UNIVERSIDAD CARLOS III DE MADRID

Procedimiento no paramétrico para la estimación del voto en España

TRABAJO DE FIN DE GRADO

Autora: Julieta Basanta Barnuevo

Tutor: Andrés M. Alonso Fernández

Grado en Estadística y Empresa

Curso 2016/ 2017- Convocatoria: Julio

Estudia el pasado si quieres pronosticar el futuro.

Confucio

TABLA DE CONTENIDO

TABLA DE FIGURAS

1 INTRODUCCIÓN

Los pronósticos electorales son controvertidos. En ocasiones nos permiten conocer con determinado grado de certeza cuál va a ser el resultado de unas elecciones con antelación, sin necesidad de realizar grandes operaciones sobre los datos de los sondeos. Otras veces parecen errar en la previsión de voto con diferencias más o menos sustanciales, y los receptores del pronóstico se sienten engañados y defraudados por un fallo aparentemente inexplicable, (M. Escobar, J. Rivière, R. Cilleros; 2014; p.35). Un ejemplo de este fenómeno lo encontramos en los sondeos previos a las elecciones en Andalucía de 2012, donde se predijo una victoria del Partido Popular (con 42% del voto) y finalmente su porcentaje de votos apenas supera en un punto al PSOE, que con el apoyo de IU obtiene la posibilidad de gobernar (García de Blas; 2012; "¿Por qué fallaron las encuestas?"; El País).

Este estudio plantea una evaluación objetiva de un procedimiento de tipo no paramétrico para la estimación de la intención del voto. Se utilizarán los datos de las encuestas electorales realizadas por el CIS (Centro de Investigaciones Sociológicas) en octubre de 2015 y enero de 2016. Dichas encuestas están dirigidas a la población española de ambos sexos y mayores de 18 años. Comprenden un total de 2493 y 2496 encuestados, respectivamente.

El estudio comenzará con un análisis descriptivo de las muestras en cuestión, con el objetivo de conocer el entorno y estructura de los datos. Dentro de este entorno, se sabe que existen **factores sociológicos** que determinan los resultados de las encuestas, por los que muchas personas se muestran indecisas o imprecisas en las encuestas. Por ello, el estudio se basará en la creencia de que los valores de no respuesta (también llamados datos faltantes) pueden ser una de las principales causas de los errores cometidos en este tipo de estimaciones. En consecuencia, se estudiará con atención y detenimiento la distribución de los datos faltantes existentes en las muestras, su tratamiento y su imputación según el resto de variables de interés.

Otro factor importante en el entorno del estudio es la **situación política** del momento. Sucesos a destacar ordenados cronológicamente:

- Octubre de 2015: realización de encuesta por el CIS.
- Diciembre de 2015: celebración de las elecciones generales. Ningún partido obtiene la mayoría absoluta.
- Enero de 2016: realización de encuesta por el CIS.

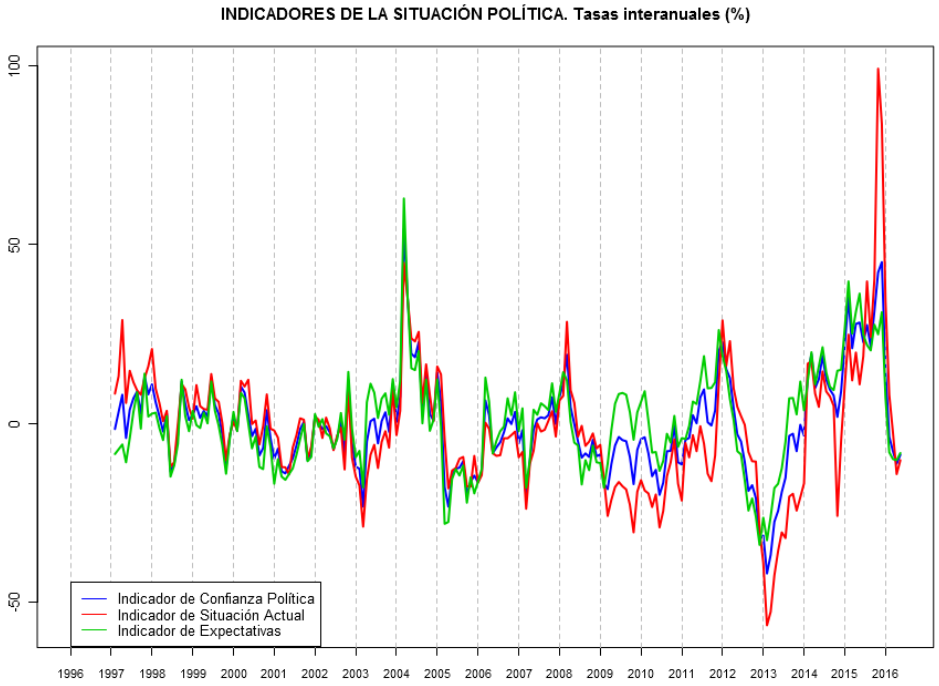


Figura 1: Tasa interanual de los indicadores de situación política. Fuente: CIS

A la vista de la *Figura 1,* se interpreta el aumento de variabilidad en los últimos años en los indicadores de confianza, de la situación actual y de expectativas, como un síntoma de incertidumbre en la situación política. Por lo que se intuye que se encontrará mayor variabilidad en los datos muestrales.

Durante este estudio se prestará especial atención a esta situación actual de incertidumbre, dado que podría contribuir a la obtención de un mayor porcentaje de error en los resultados.

Los objetivos que han presidido este estudio son: en primer lugar, el conocimiento de la población de estudio, ¿cómo son los votantes y cómo se definen los partidos?; en segundo lugar, la imputación de los datos faltantes con el procedimiento no paramétrico, con la correspondiente explicación del método; en tercer lugar, la revisión y evaluación del método empleado, ¿cuál será el grado de fiabilidad de las predicciones obtenidas?; y por último, una interpretación coherente de los resultados donde se planteará si es posible plasmar la incertidumbre existente en la predicción obtenida.

2 ANÁLISIS DESCRIPTIVO

Este estudio está basado en los resultados de entrevistas personales, que realizó el CIS *(Centro de Investigaciones Sociológicas)* en octubre de 2015 y enero de 2016. Ambos cuestionarios están dirigidos a la población española a partir de la mayoría de edad, comprendiendo 256 municipios y 50 provincias de España.

Para una simplificación del estudio se trabajará bajo una recodificación de las posibles respuestas al cuestionario que se detalla en el Anexo 1.

2.1 ANÁLISIS UNIVARIANTE

La principal variable de interés corresponde a la pregunta de *Intención de voto:*

- ***Suponiendo que mañana se celebrasen elecciones generales, ¿a qué partido votaría Ud.?***

Las posibles respuestas consideradas para esta pregunta son las siguientes:

1 Si el encuestado tuviera intención de votar al Partido Popular (PP).

2 Si su intención fuera votar al Partido Socialista Obrero Español (PSOE).

3 Si su intención fuera votar a Podemos.

4 Si la intención del encuestado fuese votar a Ciudadanos.

5 Si su intención fuera votar a Izquierda Unida (IU).

6 Si la intención del encuestado fuese votar a otro partido que no se encuentre en las opciones anteriores.

7 Si el encuestado tuviese intención de votar en blanco.

77 Si el voto resultase nulo.

97 Si el encuestado tuviera la intención de abstenerse.

Utilizamos la codificación del CIS para definir los valores faltantes. Tomaremos como valores faltantes si el encuestado seleccionara la respuesta de *No lo sabe todavía* (98) o *No contesta* (99).

El objetivo del estudio será la imputación de estos valores faltantes entre las siete primeras opciones (voto válido), a partir de un procedimiento no paramétrico. Notar que la coalición Unidos Podemos se formalizó en mayo de 2016 y por tanto no aparece en las encuestas analizadas.

A continuación, se representan los datos muestrales correspondientes a la pregunta de intención de voto:

Intención de voto	PP	PSOE	Podemos	Ciudadanos	IU	Otros Partidos	En blanco	Abstención	N.S	N.C
oct-15	15,0%	16,6%	8,8%	11,0%	3,1%	7,5%	3,0%	9,7%	22,2%	2,7%
ene-16	18,2%	14,5%	17,5%	8,3%	3,3%	7,9%	2,3%	10,1%	14,4%	3,2%

Tabla 1: Porcentajes muestrales de intención de voto

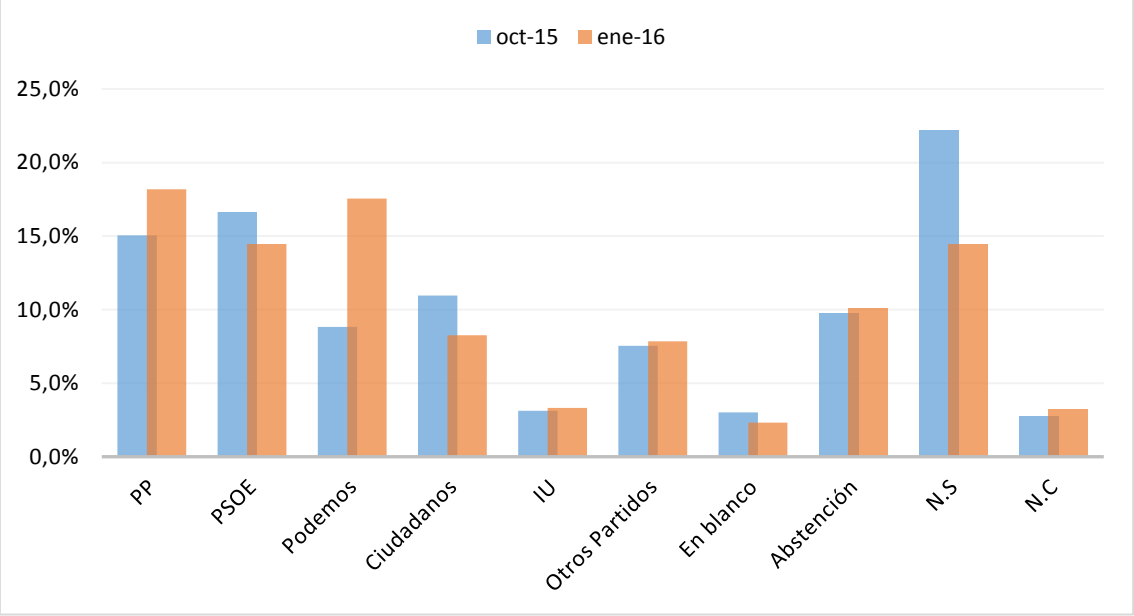


Figura 2: Porcentajes muestrales de intención de voto

En la *Figura 2*, se aprecia que el porcentaje de valores faltantes, en la pregunta de intención de voto, en octubre de 2015 es más elevado que en enero de 2016. En octubre abarcan a 24.9% de los encuestados, mientras que en la muestra de enero descienden al 17.7%.

Ordenando las frecuencias relativas para cada partido, se obtendrían los siguientes resultados muestrales, para los primeros tres partidos:

En octubre de 2015: en tercer lugar, se encontraría Ciudadanos con un 10.95%, seguido por PP con un 15.04% y, en primer lugar, PSOE con 16.65%.

En enero de 2016: en tercer lugar, estaría PSOE con un 14.46%, en segundo lugar, Podemos con 17.55%, y en cabeza de clasificación PP con 18.19%.

Es importante notar que en el periodo de tiempo previo a la encuesta de enero los partidos políticos no establecieron acuerdos para la formación de un gobierno. Éste es un hecho que podría justificar el cambio en la intención de voto de los encuestados de octubre de 2015 a enero de 2016. En este cambio, es destacable un aumento del 9% en la intención de voto a Podemos y un descenso del 8% en los encuestados que no saben a quién votar todavía.

También es importante tener en cuenta la existencia de un posible sesgo, en parte debido a que muchos encuestados prefieren no expresar directamente sus preferencias y expectativas de voto, o bien, prefieren dejar el mayor tiempo posible hasta la elección para ponderar los pros y los contras de distintas opciones políticas. Es por ello que, en la mayor parte de los casos, por defecto, el porcentaje muestral es más bajo que el porcentaje de votos finalmente emitidos. Habiendo alguna excepción: en los partidos de la izquierda menos moderados, es decir, aquellos en los que no existe oscilación del voto entre izquierda y derecha, es más probable que el voto esté decidido antes y que su expresión será más proactiva y explícita, lo cual nos podría conducir más adelante a su sobreestimación, por ejemplo, en Podemos. (M. Escobar, J. Rivière, R. Cilleros; 2014; p.35).

Es razonable la idea de que los pronósticos electorales se basen en un conocimiento exhaustivo de los factores que determinan el voto. Para una correcta estimación del voto, se realiza una selección, en cada cuestionario, de las variables más determinantes a la hora de la elección del voto o que podrían estar relacionadas con la intención de voto. Seleccionando las preguntas más significativamente influyentes en la intención de voto se puede aproximar un perfil de "identidad política" para el encuestado. Estas son las variables que utilizaremos para la imputación del voto en un partido u otro:

Octubre de 2015

- Evaluación situación económica
- Evaluación situación política
- Evaluación gobierno *(PP)**
- Evaluación oposición *(PSOE)**
- Participación en las elecciones de 2011
- Recuerdo de voto en las elecciones de 2011
- Simpatía hacia los partidos políticos
- Ideología
- Edad

Enero de 2016

- Evaluación situación económica
- Evaluación situación política
- Participación en las elecciones de 2015
- Recuerdo de voto en las elecciones de 2015
- Simpatía hacia los partidos políticos
- Ideología
- Edad

*Pregunta que solamente aparece en el cuestionario de octubre de 2015 por la inexistencia de gobierno definitivo en enero de 2016.

A partir de este momento, se trabajará con las variables que aparecen en la tabla anterior, que aportan la información necesaria para crear diferentes perfiles de votantes. El estudio de estos perfiles ayudará a la definición de patrones de respuesta existentes en nuestra selección, los cuales serán claves en el proceso de evaluación interna de la imputación.

A continuación, se expone un análisis descriptivo de las variables que componen la selección anterior.

- ***¿Cómo calificaría Ud. la situación económica general de España: muy buena, buena, regular, mala o muy mala?***

  El valor modal se encuentra en una *mala calificación* de la situación económica tanto en octubre de 2015 (38.5%) como en enero de 2016 (42.1%). El segundo valor más frecuente es *situación económica regular* (31.7% en octubre; 33.9% en enero), seguido por *muy mala* (26.1% en octubre; 16.8% en enero).

- ***¿Cómo calificaría Ud. la <u>situación política</u> general de España: muy buena, buena, regular, mala o muy mala?***

  Como en la pregunta anterior, la moda se encuentra en una *mala calificación* de la situación política con su máximo también en enero (39%). Seguido por una *muy mala calificación* de la situación política (34.3% en octubre; 30.6% en enero), y en tercer lugar una calificación *regular* (23% en octubre; 23.2% en enero).

- ***¿Cómo calificaría Ud. la gestión que está haciendo el <u>Gobierno del PP</u>: muy buena, buena, regular, mala o muy mala?***

  El valor modal de la calificación del gobierno (PP) está en *muy mala* (30.6%), precedido inmediatamente por una calificación *regular* (29.9%), y *mala* (26.9%).

- ***¿Cómo calificaría la actuación política que está teniendo el <u>PSOE en la oposición</u>: muy buena, buena, regular, mala o muy mala?***

  El valor modal de la calificación de la oposición (PSOE) está en *regular* (36.1%), precedido por *mala* (33.9%), y *muy mala* (17.9%)

- ***¿Por qué partidos siente Ud. más <u>simpatía</u> o cuál considera más cercano a sus propias ideas?***

  El valor modal, tanto en octubre como en enero, está en no sentir simpatía hacia ningún partido político, alcanzando un máximo, en octubre, del 25%. Esto podría entenderse como un reflejo del descontento de la población y por bajo nivel de identificación con la clase política actual.

  El segundo valor con mayor frecuencia es, en octubre, el PSOE (19.33%) y, en enero, el PP (17.79%).

  En una visión general, de octubre de 2015 a enero de 2016 aumenta el nivel de simpatía hacia todos los partidos menos hacia el PSOE, que disminuye cerca de un 2%. Los comentarios anteriores se obtienen de la Figura 3.

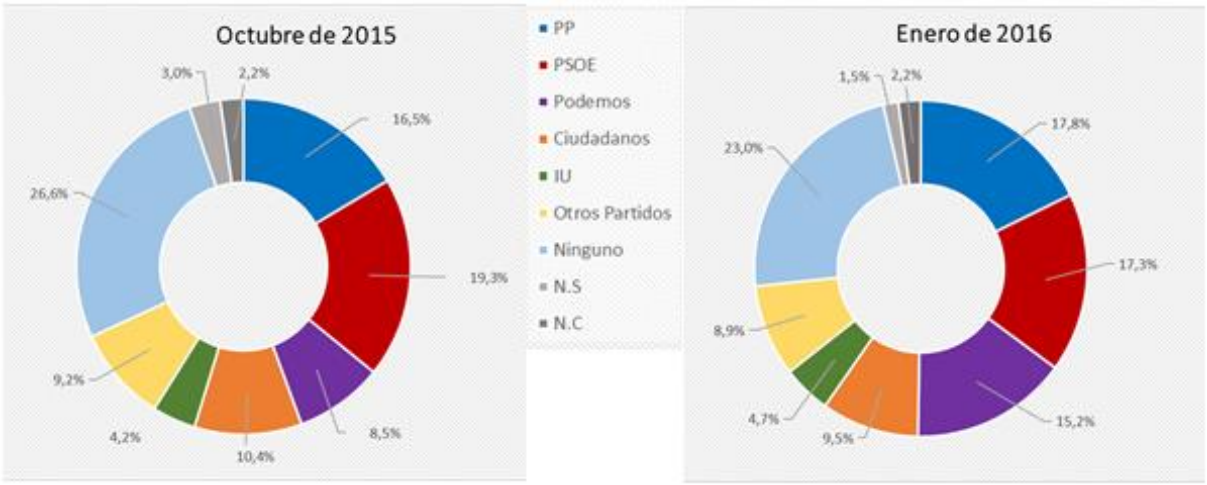


Figura 3: Diagrama de sectores de la simpatía por los partidos políticos

❖ ***Recuerdo de voto*: ¿a qué partido o coalición votó en las últimas elecciones?**

La recodificación de esta pregunta se hace a partir de la pregunta de ***participación en las últimas elecciones*** y de la pregunta de recuerdo de voto. De forma que obtenemos un porcentaje de personas que no pudieron votar por causas como: minoría de edad, o no tener posibilidad por algún motivo. A su vez, obtenemos un porcentaje de abstención que refleja las personas que en la pregunta de participación respondieron a *No quiso votar*.

A continuación, se realiza una comparación del recuerdo de voto en las elecciones de diciembre de 2015 (porcentaje observado sobre el voto válido) con el resultado electoral en las mismas elecciones.

	Recuerdo voto ene-16	Resultado dic-15	Variación (%)
PP	25,2%	**28,7%**	**-3,53%**
PSOE	21,7%	**22,0%**	-0,31%
Podemos	11,6%	**17,0%**	**-5,36%**
Ciudadanos	20,1%	**13,9%**	**6,14%**
IU	4,9%	**3,7%**	1,18%
Otros Partidos	14,2%	**13,1%**	1,13%
En blanco	2,4%	**1,7%**	0,75%

Tabla 2: Comparación resultado electoral y recuerdo de voto

Resultados electorales de 2011 y 2015: Anexo 2 y 3

La variación entre el resultado electoral de diciembre de 2015 y el recuerdo de voto sobre las mismas elecciones es elevada sobre todo en Ciudadanos, del resultado real al recuerdo muestral aumenta un 6.1% (colocando al partido en tercera posición); y Podemos, del resultado real al recuerdo muestral disminuye un 5.4%; y seguidamente, PP que del resultado real al recuerdo muestral disminuye un 3.5%.

- ***<u>Ideología</u>: Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Siendo 1: izquierda y 10: derecha. ¿En qué casilla se colocaría Ud.?***

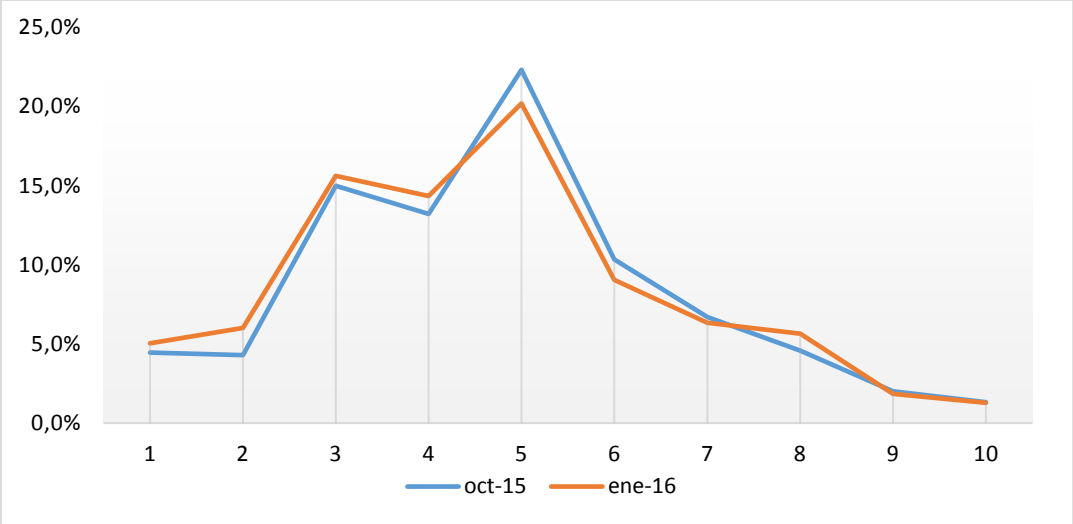


Figura 4: Auto ubicación ideológica muestral

La Figura 4 muestra los resultados obtenidos para esta pregunta en las dos encuestas consideradas.

Se sabe que cerca del 70% de los españoles se ubican entre centro-izquierda y centro-derecha.[1]

Apenas se aprecian cambios entre octubre y enero. Cerca del 75% de los encuestados se sitúan entre los valores 1 y 6, en centro e izquierda y el 25% restante se sitúa a la derecha.

El valor modal siempre está en el 5, es decir, que la frecuencia más alta de encuestados está en los que se sitúan ideológicamente en el centro.

1: Fuente artículo: Guisado, Ley (2015). *Nuevos Partidos, misma ideología* [El Mundo; 2015; URL: http://www.elmundo.es/grafico/espana/2015/12/20/5669a90346163f741f8b4587.html]

2.2 ANÁLISIS MULTIVARIANTE

2.2.1 Clúster bietápico: Perfil del votante

El objetivo del análisis clúster es clasificar a los individuos en grupos que son homogéneos internamente, pero diferentes entre sí. Se realiza este análisis para obtener información sobre las personas que componen cada muestra y así poder definir cada perfil de identidad política observado.

Generalmente, el análisis clúster está dirigido a variables cuantitativas, es por eso que se utilizará el clúster bietápico, o en dos etapas, que permite trabajar conjuntamente con variables cuantitativas y categóricas. Para la realización de éste análisis no se han tenido en cuenta los individuos con valores faltantes (*NS, NC*).

Muestra de octubre de 2015:

Las variables utilizadas para la formación de los clústeres finales son: *Intención de voto*, *Simpatía*, *Ideología*, *Evaluación del gobierno* y *Situación Política*. Se obtienen cuatro agrupaciones de calidad suficiente. Dichas agrupaciones están claramente diferenciadas por la intención de voto a los cuatro partidos más votados.

En cada casilla de la tabla se expone el valor modal de cada clúster con su frecuencia relativa al propio clúster entre paréntesis. Excepto en la variable ideología, al tratarse como cuantitativa se expone el promedio. [Aplicable también a la *Tabla 4*]

Agrupaciones. Octubre 2015				
Clúster	**3**	**2**	**4**	**1**
Tamaño	32,6%	25,7%	22,8%	18,9%
Intención de voto	Podemos (40,9%)	Ciudadanos (59,8%)	PSOE (96,6%)	PP (99%)
Simpatía	Podemos (34%)	Ciudadanos (52,7%)	PSOE (93,4%)	PP (99%)
Ideología (promedio)	3,37	5,35	3,59	7,25
Gobierno	Muy mala (63,1%)	Regular (47,6%)	Muy mala (42,7%)	Buena (49,7%)
Situación política	Muy mala (55,7%)	Mala (41,7%)	Mala (44,7%)	Regular (50,7%)

Tabla 3: Clúster bietápico. Octubre 2015

Información adicional de clústeres, octubre 2015: Anexo 4

Muestra de enero de 2016:

Las variables utilizadas para la formación de los clústeres finales son: *Ideología*, *Intención de voto, Simpatía*, *Situación Económica* y *Situación Política*. Se obtienen cinco agrupaciones de calidad suficiente. Dichas agrupaciones están claramente diferenciadas por la ideología del encuestado y su intención de voto. Con respecto a octubre, se ha creado una nueva agrupación que abarca el 17% de la muestra y su mayor intención de voto es a otros partidos.

Notar que todos los clústeres tienen la moda de situación política en *Mala.*

Agrupaciones. Enero 2016					
Clúster	**2**	**1**	**3**	**5**	**4**
Tamaño	24,5%	20,7%	19,0%	18,9%	17,0%
Ideología (promedio)	5,13	7,29	3,75	2,96	3,28
Intención de voto	Ciudadanos (39,6%)	PP (96,6%)	PSOE (94,4%)	Podemos (100%)	Otros partidos (54,7%)
Simpatía	Ninguno (52,5%)	PP (100%)	PSOE (98,1%)	Podemos (100%)	Otros partidos (58,5%)
Situación económica	Mala (44,6%)	Regular (54,0%)	Mala (48,0%)	Mala (48,4%)	Mala (55,7%)
Situación política	Mala (45,6%)	Mala (39,2%)	Mala (40,2%)	Mala (38,8%)	Mala (47,1%)

Tabla 4: Clúster bietápico. Enero 2016

Información adicional de clústeres, enero 2016: Anexo 5

2.2.2 Correspondencias múltiples: Posicionamiento de los partidos

El análisis de correspondencias múltiples es una técnica descriptiva o exploratoria aplicable a variables cuantitativas y categóricas. Se utilizará el gráfico conjunto de puntos de categoría (*Figuras 4 y 5*) para una representación bidimensional de nuestros datos que permitirá conocer el posicionamiento de los partidos políticos en función de las opiniones de los individuos encuestados.

Notar que en el gráfico conjunto de puntos de categoría cuanto menor sea la distancia entre las categorías y los ejes X e Y, menor será la información explicada por las mismas. Aquellas categorías que se encuentren cercanas al punto (0,0) no contarán con una interpretación satisfactoria. Como en el análisis clúster, en este análisis no se han tenido en cuenta los individuos con valores faltantes (NS, NC).

Muestra de octubre de 2015:

La selección de variables sobre la cual se obtiene un 94.3% de la variabilidad explicada de la muestra es la siguiente: Intención de voto, Ideología, Evaluación del gobierno, Situación Económica y Situación Política.

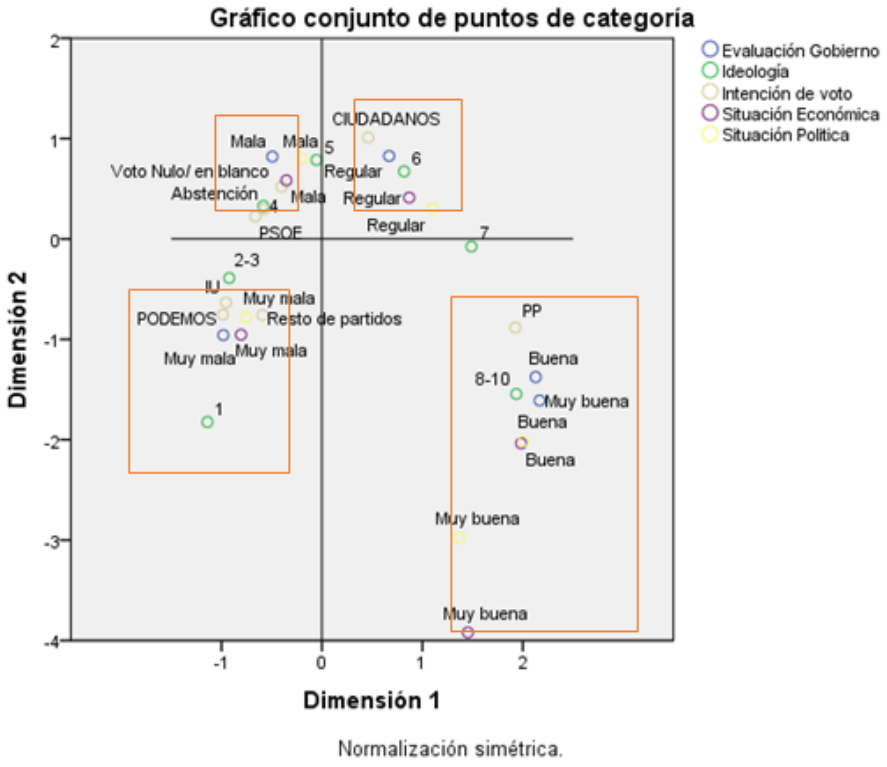


Figura 5: Correspondencias múltiples. Octubre 2015

Se obtienen cuatro agrupaciones destacables:

- Esquina inferior derecha del gráfico: los encuestados tienen una buena visión general de la situación económica y política e ideológicamente se ubican en torno 8-10 (en una escala del 1 al 10; siendo 1 izquierda y 10 derecha). Aquí se encuentra posicionado el PP.
- Esquina inferior izquierda del gráfico: los encuestados tienen una visión muy mala de la situación económica y política e ideológicamente se ubican en torno 1-3. Aquí se encuentra posicionado tanto Podemos como Izquierda Unida.

- Esquina superior derecha del gráfico: los encuestados tienen una visión regular del entorno económico y político e ideológicamente se ubican en torno al 6. Aquí se encuentra posicionado Ciudadanos.

- Esquina superior derecha del gráfico: los encuestados tienen una visión mala del entorno político y económico, y no tienen intención de votar a ningún partido.

Muestra de enero 2016:

La selección de variables de la muestra es la siguiente: *Intención de voto, Ideología, Situación Económica* y *Situación Política.* El porcentaje de variabilidad explicada obtenido es el 91.6%.

Se obtienen cuatro agrupaciones destacables bastante similares a las obtenidas con la muestra de octubre. La principal diferencia entre las muestras es que en enero el PSOE pertenece a la agrupación que se localiza en la esquina superior izquierda, cerca de los individuos que no quieren votar a ningún partido.

Notar que la respuesta "Muy buena" en las preguntas de *Situación Económica y Situación Política* ha sido considerada como dato atípico para la realización de éste análisis. Con una frecuencia relativa del 0.003%.

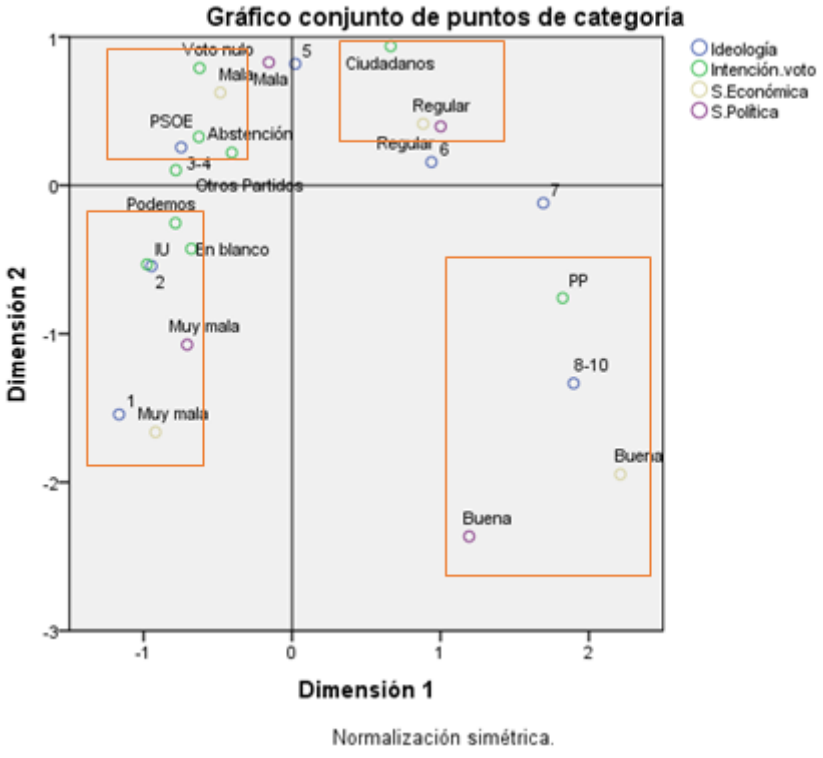


Figura 6: Correspondencias múltiples. Enero 2016

Información adicional análisis de correspondencias múltiples: Anexo 6 y 7

3 DATOS FALTANTES

3.1 TIPOS DE DATOS FALTANTES

Para decidir cómo manejar los valores perdidos, es conveniente conocer las características que los definen. Se consideran dos tipos de datos faltantes: aquellos que se consideran ignorables y aquellos que no.

3.1.1 Ignorables

En referencia a las muestras del CIS, corresponden a todos los posibles valores de una pregunta que no corresponde responder al encuestado. En una primera codificación, realizada por el CIS, son registrados como NA. Aparecen en todas aquellas preguntas/ variables que derivan de una anterior y, por tanto, dependen de ella.

Por ejemplo, si a la pregunta sobre participación en las elecciones anteriores el encuestado responde que no pudo votar porque no alcanzaba la mayoría de edad, el encuestado no debe contestar a la pregunta derivada sobre el recuerdo de voto.

También son denominados **Missing at Random (MAR)** debido a que, una observación es faltante con aleatoriedad si la probabilidad de pérdida del valor depende de información de la propia muestra, pero no de la propia variable. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos.

$$Pr\ (Y \text{ sea faltante } |X, Y) = Pr\ (Y \text{ sea faltante } |X)$$

Podemos testar si valores faltantes en Y dependen de X. Pero no podemos testar si valores faltantes en Y dependen de Y.

3.1.2 No ignorables

En referencia a las muestras del CIS son aquellos que corresponden a las respuestas *No sabe* o *No contesta* (comúnmente llamados “indecisos”). Tras la recodificación realizada en este estudio, estos valores se traducen en NA. Serán el objeto a estimar en el proceso de imputación.

Se pueden comportar como:

- **Missing Completely at Random (MCAR):**

Una observación es faltante con completa aleatoriedad si la probabilidad de ser faltante es la misma para todas las observaciones. Es decir, la ausencia de la información no está originada por ninguna variable de la muestra.

Siendo *Y* la matriz de datos y *X* el vector de variables observadas:

$$Pr(Y \text{ sea faltante} \mid X, Y) = Pr(Y \text{ sea faltante})$$

- **Missing Not at Random (MNAR):**

Los valores pueden ser faltantes sin aleatoriedad cuando la probabilidad de pérdida depende de la variable que es faltante. El mecanismo de pérdida es no ignorable.

Si el mecanismo de no respuesta depende del verdadero valor del dato perdido o de variables no observables.

3.2 ANÁLISIS DE LOS DATOS FALTANTES

Se analizarán los datos faltantes no ignorables de la muestra. Si se denominan no ignorables, es debido a que es importante tener en cuenta la forma en la que están distribuidos. Se debe determinar si los valores perdidos de la muestra están distribuidos de forma completamente aleatoria, o no.

Para conocer la distribución de los valores perdidos, lo primero que se realiza es un estudio de los patrones de respuesta, determinando si el valor observado es faltante o no.

En las *Figuras 6 y 8* se expone una representación de los distintos patrones de valor perdido. Cada patrón representa una distribución de valores perdidos en la encuesta, con una frecuencia de una o más personas.

Las *Figuras 7 y 9* representan un gráfico de barras de la proporción de observaciones que abarca cada uno de los diez patrones que aparecen con más frecuencia en cada muestra.

Muestra de octubre de 2015:

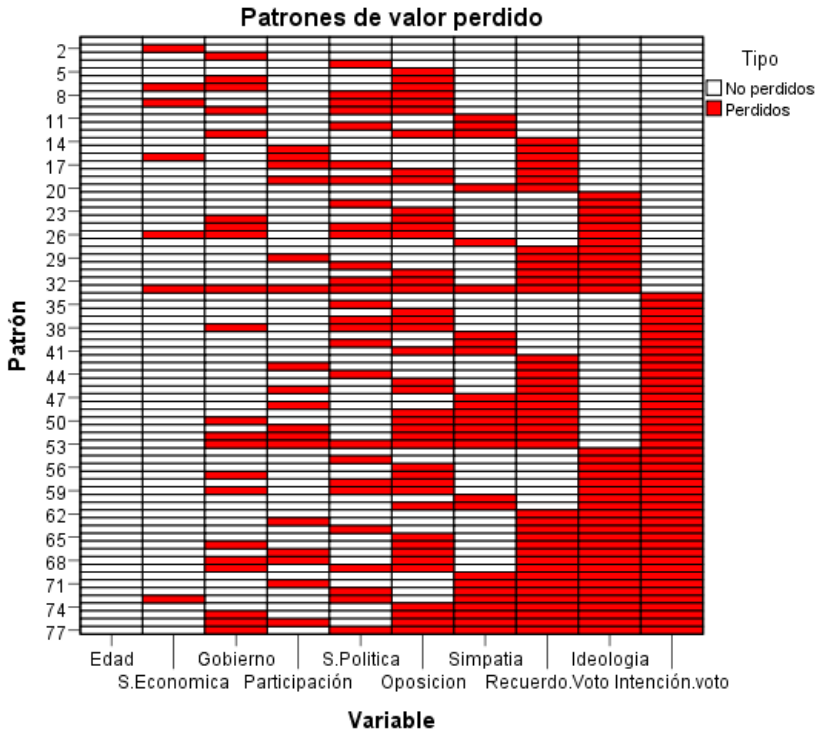


Figura 7: Patrones de valor faltante. Octubre 2015

En la muestra de octubre de 2015 se definen 77 patrones posibles de respuesta. Los valores perdidos de nuestra selección de interés siguen una distribución determinada en la que las variables que hacen referencia al voto o a la "identidad política" del encuestado son las que mayor porcentaje de valores perdidos acumulan. El caso contrario se encuentra en la variable edad, en la que contamos con la respuesta del 100% de la muestra.

En la *Figura 7* se muestran los diez patrones con mayor frecuencia relativa en la muestra. El patrón número 1, patrón con mayor frecuencia (58.32%), corresponde a todas aquellas personas encuestadas que han respondido al 100% de la encuesta. Forman un subconjunto del que poseemos toda la información posible. El siguiente patrón con mayor frecuencia relativa (11.83%) es el 34, representa a todas aquellas personas que únicamente han dejado de contestar a la pregunta de *Intención de voto*. Seguido por el patrón número 21 que representa a los individuos que no han contestado a la pregunta de *Ideología*, con una frecuencia relativa del 6.62% de los encuestados.

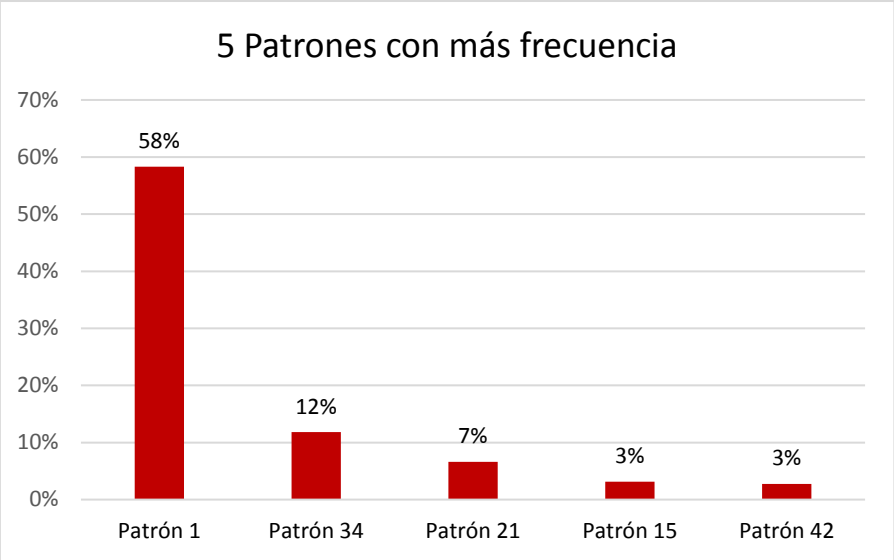


Figura 8: Gráfico de barras patrones de valor faltante. Octubre 2015

Muestra de enero de 2016:

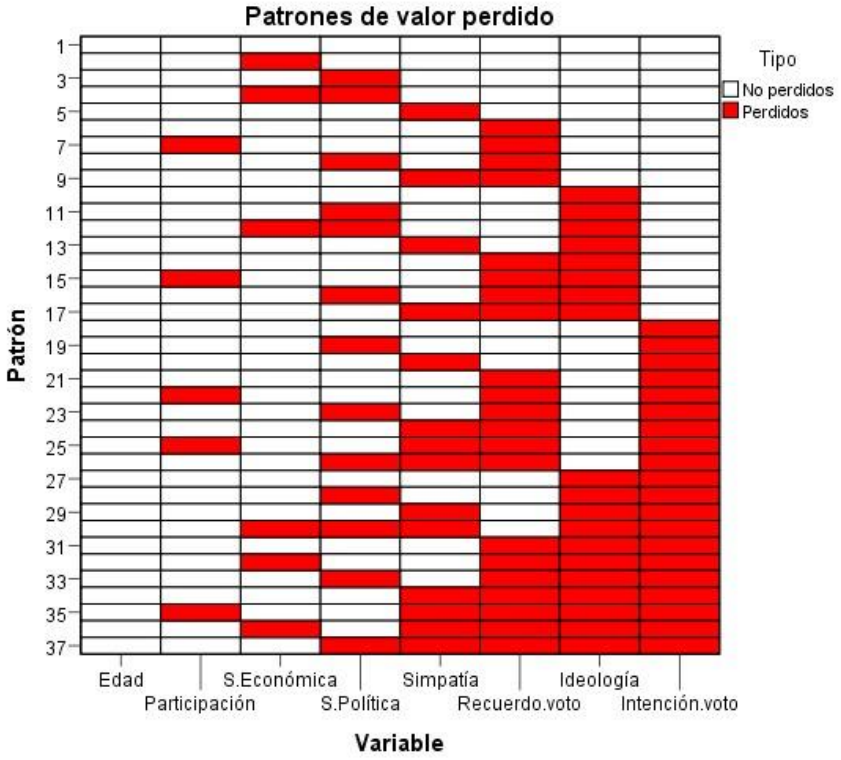


Figura 9: Patrones de valor faltante. Enero 2016

En la muestra de enero de 2016 se definen 37 patrones posibles de respuesta. El cambio más significativo en el orden de las variables entre octubre y enero se encuentra en la participación, probablemente porque en octubre de 2015 no todos los encuestados

recuerdan con claridad las elecciones de 2011. En cambio, en enero de 2016 los encuestados tienen un recuerdo reciente de las elecciones de diciembre de 2015.

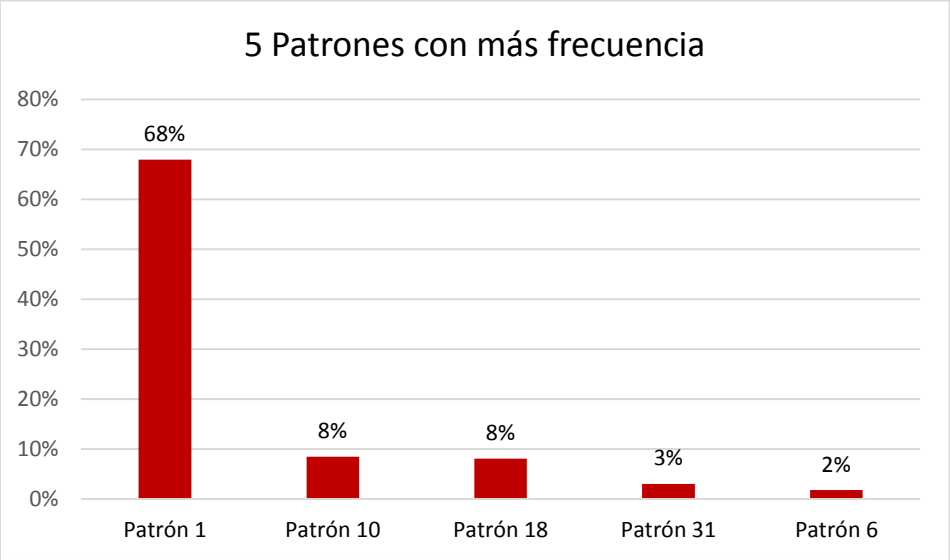


Figura 10: Gráfico de barras. Patrones de valor faltante. Enero 2016

En la muestra de enero, el patrón número 1 abarca al 68% de las personas encuestadas. El siguiente patrón de respuesta con mayor frecuencia relativa (8.49%) es el 10 que representa a aquellas personas que han contestado a todo, excepto únicamente, a la pregunta de ideología. El patrón 18 con un 8.09% de la muestra representa a aquellas personas que solo han dejado de contestar a la pregunta de intención de voto. El siguiente es el patrón 31, que representa a los individuos que no han contestado a las preguntas de recuerdo e intención de voto. Y, en quinto lugar, el patrón 6 en representación de los individuos que no han contestado solamente a la pregunta de *Recuerdo de voto.*

Información detallada sobre la distribución de valores faltantes: Anexo 8

¿Son aleatorios los datos faltantes?

Para clasificar los valores faltantes muestrales como MNAR o MCAR, se debe contrastar si estos se distribuyen con aleatoriedad o no.

Para la realización del contraste de independencia se utilizará la prueba de chi-cuadrado. Esta prueba permite determinar si dos variables cualitativas están o no asociadas. Si al final

del estudio concluimos que las variables no están relacionadas podremos decir con un determinado nivel de confianza, previamente fijado, que ambas son independientes.

Se definen las siguientes hipótesis:

Hipótesis nula (H_0): No hay asociación entre las variables.

Hipótesis alternativa (H_1): Sí hay asociación entre las variables.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde:

O_{ij} denota a las frecuencias observadas. Es el número de casos observados clasificados en la fila i y la columna j.

E_{ij} denota a las frecuencias esperadas o teóricas. Es el número de casos esperados correspondientes a la fila i y columna j.

Contraste para octubre 2015:

Oct-15	N	Media	Desviación estándar	Perdidos	
				Recuento	Porcentaje
Ideología	**2101**	**4,74**	**1,952**	**392**	**15,7**
Edad	2493	49,78	18,039	0	,0
S.Economica	2484			9	,4
S.Politica	2420			73	2,9
Gobierno	2450			43	1,7
Oposicion	2366			127	5,1
Intención.voto	1871			622	**24,9**
Simpatía	2362			131	5,3
Participación	2439			54	2,2
Recuerdo.Voto	**2155**			**338**	**13,6**

Tabla 5: Estadísticos univariados. Contraste aleatoriedad de faltantes. Octubre 2015

oct-15		Categoría intención de voto		
	X\Y	Si contesta	No contesta	Total
Resto de categorías	Si contesta a todo	1446	299	1745
	No contesta a todo	425	323	748
	Total	1871	622	2493

Tabla 6: Tabla de contingencia para contraste Chi-cuadrado. Octubre 2015

Al ser la Tabla 6 una tabla 2x2 se puede utilizar la siguiente expresión para calcular el valor del estadístico:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{2493\ x\ (1446x\ 323 - 299x425)^2}{(1446+299)x(425+323)x(1446+425)x(299+323)}$$

$$\chi^2 = 189.6$$

Como el valor obtenido del estadístico es mayor que 3.84 (valor crítico de una χ^2 con un grado de libertad para un nivel α=0.05) se rechaza la hipótesis nula, es decir, tenemos evidencias de que las dos variables no son independientes.

Contraste para enero de 2016:

Ene-16	N	Media	Desviación estándar	Perdidos	
				Recuento	Porcentaje
Ideología	2131	4,63	2,016	365	14,6
Edad	2496	49,72	17,949	0	,0
S.Económica	2486			10	,4
S.Política	2412			84	3,4
Intención.voto	2055			441	**17,7**
Simpatía	2405			91	3,6
Participación	2488			8	,3
Recuerdo.voto	2256			240	9,6

Tabla 7:Estadísticos univariados. Contraste aleatoriedad de faltantes. Enero 2016

ene-16		Categoría intención de voto		
	X\Y	Si contesta	No contesta	Total
Resto de categorías	Si contesta a todo	1697	200	1897
	No contesta a toto	358	241	599
	Total	2055	441	2496

Tabla 8: Tabla de contingencia para contraste Chi-cuadrado. Enero 2016

$$\chi^2 = \frac{2496\ x\ (1697x\ 241 - 200x358)^2}{(1697 + 200)x(358 + 241)x(1697 + 358)x(200 + 241)}$$

$$\mathrm{X}^2 = 275.78$$

Como en el caso anterior, el valor obtenido del estadístico permite rechazar la hipótesis nula de independencia.

Tanto para la muestra de octubre como para la muestra de enero se rechaza la hipótesis nula de no asociación entre variables. Por tanto, se clasifican los datos faltantes de ambas muestras como **MNAR (*Missing Not At Random)*.**

4 METODOLOGÍA

4.1 MUESTREO Y PROCEDIMIENTO

El procedimiento de muestreo seguido por el CIS es polietápico, estratificado por conglomerados, con selección de las unidades primarias de muestreo (municipios) y de las unidades secundarias (secciones) de forma aleatoria proporcional, y de las unidades últimas (individuos) por rutas aleatorias y cuotas de sexo y edad.

Los estratos se han formado por el cruce de las 17 comunidades autónomas, con el tamaño de hábitat, dividido en 7 categorías: menor o igual a 2.000 habitantes; de 2.001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000, y más de 1.000.000 de habitantes.

4.2 MÉTODO DE LOS *K* VECINOS MÁS PRÓXIMOS

4.2.1 Procedimiento

El método de los *k* vecinos más próximos (*k*-NN) es un método no paramétrico cuyo objetivo es predecir los valores de Y basándose en alguna similaridad entre las variables X.

La idea básica sobre la que se fundamenta esta metodología es que un individuo del cual no conocemos su valor en la respuesta a Y se clasificará en la clase más frecuente a la que pertenecen sus *k vecinos* más cercanos entendida la vecindad en las variables observadas, X. Para el valor de *k* se tomarán los valores 3 y 5.

En otras palabras, la observación objeto *u* recibirá un valor $\hat{y}_U$ derivado de las *k* observaciones más cercanas, donde la cercanía entre observaciones de referencia y objetivo se mide en el espacio X que tiene *p* dimensiones, donde *p* es el número de variables predictoras. En las observaciones de referencia se conoce tanto el valor de las variables X como de Y.

El procedimiento general *k*-NN puede ser descrito como sigue:

I. Calcular una distancia entre el punto objetivo y los puntos de referencia. *d(u,j); j=1, 2, … J.*

II. Seleccionar el conjunto de los *k* puntos de referencia más cercanos que denotaremos por $\{j_{u1}, j_{u2}, \ldots, j_{uk}\}$

III. Ordenar dichas distancias en orden ascendente, simbolizando el vector ordenado por $d(u) = \{d(u, j_{u1}), d(u, j_{u2}), \ldots, d(u, j_{uk})\}$

IV. Calcular las ponderaciones normalizadas para las primeras *k* distancias en el vector *d(u)*, usando una función de ponderación inversamente proporcional a la raíz cuadrada de la distancia.

V. Predecir los valores para el *u*-ésimo punto objetivo mediante el algoritmo kNN (*k-Nearest Neighbour Imputation*) basado en una variación de la Distancia de Gower, perteneciente a la librería VIM de R (*Visualization and Imputation of Missing Values*).

Código en R del procedimiento de imputación: Anexo 9

4.2.2 Distancias

Existen diferentes variaciones del método de k-vecinos en relación con el criterio a seguir a la hora de calcular las distancias que determinarán qué vecinos están más próximos:

- Distancias para variables cuantitativas: distancia euclídea, Minkowski, de ciudad o Manhattan, dominante y Mahalanobis, entre otras.
- Distancias para variables binarias: coeficientes de similaridad como Sokal y Michener; y Jaccard.
- Distancias para variables categóricas: medidas de disimilaridad como Bhattacharyya (o de Cavalli-Sforza), y la distancia de Balakrishnan-Sanghvi.
- **Distancias para variables mixtas:** en el caso de este estudio, se dispone de un conjunto de variables mixto, de decir, que se han observado tanto variables cuantitativas (como por ejemplo *Edad*), como categóricas (como por ejemplo *Intención de voto*).

  Se define la distancia de Gower como $d^2_{ij} = 1 - S_{ij}$, donde:

$$S_{ij} = \frac{\sum_{h=1}^{p1}\left(1 - \frac{|x_{ih} - x_{jh}|}{G_h}\right) + a + \propto}{p_1 + (p_2 - d) + p_3}$$

Donde:

p_1 es el número de variables cuantitativas continuas,

p_2 es el número de variables binarias,

p_3 es el número de variables cualitativas (no binarias),

a es el número de coincidencias (1,1) en las variables binarias,

d es el número de coincidencias (0,0) en las variables binarias,

α es el número de coincidencias en las variables cualitativas (no binarias) y

G_h es el rango (o recorrido) de la h-ésima variable cuantitativa.

4.2.3 Calibración del procedimiento:

En una primera imputación realizada, se obtuvieron porcentajes de acierto muy bajos. Es por ello que se ha decido asignar diferentes pesos a las variables que determinan nuestra imputación de la intención de voto.

Estos pesos han sido obtenidos mediante un análisis de clúster bietápico, que ofrece un valor de importancia para cada predictor comprendido entre 0 y 1. El valor de importancia comprende la proporción de variabilidad explicada de la muestra por cada variable, siendo 0 no importante y 1 muy importante. Los pesos obtenidos serían los expresados en la Tabla 9.

oct-15	Importancia	Peso	ene-16	Importancia	Peso
Intención de voto	1	16,6%	Intención de voto	1	19,2%
Simpatía	1	16,6%	Simpatía	1	19,2%
Participación	1	16,6%	Participación	1	19,2%
Recuerdo voto	1	16,6%	Recuerdo voto	1	19,2%
Ideología	0,72	12,0%	Ideología	0,94	18,0%
Gobierno	0,59	9,8%	Situación Económica	0,13	2,5%
Edad	0,25	4,2%	Edad	0,11	2,1%
Oposición	0,22	3,7%	Situación Política	0,04	0,8%
Situación Política	0,14	2,3%			
Situación Económica	0,1	1,7%			
Total	6,02	100,0%	Total	5,22	100,0%

Tabla 9: Pesos de variables para imputación mediante kNN.

Información adicional sobre la obtención de pesos: Anexo 10

5 EVALUACIÓN INTERNA DEL PROCEDIMIENTO DE IMPUTACIÓN:

Para evaluar el procedimiento de imputación se hace uso de las tablas de clasificación.

Para la compresión de las tablas de clasificación generadas, es necesario conocer aquello que se está comparando. Por ello, es conveniente el estudio de los pasos previos realizados. Se definen a continuación:

I. Se selecciona una submuestra que no contenga datos faltantes (observaciones correspondientes al patrón número 1).

II. Se generan datos faltantes "ficticios" distribuidos de la misma forma que la muestra original, es decir, cada patrón de respuesta definido en la muestra original debe tener la misma probabilidad de aparecer en la submuestra (de manera proporcional).

III. Se aplica el método de imputación sobre la submuestra con los datos faltantes "ficticios".

IV. Se selecciona una segunda submuestra que contenga, dentro de la variable de intención de voto, los datos faltantes "ficticios" ya imputados. En octubre esta submuestra corresponde a 358 observaciones y en enero corresponde a 309 observaciones.

V. Se compara, mediante una tabulación cruzada, los datos faltantes "ficticios" imputados con sus correspondientes observaciones en la muestra original, generando así la tabla de clasificación.

De esta forma se podrá evaluar el nivel de acierto en la imputación de los valores según el método de los k- vecinos.

Código en R de evaluación interna del proceso: Anexo 11

En las siguientes tablas de validación cruzada, las filas representan las frecuencias de los valores de intención de voto observados en la muestra sin datos faltantes. Las columnas representan los valores esperados una vez aplicado el método de *k*-vecinos (*k*=5 y *k*=3,

respectivamente), a todas aquellas observaciones que previamente se han definido como valores perdidos.

Muestra de octubre de 2015:

Tabla de clasificación. Octubre 2015. k=5										
		Valor imputado								
		PP	PSOE	Podemos	Ciudadanos	IU	Otros partidos	En blanco	Abstención	Total
Valor observado	PP	43	4	8	10	0	2	0	1	68
	PSOE	2	63	10	18	2	6	1	2	104
	Podemos	0	6	19	9	3	7	1	0	45
	Ciudadanos	9	7	10	17	2	2	0	0	47
	IU	0	3	10	1	3	3	0	0	20
	Otros partidos	2	4	10	9	3	11	0	0	39
	En blanco	1	1	1	0	1	2	0	0	6
	Abstención	1	1	7	2	0	4	0	14	29
	Total	58	89	75	66	14	37	2	17	358

Tabla 10: Validación cruzada para imputación k=5. octubre 2015

El porcentaje total de valores clasificados correctamente con k=5 (calculado como la suma de todos los valores de la diagonal principal) es de 47%. La categoría donde se acumula más error es en Izquierda Unida, seguida de Otros Partidos y voto en blanco.

Notamos que tanto PP y PSOE tienen un porcentaje de acierto en la estimación superior al 60% (63% y 61%, respectivamente).

Por otra parte, en la imputación de valores para Izquierda Unida gran porcentaje de los mismos se asignan erróneamente en Podemos. A su vez, para la imputación de los valores de Otros partidos se asignan erróneamente a Podemos y Ciudadanos principalmente.

Tabla de clasificación. Octubre 2015. k=3										
		Valor imputado								
		PP	PSOE	Podemos	Ciudadanos	IU	Otros partidos	En blanco	Abstención	Total
Valor observado	PP	41	7	4	10	0	1	1	0	64
	PSOE	8	55	18	10	0	1	0	2	94
	Podemos	1	12	20	1	5	5	1	5	50
	Ciudadanos	8	5	7	21	3	1	0	1	46
	IU	0	5	6	1	5	2	0	0	19
	Otros partidos	3	5	17	4	6	12	1	1	49
	En blanco	1	4	1	5	0	0	1	1	13
	Abstención	0	6	2	4	0	2	0	9	23
	Total	62	99	75	56	19	24	4	19	358

Tabla 11: Validación cruzada para imputación k=3. octubre 2015

El porcentaje total de acierto en la imputación con k=3 es de un 46%. Los partidos PP, PSOE, Podemos y Ciudadanos tienen un porcentaje de acierto en la imputación superior al 40%. Nuevamente se produce el mismo error para la imputación en Otros partidos donde se imputan valores erróneamente en la categoría de Podemos.

Esta evaluación del proceso es un factor determinante a la hora de la interpretación de los resultados de las estimaciones de octubre. Se deberá prestar atención a posibles sobreestimaciones en Ciudadanos y Podemos.

Muestra de enero de 2016:

Tabla de clasificación. Enero 2016. k=5										
		Valor imputado								
		PP	PSOE	Podemos	Ciudadanos	IU	Otros partidos	En blanco	Abstención	Total
Valor observado	PP	44	9	6	13	0	1	1	1	75
	PSOE	3	32	14	3	0	0	1	1	54
	Podemos	1	8	50	10	0	3	0	0	72
	Ciudadanos	1	4	7	17	0	3	0	2	34
	IU	0	1	8	1	2	1	0	0	13
	Otros partidos	0	2	11	1	1	15	0	0	30
	En blanco	0	0	1	1	0	2	2	3	9
	Abstención	2	0	11	3	1	0	0	5	22
	Total	51	56	108	49	4	25	4	12	309

Tabla 12: Validación cruzada para k=5. enero 2016

En la muestra de enero, con *k*=5 vecinos más cercanos, el porcentaje de acierto total en la imputación es de 54%. El error más considerable se encuentra en la categoría de Izquierda Unida, donde gran parte de los valores se asignan erróneamente en Podemos.

Tabla de clasificación. Enero 2016. k=3										
		Valor imputado								
		PP	PSOE	Podemos	Ciudadanos	IU	Otros partidos	En blanco	Abstención	Total
Valor observado	PP	51	2	3	14	0	3	1	1	75
	PSOE	5	25	15	5	0	1	2	1	54
	Podemos	0	12	49	7	0	4	0	0	72
	Ciudadanos	2	2	5	16	0	5	0	4	34
	IU	0	0	8	1	2	2	0	0	13
	Otros partidos	0	2	11	1	0	16	0	0	30
	En blanco	0	1	1	1	1	2	0	3	9
	Abstención	3	4	5	0	2	3	0	5	22
	Total	61	48	97	45	5	36	3	14	309

Tabla 13: Validación cruzada para imputación k=3. enero 2016

Utilizando *k*=3 vecinos más cercanos, el porcentaje total de acierto en la imputación es de un 53%, y la distribución del error es muy similar que en octubre de 2015. Las categorías PP, PSOE, Podemos y Ciudadanos cuentan con un porcentaje de acierto superior al 45% (68%, 46%, 68% y 47%, respectivamente). Destacan las categorías de voto en blanco e Izquierda

Unida por su error en la imputación, al igual que ha ocurrido en los casos anteriores los valores para Izquierda Unida son imputados erróneamente en Podemos, por lo general.

6 RESULTADOS

Resultados finales obtenidos tras los procesos de imputación y calibración explicados en los apartados de *Metodología (4)* y *Evaluación interna del procedimiento de imputación (5).*

Los resultados obtenidos serán un reflejo de las conclusiones que se han ido anticipando durante todo el estudio.

Para una primera visión general de las estimaciones resultantes, se expone una representación comparativa del promedio del resultado de las imputaciones obtenidas para la muestra de octubre de 2015 y enero de 2016:

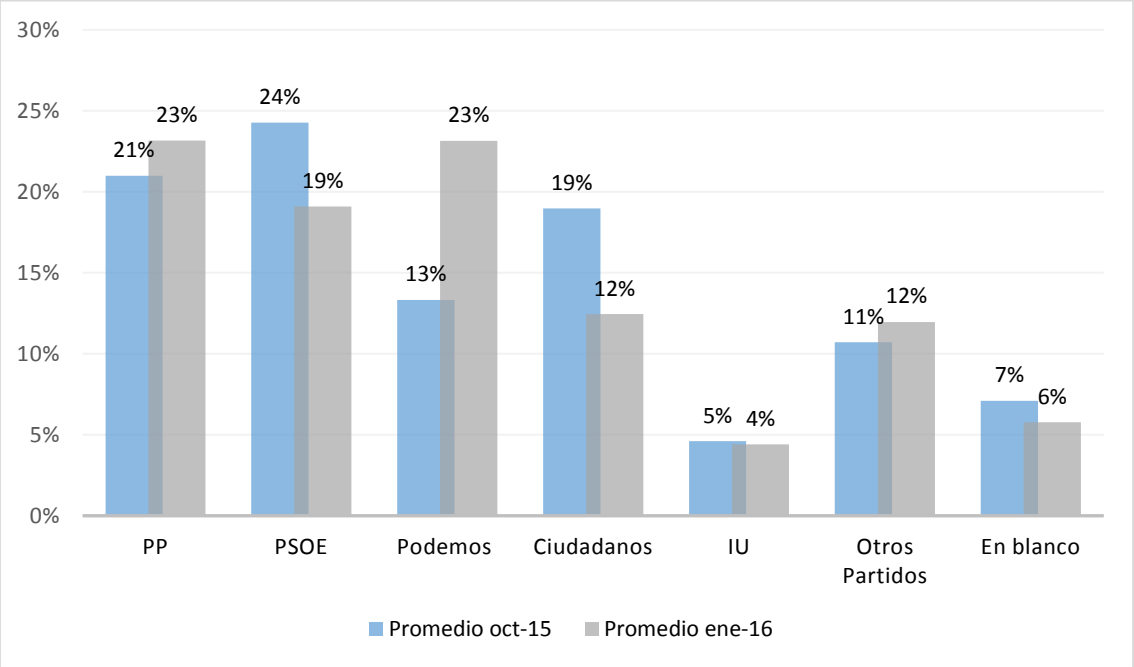


Figura 11: Diagrama de barras. Promedio de estimaciones para octubre de 2015 y enero de 2016

Se aprecia una gran evolución en la intención de voto a Podemos de octubre a enero y, en cambio, una involución tanto en el PSOE como en Ciudadanos. Se podría hablar de un sorpasso de Podemos a PSOE de octubre a enero, puesto que sus porcentajes de estimación para la intención de voto se invierten en la comparación de una encuesta a otra.

Una hipótesis posible que explique éste suceso podría ser que el periodo post-electoral ha sido muy determinante en el perfil del votante. Aunque siempre se debe tener en cuenta que, en enero de 2016, España se encuentra en un periodo de incertidumbre política, y la incertidumbre se traduce en indecisión de los votantes y, por tanto, aumento de la variabilidad de los resultados.

El orden de clasificación final obtenido (en enero) sería: en primer lugar Podemos (22%), en segundo lugar PP(18%), en tercer lugar PSOE (15%) y en cuarto lugar Ciudadanos (10%). Pero, en el caso de que se lograse corregir la probable infraestimación que se está produciendo en el PP sería razonable pensar que el orden de clasificación final cambiaría.

6.1 Comparación con las estimaciones del CIS

A continuación, se exponen gráficos comparativos del promedio de la predicción electoral resultante aplicando el método para *k*=5 y *k*=3, con la estimación realizada por el CIS:

Muestra de octubre de 2015:

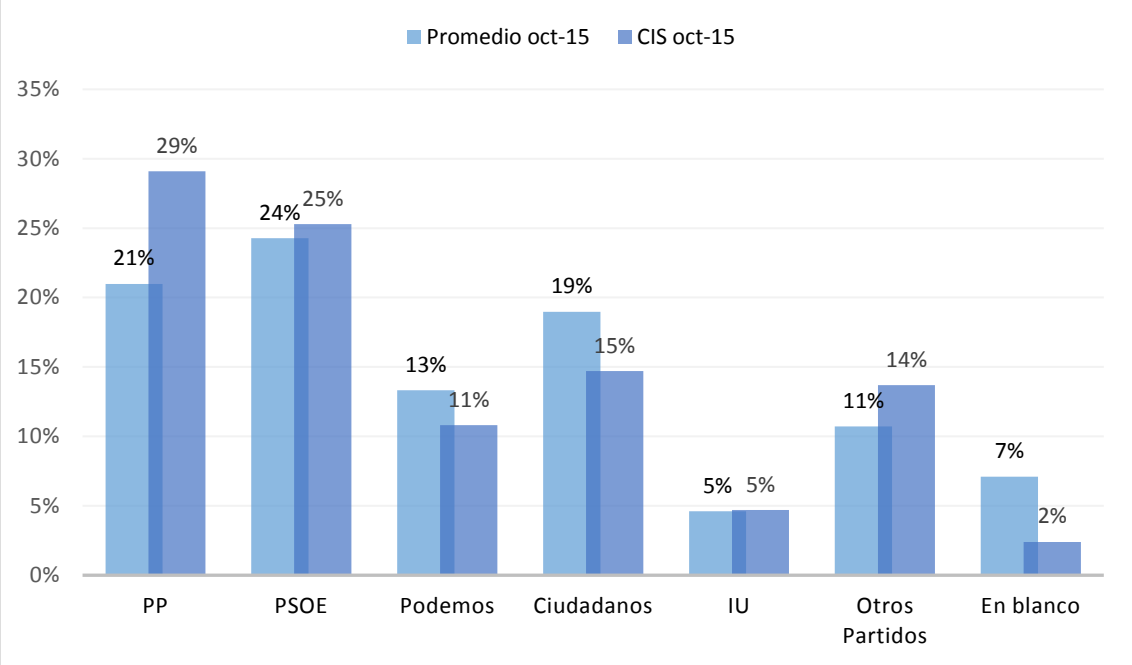


Figura 12: Diagrama de barras: comparación estimaciones octubre y resultados del CIS en octubre

En rasgos generales, las mayores diferencias con respecto a las estimaciones del CIS se encuentran en las categorías de *PP*, *PSOE*, *Otros Partidos* y *En blanco*. El resto de categorías se asemejan considerablemente a las estimaciones del CIS. A continuación se expone una tabla de frecuencias para los procesos de imputación realizados en la muestra de octubre de 2015.

	k=5		k=3		
oct-15	**Frecuencia estimada**	**% estimado sobre el voto válido**	**Frecuencia estimada2**	**% estimado sobre el voto válido2**	**Media de imputaciones**
PP	433	20,6%	447	21,3%	21,0%
PSOE	511	24,4%	507	24,2%	24,3%
Ciudadanos	414	19,7%	382	18,2%	19,0%
Podemos	276	13,2%	283	13,5%	13,3%
IU	97	4,6%	96	4,6%	4,6%
Otros partidos	220	10,5%	229	10,9%	10,7%
En blanco	147	7,0%	151	7,2%	7,1%
Abstención	391		393		
Total	**2489**	**100,0%**	**2488***	**100%**	**100%**
El porcentaje de voto nulo estimado no se tendrá en cuenta. Es por eso que el total de observaciones imputadas difiere de *k*=5 a *k=3					

Tabla 14: Resultados estimados para octubre de 2015

Muestra de enero de 2016:

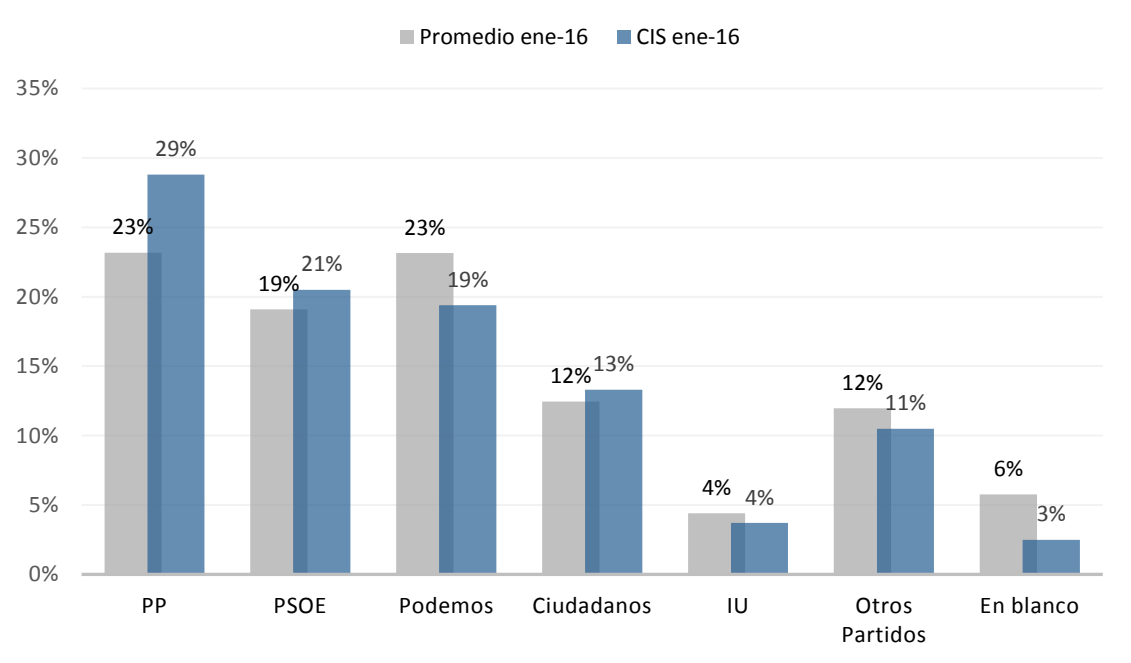


Figura 13:Diagrama de barras: comparación estimaciones enero y resultados del CIS en enero

Para los resultados de enero sigue existiendo una infraestimación considerable de PP y también, como ha sido advertido anteriormente, se da una sobreestimación de Podemos. El resto de categorías se asemejan considerablemente a las estimaciones del CIS.

	k=5		k=3		
ene-16	**Frecuencia estimada**	**% estimado sobre el voto válido**	**Frecuencia estimada2**	**% estimado sobre el voto válido2**	**Media de imputaciones**
PP	496	23,0%	499	23,3%	23,2%
PSOE	414	19,2%	406	19,0%	19,1%
Ciudadanos	267	12,4%	268	12,5%	12,5%
Podemos	490	22,7%	504	23,6%	23,1%
IU	97	4,5%	92	4,3%	4,4%
Otros partidos	256	11,9%	258	12,1%	12,0%
En blanco	135	6,3%	113	5,3%	5,8%
Abstención	334		349		
Total	**2489**	**100,0%**	**2489***	**100%**	**100%**
El porcentaje de voto nulo estimado no se tendrá en cuenta. Es por eso que el total de observaciones imputadas difiere de *k*=5 a *k=3					

Tabla 15: Resultados estimados para enero de 2016

6.2 RESUMEN IMPUTACIONES

En primer lugar, se expone la *Tabla 14* a modo de resumen de las imputaciones obtenidas en el proceso y, a su vez, comparativa con los resultados del CIS.

Intención de voto	PP	PSOE	Podemos	Ciudadanos	IU	Otros Partidos	En blanco
Imp. k=5 oct-15	20,6%	24,4%	13,2%	19,7%	4,6%	10,5%	7,0%
Imp. k=3 oct-15	21,3%	24,2%	13,5%	18,2%	4,6%	10,9%	7,2%
Promedio oct-15	**21%**	**24%**	**13%**	**19%**	**5%**	**11%**	**7%**
CIS oct-15	**29%**	**25%**	**11%**	**15%**	**5%**	**14%**	**2%**
Imp. k=5 ene-16	23,0%	19,2%	22,7%	12,4%	4,5%	11,9%	6,3%
Imp. k=3 ene-16	23,3%	19,0%	23,6%	12,5%	4,3%	12,1%	5,3%
Promedio ene-16	**23%**	**19%**	**23%**	**12%**	**4%**	**12%**	**6%**
CIS ene-16	**29%**	**21%**	**19%**	**13%**	**4%**	**11%**	**3%**

Tabla 14: Resumen de estimaciones y comparaciones

La principal diferencia encontrada con respecto a las estimaciones del CIS está en el porcentaje estimado para *PP*, disminuye considerablemente en la estimación de este estudio, y seguidamente el de Podemos, aumenta considerablemente en la estimación.

Cabe destacar que el porcentaje de votos en blanco estimado es considerablemente mayor en las estimaciones del estudio en comparación con las estimaciones del CIS.

El fenómeno de infraestimación de PP podría venir explicado por la existencia del "voto oculto": *Hay evidencias de que las personas sin ideología definida o con una ideología de derechas tienen a ser menos propensas a contestar a los entrevistadores. (M.Escobar, J.Rivière, R.Cilleros; 2014; p.56).* Es por ello, que requeriría la implementación de un algoritmo más elaborado que trate de recoger ese efecto.

En la Figura 14 representa una comparación del promedio de las imputaciones realizadas en el estudio y del promedio de estimaciones realizadas por el CIS, para octubre de 2015 y enero de 2016.

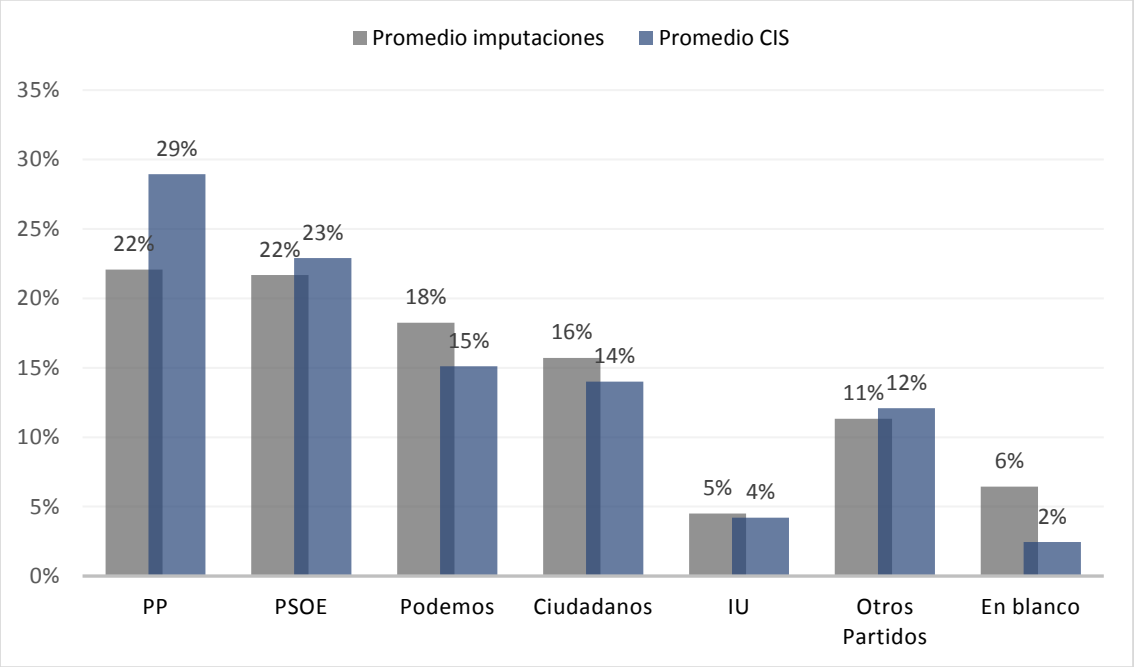


Figura 14: Comparación estimaciones con resultados CIS en promedio

6.2.1 Análisis de la incertidumbre

La incertidumbre tanto a nivel social, como político y económico es sinónimo de indecisión de la población, lo que provoca variabilidad en los datos y, por tanto, variabilidad en los resultados obtenidos. Se busca la manera de medir esa incertidumbre existente.

La incertidumbre es un índice de la calidad de la medida, la relación entre la calidad y la incertidumbre en inversamente proporcional, si la magnitud de la incertidumbre baja la calidad de la medida es mayor y viceversa.

Lo lógico sería realizar intervalos de confianza para los resultados de cada partido. Para su realización se parte de la siguiente idea:

Asumiendo que los individuos observados han contestado a todas las preguntas, es decir, que nos encontramos en el patrón de respuesta número 1. Se sabe que:

$$\widehat{P_i} \pm z_{\propto/2}\sqrt{\frac{\widehat{P_i}\,(1-\hat{P}_i)}{n}}$$

$$siendo\ \widehat{P_i}: proporción\ estimada\ de\ intención\ de\ voto\ para\ el\ partido\ i$$

$$n: número\ de\ individuos\ bajo\ cero\ no\ respuesta$$

$$z_{\frac{\propto}{2}}: valor\ para\ un\ nivel\ de\ confianza(1-\propto)$$

En nuestro caso, tenemos que

$$\widetilde{P_i} = \frac{\#\ individuos\ valor\ observado\ i + \#\ individuos\ valor\ imputado\ i}{m}$$

$$siendo\ \tilde{P}_i: proporción\ estimada\ de\ intención\ de\ voto\ para\ el\ partido\ i$$

$$m: número\ de\ individuos\ de\ la\ muestra$$

Por semejanza, se podría asumir que:

$$\tilde{P}_i \pm z_{\frac{\propto}{2}}\sqrt{\frac{\tilde{P}_i\left(1-\tilde{P}_i\right)}{m}}$$

Sin embargo, para la obtención de este intervalo se están asumiendo hechos que deberían ser contrastados previamente: 1) se está asumiendo que los valores imputados para el partido *i* siguen la misma distribución que los valores observados para el partido *i.* 2) Se supone que los individuos son independientes. Este último punto es cuestionable debido a

que, en base al procedimiento de imputación, los valores imputados dependen de los observados. Con esto, es aceptable suponer que los individuos sean independientes pero las respuestas imputadas no lo serían respecto a las respuestas observadas.

En un estudio más exhaustivo de la incertidumbre se deberían contrastar estos dos puntos para la obtención del intervalo de confianza deseado. Estos contrastes requieren algoritmos de gran complejidad que no serán desarrollados en este estudio.

7 CONCLUSIONES

En este punto, ofrece una síntesis de los puntos desarrollados en el estudio y los resultados alcanzados.

En la revisión del *entorno de la población* se ha visto: por una parte, los encuestados se diferencian principalmente por su *Ideología* e *Intención de voto*, y se pueden considerar cuatro principales agrupaciones según sean más afines al PP, PSOE, Podemos o Ciudadanos. Por otra parte, se ha localizado la posición de cada partido con respecto a las opiniones de los encuestados, destaca una clara separación de los encuestados que perciben la situación actual como buena o muy buena (autoubicados ideológicamente a la derecha) y los encuestados que perciben la situación actual como mala o muy mala (autoubicados ideológicamente a la izquierda). También destaca la proximidad existente entre los partidos Izquierda Unida y Podemos, lo que implica que están dirigidos al mismo conjunto de votantes.

La clave está en los encuestados indecisos: en primer lugar, se han definido los patrones de respuesta existentes en cada muestra, los cuales dan información sobre la distribución que siguen los datos faltantes, por ejemplo, en ambas muestras por lo menos el 58% de los encuestados ha respondido a todas las preguntas de interés; en segundo lugar, se ha contrastado la no aleatoriedad de los datos faltantes, rechazando la hipótesis nula de que los datos faltantes en la categoría intención de voto son independientes a los datos faltantes en el resto de categorías, concluyendo, que los datos faltantes no tienen una distribución completamente aleatoria, y por tanto, que son de tipo MNAR (*Missing Not At Random*).

En una descripción teórica de la *metodología* utilizada se han expuesto las razones de utilización del procedimiento *k*-ésimo vecino más cercano, su funcionamiento y los criterios utilizados para mejorar los resultados de imputación. Un primer criterio es la utilización de la distancia de Gower, por el hecho de que la muestra es un conjunto de datos de tipo mixto. Otro criterio determinante, se encuentra en la calibración del proceso, en la que se establecen pesos a cada variable de interés según el nivel de información que expliquen sobre la muestra, de forma que las variables como *Intención de voto, Simpatía o Recuerdo*

de voto, tendrán más peso a la hora de asignar un voto a un encuestado indeciso que la *Situación Política, Edad o Situación Económica.*

En la *evaluación interna del proceso* se tienen en cuenta tablas obtenidas mediante un proceso de validación cruzada, que indican en una evaluación final que existe un porcentaje de acierto en el procedimiento empleado entre el 46% y 54%. En este apartado se comparan los valores de *Intención de voto* esperados para los encuestados indecisos (artificialmente creados), con los valores reales observados en la muestra para los mismos individuos. De esta forma se ha visto cómo se distribuye el error en nuestro modelo de imputación, existe un error de infraestimación en el Partido Popular, y un error de sobreestimación en el partido de Podemos y seguidamente en Ciudadanos.

Con respecto a los *resultados finales* se concluye lo siguiente: en primer lugar, se refleja claramente la incertidumbre que existe en la población por cambios drásticos entre los resultados para octubre de 2015 y enero de 2016. Entre estos cambios se destaca un "intercambio de papeles" entre los partidos PSOE y Podemos, se invierte el porcentaje estimado de votos entre ambos partidos de octubre a enero. También se aprecia un descenso en la intención de voto a Ciudadanos de octubre de 2015 a enero de 2016. Con respecto a las comparaciones con las estimaciones del CIS se puede decir que hay similitud entre los resultados, exceptuando la estimación al PP con una diferencia del 7%.

Finalmente, se plantea la cuestión de *si sería posible cuantificar la incertidumbre* en los resultados estimados. Se defiende la idea de sí que es posible, siempre conociendo la distribución de los valores resultantes de la imputación, para ello es necesaria la realización de determinados modelos boostrap con los que obtener estimadores robustos de la varianza de los resultados.

En una visión general del estudio se concluiría diciendo que la estimación de la intención de voto es un problema complejo en el que influyen muchos factores, algunos conocidos y otros aún por determinar. Son muchas las teorías y metodologías que se han empleado en la literatura especializada para paliar la complejidad de este problema, aunque no se ha llegado a una solución definitiva. No obstante, hoy día se cuenta con un recurso en potencia como son las redes sociales (fuente de "infinita" información), y también, hay un desarrollo acelerado de técnicas para el análisis de esa información (Big Data). Sin duda, la combinación de estas nuevas técnicas y estos recursos serán una alternativa para la realización de pronósticos electorales en el futuro.

ANEXOS:

Anexo.1 Recodificación

Tabla de recodificaciones. Octubre 2015							*Cuestionario nº 3114*	
Variable	**Cód. Original**				**Variable**	**Recodificación**		
P.23	Intención de voto				P.23_recod			
	PP	01	UPN	14		PP	01	
	PSOE	02	Podemos	15		PSOE	02	
	IU	03	Ciudadanos	16		Podemos	03	
	UPyD	04	Unió	17		Ciudadanos	04	
	Convergència	05	Voto nulo	77		IU	05	
	Amaiur	06	Otro partido	95		Otros partidos	06	
	ERC	07	En blanco	96		En blanco	07	
	PNV	08	No votaría	97		*Voto nulo*	*77*	
	BNG	09	No lo sabe todavía	98		*Abstención*	*97*	
	CC	10	N.C	99		*No lo sabe todavía*	*98*	
	Compromís-Equo	11				*N.C*	*99*	
	FAC	12						
	Geroa Bai	13						
P.24	Simpatía hacia los partidos políticos				P.24_recod			
	PP	01	Compromís-Equo	10		PP	01	
	PSOE	02	FAC	11		PSOE	02	
	IU	03	Geroa Bai	12		Podemos	03	
	UPyD	04	UPN	13		Ciudadanos	04	
	Convergència	05	Podemos	14		IU	05	
	Amaiur	06	Ciudadanos	15		Otros partidos	06	
	ERC	07	Unió	16		En blanco	07	
	BNG	08	Otro partido	17		*Ninguno*	*97*	
	CC	09	En blanco	96		*N.S*	*98*	
			Ninguno	97		*N.C*	*99*	

			N.S	98			
			N.C	99			
P.29	Participación en elecciones noviembre 2011				P.29_recod		
	Fue a votar y votó	01	No recuerda	08		Fue a votar y votó	01
	No tenía edad para votar	02	N.C	09		No pudo	02
	Fue a votar pero no pudo hacerlo	03				Se abstuvo a votar	03
	No fue a votar porque no pudo	04				*No recuerda*	*98*
	Prefirió no votar	05				*N.C*	*99*
P.29ª	Recuerdo de voto en elecciones noviembre 2011				P.29A_recod		
	PP	01	BNG	09		PP	01
	PSOE	02	CC-NC	10		PSOE	02
	IU	03	Compromis-Equo	11		IU	03
	UPyD	04	FAC	12		Otros partidos	04
	CiU	05	Geroa Bai	13		No pudo votar	05
	Amaiur	06	Otros partidos	14		Abstención	06
	PNV	07	En blanco	15		En blanco	07
	ERC	08	Voto nulo	77		*Voto nulo*	*77*
			No recuerda	98		*No recuerda*	*98*
			N.C	99		*N.C*	*99*

Tabla de recodificaciones. Enero 2016						*Cuestionario nº 3124*	
Variable	**Cód. Original**				**Variable**	**Recodificación**	
P.16	Intención de voto				P.16_recod		
	PP	01	Convergència	09		PP	01
	PSOE	02	En Marea	10		PSOE	02
	Podemos	03	PNV	11		Podemos	03
	Ciudadanos	04	EH Bildu	12		Ciudadanos	04
	IU (Unidad Popular)	05	CC	13		IU	05
	En Comú Podem	06	Otros partidos	14		Otros partidos	06
	Compromís-Podemos	07	En blanco	15		En blanco	07
	ERC	08	Voto nulo	77		*Voto nulo*	*77*
			No lo sabe todavía	98		*Abstención*	*97*
			N.C	99		*No lo sabe todavía*	*98*
						N.C	*99*
P.17	Simpatía hacia los partidos políticos				P.17_recod		
	PP	01	Convergència	09		PP	01
	PSOE	02	En Marea	10		PSOE	02
	Podemos	03	PNV	11		Podemos	03
	Ciudadanos	04	EH Bildu	12		Ciudadanos	04
	IU (Unidad Popular)	05	CC	13		IU	05
	En Comú Podem	06	Otros partidos	14		Otros partidos	06
	Compromís-Podemos	07	En blanco	15		En blanco	07
	ERC	08	Ninguno	77		*Ninguno*	*97*
			N.S	98		*N.S*	*98*
			N.C	99		*N.C*	*99*
P.22	Participación en elecciones diciembre 2015				P.22_recod		
	Fue a votar y votó	01				Fue a votar y votó	01
	No tenía edad para votar	02				No pudo	02
	Fue a votar pero no pudo hacerlo	03				Se abstuvo a votar	03

	No fue a votar porque no pudo	04				*No recuerda*	*98*
	Prefirió no votar	05				*N.C*	*99*
	No recuerda	98					
	N.C	99					
P.22A	Recuerdo de voto en elecciones diciembre 2015				P.22A_recod		
	PP	01	Convergència	09		PP	01
	PSOE	02	En Marea	10		PSOE	02
	Podemos	03	PNV	11		Podemos	03
	Ciudadanos	04	EH Bildu	12		Ciudadanos	04
	IU (Unidad Popular)	05	CC	13		IU	05
	En Comú Podem	06	Otros partidos	18		Otros partidos	06
	Compromís-Podemos	07	En blanco	15		No pudo votar	07
	ERC	08	Voto nulo	77		Abstención	08
			No recuerda	98		En blanco	09
			N.C	99		*Voto nulo*	*77*
						No recuerda	*98*
						N.C	*99*

Anexo.2 Resultados electorales noviembre 2011:

Candidaturas	Votos	Diputados
PARTIDO POPULAR	10.866.566(44,63%)	186
PARTIDO SOCIALISTA OBRERO ESPAÑOL	7.003.511(28,76%)	110
ESQUERRA REPUBLICANA	256.985(1,06%)	3
EUZKO ALDERDI JELTZALEA-PARTIDO NACIONALISTA VASCO	324.317(1,33%)	5
IZQUIERDA UNIDA-LOS VERDES: LA IZQUIERDA PLURAL	1.686.040(6,92%)	11
AMAIUR	334.498(1,37%)	7
COALICIÓN CANARIA-NUEVA CANARIAS	143.881(0,59%)	2
PARTIDO ANIMALISTA CONTRA EL MALTRATO ANIMAL	102.144(0,42%)	-
UNIÓN PROGRESO Y DEMOCRACIA	1.143.225(4,70%)	5
BLOQUE NACIONALISTA GALEGO	184.037(0,76%)	2
PARTIDO COMUNISTA DE LOS PUEBLOS DE ESPAÑA	26.254(0,11%)	-
GEROA BAI	42.415(0,17%)	1
CIUDADANOS DE CENTRO DEMOCRÁTICO	1.074(0,00%)	-
ESCAÑOS EN BLANCO	97.673(0,40%)	-
FALANGE ESPAÑOLA DE LAS J.O.N.S.	2.898(0,01%)	-
POR UN MUNDO MÁS JUSTO	27.210(0,11%)	-
SOLIDARIDAD Y AUTOGESTION INTERNACIONALISTA	6.863(0,03%)	-
PARTIDO HUMANISTA	10.132(0,04%)	-
UNIDAD DEL PUEBLO	1.138(0,00%)	-
PARTIDO DE LA LIBERTAD INDIVIDUAL	2.065(0,01%)	-
DEMOCRACIA NACIONAL	1.867(0,01%)	-
PARTIDO REGIONALISTA DEL PAIS LEONES	2.058(0,01%)	-
FAMILIA Y VIDA	829(0,00%)	-
MUERTE AL SISTEMA	791(0,00%)	-
CONVERGÈNCIA I UNIÓ	1.015.691(4,17%)	16
BLOC-INICIATIVA-VERDS-EQUO-COALICIÓ COMPROMÍS	125.306(0,51%)	1
FORO DE CIUDADANOS	99.473(0,41%)	1

Fuente: Ministerio del Interior, Gobierno de España

Anexo.3 Resultados electorales diciembre 2015:

Candidaturas	Votos	Diputados
PARTIDO POPULAR	7.215.752(28,72%)	123
PARTIDO SOCIALISTA OBRERO ESPAÑOL	5.530.779(22,01%)	90
PODEMOS	3.182.082(12,67%)	42
CIUDADANOS-PARTIDO DE LA CIUDADANÍA	3.500.541(13,93%)	40
EN COMÚ PODEM	927.940(3,69%)	12
COMPROMÍS-PODEMOS-ÉS EL MOMENT	671.071(2,67%)	9
ESQUERRA REPUBLICANA DE CATALUNYA-CATALUNYA SÍ	599.289(2,39%)	9
DEMOCRÀCIA I LLIBERTAT. CONVERGÈNCIA. DEMÒCRATES. REAGRUPAMENT	565.501(2,25%)	8
EN MAREA	408.370(1,63%)	6
EUZKO ALDERDI JELTZALEA-PARTIDO NACIONALISTA VASCO	301.585(1,20%)	6
UNIDAD POPULAR: IZQUIERDA UNIDA, UNIDAD POPULAR EN COMÚN	923.133(3,67%)	2
EUSKAL HERRIA BILDU	218.467(0,87%)	2
COALICIÓN CANARIA - PARTIDO NACIONALISTA CANARIO	81.750(0,33%)	1
PARTIDO ANIMALISTA CONTRA EL MALTRATO ANIMAL	219.191(0,87%)	0
UNIÓN PROGRESO Y DEMOCRACIA	153.505(0,61%)	0
NÓS-CANDIDATURA GALEGA (BNG-CG-FOGA-PCPG-PG)	70.464(0,28%)	0
UNIÓ DEMOCRÀTICA DE CATALUNYA	64.726(0,26%)	0
VOX	57.753(0,23%)	0
RECORTES CERO-GRUPO VERDE	48.222(0,19%)	0
MÉS	33.931(0,14%)	0
PARTIDO COMUNISTA DE LOS PUEBLOS DE ESPAÑA	30.897(0,12%)	0
GEROA BAI	30.554(0,12%)	0
EL PI - PROPOSTA PER LES ILLES	12.902(0,05%)	0
CIUDADANOS DE CENTRO DEMOCRÁTICO	10.805(0,04%)	0
ESCAÑOS EN BLANCO	10.060(0,04%)	0
FALANGE ESPAÑOLA DE LAS J.O.N.S.	7.594(0,03%)	0

X LA IZQUIERDA-LOS VERDES	7.342(0,03%)	0
SOM VALENCIANS	6.084(0,02%)	0
POR UN MUNDO MÁS JUSTO	4.533(0,02%)	0
SOLIDARIDAD Y AUTOGESTIÓN INTERNACIONALISTA	4.516(0,02%)	0
LOS VERDES-ECOPACIFISTAS	3.254(0,01%)	0
PARTIDO DA TERRA	2.957(0,01%)	0
PARTIDO HUMANISTA	2.908(0,01%)	0
CANARIAS DECIDE: LOS VERDES, UNIDAD DEL PUEBLO Y ALTERNATIVA REPUBLICANA	2.874(0,01%)	0
PARTIDO LIBERTARIO	2.833(0,01%)	0
ARA, PAÍS VALENCIÀ	2.487(0,01%)	0
EXTREMADURA UNIDA-EXTREMEÑOS	1.995(0,01%)	0
PARTIDO COMUNISTA OBRERO ESPAÑOL	1.906(0,01%)	0
DEMOCRACIA NACIONAL	1.685(0,01%)	0
INICIATIVA FEMINISTA	1.594(0,01%)	0
PARTIDO REGIONALISTA DEL PAÍS LEONES	1.363(0,01%)	0
EN POSITIU	1.276(0,01%)	0
CIUDADANOS LIBRES UNIDOS	1.188(0,00%)	0
CIUDADANOS RURALES AGRUPADOS	1.027(0,00%)	0
LIBERTAD NAVARRA, LIBERTATE NAFARRA	1.022(0,00%)	0
AVANT VALENCIANS	1.001(0,00%)	0
MÁLAGA POR SÍ	924(0,00%)	0
ANDALUCES DE JAÉN UNIDOS	771(0,00%)	0
FAMILIA Y VIDA	714(0,00%)	0
INDEPENDIENTES POR ARAGÓN	673(0,00%)	0
FORO DEMÓCRATA	454(0,00%)	0
SOLUCIONA	406(0,00%)	0
JUSTIZIA SOCIAL, PARTICIPACIÓN CIUDADANA	405(0,00%)	0
MUERTE AL SISTEMA	309(0,00%)	0
PARTIDO LIBERAL DE DERECHAS	204(0,00%)	0
ONGI ETORRI	110(0,00%)	0

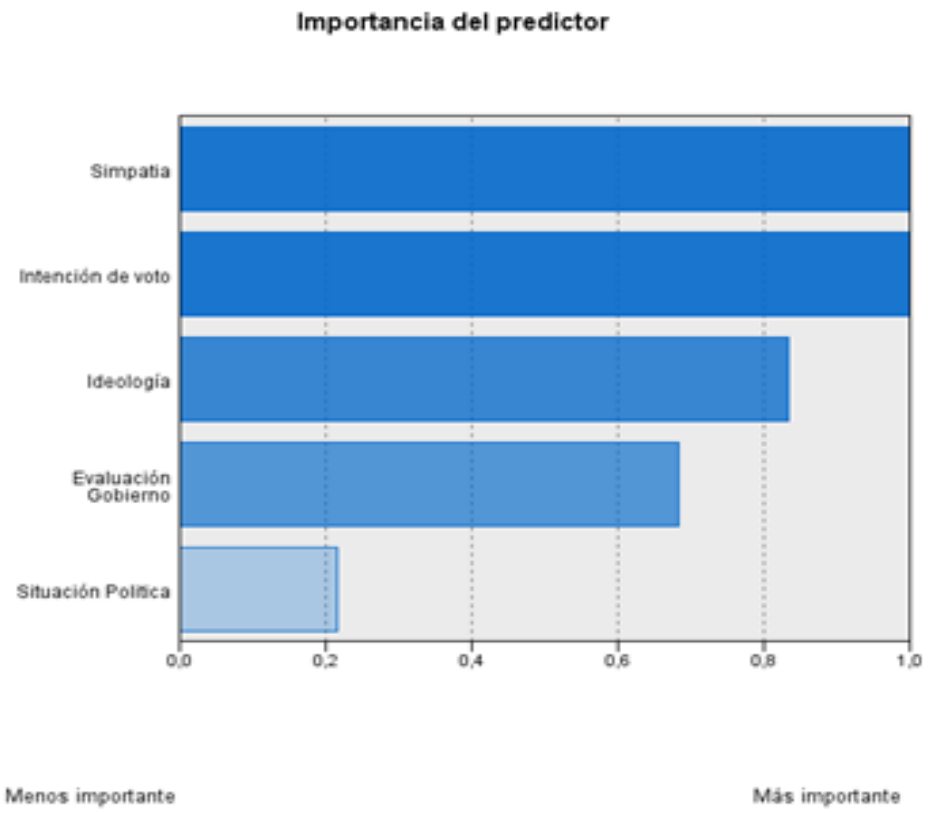
Fuente: Ministerio del Interior, Gobierno de España

Anexo.4 Clúster bietápico. Octubre 2015

Agrupaciones

Importancia de entrada (predictor)
1,0 0,8 0,6 0,4 0,2 0,0

Clúster	3	2	4	1
Etiqueta				
Descripción				
Tamaño	32,6% (499)	25,7% (393)	22,8% (349)	18,9% (290)
Entradas	Intención de voto Podemos (40,9%)	Intención de voto C´s (59,8%)	Intención de voto PSOE (96,6%)	Intención de voto PP (99,0%)
	Simpatia Podemos (34,3%)	Simpatia C´s (52,7%)	Simpatia PSOE (93,4%)	Simpatia PP (99,0%)
	Ideología 3,37	Ideología 5,35	Ideología 3,59	Ideología 7,25
	Evaluación Gobierno Muy mala (63,1%)	Evaluación Gobierno Regular (47,6%)	Evaluación Gobierno Muy mala (42,7%)	Evaluación Gobierno Buena (49,7%)
	Situación Politica Muy mala (55,7%)	Situación Politica Mala (41,7%)	Situación Politica Mala (44,7%)	Situación Politica Regular (50,7%)

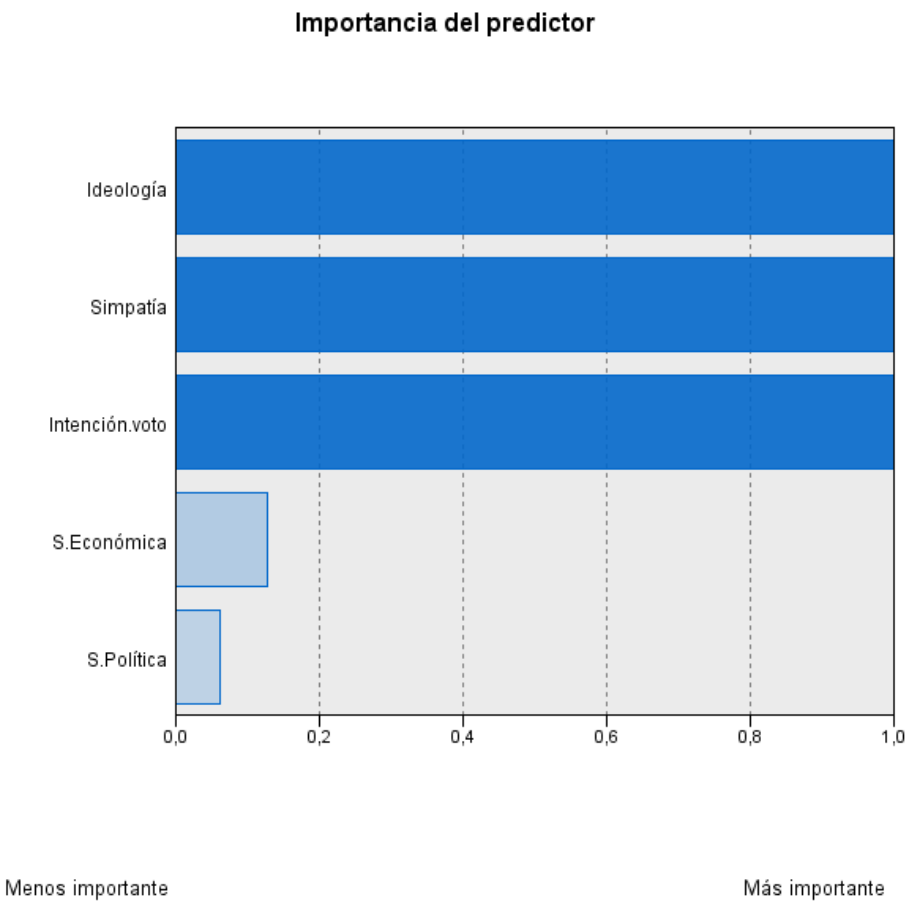

Importancia del predictor
Simpatia
Intención de voto
Ideología
Evaluación Gobierno
Situación Política
0,0
0,2
0,4
0,6
0,8
1,0
Menos importante
Más importante

Anexo.5 Clúster bietápico. Enero 2016

Agrupaciones

Importancia de entrada (predictor)
1,0 0,8 0,6 0,4 0,2 0,0

Clúster	2	1	3	5	4
Etiqueta					
Descripción					
Tamaño	24,5% (417)	20,7% (352)	19,0% (323)	18,9% (322)	17,0% (289)
Entradas	Ideología 5,13	Ideología 7,29	Ideología 3,75	Ideología 2,96	Ideología 3,28
	Intención.voto Ciudadanos (39,6%)	Intención.voto PP (96,6%)	Intención.voto PSOE (94,4%)	Intención.voto Podemos (100,0%)	Intención.voto
	Simpatía Ninguno (52,5%)	Simpatía PP (100,0%)	Simpatía PSOE (98,1%)	Simpatía Podemos (100,0%)	Simpatía
	S.Económica Mala (44,6%)	S.Económica Regular (54,0%)	S.Económica Mala (48,0%)	S.Económica Mala (48,4%)	S.Económica Mala (55,7%)
	S.Política Mala (45,6%)	S.Política Regular (39,2%)	S.Política Mala (40,2%)	S.Política Mala (38,8%)	S.Política Mala (47,1%)


Importancia del predictor
Ideología
Simpatía
Intención.voto
S.Económica
S.Política
0,0
0,2
0,4
0,6
0,8
1,0
Menos importante
Más importante

Anexo.6 Analisis de correspondencias múltiples octubre de 2015:

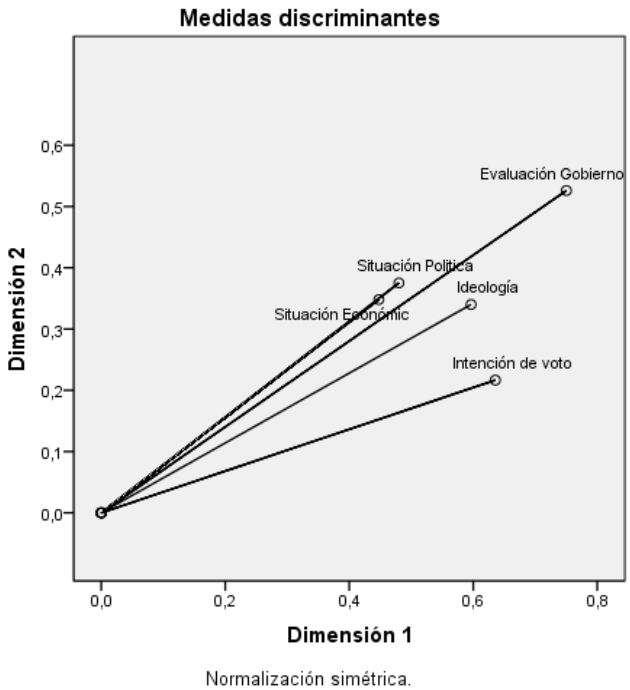
Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza contabilizada para	
		Total (autovalor)	Inercia
1	,820	2,910	,582
2	,558	1,806	,361
Total		4,716	**,943**
Media	,720[a]	2,358	,472

a. La media de alfa de Cronbach se basa en la media de autovalor.

Medidas discriminantes

	Dimensión		Media
	1	2	
Situación Económica	,448	,348	,398
Situación Politica	,480	,375	,428
Evaluación Gobierno	,750	,526	,638
Intención de voto	,635	,217	,426
Ideología	,596	,340	,468
Total activo	2,910	1,806	2,358



Normalización simétrica.

Anexo.7 Análisis de correspondencias múltiples. Enero de 2016:

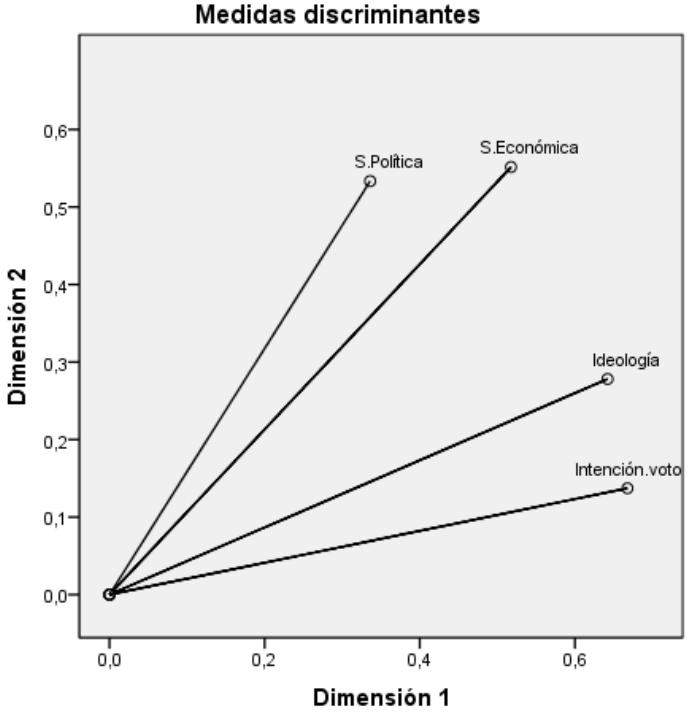
Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza contabilizada para	
		Total (autovalor)	Inercia
1	,717	2,163	,541
2	,445	1,500	,375
Total		3,664	,916
Media	,605[a]	1,832	,458

a. La media de alfa de Cronbach se basa en la media de autovalor.

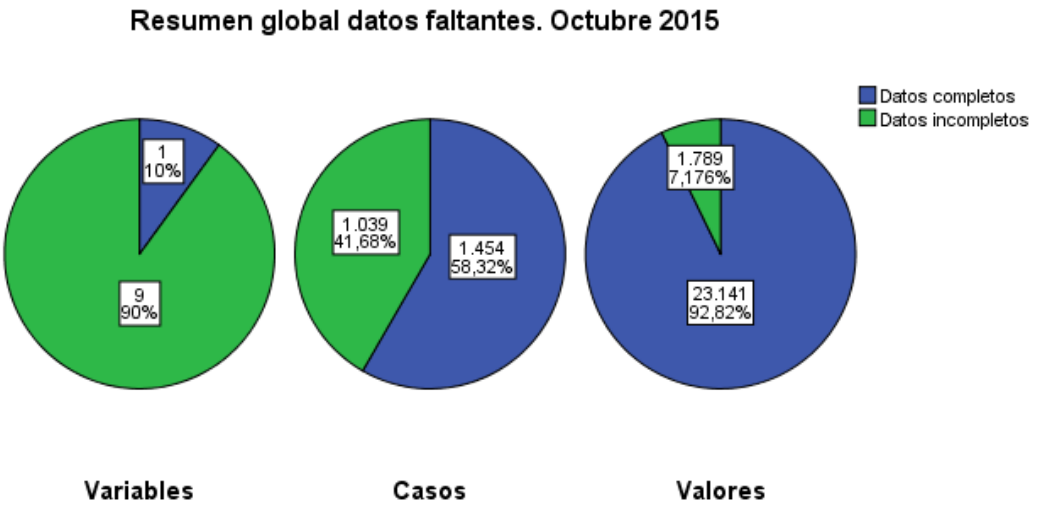
Medidas discriminantes

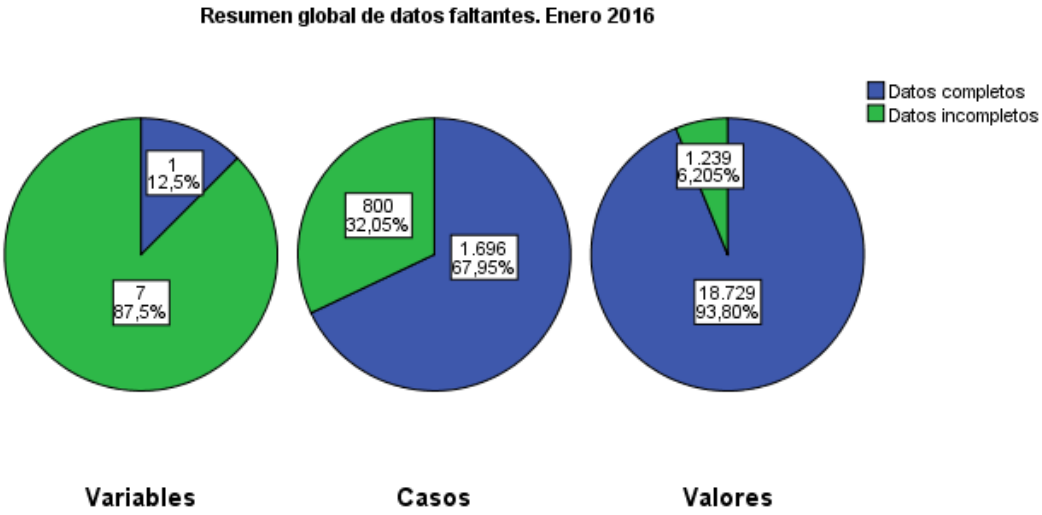
	Dimensión		Media
	1	2	
S.Económica	,518	,552	,535
Intención.voto	,668	,137	,402
Ideología	,642	,278	,460
S.Política	,336	,533	,435
Total activo	2,163	1,500	1,832



Normalización simétrica.

Anexo.8 Distribución de datos faltantes:





Edad	S.Económica	Gobierno	Participación	S.Política	Oposición	Simpatía	Recuerdo. voto	Ideología	Intención. voto	Total variables con faltantes	# individuos pattern	Prob.pattern	
1	1	1	1	1	1	1	1	1	1	0	1454	58,32%	Patrón 1
1	1	1	1	1	1	1	1	1	0	1	295	11,83%	Patrón 34
1	1	1	1	1	1	1	1	0	1	1	165	6,62%	Patrón 21
1	1	1	1	1	1	1	0	1	1	1	78	3,13%	Patrón 15
1	1	1	1	1	1	1	0	1	0	2	69	2,77%	Patrón 42
1	1	1	1	1	1	1	1	0	0	2	61	2,45%	
1	1	1	1	1	1	0	0	1	0	3	31	1,24%	
1	1	1	1	1	0	1	1	1	1	1	29	1,16%	
1	1	1	1	1	1	1	0	0	0	3	28	1,12%	
1	1	1	1	1	1	0	1	1	0	2	22	0,88%	
1	1	1	0	1	1	1	0	1	1	2	21	0,84%	
1	1	1	1	1	1	1	0	0	1	2	20	0,80%	
1	1	1	1	1	1	0	1	1	1	1	19	0,76%	
1	1	1	1	0	1	1	1	1	1	1	17	0,68%	
1	1	1	1	1	1	0	0	0	0	4	12	0,48%	
1	1	1	1	1	0	1	1	0	1	2	11	0,44%	
1	1	1	1	0	1	1	1	0	1	2	9	0,36%	
1	1	1	0	1	1	1	0	1	0	3	8	0,32%	
1	1	1	1	1	0	1	1	1	0	2	7	0,28%	
1	1	1	1	1	1	0	1	0	0	3	7	0,28%	
1	1	1	1	0	1	1	1	1	0	2	6	0,24%	
1	1	1	1	1	1	0	0	1	1	2	6	0,24%	
1	1	1	1	1	0	1	1	0	0	3	6	0,24%	
1	1	1	1	1	0	1	0	0	0	4	6	0,24%	
1	1	0	1	1	0	1	0	0	0	5	6	0,24%	
1	1	1	0	1	1	1	0	0	1	3	5	0,20%	
1	1	1	1	0	1	1	0	0	0	4	5	0,20%	
1	1	0	1	0	0	1	1	0	0	5	5	0,20%	
1	1	0	1	1	0	1	1	1	1	2	4	0,16%	
1	1	1	1	1	0	0	1	1	0	3	4	0,16%	
1	1	0	1	1	0	1	1	0	1	3	4	0,16%	
1	1	1	1	1	0	0	1	0	0	4	4	0,16%	
1	1	1	0	1	1	1	0	0	0	4	4	0,16%	
1	0	1	1	1	1	1	1	1	1	1	3	0,12%	
1	1	1	1	0	1	1	1	0	0	3	3	0,12%	
1	1	1	1	1	0	1	0	0	1	3	3	0,12%	
1	1	0	1	0	0	1	1	0	1	4	3	0,12%	
1	1	1	0	1	1	0	0	0	0	5	3	0,12%	
1	1	0	1	0	0	1	0	0	0	6	3	0,12%	
1	1	0	1	1	0	0	0	0	0	6	3	0,12%	
1	1	1	1	1	1	0	1	0	1	2	2	0,08%	
1	1	1	1	1	0	1	0	1	1	2	2	0,08%	
1	1	0	1	0	0	1	1	1	1	3	2	0,08%	
1	1	1	1	0	1	0	1	1	0	3	2	0,08%	
1	1	1	1	0	1	1	0	0	1	3	2	0,08%	
1	1	1	1	1	0	0	0	1	0	4	2	0,08%	
1	1	1	0	1	1	0	0	1	0	4	2	0,08%	
1	1	0	1	1	1	1	1	1	1	1	1	0,04%	
1	1	1	1	0	0	1	1	1	1	2	1	0,04%	
1	1	1	1	0	1	0	1	1	1	2	1	0,04%	
1	0	1	1	0	0	1	1	1	1	3	1	0,04%	
1	0	0	1	1	0	1	1	1	1	3	1	0,04%	
1	1	1	1	0	0	1	1	1	0	3	1	0,04%	
1	1	0	1	1	0	0	1	1	1	3	1	0,04%	
1	1	1	1	0	1	1	0	1	0	3	1	0,04%	
1	1	1	1	1	0	1	0	1	0	3	1	0,04%	
1	0	1	0	1	1	1	0	1	1	3	1	0,04%	
1	1	1	0	0	1	1	0	1	1	3	1	0,04%	
1	1	0	1	0	0	1	1	1	0	4	1	0,04%	
1	1	1	1	0	0	1	1	0	0	4	1	0,04%	
1	1	0	1	1	0	1	1	0	0	4	1	0,04%	
1	1	1	1	0	0	1	0	0	1	4	1	0,04%	
1	1	1	0	0	0	1	0	1	1	4	1	0,04%	
1	1	1	0	1	0	1	0	1	0	4	1	0,04%	
1	0	0	1	0	0	1	1	0	1	5	1	0,04%	
1	1	0	1	1	0	0	0	1	0	5	1	0,04%	
1	1	1	1	0	1	0	0	0	0	5	1	0,04%	
1	1	1	1	1	0	0	0	0	0	5	1	0,04%	
1	1	1	0	1	0	0	0	1	0	5	1	0,04%	
1	1	1	0	1	0	1	0	0	0	5	1	0,04%	
1	0	1	1	0	1	0	0	0	0	6	1	0,04%	
1	1	0	0	1	0	0	0	1	0	6	1	0,04%	
1	1	0	0	1	0	1	0	0	0	6	1	0,04%	
1	1	0	1	0	0	0	0	0	0	7	1	0,04%	
1	1	0	0	0	0	0	0	1	0	7	1	0,04%	
1	1	0	0	1	0	0	0	0	0	7	1	0,04%	
1	0	0	0	0	0	0	0	0	1	8	1	0,04%	

Edad	Participación	S.Económica	S.Política	Simpatía	Recuerdo. voto	Ideología	Intención. voto	Total variables con	# individuos patte	Prob. patte rn	
1	1	1	1	1	1	1	1	0	1696	68%	Patrón 1
1	1	1	1	1	1	0	1	1	212	8%	Patrón 10
1	1	1	1	1	1	1	0	1	202	8%	Patrón 18
1	1	1	1	1	0	1	0	2	76	3%	Patrón 31
1	1	1	1	1	0	1	1	1	45	2%	Patrón 6
1	1	1	1	1	1	0	0	2	40	2%	
1	1	1	0	1	1	1	1	1	36	1%	
1	1	1	1	1	0	0	0	3	32	1%	
1	1	1	1	0	0	1	0	3	27	1%	
1	1	1	0	1	1	0	1	2	20	1%	
1	1	1	1	0	0	0	0	4	17	1%	
1	1	1	1	0	1	1	0	2	15	1%	
1	1	1	1	1	0	0	1	2	15	1%	
1	1	1	1	0	1	1	1	1	11	0%	
1	1	1	0	1	1	1	0	2	5	0%	
1	1	1	1	0	1	0	0	3	5	0%	
1	1	1	0	1	0	0	0	4	5	0%	
1	1	0	1	1	1	1	1	1	3	0%	
1	1	0	0	1	1	1	1	2	3	0%	
1	1	1	1	0	1	0	1	2	3	0%	
1	1	1	0	1	1	0	0	3	3	0%	
1	1	1	0	0	0	0	0	5	3	0%	
1	1	1	0	1	0	1	1	2	2	0%	
1	1	1	1	0	0	1	1	2	2	0%	
1	0	1	1	1	0	1	1	2	2	0%	
1	1	1	0	1	0	1	0	3	2	0%	
1	1	1	0	1	0	0	1	3	2	0%	
1	0	1	1	0	0	1	0	4	2	0%	
1	0	1	1	0	0	0	0	5	2	0%	
1	1	0	0	1	1	0	1	3	1	0%	
1	1	1	1	0	0	0	1	3	1	0%	
1	0	1	1	1	0	1	0	3	1	0%	
1	0	1	1	1	0	0	1	3	1	0%	
1	1	1	0	0	0	1	0	4	1	0%	
1	1	0	1	1	0	0	0	4	1	0%	
1	1	0	0	0	1	0	0	5	1	0%	
1	1	0	1	0	0	0	0	5	1	0%	

Anexo.9 Código en R del procedimiento de imputación [Código para enero 2016]:

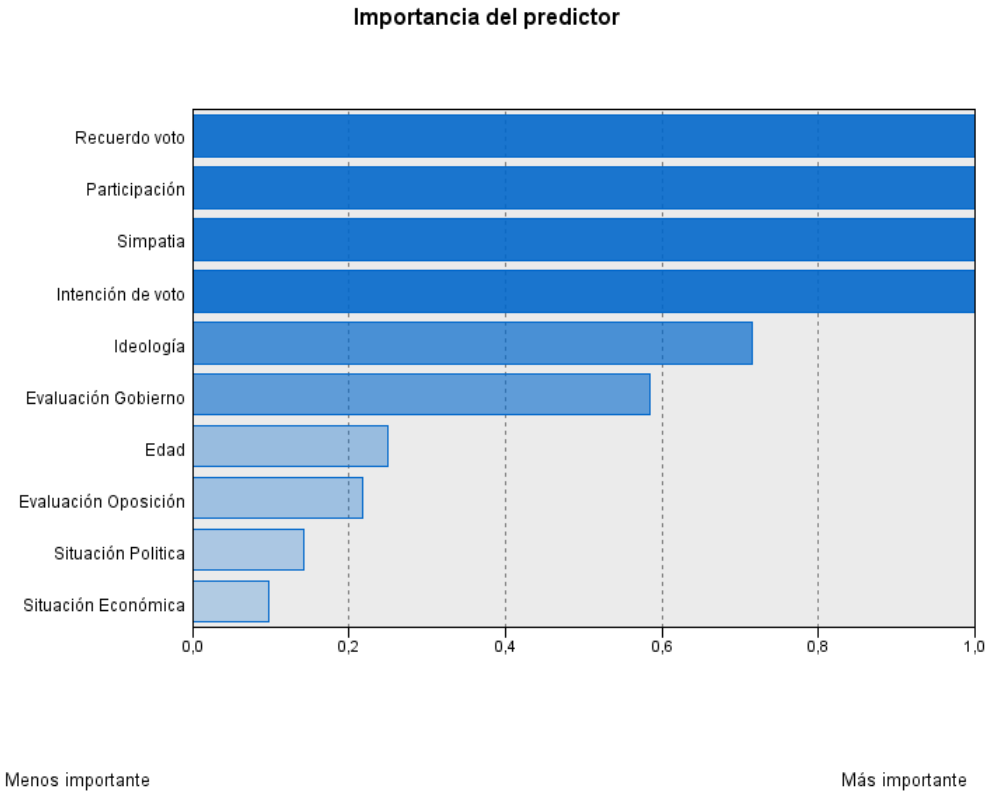
```
imp_subdataK5<-kNN(subdata,variable= colnames(subdata),metric=NULL,k=5,
            dist_var = colnames(subdata),weights = c(0.024904214559387,
          0.00766283524904215,
                                    0.191570881226054,   0.191570881226054,
                                    0.18007662835249,    0.191570881226054,
                                    0.191570881226054,   0.0210727969348659),
            numFun = median,
            catFun = maxCat, makeNA = NULL, NAcond = NULL, impNA = TRUE,
            donorcond = NULL, mixed = vector(), mixed.constant = NULL,
            trace = TRUE, imp_var = TRUE, imp_suffix = "imp", addRandom = TRUE,
            useImputedDist = TRUE)
write.csv2(imp_subdataK5, "imp_subdataK5.csv")


imp_subdataK3<-kNN(subdata,variable= colnames(subdata),metric=NULL,k=3,
            dist_var = colnames(subdata),weights = c(0.024904214559387,
          0.00766283524904215,
                                    0.191570881226054, 0.191570881226054,
                                    0.18007662835249,   0.191570881226054,
                                    0.191570881226054, 0.0210727969348659),
            numFun = median,
            catFun = sampleCat, makeNA = NULL, NAcond = NULL, impNA = TRUE,
            donorcond = NULL, mixed = vector(), mixed.constant = NULL,
            trace = TRUE, imp_var = TRUE, imp_suffix = "imp", addRandom = FALSE,
            useImputedDist = TRUE)
write.csv2(imp_subdataK3, "imp_subdataK3.csv")
```

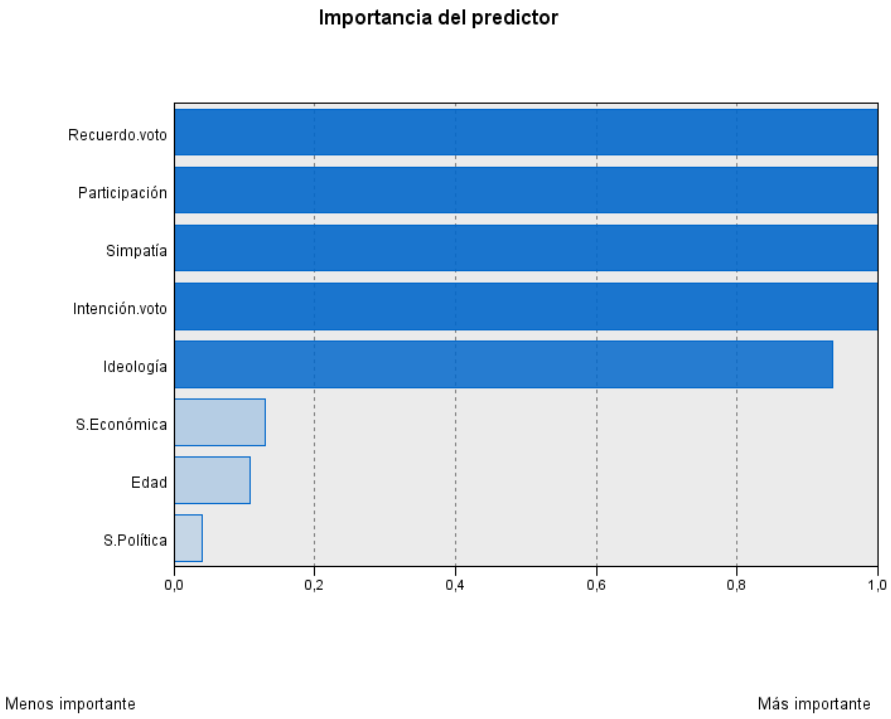
Anexo.10 C

Anexo.11 alibración del proceso: obtención de pesos de las variables mediante clúster bietápico

oct-15	Importancia predictor	Peso
Situación Económica	0,1	1,7%
Situación Política	0,14	2,3%
Gobierno	0,59	9,8%
Oposición	0,22	3,7%
Intención de voto	1	16,6%
Simpatía	1	16,6%
Ideología	0,72	12,0%
Participación	1	16,6%
Recuerdo voto	1	16,6%
Edad	0,25	4,2%
Total	**6,02**	**100,0%**



enero	Importancia predictor según análisis clúster	Peso
Situación Económica	0,13	2,5%
Situación Política	0,04	0,8%
Intención de voto	1	19,2%
Simpatía	1	19,2%
Ideología	0,94	18,0%
Participación	1	19,2%
Recuerdo voto	1	19,2%
Edad	0,11	2,1%
Total	**5,22**	100,0%



Anexo.12 Código en R de evaluación interna del proceso [Código para enero 2016]:

```
ind_sinmiss<- c(rownames(miss.P1), row.names(miss.P4), row.names(miss.P16),
            row.names(miss.P17), rownames(miss.P18), row.names(miss.P22),
            row.names(miss.P22A),rownames(miss.P25))
ind_sinmiss<-unique(ind_sinmiss)
ind_sinmiss<-sort(as.numeric(ind_sinmiss))
ind_sinmiss<-as.numeric(ind_sinmiss)
subdata_sinmiss<-subdata[-ind_sinmiss,] #muestra dentro de subdata SIN valores perdidos
library(mice)
pattern<-md.pattern(subdata) #matrix de 0 y 1, donde 1=observed y 0=missing
#pattern[,1] = # de observaciones del patrón
names(pattern)[9]<-"Total.missing"
pattern<-cbind(pattern, total.pattern= as.numeric(row.names(pattern)))
pattern<-as.data.frame(pattern)

for ( i in 1: dim(pattern)[1]){
  pattern[i,11]=pattern[i,10]/dim(subdata)[1]
}
names(pattern)[10]<-"Prob.pattern"
pattern<-pattern[-38,]
#tenemos una matriz con 37 patrones, donde la ultima columna representa la probabilidad
#para una observación de pertenecer a ese patrón.
write.csv2(pattern,"pattern.csv")

pattern_2<-data.frame(pattern$S.Económica,pattern$S.Política,pattern$Intención.voto,
                pattern$Simpatía, pattern$Ideología, pattern$Participación,
                pattern$Recuerdo.voto,pattern$Edad)
```

```
row.names(pattern)=1:37
pat<-sample(1:37, dim(subdata_sinmiss)[1], prob= pattern$Prob.pattern,replace =T)
fit<-subdata_sinmiss

for ( i in 1:dim(subdata_sinmiss)[1]){
  for( j in 1:dim(subdata_sinmiss)[2]){
    if(pattern_2[pat[i],j]==0){
      fit[i,j]=NA}  } }
#visualizacion de los valores perdidos generados
Amelia::missmap(fit,legend=TRUE, col=c("wheat","darkred"),
                y.cex=0.8,x.cex=0.8,csvar=NULL,tsvar=NULL,rank.other=TRUE)

#para la tabla de clasificación (comparación) necesitamos seleccionar únicamente
#las observaciones donde hemos generado datos faltantes.
# se excluye el patron 1 de fit en subdata_sinmiss.
pattern_fit<-md.pattern(fit) #matrix de 0 y 1, donde 1=observed y 0=missing

names(pattern_fit)[9]<-"Total.missing"
pattern_fit<-cbind(pattern_fit, total.pattern_fit= as.numeric(row.names(pattern_fit)))
pattern_fit<-as.data.frame(pattern_fit)

for ( i in 1: dim(pattern_fit)[1]){
  pattern_fit[i,11]=pattern_fit[i,10]/dim(fit)[1]
}
names(pattern_fit)[11]<-"Prob.pattern"
pattern_fit<-pattern_fit[-30,]
```

```
#debemos crear una submuestra para la evaluación del modelo que contenga:

#el número de las observaciones con datos faltantes en intención de voto de fit.

fit_voto<-data.frame(fit$Intención.voto)

fit_voto<-cbind(fit_voto, rows= as.numeric(row.names(fit_voto)))

na<-is.na(fit_voto$fit.Intención.voto)

ind<-c(0)

for(i in 1: dim(fit)[1]){if(na[i]== "TRUE"){ ind<-c(ind,fit_voto$rows[i]) } }

ind<-ind[-1]

#imputamos los valores con k=5 en la sub-submuestra con NA´s aleatorios

library(VIM)

imp_fitk5<-kNN(fit,variable= colnames(fit),metric=NULL,k=3,

          dist_var = colnames(fit),weights = c(0.024904214559387,
           0.00766283524904215, 0.191570881226054,    0.191570881226054,
0.18007662835249,      0.191570881226054, 0.191570881226054,
           0.0210727969348659),           numFun =median, catFun = sampleCat, makeNA =
NULL, NAcond = NULL, impNA = TRUE, donorcond = NULL, mixed = vector(), mixed.constant
= NULL, trace = TRUE, imp_var = TRUE, imp_suffix = "imp", addRandom = FALSE,
useImputedDist = FALSE)

sub_imp_fitk5<- imp_fitk5[c(ind),]

sub_sinmiss<- subdata_sinmiss[c(ind),]

tabla<-table(sub_sinmiss$Intención.voto,sub_imp_fitk5$Intención.voto) #x<.filas e
y.columnas

if(dim(tabla)[1]==9){

  tabla<-tabla[-8,]}

if(dim(tabla)[2]==9){

  tabla<-tabla[,-8]}

total.observado<-c(0)

total.imputado<-c(0)

for( i in 1:dim(tabla)[2]){

  total.observado<-c(total.observado, sum(tabla[i,]))

  total.imputado<-c(total.imputado,sum(tabla[,i]))

}
```

```
total.observado<-total.observado[-1]

total.imputado<-total.imputado[-1]


tabla<-cbind(tabla, total.observado)

tabla<-rbind(tabla,total.imputado)

tabla[9,9]=sum(total.observado)

row.names(tabla)=c("PP","PSOE","Podemos","Ciudadanos","IU","Otros partidos","En
blanco",

                   "Abstención","Total.imputado")

colnames(tabla)=c("PP","PSOE","Podemos","Ciudadanos","IU","Otros partidos","En
blanco",

                  "Abstención","Total.observado")


#tabla es nuestra tabla de clasificación que recoge la evaluación del proceso
```

BIBLIOGRAFÍA

- ALLISON (1999) *Multiple Imputation for Missing Data: A Cautionary Tale.* University of Pennsylvania.
- A. SOUNDERS, MORROW-HOWELL, SPITZNAGEL, DORI, K. PROCTOR, PESCARINO. (2006) *Imputing Missing Data: A Comparison of Methods for Social Work Researchers.* National Association of Social Workers
- BLACKWELL, HONAKER, KING. (2012) *Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data.*
- BUUREN, GROOTHUIS-OUDSHOORN. (2000) *MICE: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Sofware. Disponible en: http://www.jstatsoft.org/
- CIS (2016) Centro de Investigaciones sociológicas. *Barómetros*. Disponible en: http://www.cis.es/cis/opencm/ES/11_barometros/index.jsp (Consulta: 08/03/2016)
- D. ALLISON. (2012) *Moderns Methods of Missing Data.* Disponible en www.statisticalhorizons.com
- ESCOBAR, RIVIÈRE, CILLEROS. (2014) *Los pronósticos electorales con encuestas. Elecciones generales en España (1979-2011)*. Madrid: Elecciones, ISBN: 978 84 7476 650 9
- GUISADO, LEY (2015) *Nuevos Partidos, misma ideología.* El Mundo. Disponible en: http://www.elmundo.es/grafico/espana/2015/12/20/5669a90346163f741f8b4587.html (Consulta: 15/03/2016)
- HERRANZ (2015) *¿Predice el Big Data sobre redes sociales mejor que las encuestas quiénes ganan las elecciones?* Xataka. Disponible en:

http://www.xataka.com/aplicaciones/predice-el-big-data-sobre-redes-sociales-mejor-que-las-encuestas-quienes-ganan-las-elecciones (Consulta: 19/05/2016)

- HONAKER, KING. (2010) "What to Do about Missing Values in Time-Series Cross-Section Data." American Journal of Political Science, Vol. 54, No. 2, Pp. 561–581
- HONAKER, KING, BLACKWELL (2015) *Amelia II: A program for Missing Data.* R-project. Disponible en: https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf (Consulta: 20/03/2016)
- J. HORTON, P. KLEINMAN. (2007) "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." American Statistical Association. Vol. 61, No.1
- KING, HONAKER, JOSEPH, SCHEVE. (2001) "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation". American Political Science Review. Vol. 95, No.1
- KLEINKE, REINECKE. (2013) *A Multiple Imputation Package for Incomplete Count Data.* University of Bielefeld, Faculty of Sociology
- MICHY (2015) *Imputing Missing Data with R; MICE package.* R-bloggers. Disponible en: http://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/ (Consulta: 20/03/2016)
- Ministerio del Interior. Gobierno de España (2016) *Consulta de resultados electorales.* Disponible en: http://www.infoelectoral.interior.es/min/ (Consulta: 16/04/2016)
- SCHAFER. (1999) *Multiple Imputation: a primer.* University of Pennsylvania, Departament of Statistics