Trabajo Fin de Máster

Estimación en tablas de contingencia con marginales dadas

Sergio Prieto García

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación en táboas de continxencia con marxinais dadas
Título en español: Estimación en tablas de contingencia con marginales dadas
English title: Estimation in contingency tables with given marginals
Modalidad: Modalidad A
Autor/a: Sergio Prieto García, Universidade de A Coruña
Director/a: Ricardo José Cao Abad, Universidade de A Coruña;
Breve resumen del trabajo: El objetivo de este TFM es llevar a cabo una visión general de los métodos ya existentes de estimación en tablas de contingencia de las probabilidades conjuntas, proponer un algoritmo EM y finalmente aplicarlo en un problema de estimación de la probabilidad de transferencia de votos entre las distintas opciones políticas en dos elecciones consecutivas en España. En este caso, únicamente se disponen de las frecuencias de voto para cada opción política a nivel de los colegios electorales, siendo las probabilidades conjuntas y condicionales, en principio, desconocidas.

Don Ricardo José Cao Abad, Catedrático de Universidad de la Universidade de A Coruña, informa que el Trabajo Fin de Máster titulado

Estimación en tablas de contingencia con marginales dadas

fue realizado bajo su dirección por don Sergio Prieto García para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Redondela, a 05 de junio de 2023.

CAO ABAD RICARDO JOSE - DNI 32769018K

Firmado digitalmente por CAO ABAD RICARDO JOSE - DNI 32769018K
Fecha: 2023.06.05 15:31:56 +02'00'

El director:
Don Ricardo José Cao Abad

El autor:
Don Sergio Prieto García

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen

Resumen en español

La estimación de las probabilidades conjuntas en tablas de contingencia basándose únicamente en las frecuencias marginales representa un desafío ampliamente investigado. Bien es cierto, que en la actualidad, existen varios métodos que calculan las probabilidades de las celdas de una tabla de contingencia. Sin embargo, en estos, se deben cumplir con restricciones para las distribuciones marginales, pero, disponiéndose de las frecuencias de la tabla de contingencia.

El objetivo de este TFM es llevar a cabo una visión general de los métodos ya existentes, proponer un algoritmo EM y finalmente aplicarlo en un problema de estimación de la probabilidad de transferencia de votos entre las distintas opciones políticas en dos elecciones consecutivas en España. En este caso, únicamente se disponen de las frecuencias de voto para cada opción política a nivel de los colegios electorales, siendo las probabilidades conjuntas y condicionales, en principio, desconocidas. Por lo tanto, esta investigación aborda un problema mucho más desafiante y posiblemente sin precedentes, que los ya existentes, que es la estimación de las probabilidades de celda cuando solo se han observado las frecuencias marginales pero no las frecuencias de las propias celdas.

English abstract

Estimating cell probabilities in contingency tables based on marginal frequencies alone is a widely researched challenge. Admittedly, there are currently several methods that calculate the probabilities of the cells of a contingency table. However, in these, restrictions must be met for the marginal distributions, but, given the contingency table frequencies

The objective of this TFM is to provide an overview of the existing methods, propose an EM algorithm, and ultimately apply it to a problem of estimating the probability of vote transfer between different political options in two consecutive elections in Spain. In this case, only the vote frequencies for each political option at the level of electoral colleges are available, with the joint and conditional probabilities initially unknown. Therefore, this research tackles a much more challenging and possibly unprecedented problem than the existing ones, which is the estimation of cell probabilities when only the marginal frequencies have been observed but not the frequencies of the cells themselves.

Capítulo 1

Introducción

Este capítulo, aborda la presentación del problema, los objetivos de investigación y la justificación de la misma. Se destaca la importancia y relevancia del tema a investigar, y se establece el objetivo general y los objetivos específicos que guiarán el estudio. Además, se proporciona la justificación para llevar a cabo esta investigación, destacando las razones que respaldan su importancia.

1.1. Presentación del problema y objetivos de la investigación

Las tablas de contingencia son una de las principales herramientas empleadas en la estadística descriptiva e inferencial para analizar la relación entre dos o más variables categóricas. Estas tablas se presentan en forma de matriz, donde las filas representan una variable, las columnas representan otra, y los valores de las celdas indican la frecuencia conjunta o el conteo de combinaciones de categorías. Son ampliamente utilizadas en diversos campos de estudio, como la sociología, la psicología, la medicina, la biología, la economía, la investigación de mercados y la demoscopia, entre otros, debido a su sencillez en la visualización de los datos. Que se presenten en forma de matriz permite una representación clara y organizada de las relaciones entre las variables categóricas, facilitando la identificación de patrones, tendencias y asociaciones entre variables, lo que puede ayudar a los investigadores y profesionales a comprender y analizar los datos de manera eficiente.

El problema que ocupan en esta investigación las tablas de contingencia se encuentra directamente relacionado con la demoscopia. Según la RAE la demoscopia se define como el "estudio de las opiniones, aficiones y comportamiento humanos mediante sondeos de opinión". Es decir, la demoscopia es una disciplina que se centra en estudiar científicamente la opinión pública y la conducta electoral, utilizando métodos y técnicas estadísticas para recopilar, analizar e interpretar datos sobre las preferencias políticas, actitudes y comportamientos de la población en relación con los procesos políticos y electorales.

Este estudio se centrará en el proceso electoral, más concretamente, en el análisis, mediante el uso de tablas de contingencia, de dos elecciones sucedidas en espacios pequeños de tiempo, con el objetivo de tratar de estimar las probabilidades de transferencia de votos de unas elecciones a otras únicamente mediante el uso de las frecuencias marginales, que se corresponderían con el recuento de votos para cada una de las opciones políticas (partidos políticos) en cada una de las dos elecciones. Debido a que las elecciones generales en España se celebran cada cuatro años, se ha optado por trabajar con las primeras elecciones generales sucedidas en España en 2019 y las elecciones municipales sucedidas en distintos municipios españoles en ese mismo año, 2019.

Tabla 1.1: *Ejemplo de tabla de contingencia de doble entrada con frecuencias conjuntas y marginales*

ELECCIONES X \ ELECCIONES Y	Partido A	Partido B	Partido C	
Partido A	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Partido B	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
Partido C	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	

Para visualizar mejor la presentación del problema, propuesto en esta investigación, se empleará la Tabla 1.1 para mostrar el interés del estudio. Como se puede observar, la Tabla 1.1 se encuentra formada por dos variables categóricas "Elecciones X" y "Elecciones Y", teniendo cada una de ellas tres categorías "Partido A", "Partido B", "Partido C". Las frecuencias marginales de las "Elecciones X", $n_{1\cdot}$, $n_{2\cdot}$ y $n_{3\cdot}$, y las frecuencias marginales de las "Elecciones Y", $n_{\cdot 1}$, $n_{\cdot 2}$ y $n_{\cdot 3}$, son los únicos datos disponibles para tratar de estimar los distintos valores de las celdas (n_{11}, n_{12}, n_{13}, n_{21}, $\ldots$, n_{33}), o más concretamente, para tratar de estimar las probabilidades de celda.

Tabla 1.2: *Ejemplo de tabla de contingencia de doble entrada con proporciones conjuntas y marginales*

ELECCIONES X \ ELECCIONES Y	Partido A	Partido B	Partido C	
Partido A	p_{11}	p_{12}	p_{13}	$p_{1\cdot}$
Partido B	p_{21}	p_{22}	p_{23}	$p_{2\cdot}$
Partido C	p_{31}	p_{32}	p_{33}	$p_{3\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	

He aquí la presentación del problema de la investigación, tratar de estimar las probabilidades de celda en tablas de contingencia cuando únicamente se observan las frecuencias marginales. En la Tabla 1.2 puede observarse como se construiría la tabla de contingencia con las correspondientes proporciones conjuntas y marginales.

Los objetivos que tratarán de poner solución a la presentación de este problema se encuentran divididos en, por un lado, un objetivo general eje de la investigación, y, por el otro lado, una serie de objetivos específicos complementarios al general.

Objetivo general:

- Estimar las probabilidades de celda de una tabla de contingencia dadas únicamente sus frecuencias marginales

Objetivos específicos:

- Justificar el uso de las probabilidades condicionales para estimar las probabilidades conjuntas
- Contrastar la hipótesis de que las probabilidades condicionales de transferencia de voto, no dependen de la provincia, circunscripción electoral o mesa electoral
- Diseñar un algoritmo de esperanza-maximización (EM) para la estimación de las probabilidades condicionales
- Estudiar la convergencia de las probabilidades condicionales a lo largo de las iteraciones del algoritmo y su tiempo de ejecución
- Estimar mediante Monte Carlo el error cuadrático medio (MSE) de estimación de las probabilidades condicionales para cada conjunto de datos simulados

1.2. Justificación de la investigación

La estimación de las probabilidades de celda en una tabla de contingencia dadas sus marginales es una técnica estadística muy valiosa que puede tener aplicaciones en diversos ámbitos. En el contexto de la demoscopia, esta técnica podría emplearse para (1) analizar la transferencia de votos, permitiendo comprender cómo se transfieren los votos de un partido político a otro en distintas elecciones o situaciones. Esto podría proporcionar información valiosa sobre patrones de cambio en la preferencia electoral o la dinámica política, que sirvan como base para, por ejemplo, diseñar estrategias políticas y electorales. La estimación de las probabilidades conjuntas en una tabla de contingencia, dadas sus marginales, también podrían ayudar a (2) evaluar la fidelidad del voto, es decir, la tendencia de los votantes a mantener su preferencia política a lo largo del tiempo. Esto podría ser útil para medir la estabilidad o la volatilidad del electorado y comprender cómo los votantes cambian o mantienen sus preferencias políticas en diferentes contextos, y, por consiguiente, tener implicaciones significativas para la comprensión de la dinámica política, la estrategia electoral y la predicción de resultados electorales. Por último, en este contexto, (3) la demoscopia, como disciplina que estudia el comportamiento y las actitudes de los votantes, podría beneficiarse de esta técnica, sirviendo como herramienta adicional para el análisis de datos electorales y la comprensión de las preferencias y comportamientos de los votantes en diferentes contextos y momentos. Además, la estimación de las probabilidades de celda a partir de las marginales podría ser especialmente útil cuando los datos disponibles son limitados o cuando se busca una aproximación rápida y eficiente para comprender la transferencia de votos y la fidelidad del voto en estudios demoscópicos.

Además del contexto demoscópico, esta técnica también podría ser de provecho en otros campos como la investigación de mercado, las ciencias sociales y comportamentales, y los estudios de salud pública. En el campo de (1) la investigación de mercado, las tablas de contingencia son ampliamente utilizadas para analizar la relación entre variables categóricas, como las preferencias de marca, la lealtad del cliente o los patrones de compra. Esta investigación podría proporcionar información valiosa sobre cómo los consumidores se mueven entre diferentes marcas o categorías a lo largo del tiempo, lo que podría ser útil para la toma de decisiones estratégicas de marketing y el diseño de estrategias de fidelización de clientes. En campos como (2) la psicología, la sociología o la antropología, las tablas de contingencia pueden ser empleadas para analizar la relación entre variables categóricas en estudios de comportamiento humano, actitudes sociales o preferencias culturales. El empleo de esta técnica, por lo tanto, podría tener implicaciones teóricas y prácticas para la comprensión del comportamiento humano en diversos contextos. Por último, (3) en el campo de la salud pública, las tablas de contingencia pueden ser empleadas para analizar la asociación entre diferentes factores de riesgo, como la exposición a determinados agentes o a la presencia de enfermedades. Este estudio, por lo tanto, podría ser útil para comprender cómo se relacionan diferentes factores de riesgo y cómo de distribuyen en diferentes grupos

de la población, lo que podría tener implicaciones importantes para la prevención y control de enfermedades, la identificación de factores de riesgo y la toma de decisiones en políticas de salud pública.

Capítulo 2

Marco Teórico

En este capítulo se llevará a cabo una profunda revisión del material teórico fundamental y de ciertos trabajos de investigación que se encuentren directamente relacionados con el tema a abordar y que servirán como apoyo para fundamentar, contrastar y corroborar las teorías que conformarán el marco teórico de este trabajo fin de máster.

2.1. Tablas de contingencia

Si se trabaja con datos categóricos, una de las formas más comunes de resumir los datos es formando tablas de contingencia. Por lo general, el objetivo de formar tablas de contingencia se centra en estudiar si existe alguna asociación entre dos variables categóricas, una de ellas ocupando las filas de la tabla de contingencia y la otra las columnas.

Lo más común es considerar X e Y dos variables categóricas con I x J categorías respectivamente. Una observación cualquiera en la tabla de contingencia puede venir clasificada en una de las posibles I x J categorías que existen.

Cuando las casillas de la tabla contienen las frecuencias observadas, la tabla se denomina tabla de contingencia, término que fue introducido por Pearson en 1904. Una tabla de contingencia, con I filas y J columnas se denomina una tabla I x J.

2.1.1. Estructura de probabilidad para tablas de contingencia

Agresti (2007) en su libro *An Introduction to Categorical Data Analysis* trata la estructura de probabilidad para las tablas de contingencia, definiendo la distribución conjunta, la distribución marginal y la distribución condicional.

Distribución conjunta

Si π_{ij} indica la probabilidad de que (X,Y) ocurra en la casilla de la fila i y la columna j, la distribución de probabilidad $\{\pi_{ij}\}$ es la distribución conjunta de X e Y. Por lo tanto, la distribución conjunta vendría dada por:

$$\pi_{ij} = P(X = i, Y = j) \quad \text{con} \quad i = 1,...,I \quad \text{y} \quad j = 1,...,J$$

Distribución marginal

Las distribuciones marginales pueden definirse como los totales de filas o columnas que resultan al sumar las probabilidades conjuntas.

Denotando las probabilidades marginales para la variable fila como $\pi_{i\cdot}$ y las probabilidades marginales para la variable columna como $\pi_{\cdot j}$, donde el subíndice "·"denota la suma sobre ese índice. Esto es,

$$\pi_{i\cdot} = P(X=i) = \sum_{j=1}^{J} P(X=i, Y=j) = \sum_{j=1}^{J} \pi_{ij} \pi \cdot j = P(Y=j) = \sum_{i=1}^{I} P(X=i, Y=j) = \sum_{i=1}^{I} \pi_{ij}$$

Dichas probabilidades satisfacen:

$$\sum_{i} \pi i \cdot = \sum_{j} \pi \cdot j = \sum_{i} \sum_{j} \pi_{ij} = 1$$

Las distribuciones marginales únicamente proporcionan información acerca de una sola variable.

Distribución condicional

En la gran mayoría de tablas de contingencia, una de las variables, por ejemplo Y, puede considerarse como una variable de respuesta y la otra variable, X, como una variable explicativa o predictora. Cuando X es fijo en lugar de aleatorio, la noción de distribución conjunta para X e Y ya no es significativa. Sin embargo, para una categoría fija de X, la variable Y tiene una distribución de probabilidad. Esto permite estudiar cómo cambia su distribución a medida que cambia el valor de X.

Poniendo de ejemplo la distribución condicionada de Y dada X, se puede ver que, dado un sujeto que está clasificado en la fila i de X, $\pi_{j|i}$ denota la probabilidad de clasificación en la columna j de Y, $j = 1, ..., J$, es decir, $n_{j|i} = P(Y = j|X = i)$. Nótese que:

$$\sum_{j=1}^{J} \pi_{j|i} = 1 \quad \text{para todo} \quad i = 1, ..., I$$

Las probabilidades $\{\pi_{1|i}, ..., \pi_{J|i}\}$ forman la distribución condicional de Y en la categoría i de X.

2.2. Estimación de probabilidades de celda a partir de frecuencias marginales

En este apartado se realizará una revisión de algunos métodos existentes que han sido usados para estimar las probabilidades de celda de una tabla de contingencia dadas sus frecuencias marginales.

Es importante resaltar que los enfoques presentados en esta sección se utilizan para calcular las probabilidades de las celdas cuando se deben cumplir restricciones para las distribuciones marginales, pero, disponiéndose de las frecuencias de la tabla de contingencia.

2.2.1. Uso de funciones aditivas

Pelz y Good (1986) propusieron varios métodos para la estimación en las tablas de contingencia de las probabilidades de celda conocidas las probabilidades marginales y mostraron que varios de los métodos tienen una característica cualitativa en común. Esta característica es una ecuación que relaciona las probabilidades estimadas en las celdas de cada subtabla 2×2 para una tabla de contingencia ordinaria, o de cada subtabla 2^d

cuando se trata de una tabla d-dimensional, en la que todas las probabilidades marginales unidimensionales se consideran conocidas.

Pelz y Good (1986) consideraron, como ejemplo sencillo, una tabla ordinaria con r filas y s columnas y con celdas n_{ij} ($i = 1,2,...,r$; $j = 1,2,...,s$). Siendo $\sum_{ij} n_{ij} = N$, el tamaño de la muestra, y siendo π_{ij} las verdaderas probabilidades de celda. Estas probabilidades son desconocidas, pero, como se comentó anteriormente, las probabilidades marginales sí que se consideran conocidas. Es decir, se tendrían ecuaciones lineales de la forma:

$$\sum_i \pi_{ij} = b_j, \quad \sum_j \pi_{ij} = a_i \quad \text{con} \quad i = 1,2,...,r; j = 1,2,...,s \tag{2.1}$$

donde las a_i y las b_j se encuentran "fijadas de antemano". Según Pelz y Good(1986), únicamente habría $r+s-1$ restricciones lineales matemáticamente independientes porque el total de la última columna se determina cuando se asignan los totales de r filas y los totales de las primeras $s-1$ columnas.

Varios métodos de estimación implican la optimización, sujeta a las restricciones, de una función objetivo "aditiva", es decir, una función objetivo de la forma $x = \sum_{ij} f(n_{ij}, p_{ij})$, donde p_{ij} es una estimación de π_{ij}.

Según estos autores, cuando existe una función objetivo aditiva diferenciable se podría aplicar el método de multiplicadores indeterminados de Lagrange y se obtendrían enseguida ecuaciones para las estimaciones p_{ij}, de la forma:

$$g(n_{ij}, p_{ij}) = \alpha_i + \beta_j \tag{2.2}$$

donde g denota la derivada parcial de la función objetivo aditiva con respecto a su segundo argumento (p_{ij}), y α_i y β_j denotan los multiplicadores indeterminados.

Escogiendo un par de filas, indexadas por i e i', y un par de columnas indexadas por j y j', entonces:

$$g(n_{ij}, p_{ij}) - g(n_{ij'}, p_{ij'}) - g(n_{i'j}, p_{i'j}) + g(n_{i'j'}, p_{i'j'}) = 0 \tag{2.3}$$

Todas esas ecuaciones $r(r-1)s(s-1)/4$ pueden derivarse de las ecuaciones:

$$g(n_{ij}, p_{ij}) - g(n_{is}, p_{is}) - g(n_{rj}, p_{rj}) + g(n_{rs}, p_{rs}) = 0 \tag{2.4}$$

de las cuales solo hay $(r-1)(s-1)$ linealmente independientes. Cuando las ecuaciones (4) se combinan con:

$$\sum_i p_{ij} = b_j, \sum_j p_{ij} = a_i \tag{2.5}$$

se tienen un total de rs ecuaciones, que esperamos sean matemáticamente independientes, con rs incógnitas. Las formas precisas de las ecuaciones (4), para varios métodos de estimación, se encuentran recopiladas en la siguiente tabla, creada por los autores Pelz y Good (1986).

Tabla 2.1: *Varios métodos de estimación por Pelz y Good (1986)*

Method	**Objective function**	**Equation (4)**
Maximum Likelihood	$x_1 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(p_{ij})$	$n_{ij}/p_{ij} - n_{is}p_{is} - n_{rj}/p_{rj} + n_{rs}/p_{rs} = 0$
Minimum χ^2	$x_2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij} - Np_{ij})^2/(Np_{ij})$	$(n_{ij}/p_{ij})^2 - (n_{is}/p_{is})^2 - (n_{rj}/p_{rj})^2 + (n_{rs}/p_{rs})^2 = 0$
Minimum Modified χ^2	$x_3 = \sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij} - Np_{ij})^2/(n_{ij})$	$p_{ij}/n_{ij} - p_{is}/n_{is} - p_{rj}/n_{rj} + p_{rs}/n_{rs} = 0$
Minimum Discrimination Information	$x_4 = \sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}/N)\log[n_{ij}/(Np_{ij})]$	$n_{ij}/p_{ij} - n_{is}/p_{is} - n_{rj}/p_{rj} + n_{rs}/p_{rs} = 0$
Modified Minimum Discrimination Information	$x_5 = \sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij})\log(Np_{ij}/n_{ij})$	$\log(n_{ij}/p_{ij}) - \log(n_{is}/p_{is}) - \log(n_{rj}/p_{rj}) + \log(n_{rs}/p_{rs}) = 0$
Quasi-Bayes Estimation	$x_6 = \sum_{i=1}^{I}\sum_{j=1}^{J}[(n_{ij}+k)/(N+tk)]\log(p_{ij})$	$(n_{ij}+k)/p_{ij} - (n_{is}+k)/p_{is} - (n_{rj+k}/p_{rj}) + (n_{rs}+k)/p_{rs} = 0$
Maximum Entropy	$x_7 = -\sum_{i=1}^{I}\sum_{j=1}^{J}\log(p_{ij})$	$\log(p_{ij}) - \log(p_{is}) - \log(p_{rj}) + \log(p_{rs})$

2.2.2. Transformación lineal del espacio de parámetros de las probabilidades de celda

Pelz y Good (1986) mostraron cómo varios métodos de estimación de probabilidades de celda en tablas de contingencia dadas las probabilidades marginales pueden expresarse en términos de optimización de funciones objetivo aditivas. Emplearon la optimización restringida con multiplicadores de Lagrange para obtener las estimaciones de la función objetivo. Lipsitz y Zhao (1994) llegaron a la conclusión de que transformando linealmente el espacio de parámetros de las probabilidades de celda, las estimaciones de cualquiera de las funciones objetivo se pueden obtener sin necesidad de optimización restringida.

Distribución multinomial

Se supone que cada sujeto se mide en T variables categóricas, donde la t-ésima variable aleatoria Y_t tiene J_T niveles. Se denota la probabilidad conjunta de caer en la celda $j_1,...,j_T$ $(j_t = 1,...,J_T)$ de una tabla de contingencia con T variables categóricas como:

$$p_{j_1,...,j_T} = P[Y_1 = j_1,...,Y_t = j_T].$$

La tabla de contingencia tiene $C = \prod_{t=1}^{T} J_t$. Se denota el número observado de sujetos que caen en la celda $j_1,...,j_T$ como $Y_{j_1,...,j_T}$. Se supone que el vector aleatorio $y = \{y_{j1},...,y_{jT}\}$ sigue una distribución multinomial con un vector de probabilidad $p = \{p_{j1},...p_{j_T}\}$, donde p contiene todas las probabilidades de las celdas a excepción de $p_{J_1},...,p_{J_T}$. Así, el espacio paramétrico Ω_1 de p tiene dimensión $C-1$. Por último, se ordenarían los elementos de la forma $p_{j_1},...,p_{j_T}$.

Estimación para una tabla de doble entrada

Cuando la tabla de contingencia es de doble entrada, es decir, cuando está formada por dos variables categóricas, las probabilidades marginales dadas son:

$$p_{j_1+} = \sum_{j_2=1}^{J_2} p_{j_1 j_2} \quad j_1 = 1,...,J_1, \tag{2.6}$$

$$p_{j_2+} = \sum_{j_1=1}^{J_1} p_{j_1 j_2} \quad j_2 = 1,...,J_2, \tag{2.7}$$

donde $\{p_{j1+}, p_{+j_2}\}$ son conocidos y fijos. La estimación de las probabilidades de celda $p_{j_1 j_2}$ implica entonces la optimización de una función objetivo mediante el uso de multiplicadores de Lagrange, con las restricciones (2.6) y (2.7). El principal objetivo entonces de Lipsitz y Zhao (1994) sería reparametrizar el modelo para las probabilidades de celda de modo que se incorporen las restricciones.

Como ejemplo sencillo de reparametrización en tablas de contingencia, los autores consideran una muestra de una distribución multinomial con K probabilidades; se pretende estimar las probabilidades optimizando una función objetivo dada. En primer lugar, se podría optimizar la función objetivo sujeta a la restricción de que las probabilidades multinomiales sumen 1 utilizando multiplicadores de Lagrange. Alternativamente, se puede dejar que una de las probabilidades sea igual a 1 menos la suma de las otras $K-1$ probabilidades, y maximizar la función objetivo como una función de las $K-1$ probabilidades. Los métodos dados por Pelz y Good (1986) son análogos al primer método; para Lipsitz y Zhao (1994) su método es análogo al segundo. En primer lugar, los autores describen heurísticamente cómo obtener la matriz de diseño para las probabilidades de celda cuando se dan las probabilidades marginales y, a continuación, discuten sobre la transformación del espacio de parámetros que conducirá a la matriz de diseño.

2.2.3. Estimación por máxima verosimilitud y programación geométrica

Programación geométrica

Según Wang *et al.* (2015), sea $g(x)$ una función monomial definida de la siguiente forma:

$$g(x) = c x_1^{a_1} x_2^{a_2} ... x_n^{a_n}$$

donde $c > 0$, $a_i \in \mathbb{R}$ y $x_i \in \mathbb{R}^+$ para $1 \leq i \leq n$. Sea $f(x)$ una función polinómica definida como la suma de uno o más monomios:

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{k1}} x_2^{a_{2k}} ... x_n^{a_{nk}}$$

Entonces, un tipo especial de problema de optimización convexo, llamado programación geométrica (GP), puede escribirse como sigue:

$$\begin{array}{ll} \text{mín} & f_0(x) \\ \text{s.a.} & f_i(x) \leq 1 \quad i = 1,...,m, \\ & g_j(x) = 1 \quad j = 1,...,p, \end{array} \tag{2.8}$$

donde $x = (x_1,...,x_n)$ es el vector de filas que contiene todas las variables de optimización (positivas), f_i son funciones posimoniales y g_i son funciones monomiales para todo $i = 0,...,m$ y $j = 1,...,p$.

La programación geométrica es una técnica común para resolver el problema de optimización anterior. Una vez realizada esta modelización de la programación geométrica, es decir, la formulación de los problemas en forma GP (2.8), que apenas requiere conocimientos de detalles técnicos y es conceptualmente sencilla, se dispone de una forma eficaz y fiable para resolver un problema práctico.

Estimación por máxima verosimilitud y programación geométrica en tablas de contingencia

En su estudio, Wang *et al.* (2015), considera el uso de la programación geométrica con tablas de contingencia, con el objetivo de abordar dos tipos importantes de problemas que se dan con frecuencia en la práctica. El primero es la estimación de las probabilidades de celda dadas las marginales; y el segundo es la estimación de las probabilidades de celda bajo dos tipos comunes de restricciones: marginales conocidas y marginales / condicionales.

Wang *et al.* (2015) lleva a cabo su estudio examinando situaciones distintivas en tablas de contingencia de triple entrada.

En primer lugar, el autor trata el problema de estimar las probabilidades de celda en tablas de contingencia con marginales conocidas. Para ello, considera una tabla de contingencia $r \times s \times t$, donde las tres dimensiones se encuentran asociadas a las variables discretas X_1, X_2 y X_3, respectivamente. Sea p_{ijk} la probabilidad de que una observación caiga en la celda (i, j, k), cuya frecuencia observada se denota por n_{ijk} en una muestra de tamaño n. Suponiendo que estas frecuencias siguen una distribución multinomial con parámetros n y p_{ijk}, la funcion de verosimilitud entonces es proporcional a:

$$\mathscr{L}(p_{ijk}) = \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{n_{ijk}}, \tag{2.9}$$

que es una función monomial de p_{ijk}.

En las tablas de contingencia de triple entrada, existen dos tipos de probabilidades marginales que pueden ser conocidas: unidimensionales como p_{i++} (es decir, se conoce alguna información univariante sobre X_1) y bidimensionales como p_{ij+} (es decir, se conoce alguna información bivariante sobre X_1 y X_2). Wang *et al.* (2015) muestra, bajo tres situaciones distintas, como la estimación por máxima verosimilitud (MLE) de p_{ijk} puede calcularse mediante programación geométrica. En lugar de maximizar la función de verosmilitud $\mathscr{L}$, se minimiza la inversa de $\mathscr{L}$ de forma equivalente.

(I) **Los tres conjuntos de probabilidades marginales unidimensionales son conocidos**. El problema de optimización puede describirse como:

$$\begin{aligned}
\underset{p_{ijk}}{\text{minimizar}} \quad & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} \\
\text{sujeto a} \quad & \sum_{j,k} p_{ijk} = p_{i++}, \quad \text{for } i = 1,...,r, \\
& \sum_{i,k} p_{ijk} = p_{+jk}, \quad \text{for } j = 1,...,s, \\
& \sum_{i,j} p_{ijk} = p_{++k}, \quad \text{for } k = 1,...,t,
\end{aligned} \tag{2.10}$$

donde p_{i++}, p_{+j+} y p_{++k} son constantes positivas que satisfacen $\sum_i p_{i++} = \sum_j p_{+j+} = \sum_k p_{++k} = 1$.

(II) **Se conoce un conjunto de probabilidades marginales unidimensionales y el resto de probabilidades bidimensionales**. Aquí, el problema de optimización puede describirse como:

$$\begin{aligned}
\underset{p_{ijk}}{\text{minimizar}} \quad & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} \\
\text{sujeto a} \quad & \sum_{j,k} p_{ijk} = p_{i++}, \quad \text{for } i = 1,...,r, \\
& \sum_{i} p_{ijk} = p_{+jk}, \quad \text{for } j = 1,...,s, \quad k = 1,...,t,
\end{aligned} \tag{2.11}$$

donde $p_{i..}$ y $p_{.jk}$ son constantes positivas que satisfacen $\sum_i p_{i++} = \sum_{j,k} p_{+jk} = 1$

(III) **Se conocen dos conjuntos cualesquiera de probabilidades marginales bidimensionales**. Aquí, el problema de optimización puede describirse como:

$$\begin{aligned} \underset{p_{ijk}}{\text{minimizar}} \quad & \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{ijk}^{-n_{ijk}} \\ \text{sujeto a} \quad & \sum_k p_{ijk} = p_{ij+}, \quad \text{for } i = 1,...,r, \quad j = 1,...,s, \\ & \sum_j p_{ijk} = p_{i+k}, \quad \text{for } i = 1,...,r, \quad k = 1,...,t, \end{aligned} \tag{2.12}$$

donde p_{ij+} y p_{i+k} son todas constantes positivas que satisfacen $\sum_j p_{ij+} = \sum_k p_{i+k}$ para $1,...,r$ y $\sum_{ij} p_{ij+} = \sum_{ik} p_{i+k} = 1$

Wang *et al.* (2015) explica que en la práctica, es frecuente que sólo se disponga de un subconjunto de restricciones en cada una de las tres situaciones mencionadas. Por ejemplo, en la primera situación, es posible que sólo se conozcan uno o dos conjuntos de probabilidades marginales unidimensionales. O, en la segunda y tercera situaciones, sólo un conjunto de probabilidades marginales bidimensionales pueden ser conocidas. O en las tres situaciones, puede conocerse un conjunto incompleto de probabilidades marginales. Debe tenerse en cuenta que, cuando ningún conjunto está completo, la restricción $\sum_{i,j,k} p_{ijk} = 1$ debe estar presente en el problema de optimización. En tales situaciones, la modelización de la programación geométrica podría realizarse fácilmente con el mismo espíritu; y la prueba de equivalencia en la estimación de las probabilidades de todas las celdas no vacías es esencialmente la misma que antes ya que la ecistencia de la celda pivote no se vería afectada.

En segundo lugar, el autor trata el problema de estimar las probabilidades de celda bajo probabilidades marginales/condicionales ordenadas. En las tablas de contingencia, las variables y sus probabilidades marginales o condicionales pueden ordenarse de forma natural. Para ello, considera de nuevo una tabla de contingencia $r \times s \times t$ donde las tres dimensiones se encuentran asociadas a las variables cualitativas X_1, X_2 y X_3.

Suponiendo que las probabilidades marginales p_{i++} o las probabilidades condicionales $p_{jk|i}$ se ordenan de acuerdo con los valores de X_1 indexados por i, $i \in \mathscr{I} = \{1,...,r\}$. Es decir,

$$p_{1++} \leq p_{2++} \leq \cdots \leq p_{r++} \tag{2.13}$$

o

$$p_{jk|1} \leq p_{jk|2} \leq \cdots \leq p_{jk|r} \quad \text{para } (j,k) \in C \tag{2.14}$$

En (2.14) se supone que C es un subconjunto propio de $J \times K$, donde $J = \{1,...,s\}$ y $K = \{1,...,t\}$ son conjuntos de índices para X_2 y X_3. Obsérvese que si $C = J \times K$, entonces (2.14) implica $p_{jk|1} \leq p_{jk|2} \leq ... \leq p_{jk|r}$ para todo j y k porque $\sum_{jk} p_{jk|i} = 1$ para todo i, lo que representa un caso trivial. Se asume además algunas condiciones de regularidad muy suaves bajo cada restricción de orden. Así, bajo estas condiciones de regularidad, la función de verosimilitud a ser maximizada aquí viene dada por (2.9), sujeto a las restricciones de orden (2.13) o (2.14), más la restricción de igualdad $\sum_{ijk} p_{ijk} = 1$.

Para realizar la modelización GP, se vuelve a parametrizar la función de verosimilitud (2.9) con $p_{jk|i}$ y p_{i++}, es decir

$$\mathscr{L}(p_{ijk}) = \left(\prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{t} p_{jk|i}^{n_{ijk}}\right)\left(\prod_{i=1}^{r} p_{i++}^{n_{i++}}\right) \tag{2.15}$$

La restricción de igualdad $\sum_{ijk} p_{ijk} = 1$ es entonces equivalente a

$$\sum_{jk} p_{jk|i} = 1, \quad \text{para todo } i = 1,2,...,r, \quad \text{y} \quad \sum_{i} p_{i++} = 1 \tag{2.16}$$

2.3. Algoritmo Expectation-Maximization

2.3.1. Contexto

Usurralde Casas (2017) comenta que, cuando se trabaja con casos reales, en ocasiones el investigador se encuentra ante una situación con valores perdidos, es decir, con datos incompletos. El autor comenta que esta situación es relativamente frecuente y puede darse, por ejemplo, en diagnósticos médicos -ya que los historiales contienen un número limitado de pruebas-, en fallos durante transmisiones de datos, en pruebas médicas imposibles de realizar o en el enmascaramiento de señales debido al ruido.

Antiguamente, lo más común era deshacerse de todas las muestras que contasen con valores perdidos. Este procedimiento puede suponer un gran problema ya que puede darse la situación de reducir tanto el número de datos que el investigador se encuentro con una muestra demasiado pequeña y muy poco representativa. Por lo tanto, este procedimiento, únicamente podría ser aceptable cuando el número de datos faltantes es pequeño y tienen origen aleatorio.

Como alternativa, se creó el Algoritmo de Esperanza-Maximimización o Expectation-Maximization. El nombre viene dado por Dempster, Laird y Rubin en 1977, los cuales también dieron una variedad de ejemplos de su aplicabilidad y establecieron su convergencia y otras propiedades básicas bastante generales. Estos autores comentan que las situaciones en las que se puede aplicar este algoritmo no solo incluyen los casos antes mencionados en los que existe la presencia de valores perdidos, distribuciones truncadas o censuradas, sino también donde la falta de datos no es tan evidente. Estas situaciones incluirían modelos estadísticos tales como efectos aleatorios, mixturas, modelos logarítmicos lineales , y de clases latentes y estructuras variables latentes. Dempster *et al.* (1977)

2.3.2. Objetivo

Usurralde Casas (2017) señala que el Algoritmo EM es un método iterativo para realizar una estimación de máxima verosimilitud (MLE) de parámetros de problemas en los que existen datos perdidos. El objetivo principal de este algoritmo, por lo tanto, es asociar a un problema de datos incompletos otro problema con datos completos, estableciendo relaciones entre la verosimilitud de estos dos problemas.

En cada iteracción del algoritmo se realizan los dos siguientes pasos:

(I) ***Paso E o Paso de Esperanza o Expectation***

En este paso lo que se lleva a cabo es el "relleno" de los datos faltantes, es decir, se crea el nuevo problema pero esta vez con los datos completos. Además debe calcularse una función de verosimilitud para el conjunto de datos completos. La forma de construir estos datos completos artificiales es a través de la esperanza condicionada a los datos observados.

(II) ***Paso M o Paso de Maximizació***n o Maximization

En este paso se encuentran los parámetros que maximizan la función de log-verosimilitud calculada en el paso E.

Por último, estos dos pasos son repetidos iterativamente hasta alcanzar la convergencia.

2.3.3. Propiedades del algoritmo

En comparación con otros algoritmos iterativos, el algoritmo EM posee algunas ventajas pero presenta ciertas limitaciones, ya que en algunas situaciones puede darse una convergencia muy lenta. Usurralde Casas

(2017) recopila una serie de ventajas y desventajas que serán mostradas a continuación:

(I) *Ventajas*:

- Es numéricamente estable con cada iteracción en la que aumenta la verosimilitud.
- En condiciones generales, tiene una convergencia global fiable. Es decir, si empezamos de un punto arbitario $\theta^{(0)}$ en el espacio de los parámetros, convergeramos casi siempre a un máximo local. Esto no ocurre cuando hacemos una mala elección del punto $\theta^{(0)}$ o por alguna patología local de la función de verosimilitud.
- Es fácil de implementar ya que se basa en el cálculo de datos completos y también es fácil de programar.
- Requiere poco espacio de almacenamiento y se puede realizar en un ordenador pequeño.
- Ya que el problema con los datos completos es más o menos estándar, el paso M se puede realizar con frecuencia empleando paquetes estadísticos estándar en situaciones donde la estimación de máxima verosimilitud de datos completos no existe de forma cerrada.
- El trabajo analítico necesario es mucho más simple que con otros métodos.
- El coste por cada iteracción es bajo, lo que compensa el mayor número de iteracciones necesitadas respecto a otros algoritmos.
- Observando el crecimiento monótono de la verosimilitud a lo largo de las iteracciones, es fácil controlar errores de convergencia y de programación.
- El algoritmo EM puede usarse para proporcionar estimaciones de los valores de los datos perdidos.

(II) *Desventajas*:

- No contiene un procedimiento incorporado para proporcionar una estimación de la matriz de covarianza de las estimaciones de los parámetros.
- Puede converger de manera muy lenta en algunos problemas con demasiados datos incompletos o en algunos problemas aparentemente inofensivos.
- No garantiza la convergencia al máximo global cuando hay múltiples máximos. En estos casos la estimación obtenida depende del valor inicial.
- En algunos casos el paso E puede ser analíticamente intratable, aunque en estas situaciones se tiene la posibilidad de aplicarlo mediante el método Monte Carlo.

2.4. Contexto electoral España 2019

En los últimos años, España ha vivido una intensa actividad electoral que ha tenido un impacto significativo en la política del país. Desde las elecciones generales de 2015, que llevaron a un cambio en el gobierno después de décadas de alternancias entre el Partido Popular (PP) y el Partido Socialista Obrero Español (PSOE), se ha producido un aumento en la polarización política y la fragmentación del voto (Miller, 2020). En este contexto, es fundamental revisar las elecciones generales y municipales, que tuvieron lugar en un clima de incertidumbre política y social.

2.4.1. Elecciones Generales 2019

En España, las elecciones generales de 2019 se celebraron en dos ocasiones debido a la falta de acuerdo para la formación de un gobierno. La primera votación se llevó a cabo el 28 de abril de 2019 y la segunda votación el 10 de noviembre de ese mismo año. Ambas elecciones se celebraron en un clima político complejo y polarizado, con la crisis catalana, la corrupción y el auge de partidos de ”extrema derecha” como temas centrales (Miller, 2020).

Siguiendo los datos recopilados por el periódico El País, en las elecciones de abril, el Partido Socialista Obrero Español (PSOE) ganó las elecciones con el 28.7 % de los votos, pero no obtuvo suficientes escaños para formar un gobierno estable. El Partido Popular (PP) quedó en segundo lugar con el 16.7 % de los votos, seguido de Ciudadanos con el 15.9 % y Unidas Podemos con el 14.3 %. VOX, un partido que había entrado en el Parlamento a raíz de las elecciones andaluzas de diciembre de 2018, también logró un importante avance, obteniendo el 10.3 % de los votos convirtiéndose en la quinta fuerza política en el Congreso de los Diputados.

Según RTVE, después de las elecciones de abril, el PSOE intentó formar un gobierno de coalición con Unidas Podemos, pero las negociaciones fracasaron debido a desacuerdos en temas como la política fiscal, la reforma laboral y la crisis catalana. Finalmente, el 23 de septiembre, el Rey disolvió el Parlamento y convocó nuevas elecciones para el 10 de noviembre.

En las elecciones de noviembre, el PSOE volvió a ganar con el 28 % de los votos, aunque perdió tres escaños en comparación con las elecciones de abril. El PP se mantuvo en segundo lugar, aunque también perdió votos y escaños, mientras que VOX aumentó su representación en el Congreso. La formación de un gobierno estable siguió siendo difícil, y finalmente, después de varios meses de negociaciones, el PSOE y Unidas Podemos acordaron formar un gobierno de coalición en enero de 2020 (La Moncloa, 2020).

2.4.2. Elecciones Municipales 2019

En mayo de 2019, se celebraron elecciones municipales en España, en las que se eligieron a los alcaldes y concejales de los ayuntamientos de todo el país. Estas elecciones también se llevaron a cabo en un clima político polarizado y complejo (ya que se celebraron en un contexto temporal cercano a las elecciones generales de abril de ese mismo año), con la crisis catalana y la corrupción como temas centrales.

En las elecciones municipales, el Partido Socialista Obrero Español (PSOE) fue el partido más votado, obteniendo el 29.2 % de los votos y ganando en 6 de las 10 mayores ciudades de España, incluyendo Madrid y Barcelona. El Partido Popular (PP) quedó en segundo lugar con el 22.7 % de los votos y Ciudadanos obtuvo el 11.1 %. En cambio, Unidas Podemos perdió representación en los ayuntamientos, obteniendo solo el 7.8 % de los votos (El País, 2019).

A nivel regional, los partidos nacionalistas y separatistas catalanes obtuvieron una importante representación en Cataluña, con ERC (Esquerra Republicana de Catalunya) como la formación más votada en las elecciones municipales en la región. En el País Vasco, el Partido Nacionalista Vasco (PNV) mantuvo su posición dominante en la mayoría de los ayuntamientos de la región.

En resumen, las elecciones municipales de 2019 en España estuvieron marcadas por la victoria del PSOE en las principales ciudades del país, la pérdida de representación de Unidas Podemos y el fortalecimiento de los partidos nacionalistas y separatistas en Cataluña y el País Vasco. Además, como apunta Miller (2020), la corrupción y la crisis catalana continuaron siendo temas centrales en estas elecciones.

Capítulo 3

Metodología

Una vez delimitados los objetivos de la investigación y redactado el marco teórico correspondiente, es momento de desarrollar el diseño de la investigación, es decir, el diseño metodológico a seguir para dar respuesta a los objetivos. En este capítulo, se detallará la selección y obtención de los datos necesarios para llevar a cabo la investigación, así como el proceso metodológico empleado para su análisis y explotación.

3.1. Selección de los datos

Los conjuntos de datos con los que se procederá a realizar el análisis metodológico han sido, por un lado, reales, obtenidos del Centro de Investigaciones Sociológicas y del catálogo de datos abiertos del Gobierno de España, y, por el otro lado, simulados.

- **Datos reales**

Del CIS se ha seleccionado el estudio Nº 3269 titulado ”Barómetro de diciembre 2019. Postelectoral elecciones generales 2019”. Esta encuesta realizada después de las elecciones generales de noviembre de 2019 a la población española de ambos sexos de 18 años y más, con un tamaño muestral de 4.804 entrevistados, recopila en forma de pregunta (1) el recuerdo de voto en las elecciones generales de abril de 2019 y (2) el recuerdo de voto en las elecciones generales de noviembre de 2019. Ambas preguntas junto con sus respuestas por parte de los entrevistados han sido extraídas del cuestionario y se han organizados en forma de tabla donde cada fila representa un entrevistado y cada columna la correspondiente respuesta a las cuestiones: (1) *Comunidad Autónoma*, (2) *Provincia*, (3) *Recuerdo de voto en las elecciones generales de abril de 2019* y (4) *Recuerdo de voto en las elecciones generales de noviembre de 2019*.

Del catálogo de datos abiertos del Gobierno de España, se han seleccionado y obtenido los resultados de las elecciones generales de abril y municipales del año 2019 por mesa electoral del Ayuntamiento de Torrent. Ambos datos para ambas elecciones se encuentran organizados en forma de tabla donde cada fila representa cada una de las mesas electorales del Ayuntamiento de Torrent, cada columna la correspondiente opción política seleccionada por los ciudadanos, y las celdas, las frecuencias de voto. Debido a la gran cantidad de opciones políticas y las bajas frecuencias de voto en algunas de ellas se ha optado por simplificar los datos, de tal manera, que se han seleccionado las seis grandes fuerzas políticas en el año 2019 en el Ayuntamiento de Torrent (PSOE, PP, PODEMOS, CIUDADANOS, VOX y COMPROMÍS), se han sintetizado aquellos partido minoritarios con bajas frecuencias de votos en una opción denominada ”Otros Partidos”, se han sintetizado también las opciones ”Voto en Blanco”y ”Voto Nulo”en una única opción denominada ”Voto Blanco o Nulo”, y, se ha creado una nueva columna denominada ”Abstención”como resultado entre el número total de personas censadas menos aquellas que depositaron su voto en la correspondiente urna de la mesa electoral.

- **Datos simulados**

Se han simulado dos conjuntos de datos que presentan la siguiente estructura:

- *Conjunto 1*. $I = 3$, $J = 3$, $K = 10$. Es decir, 3 opciones políticas en las primeras elecciones, 3 opciones políticas en las segundas elecciones y 10 mesas electorales.
- *Conjunto 2*. $I = 7$, $J = 6$, $K = 10$. Es decir, 7 opciones políticas en las primeras elecciones, 6 opciones políticas en las segundas elecciones y 10 mesas electorales.

Para ello se ha fijado de manera controlada (1) una matriz de probabilidades condicionales adaptada a las características de cada uno de los datos, (2) una matriz de probabilidades de voto para cada una de las opciones politicas en cada una de las mesas electorales para las primeras elecciones y (3) el tamaño muestral de cada una de las mesas electorales.

A continuación, a partir de estos datos previamente fijados se han creado dos matrices que representan cada una de las dos elecciones, siendo las columnas cada una de las opciones políticas, las filas las mesas electorales y las celdas las frecuencias de voto para cada una de las opciones políticas en cada mesa electoral.

Conjunto 1 de datos simulados

Este Conjunto 1 de datos simulados presenta la estructura $I = 3$, $J = 3$ y $K = 10$.

Las probabilidades condicionales c_{ij} y las probabilidades de voto por mesa electoral para las primeras elecciones $p_{i\cdot}^{(k)}$ que se han empleado para simular los datos del Conjunto 1 se encuentran en las siguientes tablas:

Tabla 3.1: *Matriz de probabilidades condicionales. Conjunto 1*

	Partido 1	Partido 2	Partido 3
Partido 1	0.65	0.15	0.20
Partido 2	0.25	0.60	0.15
Partido 3	0.30	0.20	0.50

En esta Tabla 3.1 de probabilidades condicionales se han seleccionado valores mayores o iguales que 0.15, con el objetivo de estudiar como se comporta su estimación cuando las probabilidades condicionales no son demasiado pequeñas ($> 0{,}15$).

Tabla 3.2: *Matriz de probabilidades de voto para las primeras elecciones por mesa electoral*

	Partido 1	Partido 2	Partido 3
Mesa 1	0.45	0.25	0.30
Mesa 2	0.55	0.30	0.15
Mesa 3	0.53	0.27	0.20
Mesa 4	0.39	0.31	0.30
Mesa 5	0.42	0.35	0.23
Mesa 6	0.33	0.40	0.27
Mesa 7	0.38	0.29	0.33
Mesa 8	0.25	0.34	0.41
Mesa 9	0.41	0.34	0.25
Mesa 10	0.30	0.39	0.31

En la Tabla 3.2, se ha seguido la dinámica anterior y se han seleccionado siempre probabilidades mayores o iguales que 0.15.

Como resultado de emplear las anteriores probabilidades condicionales c_{ij}, las probabilidades de voto para las primeras elecciones por mesa electoral $p_{i\cdot}^{(k)}$ y fijar un conjunto de votantes también por mesa electoral $n_{\cdot\cdot}^{(k)}$,

se han creado dos matrices de datos simulados, una para cada una de las elecciones, con las que se pretenderá estimar la matriz fijada inicialmente de probabilidades condicionales. Pueden observarse en las siguientes dos tablas.

Tabla 3.3: *Datos simulados de votos por mesa electoral en las primeras elecciones*

	Partido 1	Partido 2	Partido 3
Mesa 1	181	93	126
Mesa 2	229	144	77
Mesa 3	268	134	98
Mesa 4	183	202	165
Mesa 5	253	215	132
Mesa 6	228	248	174
Mesa 7	280	203	217
Mesa 8	191	253	306
Mesa 9	316	281	203
Mesa 10	266	325	259

Tabla 3.4: *Datos simulados de votos por mesa electoral en las segundas elecciones*

	Partido 1	Partido 2	Partido 3
Mesa 1	177	104	119
Mesa 2	212	139	99
Mesa 3	236	150	114
Mesa 4	208	179	163
Mesa 5	259	196	145
Mesa 6	279	204	167
Mesa 7	302	191	207
Mesa 8	289	236	225
Mesa 9	326	254	220
Mesa 10	353	259	238

En cada una de las Tablas 3.3 y 3.4, se representan las frecuencias marginales de voto a cada partido en cada una de las mesas electorales.

Conjunto 2 de datos simulados

Este Conjunto 2 de datos simulados presenta la estructura $I = 7$, $J = 6$ y $K = 10$.

Las probabilidades condicionales y las probabilidades de voto por mesa electoral para las primeras elecciones que se han empleado para simular los datos del Conjunto 2 se encuentran en las siguientes tablas:

Tabla 3.5: *Matriz de probabilidades condicionales. Conjunto 2.*

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.80	0.10	0.02	0.00	0.05	0.03
VOX	0.12	0.75	0.00	0.03	0.05	0.05
PSOE	0.05	0.05	0.70	0.15	0.03	0.02
PODEM	0.00	0.00	0.30	0.65	0.00	0.05
COMUN	0.00	0.00	0.45	0.50	0.00	0.05
CS	0.15	0.10	0.05	0.00	0.65	0.05
O.P.	0.15	0.10	0.20	0.05	0.15	0.35

En esta Tabla 3.5 de probabilidades condicionales, se han seleccionado valores que puedan ser más o menos parecidos a la realidad, nombrando a las candidaturas con nombres de partidos políticos reales. El objetivo es estudiar como se comporta la estimación de c_{ij} cuando, en algunos casos, $c_{ij} < 0{,}15$.

Tabla 3.6: *Matriz de probabilidades de voto para las primeras elecciones por mesa electoral*

	PP	VOX	PSOE	PODEM	COMUN	CS	O.P.
Mesa 1	0.25	0.15	0.30	0.10	0.05	0.10	0.05
Mesa 2	0.20	0.10	0.35	0.15	0.10	0.05	0.05
Mesa 3	0.15	0.05	0.40	0.15	0.05	0.15	0.05
Mesa 4	0.20	0.15	0.25	0.10	0.05	0.15	0.10
Mesa 5	0.30	0.20	0.20	0.10	0.10	0.05	0.05
Mesa 6	0.18	0.12	0.33	0.17	0.04	0.10	0.06
Mesa 7	0.33	0.17	0.18	0.12	0.10	0.06	0.04
Mesa 8	0.30	0.10	0.25	0.15	0.10	0.05	0.05
Mesa 9	0.20	0.10	0.30	0.10	0.10	0.05	0.15
Mesa 10	0.25	0.10	0.30	0.15	0.10	0.10	0.00

En la Tabla 3.6, se ha seguido la dinámica anterior y se han seleccionado probabilidades de cualquier valor.

Como resultado de emplear c_{ij}, $p_{i\cdot}^{(k)}$ y fijar $n_{\cdot\cdot}^{(k)}$, se han creado dos matrices de datos simulados, una para cada una de las elecciones. Pueden observarse en las siguientes dos tablas:

Tabla 3.7: *Datos simulados de votos por mesa electoral en las primeras elecciones*

	PP	VOX	PSOE	PODEM	COMUN	CS	O.P.
Mesa 1	182	99	198	75	45	81	43
Mesa 2	115	51	178	88	69	34	31
Mesa 3	81	27	228	60	28	77	27
Mesa 4	164	124	191	100	40	116	82
Mesa 5	262	173	163	106	78	46	42
Mesa 6	127	76	209	118	36	78	54
Mesa 7	222	104	112	90	71	36	34
Mesa 8	260	79	194	128	105	51	48
Mesa 9	119	53	159	67	63	33	92
Mesa 10	177	63	191	108	86	81	0

Tabla 3.8: *Datos simulados de votos por mesa electoral en las segundas elecciones*

	PP	VOX	PSOE	PODEM	CS	O.P.
Mesa 1	177	117	201	107	81	40
Mesa 2	114	63	198	118	39	34
Mesa 3	91	47	206	89	63	32
Mesa 4	175	141	217	118	108	58
Mesa 5	242	177	211	132	63	45
Mesa 6	135	92	224	127	73	47
Mesa 7	198	116	154	116	51	34
Mesa 8	234	107	252	162	63	47
Mesa 9	125	70	185	107	46	53
Mesa 10	163	86	224	133	73	27

En cada una de las Tablas 3.7 y 3.8, se representan las frecuencias marginales de voto a cada partido en cada una de las mesas electorales.

3.2. Notación y estructura de los parámetros

Estipulado el objetivo general eje de esta investigación "estimar las probabilidades de celda de una tabla de contingencia dadas únicamente sus frecuencias marginales", se procederá a continuación con la explicación del diseño metodológico establecido para lograr cumplirlo.

Sea K el número de mesas electorales, I el número de candidaturas en las primeras elecciones, J el número de candidaturas en las segundas elecciones, y k, i y j los índices correspondientes de K, I y J. Se tiene que, en general:

Tabla 3.9: Tabla de contingencia con probabilidades condicionales y frecuencias marginales

i \ j	$n_{\cdot 1}^{(k)}$	$\dots$	$n_{\cdot J}^{(k)}$
$n_{1\cdot}^{(k)}$	c_{11}	$\dots$	c_{1J}
$\dots$	$\dots$	$\dots$	$\dots$
$n_{I\cdot}^{(k)}$	c_{I1}	$\dots$	c_{IJ}

En la Tabla 3.9, los datos que se presentan por mesa electoral (k) no es la información individual de cada votante sino que es el número total de votos a una opción política en cada una de las elecciones, es decir, las frecuencias marginales. Todo esto en cada una de las k-ésimas mesas electorales, siendo $n_{i\cdot}^{(k)}$ y $n_{\cdot j}^{(k)}$ el recuento de las decisiones de los individuos por las opciones políticas i y j en las primeras y segundas elecciones respectivamente.

La matriz $C = c_{ij}$ representaría las probabilidades condicionales, es decir, $P(Y = j|X = i)$, siendo X las primeras elecciones e Y las segundas. Esta matriz C es totalmente desconocida ya que únicamente se conocen las frecuencias marginales. En el caso de conocer las frecuencias conjuntas, la estimación de la matriz C sería muy sencilla. Si se observa de nuevo la estructura de la Tabla 1.2, en la que se muestran las probabilidades conjuntas p_{ij} y las probabilidades marginales $p_{i\cdot}$ y $p_{\cdot j}$, las probabilidades condicionales c_{ij} de la Tabla 3.9 se construyen como $p_{ij}/p_{i\cdot}$. Esto es así, siempre y cuando se pueda suponer que c_{ij} no depende de la mesa electoral (k). Es decir, debe suponerse que c_{ij} se mantiene constante en todas las mesas electorales. Para ello, debe contrastarse la hipótesis de que las probabilidades condicionales de transferencia de voto, no dependen de la mesa electoral, circunscripción electoral o provincia. [1]

Esta matriz C de probabilidades condicionales es lo que se pretende estimar, ya que, si se estima con éxito sería sumamente sencillo calcular las probabilidades conjuntas para ambas elecciones. En el caso de contar con los parámetros reales, $p_{ij} = p_{i\cdot}c_{ij}$, siendo $p_{i\cdot}$ la probabilidad (marginal) de escoger la opción política i en las primeras elecciones. Como ese no es el caso, no queda más opción que trabajar con las estimaciones de los parámetros teniendo en cuenta la información proporcionada por cada una de las mesas electorales.

[1] El procedimiento seguido para suponer que c_{ij} no depende de la mesa electoral k viene explicado en el apartado *Contraste de la homogeneidad de transferencia de voto*

Dado que,

$$\hat{p}_{i\cdot}^{(k)} = \frac{n_{i\cdot}^{(k)}}{\sum_{r=1}^{I} n_{r\cdot}^{(k)}} \quad \text{para todo } i = 1,...,I$$

La estimación de $p_{ij}^{(k)} = P(X = i, Y = j|z = k)$ sería

$$\hat{p}_{ij}^{(k)} = \hat{p}_{i\cdot}^{(k)} \hat{c}_{ij} \quad \text{para todo } i = 1,...,I, \quad j = 1,...,J, \quad k = 1,...,K,$$

y, puesto que, la probabilidad (marginal) de escoger la opción política j en las segundas elecciones es $p_{\cdot j}^{(k)} = P(Y = j|Z = k) = \sum_{i=1}^{I} p_{ij}^{(k)}$, la estimación de $p_{\cdot j}^{(k)}$ resulta en,

$$\hat{p}_{\cdot j}^{(k)} = \sum_{i=1}^{I} \hat{p}_{ij}^{(k)} = \sum_{i=1}^{I} \hat{p}_{i\cdot}^{(k)} c_{ij} \quad \text{para todo } i = 1,...,I, \quad j = 1,...,J, \quad k = 1,...,K,$$

Por último, y teniendo en cuenta que las frecuencias conjuntas para cada mesa electoral $n_{ij}^{(k)}$ son totalmente inobservables si únicamente se encuentran disponibles las frecuencias marginales, se puede establecer la verosimilitud para el conjunto de mesas electorales para llevar a cabo la estimación de la matriz C de probabilidades condicionales a partir de la fórmula de cálculo de la distribución multinomial para la masa de probabilidad.

$$\mathscr{L}(C) = \prod_{k=1}^{K} \frac{n_{\cdot\cdot}^{(k)}!}{\prod_{r=1}^{I}\prod_{s=1}^{J} n_{rs}^{(k)}!} \prod_{i=1}^{I}\prod_{j=1}^{J} (\hat{p}_{i\cdot} \cdot c_{ij})^{n_{ij}^{(k)}}$$

$$\alpha_N^{(k)} = \frac{n_{\cdot\cdot}^{(k)}!}{\prod_{r=1}^{I}\prod_{s=1}^{J} n_{rs}^{(k)}!}$$

Aplicando logaritmos para convertir productos en sumas y transformando un posible problema de maximización en un problema de minimización, $\max \mathscr{L}(C) \iff \min l(C)$, con $l(C) = -\ln \mathscr{L}(C)$ se tiene que,

$$l(C) = -\sum_{k=1}^{K} \left[\ln(\alpha_N^{(k)}) + \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}^{(k)} \ln\left(\hat{p}_{i\cdot}^{(k)} c_{ij}\right) \right] =$$

$$l(C) = -\sum_{k=1}^{K} ln(\alpha_N^{(k)}) - \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}^{(k)} ln(c_{ij}) - \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}^{(k)} ln(\hat{p}_{i\cdot}^{(k)})$$

donde únicamente $-\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}^{(k)} ln(c_{ij})$ depende de C, y se procede a nombrar como $\tilde{l}(C)$. Por lo tanto, se tiene que,

$$\begin{aligned} \min \quad & \tilde{l}(C) \\ \text{s.a.} \quad & c_{ij} \leq 0, \quad \text{para todo } i = 1,\ldots,I, \quad j = 1,\ldots,J \\ & \sum_{j=1}^{J} c_{ij} = 1, \quad \text{para todo } i = 1,\ldots,I \end{aligned} \tag{3.1}$$

Para resolver este problema de minimización, se empleó la técnica de los multiplicadores de Lagrange para convertir el problema de minimización con restricciones en un problema de minimización sin restricciones, el

cual, puede resolverse, mediante el cálculo de las derivadas parciales y la resolución del sistema correspondiente al igualarlas todas ellas a cero. Aplicando esta técnica se obtiene que,

$$\psi(C) = \tilde{l}(C) - \sum_{i=1}^{I} \lambda_i (\sum_{j=1}^{J} c_{ij} - 1)$$

$$\min_{C} \psi(C),$$

obteniendo el siguiente resultado de las derivadas parciales,

$$\begin{aligned} &\frac{\delta \psi}{\delta c_{rs}} = 0 \Leftrightarrow -\left(\sum_{k=1}^{K} n_{is}^{(k)}\right) \frac{1}{c_{rs}} - \lambda_r = 0 \\ &-\frac{n_{rs}^{(\cdot)}}{c_{rs}} = \lambda_r \Leftrightarrow c_{rs} = -\frac{n_{rs}^{(\cdot)}}{\lambda_r} \qquad \text{para todo } r = 1,\ldots,I, \quad s = 1,\ldots,J \end{aligned} \tag{3.2}$$

$$\begin{aligned} &\frac{\delta \psi}{\delta \lambda_r} = 0 \Leftrightarrow \sum_{j=1}^{J} c_{rj} = 1 \rightarrow \\ &\quad \rightarrow 1 = \sum_{j=1}^{J} c_{rj} = \sum_{j=1}^{J} \left(\frac{n_{rj}^{(\cdot)}}{\lambda_r}\right) = -\frac{1}{\lambda_r} \sum_{j=1}^{J} n_{rj}^{(\cdot)} = -\frac{1}{\lambda_r} n_{r\cdot}^{(\cdot)} \\ &\quad = \lambda_r = -n_{r\cdot}^{(\cdot)} \qquad \text{para todo } r = 1,\ldots,I, \end{aligned} \tag{3.3}$$

ofreciendo el siguiente resultado final

$$\lambda_r = -n_{r\cdot}^{(\cdot)} \qquad \text{para todo } r = 1,...,I, \tag{3.4}$$

$$\begin{aligned} &\hat{c}_{rs} = -\frac{n_{rs}^{(\cdot)}}{\lambda_r} = \frac{n_{rs}^{(\cdot)}}{-n_{r\cdot}^{(\cdot)}} \rightarrow \\ &\rightarrow \hat{c}_{rs} = \frac{n_{rs}^{(\cdot)}}{n_{r\cdot}^{(\cdot)}} \qquad \text{para todo } r = 1,...,I, \quad s = 1,...,J \end{aligned} \tag{3.5}$$

Con todo esto, se tiene casi todo lo necesario para poder estimar la matriz de probabilidades C. Lo único que, en principio faltaría para poder estimar c_{ij}, serían las frecuencias conjuntas para cada mesa electoral $n_{ij}^{(k)}$ que son totalmente inobservables. Para ello, se procederá con el diseño del algoritmo EM que tendrá como objetivo poner solución a este inconveniente.

3.3. Contraste de la homogeneidad de transferencia de voto

Uno de los factores clave en la estimación de c_{ij} es suponer que esta es constante en cada una de las mesas electorales. Es decir, suponer que las probabilidades condicionales o probabilidades de transferencia (o no) de votos de una opción política a otra no depende del lugar, la zona o la mesa electoral donde se ejerció el voto, sino que depende exclusivamente de lo que se votó en las primeras elecciones.

Para comprobar la veracidad de esta suposición se ha diseñado el siguiente contraste de hipótesis,

$$\begin{cases} H_0\text{: } p_{j|i}^{(k)} = p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \\ H_1\text{: } p_{j|i}^{(k)} \neq p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \end{cases}$$

En este contraste, la hipótesis nula significa que las probabilidades condicionales para cada uno de los conjuntos de provincias (o Comunidades Autónomas) a comparar son iguales, es decir, no dependen del conjunto considerado, y, la hipótesis alternativa, significa que las probabilidades condicionales no son iguales para todos los conjuntos de provincias (o Comunidades Autónomas) considerados.

Para dar respuesta a este contraste de hipótesis se ha construido un diseño metodológico empleando los datos pertenecientes al Estudio 3269 "Barómetro de diciembre 2019. Postelectoral elecciones generales 2019"del Centro de Investigaciones Sociológicas. De este estudio, como se indica al inicio de este capítulo, se han seleccionado y extraído los datos de las cuestiones: (1) *Comunidad Autónoma*, (2) *Provincia*, (3) *Recuerdo de voto en las elecciones generales de abril de 2019* y (4) *Recuerdo de voto en las elecciones generales de noviembre de 2019.*

El objetivo ha sido contrastar la hipótesis de que la transferencia de votos de unas elecciones a otras no depende de la zona donde resida el individuo. Para ello, se ha seleccionado una región española que comparta las mismas opciones políticas, como puede ser la Comunidad Autónoma de Andalucía. Una vez seleccionada esta región, se ha procedido dividiéndola en dos zonas con un tamaño muestral similar (con 4 provincias cada una de ellas), y, por último, se han estudiado en cada una de las dos zonas las proporciones de transferencia del voto (probabilidades condicionales) para posteriormente contrastar si son o no similares. Esto mismo también se ha hecho comparando por un lado, la Comunidad Autónoma de Andalucía con el conjunto de Comunidades Autónomas de Asturias, Castilla-La Mancha, Castilla y León y Extremadura.

El proceso se ha realizado de la siguiente manera,

1. Limpieza y organización de los datos

Se crean dos grupos de los que interesa comparar sus transferencias de voto y se organizan en dos matrices. Estas dos matrices tienen una idéntica estructura: dos columnas que representan el voto en las elecciones generales de abril y noviembre de 2019, y, tantas filas como personas hayan participado en el cuestionario y, por lo tanto, hayan participado en ambas elecciones generales.

2. Tabla de frecuencias conjuntas

Una vez creadas las dos matrices sea crea una tabla conjunta para cada una de ellas, siendo las filas las opciones políticas de las elecciones generales de noviembre de 2019, las columnas, las opciones políticas de las elecciones generales de abril de 2019, y las celdas, las frecuencias conjuntas de ambas elecciones.

3. Simulación de nuevos datos en base a las proporciones muestrales

Sea T_1 la primera tabla conjunta creada y T_2 la segunda tabla creada, se calcula $p_{j|i} = c_{ij}$, es decir, las probabilidades condicionales para cada tabla. A continuación, se simulan dos nuevas matrices de frecuencias ,Ts_1 y Ts_2, empleando las probabilidades condicionales c_{ij}. Por último, para Ts_1 y Ts_2, se calcula de nuevo c_{ij} creando dos nuevas matrices de probabilidades condicionales Tsc_1 y Tsc_2 donde obviamente se tiene que cumplir que $\sum_{j=1}^{J} c_{ij} = 1$.

4. Diferencia y suma de cuadrados

Sea Tsc_1 y Tsc_2 las dos nuevas matrices de probabilidades condicionales, se calcula $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Tsc_{1_{ij}} - Tsc_{2_{ij}}\right)^2$, que resultará en un único valor.

5. Proceso Bootstrap

Se repiten los pasos 3 y 4 un número elevado de veces creando un proceso Bootstrap con, por ejemplo 1000 repeticiones, para así obtener 1000 valores diferentes y poder construir una función de distribución empírica.

6. Contraste de hipótesis

Si se calcula $p_{j|i} = c_{ij}$ para cada uno de los grupos que se desea comparar y posteriormente se realiza su correspondiente diferencia y suma de cuadrados, $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(c_{1_{ij}} - c_{2_{ij}}\right)^2$, se obtendrá un único valor del que no se tendrán referencias para contrastar si la diferencia entre las probabilidades condicionales entre un grupo y otro, es suficiente o no, para aceptar la hipótesis nula. Pero, con los pasos anteriormente definidos, se ha podido construir una función de distribución empírica a partir de la diferencia y suma de cuadrados de los datos simulados a partir de los reales, que servirá para poner en contexto el valor original calculado al principio de este paso. Si el valor original se encuentra en el intervalo de valores simulados en el proceso bootstrap, se aceptaría la hipótesis nula $p_{j|i}^{(k)} = p_{j|i}^{(m)} \quad \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\}$, y, por lo tanto, se podría suponer que c_{ij} no depende de la Comunidad Autónoma, provincia, circunscripción electoral o mesa electoral.

3.4. Construcción del algoritmo EM

Partiendo de la base de que el algoritmo EM, según Usurralde Casas (2017), es un método iterativo para realizar una estimación de máxima verosimilitud (MLE) de parámetros de problemas en los que existen datos perdidos, en este problema se encuentra una situación idónea para proceder con la aplicación de este algoritmo, ya que, como se ha comentado anteriormente, las frecuencias conjuntas para cada mesa electoral $n_{ij}^{(k)}$ son totalmente inobservables.

Antes de comenzar con el diseño del algoritmo debe realizarse una primera estimación inicial de las frecuencias conjuntas para cada mesa electoral $n_{ij}^{(k)}$, para así poder asociar a un problema de datos incompletos, otro problema con datos completos, estableciendo relaciones entre la verosimilitud de estos dos problemas (Usurralde Casas, 2017).

Se procede entonces de la siguiente manera,

1. Estimación inicial de las frecuencias conjuntas para cada una de las mesas electorales.

En este primer punto, se llevará a cabo una estimación inicial bajo la hipótesis (poco realista, claro está) de que las opciones políticas por las que se decanta un elector en ambas elecciones son variables independientes.

$$\hat{n}_{ij}^{(k)(0)} = n_{\cdot\cdot}^{(k)} \hat{p}_{\cdot j}^{(k)} = \frac{n_{i\cdot}^{(k)} n_{\cdot j}^{(k)}}{n_{\cdot\cdot}^{(k)}}$$

Aclarando la notación, los exponentes a los que está elevado $n_{ij}^{(k)(0)}$, se corresponden con el índice de la mesa electoral k-ésima y la iteración l-ésima correspondiente al algoritmo EM.

2. Cálculo de $\hat{n}_{ij}^{(\cdot)(0)}$ y $\hat{n}_{i\cdot}^{(\cdot)(0)}$.

$$\hat{n}_{ij}^{(\cdot)(0)} = \sum_{k=1}^{K} \hat{n}_{ij}^{(k)(0)}$$

$$\hat{n}_{i\cdot}^{(\cdot)(0)} = \sum_{s=1}^{J} \hat{n}_{is}^{(\cdot)(0)}$$

Aclarando la notación, el primer exponente de $\hat{n}_{ij}^{(\cdot)(0)}$ y $\hat{n}_{i\cdot}^{(\cdot)(0)}$ implica sumatorio en k.

3. Estimación inicial del conjunto de probabilidades condicionales c_{ij}.

$$\hat{c}_{ij}^{(0)} = \frac{\hat{n}_{ij}^{(\cdot)(0)}}{\hat{n}_{i\cdot}^{(\cdot)(0)}} \quad \text{para todo } i = 1,...,I \quad j = 1,...,J$$

El objetivo principal de estos tres pasos ha sido realizar una primera estimación, muy probablemente alejada de los valores reales, del conjunto de probabilidades condicionales. Con esto, ahora sí se tiene todo lo necesario para proceder directamente con el algoritmo EM.

Diseño algoritmo EM (Expectation-Maximization)

4. E.1. Paso de Esperanza o Expectation.

$$\tilde{n}_{ij}^{(k)(1)} = n_{\cdot\cdot}\hat{p}_{i\cdot}^{(k)} * \hat{c}_{ij}^{(0)} \quad \text{para todo } i = 1,...,I \quad j = 1,...,J \quad k = 1,...,K$$

o

$$\tilde{n}_{ij}^{(k)(1)} = n_{i\cdot}^{(k)} * \hat{c}_{ij}^{(0)} \quad \text{para todo } i = 1,...,I \quad j = 1,...,J \quad k = 1,...,K$$

5. E.2. Paso de Esperanza o Expectation. Primero se calcula $\tilde{n}_{\cdot j}^{(k)(1)}$ y, a continuación, se calculan los cocientes r_j

$$\tilde{n}_{\cdot j}^{(k)(1)} = \sum_{i=1}^{I} \tilde{n}_{ij}^{(k)(1)} \quad \text{para todo } j = 1,...,J,$$

$$r_j = \frac{n_{\cdot j}^{(k)}}{\tilde{n}_{\cdot j}^{(k)(1)}} \quad \text{para todo } j = 1,...,J$$

6. E.3. Paso de Esperanza o Expectation.

$$\hat{n}_{ij}^{(k)(1)} = r_j * \tilde{n}_{ij}^{(k)(1)} \quad \text{para todo } i = 1,...,I \quad j = 1,...,J \quad k = 1,...,K$$

7. E.4. Paso de Esperanza o Expectation.. En el final del paso de Esperanza o Expectation se calcula primero $\hat{n}_{ij}^{(\cdot)(1)}$, y, a continuación, se calcula $\hat{n}_{i\cdot}^{(\cdot)(1)}$.

$$\hat{n}_{ij}^{(\cdot)(1)} = \sum_{k=1}^{K} \hat{n}_{ij}^{(k)(1)} \quad \text{para todo } i = 1,...,I \quad j = 1,...,J,$$

$$\hat{n}_{i\cdot}^{(\cdot)(1)} = \sum_{s=1}^{J} \hat{n}_{is}^{(\cdot)(1)} \quad \text{para todo } s = 1,...,J$$

Obsérvese que, como consecuencia de los pasos 5 y 6 (Paso E.2 y Paso E.3), se tiene que los $n_{\cdot j}^{(k)}$ se usen explícitamente en la construcción de los nuevos $\hat{n}_{ij}^{(k)(1)}$, cumpliéndose además que,

$$\hat{n}_{\cdot j}^{(k)(1)} = \sum_{i=1}^{I} = r_j \sum_{i=1}^{I} \tilde{n}ij^{(k)(1)} = r_j \tilde{n}_{\cdot j}^{(k)(1)} = n_{\cdot j}^{(k)},$$

lo que significa que se produce un adecuado ajuste de los nuevos $\hat{n}_{ij}^{(k)(1)}$ con las frecuencias de las segundas elecciones, es decir, $\hat{n}_{\cdot j}^{(k)(1)} = n_{\cdot j}^{(k)}$ para todo $j = 1,..,J$.

Una vez completado el paso de Esperanza o Expectation se procede con el paso de Maximización o Maximization.

8. M.1. Paso de Maximización o Maximization.

$$\hat{c}_{ij}^{(1)} = \frac{\hat{n}_{ij}^{(\cdot)(1)}}{\hat{n}_{i\cdot}^{(\cdot)(1)}} \quad \text{para todo } i = 1,...,I, \quad j = 1,..,J$$

9. Repetición de los pasos 4-8 del algoritmo.

Repetir los pasos 4-8 del algoritmo EM para así obtener $\hat{n}_{ij}^{(k)(l)}$ y $c_{ij}^{(l)}$ a partir de $\hat{n}_{ij}^{(k)(l-1)}$ y $c_{ij}^{(l-1)}$ para $l = 2,3...,L$. Como criterio que defina el tamaño de L se ha propuesto que $L = 20000$ a no ser que el algoritmo se detenga antes siguiendo un criterio de parada específico. Este criterio de parada será que el máximo de los valores absolutos de las diferencias de los elementos de dos matrices de probabilidades condicionadas consecutivas sea menor que 0.000001, es decir, máx$(abs(c_{ij}^{(L-1)} - c_{ij}^{(L)})) < 0{,}000001$. Esto puede parecer un criterio de parada bastante exigente que pueda derivar en un gran número de iteraciones, pero, en la práctica no supone ningún problema debido a la rapidez de ejecución del algoritmo.

Una vez establecido el diseño del algoritmo EM, es momento de determinar los conjuntos de datos sobre los que se pretende trabajar. Se ha optado por trabajar con dos tipos de datos: simulados y reales.

- **Datos simulados**

A partir de ambos conjuntos de datos simulados se ha ejecutado el algoritmo 1000 veces con muestras diferentes para ambos conjuntos de datos. Con esto lo que se pretende, es estimar por Monte Carlo el error cuadrático medio de estimación de c_{ij}, obteniendo estimaciones fiables del MSE. Por otro lado, se ha estudiado la convergencia de las probabilidades condicionales a lo largo de las iteraciones del algoritmo. Lo último, sería comparar la matriz de probabilidades condicionales original con sus estimaciones.

- **Datos reales**

Los datos reales con los que se aplicará el algoritmo EM son las, anteriormente nombradas al inicio de este capítulo, elecciones generales y municipales de 2019 celebradas en el Ayuntamiento de Torrent en los meses de abril y mayo, respectivamente.

Con la ejecución del algoritmo EM sobre estos datos lo que se pretende es, (1) estudiar la matriz de probabilidades condicionales estimada y (2) estudiar la convergencia de las probabilidades condicionales a lo largo de las iteraciones del algoritmo.

Capítulo 4

Análisis de resultados

El presente capítulo tiene como finalidad mostrar y analizar los resultados obtenidos durante la investigación, los cuales se fundamentan en los objetivos diseñados al inicio del estudio. Aquí se brindará una respuesta integral a los objetivos planteados, ofreciendo una visión detallada de los hallazgos obtenidos y su relevancia en el contexto de la investigación.

Mediante un riguroso análisis de los datos y las conclusiones derivadas de ellos, se busca proporcionar una comprensión profunda y crítica de los resultados, explorando sus implicaciones en relación con la literatura existente en este estudio.

A continuación, se irá dando respuesta a cada uno de los objetivos específicos y, finalmente, se dará respuesta al objetivo general eje de esta investigación.

Objetivo específico 1. Justificar el uso de las probabilidades condicionales para estimar las probabilidades conjuntas

Este objetivo específico se ha tratado en el apartado *Notación y estructura de los parámetros* del capítulo *Metodología*. En resumen, debido a que únicamente se cuentan con las frecuencias marginales para estimar las probabilidades de celda, optar por una estimación exitosa de las probabilidades condicionales supone que sea sumamente sencillo calcular las estimaciones de las probabilidades conjuntas para los datos de ambas elecciones, ya que, simplemente, $p_{ij} = p_{i\cdot}c_{ij}$, siendo c_{ij} el conjunto de probabilidades condicionales y $p_{i\cdot}$ la probabilidad (marginal) de escoger la opción política *i* en las primeras elecciones.

Estimar las probabilidades condicionales es posible si estas se suponen como iguales a lo largo del conjunto de mesas electorales. Es decir, las probabilidades condicionales deben suponerse que no dependen del lugar, la zona o la mesa electoral donde se ejerció el voto, sino que deben depender de qué opción política se escogió en las primeras elecciones. Por lo tanto, es importante suponer esta condición para poder diseñar un algoritmo exitoso que estime la matriz de probabilidades condicionales.

Objetivo específico 2. Contrastar la hipótesis de que las probabilidades condicionales de transferencia de voto, no dependen de la provincia, circunscripción electoral o mesa electoral

Como se ha comentado en el desarrollo del anterior objetivo específico, estimar las probabilidades condicionales es posible, si estas se suponen como iguales a lo largo del conjunto de mesas electorales y únicamente dependen de la opción política elegida en las primeras elecciones. Siguiendo el proceso comentado en la metodología en el apartado *Contraste de la homogeneidad de transferencia de voto*, para comprobar la veracidad de esta suposición se ha diseñado el siguiente contraste de hipótesis,

$$\begin{cases} H_0\text{: } p_{j|i}^{(k)} = p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \\ H_1\text{: } p_{j|i}^{(k)} \neq p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \end{cases}$$

Los grupos empleados para comparar las probabilidades condicionales han sido:

- **Andalucía1**: Formado por las provincias de Cádiz, Córdoba, Jaén y Málaga.

- **Andalucía2**: Formado por las provincias de Sevilla, Granada, Almría y Huelva.

- **Grupo ACCE**: Formado por las Comunidades Autónomas de Asturias, Castilla La-Mancha, Castilla y León y Extremadura.

- **Andalucía**: Formado por el conjunto de todas las provincias de Andalucía.

En primer lugar, se realizó este contraste de hipótesis empleando el total de opciones políticas para ambos grupos a comparar en ambas elecciones generales de abril y noviembre de 2019, sin tener en cuenta que algunas de las opciones políticas presentaban frecuencias muy bajas de voto. Que estas presenten frecuencias muy bajas de voto, puede suponer que una pequeña transferencia del voto de una opción con baja frecuencia, a otra con las mismas condiciones, suponga un valor demasiado elevado en la suma de cuadrados de la diferencia de probabilidades condicionales entre un grupo y otro.

Tabla 4.1: *Frecuencias conjuntas datos completos Andalucía1*

	PP	PSOE	Cs	VOX	UP	No derecho	O.P.	Blanco	Abstención	No recuerda	N.C
PP	44	3	2	1	0	0	0	0	4	1	0
PSOE	0	80	4	0	1	1	1	1	5	1	0
Cs	1	1	17	0	0	0	0	0	1	1	0
Pacma	1	0	0	0	0	0	0	0	0	0	0
VOX	5	3	5	20	1	0	0	1	5	1	0
UP	0	4	0	0	25	0	0	0	0	0	0
Más País	0	1	1	0	0	0	0	0	0	0	0
O.P.	0	0	0	0	0	0	1	0	0	0	0
Blanco	1	0	1	0	0	0	0	2	0	0	0
Abstención	7	7	1	0	3	0	0	2	31	1	2
No recuerda	1	2	1	0	0	0	0	0	7	3	0
N.C	1	1	0	0	0	0	0	0	2	34	0

Tabla 4.2: *Frecuencias conjuntas datos completos Andalucía2*

	PP	PSOE	Cs	VOX	UP	No derecho	O.P.	Blanco	Abstención	No recuerda	N.C
PP	23	2	5	1	0	0	1	0	0	2	0
PSOE	0	93	4	0	3	0	0	1	5	0	0
Cs	0	0	18	0	0	0	0	0	1	0	0
Pacma	0	0	1	0	0	0	4	1	0	0	0
VOX	3	1	1	15	0	0	0	1	2	0	0
UP	0	2	0	0	24	0	0	0	2	1	0
Más País	0	0	1	0	0	0	0	0	0	0	0
O.P.	0	0	0	0	0	0	1	0	0	0	0
Blanco	0	1	1	0	1	0	0	2	2	0	0
Abstención	0	12	1	0	2	1	0	0	34	1	1
No recuerda	0	0	0	0	0	0	0	0	2	6	0
N.C	0	3	0	0	2	0	0	0	11	2	33

Por ejemplo, si se observan las frecuencias conjuntas recopiladas en las tablas 4.1 y 4.2 para Andalucía1 y Andalucía2, se puede ver que algunas de las filas presentan un tamaño muestral demasiado pequeño, lo que significa que un recuento de uno o dos votos en las celdas de esa fila, derivan en probabilidades condicionales demasiado grandes. Esto puede suponer que al ser comparadas con las probabilidades condicionales del otro grupo, la existencia de un voto más o un voto menos en una celda determinada, pueda suponer una gran diferencia en $\sum_{i=1}^{I}\sum_{j=1}^{J}(C1_{ij}-C2_{ij})^2$, siendo C_1 las probabilidades condicionales del grupo 1 y C_2 las probabilidades condicionales del grupo 2.

Comparación entre Andalucía1 y Andalucía2. Datos completos.

La visualización de las probabilidades condicionales para Andalucía1 y Andalucía2 puede observarse en las siguientes dos tablas:

Tabla 4.3: *Probabilidades condicionales datos completos Andalucía1*

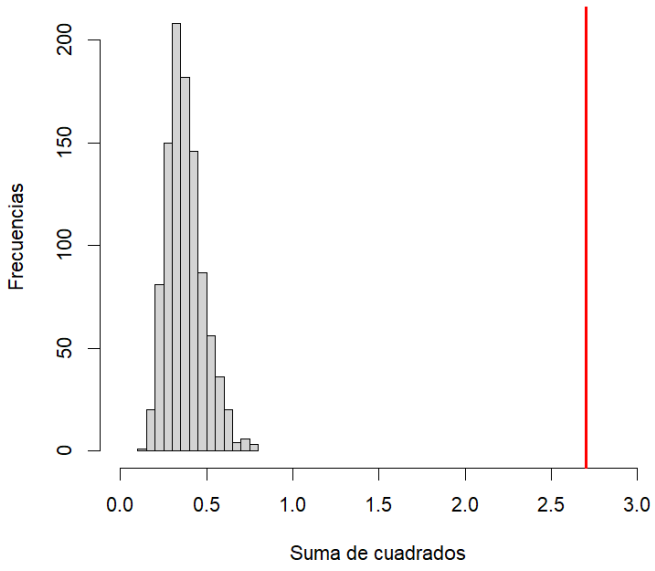
	PP	PSOE	Cs	VOX	UP	No derecho	O.P.	Blanco	Abstención	No recuerda	N.C
PP	0.800	0.055	0.036	0.018	0.000	0.000	0.000	0.000	0.073	0.018	0.000
PSOE	0.000	0.851	0.043	0.000	0.011	0.011	0.011	0.011	0.053	0.011	0.000
Cs	0.048	0.048	0.810	0.000	0.000	0.000	0.000	0.000	0.048	0.048	0.000
Pacma	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VOX	0.122	0.073	0.122	0.488	0.024	0.000	0.000	0.024	0.122	0.024	0.000
UP	0.000	0.138	0.000	0.000	0.862	0.000	0.000	0.000	0.000	0.000	0.000
Más País	0.000	0.500	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O.P.	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Blanco	0.250	0.000	0.250	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.000
Abstención	0.130	0.130	0.019	0.000	0.056	0.000	0.000	0.037	0.574	0.019	0.037
No recuerda	0.071	0.143	0.071	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.214
N.C	0.026	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.053	0.895

Tabla 4.4: *Probabilidades condicionales datos completos Andalucía2*

	PP	PSOE	Cs	VOX	UP	No derecho	O.P.	Blanco	Abstención	No recuerda	N.C
PP	0.676	0.059	0.147	0.029	0.000	0.000	0.029	0.000	0.000	0.059	0.000
PSOE	0.000	0.877	0.038	0.000	0.028	0.000	0.000	0.009	0.047	0.000	0.000
Cs	0.000	0.000	0.947	0.000	0.000	0.000	0.000	0.000	0.053	0.000	0.000
Pacma	0.000	0.000	0.167	0.000	0.000	0.000	0.667	0.167	0.000	0.000	0.000
VOX	0.130	0.043	0.043	0.652	0.000	0.000	0.000	0.043	0.087	0.000	0.000
UP	0.000	0.069	0.000	0.000	0.828	0.000	0.000	0.000	0.069	0.034	0.000
Más País	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O.P.	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Blanco	0.000	0.143	0.143	0.000	0.143	0.000	0.000	0.286	0.286	0.000	0.000
Abstención	0.000	0.231	0.019	0.000	0.038	0.019	0.000	0.000	0.654	0.019	0.019
No recuerda	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.750	0.000
N.C	0.000	0.059	0.000	0.000	0.039	0.000	0.000	0.000	0.216	0.039	0.647

Realizando $\sum_{i=1}^{I}\sum_{j=1}^{J}(C_{1_{ij}} - C_{2_{ij}})^2$ para ambas tablas 4.3 y 4.4 el resultado refleja un valor de 2.702453. Llevando a cabo el proceso bootstrap de simulación de valores se ha construido el siguiente histograma en el que se ha introducido el valor anteriormente calculado 2.702453.

Figura 4.1: *Histograma de valores bootstrap. Andalucía1 - Andalucía2.*

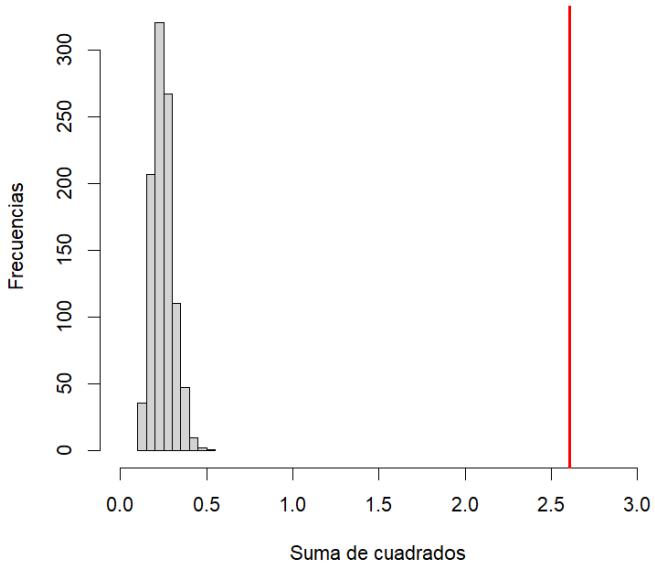


Como se puede observar, en la Figura 4.1 la línea roja que representa el valor 2.702453 se encuentra muy distanciado del conjunto de valores bootstrap obtenidos mediante simulación. Esto podría indicar que las probabilidades condicionales son distintas entre un grupo y otro.

Comparación entre Andalucía y el grupo ACCE. Datos completos.

Incluso si se lleva a cabo la misma comparación de las probabilidades condicionales entre el grupo Andalucía y el grupo ACCE el resultado es similar al que nos hemos encontrado en la Figura 4.1.

Figura 4.2: *Histograma de valores bootstrap. Andalucía - ACCE.*



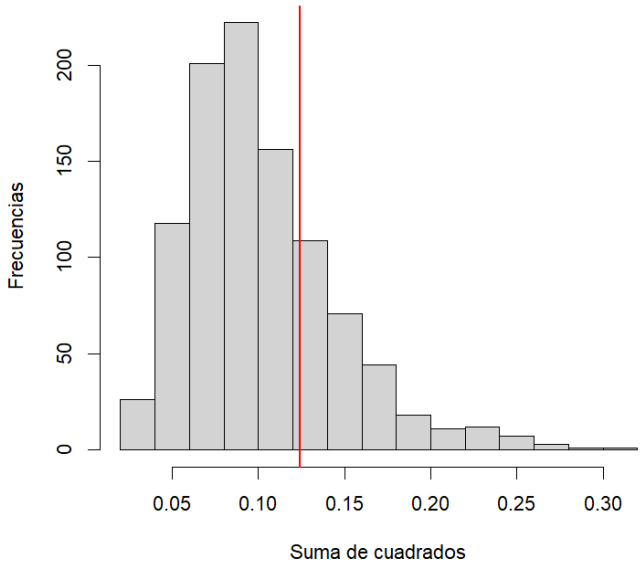
En la figura 4.2 se puede observar que la línea vertical roja que representa el valor original se encuentra también excesivamente alejado del histograma.

Llegados a este punto, conviene preguntarse que sucedería si se eliminan de los datos, aquellas categorías de respuesta que presentan bajas frecuencias como: ”PACMA”, ”Más País”, Otros Partidos”(O.P.), ”Voto en blanco”, ”No recuerda”, ”No tenía derecho a voto”y ”N.C.”. ¿Las probabilidades condicionales continuarán siendo significativamente distintas entre grupos?

Comparación entre Andalucía1 y Andalucía2. Datos reducidos.

Una vez reducido el conjunto de datos, eliminando aquellas categorías de respuesta que presentasen bajas frecuencias en ambas elecciones generales, los resultados cambian por completo. Si se vuelve a realizar el proceso bootstrap simulando datos a partir del nuevo conjunto reducido y se representa el resultado en un histograma junto con el valor original de $\sum_{i=1}^{I}\sum_{j=1}^{J}(C1_{ij}-C2_{ij})^2$, el resultado es el siguiente.

Figura 4.3: *Histograma de valores bootstrap. Andalucía1 - Andalucía2.*

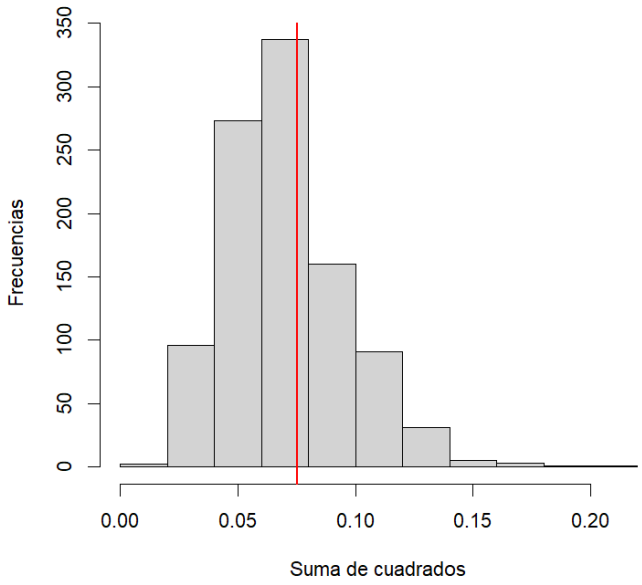


En la Figura 4.3 se observa como la línea roja que representa $\sum_{i=1}^{I}\sum_{j=1}^{J}(C1_{ij} - C2_{ij})^2$, se encuentra entre los datos simulados a través del proceso bootstrap. Esto señala que, en el nuevo conjunto de datos reducido, las probabilidades condicionales para ambos grupos pueden aceptarse como iguales.

Comparación entre Andalucía y grupo ACCE. Datos reducidos.

Repitiendo el proceso realizado para alcanzar el histograma anterior pero esta vez comparando las probabilidades condicionales entre Andalucía y el grupo de Comunidades Autónomas compuesto por Asturias, Castilla La-Mancha, Castilla y León y Extremadura se obtiene el siguiente histograma.

Figura 4.4: *Histograma de valores bootstrap. Andalucía - Grupo ACCE.*



En el histograma se muestra como $\sum_{i=1}^{I}\sum_{j=1}^{J}(C_{1_{ij}} - C_{2_{ij}})^2$, se sitúa más o menos en la mitad del intervalo de valores obtenidos mediante el proceso bootstrap.

Con esta información, mas con la información obtenida al realizar el proceso de comparación entre Andalucía1 y Andalucía2, observando que la línea roja para ambos histogramas se sitúa entre el intervalo de valores bootstrap, se puede aceptar la hipótesis nula de que, para datos electorales estructurados en tablas de contingencia con frecuencias marginales suficientemente grandes,

$$p_{j|i}^{(k)} = p_{j|i}^{(m)} \quad \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\},$$

es decir, se acepta la hipótesis nula de que las probabilidades condicionales no dependen de la Comunidad Autónoma, provincia, circunscripción electoral o mesa electoral donde se ejerció la decisión política de votar o no votar.

Objetivo específico 3. Diseñar y ejecutar al menos dos algoritmos diferentes para estimar las probabilidades condicionales

Este objetivo específico se ha tratado en el apartado de *Metodología*. En el desarrollo de los siguientes objetivos se mostrarán los resultados correspondientes a cada uno de los algoritmos.

Objetivo específico 4. Estudiar la convergencia de las probabilidades condicionales a lo largo de las iteraciones del algoritmo y su tiempo de ejecución

Como se ha explicado en el capítulo de *Metodología*, la convergencia de las probabilidades condicionales a lo largo del algoritmo EM se ha estudiado a partir de dos conjuntos de datos simulados y un conjunto de datos reales.

En primer lugar, se va a estudiar la convergencia para los conjuntos de datos simulados, y en segundo lugar, se estudiará la convergencia para el conjunto de datos real.

Datos simulados. Conjunto 1.

Una vez simulados los datos del Conjunto 1 se lleva a cabo la ejecución del algoritmo EM. Empleando el criterio de parada establecido, el algoritmo se ha detenido en la iteración número 1539 y ha tardado 1.17 segundos aproximadamente.

La primera iteración del algoritmo ha tenido como resultado la siguiente matriz de probabilidades condicionales.

Tabla 4.5: *Matriz probabilidades condicionales en la primera iteración del algoritmo EM*

	Partido 1	Partido 2	Partido 3
Partido 1	0.4376239	0.3024133	0.2599629
Partido 2	0.4168510	0.3118060	0.2713430
Partido 3	0.4088434	0.3036731	0.2874835

Mientras que la última iteración ha devuelto el siguiente resultado final.

Tabla 4.6: *Matriz probabilidades condicionales en la última iteración del algoritmo EM*

	Partido 1	Partido 2	Partido 3
Partido 1	0.6126582	0.2119134	0.1754284
Partido 2	0.3077105	0.5506888	0.1416007
Partido 3	0.3007021	0.1417654	0.5575324

Como era de esperar existe una gran diferencia entre la matriz de probabilidades condicionales de la Tabla 4.5 y la Tabla 4.6. Para estudiar la convergencia de cada uno de los valores de estas dos tablas, se han creado tres gráficos que muestran como han convergido hacia el valor final, los valores de la Tabla 4.5. En estos, se muestra en forma de línea, la convergencia del conjunto de elementos que forman la matriz de probabilidades condicionales. También se muestra el valor que da inicio al algoritmo para cada uno de los elementos, y su correspondiente valor final. Por último, una línea discontinua muestra, aproximadamente, a partir de qué iteración, los valores comienzan a ser estables.

Gráfico 4.5: *Convergencia de las probabilidades condicionales si se ha votado al Partido 1 en las primeras elecciones.*

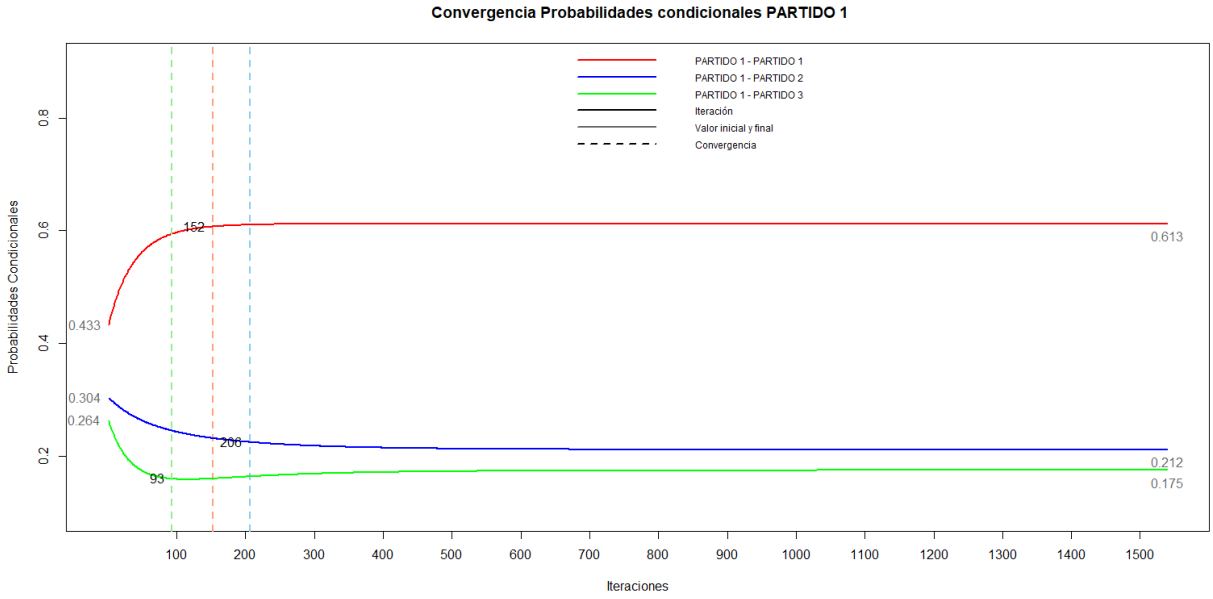


Gráfico 4.6: *Convergencia de las probabilidades condicionales si se ha votado al Partido 2 en las primeras elecciones.*

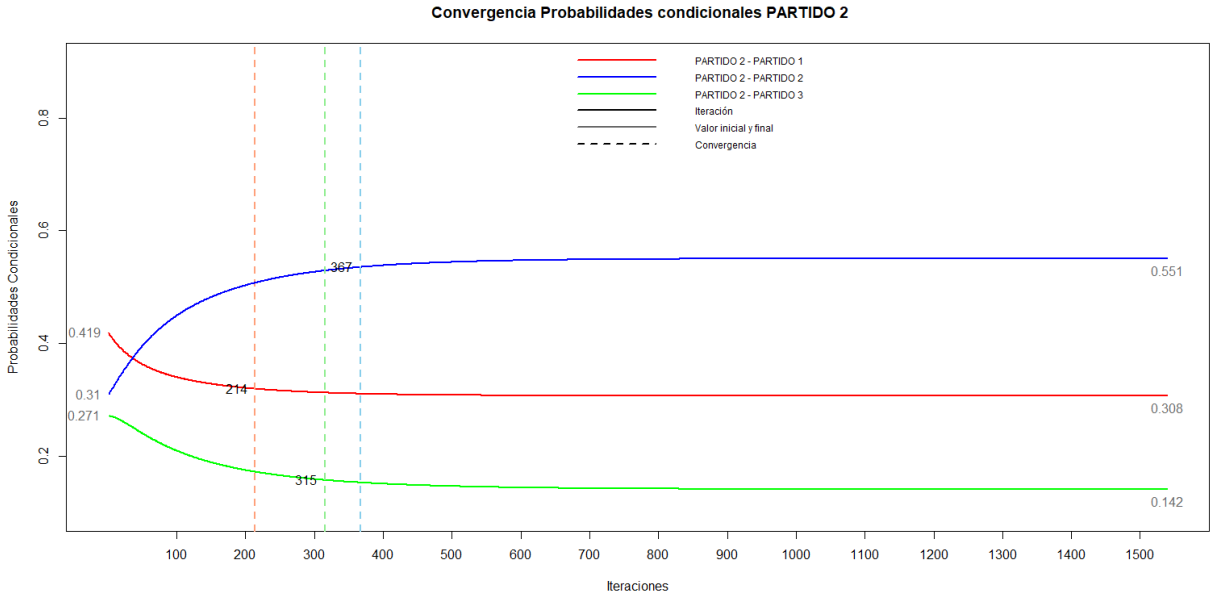
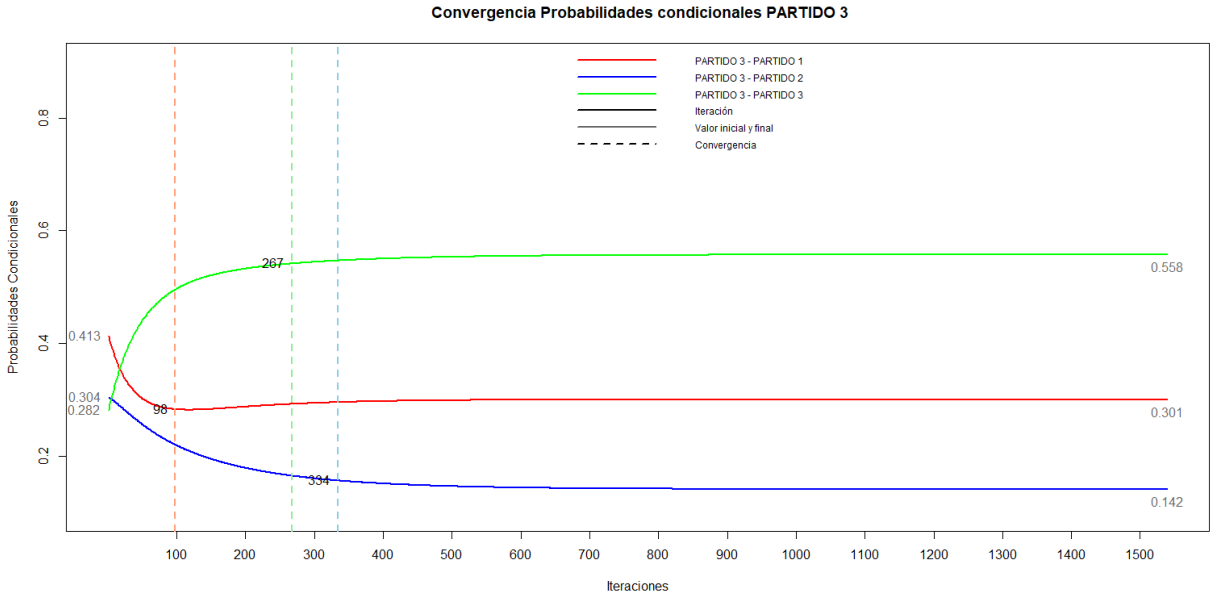


Gráfico 4.7: *Convergencia de las probabilidades condicionales si se ha votado al Partido 3 en las primeras elecciones.*



En el Gráfico 4.5, se representa la convergencia de los valores de la primera fila de la matriz de probabilidades condicionales, en el Gráfico 4.6 se representa la segunda fila, y en el Gráfico 4.7 se representa la última fila. Se puede observar que, para los tres gráficos, en líneas generales, la convergencia de los valores se hace más fuerte (es decir, los valores empiezan a ser cada vez más próximos al valor final del algoritmo) antes de, aproximadamente, la iteración 400. Además, los valores que consiguen converger antes son aquellos que se encuentran más próximos del valor con el que se inicia el algoritmo, como se puede observar claramente en el Gráfico 4.6 y el Gráfico 4.7.

Datos simulados. Conjunto 2.

Ejecutando el algoritmo EM para el Conjunto 2 de datos simulados, este se ha detenido en la iteración número 10209 y ha tardado 20.18 segundos aproximadamente.

La primera iteración del algoritmo ha tenido como resultado la siguiente matriz de probabilidades condicionales.

Tabla 4.7: *Matriz de probabilidades en la primera iteración del algoritmo EM*

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.2518729	0.1526847	0.2799819	0.1712994	0.08808633	0.05607479
VOX	0.2506410	0.1677642	0.2684733	0.1629685	0.09218712	0.05796590
PSOE	0.2191880	0.1316737	0.3156112	0.1745395	0.09835734	0.06063041
PODEM	0.2340695	0.1410285	0.2979685	0.1764027	0.09216945	0.05836142
COMUN	0.2463928	0.1395392	0.2938505	0.1810039	0.08280145	0.05641205
CS	0.2165840	0.1390533	0.3071012	0.1662115	0.11128290	0.05976703
O.P.	0.2230522	0.1442176	0.2941846	0.1683946	0.09575958	0.07439138

Mientras que la última iteración ha devuelto el siguiente resultado final.

Tabla 4.8: *Matriz de probabilidades condicionales en la última iteración del algoritmo EM*

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.765186812	0.123687335	0.002933183	0.041261132	0.065911375	0.001020163
VOX	0.111423010	0.758097343	0.034978925	0.000002188	0.048019171	0.047479364
PSOE	0.066978028	0.008915874	0.695265992	0.165773525	0.000111411	0.062955169
PODEM	0.006070905	0.000258268	0.387763654	0.529971056	0.000010718	0.075925399
COMUN	0.009430216	0.015995408	0.475301943	0.449685038	0.017862032	0.031725363
CS	0.077017172	0.156900522	0.061379340	0.000000000	0.694224599	0.010478367
O.P.	0.153068466	0.077749531	0.157500431	0.129940752	0.123047504	0.358693316

Al igual que para el Conjunto 1, existe una gran diferencia entre la matriz de probabilidades condicionales de la Tabla 4.7 y la Tabla 4.8. Para estudiar la convergencia de los valores de estas dos tablas, lo que se ha hecho es, crear un gráfico donde se muestre la convergencia de cinco valores con $c_{ij} > 0{,}15$, y, crear otro gráfico donde se muestre la convergencia de otros cinco valores con $c_{ij} \leq 0{,}15$. En estos gráficos, se muestra en forma de línea, la convergencia de los correspondientes valores y se muestra en forma de línea discontinua, el momento en el que los valores comienzan a estabilizarse.

Gráfico 4.8: *Convergencia de* $c_{ij} > 0{,}15$.

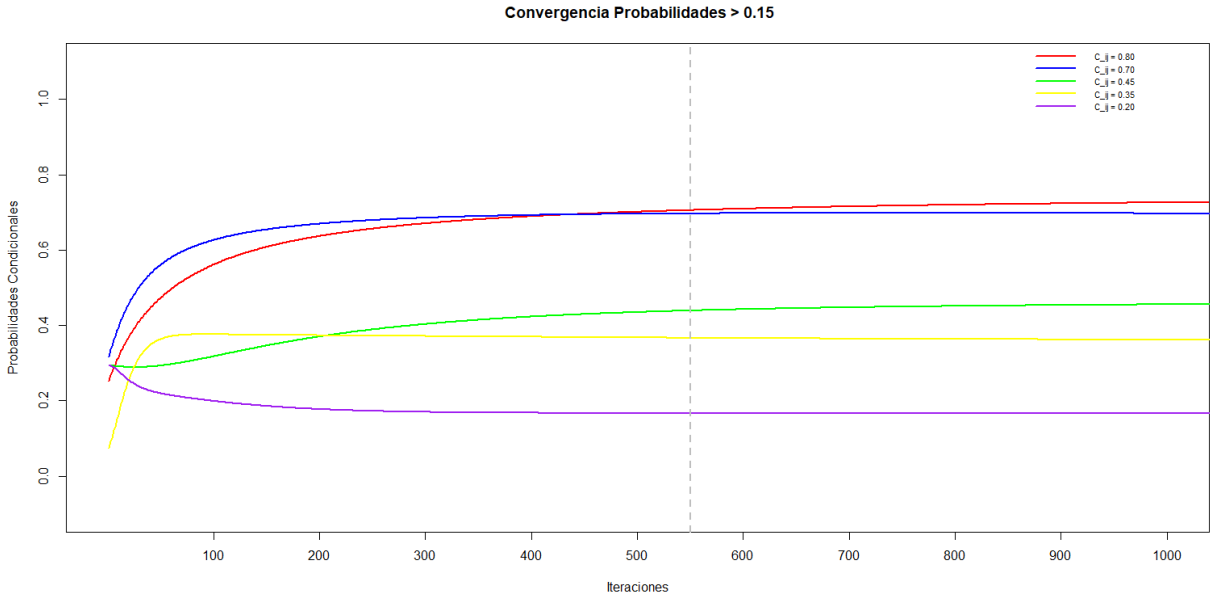
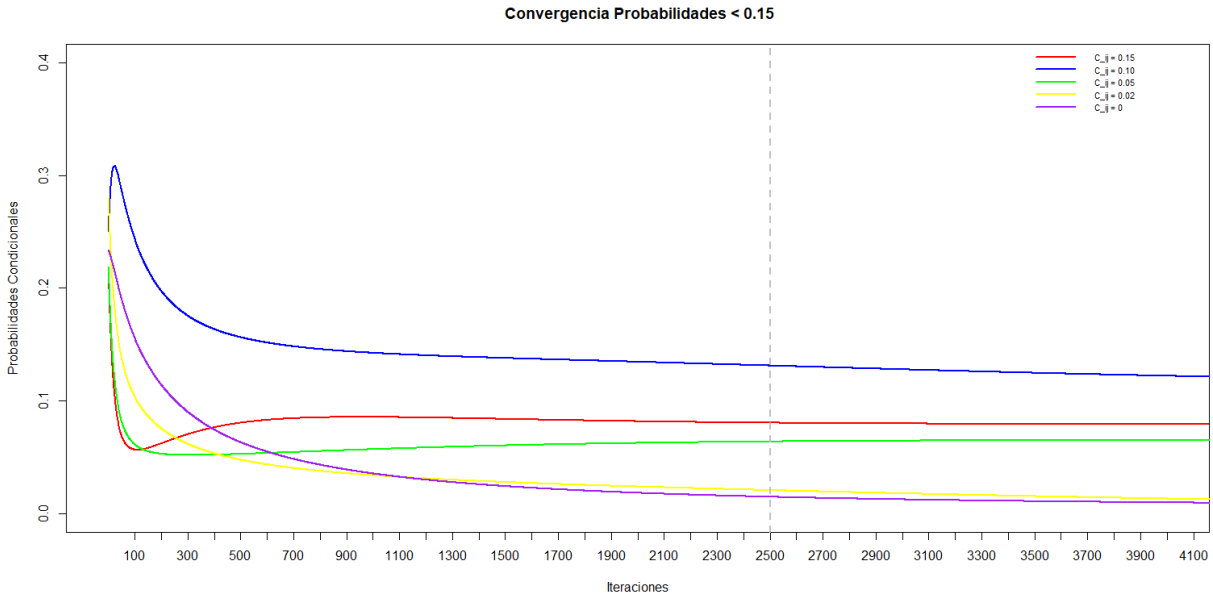


Gráfico 4.9: *Convergencia de* $c_{ij} \leq 0{,}15$.



Comparando un gráfico con otro, se puede ver que, las probabilidades condicionales menores o iguales que 0.15 presentan, en general, una convergencia más lenta que las probabilidades condicionales mayores que 0.15. Aproximadamente, para el Grafico 4.8 los valores presentan una convergencia más fuerte entorno a la iteración 550, mientras que para el Gráfico 4.9, la estabilidad de los valores comienza aproximadamente entorno a la iteración 2500.

Datos reales. Ayuntamiento de Torrent.

Ejecutando el algoritmo EM para los datos del ayuntamiento de Torrent, este se ha detenido en la iteración número 8852 y ha tardado aproximadamente 3.70 minutos.

La primera iteración del algoritmo ha tenido como resultado la siguiente matriz de probabilidades condicionales.

Tabla 4.9: *Matriz de probabilidades condicionales en la primera iteración del algoritmo EM*

	COMPROMIS	PSOE	PP	PODEM	C.s	VOX	O.P.	Blanco o Nulo	Abstencion
COMPROMIS	0.0543	0.2322	0.1638	0.0219	0.0585	0.0470	0.0259	0.0072	0.3891
PSOE	0.0469	0.2510	0.1475	0.0229	0.0509	0.0388	0.0251	0.0068	0.4101
PP	0.0480	0.2252	0.1907	0.0198	0.0513	0.0448	0.0263	0.0073	0.3866
PODEM	0.0527	0.2381	0.1459	0.0235	0.0568	0.0420	0.0258	0.0071	0.4081
C.s	0.0529	0.2326	0.1583	0.0213	0.0624	0.0451	0.0255	0.0073	0.3947
VOX	0.0516	0.2280	0.1659	0.0216	0.0575	0.0467	0.0255	0.0073	0.3958
O.P.	0.0497	0.2377	0.1493	0.0232	0.0554	0.0431	0.0250	0.0072	0.4094
Blanco o Nulo	0.0471	0.2381	0.1584	0.0223	0.0529	0.0418	0.0267	0.0082	0.4044
Abstencion	0.0448	0.2308	0.1515	0.0223	0.0470	0.0391	0.0246	0.0066	0.4332

Esta Tabla 4.9 dista mucho de lo que sería la matriz de probabilidades condicionales de la última iteración

del algoritmo EM.

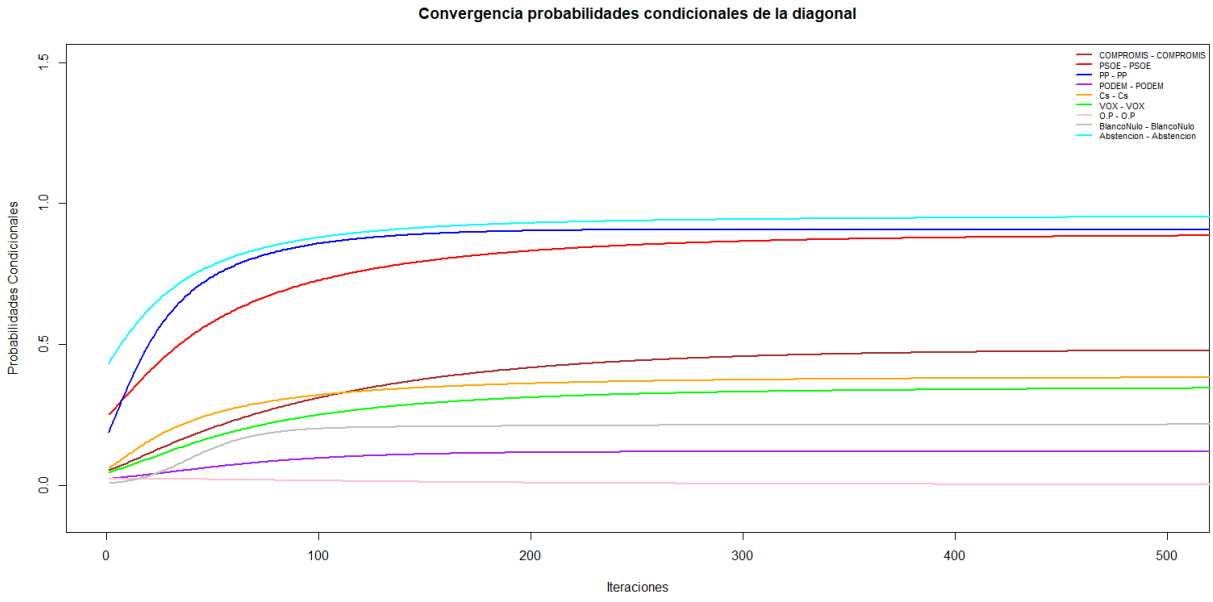
Tabla 4.10: *Matriz de probabilidades condicionales en la última iteración del algoritmo EM*

	COMPROMIS	PSOE	PP	PODEM	C.s	VOX	O.P.	Blanco o Nulo	Abstencion
COMPROMIS	0.4817	0.2933	0.0000	0.0000	0.000	0.1696	0.0554	0.0000	0.0000
PSOE	0.0000	0.8980	0.0544	0.0231	0.000	0.0000	0.0164	0.0012	0.0068
PP	0.0000	0.0319	0.9096	0.0000	0.000	0.0000	0.0568	0.0017	0.0000
PODEM	0.2178	0.0000	0.0000	0.1186	0.000	0.0000	0.1012	0.0000	0.5624
C.s	0.0665	0.1005	0.0000	0.0000	0.385	0.0000	0.0006	0.0072	0.4401
VOX	0.0190	0.1398	0.2614	0.0000	0.000	0.3529	0.0019	0.0180	0.2070
O.P.	0.0000	0.0049	0.0000	0.1399	0.000	0.0027	0.0000	0.0637	0.7888
Blanco o Nulo	0.0000	0.4149	0.0000	0.0000	0.000	0.0000	0.2155	0.2163	0.1532
Abstencion	0.0000	0.0000	0.0120	0.0127	0.000	0.0000	0.0013	0.0000	0.9740

Se puede observar que existe una gran diferencia entre la Tabla 4.9 y la Tabla 4.10, sobre todo en la cantidad de ceros que resultan en la última iteración del algoritmo. Para estudiar la convergencia de los valores de estas dos tablas, se han creado dos gráficos que representan, por un lado, la convergencia de los valores de la diagonal de la matriz de probabilidades condicionales, y, por el otro lado, la convergencia de algunos de los valores próximos a cero.

El gráfico correspondiente a los valores de la diagonal de la matriz de probabilidades condicionales se ha reducido en el eje de ”Iteraciones”, ya que, a partir de, aproximadamente, la iteración 300, hasta la última iteración, los valores se estabilizan formando prácticamente una recta.

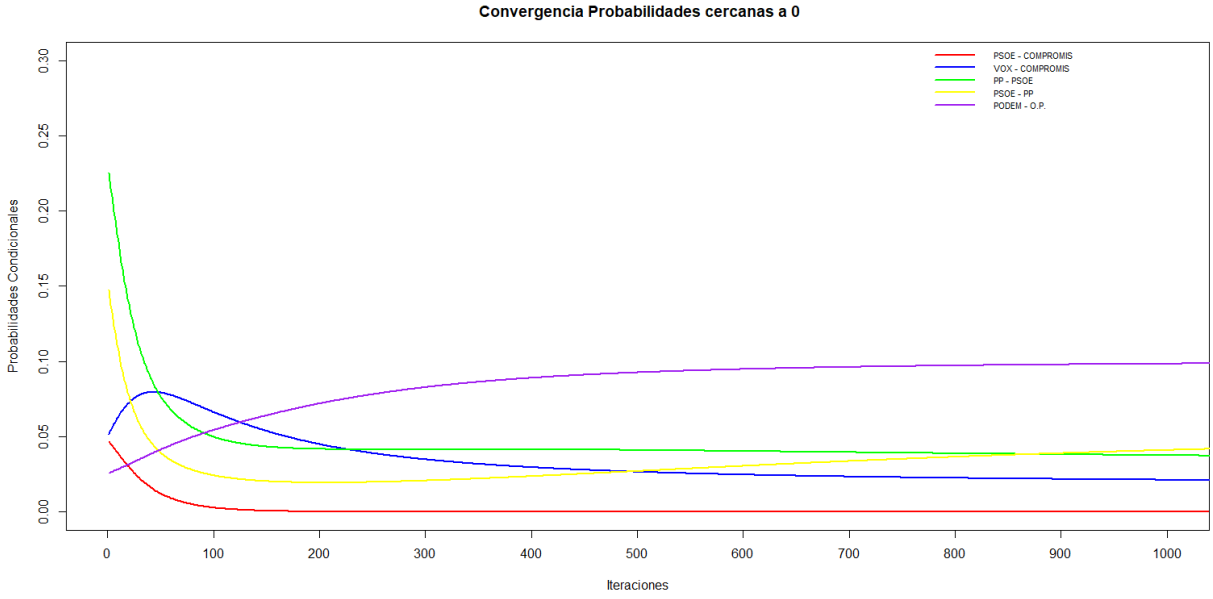
Gráfico 4.10: *Convergencia de la diagonal de la matriz de probabilidades condicionales.*



En el Gráfico 4.10 puede observarse que prácticamente todos los elementos de la diagonal logran estabilizarse entorno a la iteración 300. Sin duda, los elementos que convergen antes son los correspondientes a ”Otros

Partidos - Otros Partidos" con un valor de prácticamente 0 en la matriz final de probabilidades condicionales, a "PODEM - PODEM" con un valor de aproximadamente 0.1186 también en esta matriz y a "Voto Blanco o Nulo - Voto Blanco o Nulo" con un valor de aproximadamente 0.2163. El resto de los elementos de la diagonal, presentan una convergencia un poco más lenta en comparación.

Gráfico 4.11: *Convergencia de elementos cercanos a 0 de la matriz de probabilidades condicionales.*



Las probabilidades condicionales del Gráfico 4.11 necesitan un mayor número de iteraciones para estabilizarse en comparación con las del Gráfico 4.10. Estas logran estabilizarse entorno a la iteración 600, menos el elemento "PSOE - PP"que comienza a estabilizarse entorno a la iteración 900.

Aunque de la sensación de que, para ambos gráficos, las probabilidades condicionales se estabilicen antes de, aproximadamente, la iteración 300 o 900, dependiendo el caso, el algoritmo no termina pasadas las 8000 iteraciones debido al criterio de parada tan exigente que se ha seleccionado. Aun así, esto no es problema debido a la rapidez de ejecución del algoritmo, que apenas ha tardado aproximadamente 3.70 minutos en completarse.

En resumen, puede afirmarse que para los 3 conjuntos de datos la convergencia no es extremadamente rápida, ya que necesita alrededor de entre 400 y 900 iteraciones, o incluso 2500, como se muestra en el Gráfico 4.9, para comenzar a estabilizarse entorno a su valor final. Sin embargo, a pesar del gran número de iteraciones que necesita el algoritmo para finalizar, el tiempo de ejecución es bastante corto. Se necesitó de apenas un segundo para el Conjunto 1 de datos simulados, de 20 segundos para el Conjunto 2 (ambos con únicamente 10 mesas electorales), y, de aproximadamente 3.70 minutos para el conjunto de datos reales del Ayuntamiento de Torrent, con 99 mesas electorales.

Objetivo específico 5. Estimar mediante Monte Carlo el error cuadrático medio (MSE) de estimación de las probabilidades condicionales para cada conjunto de datos simulados

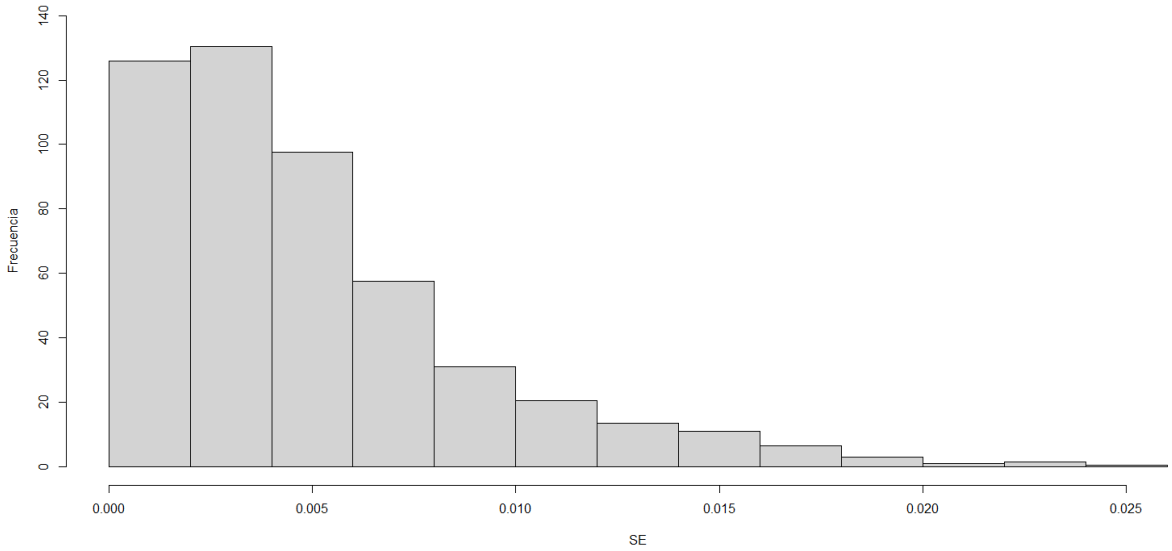
Para cada uno de los conjuntos de datos simulados se han obtenido estimaciones del error cuadrático o SE (o ASE, nombrado en inglés como *average squared error*) de las dos estimaciones de las matrices de probabilidades condicionales fijadas en cada uno de estos mediante el método de Monte Carlo. Para estudiar el SE se ha optado por la visualización de su conjunto de valores mediante la representación en un histograma y el cálculo

de sus estadísticos descriptivos básicos (media, mediana, mínimo y máximo). Con este cálculo de la media, en los estadísticos descriptivos, lo que se obtendría inmediatamente sería el MSE (o MASE, nombrado en inglés como *mean average squared error*) o error cuadrático medio.

Datos simulados. Conjunto 1

La representación en histograma del conjunto de valores estimados del SE para estos datos ha sido la siguiente,

Gráfico 4.12: *Estimación mediante Monte Carlo del SE para el Conjunto 1 de datos simulados*



Se puede observar en el Gráfico 4.12 como la gran mayoría de los errores cuadráticos se sitúan próximos a 0, exactamente entre 0 y 0.010. Observando los estadísticos descriptivos se tiene lo siguiente.

Tabla 4.11: *Resumen estadísticos descriptivos valores SE. Conjunto 1 de datos simulados*

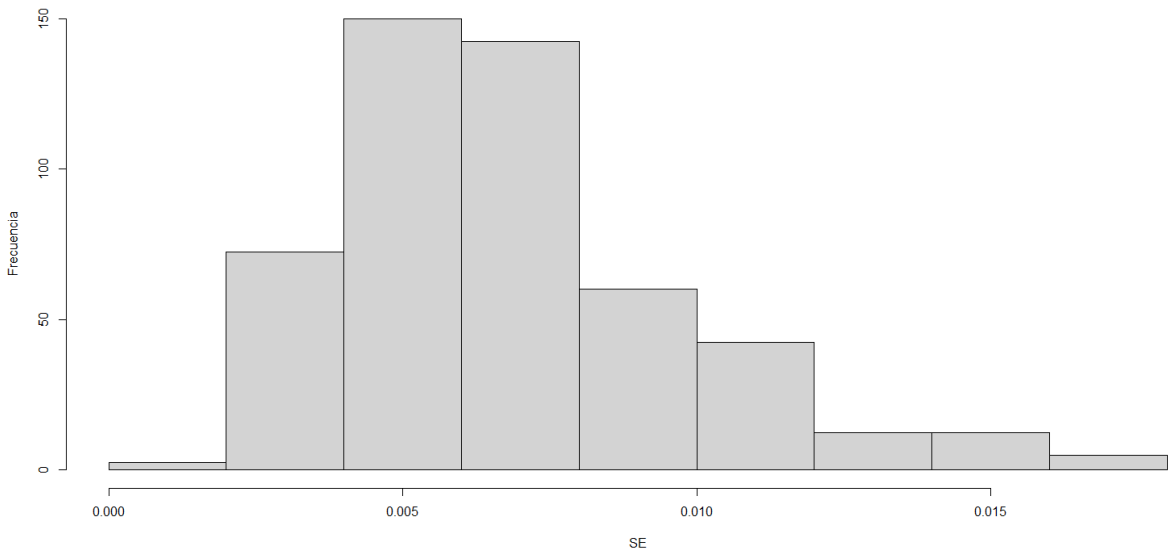
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000093	0.001987	0.003913	0.004963	0.006558	0.024118

En la Tabla 4.11 puede observarse que la media de los errores cuadrático (MSE) es de 0.004963, el error cuadrático mínimo es prácticamente 0, el máximo equivale a 0.24, aproximadamente, y la mediana recibe un valor de 0.0039, aproximadamente.

Datos simulados. Conjunto 2

La representación en histograma del conjunto de valores estimados del SE para estos datos ha sido la siguiente.

Gráfico 4.13: *Estimación mediante Monte Carlo del SE para el Conjunto 12 de datos simulados*



Se puede observar en el Gráfico 4.13 como el conjunto de errores cuadráticos se encuentra algo más alejado de 0. Observando los estadísticos descriptivos se tiene lo siguiente.

Tabla 4.12: *Resumen estadísticos descriptivos valores SE. Conjunto 2 de datos simulados*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001487	0.004659	0.006260	0.006779	0.008075	0.016364

Si comparamos esta Tabla 4.12 con la Tabla 4.11 se puede observar que el error cuadrático mínimo y la mediana son notablemente superiores, pero, para el caso del error cuadrático máximo, el valor de la Tabla 4.11, 0.024, aproximadamente, es mayor que el valor de la Tabla 4.12, 0.016, aproximadamente. Si se atiende al valor del error cuadrático medio (MSE), para estos datos, recibe un valor de 0.006260, superior al error cuadrático medio en el Conjunto 1 de datos simulados de la Tabla 4.11.

Por lo tanto, se puede afirmar que, cuanto mayor es el tamaño de la matriz de probabilidades condicionales, mayor es el error cuadrático medio , encontrándose la distribución de los errores cuadráticos más alejada de 0.

Objetivo general. Estimar las probabilidades de celda de una tabla de contingencia dadas únicamente sus frecuencias marginales

Como se ha comentado en el diseño metodológico de esta investigación, $p_{ij} = p_{i\cdot}c_{ij}$ para todo $i = 1,..,I$ y $j = 1,...,J$. Por ello, una exitosa estimación de la matriz de probabilidades condicionales, fija a lo largo de todas las mesas electorales, bastaría para obtener una estimación fiable de las probabilidades conjuntas o de celda en cada una de las mesas electorales. Por lo tanto, como el interés de esta investigación es lograr la estimación de las probabilidades condicionales, la resolución de este objetivo general se centrará exclusivamente en este hecho, dejando en segundo plano la estimación de las probabilidades conjuntas $\hat{p}_{ij}$, las cuales, sí que dependen de la mesa electoral.

En primer lugar, para garantizar que el algoritmo EM realiza buenas estimaciones de c_{ij}, se va a proceder a comentar los resultados de las estimaciones para los conjuntos de datos simulados. Como en estos conjuntos de datos simulados se parte de la base de que ya se conoce c_{ij}, porque se ha fijado con antelación, bastaría simplemente con llevar a cabo su estimación y compararla con el parámetro real.

Datos simulados. Conjunto 1.

Repasando, este Conjunto 1 de datos simulados está formado por 3 opciones políticas para las primeras elecciones, 3 para las segundas y 10 mesas electorales, es decir, $I = 3$, $J = 3$ y $K = 10$.

La matriz de probabilidades condicionales real para este caso puede observarse en la Tabla 3.1, que será impresa de nuevo a continuación.

Tabla 3.1: *Matriz de probabilidades condicionales. Conjunto 1*

	Partido 1	Partido 2	Partido 3
Partido 1	0.65	0.15	0.20
Partido 2	0.25	0.60	0.15
Partido 3	0.30	0.20	0.50

Una vez simulados los datos para las dos elecciones a partir de esta matriz de probabilidades de la Tabla 3.1, se ejecuta el algoritmo EM y se obtiene el siguiente resultado final.

Tabla 4.13: *Estimación de la matriz de probabilidades condicionales para el Conjunto 1*

	Partido 1	Partido 2	Partido 3
Partido 1	0.6126582	0.2119134	0.1754284
Partido 2	0.3077105	0.5506888	0.1416007
Partido 3	0.3007021	0.1417654	0.5575324

Observando los resultados elemento a elemento de la matriz podría afirmarse que, para un número tan pequeño de mesas electorales (únicamente 10), la estimación ha resultado exitosa. Las celdas que más se asemejan a su parámetro real han sido las correspondientes a ”Partido 3 - Partido 1”, ”Partido 2 - Partido 3”, ”Partido 1 - Partido 3ƶ ”Partido 1 - Partido 1”. Por otro lado, la celda que más se ha alejado del parámetro real ha sido la correspondiente a ”Partido 1 - ”Partido 2”, con aproximadamente un error de estimación de 0.06, siendo el valor real 0.15 y el valor estimado 0.21, aproximadamente.

Datos simulados. Conjunto 2.

De nuevo, este Conjunto 2 de datos simulados está formado por 7 opciones políticas para las primeras elecciones, 6 para las segundas y 10 mesas electorales, es decir, $I = 7$, $J = 6$ y $K = 10$.

La matriz de probabilidades condicionales real para este caso puede observarse en la Tabla 3.5, que será nuevamente impresa a continuación.

Tabla 3.5: *Matriz de probabilidades condicionales. Conjunto 2*

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.80	0.10	0.02	0.00	0.05	0.03
VOX	0.12	0.75	0.00	0.03	0.05	0.05
PSOE	0.05	0.05	0.70	0.15	0.03	0.02
PODEM	0.00	0.00	0.30	0.65	0.00	0.05
COMUN	0.00	0.00	0.45	0.50	0.00	0.05
CS	0.15	0.10	0.05	0.00	0.65	0.05
O.P.	0.15	0.10	0.20	0.05	0.15	0.35

Una vez simulados los datos para las dos elecciones a partir de esta matriz de probabilidades de la Tabla 3.5, se ejecuta el algoritmo EM y se obtiene el siguiente resultado final.

Tabla 4.14: *Estimación de la matriz de probabilidades condicionales para el Conjunto 2*

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.765186812	0.123687335	0.002933183	0.041261132	0.065911375	0.001020163
VOX	0.111423010	0.758097343	0.034978925	0.000002188	0.048019171	0.047479364
PSOE	0.066978028	0.008915874	0.695265992	0.165773525	0.000111411	0.062955169
PODEM	0.006070905	0.000258268	0.387763654	0.529971056	0.000010718	0.075925399
COMUN	0.009430216	0.015995408	0.475301943	0.449685038	0.017862032	0.031725363
CS	0.077017172	0.156900522	0.061379340	0.000000000	0.694224599	0.010478367
O.P.	0.153068466	0.077749531	0.157500431	0.129940752	0.123047504	0.358693316

Debido a la gran cantidad de celdas en la Tabla 3.5, y por consiguiente en la Tabla 4.14, se va a proceder a crear una nueva tabla con la diferencia entre ambas para visualizar mejor los resultados.

Tabla 4.15: Diferencia entre la matriz de probabilidades condicionales real y estimada

	PP	VOX	PSOE	PODEM	CS	O.P.
PP	0.034813188	-0.023687335	0.017066817	-0.04126113	-0.015911375	0.028979837
VOX	0.008576990	-0.008097343	-0.034978925	0.02999781	0.001980829	0.002520636
PSOE	-0.016978028	0.041084126	0.004734008	-0.01577353	0.029888589	-0.042955169
PODEM	-0.006070905	-0.000258268	-0.087763654	0.12002894	-0.000010718	-0.025925399
COMUN	-0.009430216	-0.015995408	-0.025301943	0.05031496	-0.017862032	0.018274637
CS	0.072982828	-0.056900522	-0.011379340	0.00000000	-0.044224599	0.039521633
O.P.	-0.003068466	0.022250469	0.042499569	-0.07994075	0.026952496	-0.008693316

Observando las diferencias entre los valores reales y los estimados, puede afirmarse que, para un número de mesas electorales tan pequeño (únicamente 10), la estimación de la matriz de probabilidades condicionales ha resultado exitosa. únicamente 6 celdas de 42 presentan un error mayor a 0.05, siendo la que mas error presenta la celda ”PODEM - PODEM”con un error de 0.12 aproximadamente, siendo el valor real 0.65, y su estimación 0.53, aproximadamente. En cuanto al resto de elementos de la matriz de probabilidades condicionales, es destacable la exitosa estimación de los parámetros.

Datos reales. Ayuntamiento de Torrent.

Los datos del Ayuntamiento de Torrent están sintetizados en 9 opciones políticas para las elecciones generales (celebradas en primer lugar), 9 opciones para las elecciones municipales (celebradas en segundo lugar), y, 99 mesas electorales, es decir, $I = 9$, $J = 9$ y $K = 99$.

Como es obvio, para este caso no se puede emplear la matriz de probabilidades condicionales real para compararla con su estimación, ya que es totalmente desconocida. Aun así, la exitosa estimación en los anteriores dos ejemplos, a pesar de únicamente contar con 10 mesas electorales, ofrece garantías de que el método funciona y estima correctamente. Con todo esto, el resultado de la estimación de c_{ij} para las elecciones generales de abril y las elecciones municipales de mayo, celebradas ambas en 2019 en el Ayuntamiento de Torrent ha sido la siguiente:

Tabla 4.16: *Estimación de la matriz de probabilidades condicionales para los datos de Torrent*

	COMPROMIS	PSOE	PP	PODEM	C.s	VOX	O.P.	Blanco o Nulo	Abstencion
COMPROMIS	0.4817	0.2933	0.0000	0.0000	0.000	0.1696	0.0554	0.0000	0.0000
PSOE	0.0000	0.8980	0.0544	0.0231	0.000	0.0000	0.0164	0.0012	0.0068
PP	0.0000	0.0319	0.9096	0.0000	0.000	0.0000	0.0568	0.0017	0.0000
PODEM	0.2178	0.0000	0.0000	0.1186	0.000	0.0000	0.1012	0.0000	0.5624
C.s	0.0665	0.1005	0.0000	0.0000	0.385	0.0000	0.0006	0.0072	0.4401
VOX	0.0190	0.1398	0.2614	0.0000	0.000	0.3529	0.0019	0.0180	0.2070
O.P.	0.0000	0.0049	0.0000	0.1399	0.000	0.0027	0.0000	0.0637	0.7888
Blanco o Nulo	0.0000	0.4149	0.0000	0.0000	0.000	0.0000	0.2155	0.2163	0.1532
Abstencion	0.0000	0.0000	0.0120	0.0127	0.000	0.0000	0.0013	0.0000	0.9740

Observando la Tabla 4.16, todo parece indicar que se ha realizado una estimación con éxito debido a la coherencia que muestra el resultado de la estimación. Como se puede observar, en la diagonal de la Tabla 4.16, que hace referencia a la fidelidad de los votantes, todas presentan valores elevados, a excepción de ”PODEM - PODEM”con una estimación de 0.1186 y de ”Otros Partidos - Otros Partidos”con una estimación de 0. Por otro lado, los dos principales partidos políticos en España, PP y PSOE, son los que conservan en mayor medida la fidelidad de sus votantes. También, si se observa la transferencia de votos entre partidos con similar espectro político, las transferencia de votos son más elevadas que con aquellos con los que su espectro político es opuesto o lejano. Un ejemplo podría ser la transferencia de votos entre VOX y el PP, en el que, sabiendo que un ciudadano que ha votado a VOX en las primeras elecciones tiene una probabilidad estimada de 0.2614 de votar al PP en las segundas elecciones. Otro ejemplo sería el caso de transferencia de votos entre COMPROMÍS y el PSOE, con una probabilidad de 0.2933 de votar al PSOE en las segundas sabiendo que se votó a COMPROMÍS en las primeras. Por último, destacar que el elemento con una probabilidad más cercana a 1 en esta matriz se corresponde con la celda ”ABSTENCIÓN - ABSTENCIÓN”, lo que significa que, sabiendo que un ciudadano se ha abstenido en las elecciones generales, hay aproximadamente un 97 % de probabilidades de que se abstenga en las posteriores elecciones municipales.

Capítulo 5

Consideraciones finales

En este capítulo, se presentan las conclusiones obtenidas durante de la investigación realizada. Se analizan los resultados obtenidos y se extraen las principales implicaciones y hallazgos relevantes. Además, se exploran las posibilidades de futuros estudios que podrían profundizar en aspectos no abordados en esta investigación, proporcionando así nuevas perspectivas y conocimientos en el campo de estudio.

Asimismo, se aborda un apartado dedicado a evaluar las fortalezas y debilidades que han surgido durante la redacción de este trabajo de fin de máster. Se examinan los aspectos positivos que han contribuido al desarrollo y la calidad de la investigación, así como las limitaciones o desafíos encontrados que podrían haber influido en los resultados obtenidos.

Este capítulo constituye una etapa crucial en el proceso de investigación, ya que permite consolidar y reflexionar sobre los resultados y su relevancia en el contexto del estudio. Asimismo, ofrece una visión crítica y constructiva sobre el trabajo realizado, identificando áreas de mejora y brindando una base sólida para futuros trabajos en esta área de investigación.

A través de estas conclusiones, las perspectivas de investigaciones futuras y la evaluación de fortalezas y debilidades, se proporciona un cierre significativo a esta investigación y se abre la puerta a nuevos horizontes en la búsqueda del conocimiento en este campo.

5.1. Conclusiones

Tras el análisis realizado para dar respuesta a cada uno de los objetivos que ha guiado el transcurso de esta investigación se puede concluir lo siguiente de cada uno de ellos.

- ***Justificar el uso de las probabilidades condicionales para estimar las probabilidades conjuntas***

  En un contexto en el que únicamente son conocidas las frecuencias marginales en una tabla de contingencia de doble entrada, el cálculo de las frecuencias conjuntas y, por lo tanto, de las probabilidades conjuntas, se complica. El cálculo de las frecuencias marginales es posible a partir de las frecuencias conjuntas, lo contrario no podría hacerse.

  Por ello, ante los datos de dos elecciones generales consecutivas, en las que únicamente se conocen $n_{i\cdot}^{(k)}$ y $n_{\cdot j}^{(k)}$, existiendo tantas tablas de contingencia como mesas electorales (k), optar por una estimación exitosa de c_{ij} supone que sea sumamente sencillo calcular las estimaciones de p_{ij} para los datos de ambas elecciones, ya que, simplemente, $p_{ij} = p_{i\cdot}c_{ij}$.

  Además, estimar las probabilidades condicionales es posible si estas se suponen como iguales a lo largo del conjunto de mesas electorales, es decir, debe suponerse que c_{ij} no depende de la Comunidad

Autónoma, provincia, circunscripción electoral o mesa electoral donde se ejerció el voto, sino que deben depender de qué opción política se escogió en las primeras elecciones.

- ***Contrastar la hipótesis de que las probabilidades condicionales de transferencia de voto, no dependen de la provincia, circunscripción electoral o mesa electoral***

Para poder cotejar esta cuestión se diseñó el siguiente contraste de hipótesis,

$$\begin{cases} H_0\colon p_{j|i}^{(k)} = p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \\ H_1\colon p_{j|i}^{(k)} \neq p_{j|i}^{(m)} & \forall i \in \{1,...,I\}, \quad \forall j \in \{1,...,J\}, \quad \forall km \in \{1,...,K\} \end{cases}$$

Mediante la construcción de un coherente diseño metodológico y el uso de unos datos específicos adaptados a resolver este contraste de hipótesis, se llegó a la conclusión de aceptar la hipótesis nula de que, para datos electorales estructurados en tablas de contingencia con frecuencias marginales suficientemente grande, las probabilidades condicionales c_{ij} no dependen de la Comunidad Autónoma, provincia, circunscripción electoral o mesa electoral.

- ***Diseñar un algoritmo de esperanza-maximización (EM) para la estimación de las probabilidades condicionales***

El diseño del algoritmo de esperanza-maximización se ha implementado en el software estadístico R, resultando en un gran éxito, como se comentará en las conclusiones de los siguientes objetivos de la investigación.

- ***Estudiar la convergencia de las probabilidades condicionales a lo largo de las iteraciones del algoritmo y su tiempo de ejecución***

Los resultados muestran que la convergencia de las probabilidades condicionales a lo largo del algoritmo EM no es extremadamente rápida, sino que tarda entre 300 y 900 iteraciones en comenzar a estabilizarse entorno a su valor final. A pesar de lograr esta estabilidad llegado ese número de iteraciones, el algoritmo no se detiene en ninguno de los casos antes de las 1500 iteraciones, como en el caso del Conjunto 1 de datos simulados, incluso alcanzando las 8800 en el conjunto de datos reales del Ayuntamiento de Torrent o las 10200 para el Conjunto 2 de datos simulados.

En cuanto al tiempo de ejecución del algoritmo, pasa todo lo contrario que con su convergencia, este es realmente rápido. Apenas basta 1 segundo para el primer conjunto de datos simulados, 20 segundos para el Conjunto 2 y 3.70 minutos para los datos reales del Ayuntamiento de Torrent. Lo que realmente ralentiza el tiempo de ejecución del algoritmo es el número de mesas electorales, cuanto mayor sea, mayor será el tiempo de ejecución.

- ***Estimar mediante Monte Carlo el error cuadrático medio (MSE) de estimación de las probabilidades condicionales para cada conjunto de datos simulados***

Para cada uno de los conjuntos de datos simulados se han obtenido estimaciones del error cuadrático o SE y del error cuadrático medio o MSE, de las dos estimaciones de las matrices de probabilidades condicionales fijadas en cada uno de estos mediante el método de Monte Carlo. La conclusión alcanzada fue que, cuanto mayor es el tamaño de las matrices condicionales, es decir, cuanto mayor es el número de celdas de la matriz, mayor es el error cuadrático medio y la distribución de los errores cuadráticos se encuentra más alejada de 0.

- ***Estimar las probabilidades de celda de una tabla de contingencia dadas únicamente sus frecuencias marginales***

  En el análisis de resultados, únicamente se estimaron las probabilidades condicionales para cada uno de los conjuntos de datos, demostrando un gran éxito en la estimación de los parámetros para los datos simulados y una gran coherencia en la estimación de las probabilidades condicionales para los datos del Ayuntamiento de Torrent.

  En este apartado de conclusiones, se va a realizar la estimación de las probabilidades de celda para los resultados electorales del Ayuntamiento de Torrent para la primera de las mesas electorales. Para ello, se necesita $\hat{p}_{i\cdot}^{(1)}$ y $\hat{c}_{ij}$.

Tabla 5.1: *Estimación probabilidades marginales elecciones generales mesa electoral 1*

COMPROMIS	PSOE	PP	PODEM	C.s	VOX	O.P.	Blanco o Nulo	Abstención
0.0429	0.1545	0.1931	0.0901	0.0837	0.1159	0.0150	0.0215	0.2833

Empleando los datos de la Tabla 5.1 de probabilidades marginales y los datos de la Tabla 4.16 de probabilidades condicionales, se realiza $p_{ij}^{(1)} = p_{i\cdot}^{(1)} c_{ij}$ para todo $i = 1,...,I$ y para todo $j = 1,...,J$. El resultado ha sido el siguiente,

Tabla 5.2: *Estimación probabilidades conjuntas para ambas elecciones mesa electoral 1*

	COMPROMIS	PSOE	PP	PODEM	C.s	VOX	O.P.	Blanco o Nulo	Abstencion
COMPROMIS	0.0207	0.0126	0.0000	0.0000	0.0000	0.0073	0.0024	0.0000	0.0000
PSOE	0.0000	0.1387	0.0084	0.0036	0.0000	0.0000	0.0025	0.0002	0.0011
PP	0.0000	0.0062	0.1757	0.0000	0.0000	0.0000	0.0110	0.0003	0.0000
PODEM	0.0196	0.0000	0.0000	0.0107	0.0000	0.0000	0.0091	0.0000	0.0507
C.s	0.0056	0.0084	0.0000	0.0000	0.0322	0.0000	0.0001	0.0006	0.0368
VOX	0.0022	0.0162	0.0303	0.0000	0.0000	0.0409	0.0002	0.0021	0.0240
O.P.	0.0000	0.0001	0.0000	0.0021	0.0000	0.0000	0.0000	0.0010	0.0118
Blanco o Nulo	0.0000	0.0089	0.0000	0.0000	0.0000	0.0000	0.0046	0.0046	0.0033
Abstencion	0.0000	0.0000	0.0034	0.0036	0.0000	0.0000	0.0004	0.0000	0.2759

Debido a que se está presenciando en la Tabla 5.2 una tabla de contingencia con probabilidades conjuntas, $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Su interpretación es muy sencilla, por ejemplo, si se atiende a la celda "Abstención - Abstención", de todos los votantes que participaron en ambas elecciones, un 27.59 % escogieron esta opción política en ambas.

5.2. Futuros estudios

El inicio de este trabajo fin de máster estuvo marcado por el planteamiento de dos métodos para la estimación en una tabla de contingencia de las probabilidades de celdas dadas únicamente sus marginales. Por un lado, el algoritmo EM, que es el que se ha logrado desarrollar en esta investigación, y por el otro lado, un algoritmo basado en la expresión comúnmente empleada en el contexto de la regresión lineal, utilizada para calcular los coeficientes de regresión mediante el método de mínimos cuadrados, $(XX^t)^{-1}XY$.

Debido principalmente a la falta de tiempo ocasionada por un presunto inconveniente en la implementación del código del algoritmo, aunque no es descartable la existencia de algún tipo de error en el desarrollo teórico del método, únicamente se ha sacado adelante la implementación del algoritmo EM para la estimación en una tabla de contingencia de las probabilidades de celda dadas únicamente sus marginales.

Queda pendiente una futura investigación para resolver los problemas que ha generado el desarrollo de este método, y así, poder comparar la eficacia de ambos y llegar a una conclusión de cuál puede ser preferible en cuestiones de convergencia, tiempo de ejecución y estimación de los parámetros, en general, o según la naturaleza de los datos.

5.3. Fortalezas y debilidades

Una de las fortalezas destacadas de este trabajo fin de máster es la relación directa que existe entre el tema seleccionado y el abordado en mi trabajo de fin de grado, titulado ”Internet y su impacto en el comportamiento electoral de los jóvenes”. Esta continuidad temática permitió aprovechar los conocimientos adquiridos durante el grado y aplicarlos en el desarrollo de esta investigación.

Además, haber cursado este máster interuniversitario proporcionó los fundamentos teóricos y metodológicos necesarios para llevar a cabo este trabajo de manera rigurosa y precisa. Los conocimientos adquiridos en el máster fueron una base sólida para comprender y aplicar los métodos estadísticos requeridos en el análisis de la tabla de contingencia estudiada.

Otra fortaleza significativa fue la excelente guía proporcionada por el director de este trabajo de fin de máster, Ricardo José Cao Abad, catedrático de Universidad. Su dedicación personalizada, atención cuidadosa y explicaciones claras fueron fundamentales para orientar y enriquecer el desarrollo de la investigación. La dirección experta y la retroalimentación constante brindaron un apoyo invaluable para alcanzar los objetivos planteados y garantizar la calidad del trabajo.

Una de las principales debilidades encontradas durante el desarrollo de este proyecto fue el desconocimiento previo del algoritmo EM. Esto implicó dedicar tiempo adicional a estudiar y comprender en profundidad dicho algoritmo para lograr su implementación adecuada en esta investigación. A pesar de este desafío inicial, se logró adquirir el conocimiento necesario y pudo aplicarse exitosamente.

Otra limitación importante fue el escaso tiempo disponible para llevar a cabo esta investigación, con un plazo de apenas 3 meses, desde principios de marzo hasta principios de junio. Esta restricción temporal representó un desafío adicional en términos de la planificación y ejecución de todas las etapas del proyecto. Sin embargo, a pesar de este contratiempo, se logró cumplir con los objetivos establecidos y obtener resultados relevantes.

Además, al provenir de una licenciatura en Sociología, surgieron dificultades relacionadas con el uso de herramientas específicas como el software estadístico R o Látex. Sin embargo, a lo largo de estos dos años de aprendizaje, se pudo superar exitosamente estas limitaciones y resolver los desafíos planteados, adquiriendo las habilidades necesarias para utilizar estas herramientas de manera efectiva en el desarrollo de la investigación.

Por último, durante la redacción de esta investigación, identifiqué una de mis debilidades particulares: mi limitada soltura y habilidad en el manejo de la notación matemática. A lo largo del proceso, me percaté de que enfrentaba dificultades al expresar de manera precisa y fluida conceptos y fórmulas matemáticas relevantes para mi investigación. Esta falta de soltura en la notación matemática no solo me generaba inseguridad al transmitir mis ideas de forma clara, sino que también afectaba mi capacidad para desarrollar argumentos sólidos y coherentes. Reconozco que, a causa de esta debilidad, mis explicaciones y derivaciones matemáticas podrían haber carecido de la claridad y rigurosidad necesarias, lo cual pudo haber impactado la calidad general de esta investigación. Sin embargo, esta experiencia me ha permitido comprender la importancia de mejorar mis habilidades en la notación matemática, y me ha motivado a trabajar en el perfeccionamiento de esta destreza fundamental para futuros proyectos académicos.

Bibliografía

[1] Agresti A (2007) Contingency tables. In *An Introduction to Categorical Data Analysis*. Wiley, Florida.

[2] Ayuntamiento de Torrent (2019a) Resultados elecciones generales 2019. Disponible en: https://datos.gob.es/es/catalogo/l01462444-resultats-provisionals-eleccions-generals-2019-resultados-provisionales-elecciones-autonomicas-2019.

[3] Ayuntamiento de Torrent (2019a) Resultados elecciones municipales 2019. Disponible en: https://datos.gob.es/es/catalogo/l01462444-resultats-eleccions-municipals-2019-resultados-elecciones-municipales-2019.

[4] Centro de Investigaciones Sociológicas (2011) Barómetro de diciembre 2019. Postelectoral elecciones generales 2019. Disponible en:$https://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=$ 14479. Estudio 3269.

[5] Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977) Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistic Society. Series B (Methodological)*, 39(1):1-38.

[6] El País (2019a) Resultados electorales generales 2019. *El País*. Disponible en: $https://resultados.elpais.com/elecciones/2019-28A/generales/congreso/$.

[7] El País (2019b) Resultados electorales municipales 2019. *El País*. Disponible en: $https://resultados.elpais.com/elecciones/2019/municipales/$.

[8] La Moncloa (2020) Gobierno de la XIV Legislatura. *La Moncloa*. Disponible en: $https://www.lamoncloa.gob.es/gobierno/gobiernosporlegislaturas/Paginas/xiv_legislatura.aspx$.

[9] Lipsitz, S. R. y Zhao, L.(1994) Estimation in contingency tables with given marginals. *Journal of the Royal Statistic Society. Series D (The Statistician)*, 43(2):223-230.

[10] Martín, Jesica(2019) España se asoma a sus cuartas elecciones generales en cuatro años y a su segunda repetición electoral. RTVE. Disponible en: https://www.rtve.es/noticias/20190917/espana-se-asoma-cuartas-elecciones-generales-cuatro-anos-su-segunda-repeticion-electoral/1979378.shtml.

[11] Miller, L. (2020) Polarización en España: más divididos por ideología e identidad que por políticas públicas. *EsadeEcPol Insight*. 18:14

[12] Pelz, W. y Good, I. (1986) Estimating probabilities from contingency tables when the marginal probabilities are known, by using additive objective functions. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 35(1):50

[13] Usurralde Casas, B (2017) Algoritmo EM, Aplicaciones y extensiones.

[14] Wang, X., Lim, J., Kim, S.-J., y Hahn, K. S. (2015) Estimating cell probabilities in contingency tables with constraints on marginals/conditionals by geometric programming with applications. *Computational Statistics*, 30:107.129