Universidad
Carlos III de Madrid

Autora: NEREA SÁNCHEZ JARA

Tutor: ANDRÉS M. ALONSO FERNÁNDEZ

IMPUTACIÓN DE DATOS FALTANTES EN ENCUESTAS DE INTENCIÓN DE VOTO

2015

UNIVERSIDAD CARLOS III (MADRID)

Grado ESTADÍSTICA Y EMPRESA

En memoria de Inocencio.

"Hija, ¿Cómo se llamaba lo que estudiabas? -Estadística y Empresa, abuelo"

Tabla de contenido

Ilustraciones

Capítulo Primero

1. INTRODUCCIÓN

En los estudios electorales la estimación del voto es, a menudo, el objetivo principal. Según la definición de Lewis-Beck (Lewis-Beck, 2005) pronosticar una elección significa anticipar el resultado antes de que ocurra. Si partimos de esta base el pronóstico electoral tiene la finalidad de ajustar su estimación lo máximo posible a los resultados. Sin embargo, en muchas ocasiones el valor estimado y el real son diferentes. La Real Academia Española define pronosticar como el conocimiento anticipado de lo que sucederá en un futuro a través de ciertos indicios. En este sentido, en este trabajo se pronosticará la intención de voto de los encuestados en unas futuras elecciones a través de diferentes variables. Creemos que la no respuesta es una de las causas del sesgo que se puede producir en las diferentes estimaciones del voto. Estas no respuestas son el objetivo principal de nuestras estimaciones. Por eso este trabajo detallará de manera exhaustiva los porcentajes de no respuesta o también llamados datos faltantes de las variables de interés, esta información se encuentra en la siguiente sección en la que se realizará un análisis descriptivo de todas ellas.

Existen varios organismos y/o empresas que se dedican a realizar encuestas de intención de voto, para su posterior análisis y estimación. Nosotros realizaremos este proyecto con los datos disponibles en el Centro de Investigaciones Sociológicas, en adelante el CIS. Este organismo de carácter público realiza las encuestas con ámbito nacional y a población española de ambos sexos de 18 años o más. Se debe tener en cuenta que los cuestionarios se han aplicado mediante entrevista personal en los domicilios de los encuestados, dichas entrevistas se han realizado en 239 municipios y 50 provincias, con un tamaño muestral de 2.480 individuos en la encuesta de Octubre de 2014 y 2.481 en la del mes de Enero de 2015, que serán las encuestas que estudiaremos en este proyecto. En cuanto al procedimiento de muestreo utilizado en estas encuestas se detalla lo siguiente:

> *"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, 2014).

Además del CIS existe un gran número de empresas que realizan este tipo de encuestas pero observamos que pocas son transparentes, es decir, no informan de cómo realizan las estimaciones, ni si las entrevistan las realizan ellos, o por lo contrario contratan a otras empresas o utilizan datos de organismos oficiales como el CIS o el Censo electoral de los Españoles Residentes ausentes que viven en el extranjero (CERA).

Algunas de las empresas que realizan estos trabajos y que hemos considerado son, entre muchas otras, las siguientes:

MyWord-Social and Market Research, tiene un proyecto junto a la Cadena Ser llamado el obSERvatorio que realiza estimaciones del voto utilizando muestras de menor tamaño que el CIS (1.000 individuos) y realizando las entrevistas online a partir de un panel de captación activa (sólo por invitación), dichos resultados según informa la propia empresa (myWord) *"son proyecciones electorales generadas a partir de un modelo de estimación de voto que se aplica a los datos de intención de voto de la encuesta".* Pero no aporta mayor información sobre el proceso de estimación. En cuanto a la información de no respuesta referida a la pregunta de intención de voto en unas supuestas elecciones electorales, en la encuesta que realizaron en Abril de 2015 un 20% de los encuestados contestaron no sabe/no contesta.

Celeste-tel. Investigación Sociológica, esta entidad realiza entrevistas telefónicas para recopilar sus datos, realiza muestras representativas de población, según provincia, ciudad, pueblo... Por ejemplo en el barómetro que realizó para Junio 2015 el tamaño de la muestra es de 1.100 entrevistas, con un *"Muestreo: Polietápico. Estratificado por conglomerados. Unidad primaria de muestreo: municipio. Selección aleatoria proporcional. Unidad secundaria de muestreo: sección electoral. Selección aleatoria proporcional. Unidad terciaria de muestreo: Ciudadanos y ciudadanas. Selección aleatoria por cuotas de edad, sexo y recuerdo de voto. Afijación no proporcional."* (Celeste-tel) En esa encuesta podemos observar que el 34.1% de las personas que han sido entrevistadas no responden a la pregunta directa de intención de voto.

Metroscopia, realiza diferentes encuestas para otras entidades. Por ejemplo, en un trabajo para el País en el que el "*sondeo fue realizado mediante entrevista telefónica, con un tamaño muestral de 2.000, estratificadas por la intersección hábitat/comunidad autónoma y distribuidas de manera proporcional al total de cada región, con cuotas de sexo y edad aplicadas a la unidad última*" (El País, 2015). En esa encuesta el 38.5% de las personas se abstienen de contestar al preguntarles por el posible sentido de su voto en unas inminentes elecciones, estos resultados fueron publicados el 6 de Junio de 2015 y pueden ser observados en la figura 1.

Figura 1. Fuente: El País.

Encuestamos es otra de las empresas que realiza encuestas e informes sobre la intención de voto, en esta ocasión no tenemos información del tamaño de la muestra ni que modelos utiliza, tan sólo podemos observar que existe un 15.2% de no respuesta en la encuesta realizada el 23 de enero de 2015. (Encuestamos, 2015).

Sigmados es una empresa que realiza encuestas y estimaciones de la intención de voto en España para otras entidades tanto públicas como privadas. Como nos ocurría con algunas de las empresas anteriores no nos muestra información de cómo realizan las estimaciones, como se obtienen los datos ni el tratamiento y porcentaje de no respuesta.

Nuestro trabajo surge con la motivación de obtener modelos estadísticos para la imputación de valores faltantes, en las encuestas realizadas en los meses de Octubre 2014 y Enero 2015 por el CIS. A partir de estas imputaciones nuestro objetivo principal es la estimación de la intención de voto, así como variables relacionadas tipo socioeconómico, político y demográfico. Además, compararemos los resultados obtenidos con las estimaciones que propone el CIS, notar que sus modelos de estimación no están especificados claramente. Por último, estudiaremos los diferentes patrones de no respuesta y realizaremos una evaluación interna del proceso llevado a cabo. Todo esto será explicado y detallado en las diferentes secciones de las que consta el trabajo. En definitiva el objetivo de este trabajo es la implementación de un modelo transparente, sencillo y utilizarlo con datos del CIS.

En la siguiente sección, como ya hemos mencionado anteriormente, se realizará un análisis descriptivo de las variables que vamos a estudiar. En el capítulo segundo nos centraremos en la metodología del proyecto, así en la segunda sección de dicho capítulo realizaremos el caso práctico, en el que detallaremos las imputaciones, realizaremos la evaluación interna citada con anterioridad y compararemos los resultados de las estimaciones con los resultados del CIS. Para finalizar, detallaremos todas las conclusiones a las que se ha llegado. El software empleado ha sido `R` (R proyect) y la librería específica `mi` (A Community Site for R), para la creación de programas de elaboración propia con el fin de realizar las imputaciones de los datos y las evaluaciones de las imputaciones, además de Excel para la creación de tablas con un formato más adecuado y legible.

2. ANÁLISIS DESCRIPTIVO

En esta sección se realizará un análisis descriptivo tanto de las variables de interés así como de las frecuencia de valores faltantes de todas ellas.

La variable principal a estudiar es la pregunta referida a la intención de voto del encuestado, que en las encuestas está formulada de la siguiente manera:

Suponiendo que mañana se celebrasen elecciones generales, es decir, al Parlamento español,

¿A qué partido votaría Ud.?

Las respuestas a esta pregunta son los diferentes partidos o coaliciones políticas del Estado Español, además de la posibilidad de votar en blanco, voto nulo y no votar. Para este proyecto hemos visto conveniente recodificar dichas respuestas en los siguientes grupos:

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

2 si su intención de voto fuera *PSOE* (Partido Socialista Obrero Español).

3 para intención de voto a *IU* (Izquierda Unida).

4 para intención de voto a *UPyD* (Unión Progreso y Democracia).

5 para intención de voto a *Podemos*.

6 para intención de voto a *Ciudadanos* (Partido de la Ciudadanía).

7 la hemos llamado *Ninguno*, e incluye aquellas respuestas de votar nulo o no votar.

8 para votaciones en *Blanco*.

9 para *Otros* partidos, aquí están incluidos aquellos partidos regionales como pueden ser UPN (Unión del Pueblo Navarro) que sólo se puede votar en Navarra o CIU (Convergencia y Unión) sólo en Cataluña, entre otros.

Por último también están las opciones de *No sabe* y *No contestan* que son los *valores faltantes* de las encuestas, esta no respuesta es el objetivo principal en este proyecto. En este apartado vamos a estudiar en detalle la frecuencia de valores faltantes de esta variable así como de las variables que explicamos a continuación.

Para la realización de las imputaciones y para poder comparar entre la encuesta de Enero 2015 y la de Octubre 2014, decidimos seleccionar las siguientes variables:

- ✓ *Valoración de la situación Económica.* Toma valores del 1 al 5 siendo 1.Muy Buena, 2.Buena, 3.Regular, 4.Mala y 5.Muy mala.
- ✓ *Valoración de la situación Política.* Los valores que toma son similares a la pregunta anterior.
- ✓ *Probabilidad de voto a los diferentes grupos políticos*. Sabiendo que el 0 significa que "con toda seguridad no le votaría nunca" y el 10 significa que "con toda seguridad, le votaría siempre". Hemos realizado la siguiente transformación: definimos p= (respuesta del encuestado que está entre 1 y 10)/10 para tener valores entre 0 y 1 valores propios de una probabilidad. Después realizamos P^*=log (1+p).
- ✓ *Evaluación del Gobierno.* Tiene la misma escala que las preguntas referentes a la valoración de la situación política y económica.
- ✓ *Evaluación de la Oposición.* Similar a la pregunta anterior.
- ✓ *Simpatía* por un determinado partido o coalición. La codificación es la misma que para la variable de intención de voto en unas supuestas elecciones.
- ✓ *Ideología* del encuestado siendo 1 de izquierda hasta el 10 que indica de derecha.
- ✓ *Recuerdo del voto* en elecciones anteriores. Esta codificación es la misma que para simpatía o intención del voto, pero debemos tener en cuenta que en las anteriores elecciones nacionales, los partidos políticos Podemos y Ciudadanos no existían o no concurrieron. Podemos surgió en Enero de 2014

y Ciudadanos se fundó en 2006 pero su expansión nacional tuvo lugar en el año 2014.

- ✓ *Sexo*, 1.Hombre 0.Mujer.
- ✓ *Edad*, a partir de los 18 años.

En primer lugar, observaremos y estudiaremos la variable ***intención de voto***. Mostramos un gráfico comparativo de las respuestas en ambas encuestas, notar que ya están recodificadas cómo hemos explicado anteriormente:

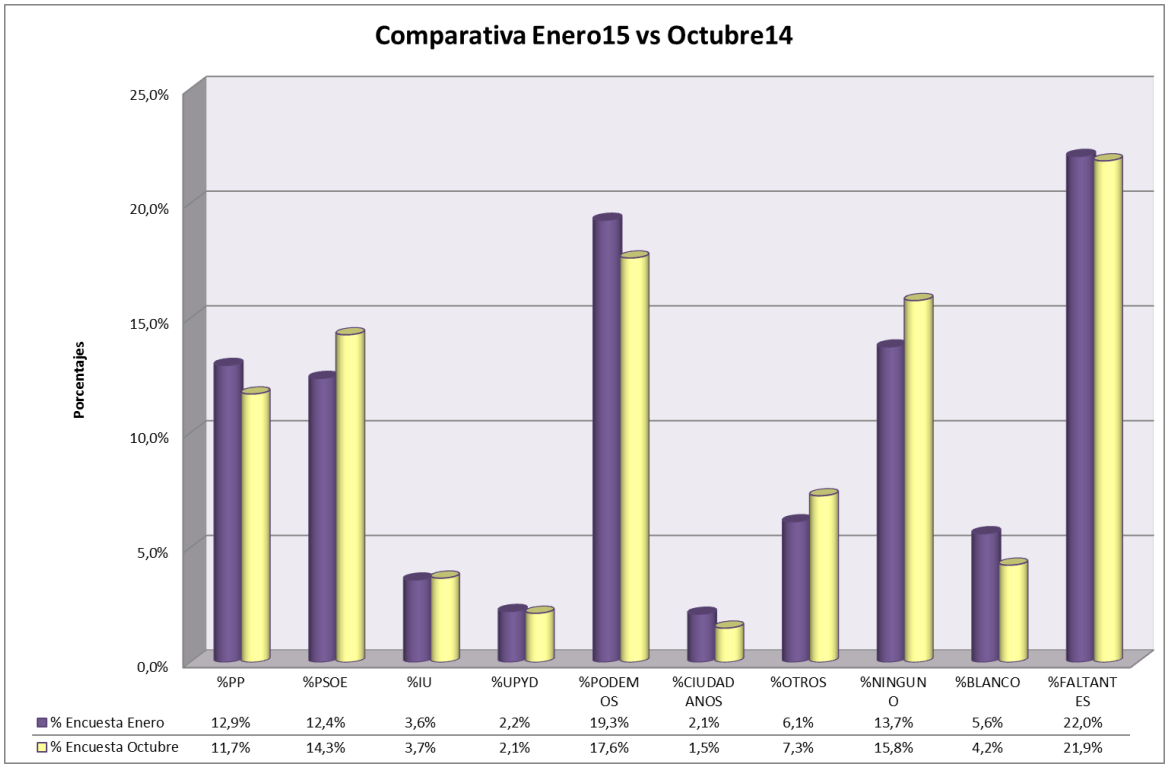

Figura 2.Comparativa entre encuestas. Elaboración propia.

En la figura 2 observamos los porcentajes de respuesta a la pregunta sobre la intención de voto, además del porcentaje de valores faltantes (%FALTANTES). En color morado se muestran los valores de Enero de 2015 y en amarillo los correspondientes a la encuesta del mes Octubre de 2014. No existen grandes diferencias entre las dos encuestas, difieren no más de 2 puntos. Nuestro objetivo como ya hemos explicado antes son los datos faltantes, ambos valores también son muy similares en la encuesta de Octubre existe un 21.9% de no respuesta y en Enero esta cifra es del 22.0%.

A continuación desagregaremos ambas encuestas, en la figura 3 se muestra el porcentaje de respuesta y no respuesta en la encuesta de Enero 2015:

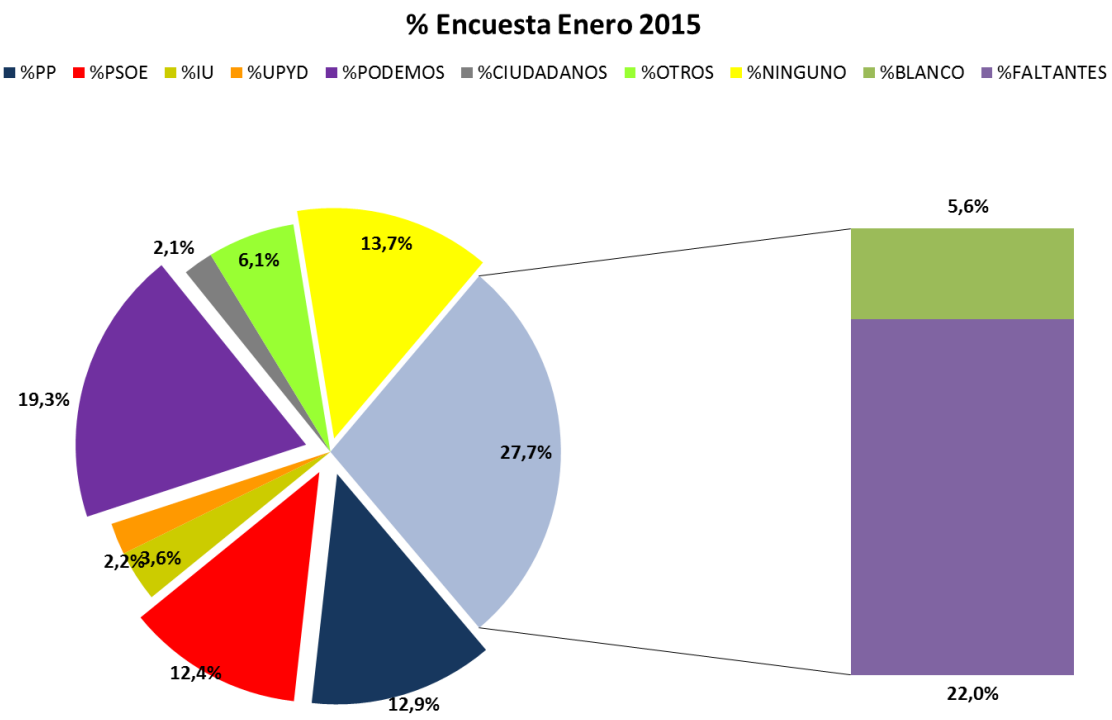

Figura 3. Encuesta enero. Elaboración propia.

La fuerza política con un porcentaje mayor de intención de voto, en unas supuestas elecciones generales en Enero de 2015, es Podemos con un 19,3%, seguida por PP y PSOE el primero con un 12,9% y segundo con un 12,4%. Si observamos el gráfico vemos como un 27,7% de los votos son votos en blanco o faltantes, junto con Ninguno 13,7%, aproximadamente el 42% de los votos no son para ningún partido o coalición concreto, esto indica que dicho porcentaje de encuestados no tiene ningún interés por un partido definido. Pero nuestro dato de interés y dónde nos centraremos es en el porcentaje de datos faltantes, es decir, el porcentaje de encuestados que no saben o no contestan a nuestra pregunta de interés. Este 22% es el que en el capítulo siguiente intentaremos estimar.

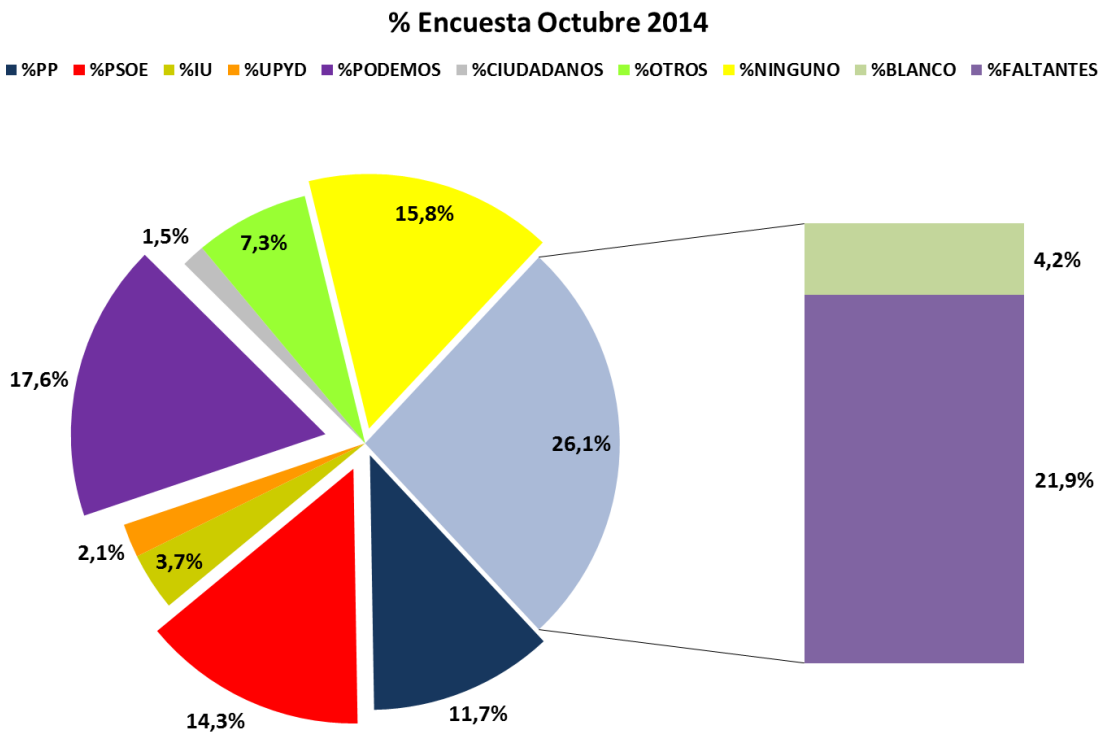

Figura 4. Encuesta octubre. Elaboración propia.

En la figura 4 observamos como la primera fuerza política, al igual que en Enero, es Podemos con un 17,6%, seguida por PSOE 14,3% y PP con 11,7%. Al igual que antes aproximadamente un 42% de los votos son para ninguno, voto en blanco y no sabe/no contesta.

Una de las variables que más nos interesa estudiar, como variable explicativa de la intención de voto, es el ***recuerdo de voto*** en las elecciones generales de 2011 pero no debemos dejar de lado la simpatía de los encuestados a un grupo político o coalición. Ambas variables las estudiaremos a continuación:

- ✓ 215 personas no contestaron en Enero de 2015 a la pregunta que hace referencia al recuerdo del voto en elecciones anteriores y 246 en la encuesta de Octubre 2014, es decir, en Enero existe un 8,7% de no respuesta y en Octubre un 9,9%.
- ✓ En cuanto a la ***simpatía*** de los encuestados hacia un partido político u otro, vemos como el porcentaje de no respuesta no es muy elevado: en Enero 106 personas no contestaron a esta pregunta lo que supone un porcentaje de no respuesta de 4,3%. Mientras que en Octubre este porcentaje aumenta hasta el 5,7%.

A continuación mostramos unas tablas con las frecuencias y porcentajes de datos faltantes o no respuesta en las variables de interés:

	INTENCIÓN VOTO		SIMPATIA		IDEOLOGIA		RECUERDO VOTO	
	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES
ENERO	547	22,05%	106	4,27%	465	18,74%	215	8,67%
OCTUBRE	542	21,85%	141	5,69%	519	20,93%	246	9,92%

	SIT.ECONOMICA		SIT.POLITICA		GOBIERNO		OPOSICION	
	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES
ENERO	5	0,20%	71	2,86%	26	1,05%	88	3,55%
OCTUBRE	9	0,36%	76	3,06%	53	2,14%	101	4,07%

	PROB.PP		PROB.PSOE		PROB.CIUDADANOS		PROB.PODEMOS	
	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES	Faltantes	%FALTANTES
ENERO	146	5,88%	159	6,41%	435	17,53%	234	9,43%
OCTUBRE	175	7,06%	188	7,58%	514	20,73%	334	13,47%

	PROB.UPYD		PROB.IU	
	Faltantes	%FALTANTES	Faltantes	%FALTANTES
ENERO	298	12,01%	194	7,82%
OCTUBRE	313	12,62%	222	8,95%

Observamos como los valores más altos de no respuesta son a la pregunta sobre la intención del voto alrededor del 22% en ambas encuestas, seguida de las variables ideología (18,74% Enero, 20,93% Octubre) y probabilidad de votar a Ciudadanos con 17,53% de no respuesta en Enero y un 20,73% en Octubre. En el otro extremo, tenemos la pregunta que referencia a la situación económica con tan solo un 0,20% y 0,36% de no respuesta (tan sólo 5 y 9 personas no contestaron), por encima de ella la valoración al gobierno con aproximadamente un 2%.

Una vez analizado de manera breve nuestras variables pasaremos a explicar la metodología seguida para la realización de nuestro caso de estudio. Como ya hemos dicho anteriormente en este trabajo se pretende estimar, mediante imputación, los datos faltantes de la variable intención de voto.

Capítulo Segundo

1. METODOLOGÍA

Los análisis que se realizan con datos recogidos a partir de encuestas se enfrentan de manera habitual a la existencia de no respuestas o datos faltantes. Ante la falta de información en algunas preguntas de las encuestas, la práctica más común que llevan a cabo los investigadores es obviar los datos faltantes y trabajar únicamente con la información disponible. Estos métodos son llamados listwise o pairwise deletion, según Schafer (1997) dependiendo de si se elimina o no el individuo con valores faltantes. Cuando el volumen de registros incompletos es un porcentaje reducido (5% o menos), la eliminación de estos registros es una solución razonable a los problemas de datos faltantes (Rivas, Martínez, & Galindo, 2010).

Pero esta eliminación trae consigo un importante inconveniente y es que la muestra se reduce, con el consiguiente problema de posibles sesgos en la muestra. Otra opción es utilizar métodos de imputación simple, en los que las no respuestas son sustituidas por un valor calculado por el propio investigador. Los métodos más utilizados son la imputación por media, imputación mediante regresión o por imputación deductiva.

Frente a estas opciones existen los métodos de imputación múltiple, este trabajo se centrará en ésta última opción para las imputaciones de los valores faltantes de la variables explicadas en secciones anteriores, pero detallaremos ambos métodos en secciones siguientes. Debemos notar que a lo largo de este trabajo se habla de la no respuesta en encuestas por muestreo y se usan de manera indistinta términos como no respuesta, datos faltantes o datos missing.

Los distintos métodos de imputación se pueden clasificar según dos criterios como bien recoge (Déborah, 2011). Por un lado, pueden ser simples o múltiples. Por otro, pueden ser determinísticos o aleatorios. En primer lugar, vamos a detallar cuál es la diferencia entre los métodos de imputación simple y múltiple:

- ✓ *Imputación simple.*

Consiste en asignar un valor por cada valor faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa.

- ✓ *Imputación múltiple.*

  Consiste en asignar a cada valor faltante varios valores (m), generando m conjuntos de datos completos. En cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.

Además, los métodos de imputación se pueden clasificar de una segunda forma:

- ✓ *Métodos de imputación determinísticos.*

  Son aquellos métodos en los que se producen las mismas respuestas cuando se repite la imputación en varias unidades, todo ello bajo las mismas condiciones.

- ✓ *Métodos de imputación estocásticos o aleatorios.*

  Son aquellos que producen resultados diferentes cuando se repite el método de imputación bajo las mismas condiciones para una unidad.

1.1. Tipos de datos faltantes.

Antes de comenzar a explicar los métodos de imputación debemos hablar de los posibles tipos de datos faltantes que existen y que se clasifican de la siguiente manera:

- ✓ *MCAR:* una variable es MCAR (missing completely at random) si la probabilidad de pérdida de una observación para todos los individuos es la misma y no depende de las medidas de otras. Es decir, la ausencia de la información no está originada por ninguna variable presente en la matriz de datos.
- ✓ *MAR:* una variable es MAR (missing at random) si la probabilidad de pérdida de la observación de un individuo depende de la información observada. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos.

Los datos faltantes mencionados también se denominan como datos ignorables, ya que producen efectos que se pueden ignorar si se controla de manera eficiente las variables que determinan la no respuesta.

- ✓ *MNAR:* una variable es MNAR (missing not at random) si la probabilidad de que la observación de un individuo esté perdida está relacionada con los valores perdidos. Este tipo de dato también es denominado no ignorable.

1.2. Métodos de imputación.

Se denomina imputación al procedimiento que utiliza la información contenida en la muestra para asignar un valor a aquellas variables que tienen registros con el valor ausente, ya sea porque se carece de información o porque se detecta que algunos de los valores recolectados no corresponden con el comportamiento esperado. La razón principal por la cual se realiza la imputación es obtener un conjunto de datos completo y consistente al cual se le puedan aplicar las técnicas estadísticas ordinarias (Déborah, 2011).

1.2.1. Imputación simple.

Como hemos introducido anteriormente los métodos de imputación simple, son aquellos en los que el investigador pretende solucionar el problema de los datos faltantes sustituyendo los mismos por valores estimados a partir de información de la muestra. Con ello se consigue una matriz de datos completa en la que es posible realizar diferentes análisis estadísticos. Existen muchos métodos de imputación simple, nosotros hablaremos sobre imputación mediante la media, mediante regresión y métodos por imputación deductiva.

- ✓ *Imputación mediante la media.*

Mediante este método se reemplaza cada valor perdido por la media de los valores observados de la matriz de datos. Este método fue propuesto en 1932 por Wilks y es posiblemente uno de los procedimientos más sencillos y antiguos de imputación. Este método tiene dos variantes imputación por media no condicional e imputación por media condicional. En la primera se asume que los datos missing siguen un patrón MCAR y consiste en estimar la media de los valores observados. El segundo imputa medias condicionadas a valores observados.

- ✓ *Imputación mediante regresión.*

Este proceso se basa en eliminar las observaciones con datos incompletos y ajustar la recta de regresión para predecir los valores de los valores faltantes. Sea Y una variable que presenta n_{pe} valores perdidos y como valores observados

n_i=n−n_{pe}. Supongamos que las K variables, $\mathbf{X}=(X_1 \ldots X_K)$, no presentan valores perdidos.

Si para el caso i el valor y no se observa, entonces siendo un modelo determinístico, el valor faltante es imputado usando la siguiente ecuación de regresión:

$$\hat{y}_i = \hat{\beta}_{0*12\ldots k} + \sum_{j=1}^{k} \hat{\beta}_{j*12\ldots k} * x_{ij} \,,$$

donde $\hat{\beta}$ y $\hat{\beta}_j$ son estimadores MCO de la regresión de Y sobre X basada en las observaciones completas.

Una alternativa para atenuar el efecto de subestimar la variabilidad consiste en añadir al valor predicho por la regresión una perturbación aleatoria. Es decir, imputaremos el valor faltante mediante:

$$\hat{y}_i = \hat{\beta}_{0*12\ldots k} + \sum_{j=1}^{k} \hat{\beta}_{j*12\ldots k} * x_{ij} + z_i \,,$$

donde $z_i \sim N(0; \tilde{\sigma}_{12\ldots k})$, siendo $\tilde{\sigma}_{12\ldots k}$ la varianza residual de la regresión de Y sobre X basada en las observaciones completas.

✓ *Imputación deductiva.*

Este método pertenece a la clasificación de métodos de imputación determinísticos y se realiza en situaciones en que los valores perdidos se pueden deducir del resto de variables o información del conjunto de datos, es decir, los valores se asignan mediante relaciones lógicas existentes entre las diferentes variables. Generalmente se tiene esta fórmula:

Si (se expresa la condición) *entonces* (se expresa la acción).

Por ejemplo, se podría decir que *si* un encuestado responde que votó al PSOE en las pasadas elecciones *entonces* su intención de voto es votar a dicho partido.

1.2.2. Imputación múltiple.

La imputación múltiple fue propuesta por (Rubin, 1987) como una alternativa a las técnicas de imputación simple. Debemos notar que los métodos de imputación

múltiples tienen como objetivo crear predicciones de cada valor ausente, no la explicación causal de dicha ausencia ni la interpretación de los parámetros obtenidos.

Este método requiere la obtención de m>1 valores aleatorios para cada valor perdido, de manera que dispondremos de m conjuntos de datos completos. El número de imputaciones realizadas suelen estar entre tres y diez, por defecto la librería `mi` realiza tres imputaciones. En cada uno de estos m conjuntos de datos los valores observados son los mismos mientras que los valores faltantes presentan valores imputados diferentes. Finalmente la imputación múltiple requiere de un modelo de análisis estadístico que combine las m imputaciones realizadas, para así obtener un único valor estimado para cada valor perdido (Muñoz Rosas & Álvarez Verdejo, 2009). Es decir, podemos resumirlo en los siguientes pasos:

1) Cada valor perdido se reemplaza por un conjunto de m>1 valores generados por simulación, con lo que se crean m conjuntos de datos completos.
2) Se aplica a cada uno de ellos el método de análisis deseado.
3) Los resultados obtenidos se combinan mediante reglas simples para producir una estimación global.

La figura 5 muestra un esquema del procedimiento de imputación múltiple.

Una condición necesaria para la imputación múltiple es que los datos faltantes deben ser perdidos al azar, es decir, en la clasificación realizada en la sección anterior los datos deben ser MAR (missing at random). Si lo aplicamos a nuestro estudio concreto, aunque las probabilidades de responder a la intención de voto puedan ser menores entre los votantes de determinados partidos políticos, el hecho de que al menos algunos individuos pertenecientes a ese partido contesten a preguntas referidas a simpatía, recuerdo de voto etc. permitirán predecir su intención de voto.

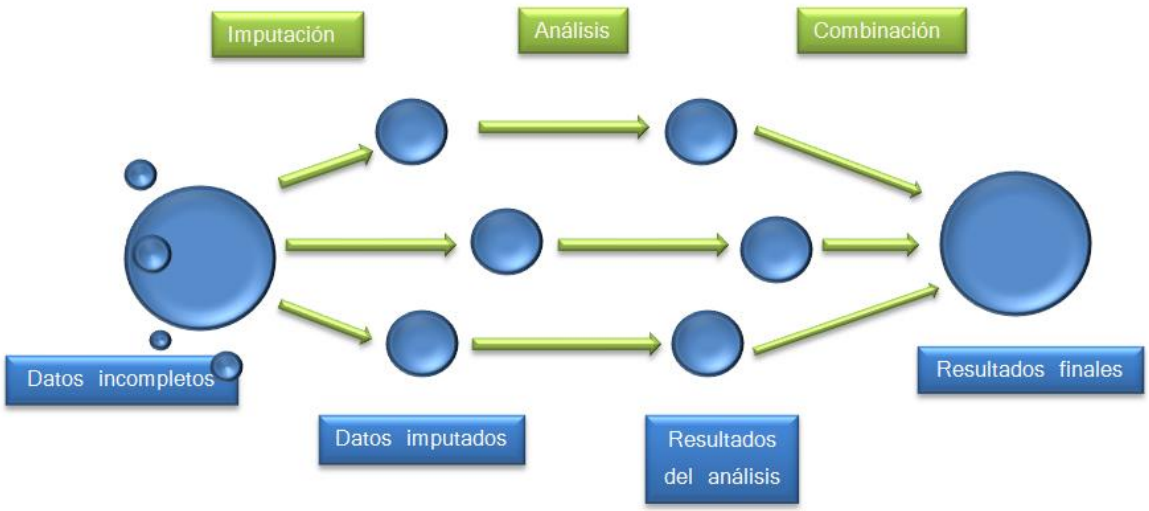

Figura 5. Proceso de imputación. Elaboración propia

Los resultados obtenidos en el análisis se combinan mediante reglas simples para producir una estimación global. De acuerdo con (Restrepo Estrada & Marín Diazaraque, 2012) para los parámetros estimados se toma la media aritmética de las estimaciones de cada conjunto de datos generado. En el caso de los errores estándar es necesario estimar tanto la varianza dentro de cada imputación como la varianza entre imputaciones. Para la varianza dentro de la imputación (Restrepo Estrada & Marín Diazaraque, 2012), proponen la siguiente expresión:

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

donde t es un conjunto de datos imputado particular y m es el número total de conjuntos de datos imputados. Por otro lado, para la varianza entre imputaciones se usa:

$$V = \frac{\sum\left(\hat{\vartheta}_t - \bar{\vartheta}\right)^2}{m-1},$$

donde $\hat{\vartheta}_t$ es el parámetro estimado del conjunto de datos imputado t y $\bar{\vartheta}$ es la media del parámetro estimado en las m imputaciones. Finalmente, el error estándar combinado se calcula a partir de la varianza dentro de la imputación y la varianza entre imputaciones (Restrepo Estrada & Marín Diazaraque, 2012):

$$SE = \sqrt{\left(H + V + (V/m)\right)}.$$

De esta manera, la imputación múltiple resuelve el problema de subestimación de los errores estándar, que se da en el caso de la imputación simple, incorporando la

varianza entre imputaciones en el error estándar. Si el porcentaje de datos faltantes es pequeño, las estimaciones y los errores estándar serán muy cercanos entre las distintas imputaciones y la estimación global y el error estándar serán casi iguales (Restrepo Estrada & Marín Diazaraque, 2012).

Como hemos mencionado nosotros implementaremos un método de imputación múltiple con el software R, utilizando la librería llamada mi, a continuación nos centraremos en el procedimiento que lleva a cabo para imputar.

Las imputaciones multivariadas, en las que se imputan los valores perdidos de más de una variable en el mismo algoritmo (anexo 1, se muestran las variables a imputar) tienen una mayor complejidad que las imputaciones simples.

En este caso, mi utiliza un algoritmo conocido como ecuaciones encadenadas (Yu-sung, Gelman, Hill, & Yajima, 2011). Esta estrategia es práctica y flexible a la hora de manipular los datos faltantes pues permite implementarlo en variables con diferentes niveles de medición. El proceso consiste en asignar valores a los datos faltantes múltiples veces, creando múltiples conjuntos de datos "completos", estimando un modelo de imputación separado para cada variable y usando como variables explicativas al resto de variables incluidas en el análisis. Una vez se han obtenido los múltiples conjuntos de datos, se analizan individualmente de forma idéntica para obtener un conjunto de estimadores de los parámetros. (Royuela Vicente, 2014).

Las ecuaciones encadenadas de este estudio tienen la siguiente forma, teniendo en cuenta que el orden es ascendente en número de datos faltantes:

Sit.econ~ voto + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia +recuerdo + sexo + edad

gobierno ~ voto + sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + oposicion + simpatia + ideologia + recuerdo + sexo + edad

sit.politica ~ voto + sit.econ + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad

.
.
.

Pciudadanos ~ voto + sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad

ideologia ~ voto + sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + recuerdo + sexo + edad

voto ~ sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad

El algoritmo a seguir es el siguiente:

> *"Inicialmente, todos los valores perdidos son sustituidos por valores extraídos aleatoriamente con re-emplazamiento del conjunto de valores observados para cada variable.*
>
> *A continuación, sobre la primera variable con el menor número de datos faltantes (llamada X_1, por ejemplo), se aplica una regresión condicionando el resto de variables (X_2, X_3,..., X_k) como términos independientes, solo sobre aquellas observaciones donde X_1 esté completa. Los valores faltantes de X_1 son reemplazados entonces por valores simulados (simulated draws) de la correspondiente distribución predictiva posterior de X_1. Sobre la siguiente variable en orden ascendente con datos faltantes, X_2, se aplica entonces otra regresión con todas las demás variables como términos independientes (X_1, X_3,...,X_k), restringiendo esta regresión a aquellas observaciones con todos los valores de X_2 observados y utilizando los valores imputados de X_1. De nuevo, los valores perdidos de X_2 son reemplazados por valores simulados (draws) de la distribución predictiva posterior de X_2. Este proceso se repite sucesivamente para todas las demás variables con datos faltantes. 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. Una característica destacada de mi reside en su habilidad para el manejo de diferentes tipos de variables (continuas, binarias, categóricas ordenadas o desordenadas)"* (Royuela Vicente, 2014).

En nuestro caso de estudio realizaremos el siguiente modelo para la estimación de la variable intención de voto:

"voto"="voto ~ sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad"

Finalmente, el modelo estimado es el siguiente:

Voto = 1.72 − 0.04 sit.econ + 0.05 sit.politica − 3.77 Ppp - 2.20 Ppsoe - 1.86 Piu - 1.65 Ppodemos + 1.36 Pciudadanos + 0.78 Pupyd + 0.16 gobierno + 0.11 oposicion + 0.60 simpatia + 0.01 ideologia + 0.08 recuerdo +0.03 sexo − 0.01 edad

Observamos como en el modelo generado los coeficientes negativos más elevados son para variables como probabilidad de votar al PP y PSOE con coeficientes negativos de 3,77 y 2,20 puntos, respectivamente. Las variables probabilidad de votar a Ciudadanos y UPYD junto con Simpatía obtienen los coeficientes positivos más elevados.

1.3. Imputación simple frente imputación múltiple.

Los métodos de imputación simple tienen la gran ventaja que no necesitan de un esfuerzo de computación elevado, mientras que los métodos de imputación múltiple centran su gran desventaja en este punto, pues el coste computacional es más elevado. Los métodos múltiples corrigen la desventaja de la imputación simple respecto a la subestimación de la verdadera varianza cuando la proporción de datos faltantes es elevada. Una desventaja de los métodos múltiples es que al producir varias respuestas, el investigador debe trabajar con diferentes bases de datos, en las cuales los valores imputados no tienen el mismo valor. Pero el poder obtener varias opciones de imputación para los valores faltantes, crea una ventaja para los métodos de imputación múltiple que los simples no pueden crear.

Una desventaja común a ambos métodos es que se asumen patrones de datos MAR, es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos y existen situaciones en las que no es posible verificar este supuesto.

2. CASO PRÁCTICO

2.1. Estudio de los datos faltantes.

En esta sección del trabajo explicaremos de manera exhaustiva nuestro caso de estudio. Cabe recordar que el objetivo es la estimación de la intención de voto, en las encuestas realizadas por el CIS en Octubre de 2014 y Enero de 2015. Para ello, como ya hemos comentado anteriormente, estudiaremos diversas variables

además de la variable de interés. Para realizar dicha estimación debemos imputar los valores faltantes de todas las variables, por ello hemos creado un código con el software estadístico llamado R. Como ya hemos comentado en apartados anteriores hemos utilizado la librería mi. Se puede observar dicho código en el Anexo, 1.

En primer lugar vamos a mostrar un gráfico en el que se pueden observar los datos faltantes que existen en la encuesta realizada en el mes de Enero de 2015.

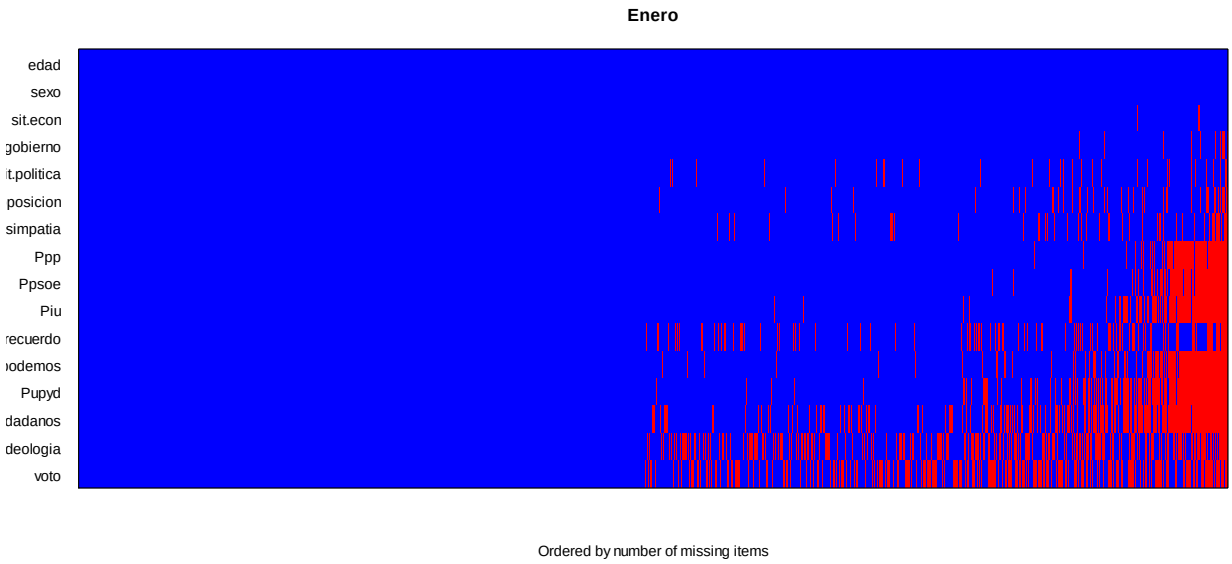

Figura 6. Faltantes encuesta Enero. Fuente: Elaboración propia con R.

La figura 6 muestra los valores faltantes en color rojo, en color azul se muestran aquellos datos que han obtenido contestación por parte del encuestado. El gráfico ha sido elaborado con R y está ordenado de variables que tienen más datos faltantes a variables que no tienen faltantes como son edad y sexo. Sabemos que en la encuesta de Enero existen 1223 personas que contestaron a todas las variables de interés. Podemos observar como la variable con mayor índice de no respuesta es la intención de voto (voto), en segundo lugar la pregunta referida a la ideología del encuestado, en tercer puesto la variable que está reseñada a la pregunta sobre la probabilidad de votar a Ciudadanos... Las variables sexo y edad tienen toda su área de color azul puesto que no tienen ninguna no respuesta.

A continuación mostramos el gráfico de faltantes para la encuesta de Octubre 2014, en esta ocasión existe un total de 1203 personas que contestaron todas las variables de estudio.

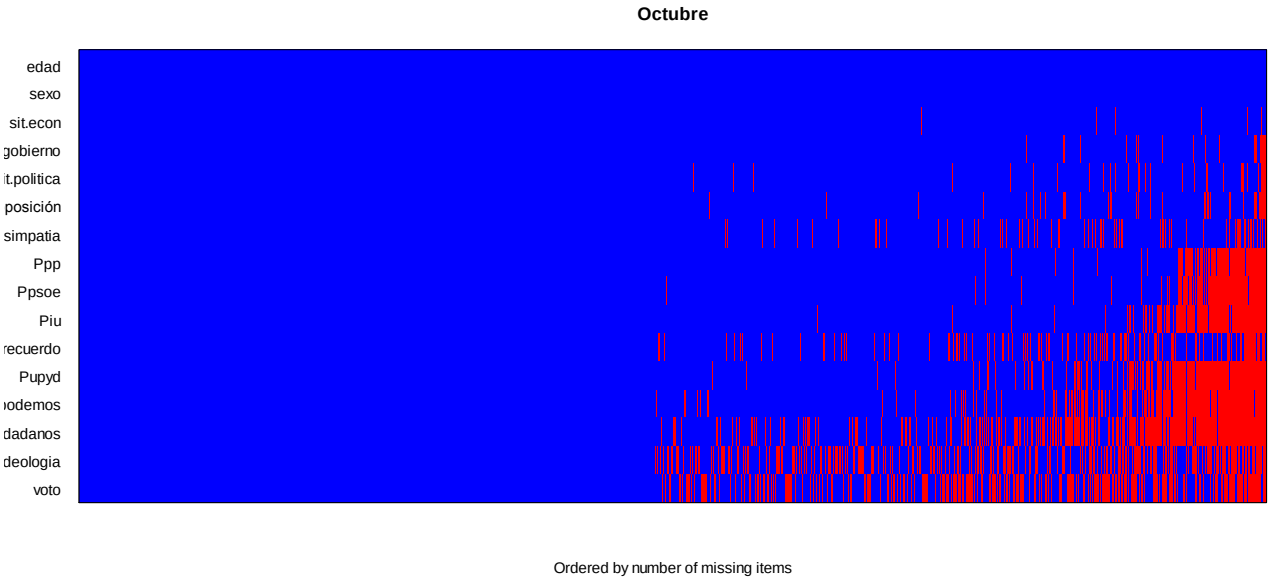

Figura 7 Faltantes encuesta Octubre. Fuente: Elaboración propia con R.

Podemos observar que ambos gráficos no difieren demasiado, en cuanto a la ordenación de las variables en Octubre tan sólo se diferencian en el orden de la variable que hace referencia a la pregunta sobre la probabilidad de votar Podemos y a UPyD, en Enero existen más datos faltantes en la que se refiere a UPyD y en Octubre en la variable llamada "Ppodemos" que hace referencia a la probabilidad de votar a Podemos.

Con estos gráficos se pueden intuir ciertos patrones a la hora de no responder a las preguntas de interés. Así pues observamos cómo hay encuestados, los que más preguntas de interés contestan, que responden a todas las preguntas menos a ideología, probabilidad de votar a Podemos y recuerdo del voto anterior. O algunas personas en el otro extremo que no responden a nada excepto cuando se le preguntan por la situación económica, edad y sexo, debemos tener en cuenta que la edad y el sexo son variables a las que contestan todas las personas encuestadas.

Si estudiamos detalladamente algunos patrones, podemos afirmar que en la encuesta de Enero existen 131 personas que no contestaron a la variable intención de voto y tampoco lo hicieron a ideología, mientras que existen 176 personas que tienen el mismo patrón de no respuesta en la encuesta de Octubre. Por ejemplo en la misma encuesta de Enero tan sólo existe una persona que no contesta a la intención de voto, la situación económica y la situación política, mientras que en la encuesta de Octubre para este mismo patrón constan dos personas que no contestan a esas preguntas.

2.2. Evaluación de las imputaciones.

Para evaluar la efectividad del método de imputación implementado en nuestros datos es necesario trabajar con bases de datos completas, es decir, el modo de conocer si un valor ha sido correctamente imputado es conociendo el valor real de dicho dato. Para ello hemos creado un código (véase Anexo, 2) que consta de varios pasos importantes a seguir. En primer lugar, seleccionamos la encuesta a evaluar, esta encuesta la dividimos en dos submuestras, una llamada 'entrenamiento' con 2000 datos y otra llamada 'prueba' con los datos restantes, esta partición se realiza de forma aleatoria. Cuando tenemos la submuestra 'prueba' guardamos los datos referentes a la intención de voto en otra variable llamada 'reales', eliminamos estos datos para poder imputarlos y después compararlos con los datos guardados en 'reales'. Así obtenemos una variable con datos imputados que llamaremos 'estimados' y otra variable con los votos 'reales'. Con estas dos variables podemos obtener los porcentajes de acierto de nuestra imputación, es decir, aquellos valores que hemos imputado y clasificado de manera correcta en su categoría, pudiendo así realizar la evaluación de nuestro proceso. A continuación mostraremos los resultados de esta evaluación en las diferentes encuestas.

ENERO 2015	Estimación								
Voto Real	PP	PSOE	IU	UPyD	Podemos	Ciudadanos	Otros	Ninguno	Blanco
PP	62	1		1	1	1	2	3	
PSOE	4	45	1		5			4	2
IU		1	9		2		1	1	4
UPyD	2	1		2		1	1	2	
Podemos	2	1	2		63		3	3	7
Ciudadanos	1				2	2		3	
Otros	1				4		23	3	
Ninguno	5	2	1	1	9	1	2	33	6
Blanco		2		1	6	1		10	10
NA	12	21	8	1	22	2	8	21	17

Comenzaremos por comentar los resultados obtenidos en la encuesta de Enero 2015, los mejores resultados que hemos obtenido son para la estimación del voto en partidos como PP 87.3% de acierto, con un 77.8% de acierto en cuanto a Podemos, seguidos por Otros con un 74.2% y PSOE con un porcentaje del 73.8. En el otro extremo tenemos a partidos como UPyD con un 22.2% de acierto, Ciudadanos con un 25%, en blanco con 33.3% y con un porcentaje medio tenemos al partido IU (50%) y Ninguno con un 55%.

Debemos tener en cuenta que los partidos con menor porcentaje de acierto son aquellos en el que el número total es muy escaso, es decir, tan sólo contamos con 8 y 9 encuestados que su intención de voto era para el partido Ciudadanos y UPyD, respectivamente. Observamos como a la hora de estimar los votos en blanco, acertamos tan sólo 10 votos, ya que, clasificamos votos en blanco como votos a ninguno en la misma cantidad, pero esto no es preocupante pues son valores en cierto sentido similares. En cambio los votos que estimamos erróneamente en Ninguno si son de mayor importancia, pues clasificamos bien 33 de 60 votos, y 21 de ellos los reemitimos a diferentes partidos. Pero los errores "más graves" son aquellos partidos de una ideología clara y opuesta a nuestras estimaciones.

Si definimos porcentaje global de aciertos como:

$$\%\, Global = \frac{aciertos}{total\ datos\ reales\ disponible} * 100\,,$$

En la submuestra realizada para la evaluación de Enero 2015, disponemos de 369 datos disponibles y 249 de aciertos. Por lo que obtenemos un porcentaje total de acierto del 67.5%, esto nos indica que a la hora de estimar la intención de voto para esta encuesta tenemos un porcentaje elevado de aciertos.

A continuación especificaremos los resultados para la encuesta de Octubre 2014 para su posterior estudio.

OCTUBRE 2014 Estimación									
Voto Real	PP	PSOE	IU	UPyD	Podemos	Ciudadanos	Otros	Ninguno	Blanco
PP	35	4		2	2	1	2	8	3
PSOE	2	54	1	1	9	1		10	1
IU		1	8	1	1	1	1	2	2
UPyD		4		2	2			2	1
Podemos	1	1			61	1		9	3
Ciudadanos	1	1		1	1	3		2	1
Otros	1	1			2		25	5	3
Ninguno	3	6	2	1	11	1	6	43	8
Blanco	2	2		1	4	2		8	4
NA	9	17	4	1	9	4	11	29	5

En esta encuesta obtenemos de la misma manera el porcentaje global de aciertos y para ello disponemos de 391 datos y 235 como suma de aciertos, por lo que tenemos que en la encuesta de octubre 2014 existe un 60.1% de acierto global, siete puntos menos que en la anterior. No tenemos una explicación para esta diferencia pero nos llama especialmente la atención que existen muchos errores, al clasificar los diferentes partidos, que van dirigidos a Podemos, tal vez sea porque

por esas fechas este partido tuvo mucha repercusión mediática y eso influyera en los encuestados.

Además observamos como para la estimación de votos de este partido se obtiene el porcentaje de acierto más elevado de todos, con un 80.3% de acierto. Seguido por PSOE, Otros y PP con porcentajes de 68.4%,67.6% y 61.4%, respectivamente. Si comparamos estos resultados con los de la encuesta de Enero observamos como existen diferencias notables, ya que en Octubre cometemos más errores sobre todo con Podemos, pues estamos imputando datos a este partido de manera errónea. En general, obtenemos resultados más bajos que en la encuesta de Enero.

2.3. Comparación de resultados.

En esta última parte del caso de estudio vamos a realizar una comparación entre nuestras estimaciones y las que realiza el CIS, debemos tener en cuenta que no se está realizando el mismo tipo de modelo estadístico por lo que no pretendemos obtener los mismos resultados. Sobre el método que utiliza éste organismo tenemos la siguiente información que aparece en los barómetros de Enero, Abril, Julio y Octubre:

> Dado que los datos de los indicadores "intención de voto" e "intención de voto + simpatía" son datos directos de opinión y no suponen ni proporcionan por sí mismos ninguna proyección de hipotéticos resultados electorales, en este anexo se recogen los resultados de aplicar un modelo de estimación a los datos directos de opinión proporcionados por la encuesta. Procedimiento que conlleva la ponderación de los datos por recuerdo de voto imputado y aplicación de modelos que relacionan la intención de voto con otras variables. Obviamente, la aplicación a los mismos datos de otros modelos podría dar lugar a estimaciones diferentes.

Realizaremos esta comparación de manera separada por un lado la encuesta de Enero 2015 y por otro la de Octubre 2014. Así bien comenzaremos por la primera citada:

INTENCIÓN VOTO	ENCUESTA ENERO	INPUT 1	INPUT 2	INPUT 3	PROMEDIO INPUT
PP	321	365	369	371	368
%PP	12,9%	14,7%	14,9%	15,0%	14,8%
PSOE	307	409	398	420	409
%PSOE	12,4%	16,5%	16,0%	16,9%	16,5%
IU	89	118	121	113	117
%IU	3,6%	4,8%	4,9%	4,6%	4,7%
UPYD	55	70	63	67	67
%UPYD	2,2%	2,8%	2,5%	2,7%	2,7%
PODEMOS	478	575	567	571	571
%PODEMOS	19,3%	23,2%	22,9%	23,0%	23,0%
CIUDADANOS	52	67	70	72	70
%CIUDADANOS	2,1%	2,7%	2,8%	2,9%	2,8%
OTROS	152	184	187	190	187
%OTROS	6,1%	7,4%	7,5%	7,7%	7,5%
BLANCO	139	197	204	211	204
%BLANCO	5,6%	7,9%	8,2%	8,5%	8,2%
NINGUNO	341	496	502	466	488
%NINGUNO	13,7%	20,0%	20,2%	18,8%	19,7%
FALTANTES	547	-	-	-	-
%FALTANTES	22,0%	-	-	-	-

En la tabla anterior, se muestran los datos reales de la encuesta referidos a la intención de voto en unas supuestas elecciones generales (color verde). Además de los resultados obtenidos en nuestras tres imputaciones (color amarillo) éstos están referidos al total de encuestados y en color azul se muestra el promedio de las imputaciones.

ENCUESTA ENERO	INPUT1	INPUT2	INPUT3	Promedio Imputaciones	ESTIMACIÓN CIS
PP	18,4%	18,6%	18,4%	18,5%	27,3%
PSOE	20,6%	20,1%	20,8%	20,5%	22,2%
IU	5,9%	6,1%	5,6%	5,9%	5,2%
UPYD	3,5%	3,2%	3,3%	3,3%	4,6%
PODEMOS	29,0%	28,7%	28,3%	28,7%	23,9%
CIUDADANOS	3,4%	3,5%	3,6%	3,5%	3,1%
OTROS	9,3%	9,4%	9,4%	9,4%	9,5%
BLANCO	9,9%	10,3%	10,5%	10,2%	4,1%

En la tabla anterior se muestran los porcentajes referidos al total sin tener en cuenta el valor Ninguno, realizamos esta clasificación para compararlos con los datos que proporciona el CIS teniendo las mismas clasificaciones. Así en color amarillo se muestran las tres imputaciones realizadas y en azul el promedio de éstas. Por último, se muestra en color naranja, la estimación que propone el CIS en sus

estudios sobre voto válido. Esta segunda tabla nos sirve para poder comparar las estimaciones obtenidas con las estimaciones realizadas por el CIS.

Notar que a diferencia del CIS nosotros incluimos una clasificación llamada Ninguno, en la que como hemos explicado durante el trabajo incluye las opciones voto nulo o no votar. Así pues, podemos observar como existen varios casos que nos llaman la atención, el primer caso es la estimación que realiza el CIS sobre el porcentaje de voto que iría destinado al Partido Popular (PP), este dato es de un 27%, mientras que nuestras estimaciones se aproximan al 15%, si bien cuando lo realizamos sin tener en cuenta el valor Ninguno obtenemos imputaciones cercanas al 18.5%. Este mismo caso ocurre con la intención de voto hacia el PSOE, nuestras estimaciones rondan el 16%-17% mientras que el CIS estima dicho valor entorno al 22%, pero son bastantes similares si no tenemos en cuenta el valor Ninguno en nuestras imputaciones pues los porcentajes son en torno al 20.5%.

Otro caso que llama la atención es el referido a Podemos, los porcentajes son similares entre nuestras imputaciones, alrededor del 23% realizadas con el total de encuestados, y las estimaciones del CIS 23.9%. Si bien cabe destacar la diferencia que existe cuando realizamos el promedio de las estimaciones sin tener en cuenta el valor Ninguno, pues el porcentaje sube casi seis puntos llegando al 28.7%.

En el extremo de porcentajes bajos vemos que apenas existen diferencias entre las estimaciones que propone el CIS referentes a los partidos IU (5.2%) y Ciudadanos (3.1%) frente a nuestras estimaciones con datos cercanos al 3% para imputaciones referidas a Ciudadanos y cercanas al 5% para las estimaciones realizadas hacia IU, obteniendo un punto más aproximadamente en los porcentajes realizadas sin el valor Ninguno.

No obstante no se puede afirmar que nuestras estimaciones sean más acertadas que las del CIS o viceversa tan sólo estamos comparando resultados. Pero cabe destacar que si realizamos los porcentajes sobre el total menos el valor Ninguno los porcentajes obtenidos se acercan a los del CIS.

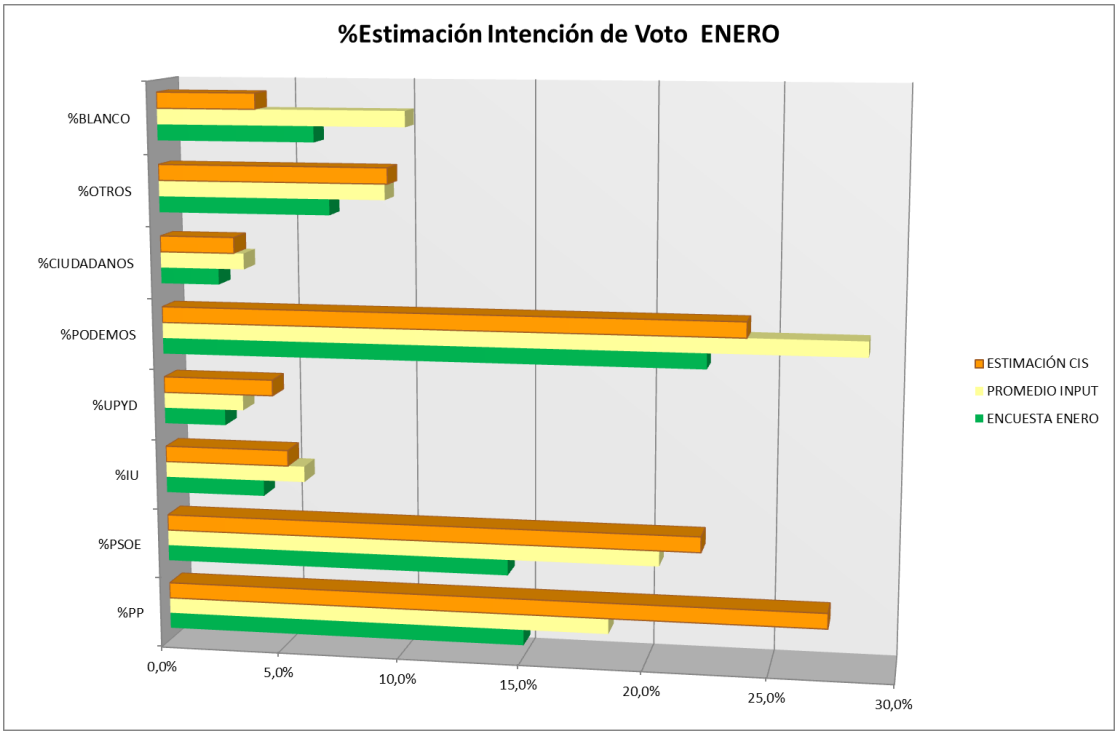

Figura 8 Comparación Enero. Fuente: Elaboración propia.

La figura 8 está elaborada con los porcentajes realizados con el total de encuestados menos el valor Ninguno, así tenemos las mismas clasificaciones que realiza el CIS. Aquí podemos observar de manera muy clara en qué valores del voto nuestras estimaciones (promedio de las imputaciones) difieren con las estimaciones realizadas por el CIS, así pues vemos que son más elevadas para el porcentaje en blanco, Podemos y en menor medida IU y Ciudadanos. Además, observamos que apenas existe diferencia cuando hablamos de las estimaciones del porcentaje a Otros partidos. Por último, vemos como en los porcentajes referidos al PSOE, PP, UPyD las estimaciones del CIS están por encima de nuestras estimaciones. Notar que hemos realizado el gráfico con el promedio de nuestras imputaciones (color amarillo), las estimaciones del CIS (color naranja) y las estimaciones obtenidas de la encuesta (color verde).

Los resultados que hemos obtenido con la encuesta de Octubre 2014 se muestran en la tabla que sigue:

INTENCIÓN VOTO	ENCUESTA OCTUBRE	INPUT 1	INPUT 2	INPUT 3	PROMEDIO INPUT
PP	290	364	351	354	356
%PP	11,7%	14,7%	14,2%	14,3%	14,4%
PSOE	354	441	446	433	440
%PSOE	14,3%	17,8%	18,0%	17,5%	17,7%
IU	91	108	113	109	110
%IU	3,7%	4,4%	4,6%	4,4%	4,4%
UPYD	53	70	82	70	74
%UPYD	2,1%	2,8%	3,3%	2,8%	3,0%
PODEMOS	437	520	505	518	514
%PODEMOS	17,6%	21,0%	20,4%	20,9%	20,7%
CIUDADANOS	37	47	52	51	50
%CIUDADANOS	1,5%	1,9%	2,1%	2,1%	2,0%
OTROS	180	222	221	223	222
%OTROS	7,3%	9,0%	8,9%	9,0%	9,0%
BLANCO	105	144	146	144	145
%BLANCO	4,2%	5,8%	5,9%	5,8%	5,8%
NINGUNO	391	564	564	578	569
%NINGUNO	15,8%	22,7%	22,7%	23,3%	22,9%
FALTANTES	542	-	-	-	-
%FALTANTES	21,9%	-	-	-	-

En la tabla anterior se muestran los porcentajes referidos al total teniendo en cuenta el valor Ninguno, mientras que en la siguiente tabla los porcentajes están realizadas sin tener en cuenta la clasificación Ninguno, para poder compararlos de forma más correcta con las estimaciones realizadas por el CIS.

ENCUESTA OCTUBRE	INPUT1	INPUT2	INPUT3	Promedio Imputaciones	ESTIMACIÓN CIS
PP	19,0%	18,3%	18,6%	18,6%	27,5%
PSOE	23,0%	23,3%	22,8%	23,0%	23,9%
IU	5,6%	5,9%	5,7%	5,8%	4,8%
UPYD	3,7%	4,3%	3,7%	3,9%	4,1%
PODEMOS	27,1%	26,4%	27,2%	26,9%	22,5%
CIUDADANOS	2,5%	2,7%	2,7%	2,6%	2,1%
OTROS	11,6%	11,5%	11,7%	11,6%	12,2%
BLANCO	7,5%	7,6%	7,6%	7,6%	3,0%

Las tablas tiene el mismo formato que las anteriores. Los resultados obtenidos son similares, existiendo diferencias notables entre los porcentajes referidos al PP y Podemos. En cuanto al PSOE los porcentajes entre las estimaciones del CIS y el promedio de las imputaciones se igualan cuando realizamos dichos porcentajes sobre

el total pero sin la clasificación Ninguno. Para comparar las estimaciones realizadas por el CIS y las obtenidas en este proyecto debemos observar la segunda tabla mostrada anteriormente. El gráfico correspondiente a dicha tabla se muestra a continuación:

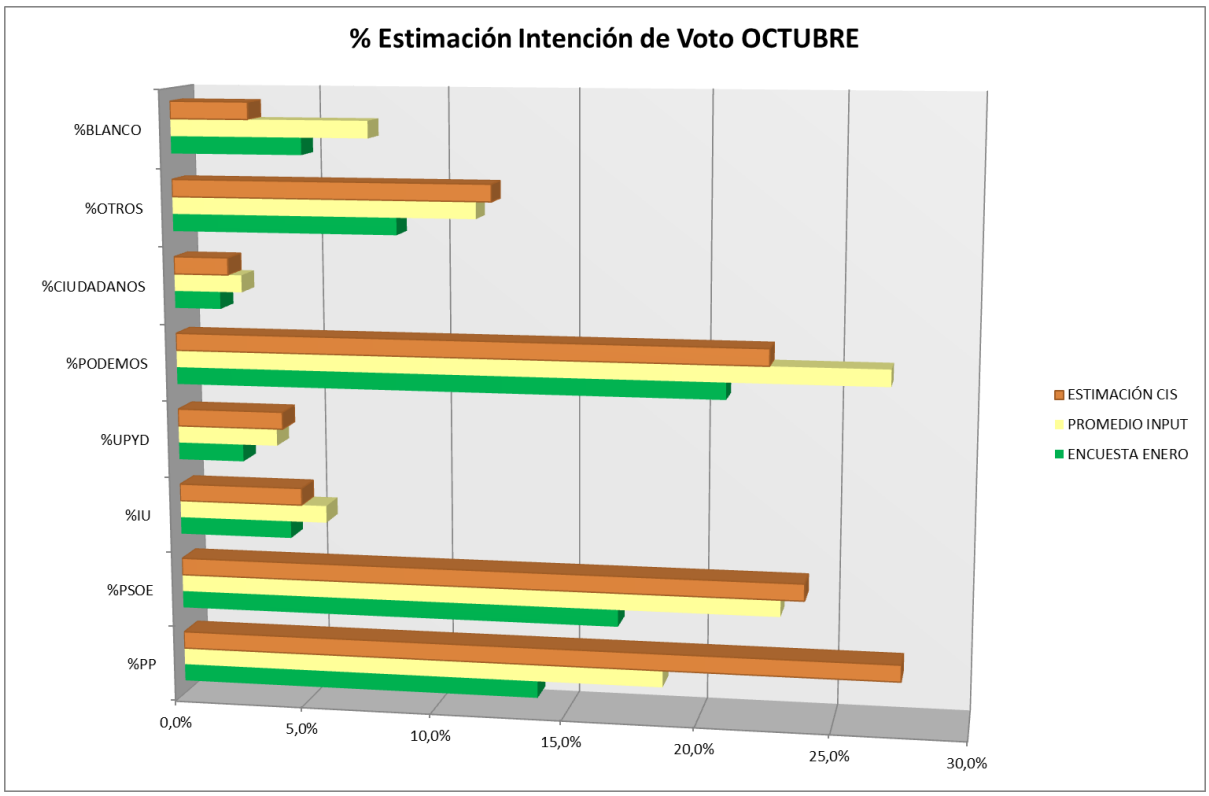

Figura 9 Comparación Octubre Fuente: Elaboración propia.

Al igual que en la figura 8 se puede observar de manera clara que los valores de voto que obtienen una estimación mayor por parte del CIS son, el porcentaje referido al PP, PSOE y Otros. En los porcentajes concerniente a voto en blanco, Ciudadanos, IU y Podemos las estimaciones que hemos realizado en el promedio de las imputaciones obtenidas tienen un porcentaje de estimación más elevado que el CIS. En cuanto a UPyD los porcentajes de estimación son similares. Cabe destacar que el porcentaje estimado por el CIS para el PP es más elevado, de manera muy sustancial, al de nuestras imputaciones. Ocurre lo contrario con el porcentaje de estimación del partido político Podemos, ya que el promedio de las estimaciones realizadas mediante nuestras imputaciones son más elevadas que las obtenidas por el CIS.

CONCLUSIONES

Las técnicas de imputación no tienen como objetivo explicar las razones de la ausencia de ciertos valores. Su principal objetivo es crear predicciones de cada valor ausente. En este trabajo hemos implementado la imputación múltiple para la estimación de valores faltantes en las encuestas de intención de voto en las encuestas de Octubre 2014 y Enero 2015 realizadas por el CIS, mediante ecuaciones encadenadas. Obteniendo las siguientes conclusiones principales:

- Los patrones de datos faltantes en ambas encuestas son muy similares. Así como, el número de no respuestas en las diferentes variables de ambas encuestas.

- Pese al coste computacional que tienen los métodos de imputación múltiple, hemos decidido implementar dicho método debido a sus principales ventajas, concluyendo que obtenemos mejores porcentajes de acierto global en la encuesta de Enero 2015 con un 67.5% que en la encuesta realizada en Octubre 2014 con un 60,1%.

- Por último, concluimos que en comparación con las estimaciones realizadas por el CIS, estas son superiores a las obtenidas en este proyecto en los casos referidos al PP, PSOE, UPyD y Otros partidos. Sucediendo al contrario con voto en blanco, Ciudadanos, Podemos e IU. Esto se observa en ambas encuestas. Podemos destacar que el porcentaje de estimación del voto referido al PP las estimaciones del CIS son muy elevadas respecto a nuestras estimaciones. Ocurre lo mismo en las estimaciones que hemos realizado para la intención de voto de Podemos, siendo más elevadas que las estimaciones realizadas por el CIS.

BIBLIOGRAFÍA

A Community Site for R. (s.f.). Recuperado el 27 de Mayo de 2015, de http://www.inside-r.org/packages/cran/mi/docs/bugs.mi

Celeste-tel. (s.f.). *Celeste-tel. Investigaciones Sociológicas.* Recuperado el 23 de Junio de 2015, de http://www.celeste-tel.es/images/Barometroelectoralsocialjunio2015.pdf

CIS. (Octubre de 2014). *Centro de Investigaciones Sociológicas.* Recuperado el 23 de Junio de 2015, de http://www.cis.es/cis/opencms/-Archivos/Marginales/3040_3059/3041/Ft3041.pdf

Déborah, O. G. (2011). *Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo.* Recuperado el 26 de Mayo de 2015, de http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_616.pdf

El País. (22 de mayo de 2015). *El País.* Recuperado el Junio de 2015, de http://elpais.com/elpais/2015/06/06/media/1433613613_640079.html

Encuestamos. (junio de 2015). *Encuestamos*. Recuperado el Junio de 23 de 2015, de http://www.encuestamos.com/termometro-politico-podemos-supera-el-30-en-intencion-de-voto/

Escobar, M., & Jaime, A. M. (2013). Métodos de imputación múltiple para predecir resultados electorales. *The Stata Journal*.

Lewis-Beck, M. (2005). Election forecasting principles and practice. *The British Journal of Politics & International Relations.*, 145-164.

Mediavilla, M. (Abril, 2012). *Método de imputación de los valores no observados.* Valencia.

Muñoz Rosas, J. F., & Álvarez Verdejo, E. (2009). Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus. *Revista de métodos cuantitativos para la economía y la empresa*, 3-30.

myWord. (s.f.). *myWord.Social and Market Research.* Recuperado el 23 de Junio de 2015, de http://myword.es/wp-content/uploads/2015/04/Informe-de-resultados-ObSERvatorio-abril-2015.pdf

R proyect. (s.f.). *R proyect*. Recuperado el 23 de Junio de 2015, de http://www.r-project.org/

Restrepo Estrada, M. I., & Marín Diazaraque, J. M. (2012). Imputación de ingresos en la Gran Encuesta Integrada de Hogares (GEIH) de 2010. *Desarrollo y Sociedad*, 219-243.

Rivas, C., Martínez, M., & Galindo, P. (2010). La imputacion múltiple como alternativa al análisis de la no respuesta en la variable intención de voto. *Revista española de ciencia política*, 99-118.

Rivero Rodríguez, G. (2011). *Análisis de datos incompletos en Ciencias Sociales.* Colección cuadernos metodológicos.

Royuela Vicente, A. (2014). *Desarrollo y aplicación de modelos pronósticos en patología lumbar.* Madrid.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York.

Yu-sung, S., Gelman, A., Hill, J., & Yajima, M. (2011). *Multiple Imputation with Diagnostics (mi) in R.* Journal of Statistical Software.

ANEXO

1. Programa para la imputación de los datos.

Mostramos el código de elaboración propia para la imputación de los datos faltantes, hemos utilizado el Software estadístico R.

```
#Instalación de la librería mi.
install.packages("mi")
library(mi)

#Carga de datos, los datos están en formato csv2.

#Realizamos unas breves comprobaciones acerca del tamaño de nuestros datos.
x=read.csv2(choose.files(),header=TRUE)
n=dim (x)[1]
p=dim (x)[2]
n
p
head(x)
summary(x)
colnames(x)

#Gráfico de los patrones sobre los datos faltantes.
mp.plot(x, x.order = TRUE,y.order = TRUE, clustered = FALSE,main="Octubre")

#Información de los datos.
info<-mi.info(x)
info

#Recodificación de las variables.
info.upd <- update(info, "type", list("voto" = "unordered-categorical","Ppp" = "unordered-categorical","Ppsoe" = "unordered-categorical","Piu" = "unordered-categorical","Pupyd" = "unordered-categorical","Ppodemos" = "unordered-categorical","Pciudadanos" = "unordered-categorical","recuerdo"="unordered-categorical","simpatia"="unordered-categorical","ideologia"="ordered-categorical"))

#Obtención de  la fórmula con la que se realiza el modelo.
info.upd$imp.formula
```

```
#Modelo.
info<-update(info.upd,"imp.formula",list("voto"="voto ~ sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad"))
info$imp.formula["voto"]

#Imputación de los datos. La imputación tiene un gran coste computacional.
IMP <- mi(x,n.iter=20,max.minutes=20,info.upd)

#Grafico interactivo de la imputación.
plot(IMP)

#Nueva imputación, para eliminar el posible ruido de los datos.
IMP <- mi(IMP, run.past.convergence = TRUE, n.iter=10,max.minutes=10)

#Guardaremos los datos.
IMP.dat.all <- mi.completed(IMP)
IMP.dat <- mi.data.frame(IMP,m=1)
write.mi(IMP,format="table")
IMP.dat["voto"]

#Regresión del modelo y las imputaciones.
fit <-glm.mi(voto ~ sit.econ + sit.politica + Ppp + Ppsoe + Piu + Ppodemos + Pciudadanos + Pupyd + gobierno + oposición + simpatia + ideologia + recuerdo + sexo + edad,IMP)
display(fit)
```

2. Programa para la evaluación de las imputaciones.

Se anexa el programa realizado para la evaluación de nuestras imputaciones. Dicho programa es de elaboración propia.

```
#Cargar los datos y realizamos comprobaciones básicas.
data=read.csv2(choose.files(),header=TRUE)
n=dim (data)[1]
p=dim (data)[2]
n
p
head(data)
summary(data)

#Realizamos la división de la muestra en varias submuestras.
#me es la submuestra de entrenamiento.
#mp es la submuestra de prueba.
#Xprueba es la submuestra en la que guardamos los votos de la
submuestra prueba.
#XNUEVO es la unión de las submuestras de entrenamiento y prueba
con voto vacío.

set.seed(134)
train<-sample(1:n,2000)
me<-data[train,]
mp<-data[-train,]
head(me)
dim (me)
summary(me)
head(mp)
dim (mp)
Xprueba=mp
mp[,"voto"]=NA
mp[,"voto"]
head(mp)
Xprueba["voto"]
head(Xprueba)
XNUEVO=rbind(me,mp)
head(XNUEVO)
dim(XNUEVO)
```

```
summary(XNUEVO)

#Gráfico de los patrones sobre los datos faltantes.
mp.plot(XNUEVO, x.order = TRUE,y.order = TRUE, clustered = FALSE,main="TEST Octubre")

#Información de los datos.
info<-mi.info(XNUEVO)
info
info.upd <- update(info, "type", list("voto" = "unordered-categorical","Ppp" = "unordered-categorical","Ppsoe" = "unordered-categorical","Piu" = "unordered-categorical","Pupyd" = "unordered-categorical","Ppodemos" = "unordered-categorical","Pciudadanos" = "unordered-categorical","recuerdo"="unordered-categorical","simpatia"="unordered-categorical","ideologia"="ordered-categorical"))
info.upd
info.upd$imp.formula

#Modelo.
info<-update(info.upd,"imp.formula",list("voto"="voto ~ sit.econ + sit.politica + Ppp + Ppsoe + Piu + Pupyd + Ppodemos + Pciudadanos + gobierno + oposicion + simpatia + ideologia + recuerdo + sexo + edad"))
info$imp.formula["voto"]

#Imputación de los datos. La imputación tiene un gran coste computacional.
IMP <- mi(XNUEVO,n.iter=20,max.minutes=10,info.upd)

#Gráfico de las imputaciones.
plot(IMP)

#Realizar una nueva imputación para eliminar el ruido.
IMP <- mi(IMP, run.past.convergence = TRUE, n.iter=10,max.minutes=5)

#Guardar los datos.
IMP.dat.all <- mi.completed(IMP)
IMP.dat <- mi.data.frame(IMP,m=1)
write.mi(IMP,format="table")
IMP.dat["voto"]
```

```
#Crear la matriz llamada votos para unir la intención de voto de
xnuevo y las imputaciones.

VOTOS=cbind(XNUEVO["voto"],IMP.dat["voto"])
dim(VOTOS)
head(VOTOS)
summary(VOTOS)

#Se guarda en un archivo CSV.
write.csv(VOTOS,file="votos.csv")
```