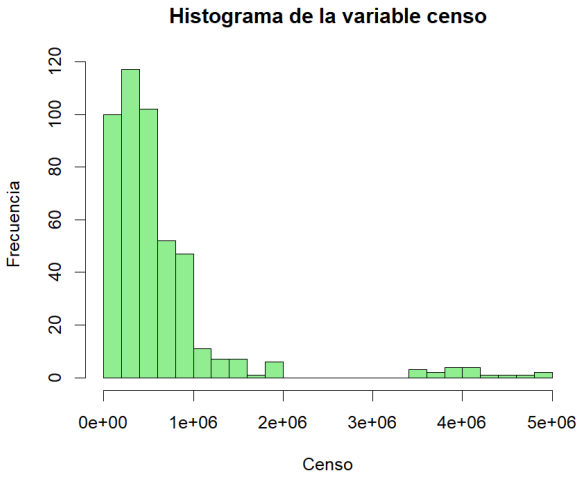


Figura 4.5: *Histograma del censo.*

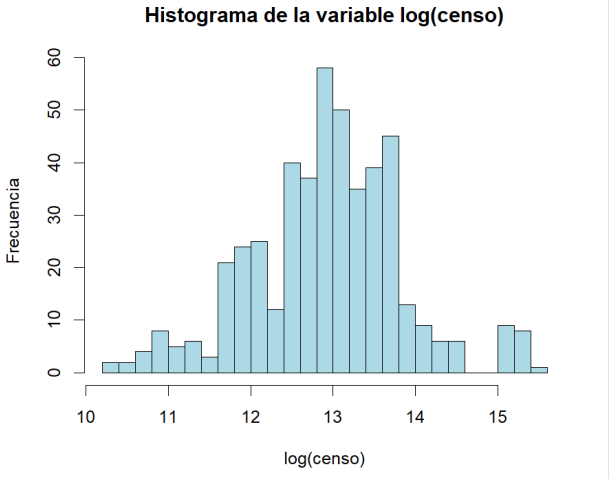


Figura 4.6: *Histograma del logaritmo del censo.*

La Figura 4.5 muestra que la variable censo presenta una fuerte asimetría positiva (hacia la derecha). Por ello, se aplica una transformación logarítmica para corregir dicha asimetría antes de ajustar los modelos de regresión. Una vez aplicada la transformación, tal y como se ve en la Figura 4.6, se corrige la asimetría.

El primer modelo que se construye es con una única variable explicativa, la cual será el logaritmo del censo de la provincia. Este modelo va a permitir evaluar si existe un efecto del tamaño del censo de las provincias en los niveles de abstención, sirviendo de base para los posteriores modelos.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.336924   0.071810   4.692 3.56e-06 ***
log(Censo)  0.007896   0.005545   1.424    0.155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 466 degrees of freedom
Multiple R-squared:  0.004332,  Adjusted R-squared:  0.002195
F-statistic: 2.027 on 1 and 466 DF,  p-value: 0.1551
```

Figura 4.7: *Resumen del primer modelo*

Tras ajustar el primer modelo se observa que el logaritmo del censo no es significativo, ya que a nivel 5 %, el *p-valor* obtenido es 0,155 , por tanto no se tienen evidencias suficientes para rechazar la hipótesis de que la pendiente sea 0, es decir, la relación de la abstención con el logaritmo del censo no parece significativa.

Se representa el modelo de forma gráfica:

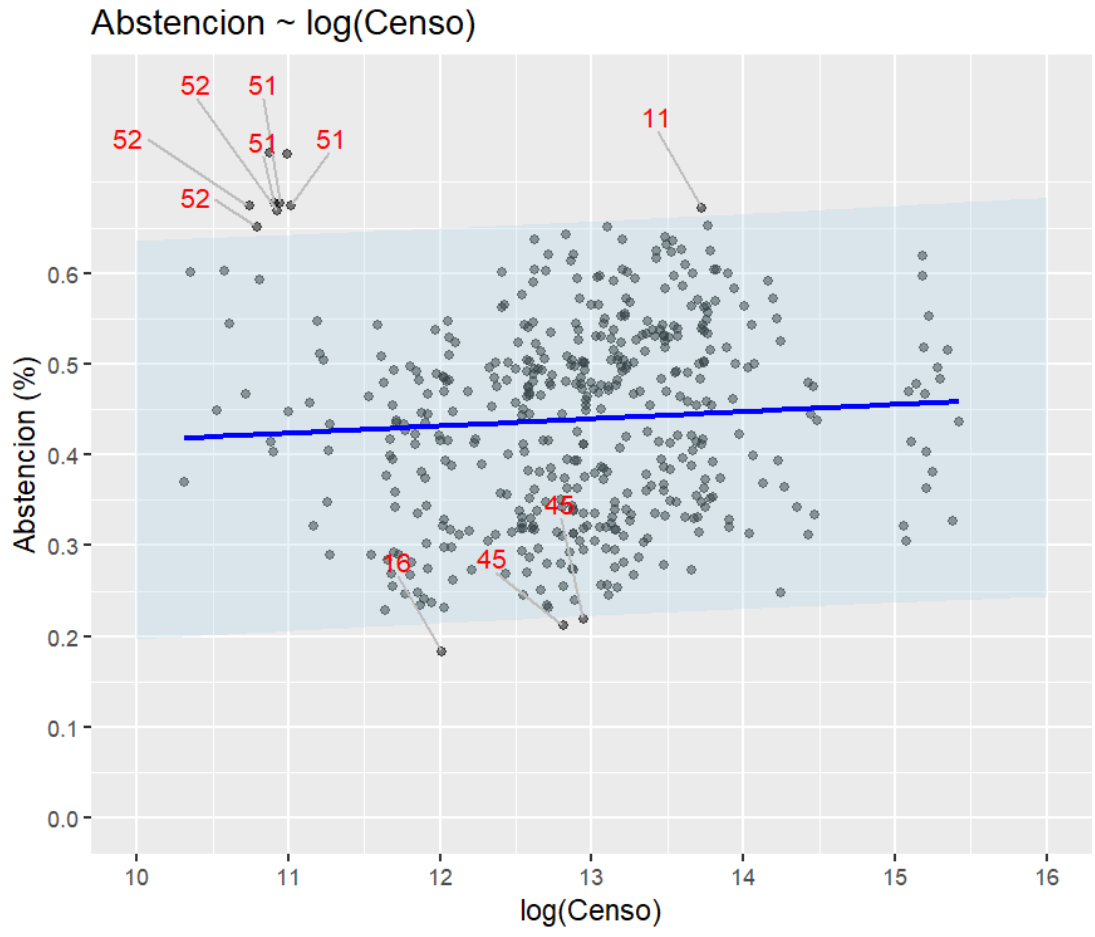


Figura 4.8: *Gráfico del primer modelo*

Observando el gráfico del modelo junto con las bandas de predicción, se identifica que tanto Ceuta (51), como Melilla (52) son valores atípicos ya que se sitúan fuera de las bandas de confianza del 95 %. Estas provincias presentan unos niveles de abstención por encima del 60 %. Al ser valores atípicos, la decisión tomada es eliminar estas provincias del conjunto de datos a partir de este momento.

Una vez eliminadas dichas observaciones se vuelve a ajustar el mismo modelo para evaluar si es o no significativo el efecto del logaritmo del censo sobre la abstención.
Los resultados obtenidos son los siguientes:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.077153   0.077228   0.999    0.318
log(Censo)  0.027398   0.005927   4.623 4.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1043 on 448 degrees of freedom
Multiple R-squared:  0.04552,   Adjusted R-squared:  0.04339
F-statistic: 21.37 on 1 and 448 DF,  p-value: 4.965e-06
```

Figura 4.9: *Resumen primer modelo sin outliers*

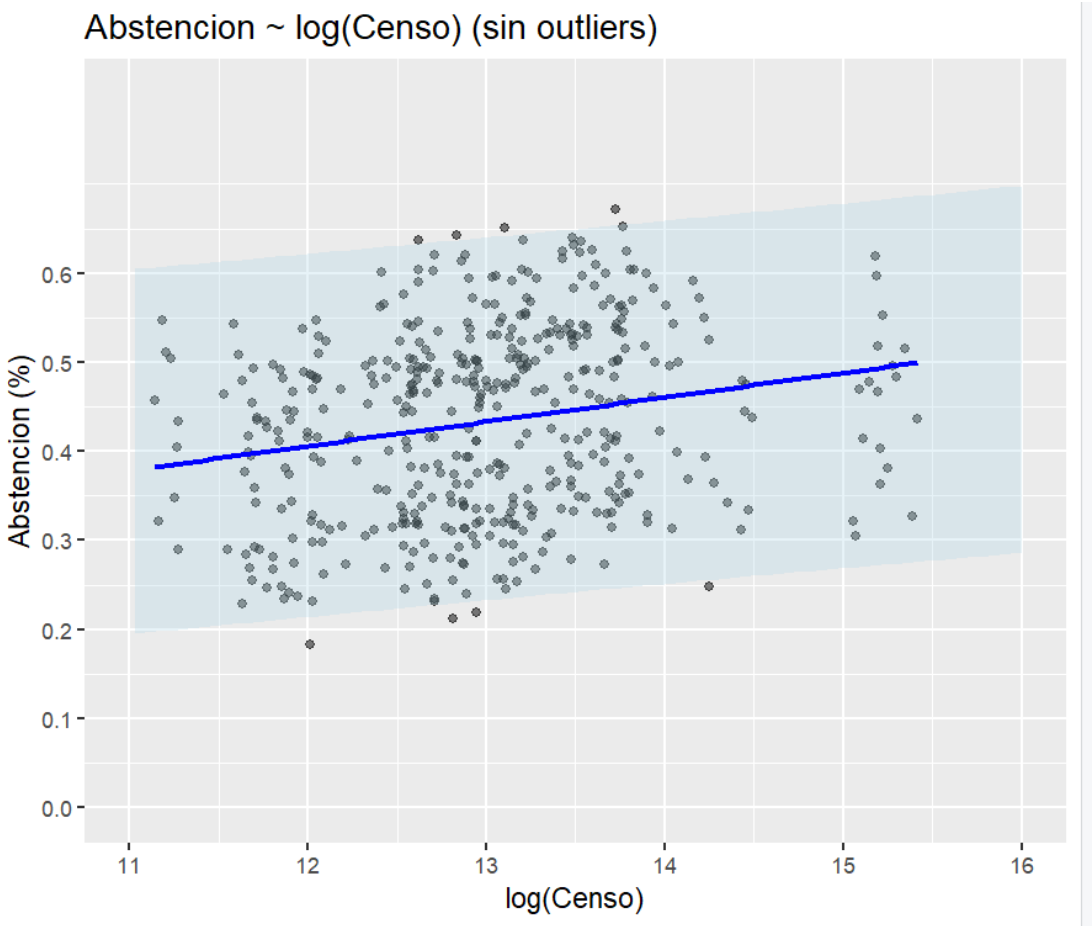


Figura 4.10: *Gráfico del primer modelo sin outliers*

Tras eliminar las observaciones correspondientes a Ceuta y Melilla se obtienen cambios significativos en el modelo como se ve en la Figura 4.9, ya que el *p-valor* obtenido es menor de 0,05, por tanto se rechaza que el efecto del logaritmo del censo sobre la abstención no sea significativo, es decir, el logaritmo del censo tiene un efecto significativo sobre la abstención.

El coeficiente estimado es 0,027398, que es positivo, por tanto a mayor tamaño del censo provincial, mayor será la abstención.

Observando el valor del coeficiente de determinación, se ve que es superior al del modelo con los valores atípicos, aunque sigue siendo bastante bajo.

Se ajusta ahora un nuevo modelo introduciendo una nueva variable explicativa junto con la interacción entre esta variable y el logaritmo del censo. La variable explicativa añadida es la convocatoria, la cual es un factor con 9 niveles. Cabe destacar que como con el anterior modelo se observó que Ceuta y Melilla eran valores atípicos, no se tienen en cuenta a la hora de realizar el ajuste de los siguientes modelos.

```
Analysis of Variance Table

Response: Porcentaje_abstencion
                 Df Sum Sq Mean Sq  F value Pr(>F)
log(Censo)        1 0.2323 0.23228  80.0674 <2e-16 ***
Agno              8 3.5866 0.44832 154.5368 <2e-16 ***
log(Censo):Agno   8 0.0302 0.00377   1.3011 0.2409
Residuals       432 1.2533 0.00290
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 4.11: *Resultados ANOVA segundo modelo*

En la Figura 4.11 se observa que el efecto de la interacción entre la variable convocatoria y el censo de la provincia no es significativo debido a que el *p-valor* (0.2409) obtenido es mayor de 0,05 , lo que indica que no hay diferencias significativas entre las pendientes de las 9 rectas de regresión, es decir, el efecto del tamaño del censo es igual en cada convocatoria.

Por tanto, se va a probar a ajustar un nuevo modelo sin tener en cuenta la interacción entre las variables explicativas, ya que se observa que tanto el efecto del logaritmo del censo, como el de la convocatoria son significativos.

Los resultados obtenidos con el ajuste del nuevo modelo son los siguientes:

```
Response: Porcentaje_abstencion
             Df Sum Sq Mean Sq F value    Pr(>F)
log(Censo)    1 0.2323 0.23228  79.632 < 2.2e-16 ***
Agno          8 3.5866 0.44832 153.696 < 2.2e-16 ***
Residuals   440 1.2835 0.00292
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 4.12: *Resultados ANOVA tercer modelo*

```
Residual standard error: 0.05401 on 440 degrees of freedom
Multiple R-squared:  0.7485,    Adjusted R-squared:  0.7433
F-statistic: 145.5 on 9 and 440 DF,  p-value: < 2.2e-16
```

Figura 4.13: *Resumen del tercer modelo*

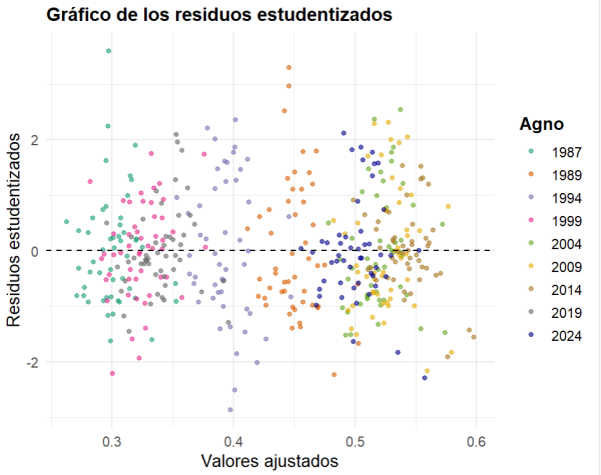


Figura 4.14: *Gráfico de los residuos estudentizados del tercer modelo.*

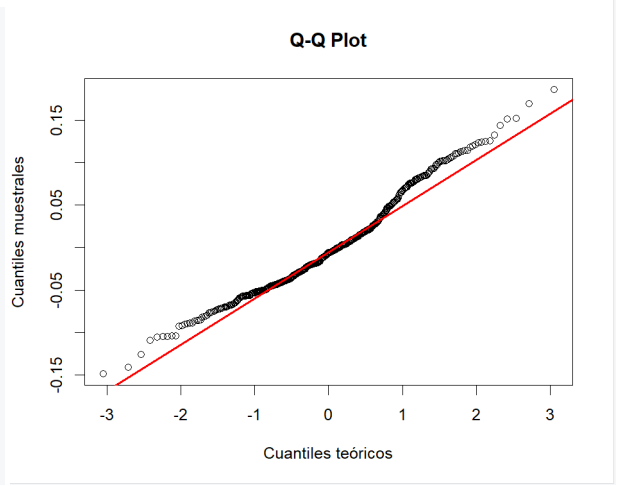


Figura 4.15: *Gráfico de normalidad del tercer modelo.*

Se revisa la Figura 4.14 para poder validar las hipótesis del modelo. Los residuos del modelo se distribuyen sin patrones ni tendencias y aleatoriamente alrededor del 0, lo que indica que la hipótesis de linealidad se valida de forma adecuada. La dispersión de los residuos es constante a lo largo de los valores ajustados, por tanto también se valida la homogeneidad de la varianza. Además en la Figura 4.15 se observa que los residuos se distribuyen según una normal. Por tanto, el comportamiento de los residuos respalda la idea de que el modelo es válido.

En la Figura 4.12 se observa que el efecto de ambas variables explicativas son significativos, dado que el *p-valor* obtenido es menor que 0,05 para ambas variables. Además en 4.13 se obtiene un valor del coeficiente de determinación bastante elevado de 0,7485, lo que indica que aproximadamente el 74,85 % de la variabilidad de la abstención es explicada por el logaritmo del censo y el año de cada convocatoria, sugiriendo un buen ajuste del modelo a los datos.

Por último, se realiza el gráfico del modelo, donde se representan las rectas de regresión de cada una de las nueve convocatorias.

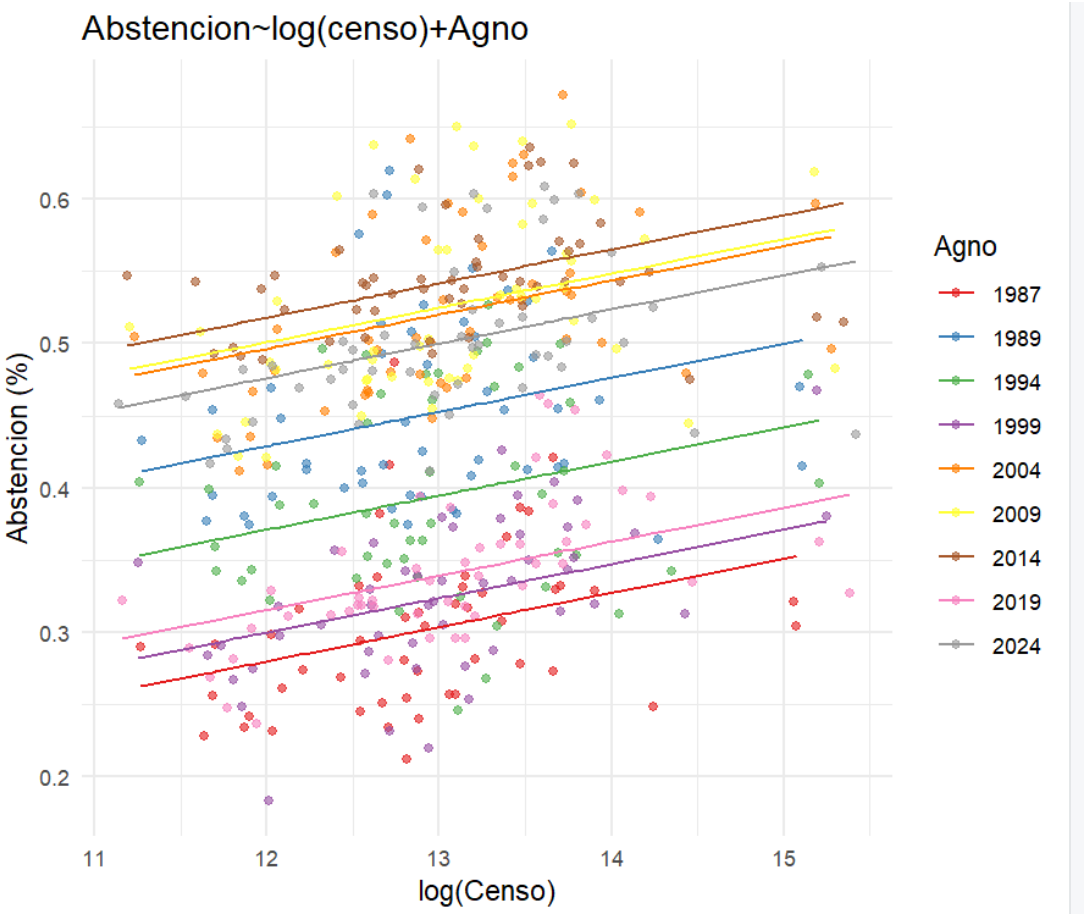


Figura 4.16: *Gráfico del tercer modelo*

Del gráfico del modelo 4.16 se puede extraer que dado que las pendientes de las rectas de regresión de cada convocatoria son positivas, cuanto mayor es el número de habitantes censados en la provincia, mayor es el porcentaje de abstención. Además, se puede asegurar que el efecto del censo es constante a lo largo de las convocatorias, debido a que no hay diferencias significativas entre las pendientes. También, en el gráfico se ve que la convocatoria de 1987 presenta los niveles más bajos de abstención, mientras que en 2009 y 2014 se tienen los niveles más altos.

4.1.3. Comparación de la abstención por convocatorias

Se realiza un análisis de la varianza entre la abstención y la convocatoria, para poder evaluar la influencia que tiene cada convocatoria en los niveles de abstención. En primer lugar, se ajusta un modelo ANOVA de un factor fijo, obteniendo los siguientes resultados:

```
            Df Sum Sq Mean Sq F value Pr(>F)
Agno         8  3.869  0.4836   120.9 <2e-16 ***
Residuals  459  1.836  0.0040
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 4.17: *Tabla ANOVA modelo con outliers*

El *p-valor* asociado a la convocatoria es $< 2.2 \times 10^{-16}$, por tanto se rechaza que el porcentaje de abstención sea igual en cada convocatoria, es decir, existen diferencias significativas entre las medias del porcentaje de abstención en las nueve convocatorias.
Se continúa representando los valores predichos y los residuos en un mismo gráfico, con el objetivo de detectar posibles outliers.

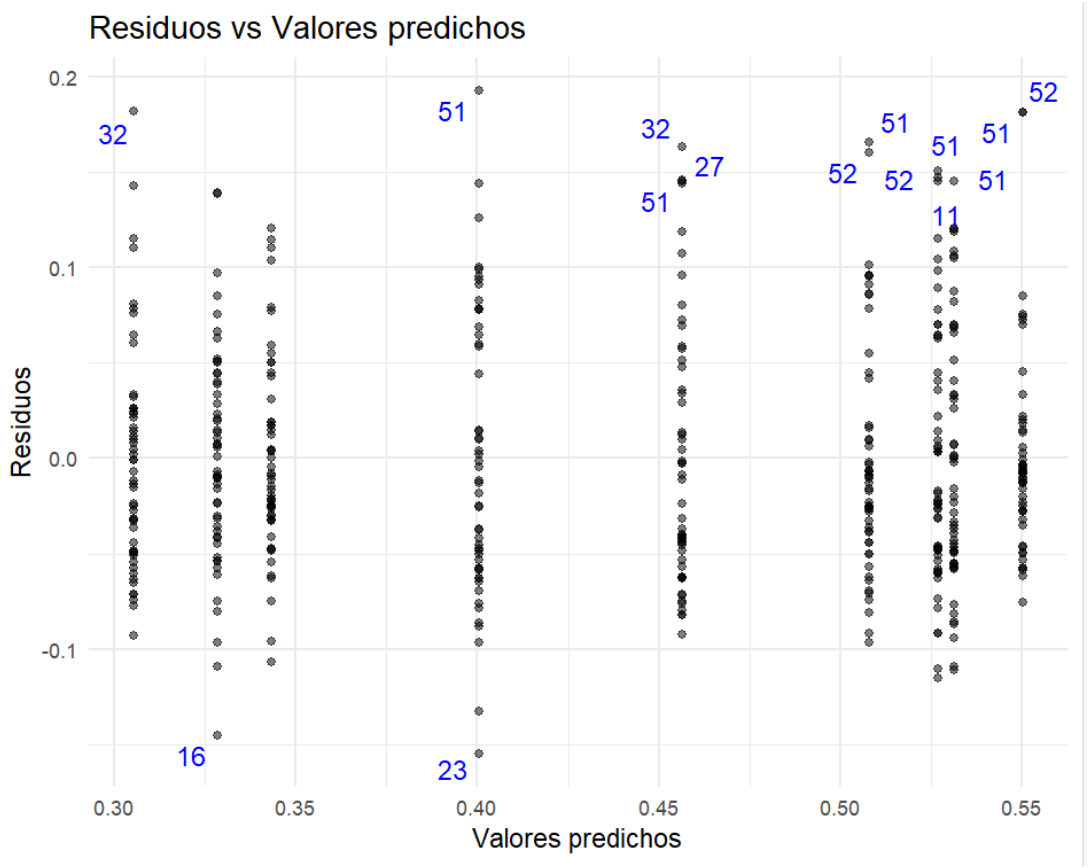


Figura 4.18: *Gráfico residuos vs predichos*

Se observa que Ceuta (51) y Melilla (52) son outliers en la mayoría de convocatorias, al igual que pasaba en los modelos de regresión ajustados anteriormente, por tanto se van a eliminar del conjunto de datos.
Una vez eliminadas del conjunto de datos, se vuelve a ajustar el mismo modelo ANOVA con las 50 provincias restantes, obteniendo:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Agno          8  3.645  0.4556   137.9 <2e-16 ***
Residuals   441  1.457  0.0033
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 4.19: *Tabla ANOVA sin outliers*

Se obtiene el mismo resultado que antes, y es que hay diferencias significativas entre convocatorias ya que p-valor $< 2.2 \times 10^{-16}$.
Se van a comprobar ahora las hipótesis del modelo.

Homogeneidad de la varianza

Se obtiene el siguiente gráfico de los residuos frente a los valores predichos, donde se observa que no hay ningún patrón y la dispersión es uniforme, por tanto no se rechaza la hipótesis de homogeneidad de la varianza.

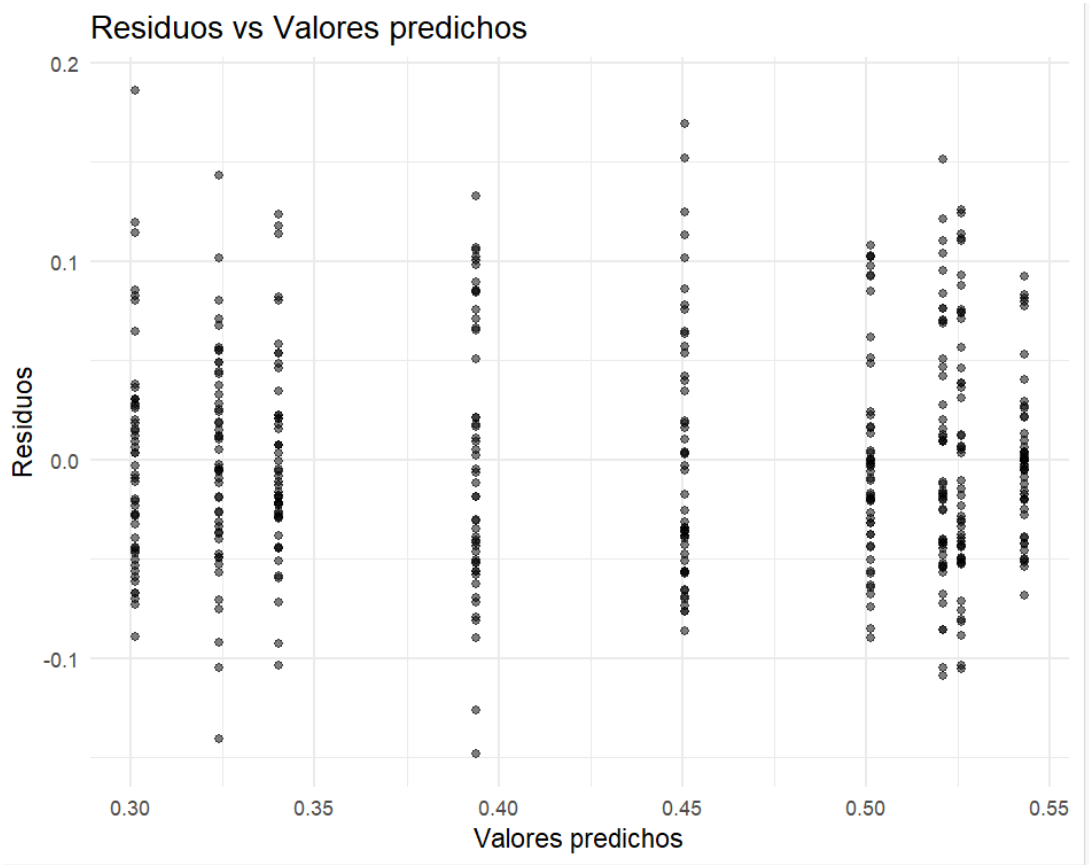


Figura 4.20: *Gráfico residuos vs predichos del ANOVA sin outliers*

Normalidad

Se realiza tanto un histograma de los residuos como el Q-Q plot para validar el supuesto.

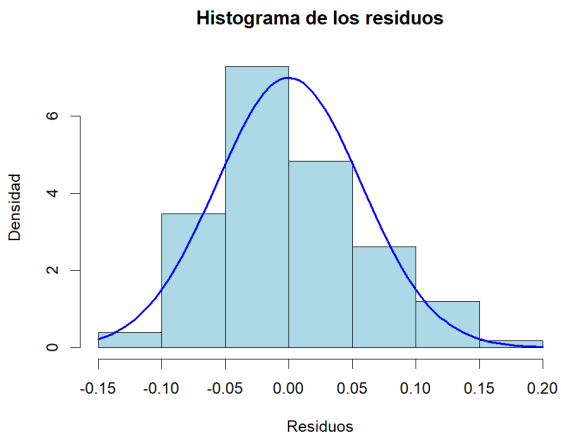


Figura 4.21: *Histograma de los residuos*

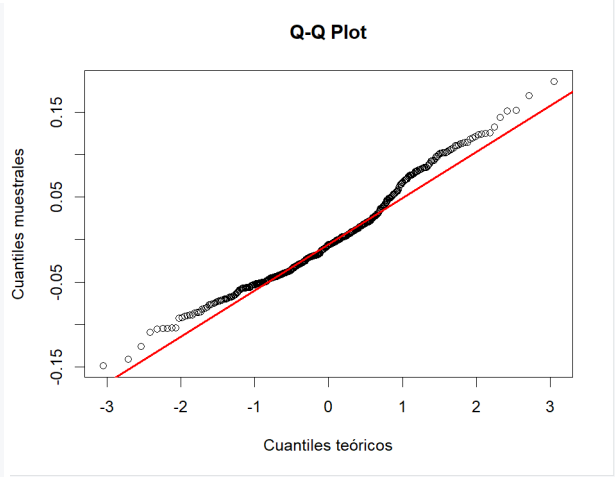


Figura 4.22: *Q-Q plot de los residuos*

Ambos gráficos muestran una distribución aproximadamente normal de los residuos, por tanto no se rechaza la hipótesis de normalidad de los residuos.

Independencia de los residuos

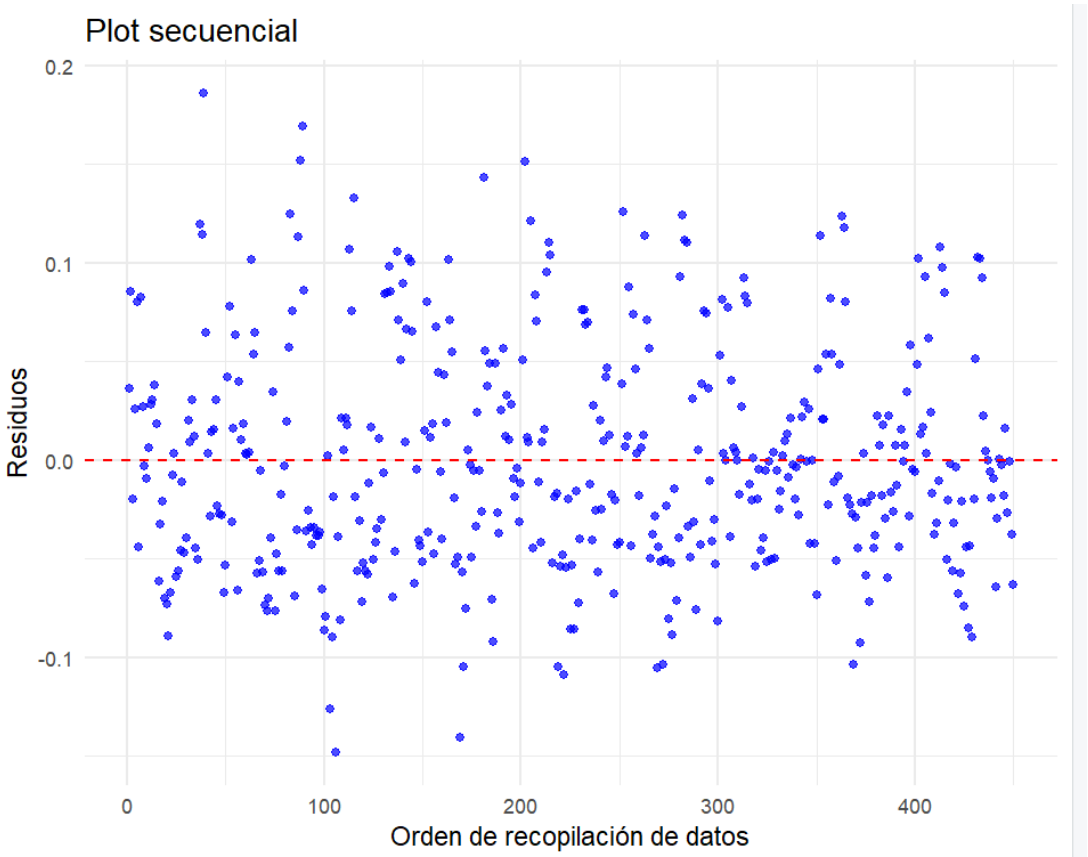


Figura 4.23: *Plot secuencial*

El gráfico realizado muestra una dispersión de los residuos aleatoria alrededor del cero, sin patrones ni tendencias. Por tanto, se cumple el supuesto de independencia de los residuos.

Dado que existen diferencias significativas entre convocatorias (Figura 4.19), y que se han validado las hipótesis del modelo, se continua con un análisis post-hoc. Para ello se realiza el test de Duncan para poder agrupar las convocatorias en grupos homogéneos [18].

	Porcentaje_abstencion	groups
2014	0.5430829	a
2009	0.5259793	a
2004	0.5207714	ab
2024	0.5012490	b
1989	0.4507336	c
1994	0.3938205	d
2019	0.3401624	e
1999	0.3238376	e
1987	0.3011156	f

Figura 4.24: *Resultados test de Duncan*

Se obtienen 3 grupos homogéneos:

- **Grupo 1:** formado por las convocatorias de 2014, 2009, 2004 y 2024, que presentan los niveles de abstención más altos, con un rango de abstención de 50,1 %–54,3 %.
- **Grupo 2:** incluye las convocatorias de 1994 y 1989. Son convocatorias donde la abstención es intermedia, con un rango de 39,4 %–45,1 %.
- **Grupo 3:** formado por las convocatorias de 1987, 1999 y 2019. Destacan por presentar una abstención más baja que el resto de convocatorias con un rango de 30,1 %–34 %.

Se puede concluir que existen diferencias significativas en la media de la abstención en las distintas convocatorias. Mediante el test de Duncan se han identificado tres grupos homogéneos. El primer grupo comprende los años de mayor crisis y polarización política en España (2004-2024), alcanzando el menor porcentaje de participación en el año 2014. En este grupo cabría esperar a la convocatoria de 2019, pero como ya se ha comentado, el motivo por el que la abstención es más baja es la celebración ese mismo año de las elecciones generales. Por otra parte, el segundo y tercer grupo reflejan las primeras convocatorias de estas elecciones, además de un periodo más estable en la política española, destacando por una menor abstención. De forma general se observa que los niveles de abstención son muy superiores al del resto de elecciones. Por ejemplo, en la convocatoria de 2019 la abstención en las elecciones europeas fue del 39,30 %, frente al 28,24 % en las elecciones al Congreso de los Diputados de ese mismo año (convocatoria de abril). Otro ejemplo llamativo es que en 2014 la abstención en las elecciones europeas alcanza el 56,19 %, mientras que en las elecciones generales de 2015 fue del 30,33 %.

4.2. Asociación de ideologías y convocatorias

El interés de este estudio radica en poder ver la relación entre las distintas convocatorias de elecciones al Parlamento Europeo y los grandes bloques ideológicos descritos en el tercer capítulo, con el fin de observar como ha ido evolucionando la distribución del voto a lo largo de los años.

Para realizar este Análisis de Correspondencias Simples entre las ideologías y las 9 convocatorias, se usa el número de votos conseguido por cada ideología en cada convocatoria.

Para comenzar, se realiza el test χ^2 de independencia para ver si son o no independientes las variables a estudiar:

$$\begin{cases} H_0 : \text{no hay asociación entre las convocatorias y la ideología} \\ H_1 : \text{hay asociación entre las convocatorias y la ideología} \end{cases}$$

A un nivel de significancia del 5 %, el *p-valor* obtenido es muy pequeño, p-valor $< 2.2 \times 10^{-16}$, por lo que se rechaza la hipótesis nula, es decir, se rechaza la no asociación entre ambas variables, por lo tanto tiene sentido continuar con el análisis.

Después de realizar el test anterior, se obtiene el Scree plot, que es el gráfico en el que se representa la variabilidad explicada por cada dimensión.

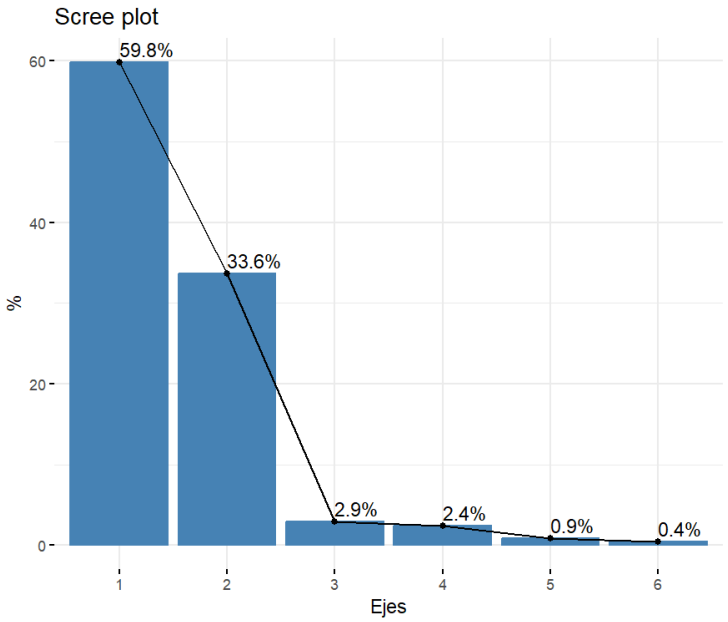


Figura 4.25: *Varianza explicada primer análisis*

En el gráfico 4.25 se observa que las dos primeras dimensiones explican conjuntamente el 93,4 % de la varianza total. El 'codo' se puede intuir a partir de la segunda dimensión, a

partir del mismo, el porcentaje de varianza total explicado por cada dimensión disminuye considerablemente, por tanto la decisión tomada es extraer las dos primeras dimensiones.

El siguiente paso del análisis es representar en dos dimensiones las ideologías y las diferentes convocatorias, obteniendo el siguiente gráfico:

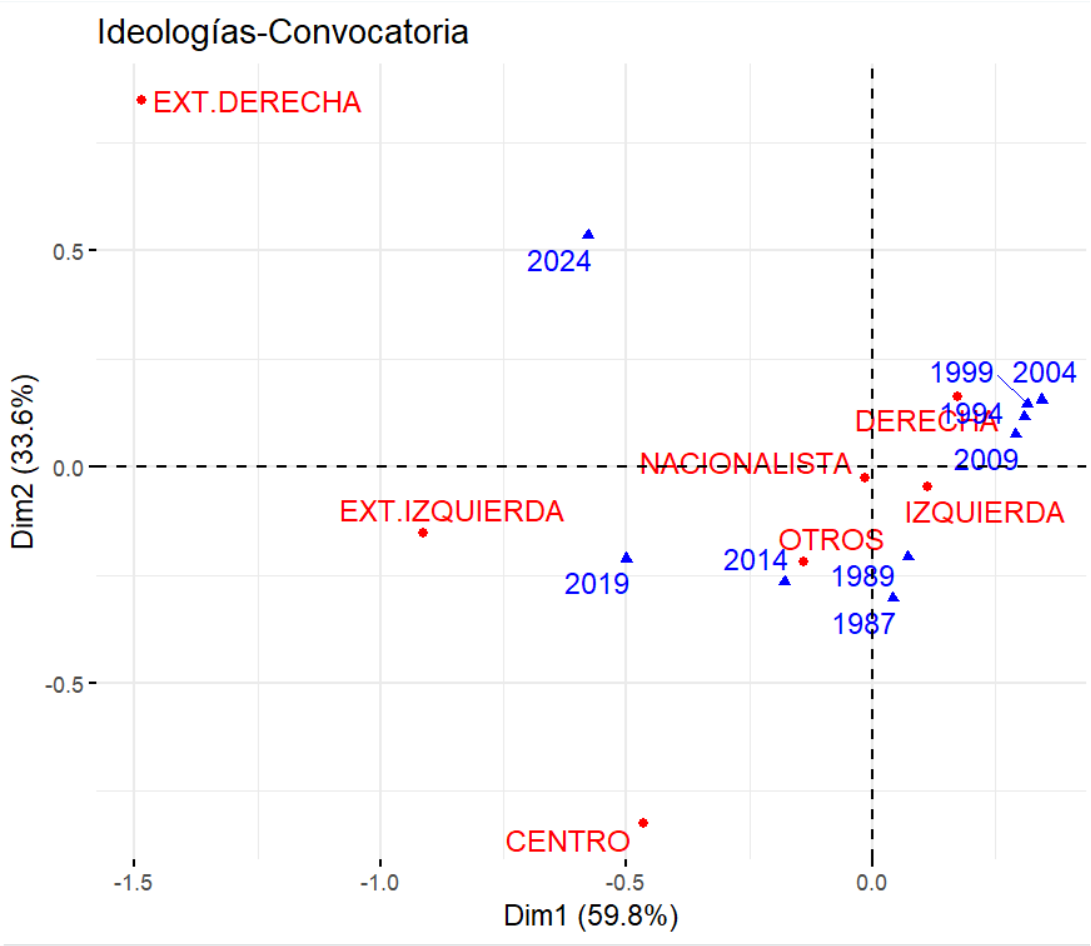


Figura 4.26: *Biplot del ACS entre ideologías y convocatorias. Primera y segunda dimensión*

De este gráfico se puede concluir que la primera dimensión separa las ideologías extremas, como son la extrema izquierda y la extrema derecha, de las ideologías más moderadas como la izquierda, derecha, centro y nacionalistas, además de separar las convocatorias más recientes (2014, 2019 y 2024), que se asocian a los extremos, de las más antiguas (1987, 1989, 1994, 1999, 2004 y 2009), que se asocian a ideologías más tradicionales.

Respecto a la segunda dimensión, se observa que separa las convocatorias más asociadas a la derecha y extrema derecha de las elecciones más asociadas a la izquierda y extrema izquierda. Esto se observa en especial en la clara diferenciación entre las dos últimas convocatorias, la de 2024, que está ligada con la extrema derecha debido al impacto reciente de partidos políticos como VOX y Se Acabó La Fiesta y la de 2019, que se asocia con la extrema izquierda, coincidiendo con el auge de Podemos.

Por otro lado, las convocatorias más antiguas se asocian a ideologías más moderadas y tradicionales como la izquierda, derecha, centro y nacionalistas, reflejando la mayor estabilidad del voto.

Cabe destacar, que se observa claramente como la convocatoria de 2014 está asociada a 'OTROS', esto es debido a que en dicho año surgieron partidos con gran impacto, como son Partido X, Partido del futuro (Partido X), el cuál es nombrado 'antipartido' [19], Movimiento de Renovación Democrática de la Ciudadanía (Movimiento RED), cuyo lider rechazaba definir el partido como de izquierdas o derechas [20] y Escaños en Blanco, cuya ideología se puede decir que es el apartidismo y es que el objetivo de este partido únicamente es mostrar el descontento de la ciudadanía con la clase política española [21].

Se puede concluir que se observa una evolución a lo largo de las convocatorias en la distribución del voto, pasando de la predominancia de ideologías tradicionales como la izquierda y la derecha, hasta el auge de los extremos en los últimos años. Esto refleja la creciente polarización y fragmentación del voto.

4.3. Análisis de Correspondencias Simples por convocatoria

En este apartado se van a estudiar las asociaciones entre las provincias y los grandes bloques ideológicos, con el fin de observar como ha ido evolucionando la distribución del voto en cada provincia a lo largo de los años.

Para llevar a cabo este análisis, se seleccionaron 4 convocatorias, la de 1987, 2014, 2019 y 2024. Esto se debe, a que en la anterior sección, se observó una clara diferenciación en los patrones ideológicos a lo largo de las convocatorias. En concreto, las convocatorias de 1987 hasta 2009 mostraban una mayor asociación a ideologías tradicionales, es decir más asociadas a políticas moderadas. Por el contrario, a partir de la convocatoria de 2014, se observó un cambio significativo, con una creciente influencia de posiciones más extremas, en concreto la extrema izquierda en 2019 y la extrema derecha en 2024. Por tanto las convocatorias seleccionadas permiten contrastar dos periodos diferenciados, uno más tradicional (1987) y otro con mayor polarización ideológica (2014-2024).

De forma previa al análisis, se realiza el test χ^2 de independencia para cada una de las 4 convocatorias, donde la hipótesis a contrastar es:

$$\begin{cases} H_0 : \text{no hay asociación entre las provincias y la ideología} \\ H_1 : \text{hay asociación entre las provincias y la ideología} \end{cases}$$

Para las 4 convocatorias analizadas, el *p-valor* obtenido es $< 2.2 \times 10^{-16}$, por tanto se confirma la asociación entre provincias e ideologías de manera muy significativa, por lo que se procede a realizar el análisis.
A continuación, se analiza cada convocatoria.

4.3.1. Convocatoria 1987

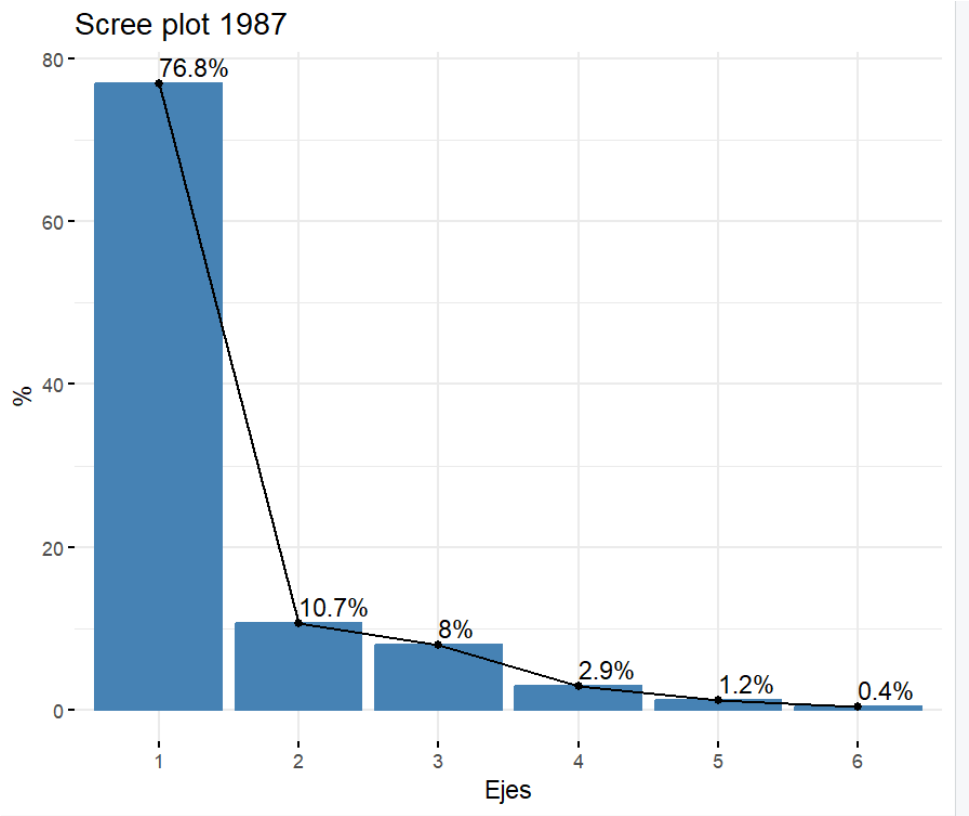


Figura 4.27: *Scree 1987*

Se observa en la Figura 4.27 que las dos primeras dimensiones explican conjuntamente un 87,5 % de la varianza total. Junto con la tercera dimensión, se explicaría un 95,5 % de la variabilidad total, por tanto serán suficientes para representar adecuadamente las asociaciones entre las provincias y las ideologías.

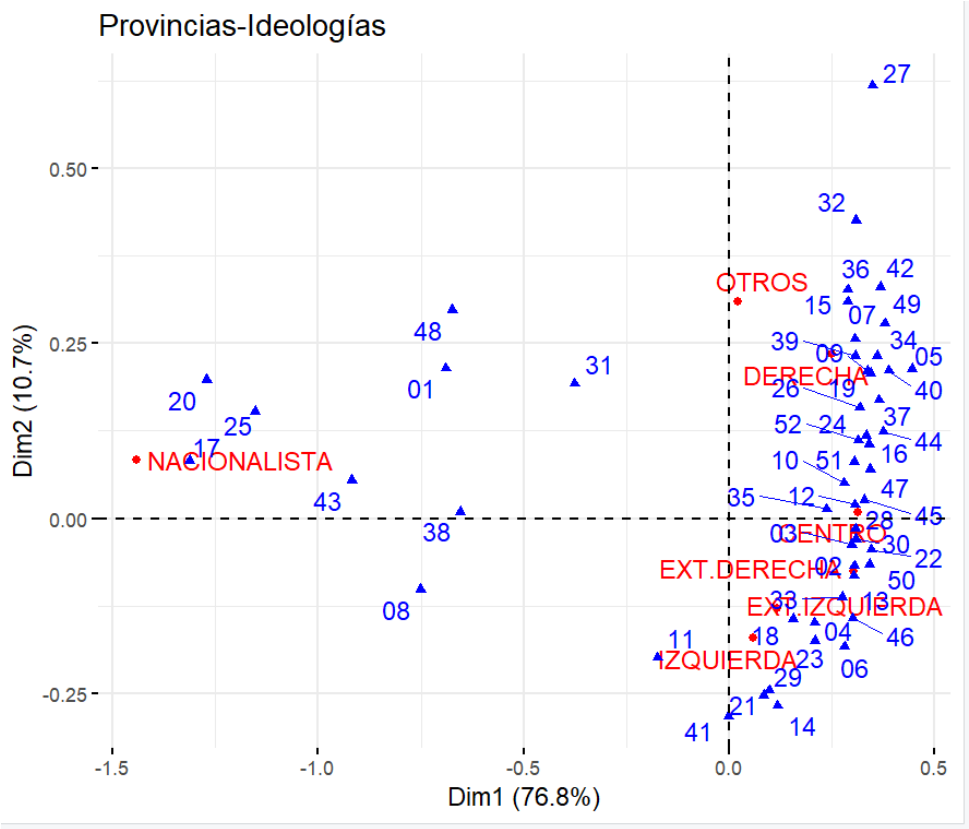


Figura 4.28: *Biplot 1987.Dimensiones 1 y 2*

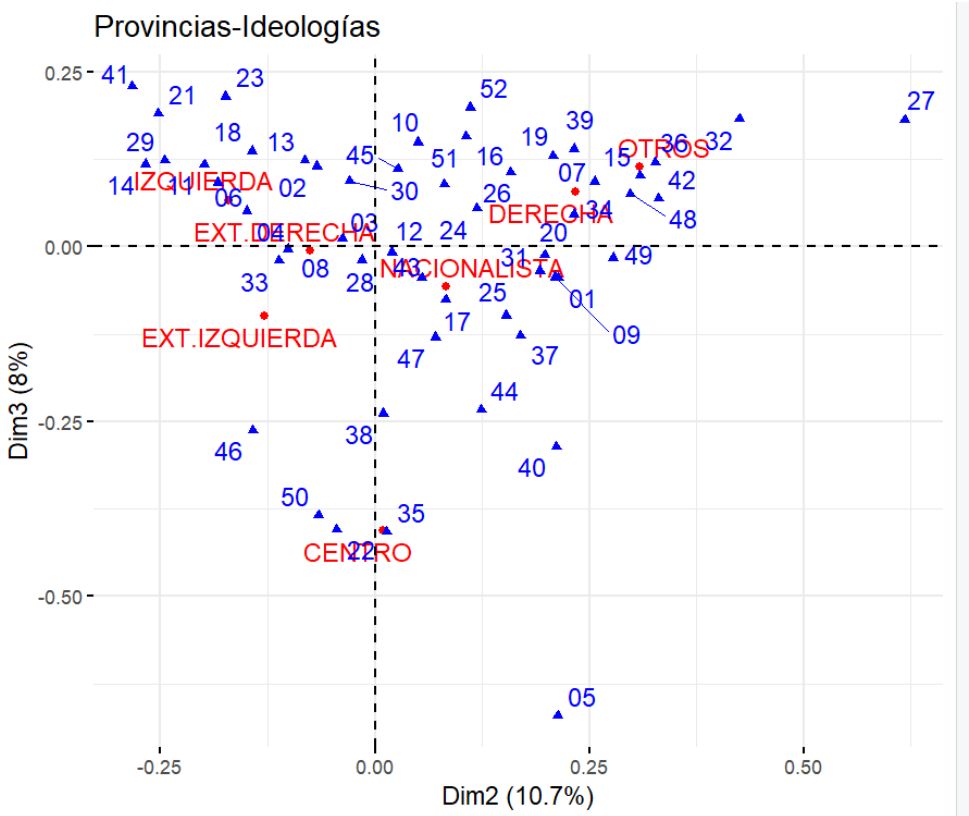


Figura 4.29: *Biplot 1987.Dimensiones 2 y 3*

En la Figura 4.28 se observa que la primera dimensión separa de manera clara las provincias asociadas al nacionalismo del resto de provincias. Las provincias más vinculadas a esta ideología son Gipuzkoa (20), Girona (17), Lleida (25), y de manera más moderada se asocian Barcelona (08), Tarragona (43), Álava (01), Bizkaia (48), Navarra (31) y Santa Cruz de Tenerife (38).

La segunda dimensión separa la izquierda y la derecha, tal y como se observa en la Figura 4.29. Se puede apreciar que las provincias más asociadas a la izquierda son Sevilla (41), Huelva (21), Jaén (23), Málaga (29), Granada (18), Ciudad Real (13), Albacete (02), Cádiz (11), Córdoba (14) y Badajoz (06). Se trata de provincias principalmente pertenecientes a la comunidad de Andalucía y Castilla-La Mancha. Respecto a las provincias más asociadas a la derecha destacan las provincias de Cantabria (39), Palencia (34), Baleares (07), Guadalajara (19), Melilla (52), Ceuta (51), La Rioja (26), Cáceres (10), Burgos (09), Salamanca (37), Cuenca (16), Castellón (12), León (24), Zamora (49) y Murcia (30).

Las ideologías extremas aparecen muy próximas en ambos gráficos realizados, y no se observa una clara separación entre las provincias asociadas a la extrema izquierda o a la extrema derecha, lo cual dificulta asociar las provincias específicamente a una ideología u otra.

Tal y como se observa en la Figura 4.29 la tercera dimensión separa las provincias asociadas al centro respecto del resto de ideologías, por tanto es claro ver cuales son las provincias más asociadas a este ideología. Destacan provincias como Valencia (46), Zaragoza (50), Las Palmas (35), Huesca (22), Ávila (05), Segovia (40), Teruel (44), Valladolid (47), Toledo (45) y Santa Cruz de Tenerife (38), esta última provincia también aparecía asociada a la ideología nacionalista, aunque en este gráfico, parece claramente más asociada al centro.

Tras realizar el Análisis de Correspondencias Simples en esta convocatoria, se puede concluir que la ideología nacionalista destaca en provincias del País Vasco y de Cataluña, mientras que la izquierda se asocia principalmente con provincias andaluzas y de Castilla-La Mancha. Respecto a la derecha y al centro, estas ideologías se asocian a provincias más variadas de diversos lugares de España. En la convocatoria del 1987, las ideologías extremas no presentaban una asociación fuerte con provincias, esto es debido a que de manera generalizada, son las ideologías con menor porcentaje de voto en todas las provincias.

4.3.2. Convocatoria 2014

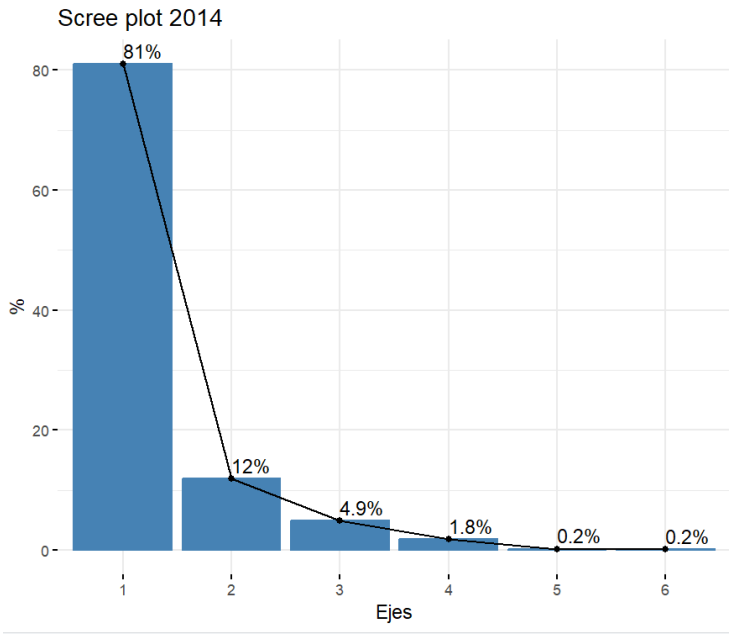


Figura 4.30: *Scree 2014*

Extrayendo las tres primeras dimensiones, se explica un 97,9 % de la variabilidad total, por tanto serán suficientes.

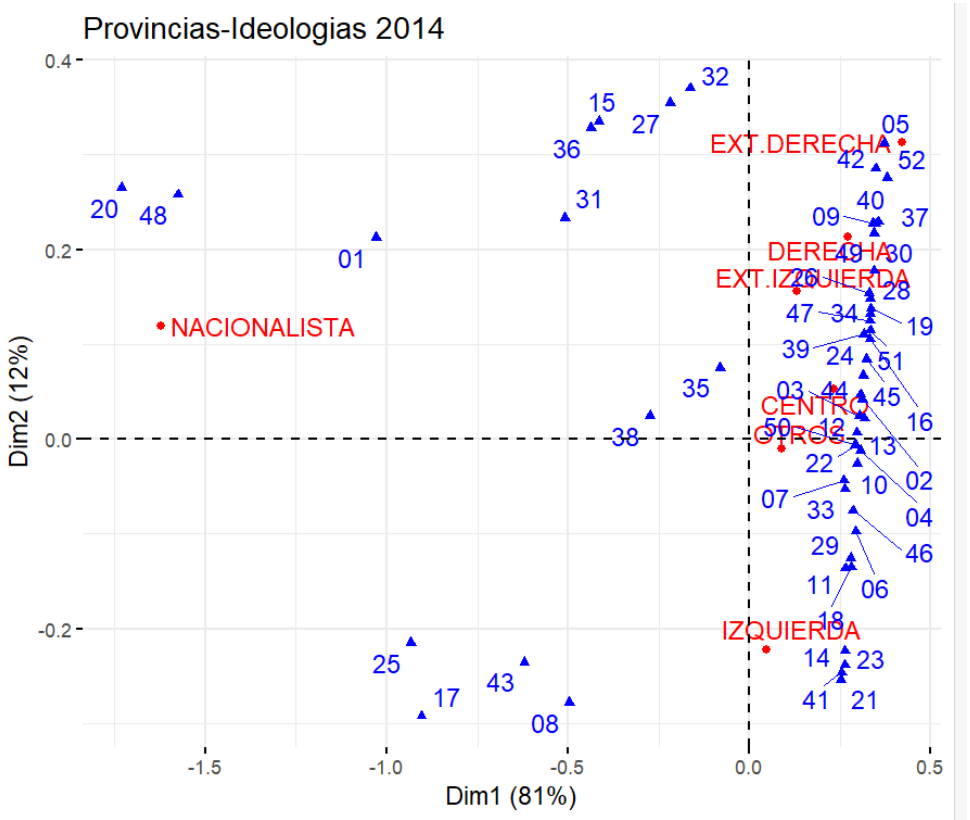


Figura 4.31: *Biplot 2014.Dimensiones 1 y 2*

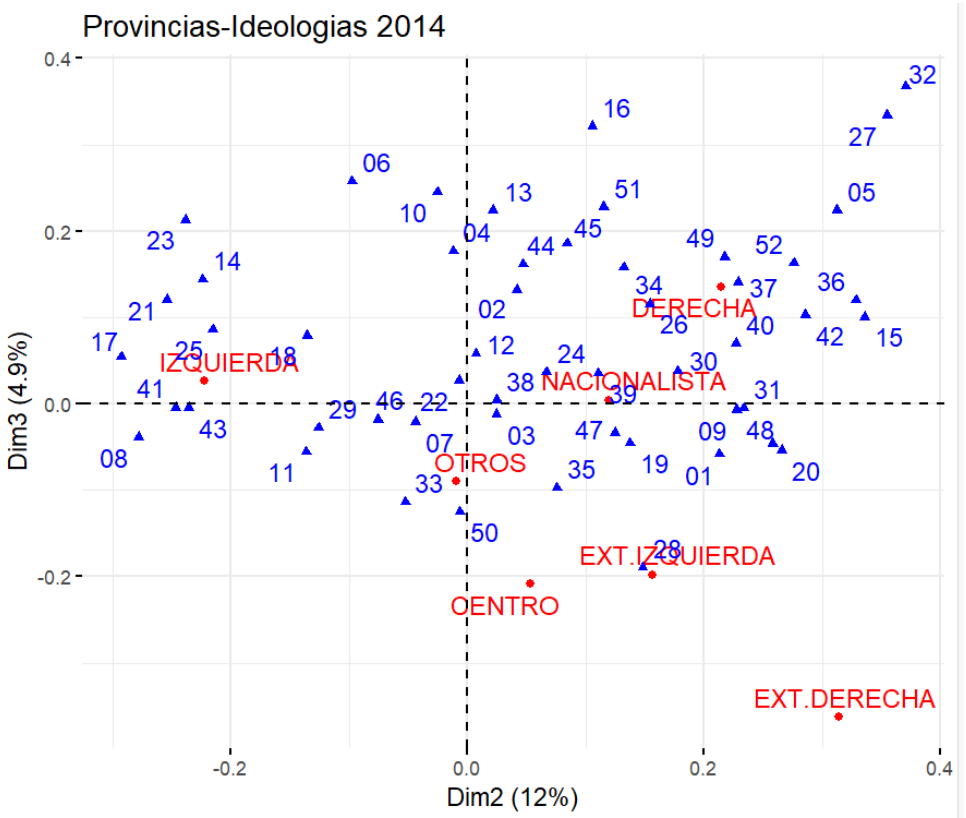


Figura 4.32: *Biplot 2014.Dimensiones 2 y 3*

En la Figura 4.31 se observa que, al igual que en la convocatoria de 1987, la primera dimensión separa de manera clara las provincias asociadas al nacionalismo del resto de provincias. Las provincias más asociadas a esta ideología son Gipuzkoa (20), Girona (17), Lleida (25), Barcelona (08), Tarragona (43), Álava (01), Bizkaia (48). Se observa que tanto Navarra (31) y Santa Cruz de Tenerife (38), que en la anterior convocatoria si que aparecían asociadas a esta ideología, en este año parecen asociadas a varias ideologías como se puede ver de forma más clara en la Figura 4.32, en el caso de Navarra, parece que tiene cierta asociación tanto con la ideología nacionalista como la extrema izquierda, mientras que Santa Cruz de Tenerife aparece asociada a los nacionalistas y a 'OTROS', que cabe recordar que recoge candidaturas minoritarias de todas las ideologías.

Respecto a la segunda dimensión, en la Figura 4.31 se puede intuir que separa la izquierda del resto de ideologías, lo cual se confirma en la Figura 4.32. Las provincias con una fuerte asociación a la ideología de izquierdas son Barcelona (08), Tarragona (43), Cádiz (11), Sevilla (41), Lleida (25), Granada (18), Girona (17), Huelva (21), Córdoba (14), Jaén (23). También las provincias de Málaga (29), Valencia (46), Huesca (22), Badajoz (06) y Cáceres (10) aparecen asociadas aunque de manera más moderada.

La tercera dimensión separa las provincias asociadas a la izquierda, derecha y nacionalista del resto de provincias, tal y como se ve en la Figura 4.32.

Las provincias asociadas a la extrema derecha se ven de forma más clara en la Figura 4.31, estas son Ávila (05), Melilla (52), Ourense (32), Lugo (27), A Coruña (15) y Pontevedra (36). Mientras que la provincia más asociada a la extrema izquierda es Madrid (28). Por otro lado, las provincias asociadas a la derecha son Burgos (09), Cuenca (16), León (24), La Rioja (26), Murcia (30), Palencia (34), Salamanca (37), Cantabria (39), Segovia (40), Soria (42), Zamora (49), Ceuta (51) y Valladolid (47), que esta última aparece en posiciones más ambiguas. Por último las provincias más asociadas al centro son Zaragoza (50), Castellón (12), Toledo (45).

Tras aplicar el ACS a esta convocatoria se observa que las provincias catalanas y vascas están claramente asociadas al nacionalismo. La izquierda tiene una fuerte presencia tanto en Andalucía como en Cataluña, mientras que la derecha aparece principalmente en zonas de Castilla y León. Un cambio relevante respecto a la anterior convocatoria analizada es la aparición de las ideologías extremas. En la convocatoria de 1987, se comentó que ambas ideologías aparecían muy próximas y por tanto era complejo asignar las provincias asociadas a cada ideología, pero en esta convocatoria se observa que ambas están bastante separadas y diferenciadas. La extrema izquierda se concentra principalmente en Madrid, mientras que la extrema derecha destaca en provincias como Ávila, Melilla, A Coruña... Esto puede explicarse debido a la aparición de partidos políticos que irrumpieron con mucha fuerza, como es el caso de Podemos para la extrema izquierda, y VOX para la extrema derecha, reflejando así una mayor polarización del voto.

4.3.3. Convocatoria 2019

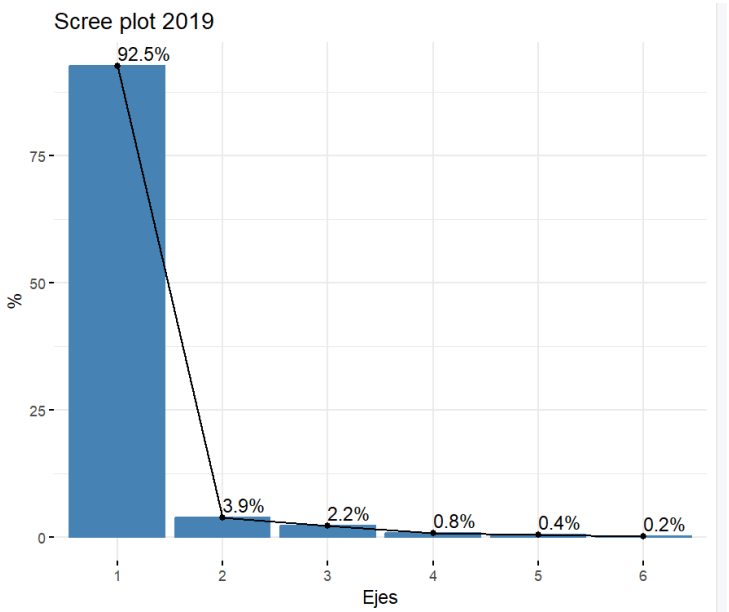


Figura 4.33: *Scree 2019*

Se observa en la Figura 4.33 que con las dos primeras dimensiones se explica conjuntamente el 96,4 % de la varianza total.

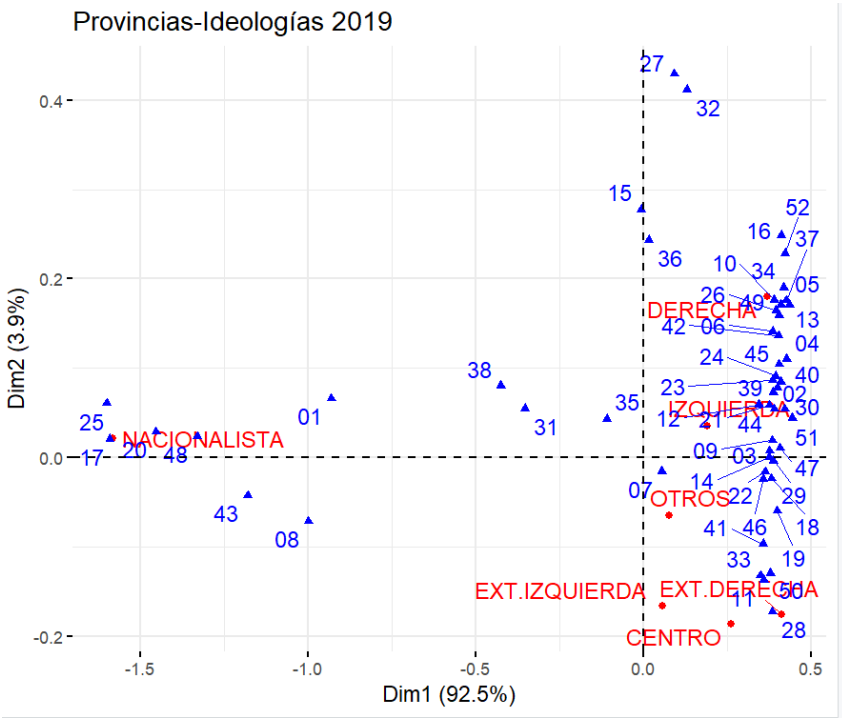


Figura 4.34: *Biplot 2019.Dimensiones 1 y 2*

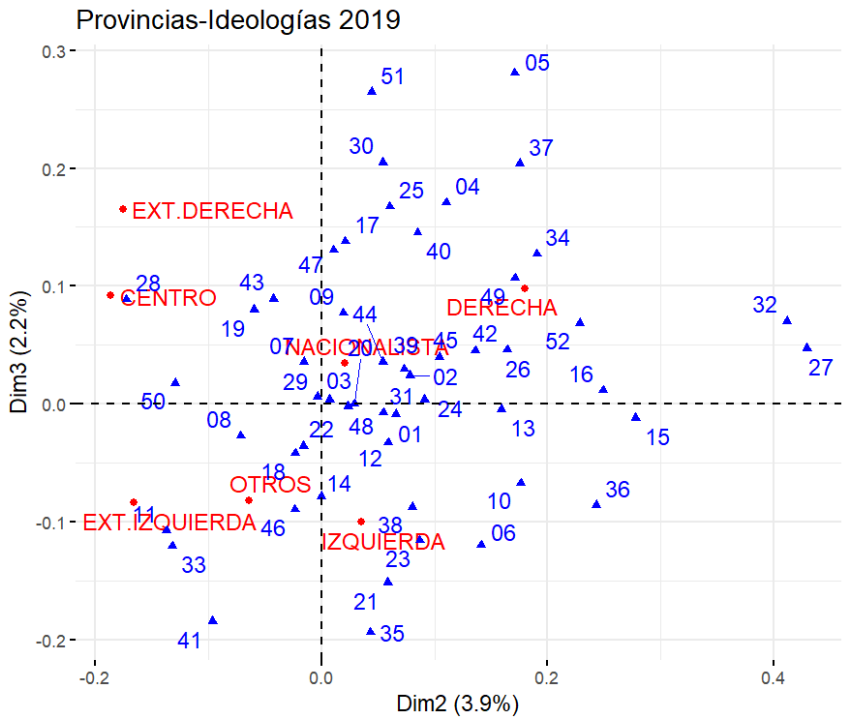


Figura 4.35: *Biplot 2019.Dimensiones 2 y 3*

Como ocurre en anteriores análisis en la Figura 4.34 se observa que la primera dimensión separa las provincias asociadas a la ideología nacionalista del resto, estas provincias son

Lleida (25), Girona (17), Gipuzkoa (20), Bizkaia (48), Álava (01), Tarragona (43), Barcelona (08), en menor medida se tienen las provincias de Navarra (31) y Santa Cruz de Tenerife (38).

La segunda dimensión separa principalmente las ideologías extremas (tanto extrema derecha y extrema izquierda), y nacionalistas, del resto de ideologías más moderadas. Mientras que la tercera dimensión separa las ideologías de izquierdas (izquierda y extrema izquierda), de las ideologías de derechas (derecha y extrema derecha), tal y como se aprecia en la Figura 4.35.

En base a lo observado en la Figura 4.35 se puede determinar que las provincias de Huelva (21), Albacete (02), Córdoba (14), Santa Cruz de Tenerife (38), Badajoz (06), Cáceres (10), Castellón (12), Baleares (07), Jaén (23) y Las Palmas (35) se asocian claramente a la izquierda, mientras que provincias como Cádiz (11), Sevilla (41) y Asturias (33), se asocian con la extrema izquierda.

Del mismo gráfico se puede extraer que las provincias asociadas a la derecha son Cuenca (16), Palencia (34), León (24), Salamanca (37), Ciudad Real (13), Cantabria (39), Segovia (40), Soria (42), Zamora (49), La Rioja (26), Toledo (45), Teruel (44), Almería (04), Málaga (29), Melilla (52), Ceuta (51), Lugo (27), Ourense (32), A Coruña (15), Pontevedra (36) y Ávila (05). Respecto a las provincias asociadas a la extrema derecha son Valladolid (47), Murcia (30) y Girona (17), que se mencionó anteriormente su asociación a la ideología nacionalista.

Respecto al centro, destacan las provincias de Madrid (28), Zaragoza (50), Burgos (09) y Guadalajara (19) y Tarragona (43), que a pesar de aparecer en la Figura 4.34 asociada a los nacionalistas, en 4.35 aparece asociada al centro.

Tras realizar el análisis para la convocatoria de 2019, se observa que las provincias catalanas y vascas siguen claramente asociadas a la ideología nacionalista. La izquierda sigue manteniendo una fuerte presencia en provincias del sur de la península, mientras que la extrema izquierda se concentra en provincias como Cádiz y Sevilla, las cuales han experimentado una transición hacia ideologías más extremas, moviéndose de la izquierda a la extrema izquierda. La derecha destaca principalmente en la zona interior de la península, en provincias de Galicia y Castilla y León. La extrema derecha destaca de forma llamativa en Girona, sugiriendo la consolidación de VOX.

Se puede concluir que claramente se va aumentando la polarización y la fragmentación del voto.

4.3.4. Convocatoria 2024

Tras representar la variabilidad explicada por cada dimensión se obtiene un gráfico muy similar a la Figura 4.33 de la anterior convocatoria, donde la primera dimensión explica gran parte de la variabilidad total, en concreto un 91,8 %, y extrayendo las dos primeras dimensiones se explica un 96,3 %. Es por esto, que no se incluye dicho gráfico para esta convocatoria.

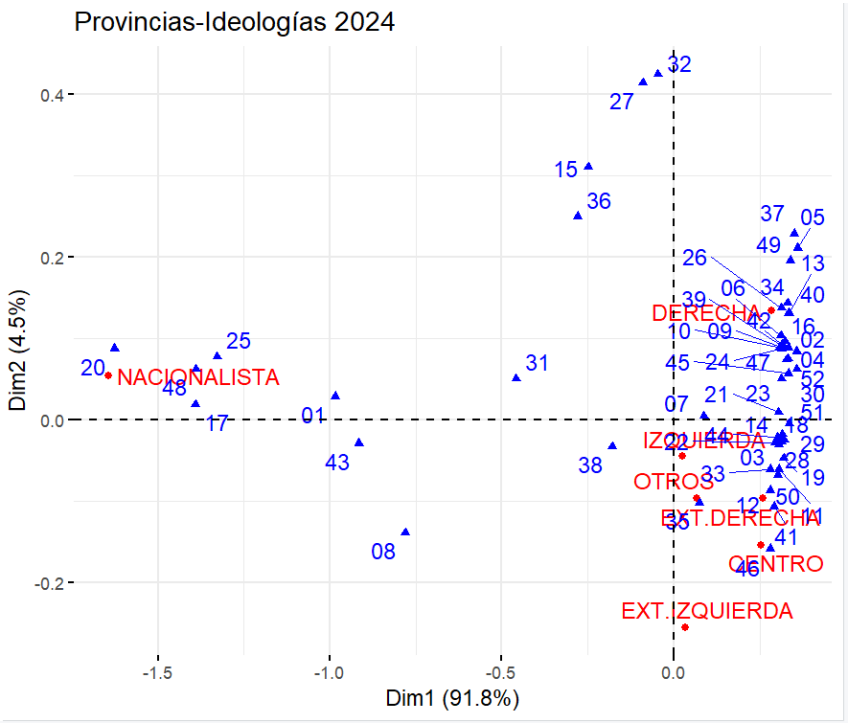


Figura 4.36: *Biplot 2024.Dimensiones 1 y 2*

La primera dimensión sigue separando las provincias asociadas a los nacionalistas del resto de provincias, tal y como ocurría en convocatorias anteriores. Las provincias más asociadas a esta ideología son Gipuzkoa (20), Bizkaia (48), Lleida (25), Girona (17), Álava (01), Tarragona (43) y Barcelona (08), mostrando la persistencia del voto nacionalista en provincias de Cataluña y País Vasco.

Mientras que la segunda dimensión parece que separa las provincias asociadas a la derecha del resto de provincias.

Se realiza el gráfico de la segunda y tercera dimensión, para poder asociar mejor las provincias a cada ideología.

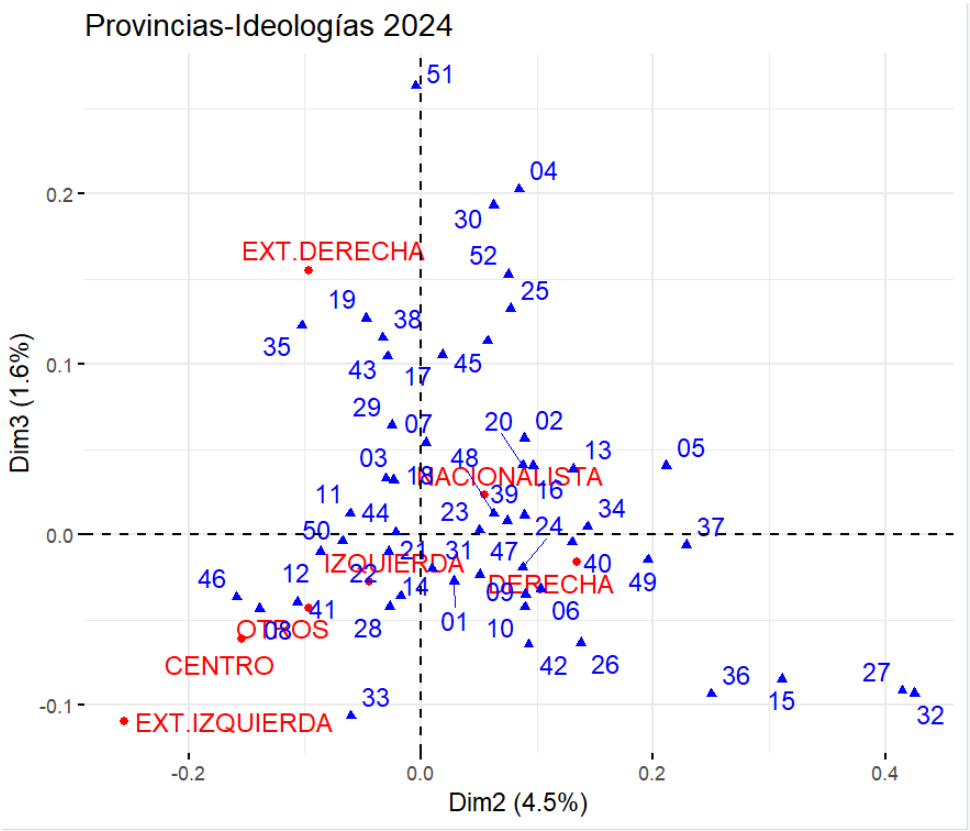


Figura 4.37: *Biplot 2024.Dimensiones 2 y 3*

Se confirma lo ya mencionado, esto es, como se podía intuir en 4.36, la segunda dimensión separa las provincias asociadas a la derecha y al nacionalismo del resto, mientras que la tercera dimensión diferencia las provincias asociadas a la extrema derecha del resto.

La extrema derecha está claramente diferenciada del resto de ideologías por la tercera dimensión, de tal manera que provincias como Ceuta (51), Melilla (52), Almería (04), Murcia (30), Toledo (45), Santa Cruz de Tenerife (38), Guadalajara (19), Las Palmas (35), Málaga (29) y Baleares (07) están asociadas a dicha ideología. Cabe destacar que provincias como Lleida (25), Tarragona (43), Girona (17) aparecen en 4.36 asociadas al nacionalismo, pero en 4.37 aparecen asociadas a la extrema derecha, observando un desplazamiento hacia posiciones más extremas debido posiblemente a cuestiones como la inmigración o insatisfacción con las políticas tradicionales.

Respecto a la extrema izquierda en esta convocatoria, únicamente se encuentra la provincia de Asturias (33) claramente asociada, observando un declive de dicha ideología en España, tras el auge de 2019.

En ambos gráficos realizados se puede observar que las provincias de Valladolid (47), Palencia (34), Burgos (09), León (24), Salamanca (37), Segovia (40), Soria (42), Zamora (49), Ávila (05), Albacete (02), Cuenca (16), La Rioja (26), Pontevedra (36), A Coruña

(15), Lugo (27), Ourense (32), Cantabria (39), Ciudad Real (13), Badajoz (06) y Cáceres (10) están asociadas de forma clara a la derecha.

Respecto a la izquierda, destacan provincias como Teruel (44), Huesca (22), Alicante (03), Huelva (21), Córdoba (14), Jaén (23), Granada (18), Cádiz (11), Zaragoza (50), Castellón (12), Valencia (46), Navarra (31) y Madrid (28).

Por último, las provincias asociadas al centro son Madrid (28), Valencia (46), Castellón (12) y Zaragoza (50). Estas provincias están mas asociadas a la izquierda, como se ha visto anteriormente, pero presentan cierta asociación también al centro.

Una vez realizado el análisis para la convocatoria de 2024, se observa que las provincias catalanas y vascas siguen manteniendo una clara asociación con la ideología nacionalista. Sin embargo, existe un cierto desplazamiento de algunas provincias catalanas (Lleida, Tarragona y Girona) hacia posiciones de extrema derecha, reflejando un descontento con las políticas tradicionales.

La izquierda mantiene su presencia en provincias del sur de España, principalmente en Andalucía, mientras que la extrema izquierda sufre un declive, estando asociada fuertemente únicamente con Asturias, lo que muestra la clara caída de Podemos.

Por el contrario, la extrema derecha parece consolidarse en gran parte de España, expandiéndose por numerosas provincias, evidenciando un gran crecimiento de la ideología, debido al apogeo de VOX y la aparición de Se Acabó La Fiesta. La derecha se mantiene en provincias del centro de España y de la zona norte.

Por último el número de provincias asociadas al centro continúa reduciéndose, confirmándose la tendencia de las anteriores convocatorias.

4.3.5. Conclusiones finales ACS por convocatoria

Tras realizar el Análisis de Correspondencias Simples entre las provincias y las ideologías en las convocatorias de 1987, 2014, 2019 y 2024 de las elecciones al Parlamento Europeo, se aprecia una clara transformación en la distribución del voto, destacando la creciente polarización política y el auge de los extremos en los últimos años.

En primer lugar, las provincias vascas (Gipuzkoa, Álava y Bizkaia) y catalanas (Barcelona, Tarragona, Girona y Lleida) han mantenido a lo largo de todas las convocatorias una fuerte asociación con el nacionalismo sin prácticamente cambios. Sin embargo, en las elecciones del año 2024, algunas provincias catalanas, como Tarragona, Girona o Lleida han mostrado un cierto acercamiento hacia la extrema derecha.

Por otra parte, la izquierda se ha mantenido a lo largo del tiempo principalmente en Andalucía, pero en la convocatoria de 2019 se empieza a ver cierta polarización de estas provincias hacia la extrema izquierda. Mientras que la derecha se ha consolidado por encima del resto en provincias del interior de la Península y Galicia, evidenciando un claro crecimiento.

La extrema derecha ha experimentado una clara evolución, partiendo de ser prácticamente inexistente en 1987, hasta consolidarse con gran fuerza en el 2024. Respecto a la extrema izquierda se observó que en 2019 alcanzó su máximo pico de popularidad entre las provincias coincidiendo con el auge de Podemos, pero a partir de dicha convocatoria se observa un claro declive. Esto, además de por una pérdida de apoyo, se debe a la ruptura entre Sumar y Podemos, lo cuál generó una división del voto.

Por último, con el centro ha surgido algo similar a la extrema izquierda, y es que se observó un claro auge en las convocatorias de 2014 y 2019, hasta convertirse en una ideología bastante minoritaria en las últimas elecciones.

Claramente se observa una polarización y fragmentación en el voto, partiendo de ideologías más moderadas y con una distribución del voto más estable, repartido principalmente entre el eje izquierda-derecha, hasta polos más extremos.

Se realiza ahora un mapa de España donde por colores se representan las provincias que mantienen la misma ideología política a lo largo de las convocatorias analizadas, además de la transición a extremos.

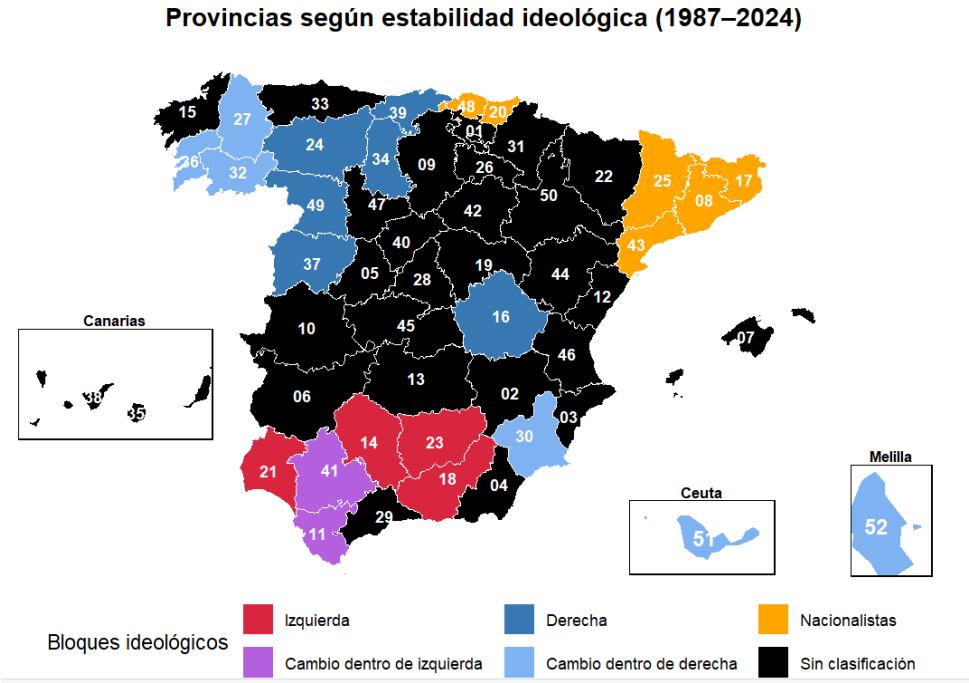


Figura 4.38: *Mapa de España con la estabilidad ideológica de provincias*

En el mapa, el color negro representa que a lo largo de las convocatorias la provincia ha experimentado variaciones en la tendencia ideológica. Por otro lado, el color azul claro y morado indican que, en la provincia ha habido un acercamiento a ideologías extremas (extrema derecha y extrema izquierda respectivamente), aunque no necesariamente hayan ganado esas opciones.

4.4. Análisis de Correspondencias Múltiple entre provincias, ideologías y convocatorias

El objetivo de este análisis es explorar las relaciones entre las provincias, ideologías políticas y las nueve convocatorias mediante un Análisis de Correspondencias Múltiples.

Al igual que en los análisis anteriores, se obtiene el gráfico donde se representa la variabilidad explicada por cada dimensión, es decir, el Scree plot.

En la Figura 4.39 se puede ver que el porcentaje de varianza total explicado por cada dimensión es bastante bajo, siendo más alto el de la primera dimensión con un 2,3 %. Al haber tantas categorías entre las 3 variables analizadas es común que esto ocurra.

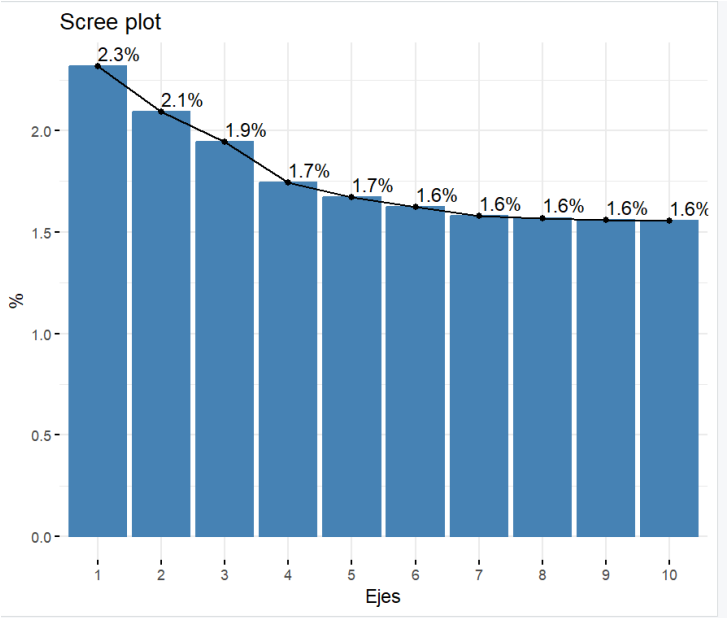


Figura 4.39: *Varianza explicada tercer análisis*

Se representan ahora los biplots correspondientes a la primera y segunda dimensión, y a la segunda y tercera dimensión.

La primera dimensión, como se ve en la Figura 4.40 separa las provincias asociadas con la ideología nacionalista respecto del resto de provincias, al igual que pasaba en los análisis de la sección anterior.

De la Figura 4.41 se extrae que la segunda dimensión distingue entre las convocatorias más antiguas y las más recientes, además de separar las provincias asociadas a las ideologías más extremas (extrema derecha y extrema izquierda) del resto.

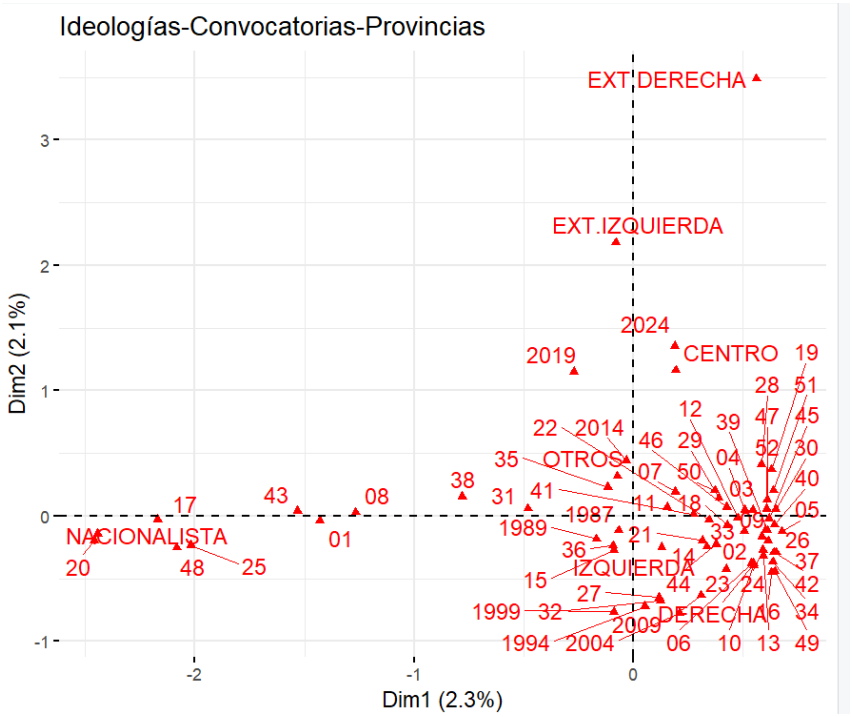


Figura 4.40: *Biplot del ACM. Primera y segunda dimensión*

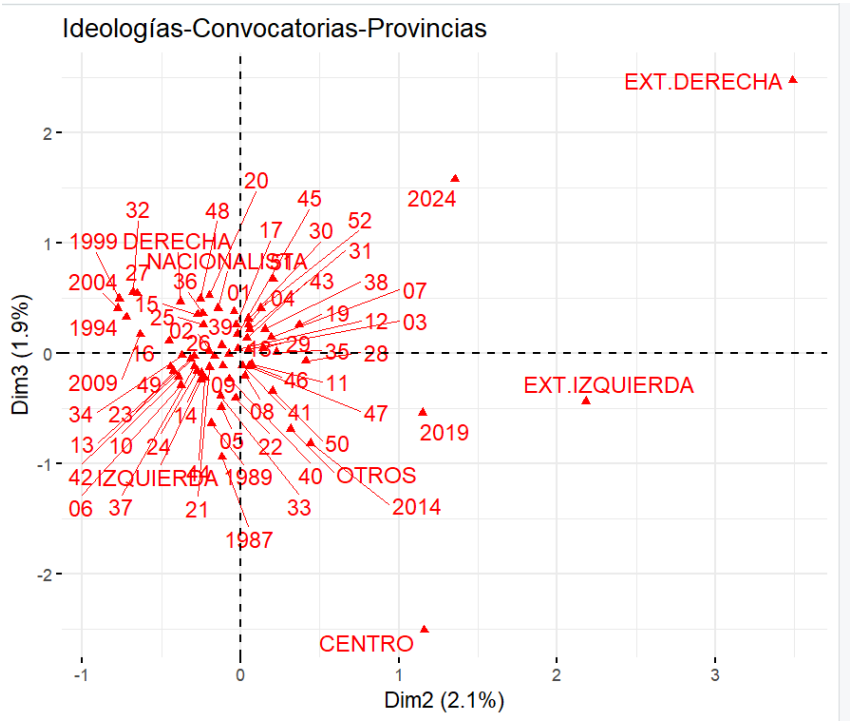


Figura 4.41: *Biplot del ACM. Segunda y tercera dimensión*

Respecto a la tercera dimensión, separa de manera general las ideologías de izquierdas (incluyendo también la extrema izquierda) de las ideologías de derechas (incluyendo la extrema derecha), diferenciando así las provincias y las convocatorias más asociadas a cada corriente ideológica.

Las primeras convocatorias (1987-2009) se asocian principalmente con ideologías como la izquierda, derecha y nacionalistas, reforzando la idea obtenida en la sección 4.2 , y es que en estos años destacaban por encima del resto las opciones más moderadas, reflejando así la menor polarización. Por otro lado, las elecciones más recientes (2014-2024) muestran el crecimiento de las opciones extremas.

Como ya se ha mencionado en el gráfico 4.40, la ideología nacionalista está claramente diferencia del resto, con una fuerte asociación a provincias como Gipuzkoa (20), Bizkaia (49), Lleida (25), Girona (17), Tarragona (43), Álava (01) y Barcelona (08). Son provincias que a lo largo de las 9 convocatorias, han estado asociadas fuertemente a esta ideología.

Respecto a la izquierda cabe destacar, que está principalmente vinculada a las elecciones de 1987 y 1989. Algunas provincias con una clara asociación son Badajoz (06), Cáceres (10), Soria (42), Zamora (49), Ciudad Real (13), Palencia (34), Jaén (23), León (24) y Huelva (21). En cambio, la derecha predomina fuertemente en las convocatorias de 1999, 2004 y 2009. Algunas de las provincias claramente asociadas son Ourense (32), Lugo (27), Pontevedra (36), A Coruña (15), La Rioja (26), Albacete (02), Cuenca (16), Salamanca (37) y Córdoba (14).

Se aprecia en numerosas ocasiones, que una misma provincia tiene una asociación ambigua entre la derecha y la izquierda, este es el caso de Alicante (03), Málaga (29), Castellón (12), Valladolid (47)...

El centro se encuentra relativamente cercano, por tanto, asociado a las elecciones de 2014 y 2019, coincidiendo justamente con el auge de Ciudadanos.

Por último, sobre las ideologías extremas se llega a la misma conclusión obtenida en el análisis de la sección 4.2 , y es que la extrema derecha se asocia de forma clara con la convocatoria de 2024 (auge de VOX y Se Acabó La Fiesta), mientras que la extrema izquierda se asocia con la del 2019 (auge de Podemos).

En resumen, el análisis muestra un evidente tránsito en la distribución del voto: partiendo de una menor polarización y predominio de opciones moderadas (1987-2009) hasta el auge de las posiciones extremas y mayor fragmentación del voto (2014-2024).

Provincias como Bizkaia, Lleida, Girona y Álava (vascas y catalanas), tienen una identidad nacionalista muy marcada y constante a lo largo del tiempo, mientras que el resto de las provincias se asocian de manera más general y ambigua al resto de ideologías.

Capítulo 5

Conclusiones y Trabajo Futuro

En este Trabajo Fin de Grado se han utilizado los datos de las nueve convocatorias celebradas en España de las elecciones al Parlamento Europeo, con el objetivo de estudiar de manera general dichas elecciones. Para ello, se ha empleado el software estadístico R. Se han usado diferentes técnicas a lo largo del estudio, como la Regresión Lineal, Análisis de Correspondencias Simples, Análisis de Correspondencias Múltiples y el Análisis de la Varianza.

En primer lugar, al realizar el estudio de la abstención se observó de manera general en prácticamente las nueve convocatorias que, cuanto mayor es el censo de la provincia, mayor es el nivel de abstención, esto se confirmó a partir de los modelos de regresión ajustados. En comparación con el TFG [22] sobre las elecciones generales en España, se observa que finalmente se obtiene un modelo con las mismas variables explicativas llegando a la misma conclusión. Tras aplicar el Análisis de la Varianza se concluyó que existen diferencias significativas en la media de la abstención provincial en las diferentes convocatorias, además de identificar tres grupos: convocatorias con abstención alta, intermedia y baja, siendo más alta en las convocatorias más recientes. En ambos trabajos se obtiene que hay diferencias significativas en la media de la abstención en las diferentes convocatorias, pero difieren en la identificación de grupos. Mientras que en este trabajo se identifican tres grupos, en el otro trabajo se obtienen dos. Una diferencia clave es que, en las elecciones al Parlamento Europeo las convocatorias de mayor abstención son las más recientes, mientras que en las elecciones generales este porcentaje va variando. Cabe destacar que el nivel de abstención en las elecciones generales es considerablemente menor que en las elecciones al Parlamento Europeo.

En segundo lugar se realizó un Análisis de Correspondencias Simples entre las convocatorias y las ideologías, donde se observó una evolución del voto partiendo del predominio de opciones moderadas, como izquierda, derecha, centro y nacionalistas en las primeras elecciones, hacia una mayor fragmentación y polarización en los últimos años. De tal manera,

que las convocatorias más antiguas se asocian principalmente al eje izquierda-derecha, mientras que las convocatorias más recientes están ligadas a posiciones extremas.

El tercer análisis realizado fue un Análisis de Correspondencias Simples entre provincias e ideologías para cuatro convocatorias, en el cual se concluyó que el bloque nacionalista ha permanecido a lo largo de los años muy estable, aunque en las últimas elecciones algunas provincias catalanas tienen un cierto acercamiento a ideologías de extrema derecha. En las provincias del sur tiene un gran predominio la izquierda, aunque en los últimos años muchas de estas provincias han experimentado una transición hacia bloques de extrema izquierda. Mientras que la derecha se asienta provincias de Castilla y León y Galicia. Por último, se ha observado una consolidación de la extrema derecha en gran parte de España, además de un declive de la extrema izquierda tras su auge en 2019.

El último análisis realizado permitió contrastar muchas de las ideas obtenidas con anteriores análisis, y es que se observa un tránsito de la distribución del voto, partiendo de una menor polarización y del predominio de opciones moderadas (principalmente la izquierda y la derecha), durante los años 1987-2009, hasta el creciente auge de bloques extremos y la fragmentación del voto en los últimos años (2014-2024).

Un posible trabajo futuro podría ser comparar de forma más detallada los resultados de las elecciones al Parlamento Europeo y las elecciones generales en España. Esto permitiría observar si las tendencias que se han identificado en este trabajo, como la polarización, fragmentación del voto, los niveles de abstención, o la evolución ideológica de las provincias, se producen de igual forma en las elecciones generales. Se podría explorar si estas tendencias son igual de marcadas en ambas elecciones.

Otra propuesta es incorporar a los datos la renta media de cada provincia, permitiendo analizar si existe una relación entre el nivel de renta y el voto a determinadas ideologías, además de observar si periodos de mayor renta se asocian a una mayor estabilidad ideológica en las provincias.

Bibliografía

[1] El Español. *Uno en Madrid vale lo mismo que uno en Zamora*. Accesible: https://www.elespanol.com/castilla-y-leon/region/20240525/madrid-vale-mismo-zamora-voto-util-europeas/857414667_0.html. Accedido el 10/02/2025.

[2] European data journalism. *¿Qué pasa con el abstencionismo en España?*. Accesible: https://www.europeandatajournalism.eu/es/cp_data_news/que-pasa-con-el-abstencionismo-en-europa/. Accedido el 10/02/2025.

[3] Elecciones generales de España de 2023. Accesible: https://es.wikipedia.org/wiki/Elecciones_generales_de_Espa%C3%B1a_de_2023. Accedido el 10/02/2025.

[4] RTVE. *Elecciones europeas: Cuando la mayoría decide no votar*. Accesible: https://www.rtve.es/noticias/20240311/interesan-elecciones-europeas-paises-nunca-mitad-poblacion/16008767.shtml. Accedido el 10/02/2025.

[5] Fernández Temprano, M. A. (2023). *Tema 2. Análisis de Correspondencias Simples*. Apuntes de la asignatura de Análisis de Datos. Universidad de Valladolid.

[6] Greenacre, M. (2008). *La práctica del análisis de correspondencias*. Fundación BBVA. Accesible: https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2008_practica_analisis_correspondencias.pdf. Accedido el 04/04/2025.

[7] Fernández Martín, I. (2023). *Tema 2. El modelo de regresión simple*. Apuntes de la asignatura Regresión y Anova. Universidad de Valladolid.

[8] Fernández Martín, I. (2023). *Tema 5. El modelo de regresión lineal en notación matricial*. Apuntes de la asignatura Regresión y Anova. Universidad de Valladolid.

[9] Fernández Martín, I. (2023). *Tema 4. Validación de los modelos de regresión lineal simple*. Apuntes de la asignatura Regresión y Anova. Universidad de Valladolid.

[10] Fernández Martín, I. (2023). *Tema 8. Modelo ANOVA con un factor*. Apuntes de la asignatura Regresión y Anova. Universidad de Valladolid.

[11] Fernández Martín, I. (2023). *Tema 9. ANOVA con un factor. Inferencias sobre los niveles del factor*. Apuntes de la asignatura Regresión y Anova. Universidad de Valladolid.

[12] Área de descarga de datos. Accesible: https://infoelectoral.interior.gob.es/es/elecciones-celebradas/area-de-descargas/. Accedido el 17/02/2025.

[13] Documentación de R. *merge*. Accesible: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/merge. Accedido el 18/02/2025.

[14] Documentación de R. *leftJoin*. Accesible: https://www.rdocumentation.org/

packages/mStats/versions/3.2.2/topics/leftJoin. Accedido el 20/02/2025.

[15] Elecciones al Parlamento Europeo de 2024 (España). Accesible: https://es.wikipedia.org/wiki/Elecciones_al_Parlamento_Europeo_de_2024_(Espa%C3%B1a). Accedido el 17/02/2025.

[16] Documentación de R. *rbindlist*. Accesible: https://www.rdocumentation.org/packages/data.table/versions/1.15.0/topics/rbindlist. Accedido el 20/02/2025.

[17] Ministerio de Agricultura, Pesca y Alimentación. *Clasificación de los municipios en función del censo*. Accesible: https://www.mapa.gob.es/es/ministerio/servicios/analisis-y-prospectiva/ayp_demografiaenlapoblacionrural2020_tcm30-583987.pdf. Accedido el 20/03/2025.

[18] Documentación de R. *duncan.test*. Accesible: https://www.rdocumentation.org/packages/agricolae/versions/1.3-7/topics/duncan.test. Accedido el 27/05/2025.

[19] Partido X. Accesible: https://es.wikipedia.org/wiki/Partido_X. Accedido el 09/05/2025.

[20] Movimiento RED. Accesible: https://www.lasexta.com/programas/al-rojo-vivo/entrevistas/elpidio-silva-movimiento-red-poner-marcha-segunda-transicion-politica_20140404572685614beb28d4460217c3.html. Accedido el 09/05/2025.

[21] Escaños en Blanco. Accesible: https://es.wikipedia.org/wiki/Esca%C3%B1os_en_Blanco. Accedido el 09/05/2025.

[22] Martín Gómez, P. *Estudio de la abstención en los resultados de las elecciones al Congreso de los Diputados en España* . Accesible: https://uvadoc.uva.es/handle/10324/74299. Accedido el 13/02/2025.

[23] Documentación de R. *ggplot2*. Accesible: https://www.rdocumentation.org/packages/ggplot2/versions/0.9.0/topics/ggplot. Accedido el 23/04/2025.

[24] Documentación de R. *data.table*. Accesible: https://www.rdocumentation.org/packages/data.table/versions/1.17.2. Accedido el 17/02/2025.

[25] Instituto Nacional de Estadística (INE). *Códigos provincias INE*. Accesible: https://www.ine.es/daco/daco42/codmun/cod_provincia.htm. Accedido el 05/04/2025.

[26] Análisis de Correspondencias Simples en R. Accesible: https://rpubs.com/Cendlozg/933519. Accedido el 04/04/2025.

[27] Análisis de Correspondencias Múltiples en R. Accesible: https://rpubs.com/gustavomtzv/1042047. Accedido el 07/04/2025.

Apéndice A

Coaliciones políticas

Se muestra una tabla con las distintas coaliciones agrupadas a lo largo de las 9 convocatorias.

Año	**Coalición**	**Partidos Integrantes**
2024	Nacionalista de izquierdas	ERC, EH-Bildu, BNG, ARA MÉS, ERPV
	Nacionalista de derechas	EAJ-PNV, CCa, GBAI, EL PI, CEUS
2019	Nacionalista de izquierdas	ERC, EH, BNG, ANDECHA ASTUR, PUYALON, ANC, AC-AR, ARA-REPUBLIS, AHORA-REPUBLICAS
	Nacionalista de derechas	EAJ-PNV, CCa, GBAI, EL PI, CxG, CEUS, DV, Compromiso por Galicia
2014	Nacionalista de izquierdas	EH, BNG, AGE, PUYALÓN, ANC, LPD, ANDECHA ASTUR
	Nacionalista de derechas	EAJ-PNV, CCa, CiU, CEU, CxG
2009	Nacionalista de izquierdas	ERC, BNG, ARALAR, ESQUERRA, CHA, EUROPA DE LOS PUEBLOS VERDES
	Nacionalista de derechas	EAJ-PNV, BLOC, CCa, CiU, CEU, PA, UM-UMe
2004	Galeusca-Pueblos de Europa	CIU, BNG, BLOC, PSM, GALEUSCA, EAJ-PNV
	Nacionalista de derechas	PAR, CC, UM, PA, EU, CEU, CHA, UV, PAS
	Europa de los Pueblos	ERC

Año	**Coalición**	**Partidos Integrantes**
1999	Nacionalista de derechas	UV, PAR, CC, PA, Coalición europea
	Nacionalista de izquierdas	EAJ-PNV, Coalición nacionalista
1994	Por la Europa de los pueblos	PEP, ERC, EA, TC-PNC
	Coalición nacionalista	UV, PAR, UM, CG, EAJ, CC
1989	Coalición Nacionalista	EAJ-PNV, AIX, CG, CN, PANCAL
	Izquierda de los pueblos	EE, PSM, UA-CHA, PSG, UPV, AC, IP
	Por la Europa de los pueblos	EA, ERC, PNG, PEP
	Federación partidos regionales	UC, PRC, EU, PRP, UPM, PRM
1987	Europa de los pueblos	EA, ERC, PNG
	Izquierda de los pueblos	EE, CIP, PSM-EN, CIP-ENE, PSG-EG
	Unión europeísta	PNV

Tabla A.1: *Coaliciones políticas en las elecciones europeas*

Apéndice B

Clasificación de candidaturas

Ahora se presenta la clasificación de cada una de las candidaturas finales según las ideologías descritas en el tercer capítulo.

CANDIDATURA	**IDEOLOGÍA**
Federación de partidos de Alianza Popular	Derecha
Partido Socialista Obrero Español	Izquierda
Partido Popular	Derecha
Centro Democrático y Social	Centro
Izquierda Unida	Izquierda
Herri Batasuna	Nacionalista
Por la Europa de los Pueblos	Nacionalista
Izquierda de los Pueblos	Izquierda
Partido Andalucista	Nacionalista
Partido Trabajadores de España-Unidad Comunista	Extrema izquierda
Unión Europeísta	Derecha
Unio Valenciana	Centro
Partido Aragonés Regionalista	Centro
Partido Demócrata Popular	Derecha
Acción Social	Derecha
Frente Nacional	Extrema derecha
Los Verdes	Extrema izquierda
Coalición Agrupaciones Independientes de Canarias	Nacionalista
Coalición Electoral Convergencia I Unio	Nacionalista
Agrupación de Electores Jose María Ruiz Mateos	Derecha
Coalición Nacionalista	Nacionalista

CANDIDATURA	**IDEOLOGÍA**
Los Verdes Ecologistas	Extrema izquierda
Federación de Partidos Regionales	Derecha
Partido Comunista de los Pueblos de España	Extrema izquierda
Coalición Foro y Centro Democrático y Social	Centro
Coalición Andalucista Poder Andaluz	Nacionalista
Bloque Nacionalista Gallego	Nacionalista
Grupo Verde	Extrema izquierda
Coalición Nacionalista de Izquierdas	Nacionalista
Coalición Nacionalista de Derechas	Nacionalista
Euskal Herritarrok	Nacionalista
Galeusca - Pueblos de Europa	Nacionalista
Europa de los Pueblos	Nacionalista
La Izquierda	Izquierda
Unión Progreso y Democracia	Centro
Iniciativa Internacionalista – La Solidaridad entre los Pueblos	Izquierda
Partido X, Partido del Futuro	Otros
Izquierda Plural	Izquierda
VOX	Extrema derecha
Partido Animalista Contra el Maltrato Animal	Otros
Podemos	Extrema izquierda
La Izquierda por el Derecho a Decidir	Izquierda
Ciudadanos	Centro
Escaños en Blanco	Otros
Movimiento de Renovación Democrática Ciudadana, Movimiento Red	Otros
Compromiso por Europa	Izquierda
Junts Per Catalunya-Lliures Per Europa	Nacionalista
SUMAR	Extrema izquierda
Agrupación de Electores “Se Acabó La Fiesta”	Extrema derecha

Tabla B.1: *Clasificación ideológica de candidaturas*

Apéndice C

Código R

C.1. Lectura de datos

```
#Fichero de candidaturas
datosC1989<-read.table("03078906.dat", header=FALSE, sep="\t",
    colClasses="character",fileEncoding = 'latin1',quote="")
#Fichero con datos de candidaturas municipales
datosM1989<-read.table("06078906.dat", header=FALSE, sep="\t",
    colClasses="character")
#Fichero con datos globales de ambito municipal
datosG1989<-read.table("05078906.dat",header=FALSE,sep="\t",
fileEncoding = 'latin1',quote="")
```

C.2. Tratamiento de las convocatorias

```
#Longitud fichero candidaturas
longC <- c(2,4,2,6,50,150,6,6,6)
#Longitud fichero candidaturas municipales
longM<-c(2,4,2,1,2,3,2,6,8,3)
#Longitud fichero datos globales
longG<-c(2,4,2,1,2,2,3,2,100,1,3,3,3,8,5,8,8,8,8,8,8,8,8,8,3,8,8,1)
split_by_size<-function(texto, sizes) {
  #Posiciones iniciales
  start<-cumsum(c(1, sizes[-length(sizes)]))
  #Posiciones finales
  end<-cumsum(sizes)
  mapply(function(i, j) substr(texto, i, j), start, end)
}
#Candidaturas
datC2024<-data.frame(t(apply(datosC2024, 1, split_by_size, sizes =
    longC)))
#Datos candidaturas municipales
```

```r
datM2024<-data.frame(t(apply(datosM2024, 1, split_by_size, sizes = longM)))
#Datos globales
datG2024<-data.frame(t(apply(datosG2024, 1, split_by_size, sizes = longG)))
#Pongo el nombre a las columnas
colnames(datC2024)<-c("T.Eleccion", "Agno", "Mes", "C.Candidatura", "Siglas", "Denominacion","C.Provincial", "C.Autonomico", "C.Nacional")
colnames(datM2024)<-c("T.Eleccion", "Agno", "Mes", "Num_vuelta", "C.Provincia", "C.Municipio","Num_distrito", "C.Candidatura", "Votos", "Num_candidatos")
colnames(datG2024) <- c("T.Eleccion", "Agno", "Mes", "Num_vuelta", "CCAA", "C.Provincia", "C.Municipio","Num_distrito", "Nombre_municipio", "C.Distrito", "PJ", "DP", "C.Comarca", "Poblacion_derecho",
"Num_mesas", "Censo", "Censo_escrutinio", "CERE", "Total_CERE", "Vot1", "Vot2","Vot_blanco", "Vot_nulo", "Vot_candidaturas", "Num_escagnos", "VAR", "VNR", "Oficial")
#Convierto las columnas a tipo numerico
datM2024$Votos <- as.numeric(datM2024$Votos)
datG2024$Censo <- as.numeric(datG2024$Censo)
datG2024$CERE <- as.numeric(datG2024$CERE)
datG2024$Total_CERE <- as.numeric(datG2024$Total_CERE)
datG2024$Vot_blanco <- as.numeric(datG2024$Vot_blanco)
datG2024$Vot_nulo <- as.numeric(datG2024$Vot_nulo)
datG2024$Vot_candidaturas <- as.numeric(datG2024$Vot_candidaturas)
#Unir datC y datM con un merge
datM2024 <- merge(datM2024, datC2024)
tabla1 <- data.table(datG2024[, c("C.Provincia",
                                  "C.Municipio", "Num_distrito",
                                   "Censo", "Vot_blanco", "Vot_nulo",
                                      "Vot_candidaturas", "CERE",
                                      "Total_CERE")])
tabla2 <- data.table(datM2024)
#Me quedo con Num_distrito=99
tabla1 <- subset(tabla1, Num_distrito == "99")
tabla2 <- subset(tabla2, Num_distrito == "99")
#Union tipo left join: C.Provincia y C.Municipio
tabla2024 <- tabla2[tabla1, on = .(C.Provincia, C.Municipio)]
#Calculo de algunas variables
tabla2024 <- tabla2024[, Porcentaje := Votos / Censo]
#Para crear la variable Tamaño
niveles<-c(0, 5000, 30000, Inf)
tabla2024<-tabla2024[, Tamano := cut(Censo, breaks = niveles, labels = nombres, right= FALSE)]
# Eliminación de columnas redundantes
```

```
tabla2024<-subset(tabla2024, select = -c(i.Num_distrito))
tabla2024<-subset(tabla2024, select = -c(T.Eleccion, Mes, C.Provincial,
C.Autonomico, Num_distrito, C.Nacional, Num_vuelta))
```

C.3. Conjunto de datos final

```
datos_final<-datos_final[, .(Votos = sum(Votos),
                             Porcentaje = sum(Porcentaje)),
by = .(Agno, C_Provincia = `C.Provincia`,
C_Municipio = `C.Municipio`, Candidatura,
    Censo,Tamano,Abstencion,Vot_blanco,Vot_nulo)]

datos_final<-data.table::rbindlist(list(tabla1987, tabla1989,
    tabla1994, tabla1999, tabla2004, tabla2009,
tabla2014, tabla2019, tabla2024), use.names = TRUE, fill = TRUE)
datos_final[, Agno := factor(Agno)]
datos_final[, C.Provincia := factor(C.Provincia)]
datos_final[, Candidatura := factor(Candidatura)]
datos_final[, C.Municipio := factor(paste(C.Provincia, C.Municipio))]
datos_final[,c("CERE","Total_CERE","Siglas_Norm","Abstencion_ext"):=NULL]
abstencion_nulo2<-datos_final[, {
  total_votos_real<-sum(Votos[Candidatura != "ABSTENCION/NULO"])
  censo_val<-unique(Censo)
  abstencion<-censo_val - total_votos_real
  voto_blanco<-unique(Vot_blanco)
  voto_nulo<-unique(Vot_nulo)
  total_extra<-abstencion
  .(
    Candidatura = "ABSTENCION/NULO",
    Censo = censo_val,
    Tamano = unique(Tamano),
    Vot_blanco = voto_blanco,
    Vot_nulo = voto_nulo,
    Abstencion = abstencion,
    Votos = total_extra,
    Porcentaje = total_extra / censo_val
  )
}, by = .(Agno, C_Provincia, C_Municipio)]
#Unir con la tabla original
datos_final<-rbind(datos_final, abstencion_nulo2, fill = TRUE)
#Borro las columnas Abstencion Vot_blanco, Vot_nulo
datos_final[,c("Abstencion","Vot_blanco","Vot_nulo"):=NULL]
datos_final[,c("Porcentaje"):=NULL]
#DETECCION DE NA
colSums(is.na(datos_final))#no hay NA
```

```r
#Votos nulos
sum(datos_final$Votos<0)
#no hay nada con votos negativos
sum(datos_final$Censo<0)
#Porcentaje mayor de 1?
sum(datos_final$Porcentaje>1)#Hay una fila
datos_final[Porcentaje>1]
#Se elimina
datos_final<-datos_final[Porcentaje<=1]

#Creación del conjunto de datos abstencion municipal
datos_final<-as.data.table(datos_final)
#Obtengo ahora el numero de votos totales por municipio
abstencion<-datos_final[,.(
  Censo=unique(Censo),
  Tamano=unique(Tamano),
  Votos_totales=sum(Votos,na.rm=TRUE)
),by=.(Agno,C_Provincia,C_Municipio)]
#Una vez obtenidos los votos totales obtengo la abstencion en cada
    municipio,
#en cada convocatoria
abstencion[,Abstencion:=Censo-Votos_totales]
abstencion[,Porcentaje_abstencion:=Abstencion/Censo]
#Posibles errores
#Abstencion negativa
abstencion[Porcentaje_abstencion<0]
#Se eliminan las 6 observaciones
abstencion<-abstencion[Porcentaje_abstencion>=0]
abstencion_municipal<-abstencion

#Creación del conjunto de datos abstencion provincial
abstencion2<-datos_final[,.(
  Censo=unique(Censo),
  Votos_totales=sum(Votos,na.rm=TRUE)
),by=.(Agno,C_Provincia,C_Municipio)]
abstencion2
abstencion3<-abstencion2[,.(
  Censo=sum(Censo),
  Votos_totales=sum(Votos_totales,na.rm=TRUE)
),by=.(Agno,C_Provincia)]
abstencion3[,Abstencion:=Censo-Votos_totales]
abstencion3[,Porcentaje_abstencion:=Abstencion/Censo]
#Posibles errores
#Abstencion negativa
abstencion3[Porcentaje_abstencion<0]
#ninguna provincia tiene este error
abstencion3[Porcentaje_abstencion>1]
```

```
#tampoco se tiene alguna provincia con este error
abstencion_provincial<-abstencion3
```

C.4. Diagramas de cajas abstención

```
abstencion_municipal<-as.data.table(abstencion_municipal)
#Diagrama de cajas segun tamaño del municipio
ggplot(abstencion_municipal,aes(x=Tamano,y=Porcentaje_abstencion))+
  geom_boxplot(fill="lightblue")+
  labs(title="Diagrama de cajas",
       subtitle="Porcentaje de abstención vs Tamaño municipio",
       x="Tamaño municipio",y="Porcentaje abstención")
#Diagrama de cajas segun cada convocatoria
ggplot(abstencion_municipal,aes(x=Agno,y=Porcentaje_abstencion))+
  geom_boxplot(fill="lightblue")+
  labs(title="Porcentaje de abstención vs Convocatorias",
       subtitle="Abstención municipal",
       x="Convocatorias",y="Porcentaje de abstención")
abstencion_provincial<-as.data.table(abstencion_provincial)
#Diagrama de cajas segun cada convocatoria (provincial)
ggplot(abstencion_provincial,aes(x=Agno,y=Porcentaje_abstencion))+
  geom_boxplot(fill="lightgreen")+
  labs(title="Porcentaje de abstención vs Convocatorias",
       subtitle="Abstención provincial",
       x="Convocatorias",y="Porcentaje de abstención")
```

C.5. Modelos de regresión para la abstención

```
m1<-lm(Porcentaje_abstencion~log(Censo),data=abstencion_provincial)
summary(m1);anova(m1)
nuevo<-data.frame(Censo=exp(seq(10,16,length.out=100)))
pred1<-predict(m1,newdata=nuevo,interval="prediction")
nuevo$fit<-pred1[,'fit']
nuevo$infe<-pred1[,"lwr"]
nuevo$supe<-pred1[,"upr"]
p1<-predict(m1,interval="prediction")
outliers<-(abstencion_provincial$Porcentaje_abstencion<p1[,'lwr']) |
  (abstencion_provincial$Porcentaje_abstencion>p1[,"upr"])
ggplot(abstencion_provincial,
   aes(x=log(Censo),y=Porcentaje_abstencion)) +
  geom_point(alpha = 0.5) +
  geom_ribbon(data = nuevo, aes(x = log(Censo), y = fit, ymin = infe,
     ymax = supe), fill = "lightblue", alpha = 0.3)+
```

```
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_x_continuous(
    limits = c(10, 16),
    breaks = 10:16
  ) +
  scale_y_continuous(
    limits = c(0, 0.8),
    breaks = seq(0, 0.6, 0.1)
  ) +
  labs(
    x = "log(Censo)",
    y = "Abstenciones",
    title = "Regresión Abstenciones ~ log(Censo)"
  )
#Elimino Ceuta y Melilla
eliminados1<-abstencion_provincial[C_Provincia!='51']
eliminados1<-eliminados1[C_Provincia!='52']
#Modelo sin outliers
m1b<-lm(Porcentaje_abstencion~log(Censo),data=eliminados1)
summary(m1b);anova(m1b)
plot(m1b$fitted.values,residuals(m1b))
abline(h=0)
mean(residuals(m1b))
qqnorm(residuals(m1b))
qqline(residuals(m1b),col="blue")
plot(residuals(m1b), type = "o", main = "Residuos vs. Orden",
     ylab = "Residuos", xlab = " ndice  del dato")
abline(h = 0, col = "red")
ggplot(eliminados1, aes(x=log(Censo),y=Porcentaje_abstencion)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_x_continuous(
    limits = c(11, 16),
    breaks = 10:16
  ) +
  scale_y_continuous(
    limits = c(0, 0.8),
    breaks = seq(0, 0.6, 0.1)
  ) +
  labs(
    x = "log(Censo)",
    y = "Abstenciones",
    title = "Regresión Abstenciones ~ log(Censo)"
  )
pred1<-predict(m1s,newdata=nuevo,interval="prediction")
nuevo$fit<-pred1[,'fit']
nuevo$infe<-pred1[,"lwr"]
```

```r
nuevo$supe<-pred1[,"upr"]
p1<-predict(m1s,interval="prediction")
outliers<-(eliminados1$Porcentaje_abstencion<p1[,'lwr']) |
  (eliminados1$Porcentaje_abstencion>p1[,"upr"])
ggplot(eliminados1, aes(x=log(Censo),y=Porcentaje_abstencion)) +
  geom_point(alpha = 0.5) +
  geom_ribbon(data = nuevo, aes(x = log(Censo), y = fit, ymin = infe,
      ymax = supe), fill = "lightblue", alpha = 0.3)+
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_x_continuous(
    limits = c(11, 16),
    breaks = 11:16
  ) +
  scale_y_continuous(
    limits = c(0, 0.8),
    breaks = seq(0, 0.6, 0.1)
  ) +
  labs(
    x = "log(Censo)",
    y = "Abstencion (%)",
    title = "Abstencion ~ log(Censo) (sin outliers)"
  )

#Segundo modelo sin outliers
eliminados2<-abstencion_provincial[C_Provincia!='51']
eliminados2<-eliminados2[C_Provincia!='52']
m2s<-lm(Porcentaje_abstencion~log(Censo)*Agno,data=eliminados2)
summary(m2s);anova(m2s)
ggplot(eliminados2, aes(x = log(Censo), y = Porcentaje_abstencion,
    color = Agno)) +
  geom_point(alpha = 0.4, size = 1) +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x) +
  labs(
    title = "Rectas de regresión por año",
    x = "log(Censo)",
    y = "Porcentaje de abstención"
  ) +
  theme_minimal() +
  theme(legend.position = "right")
#homogeneidad varianza
plot(m2s$fitted.values,residuals(m2s))
abline(h=0)
mean(residuals(m2s))
qqnorm(residuals(m2s))
qqline(residuals(m2s),col="blue")
plot(residuals(m2s), type = "o", main = "Residuos vs. Orden",
     ylab = "Residuos", xlab = " ndice  del dato")
```

```r
abline(h = 0, col = "red")
#tercer modelo sin outliers
eliminados3<-abstencion_provincial[C_Provincia!='51']
eliminados3<-eliminados3[C_Provincia!='52']
m3b<-lm(Porcentaje_abstencion~log(Censo)+Agno,data=eliminados3)
anova(m3b);summary(m3b)
eliminados3$predicciones<-predict(m3b)
ggplot(eliminados3, aes(x = log(Censo), y = Porcentaje_abstencion,
    color = Agno)) +
  geom_point(alpha = 0.6, size = 1.5) +
  geom_line(aes(y = predicciones)) +
  labs(
    title = "Abstencion~log(censo)+Agno",
    x = "log(Censo)",
    y = "Abstencion (%)"
  ) +
  scale_color_brewer(palette = "Set1") +
  theme_minimal() +
  theme(legend.position = "right")
#homogeneidad varianza
plot(m3b$fitted.values,residuals(m3b))
abline(h=0)
mean(residuals(m3b))
qqnorm(residuals(m3b))
qqline(residuals(m3b),col="green",lwd=2)
plot(residuals(m3b), type = "o", main = "Residuos vs. Orden",
     ylab = "Residuos", xlab = "Indice del dato")
abline(h = 0, col = "red")
#Gráfico de los residuos estudentizados del modelo final
eliminados3$residuos<-residuals(m3b)
eliminados3$ajustados<-fitted(m3b)
eliminados3$residuos2<-rstudent(m3b)
colores_oscuros <- c("#1B9E77", "#D95F02", "#7570B3", "#E7298A",
  "#66A61E", "#E6AB02", "#A6761D", "#666666", "darkblue"
)
ggplot(eliminados3,aes(x=ajustados, y=residuos2,color=Agno))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = 0, color = "black", linetype = "dashed") +
  scale_color_manual(values = colores_oscuros) +
  labs(title="Gráfico de los residuos estudentizados",
       x="Valores ajustados",y="Residuos estudentizados") +
  theme_minimal(base_size = 13) +
  theme(
    legend.position = "right",
    legend.title = element_text(face = "bold"),
    plot.title = element_text(face = "bold", size = 14))
```

C.6. ANOVA de la abstención con convocatoria como factor

```r
modelo<-aov(Porcentaje_abstencion~Agno,data=abstencion_provincial)
summary(modelo)#es significativo
#HOMOGENEIDAD DE LA VARIANZA
plot(modelo$fitted.values,modelo$residuals,main="Residuos vs Valores predichos",
     xlab="Valores predichos",ylab="Residuos")
#Extraer residuos y valores ajustados
residuos<-modelo$residuals
ajustados<-modelo$fitted.values
provincias<-abstencion_provincial$C_Provincia
#Crear conjunto con provincias
datos_residuos <- data.frame(
  C_Provincia = provincias,
  residuo = residuos,
  ajustado = ajustados
)
# Número de observaciones con mayores residuos a etiquetar
n<-15
top_outliers<-head(datos_residuos[order(-abs(datos_residuos$residuo)), ], n)
ggplot(datos_residuos, aes(x = ajustado, y = residuo)) +
  geom_point(alpha = 0.5) +
  geom_text_repel(data = top_outliers, aes(label = C_Provincia),
                  color = "blue", max.overlaps = Inf) +
  labs(x = "Valores predichos", y = "Residuos",
       title ="Residuos vs Valores predichos") +
  theme_minimal()
#Se eliminan Ceuta y Melilla
sin<-abstencion_provincial[C_Provincia!='51',]
sin<-sin[C_Provincia!='52',]
modelo2<-aov(Porcentaje_abstencion~Agno,data=sin)
summary(modelo2)
ggplot(sin, aes(x = modelo2$fitted.values, y = modelo2$residuals)) +
  geom_point(alpha = 0.5) +
  labs(x = "Valores predichos", y = "Residuos",
       title ="Residuos vs Valores predichos") +
  theme_minimal()
#NORMALIDAD DE LOS RESIDUOS
hist(residuals(modelo2),main="Histograma de los residuos",col="lightblue",
     xlab="Residuos",probability = TRUE,ylab="Densidad")
curve(dnorm(x,0,sd(residuals(modelo2))),add=TRUE,col="blue",lwd=2)
#qqplot de los residuos
```

```r
qqnorm(residuals(modelo2),xlab="Cuantiles teóricos",ylab="Cuantiles muestrales",main="Q-Q Plot")
qqline(residuals(modelo2),col="red",lwd=2)
#Independencia
residuos<-residuals(modelo2)
orden<-1:nrow(sin)
plot(orden, residuos,  main = "Residuos vs Orden de Observación",
     xlab = "Orden de recopilación de datos", ylab = "Residuos",pch = 19,
     col = "blue")
residuos<-residuals(modelo2)
orden<-seq_along(residuos)
residuos_df<-data.frame(orden = orden, residuos = residuos)
ggplot(residuos_df, aes(x = orden, y = residuos)) +
  geom_point(color = 'blue', size = 1.5, alpha = 0.7) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Plot secuencial",
    x = "Orden de recopilación de datos",
    y = "Residuos"
  ) +
  theme_minimal()
abline(h = 0, col = "red", lwd = 2)
#Test de Tukey
TukeyHSD(modelo2)
plot(TukeyHSD(modelo2), las = 1, cex.axis = 0.8)
#Test de Duncan
duncan_result<-duncan.test(modelo2, "Agno", console = TRUE)
```

C.7. ACS entre convocatorias e ideologías

```r
datos2<-datos_final
datos2<-datos2[,num_total:=sum(Votos),by=.(Candidatura,Agno)]
datos2<-unique(datos2)
datos3<-datos_final
datos3<-unique(datos3[,.(Agno,C_Municipio,Censo)],
               by=c("Agno","C_Municipio"))
datos3[,total_censo:=sum(Censo),by=.(Agno)]#aqui tengo el censo total
#Añado la columna total_censo calculada anteriormente
unido<-left_join(datos2, datos3[, c("Agno", "C_Municipio", "total_censo")],
                 by = c("Agno", "C_Municipio"))
#Creo para cada año una columna que indica la suma de abstencion+nulo+blanco
result<- unido %>%
```

```
  filter(Candidatura == "ABSTENCION/NULO") %>%
  group_by(Agno) %>%
  summarise(total_abstencion = sum(Votos))
#Lo añado como columna a los datos
datos2_con_abstencion <- left_join(unido, result, by = "Agno")
datos_a<-datos2_con_abstencion
datos_a[, total_abstencion := ifelse(is.na(total_abstencion),
   total_abstencion[!is.na(total_abstencion)][1], total_abstencion), by
   = C_Municipio]
datos_a[,Participan:=total_censo-total_abstencion]
result_chi<-datos_a[, .(total_votos = unique(num_total),
                        total_abstencion = unique(total_abstencion),
                        total_censo = unique(total_censo),
                        total_participan = unique(Participan)),
                    by = .(Agno, Candidatura)]
#Ahora quito de result las filas que tenga Abstencion
result_chi<-result_chi[1:56,]
#Elimino columnas innecesarias
result_chi[,c("total_abstencion","total_censo","total_participan"):=NULL]
matriz_chi<-dcast(result_chi,Candidatura~Agno,value.var
   ="total_votos",fill=0)
#Realizo el analisis
matriz_chi_numerica<-as.matrix(matriz_chi[,-1,with=FALSE]
fit<-CA(matriz_chi_numerica,graph = FALSE)
summary(fit)
fviz_screeplot(fit,addlabels = TRUE)+
ggtitle("Scree plot")+xlab("Ejes")+ylab("%")
fviz_ca_biplot(fit,repel=TRUE,title="Ideologías-Convocatoria",
               col.col="blue",col.row="red")
fviz_ca_biplot(fit,axes=c(2,3),repel=TRUE,title="Ideologías-Convocatoria"
               ,col.col="blue",col.row="red")
```

C.8. ACS entre ideologías y provincias por convocatoria

Se presenta el código únicamente para una convocatoria, en concreto la de 2024.

```
datos2<-datos_final
#Obtengo el numero de votos de cada candidatura/ideologia por cada
   provincia
datos2<-datos2[,num_total:=sum(Votos),by=.(Candidatura,C_Provincia,Agno)]
datos2<-unique(datos2)#hago que salga una vez cada uno de ellos solo
#Elimino de datos2 la provincia y tamano,
datos2<-datos2[,c("C_Municipio","Tamano","Porcentaje"):=NULL]
datos3<-datos_final
```

```r
#Quedarse con una fila por municipio por año
datos3<- unique(datos3[, .(Agno, C_Provincia, C_Municipio, Censo)])
#Ahora sí, sumar el censo por provincia y año
datos3<- datos3[, .(total_censo = sum(Censo)), by = .(Agno,
   C_Provincia)]
unido<-left_join(datos2, datos3[, c("Agno",
   "total_censo","C_Provincia")],
                 by = c("Agno", "C_Provincia"))
result<- unido %>%
  filter(Candidatura == "ABSTENCION/NULO") %>%
  group_by(Agno,C_Provincia) %>%
  summarise(total_abstencion = sum(Votos))
datos2_con_abstencion <- left_join(unido, result, by =
   c("Agno","C_Provincia"))
datos_a<-datos2_con_abstencion
datos_a[, total_abstencion := ifelse(is.na(total_abstencion),
   total_abstencion[!is.na(total_abstencion)][1], total_abstencion), by
   = C_Provincia]
datos_a[,Participan:=total_censo-total_abstencion]
#obtengo el porcentaje de votos sobre los que si participan para que de
   1
result <- datos_a[, .(total_votos = unique(num_total),
                      total_abstencion = unique(total_abstencion),
                      total_censo = unique(total_censo),
                      total_participan = unique(Participan)),
                  by = .(Agno, Candidatura,C_Provincia)]
result[,Porc_recalculado:=total_votos/total_participan]
#Ahora quito de result las filas que tenga Abstencion/
result<-result[1:(.N - 468)]
#Elimino columnas innecesarias
result<-result[,c("total_abstencion","total_censo",
                  "total_participan"):=NULL]
rownames(matriz_ca_numerica)<-matriz_ca$Candidatura
#CONVOCATORIA 2024
result2024<-result[Agno == "2024"]
matriz_ca_2024<-dcast(result2024, Candidatura ~ C_Provincia, value.var
   = "Porc_recalculado", fill = 0)
matriz_ca_2024_n<-as.matrix(matriz_ca_2024[, -1, with = FALSE])
rownames(matriz_ca_2024_n)<-matriz_ca_2024$Candidatura
total2024<-dcast(result2024,Candidatura~C_Provincia,
                 value.var="total_votos")
total2024n<-as.matrix(total2024[,-1,with=FALSE])
rownames(total2024n)<-total2024$Candidatura
grados<-(nrow(total2024n)-1)*(ncol(total2024n)-1);grados
chisq.test(total2024n)
#Realizo el analisis
fit2024 <- CA(total2024n, graph = FALSE)
```

```
fviz_screeplot(fit2024, addlabels = TRUE) +
  ggtitle("Scree plot 2024") +
  xlab("Ejes") +
  ylab("%")
fviz_ca_biplot(fit2024, repel = TRUE,
               title = "Provincias-Ideologías 2024",
               col.col = "blue",
               col.row = "red")
fviz_ca_biplot(fit2024, repel = TRUE, axes=c(2,3),
               title = "Provincias-Ideologías 2024",
               col.col = "blue",
               col.row = "red")
```

C.9. ACM entre provincias, ideologías y convocatorias

```
datos2<-datos_final
#Obtengo el numero de votos de cada candidatura/ideologia por cada
   provincia
datos2<-datos2[,num_total:=sum(Votos),by=.(Candidatura,C_Provincia,Agno)]
datos2<-unique(datos2)#hago que salga una vez cada uno de ellos solo
#Elimino de datos2 la provincia y tamano,
datos2<-datos2[,c("C_Municipio","Tamano","Porcentaje"):=NULL]
datos3<-datos_final
datos3<- unique(datos3[, .(Agno, C_Provincia, C_Municipio, Censo)])
#Ahora sí, sumar el censo por provincia y año
datos3<- datos3[, .(total_censo = sum(Censo)), by = .(Agno,
   C_Provincia)]
unido<-left_join(datos2, datos3[, c("Agno",
   "total_censo","C_Provincia")],
                 by = c("Agno", "C_Provincia"))
result<- unido %>%
  filter(Candidatura == "ABSTENCION/NULO") %>%
  group_by(Agno,C_Provincia) %>%
  summarise(total_abstencion = sum(Votos))
datos2_con_abstencion <- left_join(unido, result, by =
   c("Agno","C_Provincia"))

datos_a<-datos2_con_abstencion
datos_a[, total_abstencion := ifelse(is.na(total_abstencion),
   total_abstencion[!is.na(total_abstencion)][1], total_abstencion), by
   = C_Provincia]
datos_a[,Participan:=total_censo-total_abstencion]
result<-datos_a[, .(total_votos = unique(num_total),
                    total_abstencion = unique(total_abstencion),
                    total_censo = unique(total_censo),
```

```
                        total_participan = unique(Participan)),
                    by = .(Agno, Candidatura,C_Provincia)]

result[,Porc_recalculado:=total_votos/total_participan]
result<-result[1:(.N - 468)]
result<-result[,c("total_abstencion","total_censo",
                      "total_participan"):=NULL]
#Realizo el analisis
fit<-MCA(result[, c("Agno", "Candidatura", "C_Provincia")],
           row.w = result$total_votos,
           graph = FALSE)

fviz_screeplot(fit,addlabels = TRUE)+ggtitle("Scree
   plot")+xlab("Ejes")+ylab("%")

#PRIMERA Y SEGUNDA DIMENSION
fviz_mca_biplot(fit, repel = TRUE, ggtheme = theme_minimal(),
                title="Ideologías-Convocatorias-Provincias",invisible
                   ="ind")

#SEGUNDA Y TERCERA DIMENSION
fviz_mca_biplot(fit, repel = TRUE, ggtheme = theme_minimal(),
                axes=c(2,3),title="Ideologías-Convocatorias-Provincias",
                invisible = "ind")
```