



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN
OPERATIVA APLICADAS Y CALIDAD

ANÁLISIS DE LA INCERTIDUMBRE EN LA ESTIMACIÓN DE LA MOVILIDAD ELECTORAL MEDIANTE EL PROCEDIMIENTO LPHOM

Programa de Doctorado en Estadística y Optimización

PRESENTADA POR:
JORGE MARTÍN MARÍN



DIRIGIDA POR:
RAFAEL ROMERO VILAFRANCA
JOSE MANUEL PAVÍA MIRALLES

TUTORIZADA POR:
ELENA VÁZQUEZ BARRACHINA

VALENCIA, septiembre de 2020

AGRADECIMIENTOS

Quiero agradecer, en primer lugar, el apoyo recibido por mi familia durante todo este largo proceso y durante todos los instantes de mi vida.

En segundo lugar, y no menos importante, a mi mentor y Director, Rafa, si no hubiera sido por su amable y constante insistencia e interés, esta tesis no habría salido a la luz. A veces, uno tiene la suerte de encontrarse con maravillosas personas en su vida y, desde luego, yo he tenido ese privilegio con Rafa.

También quiero agradecer la ayuda, consejos y trabajo de mi Co-Director, Jose Manuel, siempre diligente para proponer y realizar mejoras en la investigación; ha sido un verdadero placer tenerle como director.

A mi tutora Elena, por su apoyo y confianza constante en mí, por ser siempre positiva conmigo y estar pendiente de mi

Y, por último, a mis compañeros del Departamento, por su apoyo para que esto llegara a su fin.

RESUMEN

La estimación del trasvase de votos de unos partidos a otros entre dos procesos electorales es un tema relevante dentro de la sociología política. Dadas las limitaciones de las encuestas en este campo, existe un interés creciente en los procedimientos de estimación que utilizan exclusivamente la realidad incuestionable de los propios resultados electorales. *lphom* es uno de estos métodos. *lphom* es un procedimiento, relativamente sencillo, que, tras ser aplicado en estos últimos años en algunos estudios reales, ha proporcionado, en todos los casos, resultados muy razonables.

En la presente tesis se aborda el problema, hasta ahora no investigado, de desarrollar procedimientos para estimar los márgenes de incertidumbre de los resultados obtenidos al aplicar este procedimiento en estudios reales. Tras definir unos indicadores adecuados, EI y EP, para cuantificar dicha incertidumbre, el proceso se basa en utilizar los residuos del modelo *lphom* para (i) evaluar el grado de incumplimiento de la hipótesis de homogeneidad en la que se sustenta *lphom*, y (ii) investigar, mediante un conjunto de estudios de simulación desarrollados en diferentes escenarios, las relaciones entre estos indicadores.

Las conclusiones generales de estos estudios permiten desarrollar unos procedimientos para estimar los valores previsibles de EI y EP en estudios reales, así como los límites de confianza superior para los mismos. También se desarrolla un procedimiento para obtener intervalos de confianza para las diferentes probabilidades de transición de votos p_{jk} estimadas mediante *lphom*. Los métodos desarrollados se ilustran aplicándolos a la estimación del trasvase de votos entre la primera y la segunda vuelta de las elecciones presidenciales francesas de 2017.

Se indican finalmente una serie de líneas futuras de investigación sugeridas por los resultados de esta tesis.

RESUM

L'estimació del transvasament de vots d'uns partits a altres entre dos processos electorals és un tema rellevant dins de la sociologia política. Donades les limitacions de les enquestes en aquest camp, hi ha un creixent interès en els procediments d'estimació que utilitzen exclusivament la realitat inqüestionable dels propis resultats electorals. En aquest sentit, *lphom* és un mètode senzill que després de ser aplicat aquests últims anys en alguns estudis reals, ha proporcionat en tots els casos resultats molt razonables.

En la següent tesi, s'aborda el problema, encara no investigat, de desenvolupar procediments per a estimar els margens d'incertidumbre dels resultats obtinguts al aplicar aquest procediment en els estudis reals. Després d'haber definit uns indicadors adequats, EI y EP, per a quantificar dita incertidumbre, el procés es basa en utilitzar els residus del model *lphom* per a (i) evaluar el grau de incompliment de la hipòtesis d'homogeneïtat en què es sustenta *lphom*, i (ii) investigar, mitjançant un conjunt d'estudis de simulacions desenvolupades en diferents escenaris, les relacions entre aquests indicadors.

Les conclusions generals d'aquests estudis permeten desenvolupar uns procediments per a estimar els valors previsibles de EI y EP en estudis reals, així com els límits de confiança superior per als mateixos. També s'aconsegueix un procediment per a obtenir intervals de confiança per a les diferents probabilitats de transició de vots p_{jk} estimats mitjançant *lphom*. Els mètodes desenvolupats s'il·lustren aplicats a l'estimació del transvasament de vots entre la primera y segona volta de les eleccions presidencials franceses de 2017.

S'indiquen finalment una sèrie de línies futures d'investigació suggerides pels resultats d'aquesta tesi.

ABSTRACT

The estimation of the transfer of votes from some political parties to another between two electoral processes is a relevant issue within political sociology. Given the limitations of the surveys in this field, there is a growing interest in estimation procedures that exclusively use the unquestionable reality of the electoral results themselves. In this sense, *lphom* is a simple method that, after being applied in recent years in some real studies, has provided, in all cases, very reasonable results.

This thesis addresses the problem, so far not investigated, of developing procedures to estimate the margins of uncertainty of the results obtained by applying this procedure in real studies. After defining suitable indicators, EI and EP, to quantify this uncertainty, the process is based on using the residuals of the *lphom* model to (i) evaluate the degree of non-compliance with the homogeneity hypothesis on which *lphom* is based, and (ii) investigate, through a set of simulation studies developed in different scenarios, the relationships between these indicators.

The general conclusions of these studies allow the development of procedures to estimate the predictable values of IE and PE in real studies, as well as the upper confidence limits for them. A procedure is also developed to obtain confidence intervals for the different transition probabilities of votes p_{jk} estimated by *lphom*. The methods developed are illustrated by applying them to the estimation of the transfer of votes between the first and second rounds of the 2017 French presidential elections.

A series of future lines of research suggested by the results of this thesis are finally indicated.

ÍNDICE DE CONTENIDO

Capítulo 1. Introducción, Antecedentes y Objeto de la Tesis	1
1.1 INTRODUCCIÓN	1
1.2 ANTECEDENTES	3
1.3. OBJETO Y CONTENIDO DE LA TESIS	4
Capítulo 2. El procedimiento <i>lphom</i> para estimar el trasvase de votos	6
2.1 INTRODUCCIÓN	6
2.2 EL MODELO <i>lphom</i>	6
2.2.1 Terminología utilizada y datos del modelo	6
2.2.2 El modelo básico inicial.....	7
2.2.3 La hipótesis de homogeneidad del comportamiento electoral.....	8
2.2.4 Tratamiento de las Altas y las Bajas	9
2.3 ALGUNAS APLICACIONES DE <i>lphom</i> A PROCESOS ELECTORALES ESPAÑOLES	10
2.4 APLICACIÓN DE <i>lphom</i> AL ESTUDIO DE LAS ELECCIONES PRESIDENCIALES FRANCESAS DE 2017	13
2.4.1 Introducción	13
2.4.2 Resultados del modelo	15
2.5 COMPARACIÓN CON <i>ei.MD.bayes</i>	17
2.6 CONCLUSIONES	17
Capítulo 3. Simulación de procesos electorales.....	19
3.1 INTRODUCCIÓN	19
3.2 ÍNDICES DE ERROR	20
3.2.1 Introducción	20
3.2.2 El índice EI	20
3.2.3 El índice EP	21
3.3 ÍNDICES DE HETEROGENEIDAD	21
3.3.1 Introducción	21
3.3.2 Índice de heterogeneidad real HET	22
3.3.3 Índice de heterogeneidad estimado HETe	22
3.3.4 Relación del HETe estimado con el HET real	23
3.4 METODOLOGÍA DE LA SIMULACIÓN	23
Capítulo 4. Estudio de la relación entre el índice de error EI y el índice de heterogeneidad estimado HETe	26
4.1 INTRODUCCIÓN	26
4.2 MODELACIÓN DE LA RELACIÓN ENTRE EI y HETe	27
4.3 RESULTADOS OBTENIDOS EN EL ESCENARIO 1	27
4.3.1 Descripción del Escenario 1	27
4.3.2 Resultados.....	27

4.4 RESULTADOS OBTENIDOS EN EL ESCENARIO 2	29
4.4.1 Descripción del Escenario 2	29
4.4.2 Resultados.....	30
4.5 RESULTADOS OBTENIDOS EN EL ESCENARIO 3	31
4.5.1 Descripción del Escenario 3	31
4.5.2 Resultados.....	32
4.6 RESULTADOS OBTENIDOS EN EL ESCENARIO 4	33
4.6.1 Descripción del Escenario 4	33
4.6.2 Resultados.....	34
4.7 RESULTADOS OBTENIDOS EN EL ESCENARIO 5	35
4.7.1 Descripción del Escenario 5	35
4.7.2 Resultados.....	36
4.8 SÍNTESIS DE LOS RESULTADOS OBTENIDOS EN LOS CINCO ESCENARIOS.....	37
4.9 ANÁLISIS PROVISIONAL DE LA INFLUENCIA DEL ESCENARIO SOBRE LA RELACIÓN ENTRE EL ÍNDICE DE HETEROGENEIDAD Y EL ÍNDICE DE ERROR	38
4.9.1 Introducción.....	38
4.9.2 Dos posibles características relevantes	39
4.9.2.1 Ratio REI ecuaciones/incógnitas	39
4.9.2.2 Índice DIV de diversidad de las unidades territoriales.....	39
4.9.3 Relación de REI y DIV con el índice de error EI.....	40
Capítulo 5. Un procedimiento para estimar el índice de error (EI) en estudios reales.....	42
5.1 INTRODUCCIÓN.....	42
5.2 PROCEDIMIENTO PROPUESTO.....	42
5.3 LA FUNCIÓN <i>estimei.m</i>	43
5.4 EJEMPLO DE APLICACIÓN	44
Capítulo 6. Un procedimiento para estimar el EP en estudios reales.....	46
6.1 INTRODUCCIÓN.....	46
6.2 RELACIÓN ENTRE EL EP Y EL HETe	46
6.3 ESTIMACIÓN DEL EP EN ESTUDIOS REALES	49
6.4 EJEMPLO DE APLICACIÓN	50
Capítulo 7. Intervalos de confianza para las p_{jk}	53
7.1 INTRODUCCIÓN.....	53
7.2 SESGO DE LOS ESTIMADORES \hat{p}_{jk}	54
7.3 VARIANZA DE LOS ESTIMADORES \hat{p}_{jk}	56
7.3.1 Introducción.....	56
7.4 UN PROCEDIMIENTO PARA OBTENER INTERVALOS DE CONFIANZA PARA LOS p_{jk}	57
7.4.1 Metodología.....	57
7.4.2 Ejemplos	58
Capítulo 8. Conclusiones y futuras investigaciones.....	60

BIBLIOGRAFÍA	62
Anejo 1. Software desarrollado para esta tesis	67
A1.1 FUNCIONES MATLAB	67
A1.2 SOFTWARE EN R.....	75
Anejo 2. Elecciones presidenciales francesas de 2017	77
A2.1 Resultados desagregados a nivel regional	77
Anejo 3. Descripción de los escenarios utilizados en el capítulo 4.....	79
A3.1 Escenario 1.....	79
A3.2 Escenario 2.....	80
A3.3 Escenario 3.....	81
A3.4 Escenario 4.....	83
A3.5 Escenario 5.....	84

INDICE DE TABLAS

Tabla 2. 1 Aplicaciones de <i>lphom</i> en estudios electorales españoles	10
Tabla 2. 2 Francia: resultados primera vuelta elecciones presidenciales 2017	14
Tabla 2. 3 Francia: resultados segunda vuelta elecciones presidenciales 2017.....	14
Tabla 2. 4 Porcentaje de votos que pasan de una opción electoral a otra entre la 1ª y la 2ª vuelta	15
Tabla 4. 1 Ajuste de regresión del Modelo (4.1)	28
Tabla 4. 2 Ajuste de regresión del Modelo (4.1)	30
Tabla 4. 3 Ajuste de regresión del Modelo (4.1)	32
Tabla 4. 4 Ajuste de regresión del Modelo (4.1)	34
Tabla 4. 5 Ajuste de regresión del Modelo (4.1)	36
Tabla 4. 6. Valores de EI1, los índices REI y DVI en los 5 escenarios.....	40
Tabla 4. 7 Ajuste de regresión de $\log(EI1)$	40
Tabla 5. 1 Ajuste de regresión del Modelo (4.1)	44
Tabla 6. 1 Regresión de EP en función de HETe. Escenario 1	48
Tabla 6. 2. Regresión de EP en función de HETe. Elecciones francesas	51
Tabla 7. 1 Valores medios de los p_{jk} en PMOD	55

INDICE DE FIGURAS

Figura 2. 1 Origen de la Abstención en las elecciones autonómicas de 2015	12
Figura 2. 2 Destino en 2015 del voto de los electores que votaron a UPyD en 2011	12
Figura 2. 3 Origen de los votos obtenidos en 2015 por Ciudadanos	12
Figura 2. 4 Destino de los votos de los nuevos electores incorporados al censo en 2015.....	13
Figura 2. 5 Francia: resultados de la primera vuelta elecciones presidenciales 2017	14
Figura 2. 6 Francia: resultados de la segunda vuelta elecciones presidenciales 2017.....	14
Figura 2. 7 Francia: Procedencia de los electores en las 4 opciones	15
Figura 4.1 Escenario 1: Relación entre EI y HETe	28
Figura 4.2 Escenario 1: Histograma y gráfico en papel.....	29
Figura 4.3 Escenario 1: Modelo ajustado	29
Figura 4.4 Escenario 2: Relación entre EI y HETe	30
Figura 4.5 Escenario 2: Histograma y gráfico en papel.....	31
Figura 4.6 Escenario 2: Modelo ajustado	31
Figura 4.7 Escenario 3: Relación entre EI y HETe	32
Figura 4.8 Escenario 3: Histograma y gráfico en papel.....	33
Figura 4.9 Escenario 3: Modelo ajustado	33
Figura 4.10 Escenario 4: Relación entre EI y HETe	34
Figura 4.11Figura 4.11. Escenario 4: Histograma y gráfico en papel	35
Figura 4.12 Escenario 4: Modelo ajustado	35
Figura 4.13 Escenario 5: Relación entre EI y HETe	36
Figura 4.14 Escenario 5: Histograma y gráfico en papel.....	37
Figura 4.15 Escenario 5: Modelo ajustado	37
Figura 4.16 Relación entre EI medio previsto y HETe en los cinco escenarios.....	38
Figura 4.17 EI observado frente al previsto a partir de REI y DVII	41
Figura 5.1 Elecciones francesas: Relación entre IER y HETe	44
Figura 5.2 Elecciones francesas: Histograma y gráfico en papel	45
Figura 5.3 Elecciones francesas: Modelo ajustado.....	45
Figura 6.1 Escenario 1: Relación entre sesgo absoluto medio (EP) y HETe	47
Figura 6.2 Escenario 1: Relación entre EP y HETe.....	47
Figura 6.3 Escenario 1: Histograma y gráfico en papel probabilístico normal	48
Figura 6.4 Escenario 1: Modelo ajustado de la relación del EP con HETe.....	49
Figura 6.5 Elecciones francesas: Relación entre EP y HETe	51
Figura 6.6 Elecciones francesas: Histograma y gráfico en papel	51
Figura 6.7 Elecciones francesas: Modelo ajustado	52
Figura 7.1 Escenario 1: Relación del Sesgo de \hat{p}_{jk} con el valor real de p_{jk}	54
Figura 7.2 Relación del sesgo con p_{jk} simulando con PMOD	55
Figura 7. 3 Relación entre sesgo estimación p_{11} y HETe.....	56
Figura 7.4 Relación entre varianza estimación p_{11} y HETe	57
Figura 7.5 Relación entre sesgo estimación p_{13} y HETe	57
Figura 7.6 Relación entre varianza estimación p_{13} y HETe	57

Capítulo 1. Introducción, Antecedentes y Objeto de la Tesis

1.1 INTRODUCCIÓN

El estudio de la movilidad electoral, que es como se denomina al fenómeno del trasvase de votos que se produce entre partidos o candidatos (opciones electorales, en general) entre dos procesos electorales, es un tema relevante en el campo de la sociología política que ha suscitado el interés de numerosos autores durante décadas. Gran cantidad de autores nacionales —por ejemplo, Vázquez y Romero (2001), Romero (2014), Corominas, Lusa y Valvet (2015), Puig y Ginebra (2015), Romero (2015), Pavía, Bodoque y Martín (2016), Romero (2016), Pavía y Aybar (2020), o Romero, Pavía, Martín y Romero (2020)— e internacionales —por ejemplo, Goodman (1961), Converse (1966), Goodman (1968), Hawkes (1969), Miller (1972), McCarthy y Ryan (1977), Keller y ten Cate (1977), Brown y Payne (1986), Payne et al. (1986), Thomsen (1987), Füle (1994), Gibson y Cielecka (1995), Katz y King (1999), Barbosa y Goldstein (2000), Wellhofer (2001), Ambühl (2003), Aarts, van der Kolk y Rosema (2007), van der Ploeg (2008), Park, (2008), Andreadis y Chadjipadelis (2009), Forcina y Marchetti (2011), Kellermann (2011), Park Hanmer, y Biggers (2014), Forcina, Gnaldi y Bracalente (2012), Colombi y Forcina (2016), Klima, Thurner, Molnar, Schlesinger y Küchenhoff (2016), Plescia y De Sio (2018), o Klima et al. (2019)— han contribuido mediante estudios concretos y/o a través de aportaciones metodológicas.

En España este interés se ha acrecentado estos últimos años con la aparición de dos importantes nuevas fuerzas políticas, Podemos y Ciudadanos, a las que se suma VOX más recientemente. La aparición de estos nuevos partidos ha roto el tradicional esquema bipartidista de la política española y planteado la cuestión del origen de las ingentes cantidades de votos conseguidos por estas nuevas formaciones.

Tradicionalmente la movilidad electoral se ha intentado estimar mediante encuestas postelectorales de recuerdo de voto, como acumulación de los cambios en el voto constatados en los encuestados (Benewick, Birch, Blumer y Ewbank, 1969). Sin embargo, este enfoque plantea serios problemas que cuestionan tanto su eficiencia como su eficacia. Además de presentar el inconveniente de no estar siempre disponibles.

En primer lugar, desde el punto de vista de la eficiencia, este procedimiento es, si se desean obtener resultados fiables, muy costoso pues el tamaño de la muestra debería ser muy grande para poder alcanzar una precisión razonable en la estimación de las probabilidades de transferencia. En efecto, suponiendo que hubo J opciones electorales en la primera elección y K en la segunda, se trata de estimar la probabilidad de K sucesos, el voto a cada uno de los partidos presentados en la segunda elección, en las J poblaciones diferentes definidas por los votantes de cada una de las distintas opciones en la primera elección. La obtención de estimaciones precisas de estas probabilidades exigiría una muestra con un número suficiente de electores para cada una de estas J poblaciones, con el problema adicional de que al seleccionar al azar la muestra de un determinado tamaño en la población global no se sabe cuantos individuos de esta van a pertenecer a cada una de dichas poblaciones.

Siendo la limitación anterior muy importante, tal como han puesto de manifiesto numerosos autores, el problema sin duda más relevante de este enfoque se focaliza en los problemas de fiabilidad que plantean las respuestas agregadas recolectadas en este tipo de encuestas, en las que se interroga a los encuestados por su comportamiento electoral en el pasado. Esta escasa fiabilidad se debe tanto a errores de medición como, muy especialmente, al sesgo inducido por las no respuestas. En efecto, la probabilidad de las no respuestas no se distribuye uniformemente entre los votantes de los diferentes partidos y, además, depende del contexto (Pavía, Badal y García-Cárceles, 2016). Adicionalmente las respuestas de los que sí

responden están expuestas a problemas, lo que les resta fiabilidad (Voogt, 2004; Keller & ten Cate, 1977; Upton, 1977) Es posible encontrar votantes que no recuerdan su comportamiento cuatro años antes (Weir, 1975) o cuya respuesta está condicionada por la posible aceptación social de la decisión que están tomando o tomaron en su día (Dassonneville and Hooghe, 2017).

Como un ejemplo del sesgo de las conclusiones obtenidas mediante estas encuestas, baste señalar que, en la encuesta realizada por el CIS en noviembre de 2014, solo el 28% de los encuestados afirmaron haber votado al Partido Popular en 2011 (C.I.S. 2014), cuando en realidad en dicha fecha este partido había sido respaldado por el 45% de los votantes. Constataciones como ésta hacen afirmar a Miller (1972) que “*las encuestas que tratan del cambio en el voto son especialmente poco fiables*” y que, como consecuencia de estas limitaciones, autores que han estudiado en profundidad el tema acaben afirmando que en este tipo de encuestas la imprecisión y los sesgos plantean obstáculos muy difíciles de superar (Klima et al. 2016).

Esta situación ha motivado que diversos autores hayan intentado abordar el problema con otros enfoques que utilizan únicamente los resultados efectivamente registrados en ambos procesos electorales, unos datos que son generalmente muy fiables. El problema del trasvase de votos entre dos procesos electorales aparece en este contexto como un caso particular de los abordados por la denominada Inferencia Ecológica.

Los denominados procedimientos de inferencia ecológica pueden agruparse en dos categorías principales: métodos de regresión ecológica y procedimientos de programación matemática. En esta investigación adoptamos una perspectiva de programación matemática. La literatura de regresión ecológica, sin embargo, ha sido muy fértil, indudablemente impulsada por su uso en procesos legales relacionados con el rediseño de los distritos electorales en los Estados Unidos (e.g., King, Rosen y Tanner, 2004) y por su uso en epidemiología (por ejemplo, Fisher y Wakefield, 2020). De hecho, tras los trabajos pioneros de Duncan y Davis (1953) y Goodman (1953, 1959), los métodos de regresión ecológica han experimentado un resurgimiento con el trabajo de King (1997), a pesar de las críticas de Freedman et al. (1998) y Cho (1998); siendo una referencia clave King, Rosen y Tanner (2004).¹

Desde un planteamiento puramente matemático el problema puede definirse como sigue: sobre una población (en nuestro caso todos los electores de un territorio) se definen dos particiones: una PJ con J clases y otra PK con K clases. Conociéndose los cardinales de todas las clases en ambas particiones, el objetivo es estimar los cardinales de las JxK conjuntos que resultan de la intersección de ambas particiones.

En nuestro caso las clases de la partición PJ las constituyen los electores que optaron por cada una de las J opciones ofrecidas en la Elección 1 y las de la partición PK las constituyen los electores que optaron por cada una de las K opciones ofrecidas en la Elección 2, siendo el objetivo del estudio estimar el cardinal v_{jk} de la intersección de la clase j de la primera partición con la clase k de la segunda partición. Estos v_{jk} , que son el número de electores que habiendo elegido la opción j en la primera elección optaron por la opción k en la segunda, son los elementos de la matriz \mathbf{V} de transición de votos que se desea estimar.

Obsérvese que un problema formalmente análogo aparece, por ejemplo, cuando en una única elección los electores pueden partitionarse en grupos a partir de condiciones como edad, sexo, grupo racial, etc... (ésta sería la partición PJ) y el objetivo es estimar cómo han votado cada uno de estos grupos conociendo los resultados globales de la elección (que definirían la partición PK).

¹ Los estudios electorales y la epidemiología no son las únicas disciplinas en las que se pueden utilizar los enfoques de inferencia ecológica. La inferencia ecológica es útil en todas aquellas situaciones donde el objetivo sea inferir el comportamiento a nivel individual a partir de datos agregados (Caughey y Wang, 2019).

Básicamente el problema puede plantearse, por tanto, como el de hallar los valores, no negativos, de los elementos de una matriz con J filas y K columnas de la que se conocen los totales de las filas y los totales de las columnas.

El problema fundamental que aparece en este problema, y que ha recibido el calificativo de “falacia ecológica” (Robinson, 1950), es que el sistema de ecuaciones al que conduce su formulación matemática tiene más incógnitas ($J \times K$) que ecuaciones ($J+K$), por lo que resulta básicamente indeterminado, existiendo infinitas soluciones diferentes que lo satisfacen.

En el siguiente apartado analizamos cómo se ha abordado la resolución de esta indeterminación desde diferentes enfoques.

1.2 ANTECEDENTES

La indeterminación fundamental del problema planteado, hace imprescindible la inclusión de alguna hipótesis adicional para alcanzar una solución (Greiner y Quinn, 2009). Esta hipótesis deberá ser lo más objetiva posible, para evitar que las soluciones obtenidas resulten sesgadas por la subjetividad del planteamiento en el que se ha basado su obtención.

El enfoque básico de la forma de resolver la indeterminación ha sido el mismo en todos los autores analizados: utilizar los resultados electorales, no sólo en el territorio global, sino en un conjunto de I unidades territoriales en las que se particiona dicho territorio, y asumir como hipótesis básica que las matrices de transición en las diferentes unidades son, en un cierto sentido a precisar, “similares” en todas ellas.

Dos estrategias diferentes aparecen al respecto en la literatura: la de los métodos de la denominada “regresión ecológica”, fundamentadas en un planteamiento probabilístico del problema y la de los métodos que lo abordan como una cuestión de optimización matemática. La frontera entre ambos enfoques no es, sin embargo, totalmente nítida, puesto que en los enfoques probabilísticos se acaban utilizando técnicas de optimización matemática, a veces muy complejas, para obtener las estimaciones.

La literatura en el campo de la regresión ecológica ha sido abundante desde los primeros artículos de Duncan y Davis (1953) y de Goodman (1953, 1959). Klima et al. (2016) presentan y discuten algunos de los principales métodos desarrollados con este enfoque y hacen un estudio comparativo de los mismos, poniendo de manifiesto los problemas que plantea su aplicación en la práctica. El principal de ellos en los enfoques más recientes, aparte de sus elevadas exigencias computacionales, reside en su complejidad, especialmente para los que utilizan un enfoque bayesiano, complejidad que requiere personas muy expertas para la fijación de los parámetros e hipótesis sobre las variables que se desean estimar, puesto que los resultados dependen mucho de ellas. Así, a partir de unos mismos datos pueden obtenerse resultados bastante diferentes en función de dichas hipótesis, empeorando el problema el hecho de que muchas de ellas son difíciles de interpretar o justificar en términos sociopolíticos.

Un enfoque diferente seguido por otros autores es el de plantear la cuestión como un problema de optimización matemática, en el que las incógnitas que se desea estimar son las probabilidades p_{jk} de que un elector que votó a la opción j en la primera elección haya votado a la opción k en la segunda. Estas p_{jk} deben satisfacer un conjunto de restricciones implícitas en su definición (por ejemplo, ser no negativas y sumar 1 para cada opción j) y optimizar un determinado criterio que refleja matemáticamente la idea de que las matrices de transferencia deben ser “parecidas” en la I unidades territoriales. McCarthy y Ryan (1977) proponen al respecto un modelo de programación cuadrática para minimizar las discrepancias entre los resultados obtenidos por el modelo y los constatados en cada una de las I unidades territoriales para cada una de las K opciones. Tziafetas (1986) sugiere minimizar la suma de los valores absolutos de

estas discrepancias, lo que transforma el modelo en un Programa Lineal más fácil de resolver, enfoque similar al propuesto por Vázquez y Romero (2001) que plantean diferentes alternativas para la función a optimizar, como hacen también Corominas, Lusa y Valvet (2015). Romero (2014) plantea en su modelo una alternativa novedosa para abordar el problema de los cambios en el censo entre los dos procesos electorales. Esta alternativa es profundizada en Romero, Pavía, Martín y Romero (2020)², que proponen la denominación de *lphom* para el modelo que desarrollan, modelo que se expone con detalle en el Capítulo 2 de esta tesis.

En los modelos basados en la regresión ecológica, los resultados pueden acompañarse de una estimación del margen de incertidumbre de las estimaciones obtenidas. Sin embargo, en los métodos basados en la optimización matemática no se ha propuesto hasta la fecha ningún procedimiento para estimar dicho margen de incertidumbre en los resultados que proporcionan.

Éste es, precisamente, el objetivo de la presente investigación, tal como se precisa en el siguiente apartado.

1.3. OBJETO Y CONTENIDO DE LA TESIS

Nuestra investigación se ha centrado sobre el método *lphom*, un procedimiento que, tal como se expone en el Capítulo 2, es extremadamente sencillo de aplicar y ha conducido a resultados muy razonables, desde el punto de vista sociopolítico, en todos los estudios en los que se ha utilizado para estimar la movilidad en procesos electorales reales.

El objetivo básico de la tesis es desarrollar procedimientos que permitan estimar el margen de incertidumbre de los resultados proporcionados por *lphom* cuando se aplica en estudios concretos.

Este margen de incertidumbre se va a referir tanto a la estimación global de la matriz de transferencia de votos entre las diferentes opciones de los dos procesos electorales, como a la estimación concreta de los valores de cada uno de los p_{jk} .

Tras el presente capítulo, de carácter introductorio, en el Capítulo 2 se presenta el modelo *lphom*. En el Capítulo 2 se detallan las ecuaciones que definen el modelo *lphom*, se precisan y justifican la hipótesis de homogeneidad de comportamiento electoral en la que se basa, se ilustra su aplicación con algunos estudios reales recientes correspondientes a procesos electorales españoles y se exponen con detalle los resultados que *lphom* obtiene al estudiar los trasvases de votos entre la primera y la segunda vuelta de las elecciones presidenciales francesas de 2017. El capítulo, aunque no contiene resultados originales de investigación, sí que incluye al final una comparación novedosa entre los resultados obtenidos por *lphom* en el mencionado estudio de las elecciones francesas y los que se obtienen con los mismos datos al aplicar el sofisticado procedimiento de regresión ecológica *ei.MD.bayes*, que es defendido por Klima et al. (2016) como el que mejores resultados ha obtenido en su estudio comparativo.

La metodología desarrollada en nuestra investigación se soporta en la realización de simulaciones de procesos electorales en “escenarios” similares al que se presenta en el estudio concreto que se pretende analizar. En el Capítulo 3 se estudia el procedimiento que hemos desarrollado para simular dichos procesos electorales. Previamente se proponen unos indicadores EI y EP para evaluar el grado de error de los resultados proporcionados por *lphom* al analizar un proceso electoral concreto, así como otros indicadores HET y HETe que se utilizarán para cuantificar el grado de incumplimiento en un proceso concreto de la hipótesis de homogeneidad del comportamiento electoral en las I unidades territoriales. En

² Este artículo es obra de los dos directores de esta tesis y del propio doctorando, constituyendo parte de su contenido una parte del contenido de esta tesis, especialmente en el capítulo 5.

el capítulo, tras exponer con detalle la metodología desarrollada para llevar a cabo las simulaciones, se describe la operativa de la función Matlab *simular.m* que se utiliza en el resto de la tesis con este fin.

En el Capítulo 4 la metodología anterior se aplica para establecer un modelo que relacione el coeficiente EI de un estudio con el HETe estimado en el mismo. La idea básica previa es que, dado que si en un estudio resultase HETe igual a cero EI sería necesariamente cero, parece razonable pensar que cuanto mayor sea el grado de incumplimiento de esta hipótesis de homogeneidad y, por tanto, mayor sea el valor del HETe, mayor va a ser el valor esperado de EI. Con el fin de modelar esta relación en el capítulo se llevan a cabo intensivos estudios de simulación en cinco escenarios diferentes, obteniéndose un modelo que se ajusta bien en todos los escenarios, aunque sus parámetros concretos dependen de las condiciones de cada uno de éstos. El capítulo concluye con un análisis provisional de los factores del escenario que influyen sobre la relación entre el índice de heterogeneidad y el índice de error, obteniéndose al respecto unos primeros resultados que pueden ser la base de investigaciones futuras sobre el tema.

En el Capítulo 5 los resultados del capítulo anterior se utilizan para abordar el objetivo principal de nuestra investigación: el desarrollo de un procedimiento que permite estimar el índice de error EI correspondiente a los resultados proporcionados por *lphom* al estimar la matriz de transferencia de votos en estudios electorales concretos. Para facilitar la aplicación del procedimiento propuesto se desarrolla la función Matlab *estimei.m*, que se expone con detalle, aplicándose para estimar el EI en el estudio sobre el trasvase de votos entre las dos vueltas de las elecciones presidenciales francesas de 2017 analizado en el Capítulo 2.

El Capítulo 6 es muy similar al Capítulo 5, pero centrando el interés en la evaluación del margen de incertidumbre mediante el indicador EP, que mide el error cometido al estimar la matriz \mathbf{P} de probabilidades de transferencia de votos entre las opciones de las dos elecciones. Su estructura es muy similar a la del capítulo anterior, desarrollándose una función Matlab *estimep.m* y aplicándola para estimar el EP en el estudio sobre las elecciones francesas.

En el capítulo 7 se aborda el problema de obtener intervalos de confianza para los $J \times K$ elementos p_{jk} de la matriz \mathbf{P} de transferencia estimados por *lphom*. Se estudia, en primer lugar, el posible sesgo y la varianza de dichos estimadores \hat{p}_{jk} y la relación de ambos con el HETe. Los resultados obtenidos se utilizan para elaborar un procedimiento que permite obtener un intervalo de confianza para cualquier p_{jk} en estudios reales, desarrollándose una función Matlab *estimpjk.m* para llevar a cabo esta estimación e ilustrando su aplicación con dos ejemplos relativos a las mencionadas elecciones francesas.

Finalmente, en el capítulo 8 se sintetizan las principales conclusiones y aportaciones de esta tesis y se plantean diversas líneas futuras de investigación, sobre algunas de las cuales estamos ya trabajando.

En el Anejo 1 se exponen y explican todas las funciones Matlab y scripts R utilizados en la investigación, en el Anejo 2 se dan detalles sobre la comparación realizada entre *lphom* y *ei.MD,bayes* en el Capítulo 2 y en el Anejo 3 se exponen las características de los cinco escenarios utilizados para las simulaciones realizadas en el Capítulo 4.

Capítulo 2. El procedimiento *lphom* para estimar el trasvase de votos

2.1 INTRODUCCIÓN

La presente investigación se va a centrar en el método *lphom*, un procedimiento para estimar el trasvase de votos de unas opciones a otras entre dos procesos electorales mediante la construcción de un modelo matemático de Programación Lineal. En los antecedentes analizados en el capítulo anterior se han comentado varias referencias sobre el tema, que no repetiremos aquí. Este capítulo está dedicado a profundizar en el procedimiento.

En el Apartado 2.2 del presente capítulo se expone con detalle el modelo *lphom*. En la exposición seguimos a Romero (2014) y Romero et al. (2020). Precisamente en esta última referencia los autores proponen para el modelo que desarrollan el nombre de *lphom* como un acrónimo para “**L**inear **P**rogram model based on the **HOM**ogeneity hypothesis”.

En el Apartado 2.3 se exponen algunas aplicaciones recientes del modelo *lphom* al estudio de diferentes procesos electorales españoles, resaltando el carácter muy razonable, desde el punto de vista sociopolítico, de los resultados que *lphom* produce en todos los casos.

En el Apartado 2.4 aplicamos dicho modelo para estimar el trasvase de votos entre las dos vueltas de las elecciones presidenciales celebradas en Francia en 2017 y en el Apartado 2.5 comparamos los resultados obtenidos mediante *lphom* en este estudio con los que se obtienen a partir de los mismos datos cuando se aplica el procedimiento *ei.MD.bayes* de Inferencia Ecológica.

Concluimos en el Apartado 2.6 con unas reflexiones que dan entrada a los temas de investigación abordados en esta tesis.

2.2 EL MODELO *lphom*

2.2.1 Terminología utilizada y datos del modelo

Se desea estudiar la movilidad electoral entre dos procesos electorales en un determinado territorio y se dispone de los resultados electorales diferenciados en un conjunto de unidades territoriales que constituyen una partición de dicho territorio. En adelante nos referiremos en general a estas unidades territoriales con la abreviatura u.t.

Sean³:

I: número de unidades territoriales.

³ En esta tesis utilizamos dos terminologías para referirnos a los cardinales del conjunto de las uts, del conjunto de opciones en la Elección 1 y del conjunto de opciones en la Elección 2. En general en las fórmulas, y tal como hacemos en este capítulo, utilizamos las letras I, J y K mayúsculas para referirnos a esos cardinales, para tener una notación más compacta e intuitiva. Sin embargo, en capítulos posteriores, en los que se hace continua referencia a funciones Matlab, preferimos utilizar la notación ni, nj y nk, que es la utilizada en dichas funciones, para evitar posibles confusiones entre minúsculas y mayúsculas que se presentan en algunas versiones de Matlab.

J: número de opciones electorales en la Elección 1. Dichas opciones incluyen, además de los partidos que se presentaron, un “partido” que recogerá las abstenciones, incluyendo generalmente los votos en blanco y nulos. Con el fin de tener un modelo manejable sin perder información relevante, los partidos menos importantes se agrupan como un “partido” RESTO. Tal como vemos con detalle más adelante, para tener en cuenta los nuevos votantes que se han incorporado en cada ut al censo electoral entre ambas elecciones, y que son también un origen de posibles votos en la segunda elección, se crea un partido ficticio de ALTAS.

K: número de opciones electorales en la Elección 2. Dichas opciones, que no tienen por qué ser idénticas a las de la primera elección, incluyen, como en el caso anterior, una opción con las abstenciones y votos en blanco o nulos, y una opción BAJAS que recoge los votantes que han desaparecido del censo electoral de cada ut entre ambas elecciones.

Los datos disponibles son:

x_{ij} : votos recibidos en la Elección 1 por la opción electoral j en el u.t. i. Estas x_{ij} forman una matriz con I filas y J columnas.

y_{ik} : votos recibidos en la Elección 2 por la opción electoral k en el u.t. i. Estas y_{ik} forman una matriz con I filas y K columnas.

Las incógnitas fundamentales que pretendemos estimar son los JxK valores p_{jk} definidos como sigue:

p_{jk}^i : proporción de electores del total del territorio estudiado que habiendo elegido la opción electoral j en la Elección 1 pasaron a elegir la opción electoral k en la Elección 2. Estas proporciones se expresan en tanto por uno, pudiendo interpretarse como la probabilidad de que en el territorio analizado un elector que opta por la opción j en la primera elección opte por la opción k en la segunda.

2.2.2 El modelo básico inicial

De la definición anterior se deduce inmediatamente que las p_{jk} deben cumplir las siguientes restricciones:

$$p_{jk} \geq 0 \quad \text{para } j=1 \dots J \quad k=1 \dots K \quad (1)$$

$$\sum_k p_{jk} = 1 \quad \text{para } j=1 \dots J \quad (2)$$

$$\sum_{j=1}^J (\sum_i x_{ij}) p_{jk} = (\sum_i y_{ik}) \quad \text{para } k=1 \dots K \quad (3)$$

Para que este sistema de ecuaciones sea compatible es necesario que las matrices de datos $[x_{ij}]$ e $[y_{ik}]$ verifiquen la igualdad:

$$\sum_{j=1}^J x_{ij} = \sum_k y_{ik} \quad \text{Para } i = 1 \dots I$$

o sea, que coincidan las sumas de las I filas en las dos matrices. En la mayor parte de los casos, para que se cumpla esta condición será necesario incluir las Altas como una última “opción” J en la Elección 1 y las Bajas como una última “opción” en la Elección 2.

El problema fundamental que plantea este modelo básico constituido por las ecuaciones (1), (2) y (3) es que, al ser un sistema con muchas más incógnitas que ecuaciones, resulta siempre indeterminado, existiendo por tanto infinitas soluciones diferentes para las p_{jk} que las cumplirían exactamente.

Para resolver esta indeterminación es necesario incluir alguna hipótesis adicional. Esta hipótesis deberá ser lo más objetiva posible, para no condicionar los resultados del modelo mediante apriorismos subjetivos.

2.2.3 La hipótesis de homogeneidad del comportamiento electoral

La hipótesis fundamental en la que se basa el modelo *lphom* es la de “homogeneidad del comportamiento electoral en las I unidades territoriales”, hipótesis que se desarrolla a continuación.

Sea p_{ijk} la proporción de electores de la ut i que habiendo elegido la opción j en la Elección 1 pasaron a votar a la opción k en la Elección 2. La hipótesis de homogeneidad electoral asume que estos valores de las p_{ijk} en las I unidades territoriales han sido “parecidos” a los valores promedios p_{jk} en el total del territorio, que son los que se pretende estimar.

Para que esta hipótesis sea razonable es necesario que los estímulos que han condicionado el posible cambio de votos de los electores hayan sido parecidos en las distintas unidades territoriales. Un contraejemplo en el que la hipótesis de homogeneidad electoral resultaría discutible, sería el de unas elecciones municipales en las que el voto en cada municipio está lógicamente muy condicionado por el perfil de los candidatos que presenta cada partido, perfil que puede ser muy diferente de unos municipios a otros.

Hay que resaltar que la hipótesis no implica que los resultados electorales deben ser parecidos en las diferentes unidades territoriales, sino sólo que las proporciones p_{ijk} de los trasvases de votos entre partidos han sido parecidas en dichas unidades. Por ejemplo, en el estudio de las elecciones presidenciales francesas que se analiza en el Apartado 2.4 hubo regiones que votaron preferentemente a Macron y otras que lo hicieron a Le Pen; lo que la hipótesis de homogeneidad implica es que si en el total de Francia la mayor parte de los electores que votaron a Macron en la primera vuelta volvieron a votarlo en la segunda, e igualmente los que votaron a Le Pen en la primera volvieron a hacerlo en la segunda, este comportamiento ha sido muy parecido en todas las regiones.

Para que esta hipótesis de homogeneidad sea razonable es necesario que el territorio objeto del estudio sea homogéneo para la elección considerada, en el sentido de que:

- Las opciones presentadas hayan sido básicamente las mismas en todas las uts, tanto en la Elección 1 como en la Elección 2.
- Las motivaciones que puedan haber influido en el cambio del voto de los electores hayan sido parecidas en todo el territorio, variando poco de unas uts a otras.

La hipótesis de homogeneidad del comportamiento electoral implica que si se aplican estas p_{jk} “promedio” en cada unidad territorial los resultados teóricos obtenidos deben ser “parecidos” a los observados realmente.

Denominando e_{ik} a la diferencia, que puede ser positiva o negativa, entre los resultados reales observados para el partido k en la u.t. i en la Elección 2 y los resultados teóricos esperables bajo la hipótesis de homogeneidad electoral:

$$e_{ik} = y_{ik} - \sum_j x_{ij} p_{jk} \text{ para } i=1\dots I \quad k=1\dots K \quad (4)$$

la hipótesis de homogeneidad del comportamiento electoral implica que los valores de estos e_{ik} deben ser “pequeños”.

En consecuencia, el modelo propuesto plantea obtener los valores de las p_{jk} de forma que, satisfaciendo el conjunto de restricciones (1), (2), (3) y (4), minimicen la suma de los valores absolutos de las e_{ik}

$$\text{minimizar } \sum_{i,k} |e_{ik}| \quad (5)$$

2.2.4 Tratamiento de las Altas y las Bajas

Como hemos indicado, para que el sistema de ecuaciones sea compatible es necesario que coincidan exactamente las sumas de las filas de las matrices $[x_{ij}]$ e $[y_{ik}]$, lo que obliga a incluir explícitamente en el modelo los cambios en el censo electoral entre las dos elecciones.

En el caso de dos procesos electorales muy próximos en el tiempo, como sucede en los contemplados en el Apartado 2.4, estas altas y bajas son despreciables, por lo que pueden añadirse a cualquiera de las opciones, por ejemplo, a la Abstención, sin que ello afecte en la práctica a los resultados.

Sin embargo, en la mayoría de los casos el tiempo transcurrido entre las dos elecciones es de varios años, por lo que los cambios en el censo electoral pueden ser importantes y deben ser tenidos en cuenta explícitamente porque, además, la información sobre el destino del voto de los nuevos electores, en su mayoría jóvenes que acaban de alcanzar la mayoría de edad, es relevante en el estudio.

En Romero et al. (2020) se discute detalladamente las distintas alternativas al respecto, en función de la información disponible en cada estudio. En general las Altas en cada ut i son la suma de dos colectivos: los jóvenes que se incorporan al censo por haber cumplido los 18 años de edad entre las dos fechas electorales y los electores que han inmigrado a la ut considerada en dicho periodo. Igualmente, las Bajas están formados por dos colectivos: los electores censados en la Elección 1 que han fallecido antes de la Elección 2, y aquéllos que han emigrado fuera de la ut i en el periodo interelectoral.

La información disponible sobre la importancia de estos colectivos varía de unos casos a otros, siendo responsabilidad de los que realizan el estudio utilizar esta información para estimar lo más aproximadamente posibles las cifras de Altas y Bajas en cada ut. Así, en España se publican antes de muchas elecciones las cifras provinciales de los jóvenes incorporados al censo por haber cumplido los 18 años. En otros casos estas cifras pueden estimarse con bastante aproximación a partir de los datos disponibles sobre las pirámides de edades de las poblaciones en las uts⁴. Por otra parte, las cifras de mortalidad para los mayores de 18 años son en general conocidas con bastante detalle. Combinando estas informaciones con la aportada por las diferencias entre los totales del censo de cada ut en ambas elecciones, se deberá estimar con la mayor precisión posible las cifras de Altas y Bajas entre las dos elecciones en las I uts.

Dado que las bajas electorales, es decir los electores que han desaparecido del censo electoral entre la primera y la segunda elección, se deben fundamentalmente a la mortalidad o al traslado de las personas a otras u.t., en el modelo *lphom* se suele incluir la hipótesis razonable de que estas bajas han afectado de forma similar a todas las opciones presentes en la primera elección, siendo proporcionales a los efectivos

⁴ Por ejemplo, en Pavía y Veres Ferrer (2016a) puede encontrarse una propuesta para la estimación de los nuevos electores por ut (sección censal), que es validada comparando con datos reales en Pavía y Veres Ferrer (2016b).

totales de cada uno en dicha elección. Como excepción obvia, para la opción de las ALTAS, cuyo índice es J, se fuerza a que ningún elector haya pasado a la opción BAJAS de la Elección 2.

Siendo K la opción que recoge las Bajas, y siendo $\frac{(\sum_i y_{iK})}{(\sum_{j=1}^{J-1} \sum_{i=1}^I x_{ij})}$ la proporción que suponen dichas bajas respecto al total del censo de la Elección 1, la hipótesis anterior implica las J restricciones adicionales siguientes:

$$p_{jK} = \frac{(\sum_i y_{iK})}{(\sum_{j=1}^{J-1} \sum_{i=1}^I x_{ij})} \quad \text{para } j=1 \dots J-1 \quad (6)$$

$$p_{JK} = 0 \quad (7)$$

El modelo *lphom* es por tanto un Programa Lineal con las restricciones (1), (2), (3), (4) y, en su caso, (6) y (7), en el que la función a minimizar es la dada por la ecuación (5).

En el Anejo 1 se recoge el listado de la función Matlab *lphom.m* que permite plantear y resolver el modelo descrito a partir de las matrices de datos $[x_{ij}]$ e $[y_{ik}]$. Dicha función utiliza la función auxiliar *calcbf.m*, también recogida en dicho Anejo, que permite obtener a partir de las matrices de datos los inputs para el programa lineal del modelo.

2.3 ALGUNAS APLICACIONES DE *lphom* A PROCESOS ELECTORALES ESPAÑOLES

El modelo *lphom* ha sido utilizado estos últimos años para analizar diferentes procesos electorales celebrados en España. Prácticamente en todos los casos el marco geográfico de los estudios ha sido una comunidad autónoma, marco en el que es razonable esperar el cumplimiento de la hipótesis de homogeneidad.

En la Tabla 2.1 se referencian estos estudios, algunos de los cuales no han sido publicados pero cuyos resultados son accesibles en la correspondiente web referenciada.

Tabla 2. 1 Aplicaciones de *lphom* en estudios electorales españoles.

Comunidad autónoma	Número de unidades territoriales	Elección 1	Opciones en Elección 1	Elección 2	Opciones en Elección 2	Referencia
Valencia (ciudad)	7	Europeas 2009	7	Europeas 2014	8	Romero (2014)
Aragón	15	Autonómicas 2011	8	Autonómicas 2015	10	www.est2.es
Andalucía	8	Autonómicas 2012	7	Autonómicas 2015	9	www.est2.es
Andalucía	8	Autonómicas 2015	7	Generales 20D/2015	7	www.est2.es
Andalucía	8	Generales 20D/2015	7	Generales 26J/2016	6	Romero (2016)
Cataluña	10	Autonómicas 2012	9	Autonómicas 2015	8	Romero (2015)
Comunidad Valenciana	10	Autonómicas 2011	7	Autonómicas 2015	9	www.est2.es
Comunidad Valenciana	14	Autonómicas 2015	8	Generales 20D/2015	7	www.est2.es
Comunidad	14	Generales	7	Generales	6	Romero (2016)

Valenciana		20D/2015		26J/2016		
Euskadi	8	Generales 26J/2016	6	Autonómicas 25S/2016	6	www.est2.es
Comunidad de Madrid	13	Autonómicas 2011	5	Autonómicas 2015	7	www.est2.es
Comunidad de Madrid	23	Generales 20D/2015	7	Generales 26J/2016	6	Romero (2016)

Con carácter general, en todos estos estudios los resultados obtenidos mediante *lphom* han sido "razonables", en el sentido de ser lógicos y claramente interpretables en términos de sociología política. Por ejemplo, cuando un partido se ha presentado a las dos elecciones (como sucede en general con los partidos importantes y con la opción Abstención) siempre se ha obtenido que la fuente de votos más importante para dicho partido han sido los votantes a ese partido en la elección anterior.

Esta primera constatación puede resultar incluso sorprendente, dado que los resultados obtenidos se basan en una manipulación puramente matemática de los datos, que no tiene en cuenta posibles proximidades ideológicas entre las diferentes opciones. Es obvio, por ejemplo, que cuando en los estudios se obtiene siempre que la fuente más importante de votos al PP en la Elección 2 son los votantes a este partido en la Elección 1, en ninguna de las ecuaciones que definen el modelo *lphom* aparece el hecho de que el PP de la Elección 1 y el de la Elección 2 sean el mismo partido.

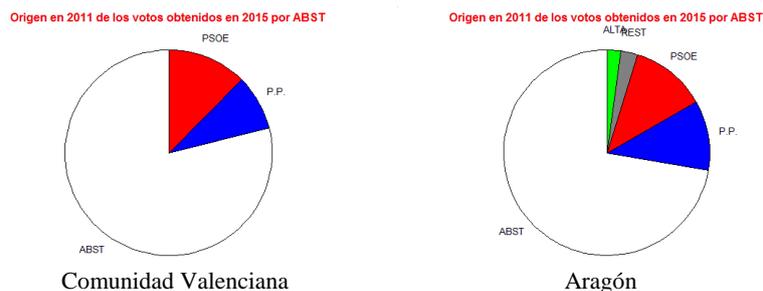
No resulta posible, ni sería quizás adecuado en el marco de esta tesis, detenerse a comentar los resultados obtenidos en todos estos estudios, pues, aunque muchos de ellos sean realmente interesantes y lógicos desde el punto de vista sociopolítico, no es éste el tema que nos ocupa en esta investigación.

Sí que resulta relevante, desde el punto de vista de la validación de la metodología, constatar que cuando se ha aplicado a procesos electorales realizados simultáneamente en distintas Comunidades Autónomas, muchos de los resultados han sido muy similares en todas ellas, a pesar de que obviamente los estudios se basaban en conjuntos de datos diferentes. Recogemos a continuación un ejemplo para poner de manifiesto este hecho.

Al estudiar en la Comunidad Valenciana y en Aragón la movilidad electoral entre las elecciones autonómicas de 2011 y de 2015, el método *lphom* obtuvo, entre otros, los resultados que se reflejan en las siguientes figuras:

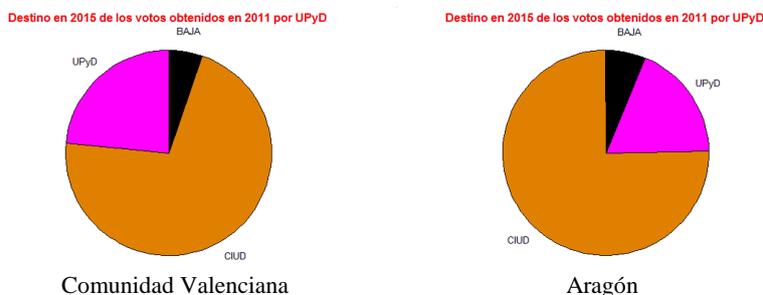
- En la Figura 2.1, relativa al origen de los electores que se abstuvieron en 2015, se aprecia de forma muy similar en ambas comunidades que, junto a una mayoría de anteriores abstencionistas, fueron muchos anteriores votantes de los dos grandes partidos tradicionales PP y PSOE los que decidieron abstenerse en 2015.

**Figura 2. 1 Origen de la Abstención en las elecciones autonómicas de 2015.
Resultados obtenidos por *lphom* para las comunidades valenciana y aragonesa.**



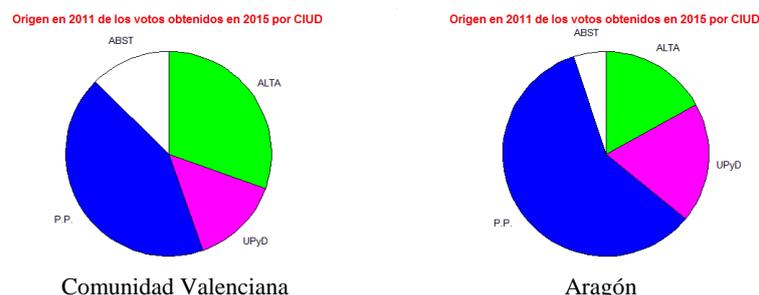
- En la Figura 2.2, relativa al destino en 2015 del voto de los electores que votaron a UPyD en 2011 se aprecia, de forma casi idéntica en ambas comunidades, que, excepto un 20% que se mantuvo fiel a dichas siglas, los restantes electores que no habían sido Bajos optaron por Ciudadanos. Resaltamos de nuevo, aun pecando de reiterativos, que para obtener estos resultados en ningún momento *lphom* ha utilizado la información de la evidente proximidad ideológica entre ambas formaciones políticas.

**Figura 2. 2 Destino en 2015 del voto de los electores que votaron a UPyD en 2011.
Resultados obtenidos por *lphom* para las comunidades valenciana y aragonesa.**



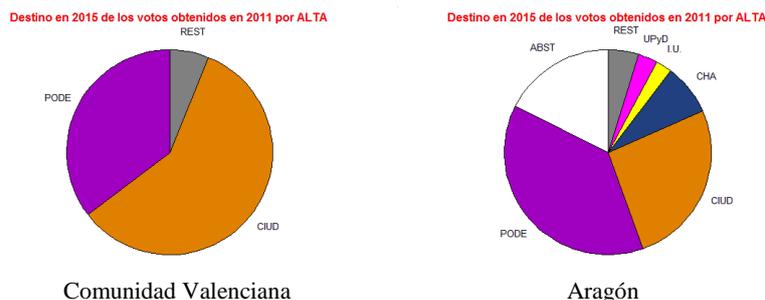
- En la Figura 2.3, relativa al origen de los votos obtenidos en 2015 por Ciudadanos se aprecia, de forma casi idéntica en ambas comunidades, que, los votos obtenidos por este partido procedían, por orden de importancia, de anteriores votantes del PP, de los jóvenes dados de alta en esas elecciones, de ex-votantes de UPyD y de anteriores abstencionistas.

**Figura 2. 3 Origen de los votos obtenidos en 2015 por Ciudadanos.
Resultados obtenidos por *lphom* para las comunidades valenciana y aragonesa.**



- En la Figura 2.4, relativa al destino de los votos de los nuevos electores incorporados al censo en 2015 se constata, de forma también casi idéntica en ambas comunidades, que, estos nuevos votantes, en su mayoría jóvenes que habían cumplido la edad de 18 años, optaron fundamentalmente por las nuevas formaciones, Podemos y Ciudadanos, en vez de hacerlo por los dos grandes partidos tradicionales.

Figura 2. 4 Destino de los votos de los nuevos electores incorporados al censo en 2015. Resultados obtenidos por *lphom* para las comunidades valenciana y aragonesa.



En el mismo sentido, en Romero (2016) se recogen los resultados de tres estudios realizados mediante *lphom* para estudiar en las comunidades valenciana, madrileña y andaluza la movilidad electoral entre los comicios del 20D de 2015 y los del 16J de 2016. En las conclusiones del artículo se recoge literalmente “Finalmente, y desde el punto de vista metodológico, merece resaltarse, por una parte, que los resultados obtenidos han sido todos razonables, en el sentido de poder ser interpretables en términos de sociología política, y, por otra, las estrechas coincidencias obtenidas en muchos casos entre los resultados de tres estudios completamente independientes”.

Del análisis de todos estos trabajos reseñados se deduce que el hecho de que *lphom*, pese a su naturaleza puramente matemática, haya proporcionado siempre resultados razonables desde el punto de vista sociopolítico, y coherentes entre estudios simultáneos en distintas comunidades basados en datos obviamente diferentes, puede considerarse como un cierto aval de su validez y utilidad para estudiar procesos reales.

2.4 APLICACIÓN DE *lphom* AL ESTUDIO DE LAS ELECCIONES PRESIDENCIALES FRANCESAS DE 2017

2.4.1 Introducción

En el presente apartado se utiliza el procedimiento *lphom* para estimar el trasvase de votos entre la primera y segunda vuelta de las elecciones presidenciales francesas celebradas en 2017.

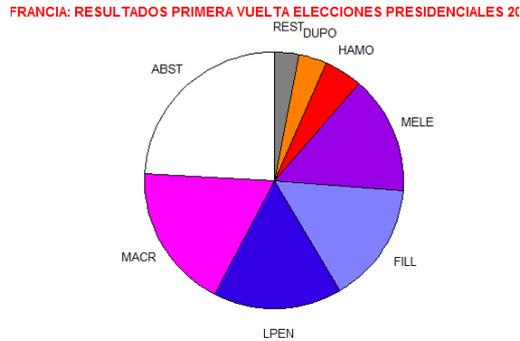
En Romero et al. (2020) se recoge un estudio similar, pero basado en los resultados desagregados recogidos en los 108 departamentos de Francia. En nuestro estudio analizaremos unos resultados más agregados, correspondientes a las doce regiones francesas, más una última “región” que aglutina todos los resultados de los territorios de ultramar. Precisamente, como comentaremos más adelante, una de las ventajas de *lphom* frente a otras alternativas de Inferencia Ecológica, propuestas para este tipo de estudios, es la de proporcionar buenos resultados incluso cuando se trabaja con datos muy agregados, o sea con un número reducido de uts.

La tabla y el gráfico siguientes recogen los resultados de la primera vuelta de las elecciones presidenciales celebradas en Francia en abril de 2017, donde los nombres de los diferentes candidatos se han abreviado a sus cuatro primeros caracteres. Los votos en blanco y nulos se incluyen en la abstención, representando el 8% de la misma.

Tabla 2. 2 Francia: resultados primera vuelta elecciones presidenciales 2017.

	TOTAL	ABST	MACR	LPEN	FILL	MELE	HAMO	DUPO	REST
VOTOS	47581118	11522305	8657326	7679493	7213797	7060885	2291565	1695186	1460561
% sobre total	100%	24.2%	18.2%	16.1%	15.2%	14.8%	4.8%	3.6%	3.1%
% sobre votos			24.0%	21.3%	20.0%	19.6%	6.4%	4.7%	4.1%

Figura 2. 5 Francia: resultados de la primera vuelta elecciones presidenciales 2017.



Al menos veinte millones de franceses, que ni se abstuvieron ni votaron a Macron o a Le Pen en la primera vuelta, tuvieron que cambiar su opción en la segunda vuelta, cuyos resultados se recogen en la tabla y el gráfico siguiente. Dada la importancia y la especial significación de los votos en blanco y nulos en esta segunda vuelta, en la que sólo competían dos candidatos, se han considerado diferenciados de la abstención.

Tabla 2. 3 Francia: resultados segunda vuelta elecciones presidenciales 2017.

	TOTAL	ABST	BLNU	MACR	LPEN
VOTOS	47582000	12076300	4079600	20760082	10666018
% sobre total	100%	25.4%	8.6%	43.6%	22.4%
% sobre votos			11.5%	58.5%	30.0%

Figura 2. 6 Francia: resultados de la segunda vuelta elecciones presidenciales 2017.



El objetivo del presente estudio es estimar, utilizando el procedimiento *lphom*, los cambios que se produjeron entre la primera y la segunda vuelta en las opciones elegidas de los votantes franceses.

2.4.2 Resultados del modelo

El procedimiento *lphom* exige disponer de los resultados electorales desagregados en un conjunto de unidades en el que se particiona el territorio estudiado. En este estudio dichas unidades han sido las doce regiones continentales francesas, englobando todos los electores restantes de ultramar en un grupo Resto. En el Anejo 3 se recogen, desagregados a nivel de regiones, los resultados de la primera y segunda vuelta de las elecciones, a partir de los que se ha construido el modelo.

Dada la proximidad de las fechas de las dos elecciones, los cambios entre las mismas en los totales de los censos de las 13 uts han sido mínimos. De todas formas, como *lphom* exige que cuadren exactamente dichos totales en ambas elecciones, dichas diferencias se han incluido en las Abstenciones, sin que esta operativa, dado el muy bajo valor de estas diferencias, pueda alterar de forma significativa los resultados del estudio.

El modelo construido para estimar los trasvases de votos ha sido un Programa Lineal con 175 variables y 91 restricciones.

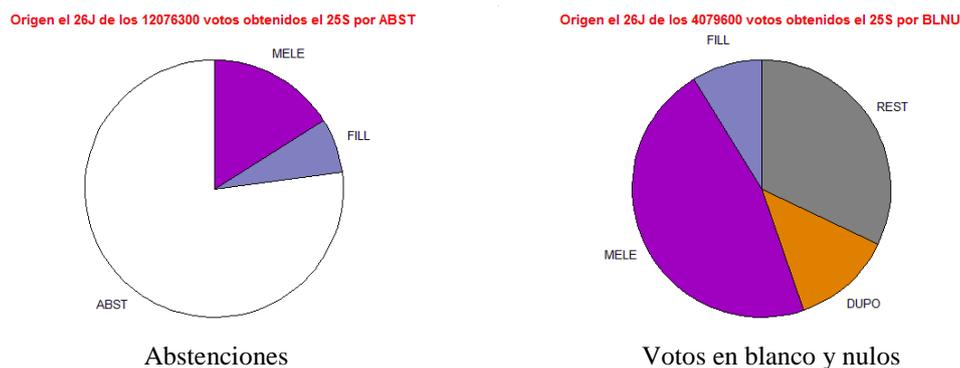
En la Tabla 2.4 se recogen los valores de los p_{jk} obtenidos al resolver el modelo *lphom*

Tabla 2. 4 Porcentaje de votos que pasan de una opción electoral a otra entre la 1ª y la 2ª vuelta en Francia, elecciones presidenciales 2017. Estimación mediante *lphom*

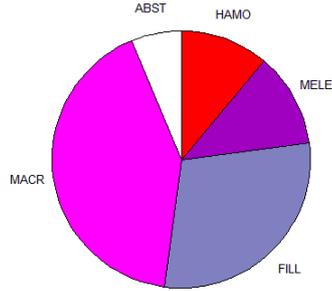
	VOTOS	ABST	BLNU	MACR	LPEN
ABST	11522305	80.0	0.3	12.1	7.6
MACR	8657326	0	0	100	0
LPEN	7679493	0	0	0	100
FILL	7213797	12.5	4.9	82.6	0
MELE	7060885	27.7	26.6	34.9	10.8
HAMO	2291565	0	0	100	0
DUPO	1695186	0	29.8	0	70.2
REST	1460561	0	88.9	0	11.1

A partir de los porcentajes recogidos en la tabla y de los totales de las ocho opciones en la primera vuelta, se calcula inmediatamente la matriz de trasvase de votos entre opciones y, a partir de ella, la procedencia de los votos recibidos por cada una de las cuatro opciones en la segunda vuelta. Estas procedencias se representan gráficamente en la figura 2.7.

Figura 2. 7 Francia: Procedencia de los electores en las 4 opciones de la segunda vuelta en las elecciones presidenciales de 2017.

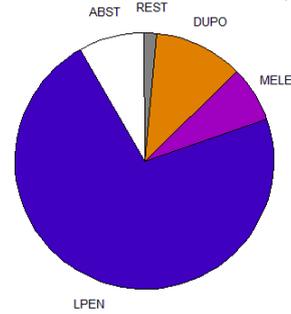


Origen el 26J de los 20760082 votos obtenidos el 25S por MACR



Macron

Origen el 26J de los 10666018 votos obtenidos el 25S por LPEN



Le Pen

Como era de esperar, la práctica totalidad de los electores que votaron a Macron o a Le Pen en la primera vuelta volvieron a votar al mismo candidato en la segunda. Hay que resaltar al respecto, y a riesgo de ser reiterativos, que, aunque este resultado parece muy lógico, en ningún momento *lphom* ha utilizado para obtenerlo la información de que “Macron en la segunda vuelta” es el mismo candidato que “Macron en la primera vuelta”, ni tampoco la información equivalente respecto a Le Pen.

El 84% de los votantes al candidato del centroderecha Fillon en la primera vuelta, lo hicieron a Macron en la segunda, repartiéndose el resto entre la abstención y votos en blanco o nulos, no votando ninguno de estos electores a Le Pen.

Los votantes al populista Melenchon en la primera vuelta dispersaron mucho su voto en la segunda, decidiéndose un 35% por Macron y repartiéndose un 54% a partes iguales entre la abstención y el voto en blanco o nulo. Hubo incluso casi un 11% de votantes a Melenchon que se decidieron en la segunda vuelta por Le Pen.

Los votantes al socialista Hamon en la primera vuelta se decidieron en su totalidad por Macron en la segunda, mientras que el 70% de los votantes al ultranacionalista DuPont en la primera vuelta se decidieron por Le Pen en la segunda, pasándose el 30% restante al voto en blanco o nulo.

Finalmente, del casi millón y medio de electores que votaron por otros candidatos minoritarios en la primera vuelta el 90% votó en blanco o nulo, mientras que el 10% restante lo hizo por Le Pen.

Como consecuencia de los anteriores resultados se deduce que sólo el 41% de los votos que obtuvo Macron en la segunda vuelta procedió de electores que votaron por él en la primera, mientras que un 29% tuvo como origen a exvotantes de Fillon, casi un 12% a los de Melenchon, una cifra muy similar a exvotantes de Hamon y un 6% de electores que se abstuvieron en la primera vuelta.

Por su parte, el 72% de los votos conseguidos por Le Pen en la segunda vuelta vino de electores que ya le votaron en la primera, procediendo el resto sobre todo de exvotantes de DuPont (11%) o Melenchon (7%) y de anteriores abstencionistas (8%).

Aunque es imposible saber hasta qué punto son correctos los resultados obtenidos por *lphom* que acabamos de comentar, puesto que los trasvases reales de votos son imposibles de conocer por la condición de secreto del voto, es indudable que todos ellos son razonables desde el punto de vista sociopolítico.

2.5 COMPARACIÓN CON *ei.MD.bayes*

En el Capítulo 1 se ha indicado la existencia de enfoques alternativos, al aquí expuesto, para abordar el problema de estimar el trasvase de votos de unas opciones a otras entre dos procesos electorales. En particular estos últimos años están adquiriendo cierto auge los procedimientos basados en los enfoques de la denominada Inferencia Ecológica.

Como se ha señalado en el Apartado 1.2, *ei.MD.bayes*

En el Anejo 3 se expone cómo hemos aplicado este modelo *ei.MD.bayes* para analizar los datos de las elecciones francesas, basándonos en los mismos datos utilizados por *lphom* en el apartado anterior, así como los resultados obtenidos.

Hemos encontrado, en primer lugar, que los tiempos de computación exigidos por *ei.MD.bayes* han sido del orden de ¡250 veces superiores! a los necesarios para *lphom*. Pero adicionalmente, y esto es lo realmente importante, los resultados proporcionados por el sofisticado procedimiento *ei.MD.bayes*, han sido totalmente absurdos desde el punto de vista político. Baste mencionar al respecto que, según estos resultados, que se recogen con detalle en el mencionado Anejo 3, menos de la mitad de los votantes a Macron en la primera vuelta habrían vuelto a votarle en la segunda y que de los que votaron a Le Pen en la primera vuelta, hubo más que votaron a Macron en la segunda que aquéllos que volvieron a votar a Le Pen.

Estos malos resultados del método *ei.MD.bayes* los hemos constatado también analizando otros de los procesos recogidos en la Tabla 1 y se deben, al parecer, al hecho de que para que este procedimiento obtenga resultados razonables es necesario que opere con un número muy elevado de unidades territoriales.

2.6 CONCLUSIONES.

El procedimiento *lphom* que hemos presentado en este capítulo es muy sencillo e inteligible y plantea unas exigencias computacionales mínimas.

lphom permite tratar con gran sencillez problemas como el de las altas y bajas en el censo entre las dos elecciones, la incorporación de nuevos partidos o la existencia de ceros estructurales en la matriz de transferencia de votos, cuestiones que resultan a veces difíciles de abordar con los otros métodos.

La hipótesis en la que se basa el modelo, la homogeneidad del comportamiento en las I unidades territoriales en que se particiona el territorio analizado, es muy razonable sociológicamente, especialmente si se elige de forma adecuada el ámbito territorial del estudio y su partición en unidades.

Para los politólogos *lphom* presenta la ventaja adicional de permitir con mucha facilidad incluir en el modelo restricciones que fuercen al resultado a cumplir ciertas condiciones que el investigador considere que deberían verificarse. El problema de actuar así, sin embargo, es que los resultados pierden en parte su carácter objetivo, dependiendo su validez de la de las hipótesis subjetivas impuestas.

A diferencia de lo que sucede con procedimientos mucho más complejos de Inferencia Ecológica, *lphom* ha proporcionado en todos los casos reales en los que se ha aplicado resultados muy razonables desde el punto de vista sociopolítico, incluso operando con un número muy reducido de unidades territoriales.

Subsiste, sin embargo, en *lphom* el problema fundamental de evaluar el margen de incertidumbre de los resultados que obtiene el procedimiento. Este margen es imposible de calcular directamente dado que en

estudios reales no es posible comparar las estimaciones logradas con los resultados que realmente se han producido, al ser estos últimos imposibles de conocer. En los procedimientos de inferencia ecológica esta cuestión (la determinación del margen de error de las estimaciones) se suele resolver explotando las propiedades estadísticas de distribuciones de probabilidad que los sustentan.

En este sentido *lphom* presenta una información adicional que podría ser explotada. La resolución del modelo, además de proporcionar las estimaciones buscadas de los p_{jk} , proporciona también los valores de los residuos e_{ik} , que cuantifican hasta qué punto los resultados realmente observados difieren de los que cabría esperar si la hipótesis de homogeneidad se cumpliera perfectamente.

La explotación de la información contenida en estos residuos e_{ik} , con el fin de evaluar el grado de incertidumbre de los resultados proporcionados por *lphom*, es el objetivo fundamental de la investigación que se desarrolla en los restantes capítulos de esta tesis.

Capítulo 3. Simulación de procesos electorales

3.1 INTRODUCCIÓN

Como hemos señalado, el objetivo de la presente investigación es el desarrollo de un procedimiento para evaluar el margen de incertidumbre asociado a los resultados que proporciona *lphom* al aplicarse en estudios concretos. Se trata, por tanto, de ofrecer una medida de la precisión de las estimaciones obtenidas mediante este procedimiento.

La primera cuestión que abordar, por tanto, consiste en determinar cómo medir el margen de incertidumbre de las estimaciones, es decir, decidir cuáles serían unos índices adecuados para la cuantificación del error asociado a los resultados. En otras palabras, es necesario definir uno o más indicadores que actúen como medidas adecuadas de la diferencia entre el resultado obtenido mediante *lphom* y el resultado real que se pretende estimar.

Una vez definidos dichos indicadores, el problema se trasladará a calcular su valor en estudios concretos. Se plantea, sin embargo, un problema fundamental, que los valores que se pretende estimar son desconocidos en los estudios reales. En efecto, en el Apartado 1.1 de la Introducción ya hemos puesto de manifiesto la escasa fiabilidad de los valores que sobre el particular proporcionan las encuestas postelectorales, cuando están disponibles.

Ante la imposibilidad práctica de conocer los valores verdaderos de trasvases de votos entre procesos electorales reales y, por consiguiente, de medir el grado de error de los resultados proporcionados por *lphom*, necesitamos idear otra estrategia que nos permita aproximar la precisión de los resultados logrados. Como alternativa práctica, emplearemos diferentes conjuntos de resultados simulados, para los que se conozca con exactitud la respuesta, para, tras aplicar *lphom*, aprender sobre las relaciones entre valores reales y estimados. En concreto, a partir de conjuntos de resultados conocidos p_{jk}^i en cada u y, por tanto, de valores p_{jk} promedios en el territorio analizado, la idea consiste en obtener los resultados que proporcionaría el modelo, tras considerar exclusivamente el tipo de información que dispondríamos en una situación real, y compararlos con los valores p_{jk} promedios efectivamente conocidos. Dado que *lphom* se basa en el cumplimiento de la hipótesis de homogeneidad del comportamiento electoral de las I unidades territoriales, los procesos deberán simularse de forma que incumplan en diferentes grados dicha hipótesis de homogeneidad, para con ello poder comprender cómo las consecuencias de dichos incumplimientos se trasladan en los resultados proporcionados por el procedimiento. En el presente capítulo se desarrolla la metodología que hemos desarrollado para resolver esta cuestión.

Previamente, se proponen unos indicadores para evaluar el grado de error de los resultados proporcionados por *lphom* al analizar un proceso electoral concreto, así como otros indicadores que utilizaremos para cuantificar el grado de incumplimiento en un proceso concreto de la hipótesis de homogeneidad del comportamiento electoral en las I unidades territoriales.

3.2 ÍNDICES DE ERROR

3.2.1 Introducción

Cuando se trata de estimar un único parámetro, el error del estimador, o sesgo, es simplemente la diferencia entre el resultado de la estimación y el verdadero valor del parámetro que se desea estimar.

En nuestro caso, sin embargo, los resultados de *lphom* son multidimensionales, consistiendo básicamente en dos matrices estrechamente relacionadas: la matriz **V** de votos trasvasados, cuyo elemento genérico v_{jk} es el número de electores que habiendo votado a la opción j en la elección 1 (E1) votaron a la opción k en la elección 2 (E2), y la matriz **P** de probabilidades de transición, cuyo elemento genérico p_{jk} es la proporción de electores que habiendo elegido a la opción j en la elección 1 (E1) optaron a la opción k en la elección 2 (E2). Ambas matrices son de dimensión $J \times K$, siendo J el número de opciones contempladas en E1 y dicho número en E2.

En esta investigación proponemos la utilización de dos índices, uno para evaluar el grado de error en la estimación de la matriz **V** de trasvase de votos (EI) y otro para evaluar el grado de error en la estimación de la matriz **P** de probabilidades de transferencia (EP).

3.2.2 El índice EI

Sea **V** la matriz $J \times K$ de movilidad electoral real, cuyo elemento genérico v_{jk} son los votos que en la elección simulada han pasado de votar a la opción electoral j en la E1 a votar a la opción electoral k de la Elección 2 y sea **V*** la matriz de movilidad estimada mediante *lphom*. Dada la metodología utilizada, las sumas por filas y las sumas por columnas coinciden en ambas matrices.

Una medida natural de la discrepancia entre **V** y **V*** es el siguiente índice:

$$EI = 100x \frac{0.5 \sum_{jk} |v_{jk} - v_{jk}^*|}{\sum_{jk} v_{jk}}$$

Este índice es básicamente idéntico al índice AD propuesto en Klima et al. (2017).

Es fácil comprobar que EI no es más que el porcentaje de votos cuyo destino ha sido mal estimado por el modelo. El hecho de multiplicar por 0.5 las diferencias en valor absoluto, tiene como objetivo evitar contabilizar dos veces cada voto erróneo.

El coeficiente EI puede variar entre 0, cuando coincidan exactamente **V** y **V***, y 100 cuando no se haya acertado ni un solo voto.

Además de este índice de error global EI, pueden definirse de forma análoga un vector EI_J con los J índices de error correspondientes a las diferentes opciones electorales de la E1 y un vector EI_K con los K índices de error correspondientes las opciones electorales de la E2. Aunque estos índices no los utilizamos en esta tesis, pueden ser interesantes en futuras investigaciones.

La función Matlab *inderr.m* para calcular el EI global, los EI_J para cada partido de E1 y los EI_K para cada partido de E2 se recoge en el Anejo 1.

3.2.3 El índice EP

Este índice se centra en las diferencias entre los valores reales de los p_{jk} y los valores p_{jk}^* estimados por *lphom*. Como hemos indicado, el valor medio de estas diferencias va a ser siempre igual a 0, puesto que la suma de las filas de las matrices \mathbf{P} y \mathbf{P}^* es siempre igual a 1.

Una forma natural de cuantificar el error global cometido al estimar las p_{jk} es hallar el valor medio de los valores absolutos de las diferencias entre las probabilidades de transición reales p_{jk} y las estimadas por el modelo p_{jk}^* .

Definimos de esta forma el índice EP mediante:

$$EP = \frac{\sum_{jk} |p_{jk} - p_{jk}^*|}{J \times K}$$

En nuestra investigación hemos comprobado que, como era de esperar, los índices EI y EP resultan estrechamente correlacionados cuando se calculan simulando diferentes procesos electorales.

Además del valor global para toda la elección, es posible definir estos valores medios para cada opción de la E1 (EPJ) y para cada opción de E2 (EPK), indicadores que podrían ser interesantes para otras investigaciones.

La función Matlab *errp.m* que calcula el EP global, los EPJ para cada opción de la elección E1 y los EPK para cada opción de la elección E2, se recoge en el Anejo 1.

3.3 ÍNDICES DE HETEROGENEIDAD

3.3.1 Introducción

El modelo que se construye en el procedimiento *lphom* se basa en la hipótesis de homogeneidad del comportamiento electoral en las I unidades territoriales (ut) en las que se particiona el territorio objeto del estudio. Dicha hipótesis asume que las matrices $\mathbf{P}^i = [p_{jk}^i]$ de movilidad entre opciones en las distintas unidades territoriales han sido “parecidas” en todas ellas.

Dado que, en la situación hipotética teórica de homogeneidad perfecta el procedimiento *lphom* obtendría con toda seguridad la estimación exacta de la verdadera movilidad, uno de los objetivos fundamentales de este estudio es investigar la influencia sobre la validez de los resultados del modelo del mayor o menor grado de incumplimiento de dicha hipótesis.

Es necesario, por tanto, cuantificar mediante un índice de heterogeneidad el grado de cumplimiento de la hipótesis en unos determinados resultados electorales reales o simulados. Para estudios ulteriores puede ser interesante obtener también de forma diferenciada los índices de heterogeneidad para cada partido origen, para cada partido destino y para cada unidad territorial.

3.3.2 Índice de heterogeneidad real HET

Este índice exige para su cálculo conocer los trasvases de votos que se han producido realmente entre partidos en cada unidad territorial. En consecuencia, sólo puede calcularse para procesos electorales simulados, dado que en los procesos reales dichos trasvases son desconocidos.

Los resultados de partida están en la hipermatriz \mathbf{W} , de dimensiones, $J \times K \times I$, cuyo elemento genérico w_{jki} es el número de votos que ha pasado de la opción electoral j en E1 ($OE1_j$) a la opción electoral k en E2 ($OE2_k$) en la unidad territorial i . En nuestros estudios \mathbf{W} se habrá obtenido mediante simulación.

A partir de \mathbf{W} es posible obtener, colapsándola por la tercera dimensión, la matriz \mathbf{V} de movilidad global y, a partir de ella, la matriz \mathbf{P} de probabilidades de transición globales, o promedias, entre cada $OE1_j$ y cada $OE2_k$. Aplicando \mathbf{P} a los datos de la elección E1 en cada unidad territorial se obtiene la hipermatriz \mathbf{W}^* de trasvase de votos teóricos en cada ut si se hubiera cumplido exactamente la hipótesis de homogeneidad electoral.

Tal como se ha hecho para el índice de error EI, el índice de heterogeneidad se define a partir de la diferencia entre los valores de \mathbf{W} y de \mathbf{W}^* mediante la expresión:

$$HET = 100x \frac{0.5 \sum_{jki} |w_{jki} - w_{jki}^*|}{\sum_{jki} w_{jki}}$$

HET mide, por tanto, el porcentaje de votos cuyo destino habría que modificar para que se cumpliera perfectamente la hipótesis de homogeneidad.

De forma análoga pueden definirse J índices de heterogeneidad HETJ para las opciones electorales de la Elección 1, K índices de heterogeneidad HETK para las diferentes opciones de la Elección 2 e I índices de heterogeneidad HETI para las distintas unidades territoriales.

La función Matlab *indet.m*, que permite calcular los índices de heterogeneidad que acabamos de definir, se recoge en el Anejo 1.

3.3.3 Índice de heterogeneidad estimado HETe

Los índices de heterogeneidad definidos en el apartado anterior tienen el inconveniente de que sólo pueden calcularse si se conocen los trasvases de votos reales que se han producido en cada unidad territorial en el proceso electoral estudiado, conocimiento que (casi) nunca se va a tener en un estudio real.

En este apartado se explota la idea de utilizar los valores de los e_{ik} , que se obtienen como un subproducto de la estimación de la movilidad electoral por el procedimiento *lphom*, para estimar la heterogeneidad existente en un determinado proceso electoral.

Se recuerda que los valores e_{ik} miden la diferencia entre los votos que ha obtenido realmente en la E2 el partido k en la unidad territorial i y los que hubiera obtenido si la matriz de movilidad $\mathbf{P}^i = [p_{jk}^i]$ en la unidad territorial i hubiera sido idéntica a la estimada como promedio para todo el ámbito geográfico, lo que sucedería si se verificase la hipótesis de homogeneidad perfecta.

Dado que es obvio que si se cumpliera exactamente esta hipótesis todos los e_{ik} serían nulos, se propone, en consecuencia, la siguiente definición para el índice de heterogeneidad estimado HETe:

$$HTEe = 100 \frac{\sum_{i,k} |e_{ik}|}{\sum_{i,k} v_{ik}}$$

A partir de esta misma idea se pueden definir de forma inmediata I índices de heterogeneidad estimados HETIe para las distintas unidades territoriales y K índices de heterogeneidad estimados HETKe para las distintas opciones electorales de la E2⁵.

El cálculo de HETe para un proceso real se lleva a cabo dentro de la función *lphom.m* que se recoge en el Anejo 1.

3.3.4 Relación del HETe estimado con el HET real

Como se desarrolla en los siguientes apartados de este capítulo, hemos llevado a cabo un conjunto de simulaciones en diferentes escenarios para analizar la relación entre el EI y el HETe. Como subproducto de estos estudios también se han obtenido en cada simulación los valores del índice de heterogeneidad real HET y del índice de heterogeneidad estimado HETe.

En los cinco escenarios diferentes analizados se ha constatado una estrecha relación entre los dos índices de heterogeneidad. En efecto, en todos los casos el coeficiente de correlación entre los logaritmos de ambas variables ha sido superior a 0.99.

En consecuencia, parece razonable utilizar el HETe, que tiene la ventaja fundamental de poderse calcular en estudios reales, para evaluar el grado de incumplimiento de la hipótesis de homogeneidad en dichos estudios.

3.4 METODOLOGÍA DE LA SIMULACIÓN

En el presente apartado se expone con detalle la metodología que hemos desarrollado para simular procesos electorales, así como la operativa de la función Matlab *simular.m* desarrollada para llevarla a cabo.

Cada simulación se realizará para un determinado “escenario” definido por:

- Número n_j de partidos en la Elección 1 y grado de importancia de estos partidos.
- Número n_k de partidos en la Elección 2.
- Número n_i de unidades territoriales e importancia relativa de éstas.
- Grado de incumplimiento de la hipótesis de homogeneidad electoral.
- Matriz $n_j \times n_k$ de probabilidades de transición reales (fijadas tentativamente al inicio).

⁵ Obsérvese que no es posible definir, a partir de las e_{ik} , índices de heterogeneidad estimados para las J opciones electorales de la E1.

Los resultados fundamentales de la función son el índice de heterogeneidad real de los datos electorales (HET), el índice de heterogeneidad estimado por el modelo (HETe) y los índices de error EI y EP de la estimación obtenida mediante *lphom* de la movilidad electoral.

Datos de entrada

Los datos de entrada, que caracterizan el “escenario” considerado, son:

- Matriz **X** ($n_i \times n_j$) de votos a cada partido en cada unidad territorial en la Elección 1. En esta matriz están implícitos los valores de n_i y de n_j , así como la importancia relativa de los n_j partidos y la de las n_i uts. Al aplicar la metodología *lphom* esta matriz es uno de los datos conocidos exactamente.
- Matriz **P** ($n_j \times n_k$) con los valores tentativos iniciales de las probabilidades de transición entre partidos. Esta matriz lleva implícito el valor de n_k . Los valores finales reales de estas probabilidades, recogidos en la matriz **PR** se calcularán en la simulación, puesto que variarán algo respecto a los iniciales previstos por el carácter aleatorio de la misma. La estimación de **PR** es el objetivo básico de *lphom*.
- Índice **d** para definir el grado de cumplimiento de la hipótesis de homogeneidad electoral. Por ejemplo, si $d = 0.05$ las p_{jk}^i se calculan añadiendo a p_{jk} un valor al azar entre -0.05 y $+0.05$, y luego reajustando para que los p_{jk} estén siempre entre 0 y 1 y sumen 1.

Operativa y resultados de la función

- En primer lugar, se generan aleatoriamente unos posibles resultados del proceso electoral. El resultado básico es la hipermatriz **VG** ($n_j \times n_k \times n_i$) de votos que han pasado de unos partidos a otros en cada ut, que se obtiene a partir de los valores simulados de la hipermatriz **PG** ($n_j \times n_k \times n_i$) de probabilidades de transición en las n_i unidades territoriales obtenida al azar a partir de **P** y de **d**. A partir de **VG** se calculan los siguientes valores:
 - Matriz **Y** ($n_i \times n_k$) de votos a cada partido k en cada unidad i en la Elección 2. Esta matriz va a ser el otro input del modelo *lphom*, dado que es conocida en los estudios reales.
 - Valor real **PR** de la matriz ($n_j \times n_k$) de probabilidades de transición de votos a nivel global. Esta matriz, que debe resultar “próxima” a la matriz **P** propuesta inicialmente, es la que se desea estimar mediante *lphom*.
 - Índices de heterogeneidad global, HET, y a nivel de los partidos, HETJ.
- Seguidamente se lanza *lphom*, utilizando como inputs **X** e **Y**, obteniéndose como resultados la matriz **PM** de probabilidades de transición estimada, el vector con los residuos e_{ik} y la suma de sus valores. A partir de esta suma se calcula el índice HETe de heterogeneidad estimado por el modelo. A diferencia del HET real, que es desconocido, el HETe es uno de los resultados proporcionados por *lphom*.
- Comparando los resultados de trasvase de votos obtenidos por el modelo con los reales se calcula el índice de error global **EI** de los mismos, así como con los índices de error diferenciados para cada partido **EIJ**.
- Comparando **PR** con **PM** se obtiene también los índices de error promedio en la estimación de las p_{jk} **EP** y **EPJ**.

- Si se desea, la función *simular.m* lista parte de los resultados obtenidos, pero todos ellos quedan guardados y son accesibles para su posterior utilización por otras funciones.

La función *simular.m* y las funciones auxiliares que utiliza (*azarpg.m*, *azarpi.m*, *azar.m*, *matvg.m*, *datsimu.m* y *calcmodel.m*) se recogen en el Anejo 1.

Capítulo 4. Estudio de la relación entre el índice de error EI y el índice de heterogeneidad estimado HETe

4.1 INTRODUCCIÓN

En el presente capítulo se utiliza la metodología de simulación descrita en el capítulo anterior para establecer la naturaleza del modelo que permita estimar el EI de un estudio concreto a partir del HETe en el mismo.

La idea básica previa es que si en un estudio resulta HETe igual a cero, o sea si se cumple estrictamente la hipótesis de homogeneidad del comportamiento electoral en las I uts, el EI será necesariamente cero. Parece razonable pensar, asimismo, que cuánto mayor sea el grado de incumplimiento de esta hipótesis de homogeneidad y, por tanto, mayor sea el valor del HETe, mayor va a ser el valor esperado del EI.

Para modelar esta relación, en este capítulo hemos investigado el comportamiento de la relación en cinco escenarios diferentes. Los escenarios están inspirados en s procesos concretos en los que se ha aplicado *lphom*.

En concreto, dado un determinado escenario, se simulan un conjunto de posibles resultados electorales con diferentes grados de incumplimiento de la hipótesis de homogeneidad, obteniéndose en cada caso los valores del índice de heterogeneidad estimado HETe y del índice de error EI de la estimación de la matriz de movilidad electoral obtenida mediante el procedimiento *lphom*. La ventaja de esta aproximación radica en que al partir de escenarios simulados todos los estadísticos básicos, incluidos los resultados reales del escenario, son conocidos.

La función Matlab *yasimular.m*, que se recoge en el Anejo 1, se utiliza para llevar a cabo las simulaciones mediante la instrucción:

$$[ie, ep, he, hr] = \text{yasimular}(x, p, id, nv)$$

Los inputs “x” y “p” definen el escenario del proceso electoral. El vector “id” contiene los nd valores de “d”⁶ que se utilizarán en cada simulación para generar aleatoriamente la hipermatriz de movilidad electoral en las I unidades territoriales. A valores crecientes de “d” le corresponden valores crecientes del grado de incumplimiento de la hipótesis de homogeneidad. Para cada valor de “d” se realizan nv , donde hemos fijado en lo que sigue $nv = 30$, simulaciones.

En total, se han estudiado en cada escenario 1380 simulaciones cuyos resultados se guardan en los vectores outputs correspondientes. En particular, los EI se guardan en *ie* y los HETe en *he*⁷.

⁶ Los valores de “d” en *id* varían entre 0.001 y 0.010 con incrementos de 0.001, y entre 0.010 y 0.100 con incrementos de 0.0025.

⁷ Datos depositados en: <https://www.openicpsr.org/openicpsr/workspace?goToPath=/openicpsr/120303&goToLevel=project#>

4.2 MODELACIÓN DE LA RELACIÓN ENTRE EI y HETe

Para analizar la relación entre el EI y el HETe se plantea un modelo de regresión clásico:

$$E(EI/HETe=x) = f(x)$$

Como siempre, se ha de verificar que cuando HETe es igual a cero, el EI es también igual a cero, la función $f(x)$ deberá necesariamente verificar que $f(0)=0$. Además, deberá también verificarse que $f(x)$ sea monótonamente creciente, al menos en el intervalo de valores considerado para HETe, pues $E(EI)$ no puede disminuir al aumentar HETe.

Una expresión analítica para $f(x)$ que satisface los requisitos exigidos y se ha adaptado muy bien a los resultados obtenidos en los cinco escenarios analizados presenta la especificación definida por la ecuación (4.1):

$$\log(f(x)) = \beta_0 + \beta_1 \log(x) + \beta_2 [\log(x)]^2 \quad (4.1)$$

Como en los cinco escenarios analizados la relación entre $\log(EI)$ y $\log(HETe)$ tiene varianza residual aproximadamente constante, el modelo (4.1) puede estimarse simplemente mediante mínimos cuadrados ordinarios utilizando $\log(EI)$ como variable dependiente (o respuesta) y $\log(HETe)$ y $[\log(HETe)]^2$ como variables explicativas (o predictores).

En los cinco casos, los coeficientes de correlación múltiple obtenidos en los ajustes han sido elevados, lo que justifica utilizar el modelo (4.1) para predecir el índice de error EI. El valor predicho por el modelo será un índice del error que cabe esperar obtener al estimar la movilidad electoral dado el valor constatado en el estudio para el índice de heterogeneidad estimado HETe.

En los siguientes apartados se describe cada uno de los escenarios propuestos y se presentan los resultados obtenidos al estudiar en cada uno de ellos la relación entre el EI y el HETe.

4.3 RESULTADOS OBTENIDOS EN EL ESCENARIO 1

4.3.1 Descripción del Escenario 1

Este escenario está inspirado en los procesos electorales analizados en Andalucía asumiendo ocho unidades territoriales ($I=8$) y seis opciones electorales tanto en la Elección 1 ($J=6$) como en la Elección 2 ($K=6$).

Los detalles que definen el Escenario 1 se recogen en el Anejo 3.

4.3.2 Resultados

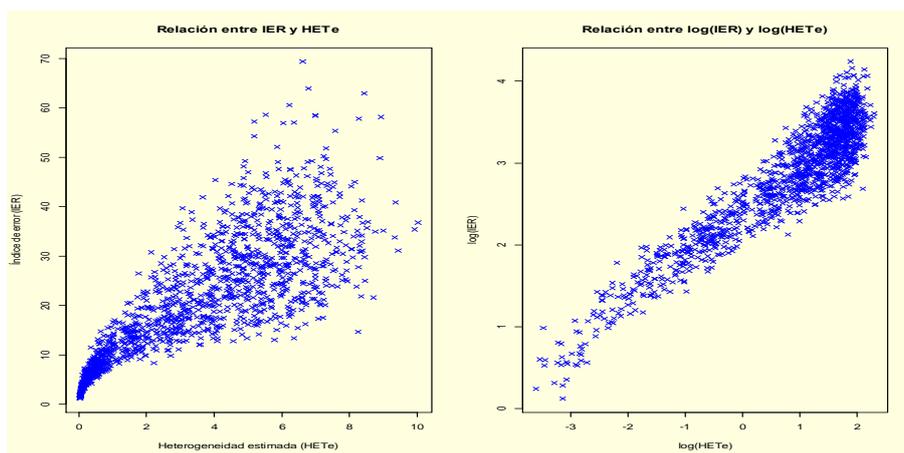
Los 1380 valores simulados de HETe han oscilado entre un mínimo de 0.03% y un máximo de 10.03%.

Como se aprecia en la siguiente figura, se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Recuérdese que al trabajar con escenarios simulados se

dispone de los valores reales y de las estimaciones obtenidas con *lphom*, con lo que es posible calcular el valor de EI asociado. Esta relación es creciente y ligeramente no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

En la figura se grafica también la relación cuando las dos variables se expresan en escala logarítmica. Se constata que, en esta escala, la no linealidad casi desaparece y que la varianza residual es aproximadamente constante.

Figura 4.1 Escenario 1: Relación entre EI y HETe.



Ajustando el Modelo (4.1) a los valores de EI y HETe obtenidos en las simulaciones se obtiene el siguiente resultado:

Tabla 4.1 Ajuste de regresión del Modelo (4.1).

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	2.4208142	0.0126369	191.567	<0.0001
log(HETe)	0.5385672	0.0057109	94.304	<0.0001
log(HETe)^2	0.0006178	0.0040740	0.152	0.879
coef. correlación múltiple: 0.9306				
desv. tip. residual: 0.2696				

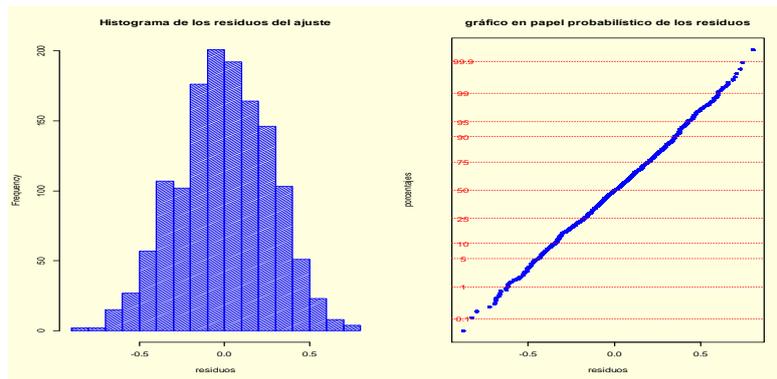
El modelo estimado es, por tanto:

$$\log(EI) = 2.421 + 0.539 \log(HETe) - 0.0006(\log(HETe))^2 + u \quad (4.2)$$

siendo la desviación típica residual 0.27 y no resultando significativo, en este caso, el término cuadrático del ajuste.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

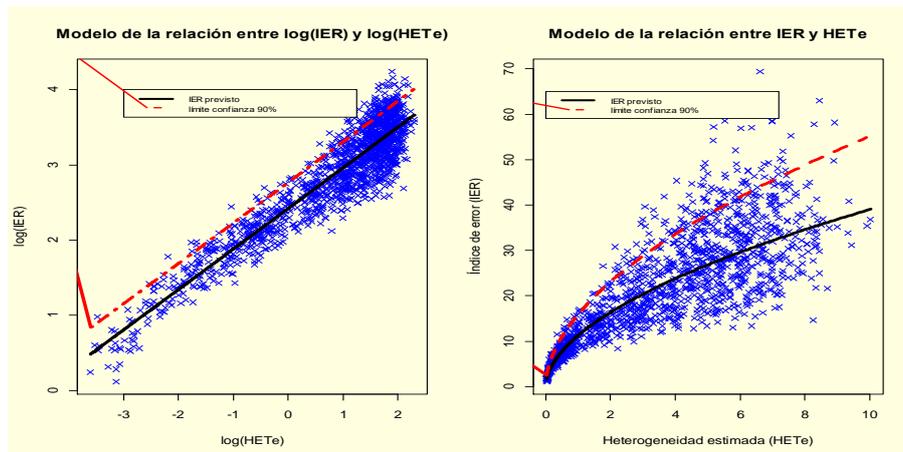
Figura 4.2 Escenario 1: Histograma y gráfico en papel probabilístico normal de los residuos.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible, por lo que la distribución normal puede utilizarse para obtener límites de confianza para las predicciones proporcionadas por el modelo para los EI en función de los valores constatados para HETe.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI, así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 4.3 Escenario 1: Modelo ajustado.



4.4 RESULTADOS OBTENIDOS EN EL ESCENARIO 2

4.4.1 Descripción del Escenario 2

Este escenario está inspirado en los procesos electorales analizados en la Comunidad Valenciana y en él se contemplan catorce unidades territoriales ($I=14$), ocho opciones electorales en la Elección 1 ($J=8$) y siete opciones electorales en la Elección 2 ($K=7$).

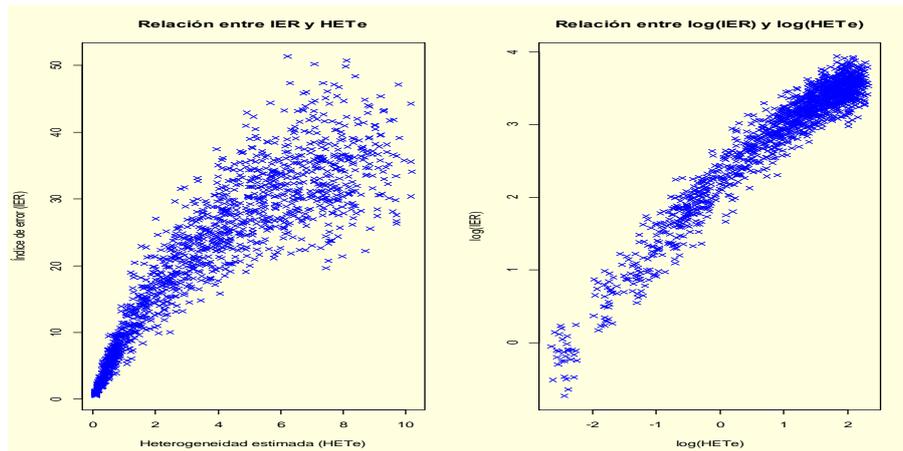
Los detalles que definen el Escenario 2 se recogen en el Anejo 3.

4.4.2 Resultados

Los 1380 valores simulados de HETe han oscilado entre un mínimo de 0.07% y un máximo de 10.20%.

En la siguiente figura se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe. Por su parte la relación obtenida entre los valores de $\log(EI)$ y los de $\log(HETe)$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es en este caso aproximadamente constante.

Figura 4.4 Escenario 2: Relación entre EI y HETe.



Ajustando los datos resultantes de la simulación al Modelo (4.1) se obtiene el siguiente resultado:

Tabla 4. 2 Ajuste de regresión del Modelo (4.1).

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	2.259505	0.009009	250.82	<0.0001
$\log(HETe)$	0.804330	0.005091	157.98	<0.0001
$\log(HETe)^2$	-0.085560	0.003648	-23.45	<0.0001

coef. correlación múltiple: 0.9759
 desv. tip. residual: 0.1981

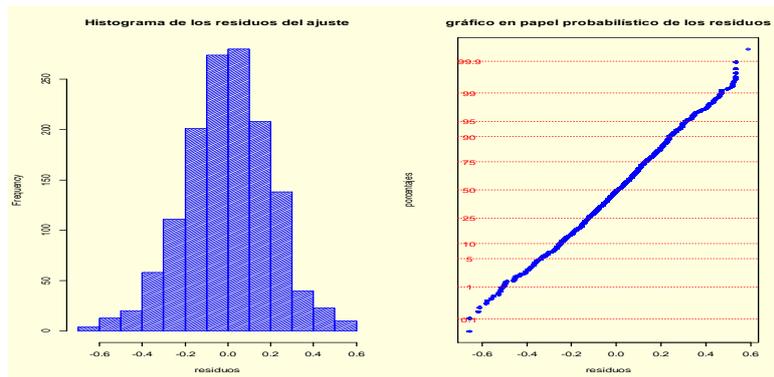
El modelo estimado es, por tanto:

$$\log(EI) = 2.260 + 0.804\log(HETe) - 0.0856(\log(HETe))^2 + u \quad (4.3)$$

siendo la desviación típica residual 0.198 y resultando muy significativo en este caso el término cuadrático del ajuste.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal,

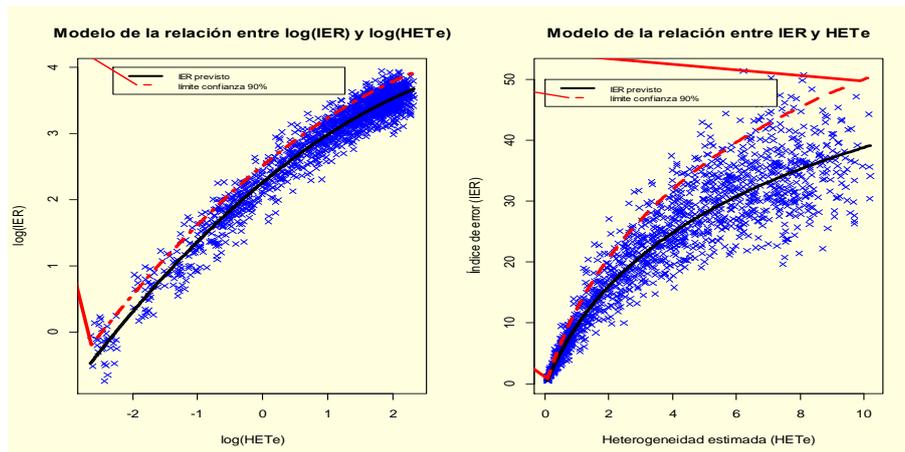
Figura 4.5 Escenario 2: Histograma y gráfico en papel probabilístico normal de los residuos.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible, por lo que la distribución normal puede utilizarse para obtener límites de confianza para las predicciones proporcionadas por el modelo para los EI en función de los valores constatados para HETe.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 4.6 Escenario 2: Modelo ajustado.



4.5 RESULTADOS OBTENIDOS EN EL ESCENARIO 3

4.5.1 Descripción del Escenario 3

Este escenario está inspirado en los procesos electorales analizados en el municipio de Madrid y tiene veinte unidades territoriales ($I=20$), siete opciones electorales en la Elección 1 ($J=7$) y seis opciones electorales en la Elección 2 ($K=6$).

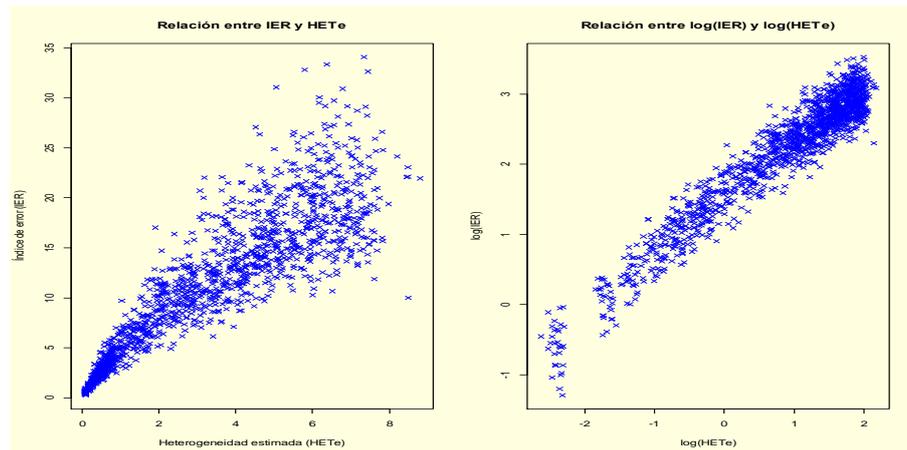
Los detalles que definen el Escenario 3 se recogen en el Anejo 3.

4.5.2 Resultados

Los 1380 valores simulados de HETe han oscilado entre un mínimo de 0.07% y un máximo de 8.79%.

Como se aprecia en la siguiente figura, se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

Figura 4.7 Escenario 3: Relación entre EI y HETe.



Por su parte la relación obtenida entre los valores de $\log(\text{EI})$ y los de $\log(\text{HETe})$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es, en este caso, aproximadamente constante.

Ajustando los datos obtenidos en la simulación al Modelo (4.1) se obtiene el siguiente resultado:

Tabla 4.3 Ajuste de regresión del Modelo (4.1)

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	1.596002	0.010793	147.869	<0.0001
$\log(\text{HETe})$	0.798867	0.006155	129.797	<0.0001
$\log(\text{HETe})^2$	-0.046938	0.004827	-9.725	<0.0001

coef. correlación múltiple: 0.9645
desv. tip. residual: 0.2342

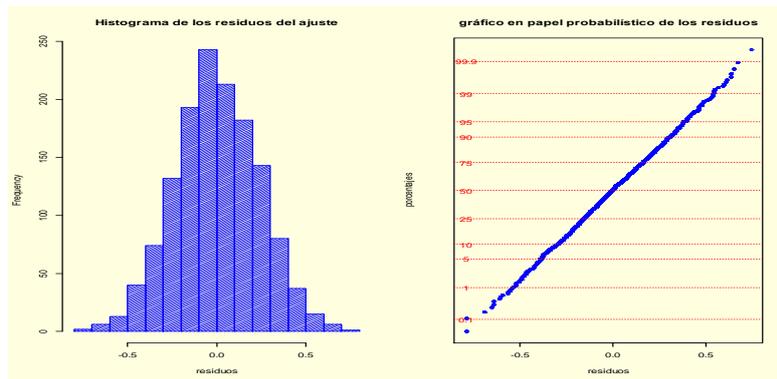
El modelo estimado es, por tanto:

$$\log(\text{EI}) = 1.596 + 0.799\log(\text{HETe}) - 0.0469(\log(\text{HETe}))^2 + u \quad (4.4)$$

siendo la desviación típica residual 0.2342 y resultando muy significativo en este caso el término cuadrático del ajuste.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

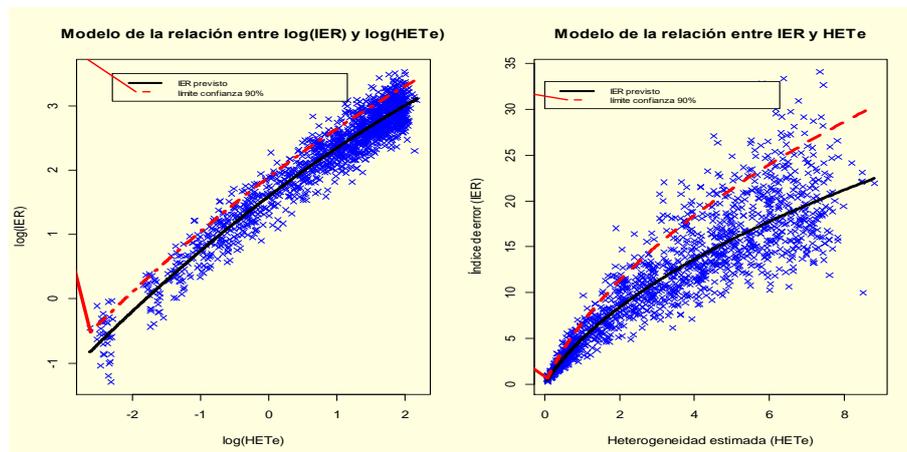
Figura 4.8 Escenario 3: Histograma y gráfico en papel probabilístico normal de los residuos.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible, por lo que la distribución normal puede utilizarse para obtener límites de confianza para las predicciones proporcionadas por el modelo para los EI en función de los valores constatados para HETe.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI, así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 4.9 Escenario 3: Modelo ajustado.



4.6 RESULTADOS OBTENIDOS EN EL ESCENARIO 4

4.6.1 Descripción del Escenario 4

Este escenario está inspirado en un estudio sobre la movilidad electoral entre las elecciones autonómicas catalanas de 2012 y 2015 y viene caracterizado por tener diez unidades territoriales ($I=10$), diez opciones electorales en la Elección 1 ($J=10$) y nueve opciones electorales en la Elección 2 ($K=9$).

Los detalles que definen el Escenario 4 se recogen en el Anejo 3.

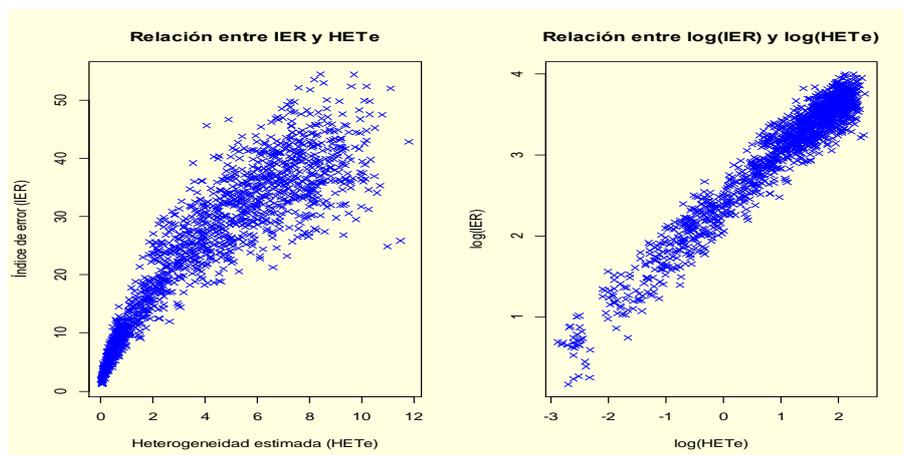
4.6.2 Resultados

Los 1380 valores simulados de HETe han oscilado entre un mínimo de 0.06% y un máximo de 11.79%.

Como se aprecia en la siguiente figura, se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

Por su parte, la relación obtenida entre los valores de $\log(EI)$ y los de $\log(HETe)$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es, en este caso, aproximadamente constante.

Figura 4.10 Escenario 4: Relación entre EI y HETe.



Ajustando los datos obtenidos en la simulación al Modelo (1) se obtiene el siguiente resultado:

Tabla 4. 4 Ajuste de regresión del Modelo (4.1).

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	2.441783	0.008497	287.367	<0.0001
$\log(HETe)$	0.637582	0.004473	142.546	<0.0001
$\log(HETe)^2$	-0.026456	0.003111	-8.504	0.879

coef. correlación múltiple: 0.9719
 desv. tip. residual: 0.1813

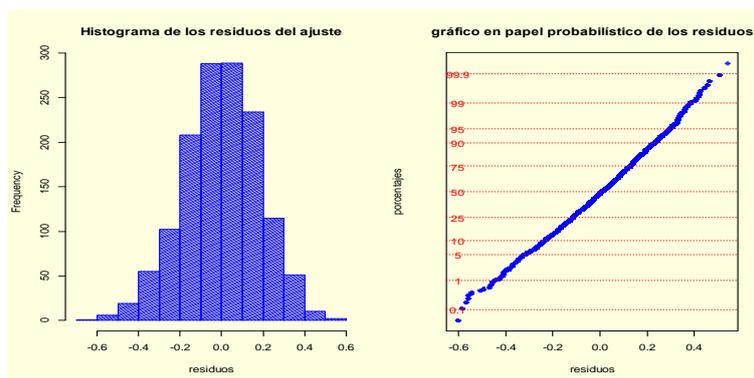
El modelo estimado es, por tanto:

$$\log(EI) = 2.442 + 0.638\log(HETe) - 0.0264(\log(HETe))^2 + u \quad (4.5)$$

siendo la desviación típica residual 0.1813 y no resultando significativo el término cuadrático del ajuste.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

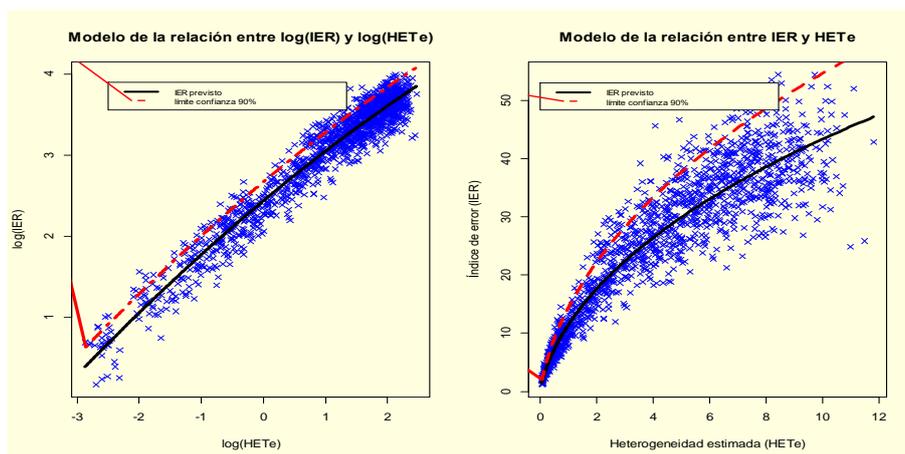
Figura 4.11 Figura 4.11. Escenario 4: Histograma y gráfico en papel probabilístico normal de los residuos.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 4.12 Escenario 4: Modelo ajustado.



4.7 RESULTADOS OBTENIDOS EN EL ESCENARIO 5

4.7.1 Descripción del Escenario 5

Este escenario está inspirado en el estudio de la movilidad electoral en Aragón entre las elecciones autonómicas de 2011 y 2015, y viene caracterizado por tener quince unidades territoriales ($I=15$), nueve opciones electorales en la Elección 1 ($J=9$) y once opciones electorales en la Elección 2 ($K=11$).

Los detalles que definen el Escenario 5 se recogen en el Anejo 3.

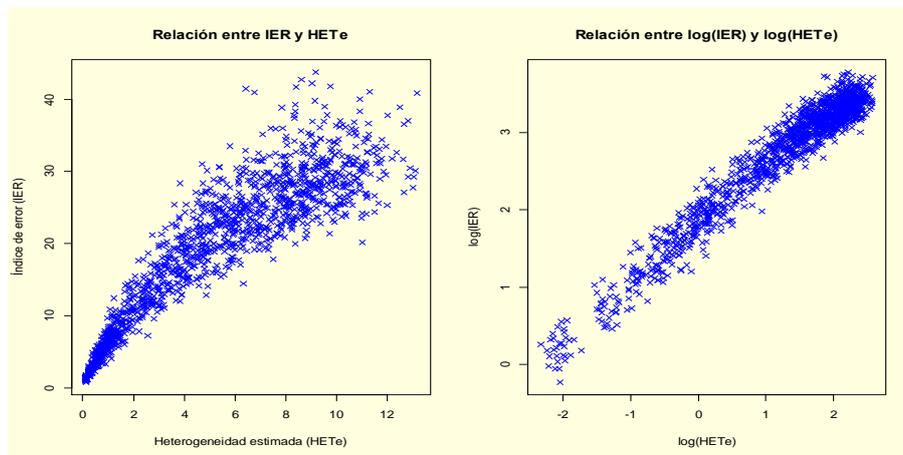
4.7.2 Resultados

Los 1380 valores simulados de HETe han oscilado entre un mínimo de 0.10% y un máximo de 13.17%.

Como se aprecia en la siguiente figura, se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

Por su parte la relación obtenida entre los valores de $\log(EI)$ y los de $\log(HETe)$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es en este caso aproximadamente constante.

Figura 4.13 Escenario 5: Relación entre EI y HETe.



Ajustando los datos obtenidos en la simulación al Modelo (4.1) se obtiene el siguiente resultado:

Tabla 4.5 Ajuste de regresión del Modelo (4.1)

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	1.887487	0.008274	228.13	<0.0001
$\log(HETe)$	0.765749	0.006058	126.41	<0.0001
$\log(HETe)^2$	-0.044170	0.003489	-12.66	0.879

coef. correlación múltiple: 0.9762
desv. tip. residual: 0.1741

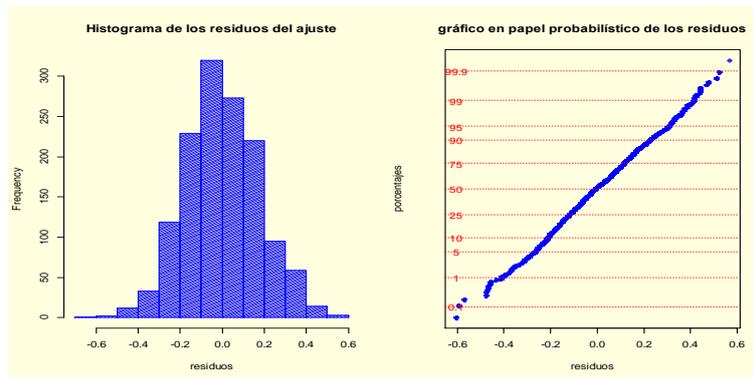
El modelo estimado es, por tanto:

$$\log(EI) = 1.887 + 0.766\log(HETe) - 0.0442(\log(HETe))^2 + u \quad (4.6)$$

siendo la desviación típica residual 0.1741 y no resultando significativo el término cuadrático del ajuste.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

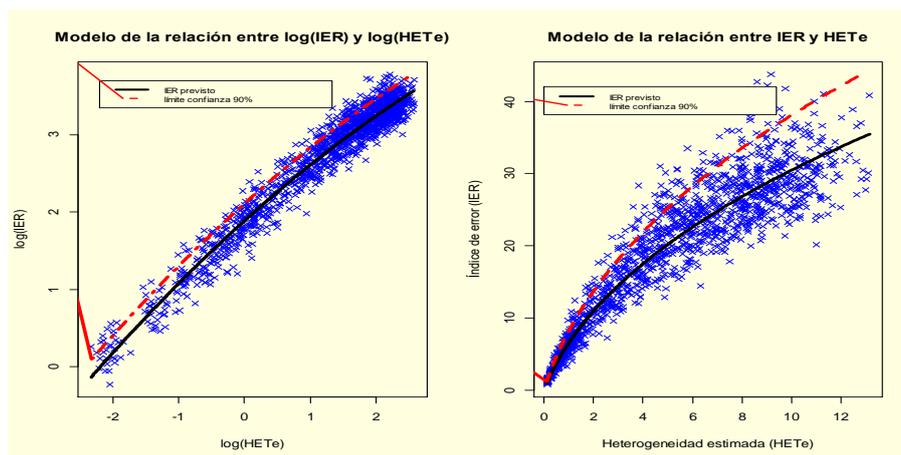
Figura 4.14 Escenario 5: Histograma y gráfico en papel probabilístico normal de los residuos.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 4.15 Escenario 5: Modelo ajustado.



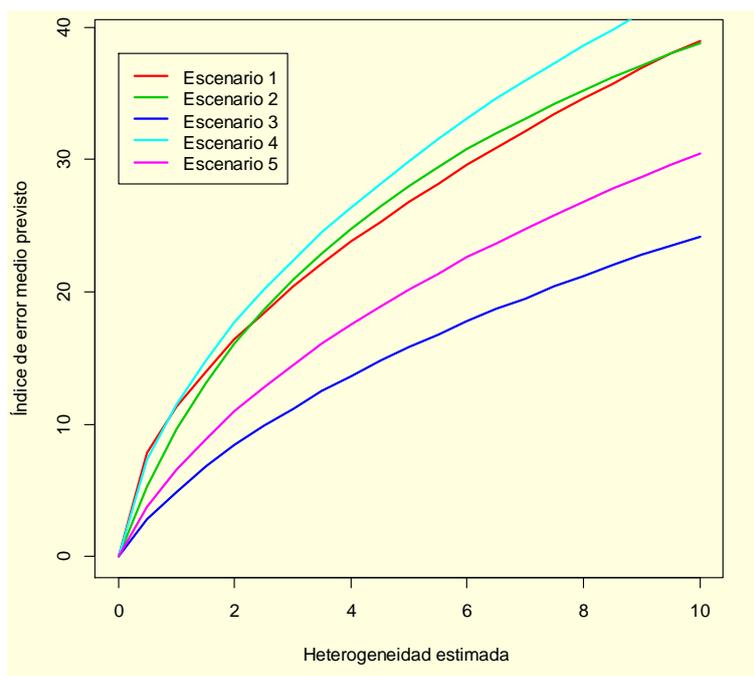
4.8 SÍNTESIS DE LOS RESULTADOS OBTENIDOS EN LOS CINCO ESCENARIOS

En los cinco escenarios analizados se ha constatado una estrecha relación entre el Índice de Error EI de los resultados de aplicar *lphom* y el valor obtenido para el Índice de Heterogeneidad Estimada HETe.

En los cinco casos, esta relación se modela satisfactoriamente mediante una ecuación de regresión utilizando como variable dependiente $\log(\text{EI})$ y como variables explicativas $\log(\text{HETe})$ y su cuadrado. Los elevados valores obtenidos para el coeficiente de correlación múltiple: 0.931 en el Escenario 1, 0.976 en el Escenario 2, 0.965 en el Escenario 3, 0.972 en el Escenario 4 y 0.976 en el Escenario 5, ponen de manifiesto que HETe es un buen predictor del EI.

La siguiente figura refleja los valores medios previstos para el EI en función de la HETe en los cinco escenarios analizados.

Figura 4.16 Relación entre EI medio previsto y HETe en los cinco escenarios.



Se constata que, aunque la forma general de la relación es muy similar en todos los escenarios, los valores concretos de la misma difieren sensiblemente de unos casos a otros.

Este resultado es importante, y será tenido en cuenta en el capítulo siguiente al proponer un procedimiento general para estimar el EI en estudios electorales concretos.

4.9 ANÁLISIS PROVISIONAL DE LA INFLUENCIA DEL ESCENARIO SOBRE LA RELACIÓN ENTRE EL ÍNDICE DE HETEROGENEIDAD Y EL ÍNDICE DE ERROR

4.9.1 Introducción

Los análisis anteriores han puesto de manifiesto en cada escenario una estrecha relación entre los EI medios previstos y el índice de heterogeneidad HETe considerado.

Sin embargo, la Figura 4.16 pone de relieve que dicha relación cambia sensiblemente en función del escenario analizado, de forma que para un mismo valor del HETe el EI medio previsto difiere bastante de unos escenarios a otros. Por ejemplo, para un HETe del 1%, el EI previsto varía desde un 5% en el Escenario 3 a más del 11% en el Escenario 4.

¿Qué características del escenario son las responsables de estas diferencias?

Desde el punto de vista del objetivo fundamental de esta tesis —estimar el EI que corresponde a los resultados obtenidos al aplicar *lphom* en un proceso electoral concreto— esta cuestión no es muy relevante, ya que tal estimación se llevará a cabo, como veremos en el capítulo siguiente, realizando un estudio de simulación partiendo un escenario similar al del proceso electoral analizado. Sin embargo, creemos que vale la pena intentar profundizar en las características que definen cada escenario que pueden influir sobre los EI resultantes. Esto permite disponer de un mejor conocimiento del funcionamiento del método *lphom* y de los factores que condicionan la precisión de los resultados que

proporciona. En este sentido, las ideas que se desarrollan en este apartado pueden ser consideradas como un avance para futuras investigaciones.

En este apartado se especula, por tanto, sobre las características del escenario que pueden afectar al índice de error correspondiente a un HETe dado. Se proponen al respecto dos características concretas y se analiza su relación con los EI obtenidos. Los resultados de este apartado deben considerarse con precaución y tienen solo un carácter provisional, dado que han sido obtenidos considerando sólo un número muy limitado de escenarios diferentes.

4.9.2 Dos posibles características relevantes

El razonamiento que ha llevado a proponer el índice de heterogeneidad como un factor relevante al estudiar el EI del método se basa en el hecho de saber que si HETe fuera cero el EI sería con toda seguridad nulo.

Un razonamiento similar, basado en situaciones extremas, puede ayudar a sugerir otros posibles factores de interés, tal como se discute a continuación.

4.9.2.1 Ratio REI ecuaciones/incógnitas

Si sólo se dispusiera de los resultados globales en el territorio analizado, el problema tendría $J \times K$ incógnitas, las p_{jk} , que deberían satisfacer las $J + K$ ecuaciones básicas del modelo. Como ya se ha comentado en el Capítulo 2, al haber más incógnitas que ecuaciones se tendría un sistema indeterminado con infinitas soluciones posibles. Al disponerse de los resultados desagregados en I unidades territoriales, se tienen $I \times K$ ecuaciones adicionales que, bajo la hipótesis de homogeneidad, aportan información adicional sobre las p_{jk} .

Es obvio que si I fuera muy pequeña, de forma que $J+K+I \times K$ resultara menor que $J \times K$, el sistema seguiría siendo indeterminado con infinitas soluciones posibles. Parece intuitivo, en consecuencia, que cuanto mayor sea I , más información se aportará sobre las p_{jk} y que, por tanto, mejores serán, todo lo demás constante, las estimaciones proporcionadas por *lphom*.

Para cuantificar la situación al respecto en un determinado escenario se propone utilizar la ratio entre el número de ecuaciones del modelo y el número de incógnitas a estimar. Provisionalmente denominamos REI (ratio ecuaciones/incógnitas) a este índice:

$$REI = \frac{J+K+I \times K}{J \times K} \quad (4.7)$$

Nuestra especulación es que valores altos de REI deben ir asociados a menores valores del EI.

4.9.2.2 Índice DIV de diversidad de las unidades territoriales

Por otra parte, es obvio que si la estructura del voto en la Elección 1, es decir, la proporción de votos obtenidos por cada uno de los J partidos, hubiera sido idéntica en las I unidades territoriales, las $I \times K$ ecuaciones no serían linealmente independientes y el sistema seguiría siendo indeterminado por elevado que fuera I .

En consecuencia, parece intuitivo que convendría que la estructura del voto en la Elección 1 fuera lo más dispar posible en las diferentes unidades territoriales, con el fin de estimar mejor los valores de las p_{jk} .

Con el fin de cuantificar este grado de disparidad se propone utilizar un índice, al que provisionalmente denominamos DIV, definido como:

$$DIV = 100 \frac{\sum_{ij} |x_{ij} - x_i p_{j.}|}{\sum_{ij} x_{ij}} \quad (4.8)$$

donde x_i es el total de electores en la unidad territorial i y $p_{j.}$ es la proporción de votos obtenida por el partido j en la Elección 1 en el total del territorio.

También en este caso nuestra especulación es que valores altos de DIV deben ir asociados a menores valores del EI.

Conviene resaltar la diferencia entre HET y DIV. El índice de heterogeneidad HET mide el grado de incumplimiento de la hipótesis de homogeneidad que postula que las matrices de movilidad $[p_{jk}^i]$ han sido parecidas en las I uts. Por su parte, DIV cuantifica el grado de disparidad en la estructura del voto en las I uts en la Elección 1. Mientras que al aumentar HET aumentan los EI, nuestra especulación es que al aumentar DIV los EI disminuyen.

4.9.3 Relación de REI y DIV con el índice de error EI

Aunque sólo se dispone por el momento de información sobre cinco escenarios, hemos querido analizar hasta qué punto se verifican en los mismos las hipótesis avanzadas en el apartado anterior.

Para ello se ha utilizado como variable dependiente EI1, que es el EI medio previsto en cada escenario cuando HETe es igual al 1%. Estos valores se han obtenido a partir de las ecuaciones estimadas en los Apartados 4.3 a 4.7 del presente capítulo.

Como posibles variables explicativas se han utilizado los valores de REI y de DIV calculados para los cinco escenarios estudiados. Esto valores se recogen en la siguiente tabla junto con los del EI previsto para HETe igual al 1%.

Tabla 4. 6. Valores de EI1, los índices REI y DIV en los 5 escenarios

	Escenario 1	Escenario 2	Escenario 3	Escenario 4	Escenario 5
EI1 [†]	11.2	9.6	5.0	11.5	6.6
Índice REI	1.67	2.02	3.17	1.21	1.87
Índice DIV	10.3	13.9	17.1	17.5	19.1

[†] Valor medio previsto para HETe = 1%

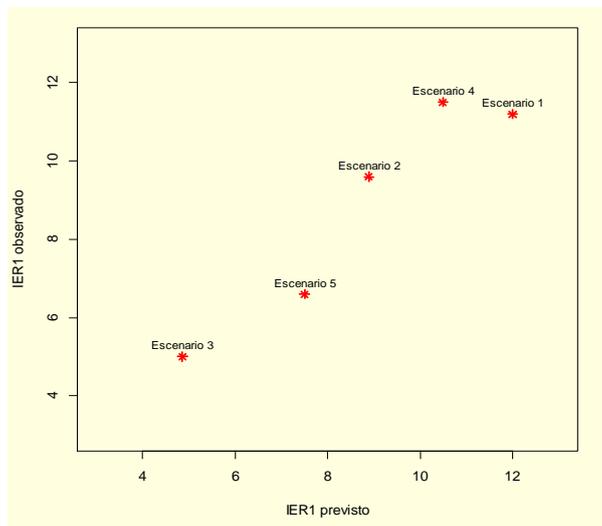
Un ajuste lineal del log(EI1) a REI y DIV ha proporcionado el siguiente resultado:

Tabla 4. 7 Ajuste de regresión de log(EI1).

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	3.60950	0.33599	10.743	0.00855
REI	-0.40097	0.09343	-4.292	0.05024
DIV	-0.04423	0.01941	-2.279	0.15033
coef. correlación múltiple: 0.9650				

El elevado coeficiente de correlación múltiple obtenido (0.965), así como el siguiente gráfico de los valores observados frente a los previstos, pone de manifiesto una clara relación del EI con los índices REI y DIV propuestos, aunque la significación estadística de la misma no sea muy elevada dado que sólo hemos dispuesto de cinco observaciones.

Figura 4.17 EI1 observado frente al previsto a partir de REI y DIV.



En la salida del ajuste se constata que los signos de los coeficientes de ambas variables explicativas resultan negativos, tal como se había especulado al razonar su posible efecto sobre el EI. Por otra parte, los valores obtenidos para los t-ratios (-4.29 y -2.28) han sido bastante elevados, aunque el escasísimo número de datos hace que los correspondientes p-valores sean claramente no significativos, especialmente en lo relativo al coeficiente asociado a DIV.

Los resultados anteriores parecen avalar la hipótesis de que las características sintetizadas por los índices REI y DIV tienen efectos relevantes sobre la precisión de las estimaciones obtenidas mediante el procedimiento *lphom*. Investigaciones adicionales, analizando un número más elevado de escenarios son necesarias para confirmar y cuantificar estos efectos.

Capítulo 5. Un procedimiento para estimar el índice de error (EI) en estudios reales

5.1 INTRODUCCIÓN

En el presente capítulo se aborda uno de los objetivos fundamentales planteados en nuestra investigación: el desarrollo de un procedimiento que permita estimar el índice de error, EI, correspondiente a los resultados proporcionados por *lphom* al estimar la matriz de transferencia de votos en estudios electorales concretos.

Tal como se expone en el siguiente apartado, el procedimiento se fundamenta en los resultados obtenidos en el capítulo anterior sobre la relación entre el índice de error EI y el índice de heterogeneidad estimado HETe.

Para facilitar la aplicación de este procedimiento se ha desarrollado la función Matlab *estimei.m*, que se expone con detalle en el Apartado 5.3.

Finalmente, en un último apartado, se aplica el procedimiento descrito para estimar el EI, en el estudio sobre las transferencias entre la primera y la segunda vuelta de las elecciones presidenciales francesas de 2017, estudio que fue analizado al final del Capítulo 2 de esta tesis.

5.2 PROCEDIMIENTO PROPUESTO

En el Capítulo 4 se ha puesto de manifiesto, en diferentes escenarios, la existencia de una fuerte vinculación entre el indicador HETe que *lphom* permite obtener en un estudio y el EI que se podría obtener, cuando los resultados son conocidos, de los resultados proporcionados por *lphom* en ese mismo estudio concreto. En el mismo capítulo se ha puesto de manifiesto, sin embargo, que, si bien la relación funcional entre EI y HETe es muy similar en todos los casos, los parámetros concretos de la ecuación que modela esta relación dependen estrechamente del escenario considerado.

Desde el punto de vista práctico, este resultado pone de manifiesto que, para evaluar el EI en un estudio concreto, a partir del índice de heterogeneidad estimado HETe obtenido, será necesario estimar la función que relaciona el EI con el HETe en el caso de estudio. Para ello, la propuesta de esta investigación es emplear un estudio de simulación similar a los cinco presentados en el Capítulo 4, pero utilizando el escenario particular que definen los datos del estudio concreto que se está analizando.

El escenario del caso de estudio vendrá caracterizado por la matriz \mathbf{X} de votos obtenidos en la Elección 1 por los J partidos en las I unidades territoriales, y por la matriz \mathbf{P} estimada en el estudio para las probabilidades de transición de votos de unas opciones a otras entre las dos elecciones. Los valores de “d” a utilizar en las simulaciones se elegirán de forma que conduzcan a índices de heterogeneidad estimados cercanos al obtenido en el estudio.

Una vez obtenida la relación de EI con HETe en el escenario considerado, el valor estimado del EI será el que corresponda a un HETe igual al obtenido en el estudio.

Con el fin de aplicar fácilmente este procedimiento, se ha desarrollado la función Matlab *estimei.m*, cuya operativa se describe a continuación.

5.3 LA FUNCIÓN *estimei.m*

El listado de la función *estimei.m* que permite en estudios concretos la estimación del EI asociado a los resultados obtenidos por *lphom*, incluyendo la obtención de un límite de confianza superior para el mismo, se recoge en el Anejo 1, junto con el de las funciones auxiliares que utiliza.

La función se lanza mediante la instrucción:

$$[ei, limei, he0, div, rei] = \mathbf{estimei}(x, y, nv, conf)$$

Los argumentos de la función son:

- x : matriz \mathbf{X} ni x nj con los resultados de la Elección 1.
- y : matriz \mathbf{Y} ni x nk con los resultados de la Elección 2.
- $conf$: nivel de confianza deseado para obtener el límite superior del intervalo de confianza.
- nv : número de simulaciones a realizar para cada uno de los 11 valores considerados de d . El valor recomendado de nv es 100.

Los resultados fundamentales que proporciona *estimei.m* son:

- ei : valor estimado del índice de error EI del estudio.
- $limei$: límite de confianza superior (para un nivel de confianza $conf$) para el EI.
- Adicionalmente *estimei.m* proporciona otros resultados, como el índice de heterogeneidad estimado $he0$ y los índices REI y DIV definidos en el Apartado 4.9.2 del capítulo anterior.

Operativa

La función *estimei.m* lanza en primer lugar *lphom(x,y)* para obtener la matriz \mathbf{P} estimada por el modelo y el valor $he0$ del índice de heterogeneidad estimado.

Seguidamente se utiliza la función *hallard0.m* para hallar, aproximadamente, el valor $d0$ que corresponde a un HETe igual a $he0$ para el escenario asociado a \mathbf{X} y \mathbf{P} y calcula el incremento $id0$ de “ d ” que conduce a un incremento de una unidad en he .

Con estos datos, la función *ajust.m* obtiene los 11 valores adecuados de d , y para cada uno de ellos realiza nv simulaciones mediante la función *yasimular.m*, obteniendo los valores del IE (ie) y HETe (he) para cada una de ellas (1100 pares de valores, si $nv=100$). A continuación, realiza el ajuste del modelo (4.1) a estos datos.

Finalmente, *estimei.m* utiliza los coeficientes y la desviación típica residual de la ecuación estimada por *ajust.m*, para obtener el valor del EI que corresponde a $he=he0$ y el límite de confianza superior para el mismo.

5.4 EJEMPLO DE APLICACIÓN

En el Apartado 2.4 del Capítulo 2 se utilizó el procedimiento *lphom* para estimar el trasvase de votos de unas opciones a otras entre la 1ª y la 2ª vuelta de las elecciones presidenciales francesas de 2017.

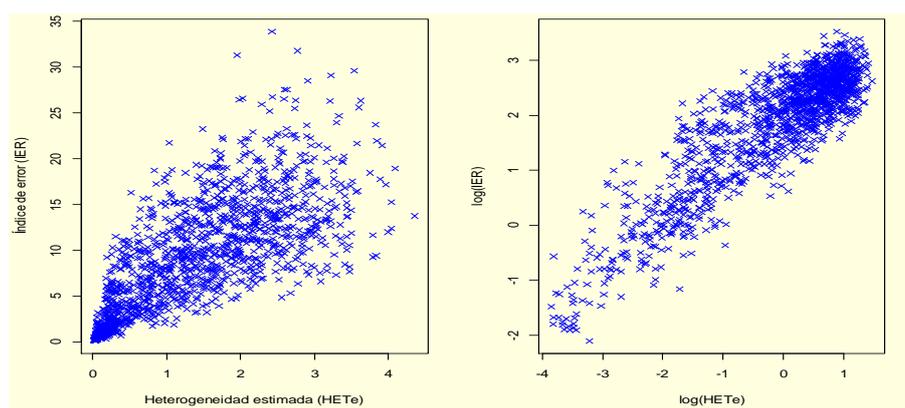
Hemos aplicado el procedimiento que acaba de describirse para estimar el EI asociado a los resultados proporcionados por *lphom*, realizando un total de 1100 simulaciones, 100 para cada uno de 11 valores diferentes de “d”, tal como se expone en el apartado anterior.

Los valores simulados de HETe han oscilado entre un mínimo de 0.02% y un máximo de 4.36%.

En la Figura 5.1 se constata una clara relación entre los valores de HETe y los índices de error EI obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

Por su parte la relación obtenida entre los valores de $\log(\text{EI})$ y los de $\log(\text{HETe})$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es en este caso aproximadamente constante.

Figura 5.1 Elecciones francesas: Relación entre EI y HETe.



Ajustando los datos obtenidos en la simulación al Modelo (4.1) se obtiene el siguiente resultado:

Tabla 5. 1 Ajuste de regresión del Modelo (4.1)

Parámetro	Estimación	Error Estándar	Estadístico-T	Valor-P
Intercept	1.773745	0.011820	150.068	<2e-16
$\log(\text{HETe})$	0.835201	0.006733	124.039	<2e-16
$\log(\text{HETe})^2$	-0.046650	0.005310	-8.7850	<2e-16
coef. correlación múltiple: 0.961				
desv. tip. residual: 0.2771				

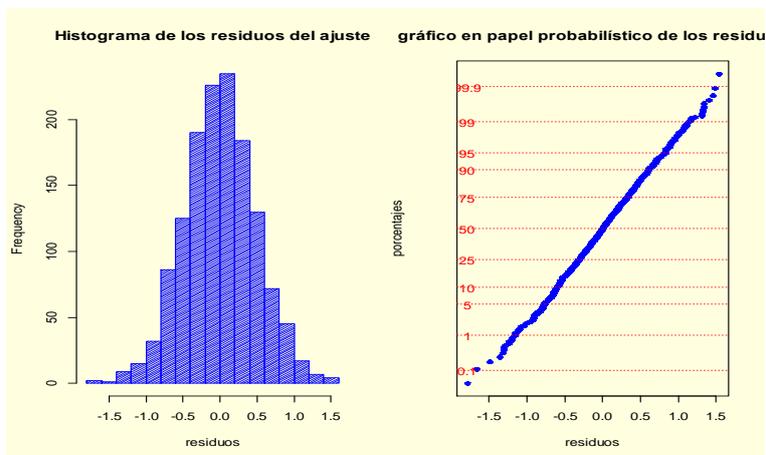
El modelo estimado es, por tanto:

$$\log(\text{EI}) = 1.773 + 0.835\log(\text{HETe}) - 0.04665(\log(\text{HETe}))^2 + u \quad (5.1)$$

siendo la desviación típica residual 0.2771 y resultando muy significativo el término cuadrático del ajuste.

Las siguientes figuras recogen el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

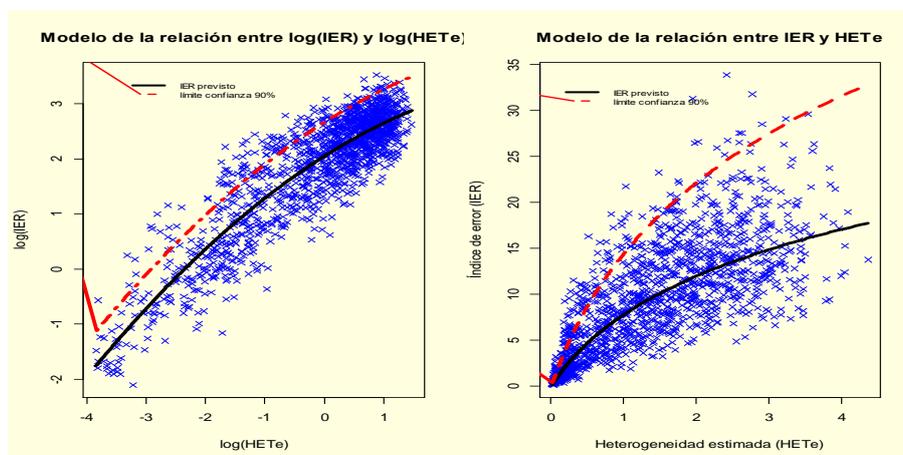
Figura 5.2 Elecciones francesas: Histograma y gráfico en papel probabilístico normal de los residuos del ajuste de log(EI) a log(HETe).



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EI así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 5.3 Elecciones francesas: Modelo ajustado.



De acuerdo con estos resultados, y dado que el valor obtenido por *lphom* para el HETe del modelo fue igual a 2.98, aplicando la ecuación (5.1) se obtiene que el valor estimado para el EI en este estudio es 13.9%, con un límite de confianza superior ($1 - \alpha = 0.90$) igual a 19.8%.

Capítulo 6. Un procedimiento para estimar el EP en estudios reales

6.1 INTRODUCCIÓN

El capítulo anterior se ha centrado en la evaluación del margen de incertidumbre de los resultados proporcionados por *lphom*, cuando esta incertidumbre se cuantifica a partir del índice de error EI definido en el Apartado 3.2.2 del Capítulo 3. Este índice mide el error cometido al estimar la matriz \mathbf{V} de trasvase de votos entre opciones electorales.

En el mismo Capítulo 3 se propuso, como una alternativa para evaluar ese margen de incertidumbre, otro indicador: el índice EP, que mide el error cometido al estimar la matriz \mathbf{P} de probabilidades de transferencia de votos entre las opciones de la Elección 1 y las de la Elección 2.

En el presente capítulo se aborda el tema de desarrollar un procedimiento para estimar el valor de este índice EP en estudios reales.

En el Apartado 6.2 se analiza la relación entre el EP y el índice de heterogeneidad estimado HETe, que resulta muy similar a la encontrada en el Capítulo 4 entre el EI y el HETe. Esta relación se utiliza en el Apartado 6.3 para proponer la metodología para estimar el EP en estudios reales concretos, estimación que se facilita mediante la función *estimep.m* que se presenta al final de ese apartado.

Finalmente, en un último apartado se aplica el procedimiento descrito para estimar el EP en el estudio sobre la segunda vuelta de las elecciones presidenciales francesas de 2017.

6.2 RELACIÓN ENTRE EL EP Y EL HETe

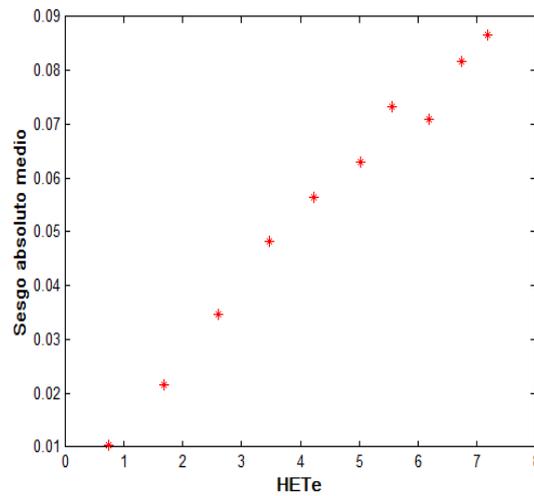
La función de Matlab *yasimilar.m*, que se presentó en el Capítulo 4, proporciona para cada simulación, además de los valores he de HETe y ei del EI, los valores ep del índice EP. A partir de estos valores es posible analizar la relación entre el EP y el HETe, de forma similar a como se hizo en dicho capítulo para estudiar la relación entre EI y HETe.

El coeficiente de correlación entre los valores de EI y EP en las 1380 simulaciones realizadas en el Escenario 1 resultó igual a 0.977. Dada esta fuerte correlación, la relación encontrada en todos los escenarios entre EP y HETe ha sido muy similar a la encontrada en dicho capítulo entre EI y HETe. Los coeficientes de correlación obtenidos entre $\log(\text{EP})$ y $\log(\text{HETe})$ han sido: 0.932 en el Escenario 1, 0.953 en el Escenario 2, 0.948 en el Escenario 3, 0.964 en el Escenario 4 y 0.965 en el Escenario 5.

Para poner de manifiesto la relación entre el EP y el índice de heterogeneidad, hemos llevado a cabo mediante la función *prueba2.m*, que se recoge en el Anejo 1, un conjunto de simulaciones, 500 en cada caso, para diferentes valores de HETe.

Los resultados obtenidos se grafican en la Figura 6.1, que pone claramente de manifiesto la relación positiva entre el sesgo absoluto medio en la estimación de las p_{jk} (EP) y el índice de heterogeneidad.

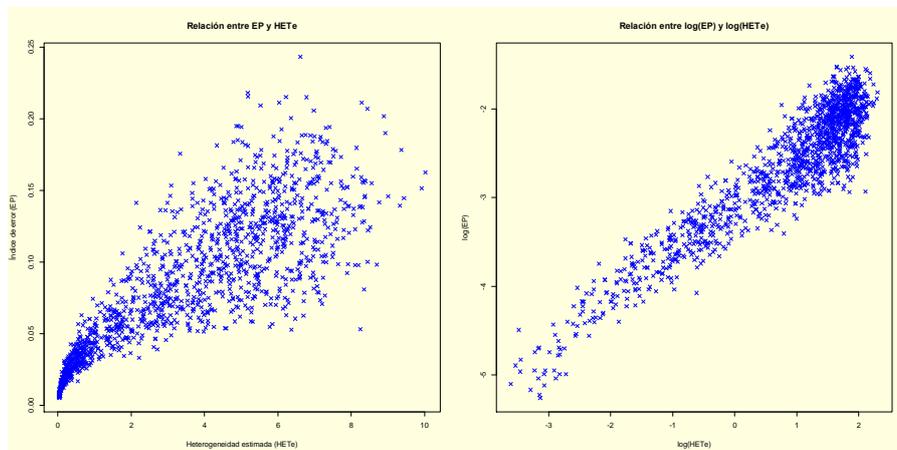
Figura 6.1 Escenario 1: Relación entre sesgo absoluto medio (EP) y HETe (cada punto es la media de 500 simulaciones).



Con el fin de no hacer excesivamente prolija esta exposición, nos limitaremos a comentar con detalle la relación entre EP y HETe encontrada en el Escenario 1.

En la Figura 6.2 se grafican los resultados individuales obtenidos en las 1380 simulaciones realizadas en el Escenario 1 utilizando la función *yasimular*. Se constata que la relación entre EP y HETe es creciente y ligeramente no lineal, con una curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe. Esta heterocedasticidad es mucho menos marcada cuando la gráfica se representa utilizando una escala logarítmica en los dos ejes.

Figura 6.2 Escenario 1: Relación entre EP y HETe.



Ajustando a los datos el Modelo:

$$\log(ep) = \beta_0 + \beta_1 \log(he) + \beta_2 [\log(he)]^2 \quad (6.1)$$

se obtiene el resultado que se recoge en la Tabla 6.1.

Tabla 6. 1 Regresión de EP en función de HETe. Escenario 1

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	-3.07326	0.01232	249.456	<0.0001
log(HETe)	0.53227	0.00557	95.6	<0.0001
log(HETe)^2	-0.00023	0.00397	-0.057	0.955

coef. correlación múltiple: 0.932
desv. tip. residual: 0.263

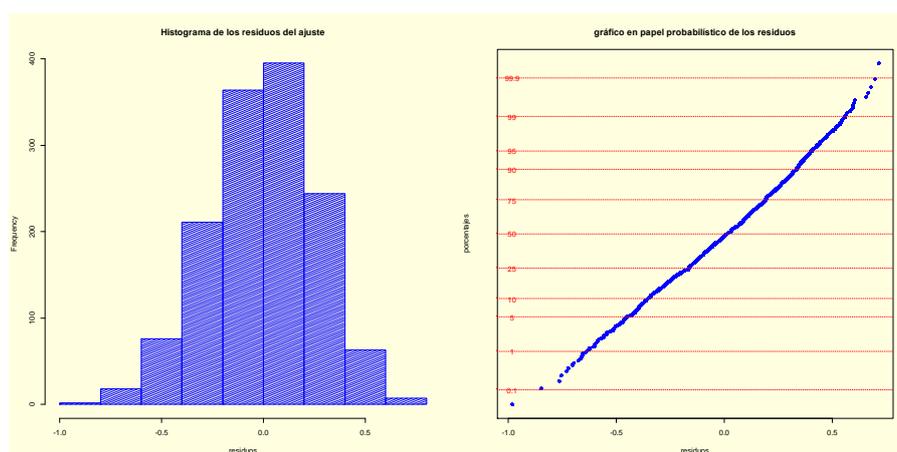
El modelo estimado es, por tanto:

$$\log(\text{Ep}) = -3.073 + 0.532\log(\text{HETe}) - 0.0002(\log(\text{HETe}))^2 + u$$

siendo la desviación típica residual 0.26 y no resultando significativo en este caso el término cuadrático del ajuste, al igual que sucedía en el ajuste del EI.

La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal.

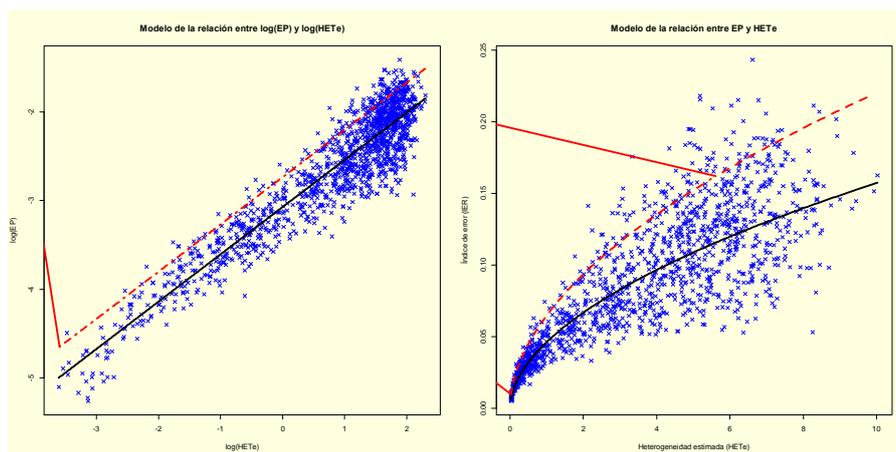
Figura 6.3 Escenario 1: Histograma y gráfico en papel probabilístico normal de los residuos del ajuste de log(EP) a log(HETe). Escenario 1.



Se constata que la hipótesis de normalidad de los residuos es perfectamente admisible, por lo que la distribución normal puede utilizarse para obtener límites de confianza para las predicciones proporcionadas por el modelo para los EP en función de los valores constatados para HETe.

En la siguiente figura se refleja, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EP, así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 6.4 Escenario 1: Modelo ajustado de la relación del EP con HETe.



Se constata, por tanto, para el Escenario 1 que la relación entre EP y HETe es muy similar a la hallada en el Capítulo 4 entre el EI y HETe. La misma conclusión se obtiene en los restantes otros cuatro escenarios considerados, cuyos resultados no exponemos en aras de la concisión.

Al igual que sucedía para el EI, en todos los casos esta relación se modela satisfactoriamente mediante la ecuación de regresión (6.1) utilizando como variable dependiente $\log(\text{EP})$ y como variables explicativas $\log(\text{HETe})$ y su cuadrado, si bien los parámetros concretos de la ecuación que modela esta relación dependen estrechamente del escenario considerado.

6.3 ESTIMACIÓN DEL EP EN ESTUDIOS REALES

La operativa que proponemos al respecto es completamente similar a la expuesta en el Capítulo 5 para estimar el EI en estudios concretos.

Para obtener el EP en un estudio a partir del índice de heterogeneidad estimado HETe obtenido en el mismo, será necesario estimar en primer lugar la función que relaciona el EP con el HETe, mediante un estudio de simulación similar al presentado en el apartado anterior para el Escenario 1, pero utilizando el escenario concreto que definan los datos del estudio que se está analizando.

Este escenario vendrá caracterizado por la matriz \mathbf{X} de votos obtenidos en la Elección 1 por los J partidos en las I unidades territoriales, y por la matriz \mathbf{P} que se haya obtenido en el estudio para las probabilidades de transición de votos de unas opciones a otras entre las dos elecciones. Los valores de “d” a utilizar en las simulaciones se elegirán de forma que conduzcan a índices de heterogeneidad estimados cercanos al obtenido en el estudio.

Con el fin de aplicar fácilmente este procedimiento en estudios concretos, se ha desarrollado la función Matlab *estimep.m*, recogida en el Anejo 1 de esta tesis, y que se describe a continuación.

La función se lanza mediante la instrucción:

$$[\text{ep}, \text{limep}] = \text{estimep}(\mathbf{x}, \mathbf{y}, \text{nv}, \text{conf})$$

Los argumentos de la función son los mismos que los de *estimei*:

- \mathbf{x} : matriz \mathbf{X} ni x_{nj} con los resultados de la Elección 1.

- y : matriz Y ni x nk con los resultados de la Elección 2.
- $conf$: nivel de confianza deseado para el intervalo de confianza superior.
- nv : número de simulaciones a realizar, para cada uno de los 11 valores considerados de d . El valor recomendado de nv es 100.

Los resultados que proporciona *estimep.m* son:

- ep : valor estimado del índice de error EP del estudio.
- $limep$: límite de confianza superior (para un nivel de confianza $conf$) para el EP.

Operativa

La operativa es muy similar a la expuesta para *estimei.m*. La función *estimep.m* lanza en primer lugar *lphom(x,y)* para obtener la matriz P y el valor $he0$ del índice de heterogeneidad estimado.

Seguidamente se utiliza la función *hallard0.m* para hallar aproximadamente el valor $d0$ que corresponde a un HETe igual a $he0$ para el escenario asociado a X y P y calcula el incremento $id0$ de “ d ” que conduce a un incremento de 1 en he .

Con estos datos, la función *ajustep.m* obtiene los 11 valores adecuados de d , y para cada uno de ellos realiza nv simulaciones mediante la función *yasimular.m* obteniendo los valores del EP y HETe para cada una de ellas (1100 pares de valores si $nv=100$). A continuación, realiza el ajuste del modelo (6.1) a estos datos.

Finalmente, *estimep.m* utiliza los coeficientes y la desviación típica residual de la ecuación estimada por *ajustep.m*, para obtener el valor del EP que corresponde a $he = he0$ y el límite de confianza superior para el mismo.

6.4 EJEMPLO DE APLICACIÓN

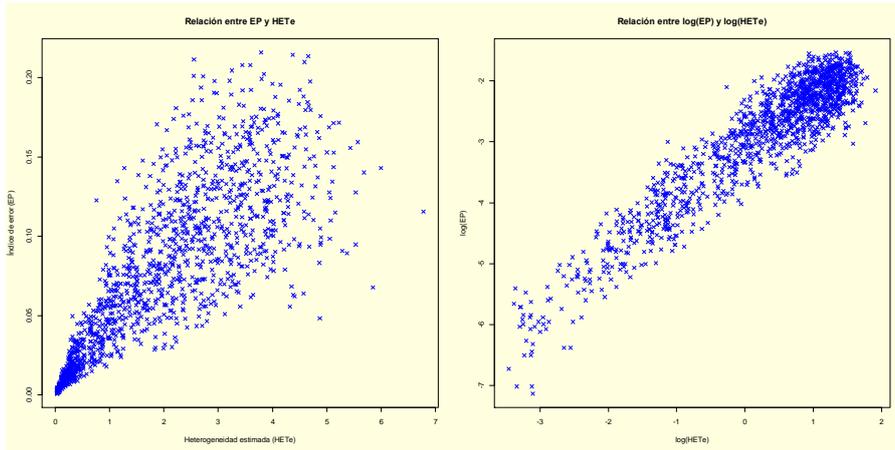
Aplicamos en este apartado el procedimiento descrito en los apartados anteriores para estimar el valor del EP en el estudio de las elecciones presidenciales francesas de 2017 que se presentó en el Capítulo 2 y cuyo EI fue estimado al final del Capítulo 5. En dicho estudio, el valor obtenido para el HETe fue 2.98.

Hemos aplicado el procedimiento que acaba de describirse para estimar el EP y el HETe asociado a los resultados proporcionados por *lphom*, realizando un total de 1100 simulaciones, 100 para cada uno de 11 valores diferentes de “ d ”, tal como se expone en el apartado anterior.

En la Figura 6.5 se constata una clara relación entre los valores de HETe y los índices de error EP obtenidos en cada simulación. Esta relación es creciente y no lineal, con curvatura negativa y una varianza residual que aumenta marcadamente al hacerlo HETe.

Por su parte la relación obtenida entre los valores de $\log(EP)$ y los de $\log(HETe)$ presenta también una ligera curvatura negativa, apreciándose que la varianza residual es en este caso aproximadamente constante.

Figura 6.5 Elecciones francesas: Relación entre EP y HETe.



Ajustando los datos obtenidos en la simulación al Modelo (6.1) se obtiene el siguiente resultado:

Tabla 6. 2. Regresión de EP en función de HETe. Elecciones francesas

	Estimación	Error Estándar	Estadístico T	Valor-P
Constante	-3.01986	0.01413	-213.787	<0.0001
log(HETe)	0.79236	0.00972	81.558	<0.0001
log(HETe)^2	-0.06131	0.00644	-9.518	<0.0001

coef. correlación múltiple: 0.939
desv. tip. residual: 0.357

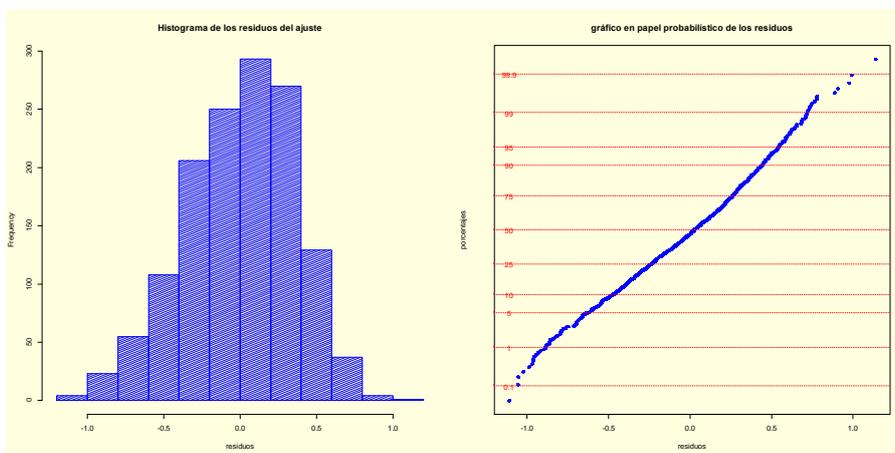
El modelo estimado es, por tanto:

$$\log(\text{IER}) = -3.0199 + 0.7924\log(\text{HETe}) - 0.0613(\log(\text{HETe}))^2 + u \quad (6.2)$$

siendo la desviación típica residual 0.3573 y resultando muy significativo el término cuadrático del ajuste.

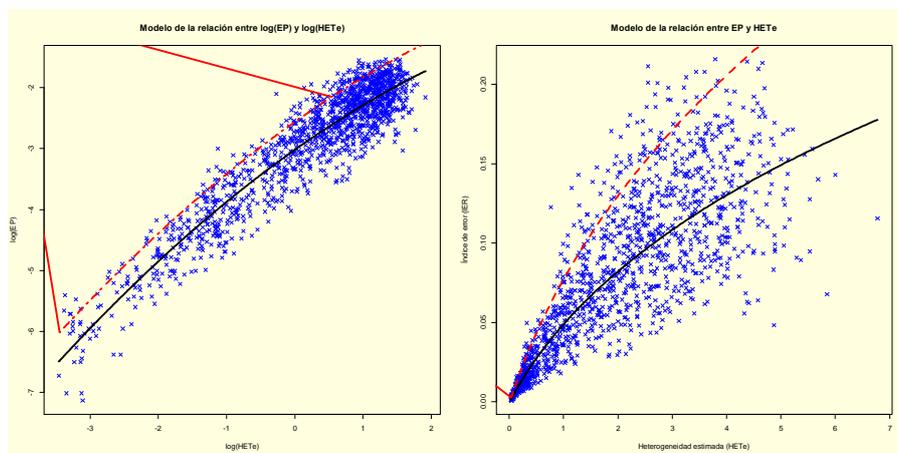
La siguiente figura recoge el histograma de los residuos del ajuste y su representación en papel probabilístico normal, constatándose que la hipótesis de normalidad es admisible.

Figura 6.6 Elecciones francesas: Histograma y gráfico en papel probabilístico normal de los residuos del ajuste de log(EP) a log(HETe).



Finalmente, en la Figura 6.7 se reflejan, tanto sobre los datos iniciales como sobre su transformación logarítmica, los valores medios previstos por el modelo para el EP, así como su límite superior de confianza para un nivel de confianza del 90%.

Figura 6.7 Elecciones francesas: Modelo ajustado de la relación del EP con HETe.



De acuerdo con estos resultados, y dado que el valor obtenido por I_{phom} para el HETe del modelo fue igual a 2.98, aplicando la ecuación (6.2) se obtiene que el valor estimado para el EP en este estudio es 0.108, con un límite de confianza superior ($\alpha=0.90$) igual a 0.171.

Capítulo 7. Intervalos de confianza para las p_{jk}

7.1 INTRODUCCIÓN

En los capítulos anteriores de esta tesis se han definido dos indicadores, EI y EP, para cuantificar el margen de incertidumbre global asociado a los resultados obtenidos tras aplicar el procedimiento *lphom* para estimar el trasvase de votos entre dos procesos electorales y, asimismo, hemos propuesto una metodología para estimar estos índices en estudios reales.

En la mayor parte de los estudios reales, sin embargo, el interés se suele centrar en los valores de determinados p_{jk} . Por ejemplo, en los análisis sobre los procesos electorales recientes de nuestro país, existía un interés especial en disponer de una estimación del porcentaje de anteriores votantes del PP que habían decidido votar por Ciudadanos. Otro ejemplo lo encontramos al estudiar la movilidad electoral en Cataluña entre las elecciones autonómicas de 2012 y 2015 (Romero, 2015). En 2015, ERC decidió primar el eje nacionalista-independentista y coaligarse con CiU para formar la opción de Junts per el Sí, siendo una cuestión de interés analizar qué porcentaje de los antiguos votantes de ERC decidieron pasarse a la CUP como repulsa a esa coalición con un partido de situado a la derecha en el eje ideológico.

Los valores de EI y EP, aunque son valiosos indicadores de la calidad global de las estimaciones de las \hat{p}_{jk} proporcionadas por *lphom*, no dicen prácticamente nada sobre el margen de incertidumbre respecto a un \hat{p}_{jk} concreto⁸, margen que podría variar sensiblemente de unos \hat{p}_{jk} a otros en función, por ejemplo, del tamaño de los partidos implicados.

En el presente capítulo abordamos el problema de estimar estos márgenes de incertidumbre, estableciendo intervalos de confianza para los valores de los diferentes p_{jk} .

Comentamos en primer lugar los resultados que hemos obtenido al estudiar en ciertos escenarios la distribución de los estimadores \hat{p}_{jk} proporcionados por *lphom*, y especialmente su posible sesgo y su varianza.

Seguidamente se establece un procedimiento para modelar, en un escenario concreto, la relación entre el sesgo y la varianza de un determinado \hat{p}_{jk} y el HETe obtenido al aplicar *lphom*.

Finalmente, se utilizan los resultados anteriores para desarrollar un procedimiento que permite obtener un intervalo de confianza para cualquier p_{jk} en estudios reales, procedimiento cuya utilización se ilustra mediante dos ejemplos reales.

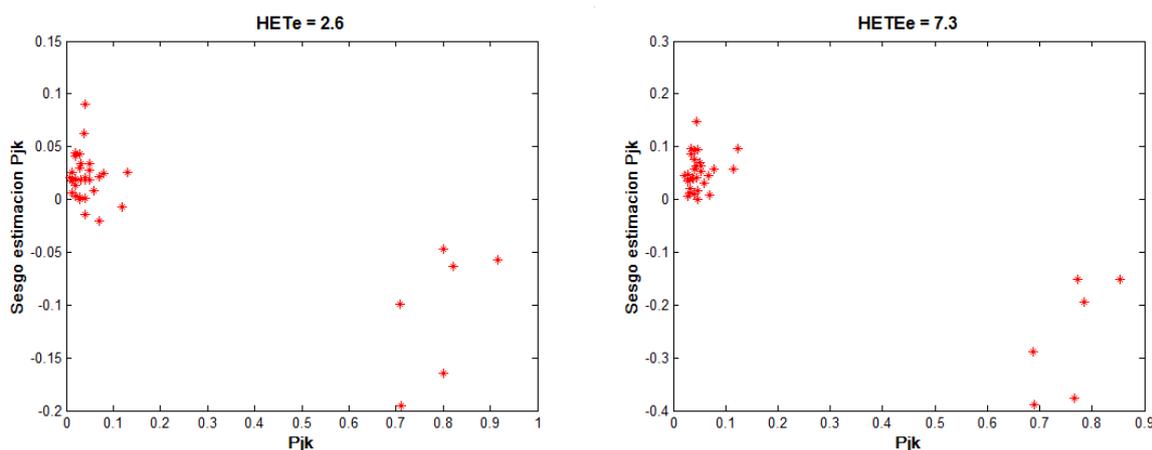
⁸ Salvo en casos muy extremos. Por ejemplo, es obvio que si el EI fuera 0 las estimaciones de las p_{jk} serían completamente exactas.

7.2 SESGO DE LOS ESTIMADORES \hat{p}_{jk}

A partir de los diferentes escenarios descritos en el Capítulo 3, hemos llevado a cabo un conjunto de estudios de simulación a fin de investigar el posible sesgo de los estimadores \hat{p}_{jk} proporcionados por *lphom*, así como, los factores que afectan a dicho sesgo. En líneas generales, los resultados obtenidos en todos los escenarios han sido similares, por lo que nos limitaremos a comentar sucintamente los hallados para el Escenario 1. Estos resultados son relevantes porque justifican la operativa que proponemos posteriormente para obtener los intervalos de confianza para los p_{jk} .

Mediante la función Matlab *prueba2.m*, que se recoge en el Anejo 1, y operando con las matrices **XAND** y **PAND** del Escenario 1 descrito en el Capítulo 4 y en el Anejo 3, hemos obtenido la relación entre el sesgo de los \hat{p}_{jk} y su verdadero valor p_{jk} en diferentes hipótesis sobre el nivel de incumplimiento de la hipótesis de homogeneidad. En concreto, la relación se ha obtenido a partir de emplear, en cada caso, 500 simulaciones y el nivel de incumplimiento de la hipótesis de homogeneidad ha sido medido por el correspondiente HETe. En la Figura 7.1 se reflejan las relaciones obtenidas para dos valores diferentes de HETe.

Figura 7.1 Escenario 1: Relación del Sesgo de \hat{p}_{jk} con el valor real de p_{jk} .



Se aprecia claramente en la figura la existencia de un ligero sesgo en la estimación de las p_{jk} , sesgo que tiende a ser negativo para valores elevados (cerca de 1) de los p_{jk} y positivo para valores bajos (cerca de 0) de estas probabilidades de transición.

La existencia de estos patrones en los sesgos puede justificarse de manera bastante intuitiva. En efecto, si un p_{jk} es muy alto, por ejemplo 0.9, el sesgo de su estimación (estimación que siempre estará entre 0 y 1) puede variar entre -0.9 y +0.1. La distribución de este sesgo, especialmente si HETe es alto, tendrá una clara asimetría negativa y su valor medio resultará negativo. Mediante un razonamiento análogo se justifica que el sesgo en la estimación de p_{jk} bajos será en promedio positivo, dado que la estimación no puede nunca resultar menor que cero.

En la Figura 7.1 también se aprecia, cuando se tienen en cuenta las escalas del eje vertical, que, como cabía esperar, los sesgos tienden a ser menores en promedio cuanto menor es el índice de heterogeneidad estimado HETe. Obviamente este promedio hay que calcularlo sobre los valores absolutos de los sesgos, tal como hicimos al definir el índice EP, puesto que al ser siempre iguales a uno la suma de las filas de la matriz de transferencia, tanto de $[p_{jk}]$ como de $[\hat{p}_{jk}]$, la suma de todos los sesgos, y por tanto su media, es siempre exactamente igual a cero.

En estudios electorales reales, como el simulado en el Escenario 1, lo más frecuente es encontrar valores bastante extremos de los p_{jk} . Esto sucede porque, en general, para cada opción j_0 en la Elección 1, hay una opción k_0 que es claramente la preferida en la Elección 2 (por ejemplo, la mayor parte de los votantes a un partido en E1 vuelven a votar al mismo partido en E2), lo que conduce a un valor elevado de p_{j_0,k_0} y, en consecuencia, a valores bajos para los restantes $p_{j_0,k}$.

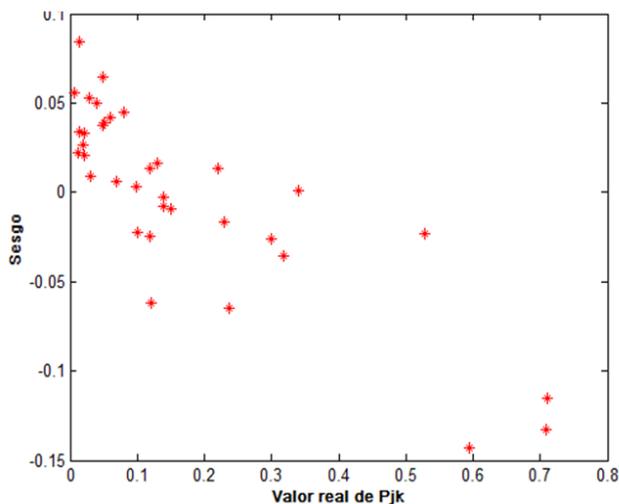
Para visualizar de forma más clara la relación entre los sesgos de los estimadores \hat{p}_{jk} y los valores reales de p_{jk} hemos realizado una nueva simulación con la misma **XAND**, pero sustituyendo la matriz **PAND** por otra matriz **PMOD** en la que aparecieran valores de las p_{jk} más homogéneamente repartidos entre 0 y 1.

Se recogen a continuación, Tabla 7.1, los valores medios de los p_{jk} en la matriz **PMOD** y el gráfico obtenido para la relación entre el sesgo de los \hat{p}_{jk} y su verdadero valor p_{jk} .

Tabla 7. 1 Valores medios de los p_{jk} en PMOD.

		Opciones electorales de destino					
		1	2	3	4	5	6
Opciones Electtorales	Origen	0.5049	0.2834	0.0769	0.0473	0.0467	0.0409
		0.1329	0.2737	0.1411	0.1025	0.2135	0.1364
		0.1318	0.0890	0.3409	0.0583	0.1465	0.2325
		0.0952	0.0824	0.0349	0.5757	0.1249	0.0869
		0.0975	0.1021	0.0633	0.1133	0.4513	0.1725
		0.0762	0.0539	0.0387	0.0898	0.1467	0.5948

Figura 7.2 Relación del sesgo con p_{jk} simulando con PMOD.



Se constata claramente en la figura la relación negativa entre los sesgos y los valores reales de los p_{jk} , resultandos sesgos positivos para los p_{jk} cercanos a 0, negativos para los valores altos de los p_{jk} , siendo más próximos a cero para los valores intermedios de las probabilidades de transición.

7.3 VARIANZA DE LOS ESTIMADORES \hat{p}_{jk}

7.3.1 Introducción

Para obtener intervalos de confianza para los p_{jk} necesitamos estimar, no sólo el sesgo previsible del estimador \hat{p}_{jk} , sino también su varianza. Previsiblemente esta varianza será tanto mayor, cuanto mayor sea el grado de incumplimiento de la hipótesis de homogeneidad, grado estimado mediante el HETe.

Para analizar la naturaleza de la relación del sesgo y la varianza de un \hat{p}_{jk} con el HETe en escenarios concretos, hemos llevado a cabo dos estudios de simulación considerando dos valores de partida distintos para p_{jk} .

1. Estimación de p_{11} en el estudio de las elecciones presidenciales francesas de 2017 que ya analizado en los Capítulos 2, 5 y 6 de esta tesis. En este ejemplo, p_{11} es la proporción de electores que habiéndose abstenido en la primera vuelta volvieron a abstenerse en la 2ª. En la estimación mediante *lphom* se obtuvo $\hat{p}_{11} = 0.80$, siendo éste un ejemplo de un caso con valor elevado de p_{jk} .
2. Estimación de p_{13} en las mismas elecciones, que es la proporción de electores que habiéndose abstenido en la primera vuelta decidieron votar a Macron en la 2ª. En la estimación se obtuvo $\hat{p}_{13} = 0.12$ y es un ejemplo de un caso con valor cercano a 0.

La función *pruebapjk.m*, que se recoge en el Anejo 1, la hemos utilizado para analizar, en estos dos casos, mediante simulación, la relación del sesgo y de la varianza de \hat{p}_{jk} con HETe.

7.3.2 Estimación de p_{11} en las elecciones presidenciales francesas

Al estudiar mediante la función *pruebapjk.m* el comportamiento del sesgo y la varianza del \hat{p}_{11} de las elecciones francesas se obtienen los gráficos que se reflejan en las Figuras 7.3 y 7.4.

Figura 7.3 Relación entre sesgo estimación p_{11} y HETe (elecciones francesas $p_{11} = 0.80$).

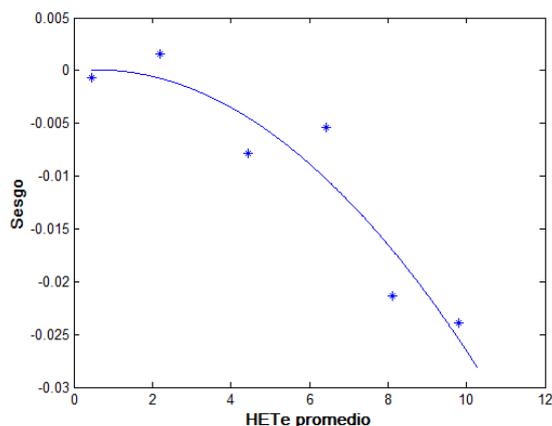
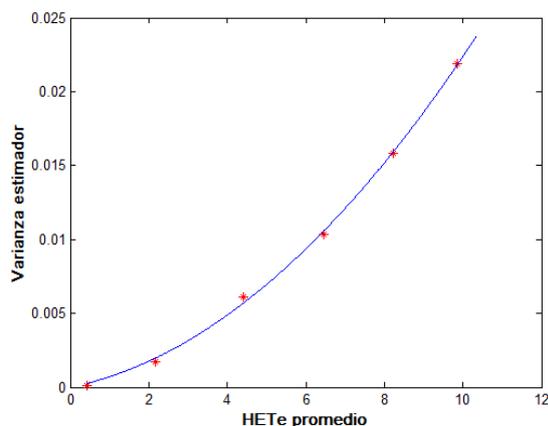


Figura 7.4 Relación entre varianza estimación p_{11} y HETe (elecciones francesas $p_{11} = 0.80$).



Se constata, por una parte, que el sesgo ($\hat{p}_{11} - p_{11}$) es ligeramente negativo y que este se incrementa, de forma progresivamente más marcada, al aumentar HETe, ajustándose bien la relación a una parábola que pasa por el origen (ver Figura 7.3). Ha de tenerse en cuenta que para HETe=0 el sesgo es necesariamente

0. Por ejemplo, para el valor del 3% del HETe obtenido en este estudio, el sesgo es aproximadamente -0.002.

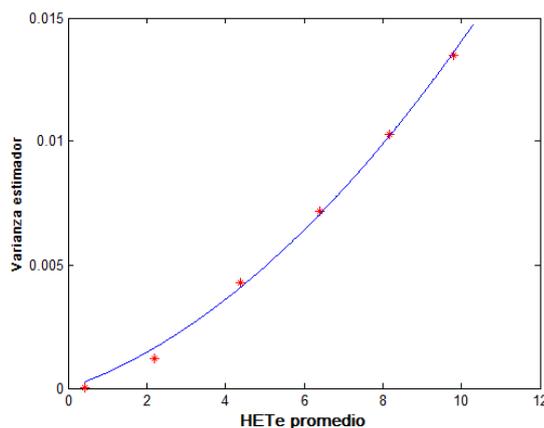
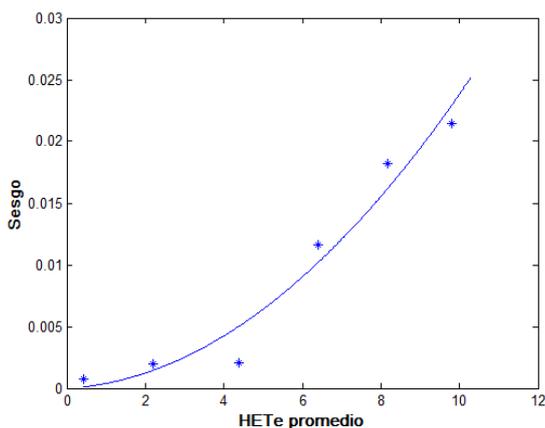
Por otro lado, se constata también que, como era de esperar, la varianza del error aumenta al hacerlo HETe, ajustándose perfectamente la relación a una parábola que pasa por el origen (ver Figura 7.4).

7.3.3 Estimación de p_{13} en las elecciones presidenciales francesas

Al estudiar mediante la función *pruebapjk.m* el comportamiento del sesgo y la varianza del \hat{p}_{13} de las elecciones francesas, se obtienen los gráficos que se reflejan en las Figuras 7.5 y 7.6.

Figura 7.4 Relación entre sesgo estimación p_{13} y HETe (elecciones francesas $p_{13} = 0.12$).

Figura 7.6 Relación entre varianza estimación p_{13} y HETe (elecciones francesas $p_{11} = 0.12$).



Se constata que los resultados son muy similares a los obtenidos en el apartado anterior, salvo que, en este caso, y tal como habíamos previsto, los sesgos resultan positivos (ver Figura 7.5). Para un HETe del 3% el sesgo medio previsto es aproximadamente +0.006.

La relación parabólica entre la varianza de \hat{p}_{13} es también muy similar a la obtenida en el apartado anterior.

7.4 UN PROCEDIMIENTO PARA OBTENER INTERVALOS DE CONFIANZA PARA LOS p_{jk}

7.4.1 Metodología

Los resultados de los apartados anteriores han puesto de manifiesto la posibilidad de obtener, mediante simulación, en un escenario concreto y para un determinado p_{jk} , las funciones que relacionan el sesgo y la varianza de \hat{p}_{jk} con el índice de heterogeneidad estimado HETe. En efecto, en los dos casos se ha constatado que estas funciones se aproximan bien mediante parábolas que pasan por el origen.

En consecuencia, se propone el siguiente procedimiento para obtener un intervalo de confianza aproximado para el parámetro p_{jk} :

Los datos serán:

- Las matrices \mathbf{X} e \mathbf{Y} con los resultados en las dos elecciones.

- Los valores j y k del p_{jk} de interés.
- El nivel de confianza “conf” para el intervalo buscado.
- El valor nv del número de simulaciones a realizar para cada uno de los 11 valores de “d” ensayados, con el fin de obtener las ecuaciones de las parábolas que relacionan el sesgo y su varianza con HETe.

A partir de estos datos, los pasos a seguir para obtener el intervalo de confianza son los siguientes:

- Paso 1: aplicar *lphom* para obtener la matriz estimada de probabilidades de transición estimada mediante el modelo **PM** y el valor $he0$ de HETe. Hacer $p_{jk0} = pm(j,k)$.
- Paso 2: obtener 11 valores de “d” para los que hay que hacer las simulaciones, eligiéndolos de forma que los HETe correspondientes estén aproximadamente centrados en torno de $he0$ (utilizar para ello la función *hallard0.m*⁹ desarrollada en Matlab).
- Paso 3: Aplicar *yasimupjk* para obtener las $11 \times nv$ parejas de valores (he, er) con las que se van a estimar las relaciones de he y var(he) con he.
- Paso 4: Estimar la parábola $er = f(he)$ que relaciona, para el escenario y p_{jk} considerados, el sesgo de \hat{p}_{jk} con HETe.
- Paso 5: Obtener la ecuación de la parábola $var(er) = g(he)$ que relaciona, para el escenario y p_{jk} considerados, la varianza de \hat{p}_{jk} con HETe, ajustando un modelo de regresión utilizando como variable dependiente los cuadrados de los residuos del ajuste obtenido en el Paso 4.
- Paso 6: Obtener, a partir de las ecuaciones estimadas, los valores del sesgo medio ($e0m$) y desviación típica del sesgo ($se0$) correspondientes a $he0$.
- Paso 7: Obtener los dos límites de confianza aproximados LS y LI mediante la expresión $e0m \pm z_{\alpha} se0$, siendo z_{α} el percentil adecuado de la distribución normal tipificado. Corregir, en su caso, haciendo $LS = \min(LS, 1)$ y $LI = \max(LI, 0)$.

La función *intconf.m*, que se recoge en el Anejo 1, realiza todos estos pasos y lista los límites de confianza obtenidos.

7.4.2 Ejemplos

Se recogen a continuación los resultados obtenidos al aplicar este procedimiento para estimar los límites de confianza de los dos p_{jk} analizados en el Apartado 7.3

7.4.2.1 Intervalo de confianza para p_{11} de las elecciones francesas (I=13 regiones).

El cálculo se lleva a cabo mediante la instrucción *intconf(xf, yf, I, I, 0.90, 400)* Esta instrucción genera los siguientes resultados:

Estimación de p_{11} e intervalo de confianza (conf: 90%).

⁹ Esta función, como todas las funciones Matlab utilizadas en esta investigación, se recoge en el Anejo 1.

Estimación: 0.800 int.conf [0.711 ; 0.896].

Se constata que el intervalo no es completamente simétrico en torno a p_{110} , como se comprueba porque $\left(\frac{0.711+0.896}{2} = 0.803 > 0.800\right)$, debido al ligero sesgo de \hat{p}_{11} . Ello deja abierta la posibilidad de corregir, si se desea el ligero sesgo de las estimaciones de las p_{jk} . Observaciones similares pueden hacerse para los resultados del siguiente ejemplo.

7.4.2.2 Intervalo de confianza para p_{13} de las elecciones francesas (I = 13 regiones).

El cálculo se lleva a cabo mediante la instrucción `intconf(xf, yf, I, 3, 0.90, 400)`. Esta instrucción genera los siguientes resultados:

Estimación de p_{13} e intervalo de confianza (conf: 90%).

Estimación: 0.121 int.conf [0.034 ; 0.196].

Capítulo 8. Conclusiones y futuras investigaciones

Se recogen en este último capítulo las principales conclusiones y aportaciones de la tesis y se comentan nuevas líneas de investigación que han surgido a lo largo de su realización.

Con carácter previo, en nuestra investigación hemos constatado que *lphom* constituye un procedimiento extremadamente sencillo e intuitivo y que conduce a resultados muy razonables siempre que se aplica a estudios reales. La comparación con el sofisticado método *ei.MD.bayes* que hemos realizado en el Capítulo 2, aplicando ambas metodologías al análisis de los resultados de las elecciones presidenciales francesas de 2017, ha puesto de manifiesto, al menos en este caso, la indudable superioridad de *lphom* frente a una alternativa con exigencias de computación 250 veces superiores.

Dado que el objetivo general de la investigación es desarrollar procedimientos para estimar el margen de incertidumbre de los resultados proporcionados por *lphom* en estudios reales, la primera cuestión que hemos tenido que abordar ha sido la de cómo medir dicho margen de incertidumbre. Con ese fin en el Capítulo 3 hemos definido dos indicadores: el EI centrado en la distancia entre la matriz real de trasvase de votos y la estimada por el modelo, y el EP que no es más que el valor medio de las diferencias en valor absoluto entre los p_{jk} reales y los estimados. Hemos constatado que, como cabía esperar, ambos índices están estrechamente correlacionados.

Dado que el fundamento del modelo *lphom* es la hipótesis de homogeneidad del comportamiento en las I unidades territoriales (uts), en el Capítulo 3 hemos propuesto también dos índices para cuantificar el grado de incumplimiento de esta hipótesis en estudios concretos. El índice HET es básicamente una suma de las distancias entre las matrices de transferencia de votos reales en la diferentes uts y las que se habrían obtenido si se hubiera cumplido perfectamente la hipótesis de homogeneidad. Este índice presenta sin embargo el inconveniente de que sólo puede obtenerse en procesos electorales simulados, puesto que exige saber los trasvases de votos que son desconocidos en los procesos reales. Para superar este inconveniente hemos propuesto un segundo índice, el HETe o índice de heterogeneidad estimado, que puede obtenerse en estudios reales a partir de los valores de los residuos e_{ik} obtenidos al estimar el modelo *lphom*. En nuestros estudios de simulación hemos constatado una estrecha relación entre estos dos índices de heterogeneidad, resultando los coeficientes de correlación entre los logaritmos de ambos indicadores superiores a 0.99, de lo que se concluye que el HETe, que tiene la ventaja fundamental de poderse calcular en estudios reales, es un indicador adecuado para evaluar el grado de incumplimiento de la hipótesis de homogeneidad en dichos estudios.

En el capítulo 3 se explica también la metodología que hemos desarrollado para simular procesos reales. Aplicando esta metodología en cinco escenarios diferentes, hemos constatado en el Capítulo 4 la existencia de una clara relación entre el indicador EI y el estimador HETe, relación que se modela muy bien mediante el modelo de regresión (4.1), utilizando como variable dependiente $\log(EI)$ y como variables explicativas $\log(HETe)$ y su cuadrado. Los coeficientes de estos modelos dependen de las características del escenario analizado, habiendo propuesto y estudiado en el Apartado 4.9 dos posibles indicadores, a los que hemos denominado REI y DIV, que parecen relevantes al respecto y cuyo estudio, junto con el de otras posibles alternativas, podría ser un tema para futuras investigaciones.

Las tres aportaciones fundamentales de esta tesis se recogen en los capítulos 5, 6 y 7, en los que se desarrollan las metodologías para estimar los márgenes de incertidumbre de los resultados proporcionados por *lphom* en estudios reales. En el capítulo 5 este margen de incertidumbre se mide mediante la estimación del indicador EI, mientras que en el Capítulo 6 se centra en estimar en estudios reales el indicador EP. Por su parte, el Capítulo 7 está dedicado a la obtención de intervalos de confianza para los valores individuales de los p_{jk} . En los tres casos las metodologías propuestas se vuelcan en

respectivas funciones Matlab, lo que permite que puedan ser utilizadas por cualquier analista en estudios reales. El uso de las funciones es ilustrado mediante su aplicación sobre ejemplos basados en los resultados de las elecciones presidenciales francesas de 2017.

A lo largo del trabajo realizado en esta tesis se han abierto campos y planteado cuestiones que merecen ser objetos de futuras investigaciones. Algunas de ellas están ya en una fase de estudio relativamente avanzada en trabajos que están llevando a cabo los dos directores de esta tesis, los profesores Romero y Pavía. Mencionaré sucintamente tres de los problemas que se están abordando.

Un primer tema es el de la comparación del sencillo procedimiento *lphom* con las técnicas mucho más complejas de regresión ecológica y, en particular, con *ei.MD.bayes* que parece ser la más prestigiosa entre ellas. Las comparaciones hechas en estudios reales con un número reducido de uts no dejan lugar a dudas al respecto, puesto que, aunque los valores reales de los trasvases de votos son desconocidos en estos estudios, los resultados obtenidos por *lphom* son siempre razonables desde el punto de vista sociopolítico, mientras que *ei.MD.bayes* proporciona resultados absolutamente inaceptables. Parece ser, por tanto, que los procedimientos de regresión ecológica exigen un número elevado de uts para funcionar correctamente. Hemos comprobado, en efecto, en algunos ejemplos reales con 50 o más uts, que *ei.MD.bayes* obtiene resultados razonables, al igual que lo hace *lphom*. La comparación entre ambos métodos en estos casos sólo puede llevarse a cabo en procesos especiales en los que los trasvases de votos son conocidos. Las investigaciones en marcha hasta el momento parecen confirmar que también en estos casos con un número elevado de uts el procedimiento *lphom*, más sencillo, se comporta mejor que el sofisticado *ei.MD.bayes*, cuando este último se utiliza con sus opciones por defecto.

Otro problema en el que se está trabajando es el que se plantea al analizar los trasvases entre las opciones de dos elecciones que se celebran simultáneamente, por ejemplo, unas elecciones autonómicas y unas generales en una misma Comunidad. Hay que tener en cuenta que, en cualquier par de elecciones a las que denominaremos E1 y E2, se pueden plantear dos problemas simétricos: la estimación de las p_{jk} , que son la proporción de electores que habiendo optado por la opción j en E1 optaron por la opción k en E2, o la estimación de las q_{kj} que son la proporción de los votos obtenidos en E2 por la opción k que procede de electores que optaron por la opción j en E1. Estimados los p_{jk} , por ejemplo, mediante *lphom*, se deducen inmediatamente de ellos los q_{kj} (llamaremos EJK a este enfoque), pero alternativamente también puede plantearse estimar los q_{kj} mediante *lphom*, considerando como origen de votos E2 y como destino E1, y deducir a partir de ellos los p_{jk} (llamaremos EKJ a este enfoque). El problema es que los resultados obtenidos mediante el enfoque EJK son diferentes de los hallados mediante EKJ, lo que plantea la cuestión de cuál de los dos escoger en un problema concreto, más aún cuando las dos posibles hipótesis de homogeneidad que se plantearían en cada caso son incompatibles entre sí. Cuando E2 es posterior a E1, la hipótesis de homogeneidad que se plantea con el enfoque EJK es mucho más razonable que la que se plantearía con el enfoque EKJ, por lo que en estos casos está claro que debe utilizarse el enfoque EJK. Sin embargo, cuando las dos elecciones son simultáneas ambas hipótesis, pese a ser incompatibles, pueden ser razonables y habría que ver qué criterios hay que utilizar para elegir entre una u otras, o intentar desarrollar una metodología alternativa a *lphom* para estos casos de elecciones simultáneas. Es con este último enfoque con el que están llevando a cabo las investigaciones.

Finalmente, el enfoque *lphom* podría generalizarse a un modelo tri-electoral cuyo objetivo sería estimar los elementos de la hipermatriz $[P_{jkl}]$ cuyo elemento genérico sería la proporción de electores que habiendo elegido la opción j en una primera elección E1 y pasado a la opción k en una segunda elección E2, se habían decidido por la opción l en una tercera elección posterior E3. Por ejemplo, en España puede haber especial interés en estimar la proporción de electores que habiendo pasado del PP a Ciudadanos o a la Abstención entre E1 y E2, han vuelto a votar al PP en una tercera elección E3. Este problema parece muy difícil de abordar con los enfoques de la regresión ecológica, pero creemos que es tratable con un enfoque que sería una generalización de *lphom*.

BIBLIOGRAFÍA

- Aarts, K., van der Kolk, H. y Rosema, M. (2007). Een verdeeld Electoraat De Tweede Kamerverkiezingen van 2006, *Spectrum*.
- Ambühl, M. (2003) Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten. *Bundesamt für Statistik* (BFS), Neuchâtel
- Andreadis, I., Chadjipadelis, T. (2009) A method for the estimation of voter transition rates. *J. Elections Public Opin. Parties* 19(2), 203–218
- Barbosa, M. F. & H. Goldstein (2000). Discrete Response Multilevel Models for Repeated Measures: An Application to Voting Intentions Data, *Quality & Quantity* 34: 323-33.
- Benewick, R.J., Birch, A.H., Blumer, H. y Ewbank, A. (1969) “The floating voter and the liberal view of representation.” *Political Studies*, vol 17, no 2, pp 177-95.
- Berkelaar, M. y others. (2014) lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs. R package version 5.6.10. <https://CRAN.R-project.org/package=lpSolve>
- Betlehem, J.G. y Kersten, H.M.P. (1986). Werken met Nonresponse (Working with Nonresponse). *Voorburg: Centraal Bureau voor de Statistiek*.
- Bosch, A., y Durán, I. M. (2017) “How does economic crisis impel emerging parties on the road to elections? The case of the Spanish Podemos and Ciudadanos.” *Party Politics*, online available.
- Brown, P. J., y Payne, C. D. (1986) “Aggregate data, ecological regression and voting transitions.” *Journal of the American Statistical Association* 81:453–460.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. y Riddell, A. (2017) “Stan: A probabilistic programming language.” *Journal of Statistical Software* 76(1):1–32.
- Caughey, D., y Wang, M. (2019) “Dynamic Ecological Inference for Time-Varying Population Distributions Based on Sparse, Irregular and Noisy Marginal Data,” *Political Analysis* 27(3):338–396.
- Cho, W. K. T. (1998) “If the Assumption Fits...: A Comment on the King Ecological Inference Solution.” *Political Analysis* 7: 143–163.
- CIS. (2014) *Estudio 3041. Barómetro octubre 2014*. Madrid: Centro de Investigaciones Sociológicas.
- Colombi, R., Forcina, A.: Latent class models for ecological inference on voters transitions. *Statistical Methods and Applications*. 1, 1–17 (2016)
- Converse, P. E., Jennings, M. K., y Zeigler, L. H. (1966.) The problema of party distances in models of votind change. In M.K, Jennings and L. H., Eds. *The Electoral Process* 175-206. Englerwood Cliffs: Prentice-Hall.
- Corominas A., Lusa, A. y Valvet, M. D. (2015) “Computing Voter Transitions: The Elections for the Catalan Parliament, from 2010 to 2012.” *Journal of Industrial Engineering and Management* 8(1):122–136.

- Dassonneville, R., y Hooghe, M. (2017) “The Noise of the Vote Recall Question: The Validity of the Vote Recall Question in Panel Studies in Belgium, Germany and the Netherlands.” *International Journal of Public Opinion Research* 29(2):316–338.
- Duncan, O. y Davis, B. (1953) “An Alternative to Ecological Correlation.” *American Sociological Review* 18:665–66.
- Fisher, L. H., y Wakefield, J. (2020) “Ecological inference for infectious disease data, with application to vaccination strategies.” *Statistics in Medicine* 39(3):220–238.
- Forcina, A., Gnaldi, M., Bracalente, B. (2012) A revised Brown and Payne model of voting behaviour applied to the 2009 elections in Italy. *Statistical Methods and Applications*. 21, 109–119
- Forcina, A. y Marchetti, G. M. (1989) “Modelling transition probabilities in the analysis of aggregate data.” In A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seber, Eds. *Statistical Modelling*. Springer-Verlag.
- Forcina A. y Marchetti, G. M. (2011) “The Brown and Payne Model of Voter Transition Revisited.” In S. Ingrassia, R. Rocci, and M. Vichi, Eds. *New Perspectives in Statistical Modeling and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer.
- Freedman, D. A., Klein, S. P., Ostland, M. y Roberts, M. R. (1998) “Review of ‘A Solution to the Ecological Inference Problem’.” *Journal of the American Statistical Association* 93:1518–1522.
- Füle, E. (1994) “Estimating Voter Transitions by Ecological Regression.” *Electoral Studies* 13:313–330.
- Galais, C. (2014) “Don’t Vote for Them: The Effects of the Spanish Indignant Movement on Attitudes about Voting.” *Journal of Elections, Public Opinion and Parties* 24:334–350.
- Gibson, J. y Cielecka, A. (1995). Economic Influences on the Political Support for Market Reforms in Post-Communist Transitions: Some Evidence from the 1993 Polish Parliamentary Elections, *Europe-Asia Studies*, 47(5): 765-785.
- Goodman, L. A. (1953) “Ecological Regressions and the Behaviour of Individuals.” *American Sociological Review* 18:663–666.
- Goodman, L. A. (1959) “Some Alternatives to Ecological Correlation.” *American Journal of Sociology* 64(6):610–625.
- Goodman, L. A. (1961) Statistical methods for the mover-stayer model, *Journal of the American Statistical Association*, 56(296): 841-868.
- Goodman, L. A., (1968). The analysis of cross-classified data: independence, quasiindependence, and interactions in contingency tables with or without missing values, *Journal American Statistical Association*, 63: 1091-1131.
- Greenwald, A., C., Carnot, R. Beach y Young, B. (1987). Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology*, 72:315-318.
- Greiner, D. J. y Quinn, K. M. (2009) “R×C ecological inference: Bounds, correlations, flexibility and transparency of assumptions.” *Journal of the Royal Statistical Society. Series A* 172:67–81.
- Greiner, D. y Quinn, K. M. (2010) “Exit Polling and Racial Bloc Voting: Combining Individual-level and RxC Ecological Data.” *The Annals of Applied Statistics* 4:1774–1796.

- Hawkes, A. G. (1969) "An Approach to the Analysis of Electoral Swing." *Journal of the Royal Statistical Society, Series A* 132: 68–79.
- Henn M., y Foard, N. (2012) *Young People and Politics in Britain: How do Young People Participate in Politics and What Can be Done to Strengthen their Political Connection?* London: Nottingham Trent/ESRC.
- Johnston, R. J. y Hay, A. M. (1983) "Voter Transition Probability Estimates: An Entropy Maximizing Approach." *European Journal of Political Research* 11:93–98.
- Katz, J.N. y King, G. (1999). A Statistical Model for Multiparty Electoral Data, *The American Political Science Review*, 93(1): 15-32.
- Keller, W.J. & A. ten Cate (1977). De verschuiving van de kiezersvoorkeur. *Economische Statistische Berichten* 26-10-1977.
- Kellermann, T. (2011) Vom Wahlergebnis zur Wählerwanderung: Welche Wähler wechselten wie ihre Entscheidung. *Stadtforsch. Stat.* 2011(1), 34–40
- King, G. (1997) *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.
- King, G., Rosen, O. y Tanner, M. A. (1999) "Binomial-beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research* 28:61–90.
- King, G., Rosen, O. y Tanner, M. A. Eds. (2004) *Ecological Inference. New Methodological Strategies*. New York: Cambridge University Press.
- Klima, A., Thurner, P. W., Molnar, C., Schlesinger, T. y Küchenhoff, H. (2016) "Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches." *AStA - Advances in Statistical Analysis* 100:133–159.
- Klima, A., Thurner, P. W., Molnar, C. y Küchenhoff, H. (2017) "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions." *Sociological Methods & Research*, online available.
- McCarthy, C. y Ryan, T. M. (1977). Estimates of Voter Transition Probabilities from the British General Elections of 1974. *Journal of the Royal Statistical Society. Series A*, 140, 78–85.
- Miller, W. L. (1972) "Measures of Electoral Change using Aggregate Data." *Journal of the Royal Statistical Society, Series A* 135:122–142.
- Núñez, L. (2016) "Expressive and Strategic Behavior in Legislative Elections in Argentina." *Political Behavior* 38(4):899–920.
- Park, W.-h. (2008) *Ecological Inference and Aggregate Analysis of Elections*. PhD Dissertation. The University of Michigan.
- Park, W., Hanmer, M.J., Biggers, D.R. (2014) Ecological inference under unfavorable conditions: straight and split-ticket voting in diverse settings and small samples. *Electoral Studies*. 36, 192–203
- Pavía, J. M., Bodoque, A. y Martín, J. (2016) "The birth of a new party: Podemos, a hurricane in the Spanish crisis of trust." *Open Journal of Social Sciences* 4:67–86.

- Pavía J. M, Larraz, B. y Montero, J. M. (2008) "Election Forecasts Using Spatiotemporal Models." *Journal of the American Statistical Association* 103:1050–1059.
- Pavía J. M. y Cantarino, I. (2017) "Dasymetric distribution of votes in a dense city." *Applied Geography* 86:22–31.
- Pavía J. M. y López-Quilez, A. (2013) "Spatial Vote Redistribution in Redrawn Polling Units." *Journal of the Royal Statistical Society, Series A* 176:655-678
- Pavía, J. M. y Veres-Ferrer, E. (2016a) "Un nuevo estimador para disgregar totales poblacionales. El caso de los nuevos electores." *Anales de Economía Aplicada XXX*:817–826.
- Pavía, J. M. y Veres-Ferrer, E. (2016b) "Desagregando Estadísticas de Población." In Herrerías, JM and Callejón J (eds.), *Investigaciones en Métodos Cuantitativos para la Economía y la Empresa*, Editorial Universidad de Granada, pp. 543-555.
- Pavía, J. M., Badal, E. y García-Cárceles, B. (2016) "Spanish exit polls: Sampling error or nonresponse bias?" *Revista Internacional de Sociología* 74(3): e043.
- Pavía, J.M. y Aybar, C. (2020). "La movilidad electoral en las elecciones 2019 de la Comunitat Valenciana". *Debats: Revista de Cultura, Poder y Sociedad*, 134, 27–51.
- Pavía-Miralles, J. M. (2005) "Forecasts from Non-Random Samples: The Election Night Case." *Journal of the American Statistical Association* 100:1113–1122.
- Payne, C., Brown, P. y Hanna, V. (1986) "By-election Exit Polls." *Electoral Studies* 5:277–287.
- Plescia, C. y De Sio, L. (2018) "An evaluation of the performance and suitability of RxC methods for ecological inference with known true values." *Quality and Quantity* 52:669–683.
- Puig, X. y J. Ginebra. (2015) "Ecological Inference and Spatial Variation of Individual Behavior: National Divide and Elections in Catalonia." *Geographical Analysis* 47(3):262–283.
- Robinson, W. S. (1950) "Ecological Correlations and the Behavior of Individuals". *American Sociological Review* 15(3):351–357.
- Romero, R. (2014) "Un modelo matemático para estimar el trasvase de votos entre partidos." *Revista Digital de la Real Academia de Cultura Valenciana*, 3–23.
- Romero, R. (2015) "Trasvase de votos entre partidos en las elecciones autonómicas catalanas del 27 de septiembre de 2015." *Revista Digital de la Real Academia de Cultura Valenciana*, 3–15.
- Romero, R. (2016) "Movilidad electoral entre las elecciones del 20D y del 26J en las comunidades autónomas valenciana, madrileña y andaluza." *Revista Digital de la Real Academia de Cultura Valenciana. Segunda Época* 1:1–25.
- Romero, R., Pavía, J.M., Martín, J. y Romero, G. (2020). "Assessing uncertainty of voter transitions estimated from aggregated data. Application to 2017 French presidential elections." *Journal of Applied Statistics*.
- Rosen, O., Jiang, W., King, G. y Tanner, M. A. (2001) "Bayesian and Frequentist Inference for Ecological Inference: The RxC Case." *Statistica Neerlandica* 55:134–56.
- Snelling, C. J. (2016) "Young People and Electoral Registration in the UK: Examining Local Activities to Maximise Youth Registration." *Parliamentary Affairs* 69(3):663–685.

- Thomsen, S.R. (1987) Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference. *Arhus: Politica*.
- Tziafetas, G. (1986) “Estimation of the Voter Transition Matrix.” *Optimization* 17: 275–279.
- Upton, G.J.G. (1977). A Memory Model for Voting Transitions in British Elections, *Journal of the Royal Statistical Society. Series A*, 140: 86-94.
- Upton, G. J. G. (1978) “A Note on the Estimation of Voter Transition Probabilities.” *Journal of the Royal Statistical Society. Series A* 141:507–512.
- van der Ploeg, C. (2008) A Comparison of Different Estimation Methods of Voting Transitions with an Application in the Dutch National Elections. *Centraal Bureau voor de Statistiek*.
- Vázquez, E. y Romero, R. (2001) “Modelos para el estudio del cambio electoral.” *Actas del XXVI Congreso Nacional de Estadística e Investigación Operativa*. Jaén, Spain.
- Voogt, R. (2004). I'm not interested- Nonresponsive bias, response bias and stimulus effects in election research, *Universiteit van Amsterdam*.
- Wakefield, J. (2004) “Ecological Inference for 2x2 Tables (with discussion).” *Journal of Royal Statistical Society, Series A* 167:385–445.
- Weir, B.T. (1975). The Distortion of Voter Recall, *American Journal of Political Science*, 14(1): 53-62.
- Wellhofer, E. S. (2001) Party realignment and voter transition in Italy, 1987-1996, *Comparative Political Studies*, 34(2): 156-86.

www.est2.es

Anejo 1. Software desarrollado para esta tesis

A1.1 FUNCIONES MATLAB

```
function [pm,hete,f,a,b,eik,hetei,tut]=lphom(x,y,rest67)
% obtencion de la matriz de movilidad mediante lphom
% x es la matriz nixnj de votos en la Eleccion 1 en las ni uts (incluye altas)
% y es la matriz nixnk de votos en la Eleccion 2 en las ni uts (incluye bajas)
% si rest67=1 se incluyen las restricciones (6) y (7) en el modelo
% pm es la matriz njxnk de probabilidades de trasvase; eik es la matriz de las eik
% hete es el indice de heterogeneidad esimado; hetei son los ETEI de las ut
% tut son los tamaños de las ut
% a es la matriz (nj+nik+nkj) x njk+2nik) de coeficientes de las restricciones
% b es el vector de 2° miembro de las restricciones

[ni,nj]=size(x); [ni,nk]=size(y); njk=nj*nk; nik=ni*nk;
tut=sum(x,2);
[f,a,b,lb,ub]=calcmoel(x,y,rest67);
opt=optimset('Display','off');
[z,fval]=linprog(f,[],[],a,b,lb,ub,[],opt);
pm=z(1:njk); pm=reshape(pm,nk,nj)';
hete=100*fval/sum(x(:));
e=z((njk+1):length(z)); e=reshape(e,2,nik)';
e=e(:,1)-e(:,2); eik=reshape(e,nk,ni)';
hetei=100*sum(abs(eik),2)./tut;
```

Descripción: resuelve el modelo *lphom* correspondiente a las matrices *x* (matriz $ni \times nj$ con los resultados de E1) e *y* (matriz $ni \times nk$ con los resultados de E2). Si $rest67=1$ se incluyen las restricciones 6 y 7 en el modelo.

```
function [f,a,b,lb,ub]=calcmoel(xc,yc,rest67);
% [f,a,b,lb,ub]=calcmoel(xc,yc); calcula inputs para el PL del modelo
% xc es la matriz nixnj de votos en la Eleccion 1 en las ni uts
% yc es la matriz nixnk de votos en la Eleccion 2 en las ni uts
% si rest67=1 se incluyen las restricciones (6) y (7) en el modelo
% a es la matriz (nj+nik+nkj) x njk+2nik) de coeficientes de las restricciones
% b es el vector de 2° miembro de las restricciones
% f es la funcion a minimizar
% lb y ub son los vectores con los bounds
[ni nj]=size(xc); [ni2 nk]=size(yc);
if ni~=ni2; error('distinto numero de filas en xc e yc');end
njk=nj*nk; nik=ni*nk; nij=ni*nj;
xt=sum(xc); yt=sum(yc);
% al es la matriz relativa a las restricciones sum(pjk)=1
al=zeros(nj,njk);
for j=1:nj; al(j,((j-1)*nk+1):(j*nk))=1;end
% se le añaden los 0 correspondientes a las eik
al=[al zeros(nj,2*nik)];
% se calculan los segundos miembros
b1=ones(nj,1);
% restricciones cuadro total de votos par los nk partidos
at=[]; for j=1:nj; at=[at xt(j)*eye(nk)];end
at=[at zeros(nk,2*nik)];
% se calculan los segundos miembros
bt=yt';
% restricciones relativas al cuadro de votos para ni uts y nk partidos
% ap inicial es la submatriz nik x njk de coeficientes de las pjk en las restricciones. b son los segundos miembros
ap=[]; bp=[];
for i=1:ni
```

```

    ai=[];for j=1:nj;ai=[ai xc(i,j)*eye(nk)];end
    ap=[ap;ai];bp=[bp;yc(i,:)'];
end
% ae es la submatriz(nik,2*nik) de coeficientes de las eik en restricciones
ae=zeros(nik,2*nik);
for f=1:nik
    ae(f,[2*f-1,2*f])=[1 -1];
end
ap=[ap ae];
% se juntan los 3 conjuntos de restricciones
a=[a1;at;ap]; b=[b1;bt;bp];
% si rest67=1 impone que los % de baja son iguales en todos los partidos y p(nj,nk)=0
if nargin==3;
    if rest67==1;
        ab=[];bb=[];pb=yt(nk)/sum(xt(1:nj-1));
        for j=1:nj
            aj=zeros(1,njk+2*nik);aj(nk*j)=1;
            ab=[ab;aj];bb=[bb;pb*(j<nj)];
        end
        a=[a;ab];b=[b;bb];
    end
end
end
% calculo bounds
lb=zeros(njk+2*nik,1);ub=ones(njk+2*nik,1)*Inf;
% calculo funcion economica: descuadre total a minimizar
f=[zeros(1,njk) ones(1,2*nik)];

```

Descripción: calcula, a partir de las dos matrices conocidas de resultados xc e yc , los inputs para el programa lineal del modelo. Estos inputs los utiliza la función *lphom.m* para resolver el modelo el modelo. Si $rest67=1$ se incluyen las restricciones 6 y 7 en el modelo.

function $y=azar(x,d)$

```

% calcula un vector de transicion y "proximo" a x
% el grado de proximidad lo define el valor de d
n=length(x);
y=x+d*2*(rand(1,n)-0.5);% se suma a x un valor al azar entre -d y +d
for k=1:n;
    if y(k)>1;y(k)=1;elseif y(k)<0; y(k)=0;end %se corrigen los valores negativos o > 1
end
if sum(y)==0;y=ones(1,n);end
y=y/sum(y);% se ajustan los valores para que sumen 1

```

Descripción: Esta es la función básica para generar al azar unos resultados electorales. El vector y que da las probabilidades de transición de un partido “origen” a los nk partidos “destino” se obtiene a partir de vector básico inicial x . Las y_k se calculan añadiendo a x_k un valor al azar entre $-d$ y $+d$ y luego reajustando para que estén siempre entre 0 y 1 y sumen 1. Para un mismo partido origen j , cuando en *azarpg* se aplique *azar* en las diferentes unidades territoriales, la heterogeneidad resultará en general tanto mayor cuanto mayor sea el valor de d . (En los estudios de simulación se puede constatar y cuantificar la relación entre d y los índices de heterogeneidad HET y HETe resultantes).

function $pi=azarpi(p,d)$

```

% calcula aleatoriamente una matriz de probabilidades de transicion "proxima" a P
% el grado de proximidad lo define el valor de d
pi=p; [nj,nk]=size(p);
for j=1:nj
    pi(j,:)=azar(p(j,:),d);
end

```

Descripción: aplica la función *azar* nj veces a las nj filas de la matriz de transición para hallar la $[p_{jk}]$ de una ut.

function $pg=azarpg(p,d,ni)$

```

% genera aleatoriamente una hipermatriz de transicion "proxima a p para ni ut
% el grado de proximidad lo define d
[nj,nk]=size(p); pg=zeros(nj,nk,ni);
for i=1:ni
    pg(:,:,i)=azarpi(p,d);
end

```

Descripción: aplica la función *azarpi* ni veces para hallar las matrices $[p_{jk}^i]$ de las ni uts.

```

function [vg]=matvg(x,pg)
% genera una hipermatriz njxnkxni de transicion de votos 'vg'
% 'pg` es la hipermatriz njxnkxni de probabilidades de transicion
% 'x' es la matriz ni x nj de votos obtenidos por cada j en cada i en la Eleccion 1
[nj,nk,ni]=size(pg); vg=zeros(nj,nk,ni);
for i=1:ni
    for j=1:nj
        vg(j,:,i)=x(i,j)*pg(j,:,i);
    end
end
end

```

Descripción: a partir de la hipermatriz de probabilidades de transición **pg** y de la matriz **x** de votos iniciales a cada partido j en cada ut, calcula los votos que pasan da j a k en cada ut.

```

function [x,y,pr,HET,HETI,HETJ]=datsimu(vg)
% a partir de vg genera "y" nixnk con los votos en cada ut a cada partido Eleccion 2
% calcula también índices de heterogeneidad reales HET, HETI y HETJ
% calcula pr njxnk que es la matriz real de probabilidades de transicion
x=squeeze(sum(vg,2))';
y=squeeze(sum(vg,1))';
[HET,HETI,HETJ]=indh(vg);
v=squeeze(sum(vg,3));
[nj,nk,ni]=size(vg);pr=zeros(nj,nk);
for j=1:nj
    pr(j,:)=v(j,:)/sum(v(j,:));
end
end

```

Descripción: a partir de vg calcula las matrices x e y, con los datos conocidos de las dos elecciones, así como la matriz real de movilidad pr y los índices de heterogeneidad reales de la elección.

```

function v=vottr(x,p)
% p es la matriz nj x nk de probabilidades de trasvase
% x es el vector nj con los votos en la Eleccion 1
% v es la matriz nj x nk de trasvase de votos
[nj,nk]=size(p);v=p;
for j=1:nj;v(j,:)=x(j)*p(j,:);end

```

Descripción: calcula el total de votos trasvasados de j a k a partir de xj y de pjk

```

function [IER,IERJ,ERP,ERPJ,HET,HETJ,HETe,pm,pr,y,vg,pg]=simular(x,p,d)
% [IER,IERJ,ERP,ERPJ,HET,HETJ,HETe,pm,pr,y,vg,pg]=simular(x,p,d)
[ni,nj]=size(x); [nj,nk]=size(p); njk=nj*nk;
pg=azarp(p,d,ni);
vg=matvg(x,pg);
[x,y,pr,HET,HETI,HETJ]=datsimu(vg);
[f,a,b,lb,ub]=calcmode(x,y);
opt=optimset('Display','off');
[z,fval]=linprog(f,[],[],a,b,lb,ub,[],opt);
pm=z(1:njk); pm=reshape(pm,nk,nj)';
[ERP,ERPJ]=errp(pr,pm);
xt=sum(x); vtr=vottr(xt,pr); vtm=vottr(xt,pm);
HETe=100*fval/sum(xt);
[IER,IERJ]=inderr(vtr,vtm);
%fprintf(1,'\n\nHET real:%6.1f   HET estimado:%6.1f   IER:%6.1f   ERP:%6.3f\n\n',
HET,HETe,IER,ERP);

```

Descripción: la operativa se ha explicado en el Capítulo 3. La función *simular.m* utiliza todas las funciones que se han descrito, además de las funciones *inderr.m* y *errp.m* que se describen a continuación

```

function [IER,IERJ,IERK,tot,r,m]=inderr(r,m,xj)
% Indices de error total y por partido origen y destino
% r es la matriz njxnk real de transferencia (en total de votos o en probabilidades)
% m es la matriz njxnk estimada de transferencia (en total de votos o en probabilidades)
% si r y m son probabilidades, dar ademas el vector xj
% IER es el indice de error total
% IERJ el vector con los IER para cada partido de la Eleccion 1
% IERK el vector con los IER para cada partido de la Eleccion 2
% tot es el total de votos incorrectos estimados
% en el output r y m son matrices de votos
[nj,nk]=size(r);
if nargin==3
    pr=r; pm=m;
    for j=1:nj; r(j,:)=xj(j)*pr(j,:); m(j,:)=xj(j)*pm(j,:);end
end
difj=0.5*sum(abs(m-r),2); %vector de votos erróneos para cada partido origen
IERJ=100*difj./sum(r,2); % indice de error para cada partido de Eleccion 1
difk=0.5*sum(abs(m-r),1); %vector de votos erróneos para cada partido destino
IERK=100*difk./sum(r,1); % indice de error para cada partido de Eleccion 2
tot=sum(difj); IER=100*tot/sum(r(:)); % indice de error total

```

Descripción: calcula los EI, EIJ y EK

```

function [HET,HETI,HETJ,HETK]=indh(v)
% HET indice de heterogeneidad global. HETI vector de HE en cada ut
% HETJ vector de HE para cada partido origen j
% HETK vector de HE para cada partido destino
% v: matriz njxnkxni de trasvase real de votos
[nj,nk,ni]=size(v);
t=teor(v); % t es la hipermatriz de trasvase de votos bajo homogeneidad perfecta
d=abs(v-t);
HET=50*sum(d(:))/sum(v(:)); %indice4 de heterogeneidad global
%indices de heterogeneidad para la ut
HETI=zeros(1,ni);
for i=1:ni
    x=v(:, :, i); tot=sum(x(:));
    di=d(:, :, i); HETI(i)=50*sum(di(:))/tot;
end
%indices de heterogeneidad para los nj partidos
HETJ=zeros(1,nj);
for j=1:nj
    x=v(j, :, :); tot=sum(x(:));
    dj=d(j, :, :); HETJ(j)=50*sum(dj(:))/tot;
end

```

```

%indices de heterogeneidad para los nk partidos
HETK=zeros(1,nk);
for k=1:nk
    x=v(:,k,:);tot=sum(x(:));
    dk=d(:,k,:);HETK(k)=50*sum(dk(:))/tot;
end

```

Descripción: calcula los índices de heterogeneidad reales a partir de la hipermatriz v de trasvase de votos

```

function [ERP,ERPJ,ERPJ]=errp(pr,pm)
% error medio en la estimacion de las pj
d=abs(pm-pr);
ERPJ=mean(d,2);ERPJ=mean(d,1); ERP=mean(ERPJ);

```

Descripción: calcula los EP, EPJ y EPK

```

function [ie,ep,he,hr,d,pr,pm]=yasimular(x,p,id,nv)
% [ie,ep,he,hr,d,pr,pm]=yasimular(x,p,id,nv) obtencion de ie he y otros indicadores en
% ns=nvxnid simulaciones
% se guardan los resultados en una matriz de texto para utilizarla en R en rel_ie_he
% cambiar en 18 el nombre del archivo de resultados para R
% pr es la matriz de prob transicion promedio real y pm es la estimada por el modelo
nid=length(id);ns=nid*nv;ie=zeros(ns,1);ep=ie;he=ie;hr=ie;d=ie;
pma=0*p;pra=pma;
for nd=1:nid
    for n=1:nv
        s=nv*(nd-1)+n;
        [IER,IERJ,ERP,ERPJ,HET,HETJ,HETe,pm,pr,y,vg,pg]=simular(x,p,id(nd));
        ie(s)=IER;ep(s)=ERP;he(s)=HETe;hr(s)=HET;d(s)=id(nd);
        pma=pma+pm;pra=pra+pr;
    end
end
pm=pma/(nid*nv);pr=pra/(nid*nv);
% se guardan los resultados para R
fid=fopen(['c:\mio\tesis_jorge\R\datheieep.txt'],'wt');
fprintf(fid,' he ie ep\n');
res=[he ie ep];fprintf(fid,['%7.3f %7.3f %7.3f\n'],res);
fclose(fid);

```

Descripción: esta función se utiliza para lanzar varias veces una simulación con diferentes niveles de incumplimiento de la hipótesis de homogeneidad. Los inputs x y p definen el “escenario” del proceso electoral. El vector id contiene los nd valores de “d” que utilizará en cada simulación *azarp* para generar aleatoriamente la hipermatriz pg; para cada valor de “d” se realizarán nv simulaciones. En total se realizan ndxnv simulaciones cuyos resultados se guardan en los vectores outputs correspondientes (los IER en ie, los ERP en ep, los HETe en he, los HET en hr).

Se guardan además los resultados en una matriz de texto en c:\mio\tesis_jorge\R para utilizarla en R (script *rel_ie_he.R*) para hacer el análisis con gráficos de la relación entre ie y he

```

function [ei,limei,he0,div,rei]=estimei(x,y,nv,conf)
% estima el ei y limey superior para el resultado de lphom(x,y)
% nv numero de simulaciones para cada uno de los 11 valores de d
% conf:nivel de confianza del intervalo
[pm,he0]=lphom(x,y,0);div=divut(x);
[ni nj]=size(x);[ni nk]=size(y);rei=(nj+nk+ni*nk)/(nj*nk);
[d0,id0]=hallard0(x,pm,he0);
[b,sr]=ajust(x,pm,d0,id0,nv);
lie=b'*[1 log(he0) (log(he0)).^2]';

```

```
lлим=lie+normal(conf)*sr;
ei=exp(lie);limei=exp(lлим);
```

Descripción: la operativa de esta función se describe con detalle en el Apartado 5.3 de la tesis.

```
function [d0,id0,he,d]=hallard0(x,p,he0)
% [d0,id0]=hallard0(x,p,he0) aproxima el valor d0 que corresponde a he0
% en el escenario definido por x y p
% id0 es el incremento de d que conduce a un incremento de 1 en he
[z z he z d]=yasimular(x,p,[.01 .03 .05 .08 .1 0.15 .2 .25 .3],2);
if he0>max(he);
[z z he z d]=yasimular(x,p,2* [.01 .03 .05 .08 .1 0.15 .2 .25 .3],2);end
% ajustar a una recta d en funcion de he
b=regre(he,d,0);
d0=he0/b(1);id0=1/b(1);
```

Descripción: halla aproximadamente el valor do que corresponde a un HETe igual a he0 para el escenario asociado a x y p. id0 es el incremento de d que conduce a un incremento de 1 en he

```
function [b,sr]=ajust(x,pm,d0,id0,nv)
% (b,sr)=ajust(x,pm,d0,id0) halla el ajuste de ie a he (ajuste cuadratico en logaritmos)
id = d0 + id0*((1:11)-6);
while id(1)<0
    id0=id0/2;id = d0 + id0*((1:11)-6);
end
[ie,z,he]=yasimular(x,pm,id,nv);
[b,sr]=regre(log(ie),[log(he),(log(he)).^2],1);
```

Descripción: halla el ajuste de ie a he (ajuste cuadrático en logaritmos)'. La función se describe con detalle en el Apartado 5.3 de la tesis.

```
function [prm,pmm,HETm,HETem]=prueba2(x,p,d,nv)
% resultados promedios de nv simulaciones para un valor de d
prm=0*p;pmm=prm;HETm=0;HETem=0;
for n=1:nv
    [IER,IERJ,ERP,ERPJ,HET,HETJ,HETe,pm,pr,y,vg,pg]=simular(x,p,d);
    prm=prm+pr;pmm=pmm+pm;HETm=HETm+HET;HETem=HETem+HETe;
end
prm=prm/nv;pmm=pmm/nv;HETm=HETm/nv;HETem=HETem/nv;
```

Descripción: función utilizada en el Capítulo 7 para estudiar la relación entre el sesgo de los p_{jk}^* y su verdadero valor p_{jk}

```
function [ep,limep]=estimep(x,y,nv,conf)
% [ep,limep]=estimep(x,y,nv,alfa) estima ep y limep superior para el resultados delphom(x,y)
% nv numero de simulaciones para cada uno de los 11 valores de d
% conf:nivel de confianza del intervalo
[pm,he0]=lphom(x,y,0);
[ni nj]=size(x);[ni nk]=size(y);rei=(nj+nk+ni*nk)/(nj*nk);
[d0,id0]=hallard0(x,pm,he0);
[b,sr]=ajustep(x,pm,d0,id0,nv);
lep=b'*[1 log(he0) (log(he0)).^2]';
lлим=lep+normal(conf)*sr;
ep=exp(lep);limep=exp(lлим);
```

Descripción: la operativa de esta función se describe con detalle en el Apartado 6.3 de la tesis.

```

function [b,sr]=ajustep(x,pm,d0,id0,nv)
% (b,sr)=ajust(x,pm,d0,id0) halla el ajuste de ie a he (ajuste cuadratico en logaritmos)
id = d0 + id0*((1:11)-6);
while id(1)<0
    id0=id0/2;id = d0 + id0*((1:11)-6);
end
[z,ep,he]=yasimular(x,pm,id,nv);
[b,sr]=regre(log(ep), [log(he), (log(he)).^2],1);

```

Descripción: halla el ajuste de ep a he (ajuste cuadrático en logaritmos)'. La función se describe con detalle en el Apartado 6.3 de la tesis.

script **pruebapjk**

```

% para estudiar la relacion de er y var(er) con he
d=[0.01 0.05 0.1,0.15,0.2,0.25];
hem=[];erm=[];vam=[];
for i=1:6
    [he,pr,pe,er]=yasimupjk(xf,pm,1,1,d(i),200);
    hem=[hem;mean(he)];erm=[erm;mean(er)];vam=[vam;var(er)];
end
b=regre(erm,[hem hem.^2],0);
x=min(hem):0.01:(max(hem)+0.5);y=b(1)*x + b(2)*x.^2;
figure(1);
plot(hem,erm,'*b',x,y);xlabel('HETe promedio');ylabel('error promedio')
title('Relacion entre Error Pjk y HETe (Pjk=0.80)','FontWeight','bold','Color',[1 0 0]);
b=regre(vam,[hem hem.^2],0);
y=b(1)*x + b(2)*x.^2;
figure(2);
plot(hem,vam,'*r',x,y);xlabel('HETe promedio');ylabel('Var(Error)')
title('Relacion entre Var(Error Pjk) y HETe (Pjk=0.80)','FontWeight','bold','Color',[1 0 0]);

```

Descripción: En realidad, más que una función *pruebapjk.m* es un script, un conjunto de instrucciones que se ejecutan cuando se lanza cogiendo los valores de las variables que hay en el espacio de trabajo y dejando allí sus resultados. Por ejemplo, tal como está estudia el p₁₁ de las elecciones francesas; para estudiar el p₁₃ hay que cambiar la sentencia 5ª.

```

function [he,pr,pe,er]=yasimupjk(x,p,j,k,id,nv)
% [he,pr,pe,er]=yasimupjk(x,p,j,k,id,nv) estudio error estimacion pjk
nid=length(id);ns=nid*nv;pr=zeros(ns,1);pe=pr;he=pr;
pma=0*p;pra=pma;
for nd=1:nid
    for n=1:nv
        s=nv*(nd-1)+n;
        [z,z,z,z,z,z,hete,pmod,preal]=simular(x,p,id(nd));
        pr(s)=preal(j,k);pe(s)=pmod(j,k);he(s)=hete;
    end
end
er=pe-pr;

```

Descripción: lanza *simular.m* nv veces para cada uno de los valores de d recogidos en id. Guarda los resultados necesarios para obtener el intervalo de confianza para p_{jk} mediante *intconf.m*

```

function [pjk0,pinf,psup]=intconf(x,y,j,k,conf,nv)
% [pjk0,pinf,psup]=intconf(x,y,j,k,alfa,nv)
% intervalo de confianza alfa para la estimacion de pjk
% nv numero de simulaciones para cada uno de los 11 valores de d
% % conf:nivel de confianza del intervalo
[pm,he0]=lphom(x,y,0);pjk0=pm(j,k);
[d0,id0]=hallard0(x,pm,he0);

```

```

id = d0 + id0*((1:11)-6);
while id(1)<0
    id0=id0/2;id = d0 + id0*((1:11)-6);
end
[he,pr,pe,er]=yasimupjk(x,pm,j,k,id,nv);
%ajuste lineal del sesgo en funcion de he
[b,sr,res]=regre(er,[he he.^2],0);eom=b(1)*he0+b(2)*he0^2;
% ajuste lineal de la varianz del sesgo en funcion de he
a=regre(res.^2,[he he.^2],0);seo=sqrt(a(1)*he0+a(2)*he0^2);
z=normal(conf+(1-conf)/2);
li=eom-z*seo;psup=pjk0-li;psup=min(1,psup);
ls=eom+z*seo;pinf=pjk0-ls;pinf=max(0,pinf);
fprintf(1,'\n\nEstimacion de P%i,%i e intervalo de confianza (conf:%3i%%)\n\n',j,k,100*conf);
fprintf(1,'Estimacion:%6.3f   int.conf [%5.3f   %5.3f]\n\n',pjk0,pinf,psup);

```

Descripción: Obtiene el intervalo de confianza de nivel de confianza $conf$ para p_{jk} en la elección definida por x e y . nv es el número de simulaciones realizadas de acuerdo con la metodología descrita en el capítulo 7 de la tesis.

A1.2 SOFTWARE EN R

script `rel_ei_he.R` para estudiar la relación entre EI y HETe

```
#####
#relación entre ie y he en un escenario#
#####

# modificar read.table en cada caso
rm(list=ls())
dat=read.table('dhrheei.txt',header=T)
names(dat)
attach(dat)
par(bg="lightyellow");par(mfrow=c(1,2))
# gráfico de ie frente a he
plot(ie~he,pch=4,col=4,main="Relación entre IER y HETe",
xlab="Heterogeneidad estimada (HETe)",ylab="Índice de error (IER)")
# gráfico de log(ie) frente a log(he)
plot(log(ie)~log(he),pch=4,col=4,main="Relación entre log(IER) y log(HETe)",
xlab="log(HETe)",ylab="log(IER)")
# Ajuste del modelo
lhe=log(he);lhe2=(log(he))^2
mod=lm(log(ie)~lhe+lhe2)
summary(mod);
sres2=sum(mod$resid^2)/mod$df.resid
# Gráficos de residuos
win.graph();par(bg="lightyellow");par(mfrow=c(1,2))
hist(mod$resid,density=50,col=4,main='Histograma de los residuos del ajuste',xlab='residuos')
plot.normal=function(x)
{n=length(x);x=sort(x)
p<-(1:n)-0.5)/n;q<-qnorm(p)
plot(x,q,xlab="residuos",ylab="porcentajes",yaxt="n",
main="gráfico en papel probabilístico de los residuos",pch=16,col=4)
pr<-c(0.1,1,5,10,25,50,75,90,95,99,99.9);q<-qnorm(pr/100)
for (i in 1:11) {abline(q[i],0,lty=3,col=2);text(min(x),q[i],pr[i],col=2)}}
plot.normal(mod$resid)
#grafico del ajuste en escala logarítmica
mi=min(lhe);ma=max(lhe);x=seq(mi,ma,(ma-mi)/500)
pred=predict(mod,newdata=data.frame(lhe=x,lhe2=x^2),se.fit=T)
win.graph();par(bg="lightyellow");par(mfrow=c(1,2))
plot(log(ie)~log(he),pch=4,col=4,main="Modelo de la relación entre log(IER) y log(HETe)",
xlab="log(HETe)",ylab="log(IER)")
lines(pred$fit~x,type="l",lwd=4,col=1)
se.obs<-sqrt(pred$se.fit^2+sres2)
lines((pred$fit+1.28*se.obs)~x,lty=4,lwd=4,col=2)
legend(min(log(he)),max(log(ie)),legend=c("IER previsto","límite confianza
90%"),cex=0.7,lty=c(1,2),lwd=c(3,2),col=c(1,2),bty="n")
#grafico del ajuste en escala natural
plot(ie~he,pch=4,col=4,main="Modelo de la relación entre IER y HETe",
xlab="Heterogeneidad estimada (HETe)",ylab="Índice de error (IER)")
x=exp(x);y=exp(pred$fit);ylim=exp(pred$fit+1.28*se.obs)
lines(y~x,type="l",lwd=4,col=1)
lines(ylim~x,lty=2,lwd=4,col=2)
legend(0,max(ie),legend=c("IER previsto","límite confianza
90%"),cex=0.7,lty=c(1,2),lwd=c(3,2),col=c(1,2),bty="n")
#Tabla de valores previstos
hete=seq(0.5,10,0.5);lx=log(hete)
pred=predict(mod,newdata=data.frame(lhe=lx,lhe2=lx^2),se.fit=T)
se.obs<-sqrt(pred$se.fit^2+sres2)
ier=exp(pred$fit);limier=exp(pred$fit+1.28*se.obs)
for (i in 1:20)
{cat('\n',sprintf('%4.1f    %4.1f    %4.1f',hete[i],ier[i],limier[i]))}
detach(dat)
```

script rel_ep_he.R para estudiar la relación entre EP y HETe

```
#####  
#relación entre ep y he en un escenario#  
#####  
# modificar read.table en cada caso  
rm(list=ls())  
dat=read.table('simuF.txt',header=T)  
names(dat)  
attach(dat)  
par(bg="lightyellow");par(mfrow=c(1,2))  
# gráfico de ep frente a he  
plot(ep~he,pch=4,col=4,main="Relación entre EP y HETe",  
xlab="Heterogeneidad estimada (HETe)",ylab="Índice de error (EP)")  
# gráfico de log(ie) frente a log(he)  
plot(log(ep)~log(he),pch=4,col=4,main="Relación entre log(EP) y log(HETe)",  
xlab="log(HETe)",ylab="log(EP)")  
# Ajuste del modelo  
lhe=log(he);lhe2=(log(he))^2  
mod=lm(log(ep)~lhe+lhe2)  
summary(mod);  
sres2=sum(mod$resid^2)/mod$df.resid  
# Gráficos de residuos  
win.graph();par(bg="lightyellow");par(mfrow=c(1,2))  
hist(mod$resid,density=50,col=4,main='Histograma de los residuos del ajuste',xlab='residuos')  
plot.normal=function(x)  
{n=length(x);x=sort(x)  
p<-(1:n)-0.5)/n;q<-qnorm(p)  
plot(x,q,xlab="residuos",ylab="porcentajes",yaxt="n",  
main="gráfico en papel probabilístico de los residuos",pch=16,col=4)  
pr<-c(0.1,1,5,10,25,50,75,90,95,99,99.9);q<-qnorm(pr/100)  
for (i in 1:11) {abline(q[i],0,lty=3,col=2);text(min(x),q[i],pr[i],col=2)}}  
plot.normal(mod$resid)  
#grafico del ajuste en escala logarítmica  
mi=min(lhe);ma=max(lhe);x=seq(mi,ma,(ma-mi)/500)  
pred=predict(mod,newdata=data.frame(lhe=x,lhe2=x^2),se.fit=T)  
win.graph();par(bg="lightyellow");par(mfrow=c(1,2))  
plot(log(ep)~log(he),pch=4,col=4,main="Modelo de la relación entre log(EP) y log(HETe)",  
xlab="log(HETe)",ylab="log(EP)")  
lines(pred$fit~x,type="l",lwd=4,col=1)  
se.obs<-sqrt(pred$se.fit^2+sres2)  
lines((pred$fit+1.28*se.obs)~x,lty=4,lwd=4,col=2)  
legend(min(log(he)),max(log(ie)),legend=c("EP previsto","límite confianza  
90%"),cex=0.7,lty=c(1,2),lwd=c(3,2),col=c(1,2),bty="n")  
#grafico del ajuste en escala natural  
plot(ep~he,pch=4,col=4,main="Modelo de la relación entre EP y HETe",  
xlab="Heterogeneidad estimada (HETe)",ylab="Índice de error (IER)")  
x=exp(x);y=exp(pred$fit);ylim=exp(pred$fit+1.28*se.obs)  
lines(y~x,type="l",lwd=4,col=1)  
lines(ylim~x,lty=2,lwd=4,col=2)  
legend(0,max(ie),legend=c("EP previsto","límite confianza  
90%"),cex=0.7,lty=c(1,2),lwd=c(3,2),col=c(1,2),bty="n")  
#Tabla de valores previstos  
hete=seq(0.5,10,0.5);lx=log(hete)  
pred=predict(mod,newdata=data.frame(lhe=lx,lhe2=lx^2),se.fit=T)  
se.obs<-sqrt(pred$se.fit^2+sres2)  
epr=exp(pred$fit);limepr=exp(pred$fit+1.28*se.obs)  
for (i in 1:20)  
{cat('\n',sprintf('%4.1f    %5.3f    %5.3f',hete[i],epr[i],limepr[i]))}  
detach(dat)
```

Anejo 2. Elecciones presidenciales francesas de 2017

A2.1 Resultados desagregados a nivel regional

Table A2.1 Resultados de la primera vuelta.

Region	Abstención¹	Macron	Le Pen	Fillon	Mélechon	Hamon	Dupont	Others
Auvergne-Rhône-Alpes	1224975	1026415	867941	846154	805862	256658	215981	170514
Bourgogne-Franche-Comté	456487	338192	387664	304380	276940	87372	87266	63058
Brittany	452907	581075	306642	380815	385734	180827	87928	77302
Centre-Val de Loire	399999	323728	329468	300316	252303	83552	82057	56174
Grand Est	902018	615681	825490	586321	484650	151311	182153	125599
Hauts-de-France	1008896	630405	1003220	521392	633288	166637	160729	117021
Île-de-France	1562804	1612736	708261	1249629	1225218	430386	226249	179855
Normandy	505109	423075	452702	370188	362535	113744	98957	70759
Nouvelle-Aquitaine	939316	851370	640226	602881	703481	240175	155600	195963
Occitanie	878653	739929	761984	565993	734054	216332	135377	161590
Pays de la Loire	513080	575831	364266	516428	403454	143491	109842	78604
Provence-Alpes-Côte d'Azur	814427	520896	774783	615451	515419	113343	119023	91819
Overseas regions and Corsica	1864516	417993	256846	353849	277947	107737	34024	72303

¹ Votos en blanco y nulos incluidos

Table A2.2 Resultados de la segunda vuelta.

Region	Abstención	Blancos/Nulos	Macron	Le Pen
Auvergne-Rhône-Alpes	1273000	488262	2452414	1200824
Bourgogne-Franche-Comté	453247	198437	816258	533417
Brittany	498380	228216	1301172	425462
Centre-Val de Loire	425879	172389	778260	451069
Grand Est	956141	327035	1500666	1089381
Hauts-de-France	1060357	352316	1497525	1331390
Île-de-France	1820439	514719	3825985	1033995
Normandy	546996	216266	1012264	621543
Nouvelle-Aquitaine	958717	440952	2011131	918212
Occitanie	976998	420844	1761223	1034847
Pays de la Loire	580157	252074	1356172	516593
Provence-Alpes-Côte d'Azur	912800	295860	1307163	1049338
Overseas regions and Corsica	1613189	172230	1139849	459947

A2.2 Análisis mediante *ei.MD.bayes*

Para realizar este análisis hemos utilizado la función *ei.MD.bayes* disponible en el paquete *ei.Pack* de R.

El análisis se ha llevado a cabo mediante el siguiente script preparado para el software R:

```
# estimacion ei.MD a elecciones francesas
library(eiPack); library(tictoc)
dat = read.table('datfrance.txt',header=T)
tic() # para empezar a medir el tiempo de cálculo
ff <- ei.MD.bayes(cbind(ABST2, BLNUL, MACR2, LPEN2) ~
                 cbind(ABST1, MACR1, LPEN1, FILL, MELE, HAMO, DUPO, REST),
                 data = dat, sample = 50000)
# obtención de la matriz de trasvase de votos a partir de las medias de las 1000 simulaciones
x = apply(ff$draws$Cell.counts, 2, mean)
v = matrix(x, 8, 4)
colnames(v) = c("ABST2", "BLNUL", "MACR2", "LPEN2")
rownames(v) = c("ABST1", "MACR1", "LPEN1", "FILL", "MELE", "HAMO", "DUPO", "REST")
# cálculo de p a partir de v
p = v; for (i in 1:8) {p[i,]=100*v[i,]/sum(v[i,])}
toc() # para finalizar la medición del tiempo de cálculo
p
```

Utilizando `sample = 1000`, que es la opción propuesta por defecto, hemos encontrado que los resultados varían mucho entre dos implementaciones consecutivas. Finalmente hemos utilizado `sample = 50000`, si bien, aun así, los resultados todavía fluctúan algo de unas aplicaciones a otras de la función. Hay que recordar que la función *ei.MD.Bayes* utiliza simulación basada en cadenas de Markov de Monte Carlo, por lo que salvo que se fije la semilla aleatoria, cada aplicación de la función conducirá a soluciones (ligeramente) diferentes.

El tiempo de cálculo necesario para realizar un análisis ha sido de 10.26 segundos, frente a los 0.04 segundos que se han necesitado para realizar el análisis mediante *lphom*.

En la siguiente tabla se recogen, expresados como porcentajes, los valores de los p_{jk} obtenidos en la estimación mediante *ei.MD.bayes*

Tabla A23 Estimaciones de los p_{jk} (%) obtenidos mediante *ei.MD.bayes*. Elecciones presidenciales francesas 2017.

		Opciones electorales de destino			
		ABST2	BLNUL	MACR2	LPEN2
Opciones Electorales Origen	ABST1	23.21190	7.950198	43.58622	25.25168
	MACR1	24.67868	10.547709	44.89457	19.87905
	LPEN1	26.56576	10.316127	38.29070	24.82742
	FILL	27.02938	7.272839	43.36187	22.33591
	MELE	30.10165	5.032586	46.27902	18.58675
	HAMO	28.00332	6.999319	46.59236	18.40500
	DUPO	16.39864	15.148756	34.71134	33.74127
	REST	16.39394	10.952440	58.40615	14.24747

En el Apartado 2.5 del Capítulo 2 se comentan estos resultados.

Anejo 3. Descripción de los escenarios utilizados en el capítulo 4

A3.1 Escenario 1

Este escenario está inspirado en el estudio de la movilidad electoral en Andalucía entre las elecciones autonómicas de 2012 y 2015 y definido por las matrices **XAND** y **PAND**, y viene definido por las siguientes características.

Total electores: 6.286.000.

Número de uts: 8 de tamaño comprendido entre 390 y 1501 miles de electores.

Opciones en Elección 1: 6 de tamaño comprendido entre 350 y 1882 miles de votos.

Opciones en Elección 2: 6 de tamaño comprendido entre 411 y 1969 miles de votos.

La matriz **PAND** de movilidad electoral básica corresponde a una situación en la que las opciones electorales de la Elección 1 y la Elección 2 son las mismas y hay un grado bastante marcado de fidelidad en el comportamiento electoral, que se traduce en valores elevados en la diagonal principal de la matriz.

Índice DIV de disparidad entre las uts en la Elección 1: 10.3.

Ratio REI entre ecuaciones e incógnitas del modelo *lphom*: 1.67.

Matriz XAND: Votos obtenido en la Elección 1 en cada unidad territorial (ut) por cada opción electoral (OE1).

	OE1 1	OE1 2	OE1 3	OE1 4	OE1 5	OE1 6
ut 1	146145	89022	117407	39482	44320	16214
ut 2	330545	180667	179054	130215	94707	53717
ut 3	172882	149628	141917	68451	55655	46552
ut 4	197933	157585	158328	83051	70597	35391
ut 5	133442	95592	74479	39349	30709	16233
ut 6	138028	148397	121850	48362	41298	24235
ut 7	345353	208405	224282	132345	132185	71375
ut 8	417850	371103	275335	207826	142301	86360

Matriz PAND: Matriz básica de trasvase de votos entre cada opción electoral de la Elección 1 (OE1) y cada opción electoral de la Elección 2 (OE2).

	OE2 1	OE2 2	OE2 3	OE2 4	OE2 5	OE2 6
OE1 1	0.93	0.02	0.03	0.00	0.01	0.01
OE1 2	0.02	0.80	0.05	0.06	0.03	0.04
OE1 3	0.04	0.05	0.82	0.04	0.03	0.02
OE1 4	0.12	0.03	0.01	0.71	0.08	0.05
OE1 5	0.04	0.07	0.02	0.03	0.80	0.04
OE1 6	0.07	0.02	0.03	0.04	0.13	0.71

A3.2 Escenario 2

Este escenario está inspirado en el estudio de la movilidad electoral en la Comunidad Valenciana entre las elecciones autonómicas de 2015 y las generales del mismo año y definido por las matrices **XVAL** y **PVAL** y viene definido por:

Total electores: 3.609.000.

Número de uts: 14 de tamaño comprendido entre 121 y 578 miles de electores.

Opciones en Elección 1: 8 de tamaño comprendido entre 107 y 1168 miles de votos.

Opciones en Elección 2: 7 de tamaño comprendido entre 112 y 864 miles de votos.

La matriz **PVAL** de movilidad electoral básica corresponde a una situación en la que las fidelidades electorales en las posibles opciones que se han presentado a las dos elecciones no son muy altas y en la que una opción presentada en la Elección 2 (OE2 4) agrupaba a dos opciones (OE1 4 y OE1 6) que se presentaron por separado en la Elección 1.

Índice DIV de disparidad entre las uts en la Elección 1: 13.9.

Ratio REI entre ecuaciones e incógnitas del modelo *lphom*: 2.02.

Matriz XVAL de votos obtenido en la Elección 1 en cada unidad territorial (ut) por cada opción electoral (OE1).

	OE 1	OE 2	OE 3	OE 4	OE 5	OE 6	OE 7	OE 8
ut 1	127306	53789	41657	26272	38573	33404	10560	11852
ut 2	72129	38839	31378	17459	20360	17336	4271	6719
ut 3	64597	47561	34938	4500	16190	11270	5368	5039
ut 4	43268	22538	23539	8044	11206	9443	5792	5728
ut 5	142896	62208	52434	45093	28395	27339	9333	9927
ut 6	55417	32798	23526	17908	16225	16224	3085	6527
ut 7	74086	53338	46387	23802	16049	17456	6122	7854
ut 8	167494	107539	55423	96151	60441	50964	17303	22687
ut 9	31276	26377	18226	24789	7668	6845	3129	3052
ut 10	39399	21419	19052	20231	10987	10291	4658	5466
ut 11	48302	28666	23928	28267	14823	14428	5334	6073
ut 12	80830	38017	40895	36228	21576	22690	7930	10662
ut 13	45973	32930	24609	32688	11130	9730	5110	4781
ut 14	175112	92593	73106	75391	35498	34969	18922	11853

Matriz PVAL: Matriz básica de trasvase de votos entre cada opción electoral de la Elección 1 (OE1) y cada opción electoral de la Elección 2 (OE2).

	OE2 1	OE2 2	OE2 3	OE2 4	OE2 5	OE2 6	OE2 7
OE1 1	0.4452	0.0926	0.1308	0.0622	0.1212	0.0052	0.1429
OE1 2	0.1829	0.8171	0.0000	0.0000	0.0000	0.0000	0.0000
OE1 3	0.2292	0.0000	0.5821	0.1802	0.0000	0.0000	0.0085
OE1 4	0.1073	0.0228	0.0807	0.6528	0.0351	0.1013	0.0000
OE1 5	0.1512	0.4293	0.0000	0.0000	0.4195	0.0000	0.0000
OE1 6	0.0000	0.0000	0.1576	0.6314	0.2111	0.0000	0.0000
OE1 7	0.0000	0.4452	0.0000	0.0000	0.0000	0.5548	0.0000
OE1 8	0.0958	0.0000	0.0000	0.2553	0.6489	0.0000	0.0000

A3.3 Escenario 3

Este escenario está inspirado en el estudio de la movilidad electoral en la Comunidad de Madrid entre las elecciones generales del 20D de 2015 y del 26J de 2016, a partir de los resultados registrados en la ciudad de Madrid. El escenario viene definido por las siguientes características:

Total electores: 2.561.000.

Número de uts: 20 de tamaño comprendido entre 65 y 179 miles de electores.

Opciones en Elección 1: 7 de tamaño comprendido entre 65 y 667 miles de votos.

Opciones en Elección 2: 6 de tamaño comprendido entre 51 y 723 miles de votos.

La matriz básica de movilidad electoral **PMAD** corresponde a una situación en la que las cinco primeras opciones electorales de la Elección 1 y la Elección 2 son las mismas y hay un grado alto de fidelidad en el comportamiento electoral, que se traduce en valores elevados en los elementos correspondientes de la matriz.

Índice DIV de disparidad entre las uts en la Elección 1: 17.1.

Ratio REI entre ecuaciones e incógnitas del modelo *lphom*: 3.17.

Matriz XMAD de votos obtenido en la Elección 1 en cada unidad territorial (ut) por cada opción electoral (OE1).

	OE 1	OE 2	OE 3	OE 4	OE 5	OE 6	OE 7
ut 1	50286	38600	25191	28451	18936	7215	4050
ut 2	40461	47401	20749	22340	21114	6190	3914
ut 3	35595	58035	21491	22590	30003	6216	4751
ut 4	44763	46091	25356	29772	19538	7517	4517
ut 5	50646	24152	30096	37657	12883	7811	3297
ut 6	27352	38932	16476	18761	22516	4783	3333
ut 7	24744	30584	14072	22264	17394	6430	3118
ut 8	21189	46986	8672	10119	18834	3134	2850
ut 9	23324	46533	8958	10488	17292	2953	2925
ut 10	22450	41195	9601	12950	16803	3833	2967
ut 11	35178	33544	25451	26838	20704	6118	3682
ut 12	32434	24574	26906	25882	20109	5716	3676
ut 13	29714	27777	27011	26650	17510	6538	3548
ut 14	30970	30927	22248	21327	19761	5330	3652
ut 15	24906	25896	22815	24955	16697	7139	2972
ut 16	26038	30500	20136	21653	17544	5535	2953
ut 17	20154	18773	13069	14385	12880	3517	2242
ut 18	17277	21780	11073	11243	12506	2632	2052
ut 19	20932	12451	12334	14844	10407	3554	2086
ut 20	11800	22323	5379	6865	14433	1715	1999

**Matriz PMAD: Matriz básica de trasvase de votos entre
cada opción electoral de la Elección 1 (OE1) y cada opción electoral de la Elección 2 (OE2).**

	OE2 1	OE2 2	OE2 3	OE2 4	OE2 5	OE2 6
OE1 1	0.9253	0.0000	0.0013	0.0000	0.0000	0.0735
OE1 2	0.0000	0.9984	0.0000	0.0000	0.0000	0.0016
OE1 3	0.2577	0.0000	0.7423	0.0000	0.0000	0.0000
OE1 4	0.0445	0.0000	0.0756	0.8799	0.0000	0.0000
OE1 5	0.0088	0.1603	0.0000	0.0000	0.8309	0.0000
OE1 6	0.0000	0.0000	0.4955	0.3585	0.1460	0.0000
OE1 7	0.2866	0.0000	0.4295	0.0000	0.1761	0.1078

A3.4 Escenario 4

Este escenario está inspirado en el estudio de la movilidad electoral en Cataluña entre las elecciones autonómicas de 2012 y 2015, con características:

Total electores: 5.578.000.

Número de uts: 10 de tamaño comprendido entre 295 y 1188 miles de electores.

Opciones en Elección 1: 10 de tamaño comprendido entre 126 y 1832 miles de votos.

Opciones en Elección 2: 9 de tamaño comprendido entre 149 y 1621 miles de votos.

La matriz de movilidad electoral básica **PCAT** corresponde a una situación muy fluida a causa de la aparición de Podemos y de la alianza en Junts per el Sí de CIU y ERC, lo que se traduce en que no existan tantos valores extremos de las p_{jk} como en otros escenarios.

Índice DIV de disparidad entre las uts en la Elección 1: 17.5.

Ratio REI entre ecuaciones e incógnitas del modelo *lphom*: 1.21.

Matriz XCAT de votos obtenido en la Elección 1 en cada unidad territorial (ut) por cada opción electoral (OE1).

	OE 1	OE 2	OE 3	OE 4	OE 5	OE 6	OE 7	OE 8	OE 9	OE 10
ut 1	205018	113657	54093	48642	53591	24538	26039	12840	18284	17089
ut 2	110531	89035	36011	21598	23338	11145	6881	6302	8333	9117
ut 3	161279	148237	61358	34688	33096	20397	12341	14499	16187	16078
ut 4	345955	235488	104651	97517	118852	95694	65910	31635	40389	52308
ut 5	157034	42303	23087	65510	48068	36623	28973	6175	21147	11290
ut 6	190684	78137	42146	79272	56333	51739	42615	9978	27790	19974
ut 7	94782	78295	31268	26018	27322	19148	14828	6988	13623	8343
ut 8	200105	115588	51044	80416	53685	51706	43697	14297	28600	24039
ut 9	90561	61105	27249	28434	23458	21489	16622	6257	12862	6002
ut 10	275647	154414	67217	42612	33938	27226	17101	17464	23139	0

Matriz PCAT: Matriz básica de trasvase de votos entre cada opción electoral de la Elección 1 (OE1) y cada opción electoral de la Elección 2 (OE2).

	OE2 1	OE2 2	OE2 3	OE2 4	OE2 5	OE2 6	OE2 7	OE2 8	OE2 9
OE1 1	0.4963	0.1249	0.1784	0.0179	0.0000	0.0640	0.0000	0.0699	0.0486
OE1 2	0.0000	0.9514	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0486
OE1 3	0.0000	0.6629	0.0000	0.0000	0.0000	0.0000	0.2885	0.0000	0.0486
OE1 4	0.0000	0.0000	0.0000	0.6746	0.2768	0.0000	0.0000	0.0000	0.0486
OE1 5	0.2989	0.0000	0.0926	0.0697	0.0000	0.4902	0.0000	0.0000	0.0486
OE1 6	0.0586	0.0000	0.0000	0.0294	0.6151	0.0000	0.2012	0.0471	0.0486
OE1 7	0.0000	0.0000	0.9514	0.0000	0.0000	0.0000	0.0000	0.0000	0.0486
OE1 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9514	0.0000	0.0486
OE1 9	0.0253	0.0000	0.4884	0.4377	0.0000	0.0000	0.0000	0.0000	0.0486
OE1 10	0.9779	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0221	0.0000

A3.5 Escenario 5

Escenario inspirado en el estudio de la movilidad electoral en Aragón entre las elecciones autonómicas de 2011 y 2015, con las siguientes características.

Total electores: 1.043.000.

Número de uts: 15 de tamaño comprendido entre 12 y 226 miles de electores.

Opciones en Elección 1: 9 de tamaño comprendido entre 16 y 340 miles de votos.

Opciones en Elección 2: 11 de tamaño comprendido entre 6 y 333 miles de votos.

La matriz de movilidad básica **PARA** corresponde a una situación en la que la aparición de las dos nuevas formaciones, Podemos y Ciudadanos, se traduce en la inexistencia de valores muy altos de las p_{jk} , siendo 0.75 el mayor valor postulado para estas probabilidades de transición.

Índice DIV de disparidad entre las uts en la Elección 1: 19.1.

Ratio REI entre ecuaciones e incógnitas del modelo *lphom*: 1.87.

Matriz XARA de votos obtenido en la Elección 1 en cada unidad territorial (ut) por cada opción electoral (OE1).

	OE 1	OE 2	OE 3	OE 4	OE 5	OE 6	OE 7	OE 8	OE 9
ut 1	55147	44320	40209	14587	7430	5655	1530	4256	6836
ut 2	31099	29685	21967	13264	4059	5236	792	4221	4457
ut 3	37569	21114	18340	3207	7757	5932	2579	1704	5931
ut 4	7360	4761	3968	719	1507	1049	497	356	724
ut 5	7406	4894	2794	720	1257	912	396	345	1355
ut 6	4987	2589	1684	311	581	438	193	171	667
ut 7	12074	8651	4337	813	1470	1853	626	513	1316
ut 8	13037	19307	4472	1231	1328	1310	1183	686	980
ut 9	30340	19369	14004	2364	5251	3598	1516	1198	963
ut 10	13225	6907	6861	1170	2279	1649	500	480	605
ut 11	19697	12639	8410	1612	3000	2292	1007	887	778
ut 12	11207	5187	4716	832	1755	1691	455	549	2138
ut 13	20674	21559	7912	1796	3405	2896	1980	1048	8651
ut 14	11551	6653	3959	897	1875	1177	599	446	2773
ut 15	64647	62094	53556	18670	12978	6186	1814	3354	2343

Matriz PARA: Matriz básica de trasvase de votos entre cada opción electoral de la Elección 1 (OE1) y cada opción electoral de la Elección 2 (OE2).

	OE2 1	OE2 2	OE2 3	OE2 4	OE2 5	OE2 6	OE2 7	OE2 8	OE2 9	OE2 10	OE2 11
OE1 1	0.7070	0.0000	0.0000	0.1953	0.0000	0.0092	0.0000	0.0014	0.0025	0.0222	0.0624
OE1 2	0.1368	0.6527	0.0000	0.0000	0.0020	0.1362	0.0000	0.0000	0.0024	0.0075	0.0624
OE1 3	0.2003	0.0000	0.6556	0.0334	0.0000	0.0000	0.0322	0.0150	0.0000	0.0011	0.0624
OE1 4	0.0000	0.0916	0.1971	0.0000	0.6490	0.0000	0.0000	0.0000	0.0000	0.0000	0.0624
OE1 5	0.0000	0.0000	0.0000	0.5671	0.0000	0.0000	0.3705	0.0000	0.0000	0.0000	0.0624
OE1 6	0.0000	0.0000	0.0000	0.3693	0.0000	0.0000	0.0000	0.5465	0.0000	0.0219	0.0624
OE1 7	0.0000	0.0000	0.0000	0.0000	0.0000	0.7524	0.0000	0.0000	0.1852	0.0000	0.0624
OE1 8	0.4451	0.0000	0.0000	0.0000	0.2317	0.0000	0.0000	0.0293	0.0000	0.2315	0.0624
OE1 9	0.1753	0.0000	0.0000	0.3799	0.0000	0.2606	0.0803	0.0255	0.0303	0.0481	0.0000

