ESTIMACIÓN DE LA RESPUESTA DE LOS «NO SABE/NO CONTESTA» EN LOS ESTUDIOS DE INTENCIÓN DE VOTO

Jesús Varela Mallou, Teresa Braña Tobío,
Alberto García Carreira y Antonio Rial Boubeta
Universidad de Santiago

Xosé Gabriel Vázquez Fernández
Universidad de A Coruña

RESUMEN

La historia electoral en España nos pone en evidencia, generalmente desfavorable, la escasa validez, exactitud y sentido de algunas predicciones hechas por las encuestas electorales. Junto con problemas de representatividad sociopolítica o errores en la estructura de la entrevista, una fuente de error en dicha predicción es el escaso tratamiento estadístico de la falta de respuesta o datos *missing*. En este trabajo utilizaremos una encuesta preelectoral llevada a cabo en la Comunidad Gallega en julio de 1997, para evaluar la eficacia de cuatro tratamientos habituales de la falta de respuesta. Los procedimientos que vamos a evaluar son: *listwise,* asignación proporcional, análisis discriminante y Hot-Deck. Los resultados muestran que el Hot-Deck, combinado con el análisis discriminante, es el método más eficaz para estimar el voto probable de los sujetos que no se han pronunciado sobre la intención de voto.

INTRODUCCIÓN

En cualquier área de investigación, a la hora de realizar el análisis de los datos es frecuente encontrarse con variables que presentan valores ausentes o perdidos *(missing data).* Cuando esta falta de información no es demasiado grande y se encuentra distribuida al azar, no hay demasiado problema, pero si esto no es así, normalmente se requiere algún tipo de estrategia de sustitución antes de proceder al análisis de los datos.

Sorprendentemente, y a pesar de que el problema de la «no respuesta» apa-

rece como algo habitual en los sondeos de opinión, este tema ha recibido poca atención en psicología (Cruz, 1990; Roth, 1994; Roth, Campion y Jones, 1996). Ante esta situación, los investigadores a menudo optan por presentar sólo la información con datos completos, o eliminar de sus estudios aquella información que no aparece de manera completa: a esto contribuye la opción de *listwise* que viene definida por defecto en los distintos paquetes estadísticos. Afortunadamente, otras áreas de investigación, como el *marketing,* la economía, la educación, la psicometría o la estadística, han ayudado al desarrollo de la metodología en el tratamiento de la falta de información.

Las razones por las cuales la gente no contesta, sin descartar la influencia de un cuestionario defectuoso, el comportamiento inadecuado del entrevistador u otros factores como la suspicacia o el temor por parte del entrevistado, hay que buscarlas en el tema de objeto de la encuesta y la naturaleza de algunas cuestiones (Martín, 1981). Los individuos responden mejor a preguntas sobre hechos o acontecimientos que a las preguntas de opinión, pero la proporción de «no respuesta» se verá incrementada si las opiniones que se desean conocer se refieren a cuestiones que podríamos denominar trascendentales, o al menos importantes, en materia económica o política.

Los valores perdidos, normalmente, originan dos tipos de problemas:

1. Una reducción del tamaño de la muestra que disminuye la precisión de los análisis. Si el número de valores perdidos es lo suficientemente amplio, estamos desestimando las respuestas de un alto porcentaje de individuos de la muestra, con lo que se puede llegar a invalidar el análisis simplemente por problemas de representatividad (Roth y col., 1996).
2. Los valores perdidos provocan sesgos en la estimación de los estadísticos independientemente del tamaño muestral. En algunos casos, los valores perdidos hacen disminuir los coeficientes de correlación, y afectan a otros estadísticos como las medidas de tendencia central, las medidas de dispersión, etc. (Donner, 1982; Little y Rubin, 1987).

Aunque disponemos de diferentes métodos para el tratamiento de la falta de respuesta, no existen guías que orienten a los investigadores sobre cuál es el procedimiento más adecuado a los datos con los que trabaja. Antes de comentar el principio básico de los más habituales, es conveniente señalar que el mejor método para trabajar con valores perdidos es evitar que se produzcan, planeando y recogiendo los datos de una forma controlada (Green y Carrol, 1986). Sin embargo, en ocasiones resulta inevitable tener que trabajar sin la totalidad de los datos, y es entonces cuando debemos proceder a su análisis.

A continuación se recogen los principales tratamientos utilizados habitualmente cuando se trabaja con valores ausentes.

Listwise

Elimina todos los casos que tengan un valor perdido en alguna de las variables seleccionadas. Aunque este procedimiento es utilizado por muchos investigadores, y es la opción por defecto de numerosos paquetes estadísticos (Little y Rubin, 1987), sacrifica una gran cantidad de datos. Un ejemplo puede verse en un trabajo de Kaufman (1988), donde investigadores de *marketing* pasan de una muestra inicial de 624 sujetos a 201 utilizando este procedimiento.

Supongamos que tenemos las puntuaciones de cinco sujetos en tres pruebas psicométricas que aparecen en la siguiente tabla, donde el 9 representa un *missing* (ver tabla 1).

TABLA 1

Puntuaciones obtenidas por los sujetos

n	*Variable 1*	*Variable 2*	*Variable 3*
1	1	5	6
2	2	4	9
3	5	9	1
4	5	1	9
5	4	2	2

Si eliminamos aquellos casos que tengan un valor perdido en alguna de las tres variables, entonces la muestra se reduciría a dos sujetos, lo que supone una gran pérdida de información.

Pairwise

Este procedimiento elimina aquellos casos con valores perdidos en alguna de las variables que intervengan en el análisis. Su principal ventaja con respecto al *listwise* es que preserva una mayor cantidad de información, aunque a cambio hace complicada la interpretación. Por ejemplo, en el caso de los coeficientes de correlación, cada coeficiente tendrá un tamaño muestral diferente (Kim y Curry, 1977; Roth, 1994).

En el ejemplo anterior, si queremos calcular el coeficiente de correlación entre la variable 1 y la variable 3, eliminaríamos aquellos casos con valores perdidos en estas variables; es decir, el sujeto 2 y el 4. Mientras que si correlacionásemos las variables 1 y 2, sólo se eliminaría el sujeto 3.

SUSTITUCIÓN POR LA MEDIA

Esta práctica permite a los investigadores utilizar el valor medio de una variable para sustituir la no respuesta en esa variable. Si estamos estudiando las puntuaciones de unos sujetos en una prueba de inteligencia, sustituiríamos el valor *missing* por la media alcanzada por el grupo.

Mientras que la sustitución por la media mantiene el tamaño de la muestra, y es fácil de utilizar, tiende a disminuir la varianza muestral, puesto que las puntuaciones desconocidas se sustituyen por la media del grupo. Por ejemplo, un investigador podría tener 30 sujetos de lo cuales 5 tienen valores perdidos. La sustitución de medias sugiere que sustituyamos los cinco valores por la media de la variable. De esta forma, estamos aumentando el tamaño muestral, pero no aumentaría la varianza de la muestra con respecto a la media, con lo que las covarianzas se verían atenuadas (Little, 1988; Roth 1994).

Cuando trabajamos con variables de tipo nominal, el procedimiento equivalente sería la asignación proporcional al tamaño de los grupos. Supongamos que estamos estudiando el tipo de prensa que leen los trabajadores de una empresa. Para eso preguntamos a 30 sujetos cuál es el periódico que compran habitualmente, y nos encontramos que 10 sujetos no responden a nuestra pregunta. Si queremos asignar proporcionalmente a los sujetos a los distintos diarios, basta con repartirlos en función de los porcentajes que tienen en el resto de los trabajadores (ver tabla 2).

TABLA 2

Procedimiento de asignación de los que no responden mediante la asignación proporiconal

n	*Periódico*	*%*	*Sujetos asignados*	*Total*
8	*La Voz de Galicia*	40	4	12
6	*El País*	30	3	9
4	*El Mundo*	20	2	6
2	*ABC*	10	1	3
10	No sabe/no contesta			

IMPUTACIÓN A TRAVÉS DE LA REGRESIÓN

La reasignación de las respuestas de los que no contestan, basándonos en las estrategias de regresión, va a depender de las escalas de medida de las variables que vayamos a utilizar en el análisis de los datos. En la tabla 3 se resume cuáles son las técnicas apropiadas, dependiendo de las escalas de medida de las variables que tengamos.

TABLA 3

Variantes de la regresión dependiendo de la escala de las variables

- Cuando tanto la variable dependiente («fuerza de intención de voto a cada partido») como las variables independientes sean continuas, el método que utilizaremos será la *regresión múltiple mínimo cuadrática.*
- Cuando la variable dependiente sea dicotómica (participación-abstención) y las variables independientes continuas, *análisis discriminante simple, modelos probit, regresión logística.*
- Variable dependiente nominal («intención de voto en las próximas elecciones») y variables independientes continuas, *análisis de discriminante múltiple.*
- Variable dependiente continua («fuerza de intención de voto a cada partido») y variables independientes nominales, *análisis de clasificación múltiple.*
- Variable dependiente dicotómica (participación-abstención») y variables independientes nominales, *modelos logit.*

Dentro de las ventajas de este procedimiento se ha señalado que mantiene una gran cantidad de datos, y no distorsiona ni varianza muestral ni la forma de la distribución (Little, 1988).

Imputación según Fichero Caliente (Hot-Deck)

El principio básico de este procedimiento consiste en sustituir el valor perdido de un sujeto por el valor que en esa misma variable obtiene otro sujeto lo más parecido a éste en otras variables del fichero. Para eso, agrupamos a los sujetos en función de la puntuación que obtienen en las variables que se hayan mostrado más relacionadas con la que queremos predecir, y asignamos al valor que desconocemos el del sujeto más próximo en esa variable.

Supongamos que queremos conocer la relación entre la ideología política medida en un continuo (extrema izquierda-extrema derecha) y la opinión acerca del aborto medida en una escala de 1 a 9, donde el 9 representa la máxima aceptación. Después de pasar los cuestionarios a diez sujetos obtenemos la siguiente matriz de datos, donde el cero representa la no respuesta (ver tabla 4).

Los individuos 4 y 9 no han contestado a la opinión sobre el aborto (0 = NS/NC). Según el procedimiento del fichero caliente, el sujeto 4 recibiría un 2 en la opinión sobre el aborto, ya que es la puntuación obtenida por el individuo más próximo a él y que también recibe un 8 en ideología política. El individuo 9, recibiría un 7, que es el valor del individuo más próximo y que también tiene un 3 en ideología política.

Los defensores de este procedimiento argumentan que se tiende a incrementar la exactitud con respecto a los métodos tradicionales (*listwise* y *pairwise*), ya que los valores perdidos son reemplazados por valores reales. En cambio, el principal inconveniente de este procedimiento es la falta de apoyo teórico y empírico que determine su fiabilidad (Roth, 1994).

TABLA 4

Puntuaciones de los sujetos en cuanto a la ideología política y a la opinión sobre el aborto

n	*Ideología política*	*Opinión sobre el aborto*
1	2	7
2	4	5
3	3	7
4	8	0
5	4	7
6	8	2
7	7	2
8	9	1
9	3	0
10	4	7

LA «NO RESPUESTA» EN LOS ESTUDIOS DE INTENCIÓN DE VOTO

Desde hace unos veinte años, en España se ha venido produciendo un auge espectacular de los sondeos realizados con motivo de la celebración de alguna elección, ya sea de tipo general, autonómica, municipal o las más recientes elecciones europeas. Este auge se ha traducido en un aumento del número de Institutos de Investigación que realizan este tipo de encuestas, así como un aumento de publicaciones de las mismas por medios de comunicación de masas (Fernández Santana, 1994; Sanz de la Tajada, 1996).

No obstante, la historia electoral en España nos pone en evidencia, generalmente desfavorable, la escasa validez, exactitud y sentido de las predicciones hechas por las encuestas electorales. Las encuestas preelectorales han sido ampliamente discutidas tras las pasadas elecciones generales de marzo de 1996, en las que los indecisos (NS/NC) han dado la vuelta a las estimaciones de las empresas demoscópicas (ver tabla 5).

La tabla anterior sugiere la siguiente pregunta: ¿qué es lo que hace que los estudios de intención de voto no acierten en los resultados finales?

Los problemas fundamentales de los estudios preelectorales podrían resumirse en los siguientes puntos:

a) La no correspondencia exacta entre la intención de voto y el voto real.

b) La posible evolución del voto entre la fecha del último sondeo y el día de las elecciones. No se puede olvidar que los resultados de la encuesta se refieren a un momento determinado y anterior en el tiempo al día de los comicios.

TABLA 5

El fallo en las encuestas preelectorales de marzo de 1996

	PP	*PSOE*	*IU*	*CiU*	*PNV*
Resultados reales	156	141	21	16	5
Colpisa	173/181	116/130	20/27	14/16	5/6
El País	170/178	118/125	24/27	14/16	5/6
ABC	176/184	117/125	22/26	13/14	5/6
El Mundo	170/179	113/123	25/29	14/15	6/7
La Vanguardia	160/170	135/145	19/21	14/15	5/6
Faro de Vigo	160/170	125/135	25/-	15/-	-/-

FUENTE: *La Voz de Galicia,* 5 de marzo de 1996.

c) Peculiaridad de cada sistema electoral. Concretamente, se ha señalado que el sistema electoral español hace muy complicada la predicción de los resultados electorales (Díez Nicolás, 1996).

d) El problema de la representatividad no sólo sociológica, sino también sociopolítica. El primer paso para poder generalizar los resultados de una encuesta consiste en garantizar que el tamaño muestral sea el apropiado, así como establecer los criterios adecuados de representatividad sociopolítica. Si bien la mayoría de los trabajos garantizan la representatividad sociológica, son pocos los trabajos que se preocupan por la representatividad sociopolítica. Sanz de la Tajada (1996) propone corregir los resultados en función del comportamiento electoral anterior.

e) El diseño del cuestionario. Es fundamental la adecuación de las preguntas a los objetivos que se persiguen. También hay que tener en cuenta factores contextuales, como la alta sensibilidad social a las cuestiones de tipo sociopolítico y el elevado desinterés por la política. A la hora de seleccionar los ítems es necesario tener en cuenta que, para el encuestado, no es lo mismo definir una preferencia que expresar una intención de voto, que supone un mayor compromiso y, por tanto, genera una mayor tasa de no respuesta.

f) El escaso tratamiento estadístico de la falta de respuesta o datos *missing.* Como ya habíamos dicho anteriormente, los individuos suelen mostrarse más reacios a dar su opinión acerca de las cuestiones políticas y, sobre todo, en el caso de «voto en las últimas elecciones», así como «la intención de voto futuro».

A nuestro juicio, la verdadera capacidad predictiva de un sondeo preelectoral debe controlar todos estos posibles sesgos. A pesar de que los tres primeros son difícilmente controlables, el esfuerzo de los investigadores debe centrarse en prestar atención a los restantes. Precisamente, el objetivo del presente trabajo se centra en el último de estos sesgos: el tratamiento de los indecisos.

OBJETIVO

En este trabajo evaluaremos la eficacia de cuatro métodos de tratamiento de los indecisos (NS/NC) con datos reales, en una encuesta preelectoral realizada en julio de 1997, sobre intención de voto en las elecciones autonómicas gallegas.

Para ello, sobre la encuesta original provocaremos un 20 por 100 de falsos *missing* con objeto de estimarlos y comparar las puntuaciones que hemos predicho con las originales. Esperamos proponer una metodología de estimación de la respuesta más posible o probable que nos permita asignar a una determinada opción política a todos aquellos sujetos que no se han pronunciado sobre la intención de voto.

MÉTODO

Procedimiento

Para la comparación de los cuatro métodos, nos basaremos en las respuestas dadas a la pregunta: «*Si mañana se celebrasen elecciones a la Xunta de Galicia, para elegir al gobierno gallego, ¿a qué partido o coalición votaría Vd.?*». Estos métodos son los siguientes:

1. Procedimiento de eliminación (*listwise* y *pairwise*). Dado que sólo se trabaja con una variable, los dos procedimientos de eliminación actuarían del mismo modo, con lo que no haremos distinción entre ambos.
2. Asignación proporcional. Como la variable dependiente es de tipo categórico, no tiene sentido utilizar la media como procedimiento de imputación, por lo que aplicaremos la asignación proporcional.
3. Análisis Discriminante. Este tipo de análisis es el más adecuado para trabajar con las escalas de medida que tenemos.
4. Fichero Caliente (Hot-Deck). Para la selección de las variables de categorización utilizaremos las cuatro que, en el método anterior, ofrezcan mayor poder discriminante.

Antes de proceder a la aplicación de los cuatro métodos de estimación propuestos es necesario depurar el fichero de datos. Para ello, en primer lugar, eliminamos aquellos sujetos de los que tenemos dudas de su sinceridad. Más concretamente, aquellos sujetos que en las preguntas filtro P1 («*En las últimas elecciones generales, concretamente las celebradas en marzo de 1996, ¿fué Vd. a votar?*») y P1a («*¿Recuerda Vd. por qué partido o candidatura lo hizo?*») hayan respondido «no recuerda» o «no sabe/no contesta», entendemos que no quieren colaborar en el estudio, por lo que mejor es eliminarlos. Y, en segundo lugar, excluiremos a los verdaderos casos *missing* en la pregunta objeto de análisis (P2: «*Si mañana se celebrasen elecciones a la Xunta de Galicia, para elegir al gobierno gallego, ¿a qué partido o coalición votaría Vd.?*»). Después de esto,

generamos aleatoriamente entre los sujetos que nos quedan un 20 por 100 de valores *missing,* guardando las puntuaciones que tenían antes de ser seleccionados en una nueva variable (NP2), con el objeto de calcular las diferencias entre las estimaciones de los cuatro métodos señalados y los valores reales.

Muestra

La muestra está compuesta por 2.904 sujetos de 18 o más años de edad, a los que se les realizaron entrevistas personales en el hogar del entrevistado durante la primera semana del mes de julio. Para la selección de los sujetos se ha seguido un muestreo polietápico por conglomerados, fijo por provincias, proporcional por tamaño municipal, y aleatorio por secciones censales para la localización del hogar y por cuotas de edad y sexo para el entrevistado.

La entrevista se basaba en la aplicación de un cuestionario de 25 preguntas que hacían referencia a cuestiones de tipo político, aspectos económicos, valoración de los líderes políticos, etc.

De los 2.904 sujetos iniciales fueron eliminados, mediante las preguntas filtro, 548 sujetos por considerar, de acuerdo con lo anteriormente expuesto, que no quisieron colaborar en el estudio. Después de esto, la muestra quedó formada por 2.336 sujetos. De estos sujetos eliminamos aquellos que no respondieron a la pregunta P2 («S*i mañana se celebrasen elecciones a la Xunta de Galicia, para elegir al gobierno gallego, ¿a qué partido o coalición votaría Vd.?*»), con lo que la muestra queda finalmente compuesta por 1.829 sujetos (ver tabla 6). En este caso, la eliminación de todos estos sujetos no es un problema, ya que el objetivo de este estudio no es realizar una predicción del voto, sino comparar la eficacia de los distintos métodos de tratamiento de los valores perdidos, por lo que es necesario tener la respuesta de todos los sujetos en la pregunta sobre intención de voto.

TABLA 6

Distribución de frecuencias de la variable NP2 con un 18,9 por 100 de falsos casos «missing»

		Variable P2		*Variable NP2*	
	Valor	*N.º real*	*%*	*N.º simulado*	*%*
BNG	1	379	20,7	297	16,2
PSOE/EU/VERDES	2	361	19,7	300	16,4
PP	3	890	48,7	720	39,4
Otros	4	11	0,6	10	0,5
Abstención (no voto)	5	188	10,3	157	8,6
Missing	·	0	0	345	18,9
Total		1.829	100	1.829	100

Análisis de datos

1. *Procedimiento de eliminación*

Desde esta perspectiva, el tratamiento que se les va a dar a los no sabe/no contesta *(missing)* es, simplemente, no considerarlos. Como podemos ver en la tabla 7, cuando utilizamos el procedimiento de eliminación, de los 1.829 sujetos que teníamos para hacer el estudio, nos quedamos con 1.484 sujetos. Esta reducción de la muestra tiene dos tipos de implicaciones. En primer lugar, la pérdida de representatividad sociológica (en cuanto que se reduce el tamaño muestral) y, sobre todo, pérdida de representatividad política al asumir que los 345 valores *missing* se distribuyen de la misma manera que los 1.484 restantes. Y, en segundo lugar, que estamos prescindiendo del resto de la información de esos sujetos, que podría resultar útil para estructurar el electorado, conocer cómo valoran a los líderes políticos y a los diferentes partidos, etc.

TABLA 7

Variable NP2 analizada mediante un procedimiento de eliminación

	Antes de eliminar		*Después de eliminar*	
	Número	*%*	*Número*	*%*
BNG	297	16,2	297	20
PSOE/EU/VERDES	300	16,4	300	20,2
PP	720	39,4	720	48,5
Otros	10	0,5	10	0,7
Abstención (no voto)	157	8,6	157	10,6
Missing	345	18,9	0	0
TOTAL	1.829	100	1.484	100

2. *Asignación proporcional al número de votos*

En el caso de la asignación proporcional, vamos a repartir el número de *missing* entre las diferentes candidaturas dependiendo del número de votos de cada una de ellas. Con este procedimiento resulta más complicado conocer qué error estamos cometiendo, ya que a pesar de que los porcentajes finales que obtenemos son prácticamente idénticos a los que teníamos en P2 (ver tablas 8 y 9), sin embargo, el procedimiento no garantiza que a cada caso le corresponda el valor que tenía antes de convertirlo en falso *missing,* sino que se asignan en función del porcentaje válido, tal y como veremos en el apartado de resultados.

TABLA 8

Asignación proporcional de los indecisos a la pregunta NP2

	n	*Porcentaje válido*	*Asignados*	*n total*
BNG	297	20,0	69	366
PSOE/EU/VERDES	300	20,2	70	370
PP	720	48,5	167	887
Otros	10	0,7	2	12
Abstención	157	10,6	37	194
Missing	345			
TOTAL	1.829	100	345	1.829

TABLA 9

Diferencia entre los valores reales y los predichos mediante la reasignación proporcional

	Voto real P2		*Votos asignados NP2*		*Diferencia*	
	n	*%*	*n total*	*%*	*n*	*%*
BNG	379	20,7	366	20,0	13	0,7
PSOE/EU/VERDES	361	19,7	370	20,2	–9	–0,5
PP	890	48,7	887	48,5	3	0,2
Otros	11	0,6	12	0,7	–1	–0,1
Abstención	188	10,3	194	10,6	–6	–0,3
TOTAL	1.829	100	1.829	100	0	0

Como podemos ver en la tabla 9, las diferencias entre el voto real (P2) y el voto estimado de los *missing* no son muy grandes, excepto en el BNG y la coalición. No obstante, hay que señalar que si bien estos datos son válidos para el conjunto de la muestra, falta por analizar qué sucede cuando la estimación se produce a nivel de provincias.

3. *Imputación a través del Análisis Discriminante*

En primer lugar, tuvimos que seleccionar las variables que entran a formar parte del Análisis Discriminante. Para ello, efectuamos diversos análisis de varianza entre la intención de voto y aquellas variables de tipo continuo del cuestionario. Los estadísticos Chi-cuadrado y Coeficiente de Contingencia fue-

ron utilizados para seleccionar las variables de tipo nominal que están más relacionadas con la intención de voto (ver tablas 10 y 11, respectivamente).

TABLA 10

Valores de F de las variables del cuestionario

Variable	*Valor de F*	*Significatividad*	*Variable*	*Valor de F*	*Significatividad*
P4a	99,3	0,0001	P54a	159,5	0,0001
P4b	19,13	0,0001	P54b	35,0	0,0001
P4c	293,82	0,0001	P54c	54,8	0,0001
P4d	126,01	0,0001	P55a	240,24	0,0001
P51a	82,9	0,0001	P55b	51,71	0,0001
P51b	38,88	0,0001	P55c	66,47	0,0001
P51c	49,54	0,0001	TABEL	55,02	0,0001
P52a	225,41	0,0001	TBEIRAS	72,55	0,0001
P52b	43,36	0,0001	TFRAGA	257,85	0,0001
P52c	63,71	0,0001	P11	208,82	0,0001
P53a	166,45	0,0001	P12	25,07	0,0001
P53b	38,06	0,0001	P18	31,56	0,0001
P53c	48,44	0,0001			

TABLA 11

Valores de Chi-cuadrado y Coeficientes de Contingencia

Variable	*Chi-cuadrado*	*Significatividad*	*Coef. de cont.*	*Significatividad*
P6	665,72	0,0001	0,5721	0,0001
P7	260,05	0,0001	0,4221	0,0001
P8	208,20	0,0001	0,4260	0,0001
P9	690,07	0,0001	0,5771	0,0001
P10	126,67	0,0001	0,2917	0,0001
P1cod	218,81	0,0001	0,3584	0,0001

Con objeto de hallar una función de clasificación, se aplicó el Análisis Discriminante tomando como variables predictoras todas las consideradas en las tablas 10 y 11, teniendo en cuenta que algunas tuvieron que ser transformadas ya que el Análisis Discriminante Múltiple requiere que las variables predictoras sean métricas o ficticias.

En la tabla 12 se ofrecen, por orden de su capacidad predictiva, las diez variables que han sido seleccionadas para formar parte de la Función Discriminante. Dicha función es capaz de clasificar correctamente el 80,19 por 100 de todos los casos.

TABLA 12

Variables seleccionadas en el Análisis Discriminante de la intención de voto

Variable	*F de salida*	*Lambda de WILKS*
Labor del PP	11,6858	0,42780
Labor del PSOE	43,6494	0,26643
Labor del BNG	18,1483	0,19969
Satisfecho con la política del gobierno	11,1335	0,15480
Participación en las generales del 96	13,6301	0,14174
Valoración global de Fraga	6,8634	0,13465
Ideología izquierda-derecha	6,3235	0,12864
Aprueba la labor de Fraga	6,0893	0,12372
Utilidad para Galicia de Abel Caballero	7,2238	0,12009
Utilidad para Galicia de Beiras	5,4938	0,11584

A continuación, se muestran los valores de cada opción política como resultado de aplicar el análisis discriminante (ver tabla 13).

TABLA 13

Asignación a través del procedimiento del Análisis Discriminante.

	n real	*n simulado*	*Asignados*	*n total*	*Diferencia*
BNG	379	297	67	364	15
PSOE/EU/VERDES	361	300	65	365	–4
PP	890	720	185	905	–15
Otros	11	10	3	13	–2
Abstención	188	157	25	182	6
Missing		345			
TOTAL	1.829	1.829	345	1.829	0

Como podemos observar en la tabla anterior, existen diferencias bastante grandes en la estimación del voto de dos partidos políticos, el PP y el BNG, en tanto que las estimaciones para las restantes opciones aparecen como buenas.

4. *Imputación a través del Hot-Deck*

Para realizar la imputación o sustitución de valores, a través del fichero caliente, seleccionamos las cuatro variables que se han mostrado como mejores predictoras de la intención de voto en el análisis discriminante de la tabla 12. A continuación, ordenamos el fichero de datos y para cada *missing* asignamos el valor de intención de voto del sujeto más próximo a él, con las mismas puntuaciones en las variables discriminantes. Cuando decimos que le asignamos la misma intención de voto manifestada por el sujeto «más próximo» hay que señalar que la «proximidad» viene dada por la pertenencia a los conglomerados muestrales definidos en función de las variables: *provincia, municipio, tamaño de municipio, sección y distrito censal.* En la tabla 14 aparece la asignación a los valores *missing* a través de este procedimiento.

TABLA 14

Asignación a través del procedimiento del fichero caliente (Hot-Deck)

	n real	*n simulado*	*Asignados*	*n total*	*Diferencia*
BNG	379	297	83	380	–1
PSOE/EU/VERDES ..	361	300	61	361	0
PP	890	720	169	889	1
Otros	11	10	1	11	0
Abstención	188	157	31	188	0
Missing	345				
TOTAL	1.829	1.829	345	1.829	0

RESULTADOS

En la tabla 15 aparece un resumen por provincias de las diferencias entre los valores reales de la intención de voto y los valores predichos a través de los diferentes tratamientos discutidos. Puesto que las tablas se refieren únicamente a los 345 sujetos, no se pueden presentar los datos referidos al procedimiento de eliminación, en el que se han desestimado estos valores.

Como habíamos comentado en el epígrafe anterior, los diferentes tratamientos que hemos evaluado han mostrado resultados bastante aceptables en cuanto a los porcentajes totales y número de votos de cada partido (ver tablas 9, 13 y 14). Pero para el objetivo de este trabajo vamos a buscar otros indicadores acerca de su eficacia.

TABLA 15

Diferencias en frecuencias y porcentajes de los diferentes métodos de estimación por provincias

	Valor real		*Asignación proporcional*			*Análisis Discriminante*			*Fichero Caliente*		
	n	*%*	*n*	*%*	*d*	*n*	*%*	*d*	*n*	*%*	*d*
A Coruña											
BNG	22	27,8	14	17,7	8	16	20,3	6	20	25,3	2
PSOE/EU/VERDES	16	20,3	15	19	1	13	16,5	3	16	20,3	0
PP	33	41,8	34	43	–1	43	54,4	–10	34	43,0	–1
Otros			1	1,3	–1				1	1,3	–1
Abstención	8	10,1	15	19,0	–7	7	8,9	1	8	10,1	0
TOTAL	79	100	79	100	|18|	79	100	|20|	79	100	|4|
Lugo											
BNG	13	15,5	17	20,2	–4	12	14,3	1	15	17,9	–2
PSOE/EU/VERDES	12	14,3	16	19,0	–4	14	16,7	–2	13	15,5	–1
PP	51	60,7	38	45,2	13	50	59,5	1	48	57,1	3
Otros						1	1,2	–1			
Abstención	8	9,5	13	15,5	5	7	8,3	1	8	9,5	0
TOTAL	84	100	84	100	|26|	84	100	|6|	84	100	|6|
Ourense											
BNG	21	23,3	19	21,1	2	18	20	3	25	27,8	–4
PSOE/EU/VERDES	19	21,1	19	21,1	0	21	23,3	–2	15	16,7	4
PP	44	48,9	45	50,0	–1	46	51,1	–2	46	51,1	–2
Otros			1	1,1	–1	2	2,2	–2			
Abstención	6	6,7	6	6,7	0	3	3,3	3	4	4,4	2
TOTAL	90	100	90	100	|4|	90	100	|12|	90	100	|14|
Pontevedra											
BNG	26	28,3	19	20,7	7	21	22,8	5	23	25,0	3
PSOE/EU/VERDES	14	15,2	20	21,7	–6	17	18,5	–3	17	18,5	–3
PP	42	45,7	50	54,3	–8	46	50,0	–4	41	44,6	1
Otros	1	1,1			1			1			1
Abstención	9	9,8	3	3,3	6	8	8,7	1	11	12,0	–2
TOTAL	92	100	92	100	|28|	92	100	|14|	92	100	|10|
Galicia											
BNG	82	23,8	69	20,0	13	67	19,4	15	83	24,1	–1
PSOE/EU/VERDES	61	17,7	70	20,3	–9	65	18,8	–4	61	17,7	0
PP	170	49,3	167	48,4	3	185	53,6	–15	169	49,0	1
Otros	1	0,3	2	0,6	–1	3	0,9	–2	1	0,3	0
Abstención	31	9,0	37	10,7	–6	25	7,2	6	31	9,0	0
TOTAL	345	100	345	100	|30|	345	100	|41|	345	100	|2|

En primer lugar, si comparamos las diferencias en valores absolutos *(d)* entre los distintos métodos en el total de la muestra (Galicia), vemos que los mejores resultados corresponden al procedimiento Hot-Deck. Del mismo modo, cuando la unidad de análisis es la circunscripción electoral (provincia), las diferencias siguen apareciendo más bajas en el caso de los procedimientos de fichero caliente y, en segundo lugar, del análisis discriminante, aunque sin mostrar diferencias significativas entre ambos. Sólo en el caso de la provincia de Ourense aparecen mejores resultados para el análisis proporcional. Esto se debe a que el porcentaje de voto de los 90 falsos *missing* seleccionados en esta provincia se distribuye de igual modo que en Galicia (23 por 100 BNG, 21 por 100 Coalición y 48 por 100 PP), con lo que la asignación proporcional resulta eficaz. No ocurre lo mismo, como es de esperar, en el resto de las provincias.

Un segundo indicador lo buscaremos en la correlación existente entre los valores reales y los predichos a través de los diferentes métodos de estimación para esos 345 sujetos. Para ello calculamos los coeficientes de contingencia entre la variable P2 (intención de voto de los 345 sujetos) y cada una de las nuevas variables de sustitución creadas con cada procedimiento (ver tabla 16).

TABLA 16

Correlación entre los valores reales y los predichos por los distintos procedimientos

Variables	*Coef. de Contingencia*	*Significatividad*
Voto real - Asignación Proporcional	0,1960	0,6208
Voto real - Análisis Discriminante	0,7572	0,0001
Voto real - Hot-Deck	0,7010	0,0001

Como podemos ver en la tabla anterior, los coeficientes de correlación entre el voto real y los procedimientos de Análisis Discriminante y Hot-Deck muestran valores relativamente altos, en tanto que el procedimiento de asignación proporcional no es significativo.

Por último, el indicador final acerca de la efectividad de los distintos métodos lo buscaremos en el porcentaje de casos bien clasificados. Para ello comparamos las puntuaciones reales de cada sujeto (P2) y las que hemos predicho con cada uno de los procedimientos.

- Si comparamos la variable P2 con la obtenida después de realizar la asignación proporcional, vemos que este procedimiento clasifica correctamente 125 de los 345 sujetos, es decir, un 36,23 por 100.

- El análisis discriminante ha clasificado correctamente 278 de los 345 sujetos, es decir, un 80,58 por 100.
- Y el procedimiento de Hot-Deck ha clasificado 255 de los 345 sujetos, lo que supone que sólo un 73,91 por 100 de los sujetos declarados falsos *missing* han sido correctamente clasificados.

Estos resultados confirman que los mejores métodos para la estimación de la respuesta de los «no sabe/no contesta» en los estudios de intención de voto son los que se basan en las respuestas de esos sujetos a otras preguntas del cuestionario (Análisis Discriminante y Hot-Deck), mientras que tanto los procedimientos de eliminación (*listwise* y *pairwise*) como los de asignación proporcional provocan sesgos en los resultados, ya que asumen que los que no contestan se distribuyen de igual modo que los que sí lo hacen, cuestión absolutamente falsa.

DISCUSIÓN

Tal y como hemos visto, en los sondeos de opinión es frecuente encontrarse valores ausentes en alguna variable. Independientemente de su origen, este tipo de valores son codificados con la categoría «no sabe/no contesta» y, en caso de que su proporción sea elevada, pueden llegar a invalidar los resultados.

Ante esta situación, el investigador puede optar por tres alternativas: en primer lugar, dejarlos como una categoría aparte y no tenerlos en cuenta en el análisis. En segundo lugar, distribuirlos proporcionalmente entre las restantes categorías con el fin de mantener el tamaño muestral. Y, finalmente, estimar las respuestas posibles a partir de otros datos obtenidos en el cuestionario. En este trabajo pudimos comprobar que esta última alternativa es la más eficaz, sobre todo cuando sabemos que en los estudios electorales las respuestas de los que no responden no se distribuyen de igual modo que los que han contestado.

El método de imputación basado en el Análisis Discriminante se mostró especialmente eficaz cuando lo que nos interesan son las estimaciones individuales, sujeto a sujeto. Hay que decir que si bien este método puede ser de gran interés cuando estamos analizando los «NS/NC» en psicología aplicada, sin embargo, cuando se trata de estimar la intención de voto el procedimiento de Hot-Deck resulta un serio competidor, ya que se muestra más eficaz cuando se considera la circunscripción electoral. A partir de esto, y a modo de conclusión, nuestra propuesta metodológica final consiste en utilizar el Análisis Discriminante para buscar las variables que mejor explican la pertenencia de los individuos a los grupos establecidos *a priori* (votantes de los distintos partidos políticos que se presentan a unas elecciones concretas) y, a continuación, recurrir al procedimiento de Hot-Deck para realizar las estimaciones. Dado que dichos métodos son formalmente correctos, creemos que para mejorar la bondad de ajuste de las estimaciones sería conveniente investigar

qué otras variables de tipo psicosocial deben ser incluidas en el cuestionario con el fin de aumentar la capacidad predictiva de la función discriminante propuesta.

Por último, quisiéramos señalar un aspecto que, por su frecuente utilización por parte de los Institutos de Investigación, creemos importante destacar. El método de imputación a través del análisis discriminante puso de manifiesto que la práctica común de estimar el «NS/NC» a partir de una única variable explicativa: *«la posición en una escala ideológica de izquierda-derecha«*, no es correcto. Si bien es cierto que dicha variable es una de las que aparecen en la función discriminante, debemos señalar que es insuficiente si se considera aisladamente. Ello quizá esté poniendo de manifiesto que, hoy en día, el aspecto ideológico de los partidos no se restringe a una sola posición en la escala ideológica.

BIBLIOGRAFÍA

BEALE, E. M. L., y LITTLE, R. A. J. (1975): «Missing values in multivariate-analysis», *Journal of the Royal Statistical Society,* series B, 37, 129-145.

CHAN, L. S.; GILMAN, J. A., y DUNN, O. J. (1976): «Alternative approaches to missing values in discriminant analysis», *Journal of the American Statistical Association,* 71, 842-844.

CRUZ CANTERO, P. (1990): «Del no sabe al no contesta: un lugar de encuentro para diversas respuestas», *REIS,* 52, 139-156.

DÍEZ NICOLÁS, J. (1996): «Predicción de escaños electorales mediante encuestas», *REIS,* 74, 269-289.

DONNER, A. (1982): «The relative effectiviness of procedures commonly used in multiple regression analysis for dealing with missing values», *The American Statiscian,* 36, 378-381.

FERNÁNDEZ SANTANA, J. O. (1994): *Diseño y utilidad de las encuestas preelectorales,* Servicio Central de Publicaciones del Gobierno Vasco, Victoria-Gasteiz.

GREEN, P., y CARROL, I. D. (1986): *Matematical tools for Applied Multivariate Analysis,* New York: Academic Press.

KAUFMAN, C. J. (1988): «The application of logical imputation to household measurement», *Journal of the Market Research Society,* 30, 453-466.

KIM, J., y CURRY, J. (1977): «The treatment of missing data in multivariate analysis», *Sociological Methods & Research,* 6 (2), 215-240.

LITTLE, R. J. A. (1988): «Missing data adjustments in large surveys», *Journal of Business & Economic Statistics,* 6, 296-297.

LITTLE, R. J. A., y RUBIN, D. B. (1987): *Statistical Analysis with missing data,* New York: Wiley.

MARTÍN MARTÍNEZ, J. L. (1981): «Ensayo de tipificación de los sin opinión», *REIS,* 16, 9-37.

ROTH, P. H. (1994): «Missing data: a conceptual Review for Applied psychologists», *Personnel Psychology,* 47, 537-560.

ROTH, P. H.; CAMPION, J. E., y JONES, S. D. (1996): «The impact of four missing data techniques on validity estimates in human Resource Management», *Journal of Business and Psychology,* 11, 101-112.

SANZ DE LA TAJADA, L. A. (1996): «La predicción de los resultados electorales a partir de las encuestas de intención de voto: una metodología evolucionada», *AEDEMO,* 26, 21-60.

ABSTRACT

Electoral history in Spain shows us, usually in an unfavourable light, the limited validity, exactitude and meaning of some of the predictions made by electoral opinion polls. Together with problems of sociopolitical representativeness or errors in the structure of the interview, one source of error in such prediction is the scarcity of statistical processing of the absence of replies or missing data. In this paper we shall use a pre-electoral opinion poll that was carried out in the Autonomous Region of Galicia in July 1997, in order to assess the efficiency of the four most usual ways of processing the absence of replies. The procedures we shall assess are: listwise, proportional allocation, discriminatory analysis and Hot-Deck. The results demonstrate that Hot-Deck combined with discriminatory analysis is the most efficient method for estimating the probable vote of those who do not express their voting preference.

TEXTO CLÁSICO