Autor: HÉCTOR ÁLVARO SAN JOSÉ

Tutor: ANDRÉS M. ALONSO FERNÁNDEZ

UTILIZACIÓN DE ENCUESTAS PREVIAS EN LA IMPUTACIÓN DE DATOS FALTANTES

2017

uc3m | Universidad Carlos III de Madrid

Grado en ESTADÍSTICA Y EMPRESA

Tabla de contenido

ILUSTRACIONES

Dedicado a *Canito*, la persona más importante en mi vida, mi referencia y mi gran amor. Eternamente agradecido por los valores que me ha transmitido, por la persona que ha hecho de mí.

Los abuelos deberían ser eternos.

Dedicado también a Yeray y a los que como él luchan cada día contra el monstruo del cáncer.

No estáis solos.

CAPÍTULO I

1. INTRODUCCIÓN

Saber qué fuerza política podría llegar a gobernar el país es uno de los principales temas de interés en cualquier sociedad democrática, especialmente cuando se acercan periodos electorales. Sin embargo, es habitual ver todos los meses diferentes sondeos que intentan pronosticar cuál sería el resultado en el caso de que se celebrasen elecciones.

Según Lewis-Beck (2005) pronosticar una elección significa anticipar el resultado. El principal objetivo de todos los organismos interesados en la predicción del voto es acertar, en el mayor porcentaje posible, esos resultados de manera anticipada.

Sin embargo, por todos es sabido, que no siempre los sondeos logran ese objetivo. Sin ir más lejos, en los últimos tiempos la mayor parte de las empresas de demoscopia (palabra que la Real Academia Española define como estudio de las opiniones, aficiones y comportamientos humanos mediante sondeos de opinión) han errado a la hora de pronosticar el éxito del *Brexit*, la victoria de Trump en EEUU, el famoso *Sorpasso* en España que quedó lejos de producirse, la reelección de David Cameron como primer ministro británico o el rechazo de los colombianos al acuerdo de paz con las FARC, entre otros casos.

Aunque hay quien defiende que los sondeos no se equivocan del todo, como es el caso del director de Metroscopia, José Juan Toharia (2016) que afirma que decir que los sondeos están fallando es una absoluta exageración y que, por ejemplo, en el caso de EEUU acertaron rotundamente que el candidato que más votos iba a obtener era Hillary Clinton.

José Miguel de Elías (2016) señala que una de las causas es la dificultad de predecir los rápidos cambios de tendencia que se dan, por ejemplo, si se producen noticias de gran calado social. Según el sociólogo Ignacio Urquizu (2016) otra posible causa es el llamado "voto oculto", que hace referencia a la infraestimación de una opción política, dado que el encuestado no quiere decir a quién va a votar porque es consciente de que esa opción está socialmente mal vista.

Otra idea que ponen los expertos en la palestra es que los sondeos no son predicciones de lo que va a pasar, sino que solamente recogen un estado de ánimo, que puede variar desde el día de la encuesta hasta el día de la votación (Toharia Cortés, 2016). El ser humano es impulsivo e impredecible y "*las encuestas no son bolas de cristal, el futuro no lo escriben los sondeos, sino las personas*" (Urquizu Sancho, 2016).

Hay diferentes metodologías para intentar predecir la intención del voto y con las que trabajan las diferentes empresas de demoscopia, además de diferentes variables interesantes que pueden ayudar a la hora de estimar la proporción de votos que recibiría cada partido.

Sin embargo, "*la sociedad es cada vez más compleja, heterogénea y plural, lo que hace cada vez más difícil esta labor. Por ejemplo, dos personas de la misma edad, del mismo sexo, del mismo nivel de estudios, del mismo barrio y con salarios similares pueden votar cosas completamente opuestas*" (Toharia, 2016).

En este trabajo se va a detallar cómo se desarrollaría ese proceso de estimación utilizando el método de imputación múltiple y los datos recogidos por el CIS (Centro de Investigaciones Sociológicas) y publicados de manera periódica en su web oficial. Hay que destacar la transparencia, la facilidad de acceso a los datos y la sencillez a la hora de tratarlos. Sin embargo, cabe señalar que existen otras muchas empresas dedicadas a la demoscopia. En particular, algunas empresas de demoscopia españolas son las siguientes.

Sigmados: es la primera empresa de demoscopia de España y se creó en 1982. Realiza encuestas de manera online con el objetivo de estimar la intención del voto y otras variables de interés para empresas tanto públicas como privadas. Sin embargo, al tratarse de un organismo privado no muestra ni da información de su manera de obtener los datos, de estimación y de tratamiento.

Metroscopia: Es una empresa que realiza estudios sociales y de opinión, entre ellos, políticos y preelectorales. Colabora desde 2008 con el diario El País y al igual que el CIS publica cada cierto tiempo un barómetro electoral.

MyWord-Social and Market Research: Es una empresa privada que tiene un proyecto junto a la Cadena Ser, llamado obSERvatorio, dedicado entre otras cosas a la estimación de la intención de voto. Realizan encuestas online mensualmente con muestras muy cercanas a los 1000 individuos. No ofrecen mucha información sobre los datos recopilados ni sobre el proceso de estimación.

Encuestamos: Es otra de las empresas que realizan encuestas y estimaciones de la intención de voto en España para entidades públicas y privadas. Recopila información de otras entidades que publica en su web Electomanía, pero no muestra información de los datos recogidos, del muestreo o del tratamiento.

Hemos elegido trabajar con los datos facilitados por el CIS porque, en adición a lo que se ha comentado anteriormente, se considera una fuente fiable y, por tanto, que se reducen los errores. Según la clasificación que hace Groves (1989) existen dos tipos de errores, varianza y sesgo, dependiendo de la constancia del error, que a su vez divide en error de no observación y de observación. Los primeros hacen referencia al muestreo y a las tasas de no cobertura y de no respuesta. Los segundos abarcan todo lo referente a la realización de la encuesta, encuestador, encuestado, instrumentos utilizados y modo de realización.

Se considera que utilizando el CIS como fuente de datos estos errores se reducen.

Por otro lado, como ya hemos comentado, el voto oculto es una de las principales causas del sesgo en las estimaciones de intención de voto y será objetivo principal de nuestro trabajo dar una estimación de dicha intención para aquellos encuestados que no respondan a nuestra variable de interés y, en particular, valorar cómo influye la utilización de encuestas previas en nuestras estimaciones.

Como curiosidad, se puede decir que trabajar con el voto oculto puede dar resultados positivos. En Estados Unidos, Arie Kapteyn acertó de pleno que Trump sería el presidente de los Estados Unidos. Para ello, enviaron miles de cartas con una encuesta y un billete de cinco dólares y a quien respondía le premiaban con otros quince. Con una muestra entre 3000 y 4000 personas bien estratificada, sabiendo que la mayoría de los que iban a votar a Trump no lo reconocerían y trabajando con preguntas sobre la posibilidad de ir a votar y de votar a un candidato u otro logró acertar el resultado de esas elecciones.

Volviendo a nuestro estudio, antes de nada, hay que detallar que el CIS realiza las encuestas con ámbito nacional y a población española mayor de 18 años, de ambos sexos, en 239 municipios de las 50 provincias (incluyendo Ceuta y Melilla). Además, aplica sus cuestionarios mediante entrevistas personales en los domicilios de los encuestados. En este trabajo utilizaremos las encuestas publicadas en julio y octubre de 2015 y de 2016. Los tamaños muestrales de esas encuestas fueron 2486, 2493, 2479 y 2491 personas, respectivamente. En cuanto al muestreo utilizado nos detallan que 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*." (CIS, 2016).

El objetivo y motivación principal por el cual surge este trabajo es valorar cómo afecta el uso y la inclusión de encuestas previas, es decir, realizadas con anterioridad al momento en el que se quiere estimar la variable de interés, mediante la comparación de los resultados con las estimaciones publicadas por el CIS y con los resultados electorales de las dos últimas elecciones generales en España en los años 2015 y 2016. Son periodos políticamente convulsos donde el sesgo y los errores de observación, comentados anteriormente, pueden desvirtuar o influir negativamente en las estimaciones. También es importante dejar claro que no se busca comparar modelos de estimación, ya que los del CIS no están detallados claramente.

Para el desarrollo del trabajo se utilizará el método de imputación de datos, con el objetivo de tratar los valores faltantes (no respuesta o indecisos) y lograr unas estimaciones para cada una de las variables de interés (de tipo económico, político y demográfico), pero en particular, para la intención del voto. Además, se estudiarán los posibles patrones de las no respuestas y se detallará todo el proceso llevado a cabo para lograr el objetivo. También se añadirán reflexiones y valoraciones en cada una de las secciones del trabajo con el fin de explicar de la mejor manera posible el trabajo realizado,

para que sea sencillo de entender, transparente y reproducible. En la siguiente sección de este capítulo, se va a desarrollar un análisis descriptivo de las variables de interés del estudio, con el objetivo de facilitar las interpretaciones. En el segundo capítulo se detallará la metodología utilizada para el estudio, centrándonos en el tipo de datos a utilizar, los tipos de imputación, así como las ventajas y desventajas de cada uno de ellos. Por otro lado, el trabajo se centra en dos casos prácticos con la intención de comparar los resultados, evaluar el método de imputación y la utilización de encuestas previas. Para ello se utilizarán las encuestas de octubre y julio de 2015 y de 2016 publicadas por el CIS y el software de programación *R*, en particular, el paquete estadístico *mi* (siglas que provienen de *multiple imputation*) con el que se crearán códigos de elaboración propia para realizar todo el proceso detallado anteriormente. Además de Excel que servirá como apoyo para la creación de tablas.Finalmente, se incluirán las conclusiones del estudio y un anexo con el código utilizado.

2. ANÁLISIS DESCRIPTIVO

En este apartado se desarrolla un análisis descriptivo de las variables utilizadas en el estudio, así como de la proporción de valores faltantes en cada una de ellas, incluyendo como ejemplo, gráficos de los datos de las encuestas de octubre 2015 y 2016, con la intención de realizar una comparativa entre ambas. Hay que destacar que se han elegido las encuestas de julio y octubre de ambos años ya que corresponden a periodos preelectorales y poselectorales, respectivamente, por lo que son momentos convulsos políticamente hablando, donde el encuestado puede ser más sensible o reacio a responder:

1. **Intención del voto** Es la variable principal y en la que se basa el estudio. En ella se pide al encuestado que diga a qué partido político votaría en el hipotético caso de que al día siguiente se celebrasen elecciones. En el código R (ver Anexo) se le denota como `Intención.Voto`.

   Las respuestas a esta pregunta abarcan un gran abanico de partidos políticos, así como la opción de voto nulo, voto en blanco, no votar y NS/NC. Esta última se consideran valor faltante y serán a las que se les impute un valor. Para la realización del trabajo, que se centra en ver cómo afecta la utilización de encuestas previas en la imputación de datos, se ha considerado recodificar la variable con las siguientes opciones:

   1.1. Intención de voto al PP (Partido Popular)
   1.2. Intención de voto al PSOE (Partido Socialista Obrero Español)
   1.3. Intención de voto a IU (Izquierda Unida)
   1.4. Intención de voto a Podemos
   1.5. Intención de voto a Ciudadanos
   1.6. Abstención. Aquellos que señalan que no van a votar
   1.7. Otra opción. Resto de fuerzas políticas, voto en blanco y voto nulo

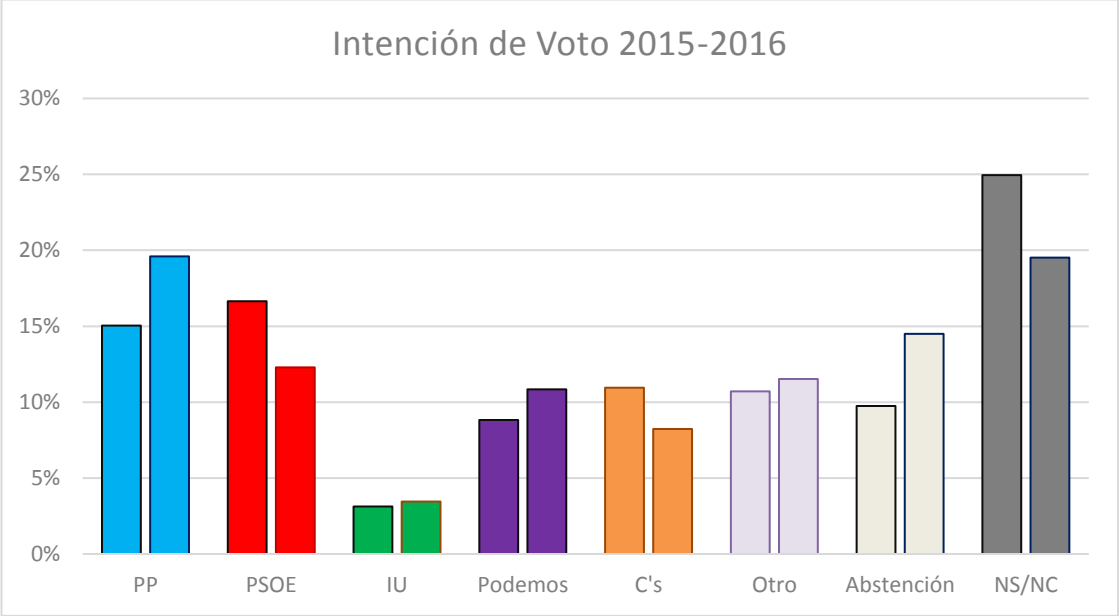


Figura 1. Comparativa Intención del voto sobre el total. Octubre 2015-2016. Elaboración propia.

El objetivo del estudio será estimar el voto de aquellos que no responden mediante imputación múltiple, por lo tanto, se tendrá que estimar un valor para el casi 25% de no respuestas en octubre de 2015 y el 20% en octubre de 2016.

2. **Situación económica actual:** Esta variable recoge las valoraciones de los encuestados en referencia a la situación económica actual. Toma valores de 1 a 5 considerando 1 como muy buena y 5 como muy mala. En el código R (ver Anexo) se le denota como `Sit.Econ`.

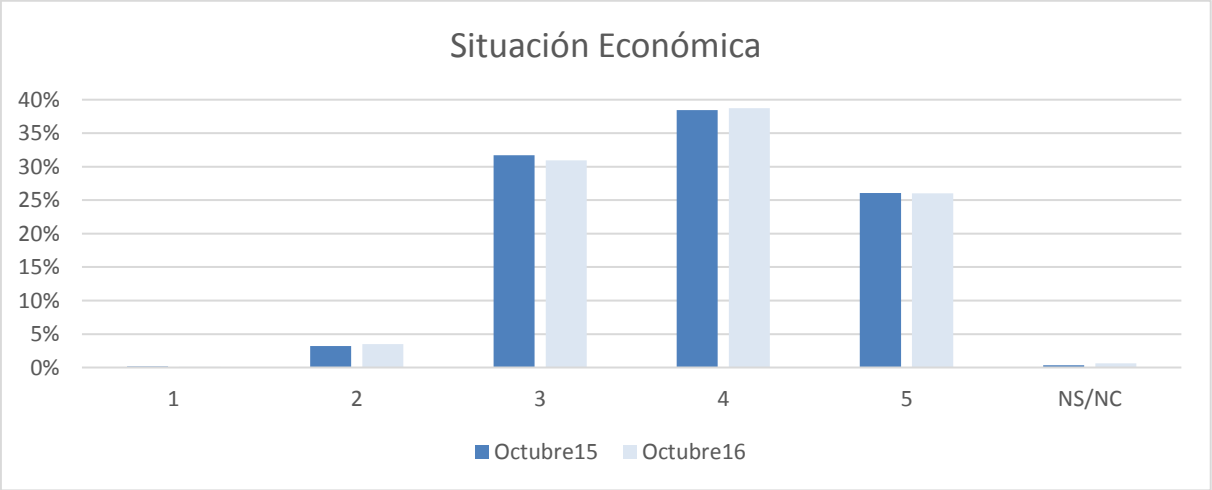


Figura 2. Evaluación de la situación económica. Octubre 2015 y 2016. Elaboración propia.

Se observa que la mayoría de los encuestados consideran que la situación económica era mala y es una apreciación que cambia muy poco con el paso de un año. Ni siquiera un 5% de los encuestados consideraba buena o muy buena la situación económica en España en esos años. Hay que destacar que es una variable con muy poco porcentaje de no respuesta, lo que facilitará el estudio.

3. **Situación política actual:** Esta variable recoge las valoraciones de los encuestados en referencia a la situación política actual. Toma valores de 1 a 5 cuya consideración es la misma que en la variable anterior. En el código R (ver Anexo) se le denota como `Sit.Polit`.

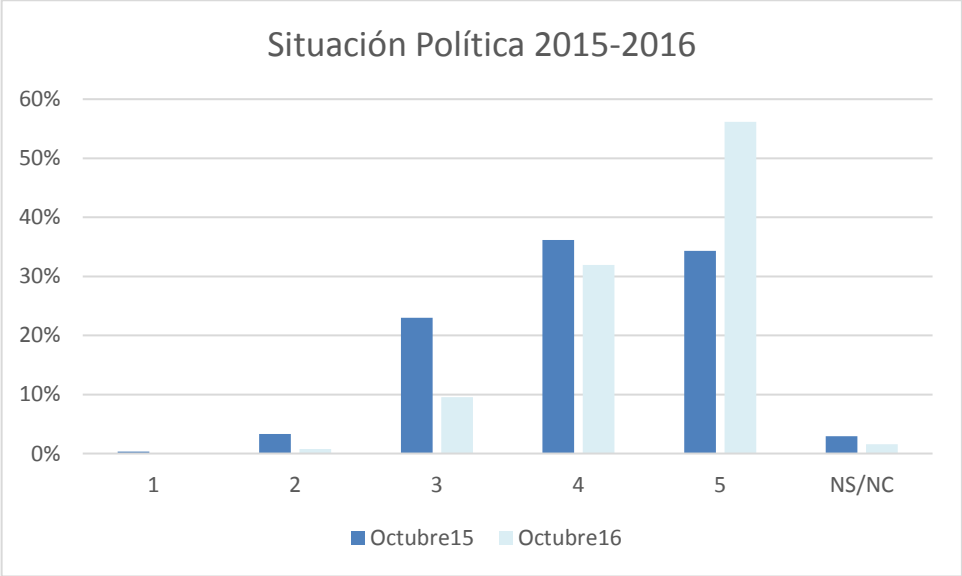


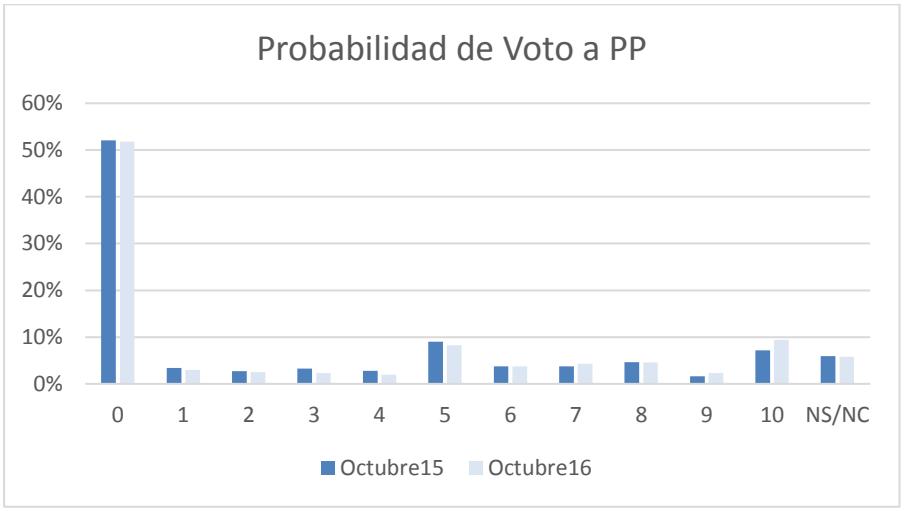
Figura 3. Comparativo de Valoración de la situación política. Octubre 2015 y 2016. Elaboración propia.

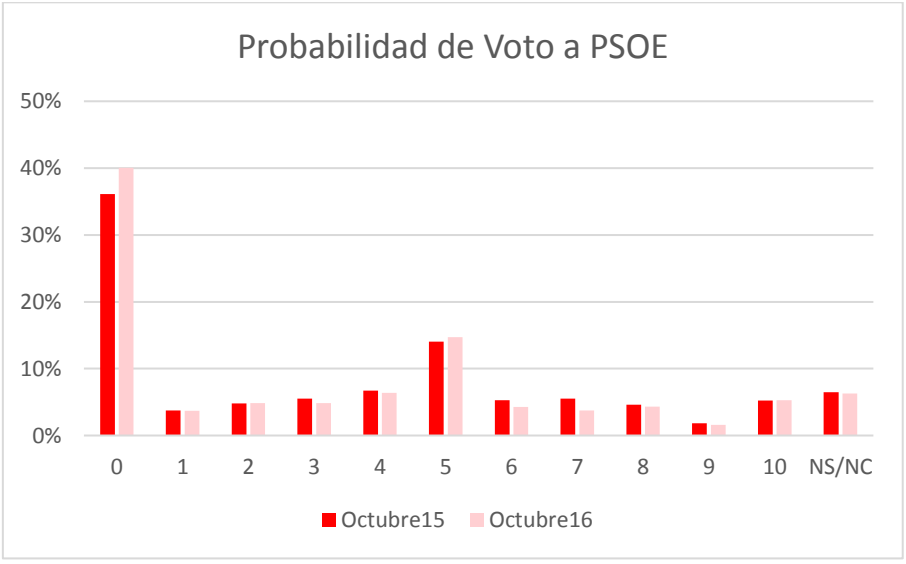
En este aspecto se observa una distribución de respuestas similar a la situación económica en octubre de 2015, pero destaca altamente que en octubre de 2016 la situación política fuera considerada mala por más del 50% de los encuestados. Es un cambio transcendente pero entendible ya que el contexto político en octubre de 2016 venía de dos elecciones en menos de siete meses y se preveían unas posibles terceras elecciones que, como se ve en el aumento de la abstención y de la consideración negativa de la situación política en octubre del 2016, no eran bien vistas por gran parte de los encuestados.

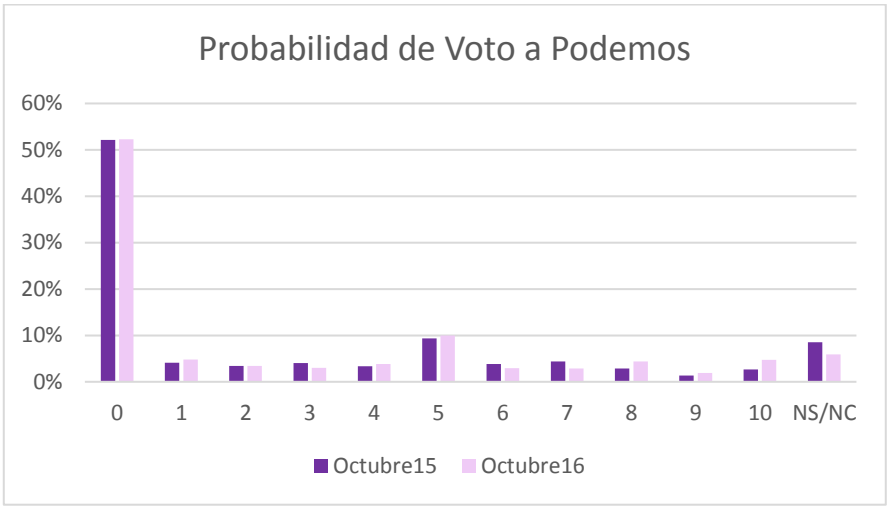
También es destacable el bajo porcentaje de no respuestas.

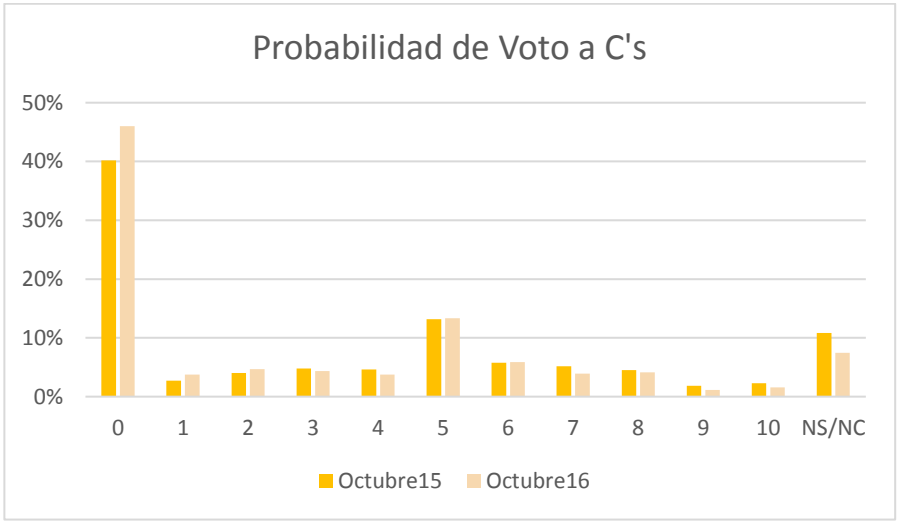
4. **Probabilidad de voto a cada partido:** Para el estudio se tienen en consideración cinco variables en las que se recogen la probabilidad, de 0 a 10, de que los encuestados votasen al día siguiente a cada uno de los partidos políticos que se muestran a continuación, en el hipotético caso de que se celebraran elecciones al Parlamento:

   4.1. Partido Popular (`Prob.PP`)
   4.2. Partido Socialista Obrero Español (`Prob.PSOE`)
   4.3. Izquierda Unida (`Prob.IU`)
   4.4. Podemos (`Prob.Podemos`)
   4.5. Ciudadanos (`Prob.Cs`)









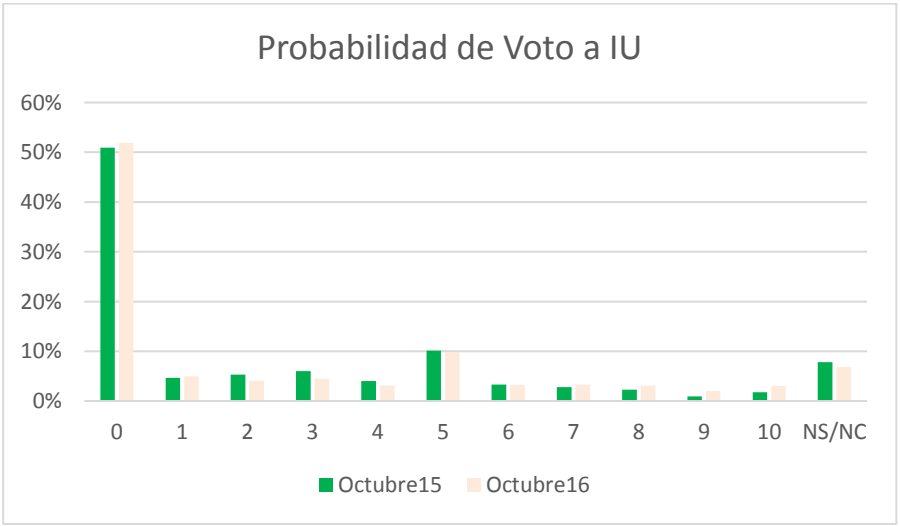
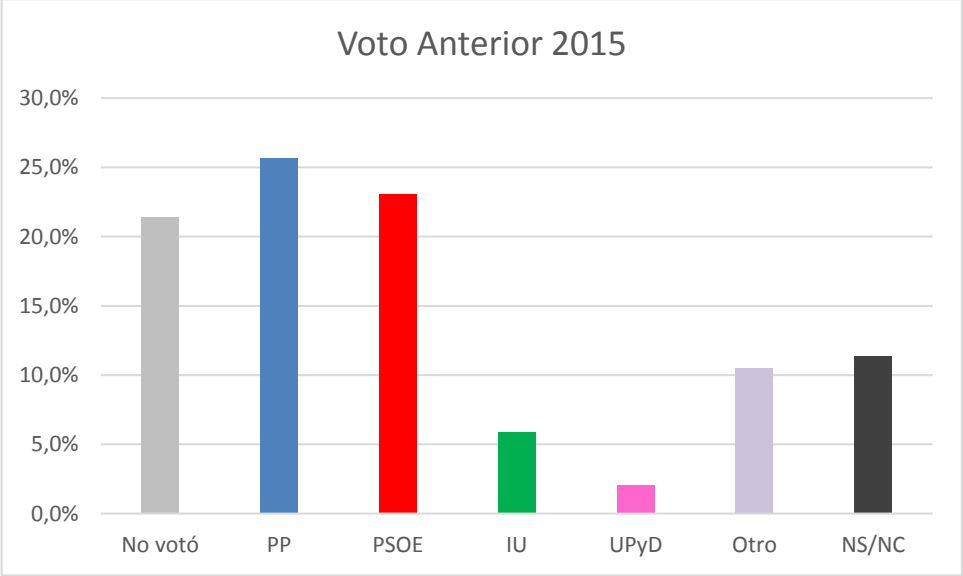


Figura 4. Comparativa probabilidad de voto. Octubre 2015-2016. Elaboración propia

Lo más destacable a primera vista es que en octubre de 2015 los encuestados consideraban al PSOE como la opción con más posibilidades para darle su voto. Tomando los porcentajes de las barras de probabilidad de 5 a 10, sumaría un total de 36.3%, por encima de C's (32.7%), PP (29.9%), Podemos (24.5%) e IU (21.2%). Sin embargo, si consideramos los resultados de las elecciones generales de diciembre de 2015, no deja de ser sorprendente. Hay que tener en consideración el resto de las variables del estudio, pero en un primer análisis es interesante intentar entender por qué el PSOE parecía la opción más atractiva en aquel momento y no obtuvo los resultados que se podrían haber esperado.

En cuanto a los valores perdidos se puede decir que en ninguna de las 5 variables se superan el 10% de no respuesta, a excepción de `Prob.Cs` para la encuesta de octubre de 2015, por lo que pueden ser de ayuda en el estudio.

5. **Voto anterior:** Esta variable recoge la opción por la que optaron los encuestados a la hora de votar en las últimas elecciones al Parlamento español que se celebraron, en la que se recogen los distintos partidos políticos y una opción para los que no lo recordaban. En el código R (ver Anexo) se le denota como `Voto.Ant`.



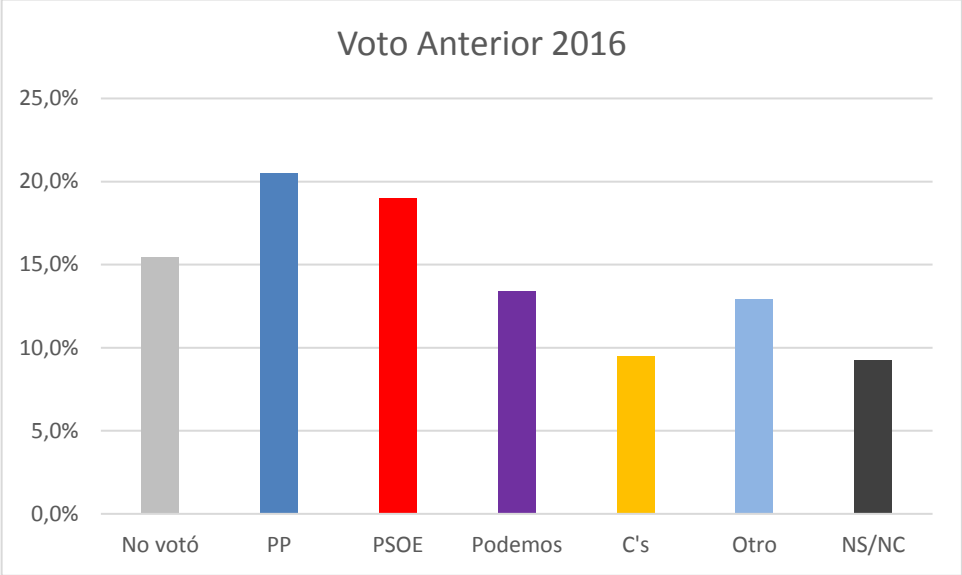


Figura 5. Recuerdo del voto anterior. Octubre 2015 y 2016. Elaboración propia.

Donde no votó corresponde a no haber votado en las últimas elecciones o haber votado en blanco o nulo, otro a las otras opciones posibles y NS/NC a los valores perdidos.

En esta variable hay que considerar que en octubre de 2015 no se encuentran ni Podemos ni Ciudadanos como posibles opciones ya que en las elecciones de 2011 no existían o no tuvieron representación parlamentaria y que en octubre de 2016 no se encuentra IU ya que en junio de 2016 se presentaron en una candidatura conjunta con Podemos.

6. **Sexo:** Esta variable binaria asigna 1 si el encuestado es hombre y 2 si es mujer. En el código R (ver Anexo) se le denota como `Sexo`.

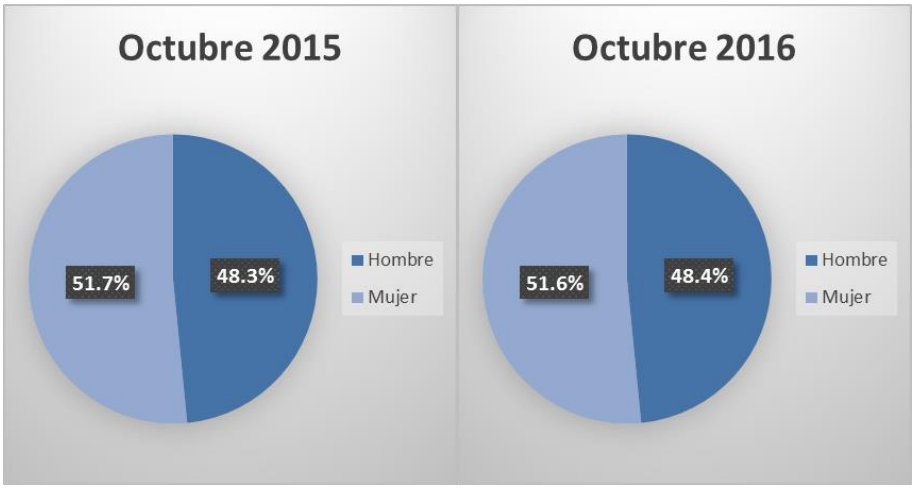


Figura 6. Sectograma del sexo de los encuestados. Julio 2015. Elaboración propia.

7. **Edad:** Esta variable recoge la edad del encuestado en el momento de realizar la encuesta. En el código R (ver Anexo) se le denota como `Edad`.

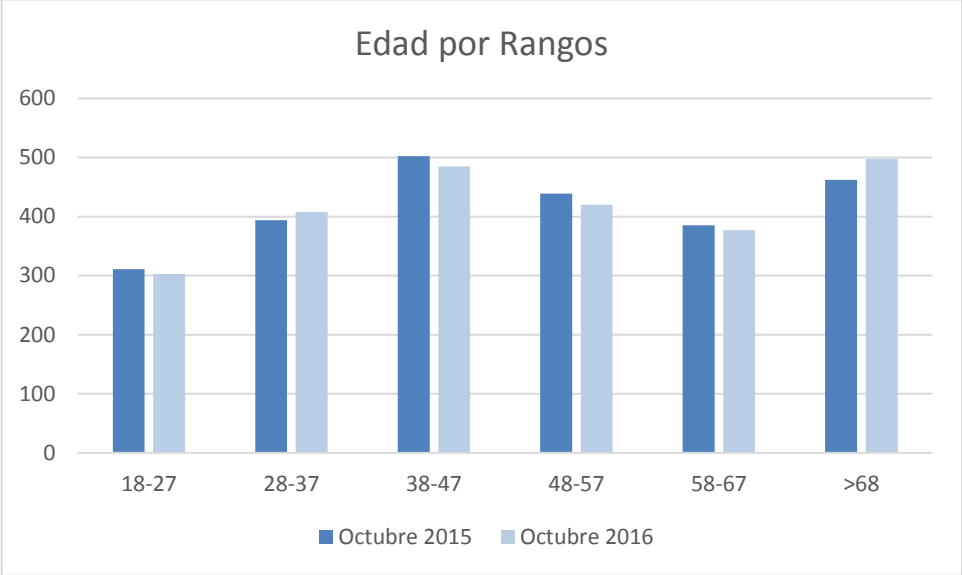


Figura 7. Distribución de la edad de los encuestados. Octubre 2015 y 2016. Elaboración propia.

Estas variables se han seleccionado para poder comparar las encuestas de octubre y julio de 2015 con las de los mismos meses de 2016 Se ha considerado la inclusión de más variables que finalmente se han descartado por no encontrarse en todas las encuestas utilizadas o por el alto coste computacional de las imputaciones.

A continuación, se realiza una comparación de la variable intención de voto en las cuatro encuestas que se utilizarán en el estudio, así como el porcentaje de los valores faltantes en dicha variable:

Tabla 1. Evolución del porcentaje de valores faltantes. Elaboración propia.

Año	2015		2016	
Mes	Julio	Octubre	Julio	Octubre
Faltantes	837	622	393	487
Porcentaje	35,12%	24,95%	15,85%	19,55%

Como se puede observar en la tabla 1, en el 2016 hay una notoria disminución de los datos faltantes en la variable de interés. Se podría decir que los encuestados tienen una predisposición mayor a la hora de contestar a la pregunta de intención de voto, pero la realidad es que los encuestados, por la situación política del momento (tras las segundas elecciones generales), se decantaron con más ocurrencia por la opción de no votar, como se puede observar en la figura 1, donde queda patente el aumento porcentual en dicha respuesta. Según ObSERvatorio, en una publicación de la Cadena Ser (2016) los encuestados definían en aquel momento su estado de ánimo por la situación política como "hartazgo" y "aburrimiento", y apuntaban que la participación en unas hipotéticas terceras elecciones habría caído casi un 10%.

Por lo tanto, se puede afirmar que la disminución en el porcentaje no es tan buena como parece a simple vista, sino que viene dada por la situación política del momento. En 2015 el periodo era preelectoral, por lo que suscitaba más interés entre el electorado, sin embargo, en julio y octubre del 2016 el periodo era poselectoral, por lo que el desinterés, aumentado por la circunstancia de que se habían celebrado segundos comicios en un periodo de tiempo menor a siete meses, podía ser normal. Esto reafirma la idea de que los sondeos no predicen lo que va a pasar, sino que recogen estados de ánimo.

Antes de empezar con el estudio también es adecuado analizar un pequeño resumen de las variables con las que vamos a trabajar y que obtenemos utilizando la función `summary()` de R:

Como se puede observar, la variable que más valores faltantes presenta es la referente a la intención del voto. También se observa que las variables sexo y edad son contestadas por todos los encuestados. Además, las variables de evaluación de la situación económica y política apenas presentan valores faltantes. Esto es una ventaja a la hora de realizar la imputación.

```
   Sit.Econ       Sit.Polit        Prob.PP       Prob.PSOE      Prob.IU      Prob.Podemos
Min.   :1.000   Min.   :1.000   0      :1298   0      :901   0      :1269   0      :1300
1st Qu.:3.000   1st Qu.:3.000   5      : 224   5      :350   5      : 252   5      : 234
Median :4.000   Median :4.000   10     : 178   4      :167   3      : 150   7      : 109
Mean   :3.874   Mean   :4.038   8      : 115   3      :138   2      : 133   1      : 102
3rd Qu.:5.000   3rd Qu.:5.000   7      :  94   7      :138   1      : 116   3      : 100
Max.   :5.000   Max.   :5.000   (Other): 436   (Other):637   (Other): 378   (Other): 435
NA's   :9       NA's   :73      NA's   : 148   NA's   :162   NA's   : 195   NA's   : 213
   Prob.Cs     Intención.Voto    Voto.Ant    Sexo          Edad
0      :1002   2      :415    1      :640   1:1205   Min.   :18.00
5      : 329   1      :375    2      :575   2:1288   1st Qu.:36.00
6      : 144   16     :273    0      :533            Median :48.00
7      : 129   19     :243    3      :147            Mean   :49.78
3      : 120   15     :220    4      : 52            3rd Qu.:64.00
(Other): 499   (Other):345    (Other):262            Max.   :94.00
NA's   : 270   NA's   :622    NA's   :284
```

Figura 8. Resumen de las variables. Julio 2015. Elaboración propia

CAPÍTULO II

1. METODOLOGÍA

En este capítulo se detalla qué es un valor faltante (*missing value)* y los tipos de datos faltantes que hay (MCAR, MAR y MNAR), en qué consiste el método de imputación, cuáles son lo métodos más conocidos y cuál se ha utilizado para la realización del trabajo (y que es usado por el paquete *mi* de R).

1.1 Datos faltantes

En los estudios que se realizan utilizando datos recogidos mediante encuestas es frecuente encontrar la existencia de datos faltantes o no respuestas, bien porque los encuestados no entiendan bien la pregunta, no se acuerden de determinados aspectos o, directamente, no quieran responder a ciertas preguntas.

Según Rubin (1976) existen varios tipos de datos faltantes:

- **MCAR (*missing completely at random*):** los datos de una variable son MCAR si la probabilidad de no respuesta es la misma para todos los encuestados y es independiente del resto de las variables y de la propia variable. En otras palabras, que los valores faltantes aparecen completamente al azar.
- **MAR (*missing at random*):** se da este caso si la ausencia de respuesta depende de alguna otra variable utilizada, es decir, si la probabilidad de no respuesta está asociada a la información observada.
- **MNAR (*missing not at random*):** los datos de una variable son MNAR si la probabilidad de no respuesta depende de los valores perdidos. Este tipo de datos también se denomina como no ignorable.

Ante esta falta de información es habitual trabajar simplemente con la información disponible y obviar los datos de aquellos que no responden a la variable de interés. Estos métodos se denominan *listwise* o *pairwise* deletion, cuya diferencia es la eliminación o no de los individuos con valores faltantes (Peugh y Enders, 2004). Para utilizar alguna de estas técnicas es necesario que los datos sean MCAR, lo cual es asumir una hipótesis que no siempre se cumple. Además, eliminar los registros que presentan datos faltantes, salvo que sea en un porcentaje muy reducido (menos del 5%), puede reducir considerablemente la muestra e introducir sesgo en las estimaciones.

1.2 Imputación de datos

Una solución a la eliminación de los registros con datos faltantes es el uso de técnicas de imputación para estimar el valor de esos datos faltantes. Existen varias técnicas de imputación:

- **Imputación simple:** Consiste en asignar un único valor para cada dato faltante, basándose en los datos de la propia y de las demás variables, obteniendo así una matriz de datos completa. Los métodos de imputación simple más comunes son:
    - Imputación mediante la media: este método, propuesto por Wilks en 1932 consiste en asignar a cada dato faltante la media de los datos observados.
    - Imputación mediante regresión: esta técnica consiste en asignar a cada valor faltante el obtenido mediante regresión, utilizando estimadores MCO (mínimos cuadrados ordinarios).
    - Imputación deductiva: esta técnica consiste en asignar a cada dato faltante un valor obtenido mediante deducción, es decir, mediante una relación lógica con otras variables. Por ejemplo, si un encuestado votó al PP en las anteriores elecciones y asigna una probabilidad relativamente alta a la posibilidad de volver a votar al PP, entonces su intención de voto será el mismo partido.

- **Imputación múltiple:** Consiste en asignar varios valores a cada respuesta faltante, generando varias matrices de datos para, posteriormente, combinar los resultados y conseguir un único valor estimado. Es habitual conseguir esa estimación global mediante reglas simples como, por ejemplo, tomar la media aritmética.

Según Otero García (2011) existe otra manera de clasificar las imputaciones:

- **Métodos de imputación determinísticos**: en los cuales se obtiene la misma respuesta cuando se repite la imputación, siempre y cuando las condiciones sean las mismas.
- **Métodos de imputación estocásticos**: en los cuales se obtienen diferentes respuestas cuando se repite la imputación con las mismas condiciones.

1.3 Ventajas y desventajas de la Imputación múltiple frente a la imputación simple

Como ya hemos comentado, la imputación simple tiene un gran problema ya que subestima los errores estándar e introduce sesgo en las estimaciones. Esto se soluciona con la imputación múltiple, pero por contrapartida, el coste computacional es más elevado.

Si se cuenta con una base de datos con un porcentaje reducido de valores faltantes es ventajoso utilizar la imputación simple, ya que la subestimación del error estándar no es tan elevada. Sin embargo, como ya hemos visto anteriormente, en nuestra variable de interés estamos trabajando con un porcentaje de valores faltantes entre el 15% y el 35%, por lo que es oportuno utilizar la imputación múltiple.

Por otra parte, bases de datos grandes con porcentajes elevados de no respuestas también produce una desventaja en la imputación múltiple ya que aumenta el coste computacional, más aún si la mayoría de las variables son categóricas. Por esta razón en este trabajo se realizan cuatro cadenas de imputaciones que posteriormente se unen en dos conjuntos de datos imputados finales promediando los resultados para cada conjunto de datos, utilizando únicamente 11 variables.

Adicionalmente, ambas metodologías asumen la hipótesis de que los datos son MAR (missing at random), es decir, que la ausencia de respuestas está asociada a otras variables incluidas en la matriz de datos y esto no es verificable en muchas ocasiones.

1.4 Imputación múltiple

Para este estudio hemos utilizado el método de imputación múltiple, una técnica de imputación estocástica propuesta por Rubin en 1987 cuya teoría se basa en la estadística bayesiana y que busca realizar inferencia respecto de los parámetros utilizando la información de las variables de la muestra. Rubin plantea combinar los resultados de varias imputaciones (bases completas) mediante reglas simples para ajustar el error estándar de dichas estimaciones y considera que con al menos tres bases imputadas es suficiente para estimar la incertidumbre asociada a los valores faltantes.

Por otro lado, para la realización de la imputación múltiple es condición necesaria que los datos sean MAR (missing at random), que como se describe anteriormente, significa que los datos faltantes dependen de otras variables. Poniendo como ejemplo nuestro estudio, que varios encuestados que reconozcan ser votantes de Ciudadanos nos respondan a otras variables (voto anterior, probabilidad de voto a otros partidos, valoración de la situación política, etc.) nos permite estimar si otros encuestados votarían también a ese partido político.

A la hora de combinar las simulaciones, para los parámetros estimados basta con tomar la media de los estimadores de cada una de las bases completas generadas (ver, por ejemplo, en Marín Diazaraque y Restrepo Estrada, 2012). Por otra parte, los errores estándar se calculan mediante otro proceso más complicado:

- Varianza dentro de la imputación: media aritmética de los errores estándar al cuadrado:

$$H = \frac{\sum_{t=1}^{m} SE^2}{m}$$

Donde t es una base imputada en particular, que en este trabajo por razones de coste computacional será t = {1, 2}, y m que es el conjunto total de bases imputadas (en este trabajo, m=2).

- Varianza entre imputaciones: cuantifica la variación de las estimaciones a lo largo de las bases imputadas.

$$B = \frac{\sum(\hat{\vartheta}_t - \bar{\vartheta})^2}{m - 1}$$

Donde $\hat{\vartheta}_t$ es el parámetro estimado (en el caso de nuestro estudio, de la intención de voto) del conjunto de bases completas y $\bar{\vartheta}$, la media de los parámetros estimados en las imputaciones.

Para concluir, el error estándar combinado se calcula mediante la combinación de los dos parámetros anteriores (Marín Diazaraque y Restrepo Estrada, 2012):

$$SE = \sqrt{(H + B + \frac{B}{m})}$$

Así, incorporando la varianza entre imputaciones en el error estándar, es como la imputación múltiple resuelve el principal problema de la imputación simple, la subestimación de los errores estándar.

Además, si el porcentaje de datos faltantes con respecto al total de respuestas es pequeño, las estimaciones y los errores estándar serán similares entre las bases imputadas, y la estimación global y el error estándar serán prácticamente idénticos (Marín Diazaraque y Restrepo Estrada, 2012)

2. CASO PRÁCTICO

Como ya se ha mencionado en anteriores ocasiones este trabajo está realizado mediante la implementación del método de imputación múltiple en el software R, utilizando la librería *mi*. En el Anexo 1 está detallado todo el código que se ha creado para la realización del estudio. Este paquete utiliza un algoritmo denominado ecuaciones encadenadas (Gelman, Hill, Yajima y Yu-sung, 2011).

Esta estrategia es flexible y práctica ya que permite la implementación en variables de distinto tipo (binarias, categóricas, continuas, etc.) y con distintos niveles de medición, es decir, con diferentes unidades de medida.

Las ecuaciones encadenadas de este estudio, que se generan partiendo de la variable con menos datos faltantes y yendo en orden ascendente por ese criterio para el resto de las variables, tienen la siguiente forma:

```
Intención.Voto ~ Sit.Econ + Sit.Polit + Prob.PP + Prob.PSOE + Prob.IU
+ Prob.Podemos + Prob.Cs + Voto.Ant + Sexo + Edad
```

El algoritmo, según Royuela Vicente (2014) sigue la siguiente metodología:

"Al inicio, se sustituyen todos los valores perdidos por valores extraídos aleatoriamente con reemplazamiento del conjunto de valores observados para cada variable.

A continuación, se aplica una regresión sobre la primera variable con el menor número de datos faltantes (por ejemplo, X1), condicionando el resto de variables ($X_2, X_3, \dots, X_n$) como términos independientes, solo sobre aquellas observaciones donde X_1 esté completa. Los valores faltantes de X_1 son reemplazados entonces por valores simulados de la correspondiente distribución predictiva posterior de X_1.

Este proceso se repite sucesivamente para todas las demás variables con datos faltantes, en orden ascendente de no respuestas y utilizando los valores imputados en las variables anteriores. A esto se le conoce como "ciclo".

Para estabilizar los resultados, el procedimiento suele repetirse varios ciclos, normalmente 10 o 20 para producir un único archivo de datos imputado promediando los resultados individuales. El proceso completo se repite m veces para conseguir m conjuntos de archivos de datos con datos imputados (sin datos faltantes) en todos ellos." (Royuela Vicente, 2014).

2.1 Resumen y análisis inicial

En esta y en las próximas secciones se explica de manera detallada el desarrollo del caso práctico realizado y cada uno de los pasos que se han ido ejecutando para conseguir el objetivo.

En primer lugar, se han seleccionado las encuestas de julio y octubre de 2015 y 2016 que el CIS (Centro de Investigaciones Sociológicas) realiza con el objetivo de estimar la intención de voto de los ciudadanos españoles. Nuestro objetivo es ver cómo afecta la utilización de encuestas anteriores en la imputación de datos faltantes y, en consecuencia, si mejora la estimación o no de nuestra variable de interés, la intención del voto.

Antes de empezar, y una vez cargados los datos, es oportuno analizar la base de datos con la que vamos a trabajar y un análisis descriptivo de las variables, como se detalla en el Capítulo I, para conocer las dimensiones de la matriz de datos, las variables a las que nos enfrentamos y visualizar un pequeño resumen estadístico de ellas.

A continuación, el primer paso a dar es convertir nuestra matriz de datos en un `missing_data.frame`. Esta clase es similar a un `data.frame`, pero se ha creado particularmente para la imputación múltiple, ya que se trabaja con datos faltantes.

Con `missing_data.frame` se crea una lista de variables con los datos faltantes y otras celdas que contienen metadatos y que indican como se relacionan entre sí dichas variables.

El aspecto más importante de una variable faltante es su clase, como continua, binaria o categórica. La función `missing_data.frame` intentará adivinar la clase apropiada para cada variable, pero no siempre se va a corresponder con la realidad. Por lo tanto, es importante ver cómo han sido clasificadas inicialmente las variables, utilizando la función `show()`, y modificarlas si fuera necesario.

Por ejemplo, la variable intención del voto es inicialmente clasificada como continua, lo que es erróneo, ya que la variable es categórica. Una vez que todas las variables estén configuradas apropiadamente, es útil realizar un análisis gráfico más exhaustivo. Con la función `image()` podemos observar los datos faltantes que hay en la encuesta:

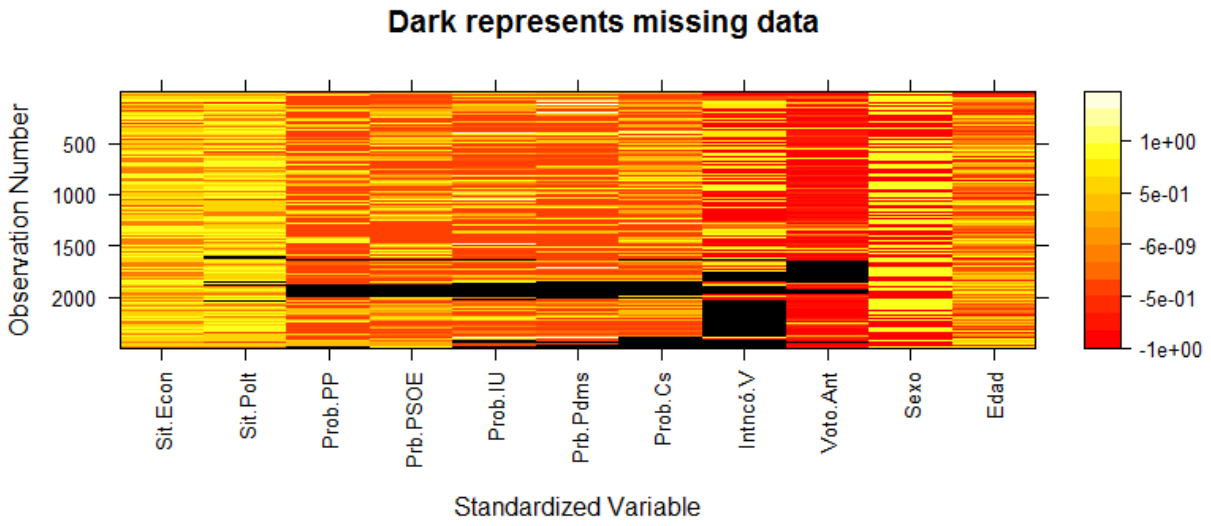


Figura 9. Datos faltantes encuesta octubre 2015. Elaboración propia.

Como indica el título del gráfico, los espacios en negro indican valores faltantes en la variable. Se observa con claridad que las variables `Sexo`, `Edad` y `Sit.Econ` son contestadas por prácticamente todos los encuestados y, sin embargo, la intención del voto es la que más valores faltantes presenta.

Se observa también que más de 1500 encuestados contestaron a todas las variables de interés. Uno de los patrones más claros que se observan es que hay una serie de encuestados que no contestan a ninguna variable donde hay relación directa con los partidos políticos (probabilidad de voto a cada partido, intención de voto y voto anterior). Otra información que se obtiene a primera vista es que hay un gran grupo de encuestados que contestan a todas las variables a excepción de la intención del voto.

2.2 Imputación

En este apartado se detalla el grueso del trabajo, la realización de las imputaciones y el proceso de comparación. En primer lugar, hay que destacar que la función que se utiliza para este proceso es `mi()`, que tiene varios argumentos adicionales como por ejemplo cuántas cadenas independientes utilizar, cuántas iteraciones a realizar y la cantidad máxima de tiempo que estamos dispuesto a esperar para que el proceso acabe.

Un tema importante a destacar es que el proceso de imputación puede ser bastante lento, sobre todo si hay muchas variables faltantes y si muchas de ellas son categóricas. En nuestro caso, a excepción de `Sexo` y `Edad`, todas las variables son categóricas, por lo que se ha optado por reducir lo máximo posible el número de variables y por realizar cuatro cadenas.

Cabe recordar que este proceso se va a repetir hasta cuatro veces. En primer lugar, con la encuesta de octubre de 2015 y a continuación, añadiendo las respuestas de los encuestados que respondieron a la variable de interés en julio de 2015. Posteriormente se repetirá este proceso con las encuestas de

octubre y julio de 2016, para poder evaluar de mejor manera si la utilización de encuestas previas tiene un efecto positivo.

Siguiendo con el proceso, llegamos a un paso que es realmente importante. Se trata de verificar si se realizaron suficientes iteraciones. En este caso se quiere que la media y varianza de cada variable completada sea aproximadamente la misma para cada una de las cuatro cadenas que hemos realizado. Además, nos tenemos que fijar en el coeficiente de convergencia para corroborar que efectivamente las iteraciones son suficientes. Se utilizan las funciones `mipply()` y `rhats()` para visualizar si el algoritmo a convergido:

	[,1]	[,2]	[,3]	[,4]
Sit.Econ	3.873	3.875	3.874	3.874
Sit.Polit	4.032	4.029	4.035	4.029
Prob.PP	-0.012	-0.004	-0.008	-0.005
Prob.PSOE	-0.006	-0.004	-0.004	-0.006
Prob.IU	-0.011	-0.009	-0.007	-0.001
Prob.Podemos	-0.012	-0.009	-0.015	-0.010
Prob.Cs	-0.019	-0.012	-0.014	-0.020
Intención.Voto	9.078	9.142	9.177	9.036
Voto.Ant	3.238	3.252	3.242	3.258
Sexo	1.517	1.517	1.517	1.517
Edad	0.000	0.000	0.000	0.000
missing_Sit.Econ	0.004	0.004	0.004	0.004
missing_Sit.Polit	0.029	0.029	0.029	0.029
missing_Prob.PP	0.059	0.059	0.059	0.059
missing_Prob.PSOE	0.065	0.065	0.065	0.065
missing_Prob.IU	0.078	0.078	0.078	0.078
missing_Prob.Podemos	0.085	0.085	0.085	0.085
missing_Prob.Cs	0.108	0.108	0.108	0.108
missing_Intención.Voto	0.249	0.249	0.249	0.249
missing_Voto.Ant	0.114	0.114	0.114	0.114

Figura 10. Media de las variables de las 4 cadenas de la imputación. Elaboración propia

mean_Sit.Econ	mean_Sit.Polit	mean_Prob.PP
0.9867789	0.9953227	0.9861505
mean_Prob.PSOE	**mean_Prob.IU**	**mean_Prob.Podemos**
0.9864906	0.9946295	0.9880993
mean_Prob.Cs	**mean_Intención.Voto**	**mean_Voto.Ant**
0.9868050	0.9901526	0.9874935
sd_Sit.Econ	**sd_Sit.Polit**	**sd_Prob.PP**
0.9879184	0.9913360	0.9898559
sd_Prob.PSOE	**sd_Prob.IU**	**sd_Prob.Podemos**
0.9881899	0.9929889	0.9908378
sd_Prob.Cs	**sd_Intención.Voto**	**sd_Voto.Ant**
0.9873356	0.9857912	0.9865608

Figura 11. Convergencia de la media y la varianza de las variables.

Se observa como prácticamente sí se ha producido la convergencia, pero optamos por realizar cinco iteraciones más para mejorar la estimación. Es posible realizar tantas iteraciones como sean necesarias para conseguir la convergencia:

A continuación, se realiza un histograma de los datos observados, imputados y completados, una comparación de los datos completos con los valores ajustados por el modelo y un gráfico de los residuos con la función *image()*:

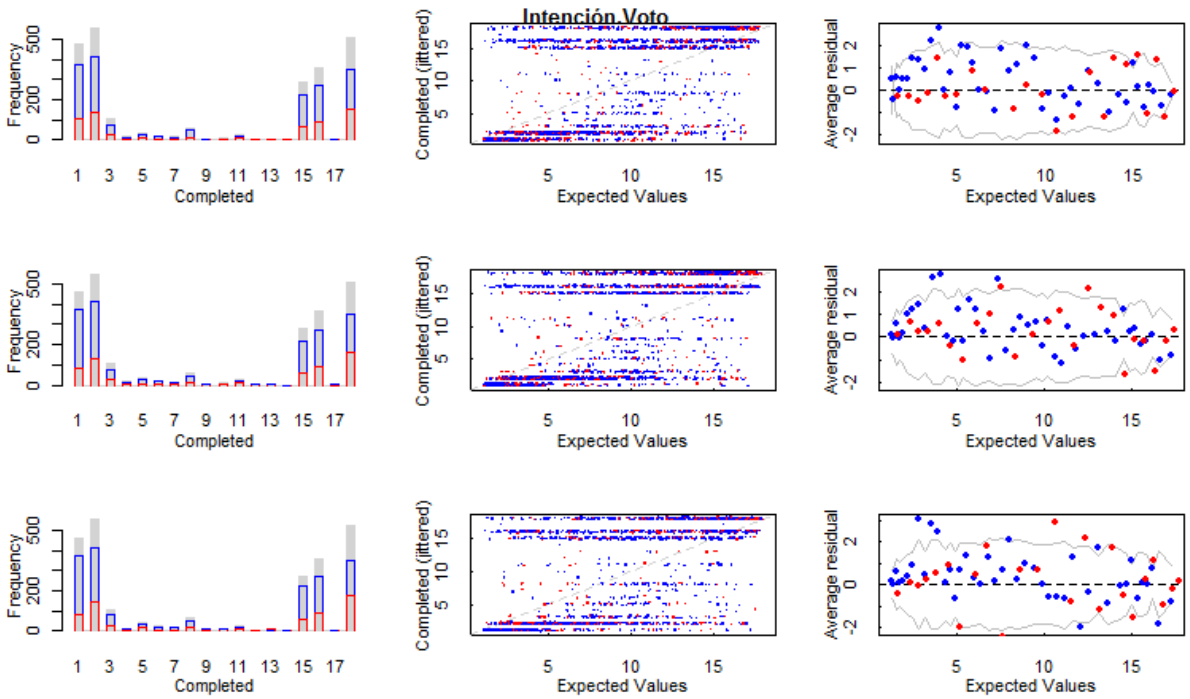


Figura 12. Gráfico de las imputaciones para intención del voto. Octubre 2015. Elaboración propia.

Se visualizan los gráficos para la variable de interés (intención de voto) de las tres primeras cadenas, que nos ayuda a tener un cierto sentido de la variabilidad de muestreo de las imputaciones. En la figura 12 en azul se observan los valores con los que contábamos inicialmente en la encuesta y en rojo, los valores imputados.

Se observa que para las tres primeras cadenas los resultados son muy similares. El gráfico de los errores es correcto, dado que no se visualizan formas específicas, sino que los puntos están dispersos, por lo que se considera que la imputación es correcta.

Finalmente, se pueden agrupar los conjuntos de datos imputados para estimar una regresión lineal descriptiva de la intención de voto.

2.3 Comparativa de los resultados

En este apartado vamos a realizar diferentes comparativas tanto con las estimaciones del CIS como con los resultados de las elecciones de diciembre de 2015.

En primer lugar, se ha creado una tabla con los porcentajes de intención de voto para las diferentes opciones. Se quiere comparar los resultados de las elecciones, que se han obtenido del periódico El País, la estimación del voto del CIS y el porcentaje de voto directo de la encuesta de octubre de 2015 (sin tener en cuenta los valores nulos, es decir, las respuestas NS/NC), con las estimaciones recogidas por la imputación:

Tabla 2. Comparativa de intención de voto y resultado de las elecciones 2015. Elaboración propia

	Encuesta	Imput 1	Imput 2	Elecciones 15	CIS
PP	23,03%	21,94%	22,34%	28,70%	29,10%
PSOE	25,49%	27,17%	26,61%	22,00%	25,30%
IU	4,79%	4,82%	5,02%	3,70%	4,70%
Podemos	13,51%	12,72%	12,95%	20,70%	10,80%
Cs	16,77%	17,03%	16,42%	13,90%	14,70%
otros	16,40%	16,32%	16,66%	11,00%	15,40%

Se observa a primera vista que el porcentaje de voto para PSOE, IU, Ciudadanos y otras opciones políticas es sobrestimado por el modelo. Sin embargo, para Podemos y PP se infraestima el porcentaje. Puede haber diversas causas.

A continuación, vamos a generar huecos aleatorios en la variable intención del voto en aquellos individuos que sí contestaron a la pregunta para ver el porcentaje de acierto de nuestra estimación. En total generamos 500 huecos aleatorios que guardamos en otra variable para comparar los resultados.

Tabla 3. Comparativa de aciertos y errores de la imputación

	PP	PSOE	IU	Otro	Podemos	Cs	No vota	Blanco	IMPUTADO
PP	88	10	0	1	1	8	0	0	**108**
PSOE	17	82	3	3	4	7	3	0	**119**
IU	1	6	10	3	1	0	1	0	**22**
Otro	3	6	2	31	8	8	12	1	**71**
Podemos	2	0	3	6	25	2	0	0	**38**
Cs	5	17	0	4	8	35	11	0	**80**
No vota	1	1	2	4	3	0	49	0	**60**
Blanco	0	0	0	1	0	1	0	0	**2**
TOTAL	**117**	**122**	**20**	**53**	**50**	**61**	**76**	**1**	**500**

El total de las filas corresponde a lo observado en la encuesta y el total de los votos imputados por opción política se encuentra en la columna *IMPUTADO*. Como se puede observar, en número total la imputación apenas difiere en unos cuantos valores, pero si nos fijamos en los aciertos por opción política podemos ver también que la imputación ha sido ciertamente efectiva. El porcentaje de acierto en el caso del PP llega hasta el 75.2%, en el caso del PSOE el 67.2%. El resto de opciones políticas están en torno al 50% y 65%, a excepción del voto en blanco que solo cuenta con un caso. Se observa que a menor número de individuos que dicen votar por una opción política, peor es la estimación. Aunque la estimación parece buena, podría haber otros modelos con más variables, más cadenas y más iteraciones que pudiesen conseguir mejores resultados.

Sin embargo, el objetivo del estudio es ver cómo afectan la utilización de encuestas previas en la imputación de datos, por lo que se procede a incluir en el modelo los datos de la encuesta de julio del 2015 de los encuestados que respondieron a la encuesta que respondieron a la pregunta de intención de voto.

Tabla 4. Comparativa de aciertos de la imputación con los datos de octubre y julio de 2015

	PP	PSOE	IU	Otro	Podemos	Cs	No Vota	Blanco	IMPUTADO
PP	95	7	0	3	0	6	4	0	115
PSOE	8	87	4	0	2	2	0	0	103
IU	0	4	12	1	2	0	0	0	19
Otro	5	5	1	39	9	7	6	1	73
Podemos	0	11	3	3	32	2	11	0	62
Cs	8	6	0	2	2	42	4	0	64
No Vota	1	2	0	5	3	2	50	0	63
Blanco	0	0	0	0	0	0	1	0	1
TOTAL	117	122	20	53	50	61	76	1	500

Se observa a primera vista una mejora en todos los partidos políticos y resto de opciones, a excepción del voto en blanco. Hay algunos casos que, de todas maneras, pueden ser preocupantes. Por ejemplo, en el caso de Podemos es normal que se puedan imputar sus votantes como votantes de IU (se presentaron juntas ambas fuerzas políticas en 2016) o del PSOE (que en muchos lugares de España gobiernan en coalición). Y en el caso de Ciudadanos es normal que sus votantes se puedan imputar como votantes de PP o PSOE, ya que en bastantes zonas de España son el apoyo de una o de otra fuerza política. Sin embargo, el caso de los votantes de Ciudadanos imputados como votantes de Podemos y viceversa no tiene fácil explicación.

Pero centrándonos en el caso de nuestro estudio hay que destacar que en el caso del PP hemos pasado de un 75.2 a un 81.2% de acierto y en el caso de PSOE hemos aumentado el acierto en un 5.1%.

Si tomamos el total general como:

$$\%\ General = \frac{Aciertos}{Total\ datos} * 100$$

Podemos decir que hemos pasado de 320 a 357 aciertos, lo que supone un aumento del 64% al 71.4%. Estas mejoras no son muy elevadas, pero no por eso dejan de ser mejoras. Es evidente que el número de variables utilizadas para el estudio, si tenemos en consideración todas las preguntas que realiza el CIS en sus encuestas, son una mínima parte, y el modelo empleado para estimar la intención de voto se queda corto. Sin embargo, este no es nuestro objetivo, recordemos que nuestro objetivo era valorar cómo afectaba la utilización de encuestas previas en la imputación y podemos decir que lo hace positivamente. Claro está que con modelos más completos se podrían obtener resultados mejores.

Una vez realizado el estudio con los datos de 2015 se va a repetir todo el procedimiento para verificar que con los datos de 2016 el resultado es similar.

2.4 Imputación con datos de julio y octubre de 2016

En esta sección se va a desarrollar el procedimiento detallado en los puntos 2.1 y 2.2 con los datos de las encuestas de 2016, para verificar que todo lo comentado anteriormente se cumple.

En primer lugar, empezamos con un pequeño resumen de las variables, un análisis gráfico que está introducido ya en el apartado 1 del trabajo, y un análisis de los datos faltantes en la encuesta de octubre de 2016:

	type	missing	method	model
Intención.Voto	continuous	487	ppd	linear
Sit.Econ	ordered-categorical	16	ppd	ologit
Sit.Polit	ordered-categorical	39	ppd	ologit
Prob.PP	continuous	146	ppd	linear
Prob.PSOE	continuous	157	ppd	linear
Prob.Podemos	continuous	148	ppd	linear
Prob.Cs	continuous	186	ppd	linear
Prob.IU	continuous	172	ppd	linear
Voto.Ant	continuous	231	ppd	linear
Sexo	binary	0	<NA>	<NA>
Edad	continuous	1	ppd	linear

Figura 13. Codificación previa de las variables del missing_data.frame.Octubre 2016. Elaboración propia

Como se comentó en la sección 2.1, es necesario ver cómo codifica el algoritmo las variables, para ver si hay algún error y corregirlo. Se observa en la figura 10 que las variables `Intención.Voto`,

`Prob.PP`, `Prob.PSOE`, `Prob.Podemos`, `Prob.Cs` y `Voto.Ant` están mal codificadas. Para ello, usamos la función `change()` y lo corregimos.

	type	missing	method	model
Intención.Voto	unordered-categorical	487	ppd	linear
Sit.Econ	ordered-categorical	16	ppd	ologit
Sit.Polit	ordered-categorical	39	ppd	ologit
Prob.PP	ordered-categorical	146	ppd	linear
Prob.PSOE	ordered-categorical	157	ppd	linear
Prob.Podemos	ordered-categorical	148	ppd	linear
Prob.Cs	ordered-categorical	186	ppd	linear
Prob.IU	ordered-categorical	172	ppd	linear
Voto.Ant	unordered-categorical	231	ppd	linear
Sexo	binary	0	<NA>	<NA>
Edad	continuous	1	ppd	linear

Figura 14. Recodificación de las variables. Octubre 2016. Elaboración propia.

Una vez categorizadas correctamente, se dibujan los patrones de datos faltantes para tener más información sobrela base de datos.

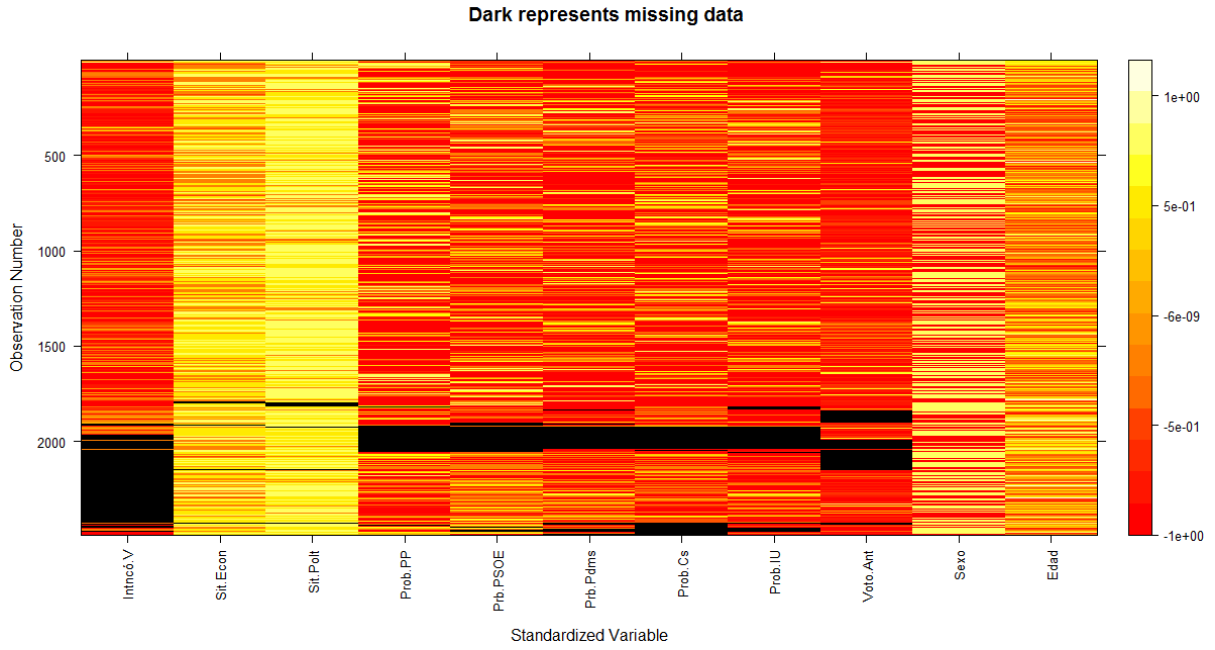


Figura 15. Representación de los valores faltantes. Octubre 2016. Elaboración propia.

Al igual que comentamos para la imputación en la encuesta de octubre de 2015, la variable que más datos faltantes presenta en la intención de voto. En segundo lugar, se encuentra la variable referente

al voto de las últimas elecciones. Por otro lado, la variable sexo, edad, situación económica y situación política son contestadas por casi todos los encuestados.

El siguiente paso es realizar la imputación y, una vez terminada, evaluar si son necesarias más iteraciones:

mean_Intención.Voto	mean_Sit.Econ	mean_Sit.Polit	mean_Prob.PP
1.0090454	0.9917258	1.0045359	1.0586505
mean_Prob.PSOE	**mean_Prob.Podemos**	**mean_Prob.Cs**	**mean_Prob.IU**
1.3270559	1.0339080	1.4040470	1.1838673
mean_Voto.Ant	**mean_Edad**	**sd_Intención.Voto**	**sd_Sit.Econ**
1.0097410	0.9976986	0.9919960	0.9906875
sd_Sit.Polit	**sd_Prob.PP**	**sd_Prob.PSOE**	**sd_Prob.Podemos**
1.0022871	1.0492495	1.2427134	1.0489601
sd_Prob.Cs	**sd_Prob.IU**	**sd_Voto.Ant**	**sd_Edad**
1.1539990	1.0865710	0.9911294	0.9978614

Figura 16. Convergencia medias y varianzas. Octubre 2016. Elaboración propia.

	chain:1	chain:2	chain:3	chain:4
Intención.Voto	6.430	6.297	6.392	6.388
Sit.Econ	3.879	3.879	3.877	3.878
Sit.Polit	4.455	4.456	4.457	4.454
Prob.PP	3.809	3.862	3.750	3.825
Prob.PSOE	4.007	4.082	4.111	3.941
Prob.Podemos	3.312	3.258	3.283	3.275
Prob.Cs	3.467	3.517	3.570	3.482
Prob.IU	3.086	3.090	3.115	3.141
Voto.Ant	3.818	3.819	3.820	3.776
Sexo	1.516	1.516	1.516	1.516
Edad	0.000	0.000	0.000	0.000
missing_Intención.Voto	0.196	0.196	0.196	0.196
missing_Sit.Econ	0.006	0.006	0.006	0.006
missing_Sit.Polit	0.016	0.016	0.016	0.016
missing_Prob.PP	0.059	0.059	0.059	0.059
missing_Prob.PSOE	0.063	0.063	0.063	0.063
missing_Prob.Podemos	0.059	0.059	0.059	0.059
missing_Prob.Cs	0.075	0.075	0.075	0.075
missing_Prob.IU	0.069	0.069	0.069	0.069
missing_Voto.Ant	0.093	0.093	0.093	0.093
missing_Edad	0.000	0.000	0.000	0.000

Figura 17. Medias de las variables imputadas. Octubre 2016. Elaboración propia

Se puede observar en la figura 15 y en la figura 16 que son necesarias más iteraciones para que el algoritmo llegue a converge, tantas como sean necesarias. En nuestro trabajo se realizan únicamente diez iteraciones más para mejorar la convergencia dado el gran coste computacional.

	[,1]	[,2]	[,3]	[,4]
Intención.Voto	6.378	6.241	6.232	6.382
Sit.Econ	3.880	3.876	3.879	3.878
Sit.Polit	4.457	4.456	4.455	4.458
Prob.PP	3.849	3.839	3.839	3.772
Prob.PSOE	4.051	3.945	3.984	3.939
Prob.Podemos	3.297	3.288	3.313	3.291
Prob.Cs	3.514	3.446	3.532	3.505
Prob.IU	3.095	3.145	3.148	3.132
Voto.Ant	3.827	3.841	3.854	3.804
Sexo	1.516	1.516	1.516	1.516
Edad	0.000	0.000	0.000	0.000
missing_Intención.Voto	0.196	0.196	0.196	0.196
missing_Sit.Econ	0.006	0.006	0.006	0.006
missing_Sit.Polit	0.016	0.016	0.016	0.016
missing_Prob.PP	0.059	0.059	0.059	0.059
missing_Prob.PSOE	0.063	0.063	0.063	0.063
missing_Prob.Podemos	0.059	0.059	0.059	0.059
missing_Prob.Cs	0.075	0.075	0.075	0.075
missing_Prob.IU	0.069	0.069	0.069	0.069
missing_Voto.Ant	0.093	0.093	0.093	0.093
missing_Edad	0.000	0.000	0.000	0.000

Figura 18. Medias y varianzas de las variables imputadas tras otras 10 iteraciones. Elaboración propia

mean_Intención.Voto	mean_Sit.Econ	mean_Sit.Polit	mean_Prob.PP
0.9882052	0.9907117	0.9888685	0.9900276
mean_Prob.PSOE	mean_Prob.Podemos	mean_Prob.Cs	mean_Prob.IU
0.9919660	0.9911145	0.9893084	0.9896903
mean_Voto.Ant	mean_Edad	sd_Intención.Voto	sd_Sit.Econ
0.9882402	0.9924734	0.9883157	0.9886672
sd_Sit.Polit	sd_Prob.PP	sd_Prob.PSOE	sd_Prob.Podemos
0.9883098	0.9913316	0.9923019	0.9915960
sd_Prob.Cs	sd_Prob.IU	sd_Voto.Ant	sd_Edad
0.9886294	0.9890901	0.9877572	0.9905253

Figura 19. Convergencia de medias y varianzas tras 10 iteraciones más. Elaboración propia.

Se observa que la convergencia es casi completa, aunque sería apropiado realizar más iteraciones. En este trabajo se da por válido ya que el coste computacional es muy elevado y el objetivo principal no es estimar el voto.

A continuación, hay que realizar un análisis gráfico de las imputaciones para validar fehacientemente. El histograma de los datos de la encuesta (en azul) y de los imputados (en rojo), una comparación de los datos completos y esperados y el gráfico de los errores.

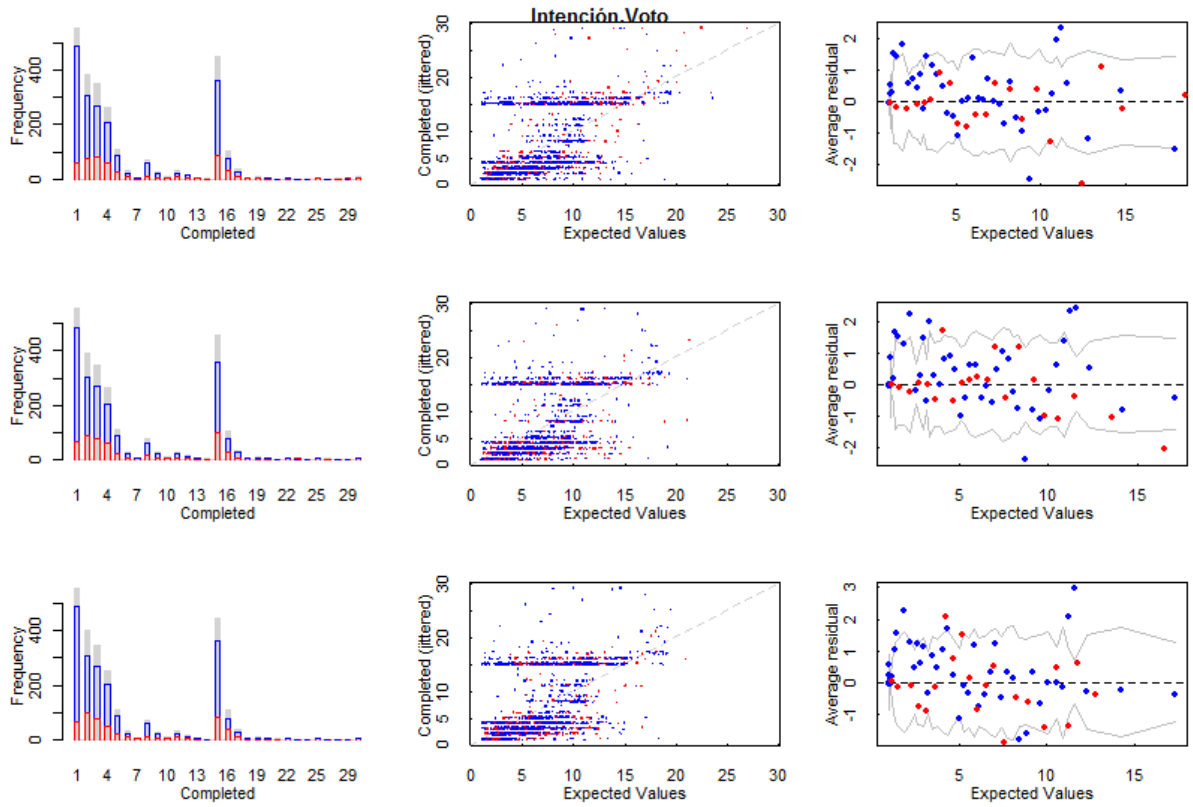


Figura 20. Gráficos de las imputaciones para Intención del Voto. Octubre 2016. Elaboración propia

En la figura 20 se visualizan solo las imputaciones para la variable intención del voto, pero R dibuja los gráficos para todas las variables del modelo. En este estudio dan una información similar y para no sobrecargar el trabajo con demasiadas imágenes nos centramos en la variable de interés, que además es la que más datos faltantes tiene y, por tanto, la que más imputaciones necesita y más información nos da.

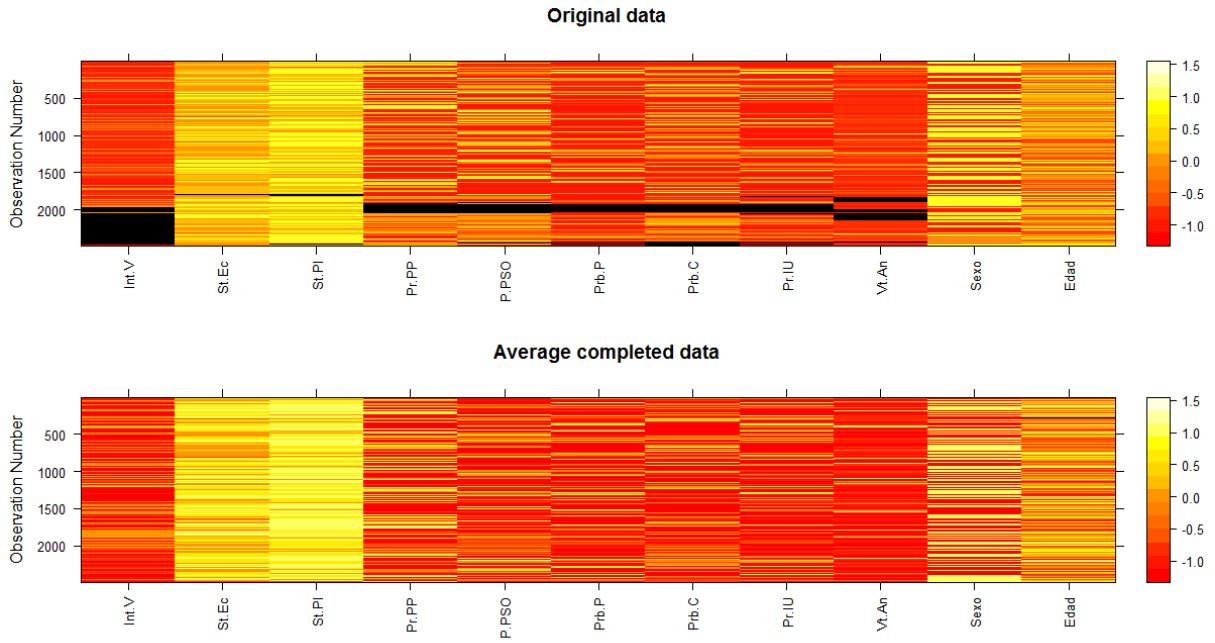


Figura 21. Comparativa de los datos faltantes antes y después de la imputación. Octubre 2016. Elaboración propia

Como se observa en la figura 21, el algoritmo ha concluido y se han imputado todos los valores faltantes de cada una de las variables, por lo que solo queda crear el modelo de regresión, guardar los datos y evaluar los resultados. Posteriormente se utilizará la encuesta de julio de 2016 para comprobar el porcentaje de acierto del modelo.

Según las imputaciones, el 18.1% de los encuestados no votaría en unas supuestas terceras elecciones. Esto es algo entendible pues, como ya hemos comentado, la participación en unos terceros comicios en 2016 o 2017 hubiera caído en un gran porcentaje.

Excluyendo esos encuestados y tomando solo el total de votos válidos, calculamos los porcentajes de intención de voto para cada fuerza política. Además, tomamos los resultados de las elecciones de 2016 y las estimaciones del CIS para octubre de 2016. Se ha considerado la suma entre Podemos e IU para el cálculo de porcentajes dado que se presentaron juntos a dichas elecciones.

Tabla 5. Comparación porcentajes voto. Octubre 2016. Elaboración propia

	Encuesta	IMPUT 1	IMPUT 2	Elecciones 2016	CIS
PP	29,70%	28,9%	29,1%	33,03%	34,50%
PSOE	18,62%	19,6%	19,9%	22,66%	17,00%
Podemos+IU	21,67%	21,9%	21,7%	21,10%	21,80%
C's	12,48%	12,9%	12,8%	13,10%	12,80%
Blanco	4,63%	4,2%	4,5%	0,75%	3,80%
Otros	12,9%	12,5%	12,0%	9,36%	10,10%

Como se puede observar en la tabla 5, las imputaciones se aproximan bastante al resultado de las elecciones y a las estimaciones del CIS. Nuestro modelo parece adecuado, aunque el objetivo no es comparar estimaciones y resultados, sino ver que las imputaciones se han realizado correctamente.

En comparación con la tabla 2, los resultados de la imputación parecen mejores. Esto puede ser debido a que, como ya se ha comentado a lo largo del estudio, en épocas preelectorales el voto oculto es mucho mayor y los encuestados son más reacios a la hora de confesar su intención de voto.

Para finalizar, hay que comprobar el porcentaje de acierto del modelo de imputaciones, por lo que se seleccionarán los datos de los encuestados que respondieron a la intención de voto y, generando aleatoriamente valores perdidos en 500 de dichos encuestados utilizando Excel (ver Anexo), se compararán los valores reales con los imputados por el modelo.

En la encuesta de octubre de 2016 encontramos 2004 individuos que dieron respuesta a la variable de interés, por lo que se va a trabajar en torno a un 25% de datos faltantes. Se realiza todo el procedimiento descrito anteriormente, con el mismo análisis gráfico y comprobando la convergencia del modelo. Una vez que todo está verificado, se guardan los datos imputados y se realizan las comparaciones con los correspondientes datos completos de los que disponíamos y habíamos seleccionado aleatoriamente.

```
                                type missing method  model
Intención.Voto unordered-categorical     500    ppd mlogit
Sit.Econ         ordered-categorical      11    ppd ologit
Sit.Polit        ordered-categorical      27    ppd ologit
Prob.PP          ordered-categorical      57    ppd ologit
Prob.PSOE        ordered-categorical      64    ppd ologit
Prob.Podemos     ordered-categorical      64    ppd ologit
Prob.Cs          ordered-categorical      85    ppd ologit
Prob.IU          ordered-categorical      79    ppd ologit
Voto.Ant       unordered-categorical      74    ppd mlogit
Sexo                         binary       0   <NA>   <NA>
Edad                      continuous      0   <NA>   <NA>
```

Figura 13. Codificación de variables. Julio y octubre 2016. Elaboración propia

Una vez codificadas las variables correctamente se procede a la realización de las imputaciones.

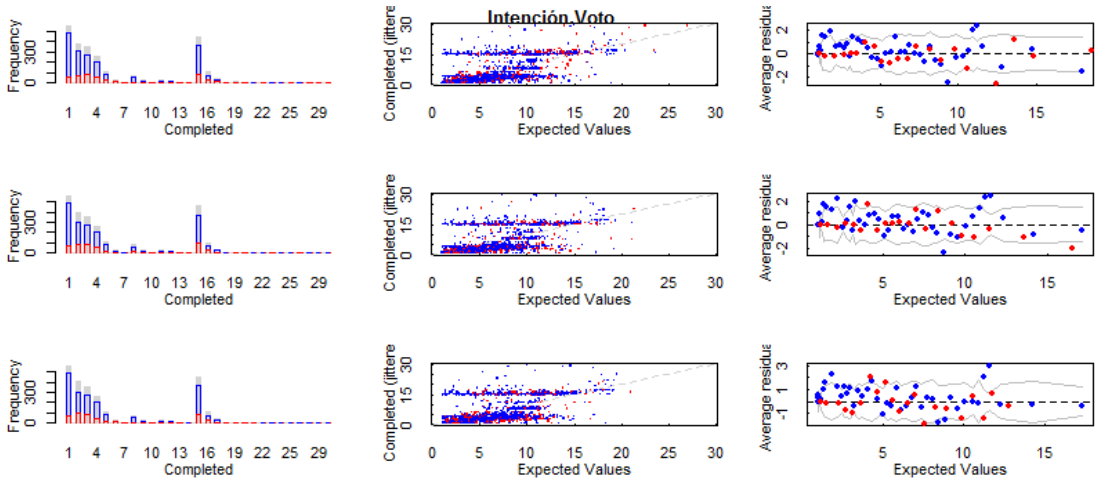


Figura 22. Gráficos imputaciones Intención Voto para comparación. Julio y octubre 2016. Elaboración propia.

La figura 22 muestra que las imputaciones se han realizado correctamente, por lo que podemos pasar a realizar las comparaciones.

Tabla 6. Comparativa de la estimación en Intención de Voto. Encuesta octubre 2016. Elaboración propia

	PP	PSOE	Otro	Podemos+IU	Cs	No Vota	Blanco	IMPUTADO
PP	87	10	4	4	14	4	1	124
PSOE	16	68	2	11	3	3	0	103
Otro	4	4	19	7	8	6	1	49
Podemos+IU	6	15	8	52	4	11	0	96
Cs	10	6	4	3	40	6	0	69
No Vota	6	4	7	5	5	29	1	57
Blanco	1	0	0	0	0	1	0	2
TOTAL	130	107	44	82	74	60	3	500

En la tabla 6 el total de las filas corresponde a lo observado en la encuesta y el total de los votos imputados por opción política se encuentra en la columna *IMPUTADO*. Volvemos a observar, como ocurría en la tabla 3 que, en cuanto a número de votos por partido político, la imputación está muy cerca de lo observado. Sin embargo, tomando el porcentaje de acierto entre datos observados e imputados, el resultado es peor que en el año 2015. Para PP, PSOE y la confluencia entre Podemos e Izquierda Unida (Unidos Podemos) el porcentaje de acierto está entre el 63 y el 67%, pero el resto de opciones rondan el 45% y el 55%. Se ha comentado en apartados anteriores que, en el periodo poselectoral, la incertidumbre es menor, pero, por otro lado, la idea de votar en unas hipotéticas terceras elecciones (como hacía suponer la encuesta del CIS de octubre del 2016) no era recibida de buen grado por gran parte del electorado.

Esta puede ser una de las razones por la obtención de un resultado aparentemente peor. Otra razón puede ser que la muestra tomada aleatoriamente no sea del todo buena e introduzca sesgo en las estimaciones.

Para concluir, seleccionamos los datos de julio de 2016 para unirlos a los que se han tratado y ver si se produce la mejora esperada en las estimaciones. El procedimiento ya se ha explicado en el apartado 2.3 y también se recoge en el anexo.

Tabla 7. Comparativa de la estimación de voto. Julio y octubre 2016. Elaboración propia

	PP	PSOE	Otro	Podemos+IU	Cs	No Vota	Blanco	IMPUTADO
PP	105	3	5	3	7	2	0	125
PSOE	8	83	3	7	3	2	0	106
Otro	3	2	26	1	3	6	1	42
Podemos+IU	2	12	3	67	5	8	1	98
Cs	10	3	2	0	51	4	0	70
No Vota	1	4	5	4	5	37	0	56
Blanco	1	0	0	0	0	1	1	3
TOTAL	130	107	44	82	74	60	3	500

La tabla 7 muestra una notable mejora en la imputación. Hemos pasado de un 67% de acierto en el PP a un 80.7%. En el caso del PSOE de un 63.6% a un 77.6%. En el caso de Podemos e IU, el porcentaje de acierto es del 81.7%. Las mejoras son notables en este caso. Cierto es que en julio de 2016 apenas había incertidumbre dado que se habían celebrado las elecciones apenas unas semanas antes.

Se puede concluir, por tanto, que la utilización de encuestas previas mejora la estimación de la intención del voto. Puede ser, entre otras cosas, que al aumentar el número de respuestas y no el de datos faltantes influya en el resultado, pero no hay motivo para aumentar dichos valores nulos porque en las encuestas el porcentaje de no respuesta es el que es y es perfectamente válido añadir más datos completos de otras encuestas anteriores.

3. CONCLUSIONES

Estimar la intención del voto de los encuestados que no quieren responder a esa pregunta en las encuestas, es un objetivo primordial en muchas empresas y organismos públicos y privados, más aún en periodos preelectorales. En este trabajo se ha intentado realizar esa tarea utilizando el método de imputación múltiple en el software R con los datos publicados por el CIS en octubre de 2015 y 2016, pero, sin embargo, el objetivo principal era ver si afecta o no y en qué cuantía, utilizar los datos de encuestas de periodos anteriores al tratado. Para ello hemos utilizado las encuestas de julio de 2015 y 2016. En este aspecto tenemos que volver a destacar que se han escogido esas fechas dado que correspondían a periodos preelectorales y poselectorales donde la variabilidad y la sensibilidad del encuestado a la hora de responder era mayor.

Hemos podido comprobar que el porcentaje de no respuestas en periodos preelectorales es mayor (35% y 25% en las encuestas de 2015 frente a 15.9% y 19.6% en 2016) y hemos intentado analizar las posibles causas (mucha indecisión, "voto oculto", …). Conjuntamente se ha explicado que en el periodo poselectoral de 2016 no se daba tanto el problema de la no respuesta, sino que se mostraba un aumento en la intención de los encuestados de no votar, en este caso probablemente más agudizado por la celebración de segundas elecciones en junio de 2016.

Con todas esas premisas y contando con un modelo sencillo con pocas variables para reducir lo máxima posible el coste computacional, pero sin desvirtuar las estimaciones demasiado, se ha procedido al desarrollo del trabajo.

Hemos observado un mejor resultado para las encuestas de 2016 (una vez incluida la encuesta de julio de ese año), probablemente por el contexto político del momento. A pesar de ser un periodo convulso dadas las posibilidades que había de unas terceras elecciones, la indecisión del encuestado era menor, ya que sabía lo que había votado hacía unos meses antes. También 2015 era un periodo con mayor incertidumbre no solo por ser un periodo preelectoral, sino por el gran crecimiento de dos fuerzas políticas que parecían desafiar al bipartidismo y que hicieron surgir muchas dudas entre los votantes.

Con todo ello y para concluir, hay que destacar que hemos obtenido una conclusión acerca del principal objetivo del estudio. Se ha comprobado que, para los dos periodos estudiados, utilizar encuestas realizadas con anterioridad mejora el resultado de las imputaciones en intención de voto. Cabe señalar que esa mejora ha sido más significativa para 2016, no solo por la menor incertidumbre del momento, como ya hemos comentado, sino porque partidos como Podemos y Ciudadanos eran mucho más conocidos y, además, a diferencia de las encuestas de 2015, aparecían en la pregunta sobre voto anterior, por lo que la información era más completa.

4. BIBLIOGRAFÍA

ALFARO, R y FUENZALIDA, M., 2009. Imputación Múltiple en Encuestas Microeconómicas. En: *Scielo,* no 46, pp.273-288. ISSN 0717-6821

Barómetros. CIS, 2016 [consulta: 20/07/2017]. Disponible en: http://www.cis.es/cis/opencm/ES/11_barometros/depositados.jsp

Centro de Investigaciones Sociológicas. CIS, 2017 [consulta: 23 marzo 2017]. Disponible en: http://www.cis.es/cis/opencms/ES/index.html

CILLEROS CONDE, R., ESCOBAR MERCADO, M. y RIVIÉRE GÓMEZ, J. *Los pronósticos electorales con encuestas.* No 6. Madrid: Centro de Investigaciones Sociológicas, 2014, pp. 39-51.

Destino: Moncloa. Cadena Ser, 2016. [Consulta: 20 julio 2017]. Disponible en: http://cadenaser.com/ser/2016/09/18/politica/1474219427_671550.html

Elecciones Generales 2015. El País, 2015 [consulta: 18 marzo 2017]. Disponible en: http://resultados.elpais.com/elecciones/2015/generales/congreso/

GELMAN, A., HILL, J., YAJIMA, M. y YU-SUNG, S., 2011. Multiple Imputation with Diagnostics. En: *Journal of Statistical Software,* no 45, pp. 3-25.

GELMAN, A., GOODRICH, B., HILL, J., KROPKO, J., PITTAU, M., YAJUAN, S. y YAJIMA, M., 2015. Missing Data Imputation and Model Checking. En: *cran.r-project, no 1, pp.3-35.*

GROVES R. M. *Survey Errors and Survey Costs*. New Jersey: Wiley-Interscience, 2004, pp. 6-15. ISBN 0-471-67851-1

HERNÁNDEZ VELASCO, I. ¿Por qué se equivocan las encuestas? En: *El Mundo*. 24 noviembre 2016 [consulta: 18 julio 2017] Disponible en: http://www.elmundo.es/sociedad/2016/11/24/5835b946268e3eab498b45af.html

Imputing missing data with R; MICE package. R-bloggers, 2015 [consulta: 7 mayo 2017]. Disponible en: https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

MARIN DIAZARAQUE J. M. y RESTREPO ESTRADA M. I., 2012. Imputación de ingresos en la Gran Encuesta Integrada de Hogares (geih) de 2010. En: *Universidad de los Andes*, no 70, pp. 227-229.

Missing Data: Listwise vs Pairwise. StatisticsSolutions ©2017 [consulta: 18 julio 2017]. Disponible en: http://www.statisticssolutions.com/missing-data-listwise-vs-pairwise/

¿Qué es metroscopia? Metroscopia, ©2011 [consulta: 18 julio 2017]. Disponible en: http://metroscopia.org/quienessomos/

ObSERvatorio. MyWord, ©2014 [consulta: 18 julio 2017]. Disponible en: http://myword.es/actividad/el-observatorio-de-la-cadena-ser/

OTERO GARCÍA, D., 2011. Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo. En: *Universidad de Santiago de Compostela*, pp.11-31.

PEUGH, J. L. y ENDERS, C. K., 2004. Missing data in educational research: A review of reporting practices and suggestions for improvement. En: *Review of Educational Research*, no 74, pp. 525-556.

ROYUELA VICENTE, A., 2014. Desarrollo y aplicación de modelos pronósticos en patología lumbar. En: *Biblos-e Archivo,* pp.38-43.

RUBIN, D. B. *Inference and missing data.* New Jersey: Biometrika, 1976, no 3, pp. 583-590

RUBIN, D. B. Mu*ltiple imputation for nonresponse in surveys*. New York, 1987, pp. 1-11.

Sondeos. Electomanía, ©2015 [consulta: 18 julio 2017]. Disponible en: http://electomania.es/category/sondeos/

WILKS, S. 1932. Moments and distributions of estimates of population parameters from fragmentary simple*,* En: *Annals of Mathematical Statistics, B, pp. 163-195.*

1. Imputaciones

En este apartado se muestra el código utilizado en el software R para la realización de las imputaciones.

#Instalación de la librería mi.

```
install.packages("mi")
```

```
library(mi)
```

#Se cargan los datos, que están en formato csv2.

#Realizamos un breve resumen acerca de las variables.

```
x=read.csv2(choose.files(),header=TRUE)
```

```
n=dim (x)[1]
```

```
p=dim (x)[2]
```

```
head(x)
```

```
summary(x)
```

#Transformamos los datos a missing_data.frame (versión de data.frame necesaria para trabajar con datos faltantes, que incluye metadatos de las variables).

```
mdf <- missing_data.frame(x)
```

#Vemos un pequeño resumen de cómo ha codificado las variables el algoritmo

```
show (mdf)
```

#Cambiamos aquellas variables mal codificadas. Aparecen como continuas y realmente son ordenadas categóricas

```
mdf<-change(mdf, y =c("Prob.PP", "Prob.PSOE", "Prob.IU",
"Prob.Podemos", "Prob.Cs", "Intención.Voto", "Voto.Ant"),
what="type", to = c("ord","ord","ord","ord","ord","un","un"))
```

#Vemos un resumen del missing_data.frame creado

```
summary(mdf)
```

#Visualizamos un gráfico para entender el patrón de datos faltantes

```
image(mdf)
```

#Dibujamos un histograma de las variables, para ello hay que cambiar el tamaño de la ventana

```
par(mar=c(1,1,1,1))
```

```
hist (mdf)
```

#Para ahorrar espacio eliminamos los objetos innecesarios

```
rm(x)
```

#Realizamos las imputaciones, tienen un gran coste computacional.

```
imputations <- mi(mdf, max.minutes = 10)
```

#Realizamos las imputaciones, tienen un gran coste computacional. Se toman 30 iteraciones, 4 cadenas y como máximo 10 minutos para realizar las imputaciones

```
imputations <- mi(mdf, n.iter = 30, n.chains = 4, max.minutes = 10)
```

#Vemos el resultado obtenido, para comprobar que está bien hecha la imputación y ha convergido el algoritmo

```
show(imputations)
```

```
round(mipply(imputations, mean, to.matrix = TRUE), 3)
```

```
Rhats(imputations)
```

#Aumentamos el número de iteraciones hasta la convergencia. En este caso se realizan 5 iteracciones extra a las 30 ya realizadas. Es válido hacerlo de esta manera y disminuye el coste computacional

```
imputations <- mi(imputations, n.iter = 5)
```

#Dibujamos las imputaciones, para verificar que el proceso ha sido correcto

```
plot(imputations)
```

```
image (imputations)
```

#Realizamos la regresión del modelo y las imputaciones

```
analysis <- pool(Intención.Voto ~ Sit.Econ+ Sit.Polit+ Prob.PP+
Prob.PSOE+ Prob.IU+ Prob.Podemos+ Prob.Cs+ Voto.Ant+ Sexo+ Edad,data
= imputations, m = 4)
```

```
display(analysis)
```

#Guardamos los datos imputados tras las 4 cadenas y todas las iteraciones en dos conjuntos de datos imputados para comparar a posteriori

```
dfs <- complete(imputations, m = 2)
```

2. Evaluación

#El siguiente paso es evaluar las imputaciones, para ello se seleccionan aquellas filas en las que los encuestados sí respondieron a la variable intención de voto. Este paso previo se realiza con la ayuda de Excel.

```
X1=read.csv2(choose.files(),header=TRUE)
```

#Se seleccionan al azar 500 filas del total (n) de ellas para realizar la imputación. Se guarda en la variable `t` para garantizar que se seleccionan los mismos individuos.

```
t<-sample(1:n, 500)
```

#Se guardan en me esas 500 respuestas para después compararlas con las imputaciones

```
me<-x1[t,]
```

#En mp se guardan el resto de datos para realizar la imputación

```
mp<-x1[-t,]
```

#Quitamos intención del voto en me

```
me[,"Intención.Voto"]=NA
```

#Juntamos me y mp en la misma matriz

```
xp=rbind(me,mp)
```

#Convertimos la matriz en un missing_data.frame

```
mdfxp<-missing_data.frame(xp)
```

#Vemos la codificación de las variables y cambiamos las que están mal codificadas

```
show (mdfxp)
```

```
mdfxp<-change(mdfxp, y =c("Prob.PP", "Prob.PSOE", "Prob.IU",
"Prob.Podemos", "Prob.Cs", "Intención.Voto", "Voto.Ant"),
what="type", to = c("ord","ord","ord","ord","ord","un","un"))
```

#Vemos un resumen y graficamos

```
summary(mdfxp)

image(mdfxp)

par(mar=c(1,1,1,1))

hist (mdf)
```

#Se realizan las imputaciones, vemos si se ha hecho bien y añadimos iteraciones hasta la convergencia

```
imputations1 <- mi(mdfxp, max.minutes=10)

show(imputations)

round(mipply(imputations, mean, to.matrix = TRUE), 3)
```

#Guardamos los datos imputados para proceder a la comparación

```
dfsxp <- complete(imputations, m = 2)

write.table(dfsxp, file = "CompOct15.csv")
```

#Una vez realizado el proceso anterior hay que repetirlo desde el punto 2 con los datos de Octubre y los completos de Julio.

#Tras acabar con los datos del 2015 se repite el mismo procedimiento para las encuestas de 2016.