Revista Internacional de Sociología

Revista Internacional de Sociología RIS
vol. 79 (1), e180, enero-marzo, 2021, ISSN-L:0034-9712
https://doi.org/10.3989/ris.2021.79.1.19.350

DATOS ADMINISTRATIVOS AGREGADOS Y ESTIMACIÓN A PARTIR DE MUESTRAS NO PROBABILÍSTICAS

AGGREGATE ADMINISTRATIVE DATA AND ESTIMATION FROM NONPROBABILITY SAMPLES

Pablo Cabrera-Álvarez
Universidad de Salamanca
pablocal@usal.es
ORCID iD: https://orcid.org/0000-0001-8105-5908

Cómo citar este artículo / Citation: Pablo Cabrera-Álvarez. 2021. "Datos administrativos agregados y estimación a partir de muestras no probabilísticas". *Revista Internacional de Sociología* 79(1):e180. https://doi.org/10.3989/ris.2021.79.1.19.350

Recibido: 12/03/2019. **Aceptado:** 30/06/2020.
Publicado: 06/04/2021

Resumen
En los últimos años, la investigación con encuestas ha estado marcada por el uso más frecuente de muestras no probabilísticas fruto de la expansión de internet y la caída sostenida de las tasas de respuesta. Para garantizar el proceso de inferencia cada vez son necesarios ajustes más complejos para los que se precisan variables auxiliares, es decir, información acerca de toda la población. En este trabajo se comprueba el potencial de los datos administrativos agregados a nivel de municipio para ajustar dos encuestas provenientes de un panel de internautas, el panel AIMC-Q, promovido por la Asociación Española para la Investigación de los Medios de Comunicación (AIMC). Los resultados muestran que la capacidad de las variables administrativas agregadas para reducir el sesgo de las estimaciones es mínima.

Palabras clave
Metodología de encuestas, muestras no probabilísticas, aprendizaje automático, sesgo de selección, datos administrativos.

Abstract
In the last two decades survey research has faced two main challenges: the spread of online research using non-probability samples and the general drop of response rates. In this scenario complex adjustments are needed to preserve the inference process. These adjustments require auxiliary information, this is variables available for the whole population. In this paper I test the use of aggregate administrative data at municipality level to adjust estimates from two web panel surveys promoted by the Spanish Association for Media Research (AIMC). Results show that the administrative variables are unable of tackling the bias of the survey estimates.

Keywords
Survey methodology, non-probability sample, machine learning, selection bias, administrative data.

Introducción

En los últimos años, dos fenómenos han afectado la deriva de la investigación con encuestas: el uso más frecuente de muestras no probabilísticas fruto de la expansión de internet y la caída sostenida de las tasas de respuesta. Ambos fenómenos son un riesgo para el proceso de inferencia —la posibilidad de conocer las características de la población a partir del estudio de la muestra— en el que se basa la encuesta. Sin embargo, igual que la inferencia se puede realizar a partir de la selección de una muestra probabilística, también es posible utilizar modelos estadísticos para eliminar o reducir el sesgo presente en las estimaciones una vez que los datos se han recogido.

Para corregir los sesgos presentes en las estimaciones, estos modelos precisan de variables auxiliares, es decir, información que esté disponible para el conjunto de la población. Además, para ser efectivas, estas variables tienen que estar relacionadas con las variables de interés de la encuesta y la probabilidad de participar. Este trabajo se centra en una fuente de variables auxiliares: los datos administrativos agregados.

Los datos administrativos agregados, frente a otro tipo de datos, como los microdatos administrativos o los datos comerciales, son más numerosos, accesibles y variados. De hecho, algunas fuentes de datos administrativos, como el censo, han sido utilizadas durante décadas para obtener totales poblacionales con los que ponderar las encuestas. Sin embargo, apenas se ha investigado la utilización de variables contextuales, como pueden ser las características del barrio o el municipio de la persona entrevistada, para ajustar las desviaciones de la muestra.

Esta investigación tiene como objetivo principal explorar el potencial de los datos administrativos agregados utilizados como variables contextuales para corregir el sesgo de las estimaciones realizadas a partir de muestras no probabilísticas. Para ello, se comparan tres conjuntos de variables auxiliares y tres métodos de estimación por modelo para ajustar dos encuestas web realizadas a partir de un panel de internautas promovido por la Asociación para la Investigación de los Medios de Comunicación (AIMC).

Este trabajo cuenta con cinco apartados. En el primero, se discuten los antecedentes teóricos y empíricos. En el segundo, se presentan una serie de hipótesis y posteriormente, en el tercero, se exponen los datos y la metodología empleada. En el cuarto, se presentan los resultados. Finalmente, se discuten los resultados y se presentan las conclusiones.

Marco teórico

En los últimos años, hemos asistido a un crecimiento exponencial del número de encuestas realizadas por internet en los ámbitos de la investigación social y de mercados (Blom *et al.* 2016; Hays, Liu y Kapteyn 2015). Una parte importante de esas encuestas se realizan a partir de muestras extraídas de paneles de internautas reclutados mediante métodos no probabilísticos (Callegaro, Manfreda y Vehovar 2015). El uso de estos procedimientos puede causar la aparición del sesgo de selección, que se refiere a la existencia de diferencias sistemáticas entre quienes forman parte del panel y quienes no. El sesgo de selección está compuesto por dos fenómenos diferenciados: el sesgo de cobertura y el de autoselección. El sesgo de cobertura se produce cuando una parte de los elementos de la población no tienen posibilidad de ser elegidos para participar en el estudio (Weiseberg 2005), como son los hogares sin acceso a internet en una encuesta web a la población general. Por su parte, el sesgo de autoselección se refiere a la probabilidad diferencial que tienen los elementos poblacionales de sumarse voluntariamente, por ejemplo, a un panel de internautas (Blom, Gathmann y Krieger 2015; Bethlehem y Biffignandi 2011).

La caída generalizada de las tasas de respuesta también arroja dudas sobre si el uso de muestras probabilísticas es suficiente para garantizar el proceso de inferencia (de Leeuw, Hox y Luiten 2018; Elliott y Valliant 2017). El problema de la no respuesta radica en que algunos grupos tienen una probabilidad más alta de participar en los estudios y la existencia de esa diferencia sistemática provoca que las estimaciones estén sesgadas (Groves y Couper 1998).

De la inferencia de diseño a la inferencia de modelo

Estos dos elementos, la expansión de la investigación por internet y el deterioro de la calidad de las muestras probabilísticas, hacen que la inferencia basada en la aleatoriedad de la muestra esté cuestionada (Pasek 2015). En consecuencia, se necesitan ajustes cada vez más complejos para garantizar la calidad de los datos. Brick (2011), en su trabajo sobre el futuro del muestreo en el ámbito de las encuestas, distingue entre dos tipos de inferencia, la que está basada en el diseño probabilístico de la muestra, llamada *inferencia de diseño*, y la que se asienta en modelos estadísticos ajustados tras la recogida de los datos, denominada *inferencia a partir de modelos*.

La *inferencia de diseño* se basa en el mecanismo probabilístico que subyace a la selección de una muestra aleatoria (Kish 1965; Neyman 1934). Desde que se desarrolló el grueso de la teoría del muestreo a mediados del siglo XX, la mayoría de las encuestas han confiado en los principios de la probabilidad para seleccionar muestras representativas de la población (Baker *et al.* 2013). Una muestra es probabilística en la medida que todos los miembros de la población tienen una probabilidad conocida de ser seleccionados que es distinta de cero (Levy y Lemeshow 2013).

Si, además, todos los elementos muestrales responden a la encuesta, las estimaciones realizadas a partir de la muestra podrán ser inferidas a la población con un cierto grado de precisión.

Sin embargo, cada vez en más ocasiones el proceso de inferencia no puede ser garantizado a partir del diseño de una muestra probabilística, ya sea porque la muestra ha sido seleccionada empleando técnicas no probabilísticas o debido a la presencia de sesgos producidos por la autoselección o la no respuesta. Un ejemplo recurrente son las encuestas realizadas a partir de paneles de internautas reclutados empleando métodos no probabilísticos. En estos casos, se puede optar por apoyar el proceso de inferencia en modelos —*inferencia a partir de modelos*— en que el aparato estadístico se encarga de controlar los sesgos (Valliant, Dorfman y Royall 2000). Dentro de la inferencia a partir de modelos, se diferencian varios mecanismos: los modelos de cuasi aleatorización, los modelos de superpoblación y una combinación de ambos: el doble ajuste (Elliott y Valliant 2017; Valliant 2019).

El método de cuasi aleatorización consiste en hallar mediante un modelo estadístico las pseudo probabilidades de selección de los elementos de la muestra no probabilística usando los datos de una encuesta probabilística como referencia (Gummer y Roßmann 2018; Pasek 2016; Valliant y Dever 2011; de Pedraza *et al.* 2010; Lee y Valliant 2009). En otros casos, las pseudo probabilidades de selección se han calculado a partir de emparejar los casos en la encuesta no probabilística con los de una muestra probabilística utilizando técnicas de *propensity score matching* (Mercer, Lau y Kennedy 2018; Ferri-García y Rueda 2018; Elliott y Valliant 2017). También se han utilizado métodos de calibración o postestratificación para calcular las pseudo probabilidades de inclusión utilizando información auxiliar agregada como totales poblacionales (Peytchev, Presser y Zhang 2018; Pasek 2016; Dever, Rafferty y Valliant 2008).

El método de superpoblación consiste en ajustar un modelo para predecir la variable de interés en la muestra no probabilística y proyectarlo al conjunto de la población (Buelens, Burger y Brakel 2018; Wang *et al.* 2015; Dorfman y Valliant 2005). Para que este método sea efectivo, los datos en la muestra y la población deben seguir un modelo común que puede ser descubierto a partir de la encuesta. Los ajustes a partir de modelos de superpoblación son menos flexibles que el uso de la cuasi aleatorización, ya que, en teoría, es necesario generar un peso para cada variable de interés. En los últimos años, se han empleados técnicas de aprendizaje automático para elaborar modelos de superpoblación (Chen, Valliant y Elliot 2018).

También existe la posibilidad de combinar las dos estrategias anteriores y realizar un doble ajuste. Se trata de calcular las pseudo probabilidades de selección que, a su vez, son utilizadas para ajustar el modelo de superpoblación (Kang y Schafer 2007). El sesgo de las estimaciones se reducirá en la medida en que uno o ambos modelos estén correctamente especificados. En los últimos años, varias investigaciones han comparado algunas de estas estrategias de ajuste. Ferri-García y Rueda (2018) hallaron que la combinación de los métodos de *propensity score* y calibración generaba ajustes más eficaces. Por su parte, Valliant (2019), a partir de simulaciones, comparó la eficacia de varias estrategias de estimación en encuestas no probabilísticas, como la cuasi aleatorización, los modelos de superpoblación o la regresión multinivel con postestratificación, encontrando que una combinación de la cuasi aleatorización con los modelos de superpoblación era la mejor opción para reducir el nivel de sesgo de las estimaciones.

En cualquier caso, todas las estrategias de ajuste tienen algo en común, precisan de variables auxiliares que estén correlacionadas tanto con las variables de interés como con la probabilidad de participar en la encuesta (West y Little 2013). De hecho, un estudio reciente utilizando encuestas de un panel de internautas en Estados Unidos muestra que la especificación de los modelos es más relevante que la técnica utilizada para ajustarlos (Mercer, Lau y Kennedy 2018).

Variables auxiliares para corregir los sesgos de autoselección y no respuesta

Tradicionalmente, la información necesaria para construir los ajustes de las encuestas provenía de estadísticas y encuestas oficiales, como el censo de población. Sin embargo, en los últimos años han aparecido múltiples fuentes de datos que potencialmente pueden ser utilizadas para corregir sesgos de encuestas: datos comerciales (West *et al.* 2015; Peytchev y Raghunathan 2013), paradatos (Kreuter 2013), datos georreferenciados (Lahtinen, Kaisa y Butt 2015) y administrativos (Couper 2013). Este trabajo se centra en una de esas fuentes: los datos administrativos agregados.

Los datos administrativos son productos o subproductos generados en la interacción de la Administración Pública con los ciudadanos, empresas u otras organizaciones (Playford *et al.* 2016). Woollard (2014) establece que los datos administrativos son recogidos para organizar, gestionar o monitorizar servicios, pero también pueden ser útiles para responder preguntas de investigación en el ámbito de las ciencias sociales. Estos datos tienen una serie de características que los hacen buenos candidatos para ser variables auxiliares.

En primer lugar, tienden a estar sujetos a menos error que los datos de encuesta, aunque la definición

de los conceptos y los instrumentos utilizados para recoger la información puedan diferir (Connelly *et al.* 2016). En segundo lugar, los datos administrativos tienen una amplia cobertura que, en muchos casos, alcanza a la totalidad de la población (Künn 2015). Como contrapunto, el acceso a los datos depende de la voluntad de la administración y puede estar restringido para garantizar la privacidad de los ciudadanos o de las organizaciones (Dibben *et al.* 2015; Stevens y Laurie 2014). Pero resulta evidente que esta desventaja afecta en menor medida a los datos administrativos agregados.

Los datos administrativos agregados han tenido un papel relevante en la corrección de los sesgos de selección y de no respuesta desde hace décadas, ya que, entre otros usos, los datos del censo suelen emplearse para ajustar la distribución de la muestra con respecto al sexo y la edad (p. ej., Morris *et al.* 2016; Park *et al.* 2013). Para poder realizar estos ajustes, el investigador tiene que saber de antemano las variables que va a utilizar para calibrar la muestra final, con el fin de incluirlas en el cuestionario.

Recientemente, ante una mayor variedad de datos administrativos disponibles, ha habido un interés renovado en combinarlos con datos de encuestas (Lohr y Raughnathan 2017; Smith y Kim 2013; Smith 2011). Una posibilidad es utilizar los datos administrativos agregados como variables contextuales, información resumida acerca del entorno de los elementos incluidos en la muestra, como puede ser el barrio o municipio. Se trata de utilizar, por ejemplo, el porcentaje de coches de lujo, la prevalencia de voto a un partido determinado o el valor de las edificaciones en el área donde reside la unidad muestral. Esta información contextual, además de ser muy variada en su temática y de fácil acceso, podría ser efectiva a la hora de ajustar los modelos que corrigen el sesgo de las estimaciones de la encuesta.

No obstante, existen pocos trabajos en los que se hayan utilizado los datos administrativos agregados como variables contextuales para tratar el sesgo de selección o no respuesta. Biemer y Peytchev (2012; 2013) utilizaron datos administrativos agregados a nivel de sección censal, municipal y de condado para detectar y corregir el efecto de la no respuesta en una encuesta telefónica en los Estado Unidos —*the National Comorbidity Survey Replication*—, concluyendo que el uso de datos administrativos no era efectivo para mejorar las estimaciones. Más recientemente, en el Reino Unido, se probó la eficacia de los datos administrativos agregados a nivel de sección censal o municipio como variables contextuales para ajustar el sesgo de no respuesta presente en la muestra de la Encuesta Social Europea en ese país (Lahtinen, Kaisa y Butt 2015). Los resultados de la investigación concluyeron que los datos agregados no estaban relacionados con la probabilidad de respuesta en esta encuesta. A pesar de los pobres resultados de estas investigaciones, cabe destacar que, en ambos casos, se utilizaron encuestas probabilísticas con cuidados procedimientos de recogida de datos en las que la presencia de sesgos suele estar atenuada.

Hipótesis

A partir de las teorías e investigaciones mencionadas en el apartado anterior, se presentan las hipótesis en relación con los dos estudios que se emplean en este artículo.

H1. *El uso como variables auxiliares derivadas a partir de la información administrativa agregada a nivel de municipio, en comparación con el uso de variables sociodemográficas, da como resultado una mayor reducción del nivel del sesgo que presentan las estimaciones*.

La ventaja de usar datos administrativos agregados estriba en su disponibilidad y variedad. En este trabajo, como se detalla en la siguiente sección, se comparan tres conjuntos de variables auxiliares para generar las ponderaciones: sociodemográficas, administrativas y la combinación de ambas. Las variables sociodemográficas son las que habitualmente se utilizan para ponderar esta encuesta (sexo, edad, comunidad autónoma de residencia y tamaño del municipio). Las variables administrativas son más numerosas y cubren un amplio abanico de temas, desde los ingresos al comportamiento electoral. Esa mayor variedad induce a pensar que algunas de las variables serán efectivas a la hora de reducir el nivel de sesgo de las estimaciones.

H2. *La combinación de las variables auxiliares administrativas y las sociodemográficas son la alternativa más efectiva para reducir el sesgo de las estimaciones*.

H3. *La efectividad de las variables auxiliares a la hora de reducir el sesgo en las estimaciones es independiente de la estrategia que se utilice para generar la ponderación.*

Para analizar la efectividad de las variables administrativas agregadas se han generado una serie de ponderaciones utilizando tres estrategias diferentes: la cuasi aleatorización, la estimación a partir de los modelos de superpoblación y el método de doble ajuste. A pesar de las diferencias entre las estrategias de estimación, se espera observar pautas similares: los datos administrativos, al ser más variados, dan lugar a ponderaciones más efectivas.

H4. *Los errores típicos de las estimaciones serán menores cuando se utilicen las variables administrativas en el cálculo de las ponderaciones*.

Se han utilizado dos procedimientos para calcular los errores de las estimaciones. Por un lado, un méto-

do linealizado que se utiliza en los muestreos con reemplazo. Por el otro, un método de replicación de tipo *jackknife* en el que el error se calcula excluyendo parte de la muestra en el cálculo de las estimaciones en cada réplica. Las ponderaciones tienen la capacidad de reducir el error de las estimaciones si las variables utilizadas están correlacionadas con la probabilidad de responder y la variable de interés. En este caso, el uso de las variables administrativas para calcular los pesos hará que las estimaciones sean más eficientes.

Datos y metodología

En este apartado se presentan los datos y la metodología del análisis que se ha llevado a cabo utilizando dos encuestas del panel AIMC-Q, el Estudio General de Medios (EGM) y datos administrativos agregados.

Fuentes de datos

Para realizar este análisis, se han utilizado tres tipos de fuentes: datos administrativos agregados, que se utilizan como variables contextuales; datos del Estudio General de Medios, que se emplean como referente poblacional para calcular el sesgo de las estimaciones, y dos encuestas realizadas en el marco del panel de internautas AIMC-Q.

La recogida de datos administrativos se realizó a partir del directorio nacional de operaciones estadísticas del Instituto Nacional de Estadística (INE), que agrupa, por nivel de agregación, todos los datos recogidos y producidos por el gobierno central. En este trabajo se utilizan los datos agregados a nivel municipal por dos motivos: el primero es que las fuentes de datos disponibles a un nivel inferior, como el censo, son escasas. El segundo, y más importante, es que las encuestas empleadas en este trabajo solo contenían la identificación del municipio del entrevistado, por lo que era imposible utilizar información agregada a un nivel inferior. Los datos agrupados a nivel de municipio fueron incluidos en una base de datos con 1099 variables, entre las cuales figura el censo de 2011, el padrón de habitantes, estadísticas del impuesto sobre la renta, datos electorales, datos de desempleo o información sobre la marca de los vehículos matriculados en el municipio, entre otras, como se muestra en la tabla 1.

Tabla 1
Variables administrativas incluidas en la investigación.

Fuente de datos	Institución	Año (periodo)	Variables	Número de variables
Nomenclator	Instituto Nacional de Estadística	2018 (semestral)	Municipios, población total, población por sexo.	2
Censo de población y viviendas	Instituto Nacional de Estadística	2011 (cada diez años)	Sexo, edad, estado civil, nivel educativo, país de nacimiento, nacionalidad. Viviendas según tamaño, tipo de propiedad, número de habitaciones y personas residiendo.	145
Padrón	Instituto Nacional de Estadística	2018 (semestral)	Sexo, edad, nacionalidad, país de nacimiento, relación lugar de nacimiento y residencia.	303
Catastro	Oficina del Catastro	2016 (anual)	Superficie según uso, valor medio del suelo, tipología del suelo.	20
Impuestos municipales	Ministerio de Hacienda	2016 (anual)	Datos IBI, IAE, IVTM, IVTUN, ICIO.	21
IRPF	Ministerio de Hacienda	2016 (anual)	Base imponible, declarantes, titulares, deducciones, renta bruta media y disponible media.	32
Liquidación del presupuesto municipal	Ministerio de Hacienda	2016 (anual)	Derechos liquidados y obligaciones reconocidas.	32
Paro y contratos registrados	Ministerio de Trabajo y Seguridad Social	2018 (mensual)	Parados y contratos celebrados.	23
Censo de conductores	Dirección General de Tráfico	2017 (anual)	Conductores, sexo.	2
Parque de vehículos	Dirección General de Tráfico	2017 (anual)	Turismos, ciclomotores, motocicletas, marca de turismos.	58
Matriculaciones	Dirección General de Tráfico	2017 (anual)	Matriculaciones.	1
Accidentes	Dirección General de Tráfico	2017 (anual)	Víctimas de accidentes.	1
Elecciones	Ministerio del Interior	2016 (anual)	Resultados de las elecciones municipales, europeas y generales (1977-2016).	459

El EGM es un estudio que se realiza en tres oleadas cada año, cuya población son los residentes en España mayores de 14 años. El objetivo principal del estudio es recabar datos sobre el consumo de medios de comunicación en España, para lo que se realiza una muestra multimedia y otras tres de un solo medio (radio, prensa o televisión). La muestra multimedia cuenta con 30 000 entrevistas, mientras que las especializadas oscilan entre las 13 000 (televisión) y las 45 000 (prensa). En el estudio multimedia, los datos se recogen entrevistando a los informantes en los domicilios seleccionados mediante muestreo probabilístico. En el caso de las encuestas de un solo medio, los datos se recogen mediante entrevista telefónica combinando fijos y móviles. Los datos provenientes de las diferentes fases son ajustados para preservar la representatividad de la muestra. En este trabajo, se utilizan las estimaciones poblacionales de la primera (enero-marzo) y segunda (abril-junio) oleadas de 2017 como puntos de referencia para calcular el sesgo de las estimaciones provenientes de las encuestas del panel AIMC-Q que se presentan a continuación. Esta estrategia presenta el inconveniente de que, a pesar del elevado número de entrevistas y de que los elementos muestrales fueron seleccionados mediante técnicas probabilísticas, las estimaciones también pueden presentar desviaciones. Sin embargo, ante la falta de referentes poblacionales, es común asumir que las estimaciones de una encuesta como el EGM presentarán un menor nivel de sesgo que las provenientes de las encuestas del panel de internautas (Yeager *et al.* 2011; Schonlau *et al.* 2009).

Por otro lado, se utilizaron dos encuestas provenientes de un panel probabilístico de internautas gestionado por la AIMC. Este panel de internautas experimental, que lleva en marcha desde 2013, está compuesto por entrevistados del EGM con acceso a internet en el hogar que accedieron a participar. En 2017, el panel contaba con 4514 miembros que eran invitados a completar encuestas periódicamente. Este trabajo se centra en dos encuestas: la primera es sobre consumo de prensa (n = 2.013), cuyo trabajo de campo tuvo lugar en el mes de marzo de 2017 y la segunda es sobre consumo de radio (n = 2.058), cuyos datos fueron recogidos en junio de 2017. Es preciso señalar que este panel está compuesto por una submuestra reclutada de forma probabilística, pensada para estudiar la población de internautas residentes en España. Sin embargo, en esta investigación se asume que, utilizando ajustes basados en modelos, el mismo panel puede ser utilizado para estudiar la población general española, como ocurre con otros paneles de internautas cuyos miembros son reclutados con métodos no probabilísticos.

La tabla 2 presenta los perfiles de las muestras de ambos estudios junto con la distribución de las mismas variables en la población. En las dos encuestas destaca la subrepresentación de los mayores de 64 años, mientras que el grupo de individuos entre 35 y 54 años está sobredimensionado. También está sobredimensionado el grupo de personas residentes en una capital de provincia —41 % de la muestra en ambos estudios, pero solo el 32 % de la población—.

Tabla 2

Perfil de la muestra de los estudios de prensa y radio del panel AIMC-Q y distribución poblacional1

		Población	Prensa	Radio
Sexo	Hombre	48,6	55,0	55,9
	Mujer	51,4	45,0	44,1
Edad	14-19	6,7	5,3	5,3
	20-24	5,7	5,8	6,7
	25-34	13,9	13,5	13,9
	35-44	19,3	27,4	26,0
	45-54	18,1	27,3	26,4
	55-64	14,4	14,2	15,7
	65 o mas	21,9	6,7	5,9
Tamaño municipio	Menos de 2000	6,1	4,6	3,7
	De 2001 a 5000	6,6	5,0	4,8
	De 5001 a 10 000	8,3	6,2	6,4
	De 10 001 a 50 000	26,5	22,4	23,3
	De 50 001 a 200 000	15,3	15,1	14,5
	De 200 001 a 500 000	5,0	5,4	5,9
	Capital de provincia	32,2	41,3	41,5

Metodología

Con el fin de comprobar la efectividad de los datos administrativos agregados para ajustar las estimaciones de las encuestas se han generado nueve ponderaciones. Estos pesos son el producto de utilizar tres conjuntos de variables auxiliares y tres métodos de estimación. Los tres conjuntos de variables auxiliares son una selección de variables sociodemográficas (SD), los datos administrativos agregados (AD) y una combinación de ambos (SD+AD). Cada conjunto ha servido para estimar un modelo de cuasi aleatorización (CA), un modelo de superpoblación (SP) y otro de doble ajuste (DA). En el caso de los modelos de superpoblación y doble ajuste, se ha generado un peso para cada una de las 13 variables de interés de la encuesta.

Variables auxiliares

El primer conjunto de datos auxiliares se corresponde con las variables sociodemográficas (SD) que

generalmente son utilizadas para ajustar las encuestas del panel. Se trata de un conjunto básico que incluye grupos de sexo y edad, tamaño de hábitat en siete categorías y la comunidad autónoma de residencia. Estas variables se utilizan de forma habitual porque los totales poblacionales están a disposición de los investigadores y la información suele estar completa para todos los elementos de la muestra. Sin embargo, la eficacia de este conjunto de variables para reducir el sesgo de las estimaciones no está garantizada, ya que puede haber una relación débil entre las variables sociodemográficas, la probabilidad de participar en el estudio y las variables de interés. Se trata de un escenario básico con el que comparar la eficacia de las variables administrativas agregadas.

El segundo conjunto de datos auxiliares son las 1099 variables administrativas agregadas (AD) que fueron recabadas a partir del directorio nacional de operaciones del INE. Se trata de un amplio conjunto de variables contextuales de fácil acceso que, en caso de demostrarse útiles para ajustar las estimaciones de la encuesta, podrían contribuir a mejorar las estimaciones de otros estudios. Antes de ser utilizadas en los modelos, estas variables fueron tratadas en tres pasos: 1) en algunos casos la información poblacional no estaba disponible para todos los municipios, por lo que los valores perdidos fueron imputados utilizando el método del vecino más próximo; 2) las variables, que por lo general eran totales, fueron convertidas en porcentajes, y 3) para el correcto funcionamiento de los modelos de regresión regularizada, fueron escaladas y estandarizadas. Por último, también se ha incluido en el diseño una combinación de ambos conjuntos, de las variables administrativas y las sociodemográficas (SD+AD).

Técnicas de estimación

Las tres técnicas de estimación utilizadas —cálculo de las pseudo probabilidades de selección, modelos de superpoblación y modelos de doble ajuste— suelen basarse en modelos lineales como la regresión logística, dado que el número de variables auxiliares a disposición de los investigadores suele ser reducido. Sin embargo, en esta ocasión se plantea el uso de más de mil variables auxiliares a la vez, lo que ha derivado en la sustitución de los modelos lineales generalizados por una técnica de aprendizaje automático: la regresión regularizada.

Modelos de regresión regularizada

En los últimos años, se han empleado diferentes técnicas de aprendizaje automático para ajustar las estimaciones de las encuestas: *random forest* (Valliant, Dever y Kreuter 2018), máquinas de soporte vectorial y redes neuronales (Buelens, Burger y Brakel 2018; Buskirk *et al.* 2018) o regresión regularizada (Chen, Valliant y Elliott 2018). La decisión de utilizar regresiones regularizadas en este trabajo responde a que, siendo un modelo de tipo lineal, se ha demostrado efectiva para la selección automática de predictores en presencia de multicolinealidad.

Cuando existe un elevado número de predictores, los modelos lineales generalizados pueden presentar problemas; es probable que algún supuesto como el de ausencia de multicolinealidad sea quebrantado. La regresión regularizada se basa en la idea de que una selección de las variables independientes contiene los efectos más relevantes del modelo (Hastie, Tibshirani y Wainwright 2015). La identificación de esas variables se lleva a cabo mediante la inclusión de un término de penalización en la función objetivo que limita la magnitud de los coeficientes, de forma que estos solo pueden aumentar si se experimenta un descenso comparable en la función objetivo.

Las penalizaciones más extendidas son las *ridge*, *lasso* y *elastic net*, conteniendo la última una combinación de las otras dos. Aquí se describe la penalización *elastic net* aplicada a la regresión logística, que es el modelo utilizado en este trabajo (Friedman, Hastie y Tibshirani 2010). Para modelar la variable y, que toma valores 0 y 1, a partir de un vector de predictores $\boldsymbol{x_i}$:

$$ln\left[\frac{P(y_i = 1|\boldsymbol{x_i})}{P(y_i = 0|\boldsymbol{x_i})}\right] = \beta_0 + \beta_0^T \boldsymbol{x_i}$$

En el modelo con penalización *elastic net*, la función objetivo utilizada para ajustarlo incluye la penalización $\boldsymbol{\lambda}$, que puede variar entre 0 y $+\infty$, y un parámetro $\boldsymbol{\alpha}$ que varía entre 0 y 1 y determina en qué medida se aplican las penalizaciones *ridge* $\frac{\|\beta\|_2^2}{2}$, *lasso* $\|\beta\|_1$ o una combinación de ambas:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i\,(\beta_0 + x_i^T\beta) - \ln\left(1 + e^{(\beta_0 + x_i^T\beta)}\right)\right] + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$$

Los modelos con penalización *elastic net* se han calculado utilizando el paquete *glmnet* de R.

Modelo de cuasi aleatorización (CA)

En el método de cuasi aleatorización, para calcular las pseudo probabilidades de selección se ha utilizado una muestra seleccionada a partir de los datos poblacionales que ha actuado como la encuesta de referencia (Valliant, Dever y Kreuter 2018). La encuesta de referencia (n = 10 000) fue combinada con las muestras de las encuestas de prensa y radio del panel. El conjunto de datos combinado contaba con una variable que indicaba si cada caso procedía de la encuesta del panel, en cuyo caso tomaba el valor de 1, o si provenía de la encuesta de referencia, en que tomaba el valor de 0.

La variable que indicaba la procedencia de la muestra se modeló utilizando una regresión logística

regularizada. Para cada encuesta se realizaron tres modelos, uno con cada conjunto de variables auxiliares (SD, AD, SD+AD). Para determinar los valores de α y λ, necesarios para ajustar el modelo, se procedió a calcular diez validaciones cruzadas para seis valores diferentes de α. Por defecto, *glmnet* calcula el modelo para un conjunto de 100 valores de λ. Los valores de α y λ que dieron como resultado un menor error de clasificación fueron utilizados para computar el modelo final con el que se predijo la probabilidad de formar parte de la muestra del panel. La ponderación final fue calculada como el inverso de esa probabilidad:

$$w_i^{CA} = \frac{1}{\hat{\pi}(\boldsymbol{x}_i)}$$

en la que $\hat{\pi}(\boldsymbol{x}_i)$ representa la probabilidad estimada de formar parte de la encuesta del panel a partir de un vector de variables auxiliares $\boldsymbol{x}$.

Modelo de superpoblación (SP)

Para el cálculo de las ponderaciones con modelos de superpoblabión w^{SP} se ha adaptado el método de calibración a partir de un modelo *lasso* adaptativo propuesto por Chen, Valliant y Elliott (2018). El método utilizado pasa por: 1) ajustar un modelo de regresión regularizada con penalización *elastic net* para predecir la variable de interésen la muestra; 2) proyectar dicho modelo en la población para predecir los valores de la variable de interés, y 3) generar las ponderaciones a partir de un modelo de calibración utilizando como referente el total de la variable predicha en la población.

En la calibración asistida por modelo (Särndal y Lundström 2005; Wu y Sitter 2001), las distancias entre los pesos de diseño d_i y las ponderaciones finales w_i se obtienen minimizando la función g en la que q_i es una constante independiente del peso de diseño:

$$E\left[\sum_{i\in S} g(w_i^{SP}, d_i)/q_i\right]$$

cumpliéndose las condiciones de que $\sum_{i\in s} w_i^{SP} = N$ y $\sum_{i\in s} w_i^{SP}\hat{y}_i = \sum_i^N \hat{y}_i$. Asumiendo q_i=1 y que g corresponde a la distancia chi-cuadrado $g(w_i^{SP}, d_i) = (w_i^{SP} - d_i)^2/d_i$:

$$\mathbf{w}^{\mathrm{SP}} = \mathbf{d} + \mathbf{DM}(\mathbf{M}^{\mathrm{T}}\mathbf{DM})^{-1}(\mathbf{T}^{\mathrm{M}} - (\mathbf{d})^{\mathrm{T}}\mathbf{M})^{\mathrm{T}}$$

donde d son los pesos de diseño de la muestra, D corresponde a la matriz en cuya diagonal se encuentran los pesos de diseño, $\mathbf{M} = [\mathbf{d}, \sum_{i\in A}\hat{y}_i]$ y $\mathbf{T}^{\mathrm{M}} = (N, \sum_i^N \hat{y}_i)$.

En este caso, para cada variable dependiente y conjunto de variables auxiliares —SD, AD y SD+AD— se generó una ponderación. Para elaborar los modelos de regresión logística regularizada se siguió el procedimiento descrito en la sección anterior.

Modelo de doble ajuste (DA)

Esta tercera estrategia de estimación consiste en combinar las dos anteriores, el modelo de cuasi aleatorización y el de superpoblación. Para aplicar esta estrategia se han seguido dos pasos: en primer lugar, se ha utilizado como base el peso procedente del modelo de cuasi aleatorización para cada conjunto de variables auxiliares. Posteriormente, ese peso se ha utilizado para ponderar el modelo de superpoblación y derivar los pesos finales siguiendo la metodología desarrollada en la sección anterior.

Error típico de las estimaciones

Para calcular el error típico de las estimaciones se han utilizados dos procedimientos: uno de tipo linealizado, diseñado para el muestreo aleatorio con reemplazo, y otro de replicación de tipo *jackknife* abreviado. El método linealizado para el muestreo con reemplazo ha sido propuesto como una alternativa para aproximar el error de las estimaciones cuando estas se realizan a partir de modelos de superpoblación o del cálculo de las pseudo probabilidades de selección (Valliant, Dever y Kreuter 2018). La ventaja de utilizar este estimador es que, aparte de estar implementado en la mayoría de los programas estadísticos, no requiere excesivos recursos de computación. La estimación del error típico de una media es:

$$se_r(\hat{\bar{y}}) = \sqrt{\hat{N}^{-2}\,\frac{n}{(n-1)}\sum_{i\in s}(\hat{z}_i - \hat{\bar{z}})^2}$$

en el que $\hat{z}_i = w_i z_i$ y z_i es una medida de desviación asociada con $\hat{y}$ y $\hat{N}$ equivale a $\sum_{i\in s} w_i$ (Valliant 2019). El principal inconveniente de este método es que no tiene en cuenta que las ponderaciones han sido elaboradas a partir de estimaciones. Por ejemplo, en el caso de la cuasi aleatorización, las pseudo probabilidades de selección son estimaciones derivadas de una muestra que combina una encuesta de referencia y la encuesta no probabilística.

El procedimiento *jackknife* consiste en replicar la estimación veces, excluyendo un caso cada vez, para, a partir de las múltiples estimaciones, hacer un cálculo de la desviación de la media del estimador. Existen investigaciones en las que se ha utilizado este método de replicación para calcular la varianza de las estimaciones realizadas a partir de muestras no probabilísticas (Valliant 2019). El principal inconveniente de este método es que es intensivo en el uso de computación —hay que ajustar cada modelo de regresión regularizada *n* veces para estimar las variables de interés—, por lo que Valliant (2019) propone utilizar una versión abreviada en la que los ca-

sos se agrupan aleatoriamente en conjuntos y uno es excluido cada vez. El error típico se estimaría mediante la fórmula:

$$se_j(\hat{\bar{y}}) = \sqrt{\frac{J-1}{J}\sum_{j=1}^{J}(\hat{\bar{y}}_{(j)} - \hat{\bar{y}})^2}$$

en la que $\hat{\bar{y}}_{(j)}$ es la estimación de la media de la variable de interés excluyendo las unidades del grupo j. El número de grupos J fue establecido en 50, lo que implica que todos los ajustes fueron calculados 50 veces con el fin de calcular el se_j de cada estimación.

Evaluación del impacto de las ponderaciones

Para evaluar la eficacia de las ponderaciones a la hora de reducir el sesgo de las estimaciones se utilizaron 13 variables factuales, presentes tanto en el EGM como en las encuestas del panel AIMC-Q. Las estimaciones se evaluaron calculando una medida ponderada del sesgo relativo, que compara la estimación del EGM con la estimación de la encuesta:

$$\bar{B}_{wr} = \frac{\sum B_r \hat{\bar{y}}_{EGM}}{\sum \hat{\bar{y}}_{EGM}}$$

en la que $\hat{\bar{y}}_{EGM}$ es la estimación de la media o proporción de la variable en el EGM y B_r es una medida del sesgo relativo de cada estimación:

$$B_r = \left| \frac{\hat{\bar{y}}_s - \hat{\bar{y}}_{EGM}}{\hat{\bar{y}}_{EGM}} \right| 100$$

en la que $\hat{y}_s$ es la estimación de la media de la variable de interés en la muestra.

Resultados

Las dos encuestas del panel AIMC-Q —radio y prensa— fueron utilizadas para comprobar el potencial de los datos administrativos agregados a la hora de ajustar los sesgos. La tabla resumen de los estadísticos descriptivos de las ponderaciones se puede consultar en el anexo I (tabla 3). La figura 1 presenta, para cada variable, la estimación poblacional del EGM, la estimación de la encuesta del panel AIMC-Q sin ponderar y nueve estimaciones ponderadas. Las nueve ponderaciones corresponden a los tres conjuntos de variables auxiliares —SD, AD y SD+AD— utilizados por cada método de estimación —pesos calculados con el método de cuasi aleatorización (CA), los modelos de superpoblación (SP) y el doble ajuste (DA)—. Además, el primer gráfico presenta un promedio del nivel de sesgo relativo que presentan las estimaciones para cada combinación de método y conjunto de variables auxiliares.

Las variables auxiliares administrativas (AD) han demostrado una capacidad mínima para reducir el nivel de sesgo de las estimaciones. En promedio, la mejora del sesgo relativo de las estimaciones (gráfico 1 de la figura 1) apenas alcanza el punto porcentual si se toman como referencia las estimaciones sin ponderar. En el mejor de los casos, cuando los datos administrativos se emplean con el método de superpoblación (SP), la reducción del sesgo relativo es de 1 punto porcentual. Prueba de ellos es que, en la mayoría de las estimaciones (gráficos 2 a 14) ajustadas con datos administrativos, apenas hay diferencias con respecto a la ausencia de ponderación. Solo en cinco de las variables se observa cierto efecto de los ajustes, aunque en tres de ellas se produce en el sentido contrario al esperado, resultando un ligero aumento del sesgo. Estos son los casos de *prensa deportiva papel ayer*, *radio TDT ayer* y *radio en el trabajo ayer*. Por ejemplo, las tres estimaciones de la proporción del consumo de *radio en el trabajo ayer* arrojan aumentos en el nivel de sesgo que alcanzan los 0,9 puntos porcentuales (modelos SP y DA). También en los casos en los que la ponderación con datos administrativos tiene un efecto reductor sobre el sesgo, la magnitud de este es mínima. El caso más destacado es el de la variable que mide el consumo de *radio en la casa ayer*, en el que el nivel de sesgo se reduce en 2,8 puntos porcentuales (SP) desde la estimación sin ponderar, aunque sigue existiendo una diferencia de 7,1 puntos porcentuales con respecto a la estimación del EGM.

Por su parte, las variables sociodemográficas (SD) se muestran ligeramente más efectivas que las administrativas cuando se emplea el método de cuasi aleatorización. Sin embargo, lo contrario ocurre si nos referimos a los modelos de doble ajuste y superpoblación. En ambos casos, además, el promedio del sesgo relativo de las estimaciones aumenta en un punto porcentual con respecto a las estimaciones sin ponderar. El uso de las variables sociodemográficas en la ponderación, frente a las administrativas, produce mayores variaciones en las estimaciones, lo que causa un incremento del sesgo en cinco de las variables, como son los casos de la lectura de algún *diario en papel ayer* o de *suplementos en los últimos 7 días*. Los efectos positivos se observan sobre todo en las variables del estudio sobre el consumo de radio. Por ejemplo, la variable de escuchar la *radio en el coche ayer*, ponderada por el peso de superpoblación, reduce el sesgo de la estimación en 4,6 puntos porcentuales; y el sesgo de escuchar la *radio en el trabajo ayer* se reduce en 2,3 puntos usando la ponderación del doble ajuste.La combinación de las variables sociodemográficas y administrativas (SD+AD) es, en promedio, el conjunto de datos auxiliares más efectivo

Figura 1

Sesgo de las estimaciones según el método de ajuste y el conjunto de variables auxiliares. El gráfico 1 presenta el promedio del sesgo relativo y los gráficos 2-14 representan las estimaciones en porcentaje sin ponderar y ponderadas para cada variable, junto con el valor de la estimación del EGM.

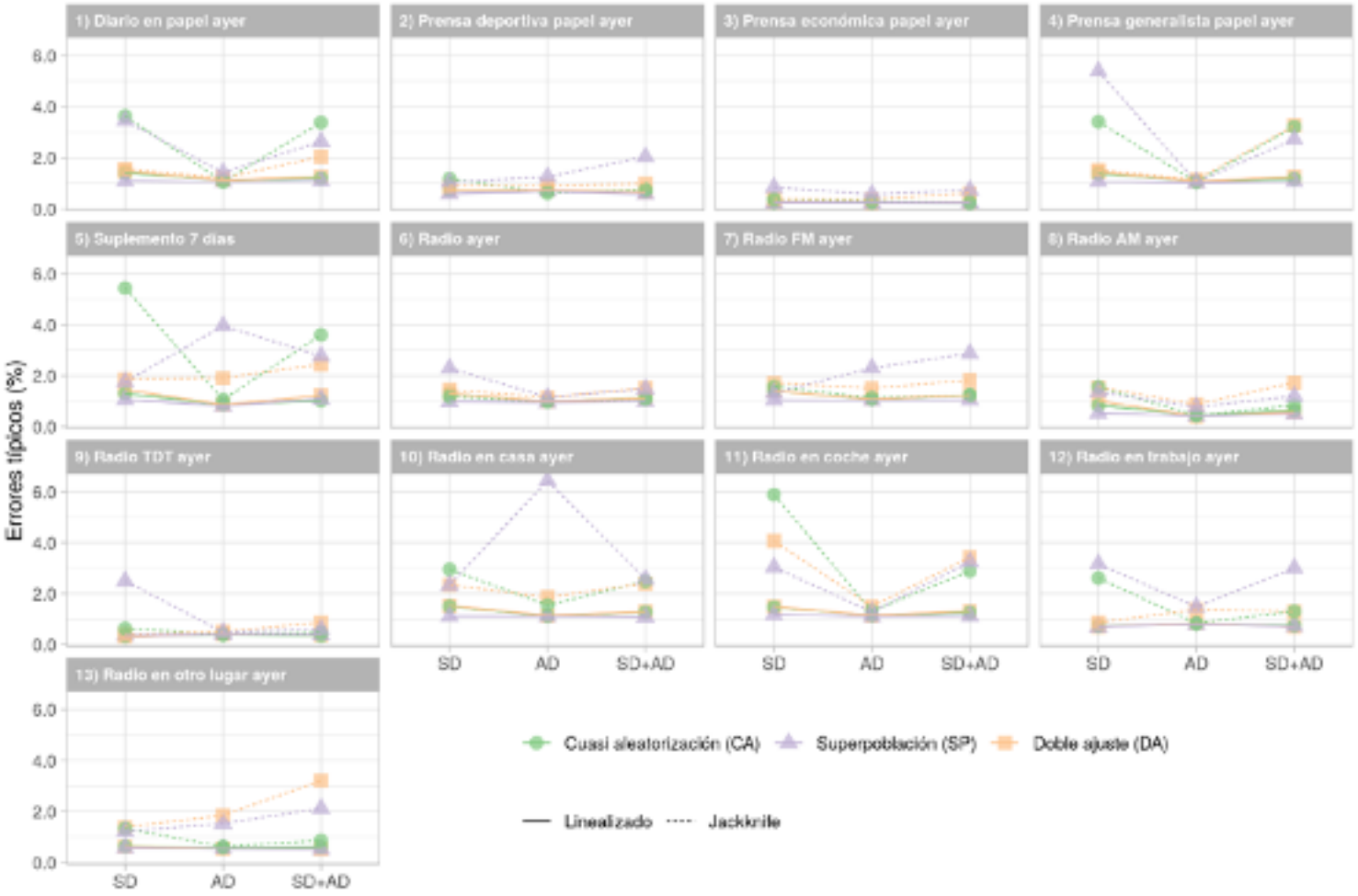
a la hora de reducir el sesgo de las estimaciones. Además, este efecto se acrecienta cuando se aplica el método de cuasi aleatorización, aunque solo supone una mejora de 2,2 puntos porcentuales sobre la estimación sin ponderar. Con respecto a la estimación, el efecto de emplear las variables auxiliares SD+AD en el proceso de estimación, en la mayoría de los casos, es muy similar al efecto producido por las sociodemográficas: solo en las estimaciones de *radio TDT ayer* y *radio en casa ayer* se observa una pauta más parecida a la seguida por las ponderaciones con datos administrativos.

Los modelos de estimación empleados en esta investigación —CA, SP y DA— aportan cierta variabilidad a las estimaciones, sobre todo cuando se tiene en cuenta el conjunto de variables sociodemográficas, si se observa el promedio del sesgo relativo. Sin embargo, este dato se difumina cuando se observan las estimaciones por separado. Para la mayoría de las variables, el uso de los diferentes métodos de estimación arroja unos resultados similares cuando se emplea el mismo conjunto de información auxiliar.

La figura 2 presenta los errores típicos de las estimaciones calculados de dos formas, mediante un método linealizado utilizado en el muestreo con reemplazo y mediante replicaciones de la muestra. Es de notar que, en líneas generales, el uso de las variables auxiliares AD tiene un menor impacto en la varianza de las estimaciones, lo que está en línea con los resultados observados en el análisis del sesgo. La clave radica en que el conjunto de variables administrativas no está relacionado con la mayoría de las variables de interés ni con la probabilidad de participar en el estudio, lo que impide que se reduzca el sesgo, pero también mantiene en niveles inferiores el error típico de las estimaciones. Por el contrario, el uso de las variables sociodemográficas tiende a incrementar la varianza de la mayoría de las estimaciones, lo que también se refleja cuando se utiliza el conjunto de variables SD+AD. Además, los errores calculados con el método *jackknife* resultan ser mayores que los linealizados, en buena medida debido a que los errores linealizados no tienen en cuenta la variabilidad derivada de utilizar estimaciones para determinar los coeficientes de ajuste.

Figura 2

Errores típicos (%) de las estimaciones de cada variable de interés según el conjunto de variables auxiliares y el método de estimación de la varianza.



Discusión y conclusiones

Las dos primeras hipótesis planteadas en este trabajo trataban sobre el conjunto de variables más efectivo para reducir el sesgo de las estimaciones. Esta investigación comprueba el efecto de usar tres conjuntos de variables auxiliares —sociodemográficas, administrativas agregadas a nivel de municipio y la combinación de ambas— para reducir el sesgo de las estimaciones de dos encuestas provenientes de un panel de internautas. Los datos administrativos agregados, por su gran número y variedad, podrían ser utilizados como variables auxiliares para ajustar las estimaciones. Sin embargo, los resultados de esta investigación no avalan esa posibilidad; las ponderaciones basadas en datos administrativos solo reducen levemente el nivel de sesgo de las estimaciones. Este hallazgo está en la línea de lo observado por Biemer y Peytchev (2013) y Lahtinen, Kaisa y Butt (2015), que no encontraron útiles las variables administrativas agregadas para corregir el sesgo producido por la falta de respuesta. Al contrario que en los trabajos anteriores, en los que se usaron muestras probabilísticas como la Encuesta Social Europea, en esta investigación se ha simulado una muestra no probabilística. Sin embargo, incluso en ese escenario, los datos administrativos agregados no han sido de gran utilidad para reducir el sesgo de las estimaciones.

En la comparación propuesta, los datos más efectivos a la hora de ajustar la muestra son, en promedio, la combinación de las variables sociodemográficas con los datos administrativos agregados y el método de cuasi aleatorización. Pero este resultado debe ser tomado con cautela por dos motivos. En primer lugar, porque las diferencias entre los promedios del sesgo de las estimaciones ponderadas y sin ponderar es de apenas 1,6 puntos porcentuales. Y, en segundo lugar, porque la efectividad de esa ponderación se debe principalmente a las variables sociodemográficas, como se deduce del análisis de las estimaciones que usan los pesos elaborados a partir de las variables administrativas y sociodemográficas por separado.

Queda claro que el motivo por el que las variables administrativas agregadas no sirven para reducir el sesgo de las estimaciones es que no están correlacionadas con la probabilidad de formar parte de la muestra ni con las variables de interés. Sin embargo, queda por discernir si el problema radica en qué miden las variables —desde el comportamiento electoral hasta la proporción de coches de lujo— o en la naturaleza agregada de los datos. A este respecto, Biemer y Peytchev (2013) plantean la necesidad de que los datos agregados estén correlacionados con las características individuales de los elementos de la muestra para ser efectivos. Este planteamiento, no obstante, necesita ser comprobado. Es necesaria más investigación para saber si existe algún contexto en el que los datos agregados, dada su naturaleza, puedan servir para ajustar estimaciones realizadas a partir de encuestas.

Por otro lado, trabajos como el de Peytchev, Presser y Zhang (2018) han intentado replantear la selección de las variables auxiliares en el tiempo del *big data*. Según los autores, es necesario contar con datos que teóricamente tengan encaje con las variables a estimar y la probabilidad de responder, en lugar de utilizar un elevado número de variables que pueden no estar relacionadas con el objeto de estudio. Los resultados de esta investigación refuerzan el planteamiento de los autores: la teoría es necesaria a la hora de seleccionar las variables auxiliares.

La tercera hipótesis versaba sobre la interacción entre los datos utilizados y la técnica de estimación. En esta investigación se han utilizado las variables auxiliares en tres modelos de estimación: la cuasi aleatorización, los modelos de superpoblación y el doble ajuste. Los modelos de cuasi aleatorización y doble ajuste han conseguido optimizar la información auxiliar en cierta medida. Sin embargo, en el caso de los modelos de superpoblación, en conjunción con las variables administrativas, ha resultado en un ligero aumento del sesgo medio de las estimaciones. Por lo general, la variabilidad de las estimaciones se debe en mayor medida a las variables auxiliares utilizadas que al método de estimación.

Por último, la cuarta hipótesis hacía alusión a los errores de las estimaciones. Se esperaba que el uso de las variables administrativas, además de reducir el sesgo en mayor medida, también pudiera incidir de forma positiva en la reducción de la varianza de las estimaciones. Los errores estimados señalan que los pesos generados a partir de las variables administrativas producen unos errores típicos más reducidos. Sin embargo, este efecto tiene que ver con la mínima variabilidad de los pesos en sí y no con la capacidad de las ponderaciones de ajustar las estimaciones.

Para concluir, hay que incidir en las limitaciones de esta investigación, que tienen que ver con la comparabilidad de las calibraciones individuales con las realizadas a partir de datos agregados y la capacidad de extrapolar los resultados. En primer lugar, el diseño ideal para esta investigación habría consistido en que el mismo conjunto de variables estuviera presente tanto en el conjunto de variables auxiliares en el ámbito individual —aquí llamadas sociodemográficas— como en el conjunto de variables contextuales. Sin embargo, limitar las variables al sexo y la edad suponía dejar fuera la posible ventaja de utilizar datos agregados, que son más accesibles y de los que hay una gran variedad. La segunda limitación tiene que ver con la capacidad de generalizar las conclusiones que se han extraído del análisis de las dos encuestas del panel AIMC-Q a otras situaciones en las que puedan utilizarse datos administrativos. Aunque se admite que dos encuestas de un panel de internautas no permiten extrapolar las conclusiones a la totalidad de investigaciones con muestras no probabilísticas, sí aportan una nueva evidencia que contribuye a crear conocimiento sobre el uso de datos agregados en el tratamiento de la falta de respuesta y en los problemas de cobertura.

Financiación

Esta investigación está financiada por un contrato predoctoral de la Fundación Bancaria “La Caixa” LCF/BQ/ES16/11570005.

Agradecimientos

El autor desea agradecer a la Asociación para la Investigación de los Medios de Comunicación (AIMC) el acceso a los datos del AIMC-Q panel, así como al Instituto de Opinión Pública (IMOP Insights) y a Sara Varela su inestimable ayuda a la hora de documentar la base de datos. También desea agradecer a Modesto Escobar el apoyo en el transcurso de esta investigación.

Notas

[1] Los datos poblacionales corresponden al año 2017 y fueron extraídos del padrón de habitantes del Instituto Nacional de Estadística.

Bibliografía

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile y Roger Tourangeau. 2013. "Summary report of the aapor task force on non-probability sampling". *Journal of Survey Statistics and Methodology* 1(2): 90-105. https://doi.org/10.1093/jssam/smt008.

Bethlehem, J. y S. Biffignandi. 2011. *Handbook of Web Surveys*. Londres: Wiley. https://doi.org/10.1002/9781118121757.

Biemer, Paul y Andy Peytchev. 2012. "Census geocoding for nonresponse bias evaluation in telephone surveys". *Public Opinion Quarterly* 76(3): 432-52. https://doi.org/10.1093/poq/nfs035.

Biemer, Paul y Andy Peytchev. 2013. "Using geocoded census data for nonresponse bias correction: An assessment". *Journal of Survey Statistics and Methodology* 1(1): 24-44. https://doi.org/10.1093/jssam/smt003.

Blom, Annelies G., Michael Bosnjak, Anne Cornilleau, Anne Sophie Cousteaux, Marcel Das, Salima Douhou y Ulrich Krieger. 2016. "A comparison of four probability-based Online and mixed-mode panels in Europe". *Social Science Computer Review* 34(1): 8-25. https://doi.org/10.1177/0894439315574825.

Blom, Annelies G., Christina Gathmann y Ulrich Krieger. 2015. "Setting up an online panel representative of the general population: The German Internet Panel." *Field Methods* 27(4): 391-408. https://doi.org/10.1177/1525822X15574494.

Brick, J. Michael. 2011. "The future of survey sampling". *Public Opinion Quarterly* 75(5 SPEC. ISSUE): 872-88.

Buelens, Bart, Joep Burger y Jan A. van den Brakel. 2018. "Comparing inference methods for non-probability samples". *International Statistical Review* 86(2): 322-43. https://doi.org/10.1111/insr.12253.

Buskirk, T. D., A. Kirchner, A. Eck y C.S Signorino. 2018. "An introduction to machine learning methods". *Survey Practice* 11: 1-36. https://doi.org/10.1007/978-1-4615-5289-5.

Callegaro, M., K. L. Manfreda y V. Vehovar. 2015. *Web survey methodology*. Londres: SAGE Publications.

Chen, Kuang, Richard L. Valliant y Michael R. Elliott. 2018. "Model-assisted calibration of non-probability sample survey data using adaptive LASSO". *Survey Methodology* 44(1). Consulta 11 de Marzo del 2019 (https://www150.statcan.gc.ca/n1/pub/12-001-x/2018001/article/54963-eng.pdf).

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle y Chris Dibben. 2016. "The role of administrative data in the big data revolution in social science research". *Social Science Research* 59: 1-12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Couper, Mick P. 2013. "Is the sky falling? New technology, changing media, and the future of surveys". *Survey Research Methods* 7(3): 145-56.

Dever, Jill, Ann Rafferty y Richard Valliant. 2008. "Internet surveys: can statistical adjustments eliminate coverage bias?". *Survey Research Methods* 2(2): 47-60.

Dibben, Chris, Mark Elliot, Heather Gowans y Darren Lightfoot. 2015. "The data linkage environment". Pp. 36-62 en *Methodological Developments in Data Linkage*. Nueva Jersey: John Wiley & Sons. https://doi.org/10.1002/9781119072454.ch3.

Dorfman, Alan H. y Richard Valliant. 2005. "Superpopulation models in survey sampling". Pp. 1575-77 en *Encyclopedia of Biostatistics*. Chichester: John Wiley & Sons. https://doi.org/10.1002/0470011815.b2a16076.

Elliott, Michael R. y Richard Valliant. 2017. "Inference for nonprobability samples". *Statistical Science* 32(2): 249-64. https://doi.org/10.1214/16-STS598.

Ferri-García, R. y M. D. M. Rueda. 2018. "Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys". *SORT: statistics and operations research transactions* 42(2): 159-182.

Friedman, J., T. Hastie y R. Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent". *Journal of statistical software* 33(1).

Groves, Robert M. y M. Couper. 1998. *Nonresponse in household interview surveys*. Nueva York: John Wiley & Sons. https://doi.org/10.1002/9781118490082.

Gummer, Tobias y Joss Roßmann. 2018. "The effects of propensity score weighting on attrition biases in attitudinal, behavioral, and socio-demographic variables in a short-term web-based panel survey". *International Journal of Social Research Methodology* 22(1): 81-95. https://doi.org/10.1080/13645579.2018.1496052.

Hastie, T., R. Tibshirani y M. Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hays, Ron D., Honghu Liu y Arie Kapteyn. 2015. "Use of internet panels to conduct surveys". *Behavior Research Methods* 47(3):685-90. https://doi.org/10.3758/s13428-015-0617-9.

Kang, J. D. Y. y J. L. Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". *Statistical Science* 22: 523-539.

Kish, Leslie. 1965. *Survey sampling*. Nueva Delhi: John Wiley & Sons.

Kreuter, Frauke. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information.* Nueva York: John Wiley & Sons. https://doi.org/10.1002/9781118596869.

Künn, Steffen. 2015. "The challenges of linking survey and administrative data". *IZA World of Labor* 1-10.

Lahtinen, Kaisa y Sarah Butt. 2015. "Using auxiliary data to model nonresponse bias The challenge of knowing too much about nonrespondents rather than too little?". Artículo presentado en el International Workshop on Household Nonresponse, 2 de septiembre, Leuven, Belgium.

Lee, Sunghee y Richard Valliant. 2009. "Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment". *Sociological Methods & Research* 37(3): 319-43. https://doi.org/10.1177/0049124108329643.

de Leeuw, Edith, Joop Hox y A. Luiten. 2018. "International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data". *Survey Insights: Methods from the Field* 1-11. Consulta 11 de Marzo del 2019 (https://surveyinsights.org/?p=10452).

Levy, Paul S. y Stanley Lemeshow. 2013. *Sampling of Populations: Methods and Applications*. Nueva York: John Wiley & Sons.

Lohr, Sharon L. y Trivellore E. Raghunathan. 2017. "Combining survey data with other data sources". *Statistical Science* 32(2): 293-312. https://doi.org/10.1214/16-STS584.

Mercer, Andrew, Arnold Lau y Courtney Kennedy. 2018. *For Weighting Online Opt-In Samples, What Matters Most?* Washington: Pew Research. Consulta 11 de Marzo del

2019 (http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most).

Morris, Sarah, Alun Humphrey, Pablo Cabrera Álvarez y Olivia D'Lima. 2016. *The UK Time Diary Study 2014-2015. Technical Report*. Londres: NatCen Social Research. Consulta 11 de Marzo del 2019 (http://doc.ukdataservice.ac.uk/doc/8128/mrdoc/pdf/8128_natcen_reports.pdf).

Neyman, Jerzy. 1934. "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". *Journal of the Royal Statistical Society* 97(4): 558. https://doi.org/10.2307/2342192.

Park, A., C. Bryson, E. Ciery, J. Curtice y M. Phillips. 2013. *British Social Attitudes 30th Report*. Londres: NatCen Social Research. Consulta 11 de Marzo del 2019 (http://www.bsa.natcen.ac.uk/media/38723/bsa30_full_report_final.pdf).

Pasek, Josh. 2015. "Beyond probability sampling: population inference in a world without benchmarks". *SSRN Electronic Journal* X(8):133-42. https://doi.org/10.2139/ssrn.2804297.

Pasek, Josh. 2016. "When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence" *International Journal of Public Opinion Research* 28(2): 269-91. https://doi.org/10.1093/ijpor/edv016.

de Pedraza, Pablo, Kea Tijdens, Rafael Muñoz de Bustillo y Stephanie Steinmetz. 2010. "A Spanish continuous volunteer web survey: sample bias, weighting and efficiency". *Revista Española de Investigaciones Sociológicas* 131(1): 109-30.

Peytchev, Andrey y Trivellore Raghunathan. 2013. "Evaluation and use of commercial data for nonresponse bias adjustment". Ponencia presentada en American Association for Public Opinion Research annual conference, Boston, EE.UU.

Peytchev, Andrey, Stanley Presser y Mengmeng Zhang. 2018. "Improving traditional nonresponse bias adjustments: combining statistical properties with social theory". *Journal of Survey Statistics and Methodology* (January): 1-25. https://doi.org/10.1093/jssam/smx035.

Playford, Christopher J., Vernon Gayle, Roxanne Connelly y Alasdair JG Gray. 2016. "Administrative social science data: The challenge of reproducible research". *Big Data & Society* 3(2): 1-13. https://doi.org/10.1177/2053951716684143.

Särndal, Carl-Erik y Sixten Lundström. 2005. *Estimation in surveys with nonresponse*. Nueva York: John Wiley & Sons. https://doi.org/10.1002/0470011351.

Schonlau, M., A. Van Soest, A. Kapteyn y M. Couper. 2009. "Selection bias in web surveys and the use of propensity scores". *Sociological Methods and Research* 37: 291-318. https://doi.org/10.1177/0049124108327128.

Smith, Tom W. 2011. "The report of the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys". *International Journal of Public Opinion Research* 23(3): 389-402. https://doi.org/10.1093/ijpor/edr035.

Smith, Tom W. y Jibum Kim. 2013. "An assessment of the multi-level integrated database approach". *The ANNALS of the American Academy of Political and Social Science* 645(1): 185-221. https://doi.org/10.1177/0002716212463340.

Stevens, Leslie A. y Graeme Laurie. 2014. "The administrative data research centre scotland: a scoping report on the legal & ethical issues arising from access & linkage of administrative data". Research Paper 2014/35. Edinburgh School of Law.

Valliant, R., A. H Dorfman y R. M. Royall. 2000. *Finite population sampling and inference: A prediction approach*. Nueva York: Wiley Series In Probability And Statistics.

Valliant, Richard y Jill A. Dever. 2011. "Estimating propensity adjustments for volunteer web surveys". *Sociological Methods & Research* 40(1): 105-137. https://doi.org/10.1177/0049124110392533.

Valliant, Richard, Jill A. Dever y F. Kreuter. 2018. *Practical tools for designing and weighting survey samples*. New York: Springer.

Valliant, Richard. 2019. "Comparing alternatives for estimation from nonprobability samples". *Journal of Survey Statistics and Methodology*: 1-33. https://doi.org/10.1093/jssam/smz003.

Wang, Wei, David Rothschild, Sharad Goel y Andrew Gelman. 2015. "Forecasting elections with non-representative polls". *International Journal of Forecasting* 31(3): 980-91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Weiseberg, Herbert. 2005. *The total survey error approach*. Chicago: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226891293.001.0001.

West, Brady T. y Roderick J. A. Little. 2013. "Non-response adjustment of survey estimates based on auxiliary variables subject to error". *Journal of the Royal Statistical Society. Series C: Applied Statistics* 62(2): 213-31. https://doi.org/10.1111/j.1467-9876.2012.01058.x.

West, Brady T., James Wagner, Frost Hubbard y Haoyu Gu. 2015. "The utility of alternative commercial data sources for survey operations and estimation: evidence from the national survey of family growth". *Journal of Survey Statistics and Methodology* 3(2): 240-64. https://doi.org/10.1093/jssam/smv004.

Woollard, Matthew. 2014. *Administrative data: Problems and benefits: A perspective from the United Kingdom*. Editado por A. Dusa, D. Nelle, G. Stock y G. Wagner. Berlin: SCIVERO.

Wu, C. y R. R. Sitter. 2001. "A model-calibration approach to using complete auxiliary information from survey data". *Journal of the American Statistical Association*, 96(453):.185-193.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser y R. Wang. 2011., "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples". *Public Opinion Quarterly* 75: 709-747. https://doi.org/10.1093/poq/nfr020.

Anexo I. Tablas

Tabla 3

Estadísticos descriptivos de las ponderaciones

		Media	Desv. Típica	Mín.	Max.	DEFF
Estudio prensa						
Cuasi aleatorización						
	SD	19700,2	15728,3	5477,8	215465	1,64
	AD	19526,7	7216,2	697,8	77760,1	1,14
	SD+AD	18050,5	9769,4	132	101971,7	1,29
Superpoblación						
Diario en papel ayer	SD	20553	2355,2	15696,1	26561,8	1,01
	AD	20553	206,3	19310,5	21255,8	1,00
	SD+AD	20553	2041,8	16357,4	27387,1	1,01
Prensa deportiva papel ayer	SD	20553	3367	10426,3	26924,2	1,03
	AD	20553	2807,1	10955	30097,3	1,02
	SD+AD	20553	2864,8	-8817,6	24644,1	1,02
Prensa económica papel ayer	SD	20553	1009,8	12255,1	21532,8	1,00
	AD	20553	2020,8	-3868,5	22456,4	1,01
	SD+AD	20553	993,1	806,3	21205,7	1,00
Prensa generalista papel ayer	SD	20553	2507,9	15327,2	28246,5	1,01
	AD	20553	164	19648,8	21007,2	1,00
	SD+AD	20553	2117	16025,5	28321,2	1,01
Suplemento 7 días	SD	20553	8063,9	9249	53611,9	1,15
	AD	20553	1205,1	9386,9	22509,5	1,00
	SD+AD	20553	8687,7	9130,7	64431,4	1,18
Doble ajuste						
Diario en papel ayer	SD	20553	11579,1	134,2	119234,4	1,32
	AD	20553	17630,4	5938,1	234088,6	1,74
	SD+AD	20553	7612,7	737,8	81066,5	1,14
Prensa deportiva papel ayer	SD	20553	11601,6	161,9	119829,6	1,32
	AD	20553	16489,6	5658	225484	1,64
	SD+AD	20553	7895	783,4	95876,9	1,15
Prensa económica papel ayer	SD	20553	11058,6	149,6	114514,9	1,29
	AD	20553	16356,4	5704,2	222912,8	1,63
	SD+AD	20553	7559,9	736,9	82205,4	1,14
Prensa generalista papel ayer	SD	20553	11723,5	132,3	122833,2	1,33
	AD	20553	17382,6	6009,5	231934,7	1,71
	SD+AD	20553	7598,1	735,2	81851	1,14
Suplemento 7 días	SD	20553	14446,6	125,5	172818,7	1,49
	AD	20553	19093,4	5557,3	254968,7	1,86
	SD+AD	20553	7602,5	731,6	81980,2	1,14
Estudio radio						
Cuasi aleatorización						
	SD	19171,3	17770,7	5033,8	353573,1	1,86
	AD	19313,3	5264,8	349	53998,2	1,07
	SD+AD	17828,5	10545	1,9	115613,7	1,35
Superpoblación						
Radio ayer	SD	20103,6	1636,6	16782,9	26952	1,01
	AD	20103,6	535,3	18060	22798,3	1,00
	SD+AD	20103,6	1623,3	15625	27274,8	1,01

		Media	Desv. Típica	Mín.	Max.	DEFF
Estudio prensa						
Radio FM ayer	SD	20103,6	2268,3	15987,8	33389,7	1,01
	AD	20103,6	1267,5	15923,6	24447,9	1,00
	SD+AD	20103,6	1477	16259	26733,9	1,01
Radio AM ayer	SD	20103,6	4286,7	13518	48323,3	1,05
	AD	20103,6	24,7	20088	20421,9	1,00
	SD+AD	20103,6	1567,2	18739,3	46712	1,01
Radio TDT ayer	SD	20103,6	271,8	19726	23341	1,00
	AD	20103,6	1327	19136,4	44931,3	1,00
	SD+AD	20103,6	1085,7	19390	41813,1	1,00
Radio en casa ayer	SD	20103,6	2976,8	14768,8	30200,9	1,02
	AD	20103,6	5927,3	-497,7	47021	1,09
	SD+AD	20103,6	629	17740,8	21420,6	1,00
Radio en coche ayer	SD	20103,6	8191,8	-2142	48087,4	1,17
	AD	20103,6	1001,4	18018,9	24308,6	1,00
	SD+AD	20103,6	5179,9	-14,2	33824,1	1,07
Radio en trabajo ayer	SD	20103,6	2575	7868,4	25859,9	1,02
	AD	20103,6	4682,4	-4766,3	30373,6	1,05
	SD+AD	20103,6	1230,6	7709,4	22189,6	1,00
Radio en otro lugar ayer	SD	20103,6	1728,7	15781,6	30163,5	1,01
	AD	20103,6	1026,7	7030,5	22430,3	1,00
	SD+AD	20103,6	1431,5	3116,4	23649,6	1,01
Doble ajuste						
Radio ayer	SD	20103,6	18634,9	5278,6	370772,7	1,86
	AD	20103,6	5519,5	350	57355,7	1,08
	SD+AD	20103,6	11833,3	2,2	128971,2	1,35
Radio FM ayer	SD	20103,6	18634,8	5285	371218,2	1,86
	AD	20103,6	5684,6	333,5	61692,9	1,08
	SD+AD	20103,6	12054,3	2	129349	1,36
Radio AM ayer	SD	20103,6	19760,4	5095	358880,3	1,97
	AD	20103,6	5489,5	366,9	56743,5	1,07
	SD+AD	20103,6	11733	2,2	130935,2	1,34
Radio TDT ayer	SD	20103,6	19476,1	5343,4	387245,1	1,94
	AD	20103,6	5683,9	363,1	55622,7	1,08
	SD+AD	20103,6	12093,7	2,5	125742	1,36
Radio en casa ayer	SD	20103,6	18998,7	5343,4	355236,3	1,89
	AD	20103,6	7738,6	478,7	89497,6	1,15
	SD+AD	20103,6	12114,4	2,2	133443,2	1,36
Radio en coche ayer	SD	20103,6	19371,9	5295,2	357219	1,93
	AD	20103,6	5668,7	414,9	54735,1	1,08
	SD+AD	20103,6	15131,7	1	172882	1,57
Radio en trabajo ayer	SD	20103,6	19548,9	4973,3	388906,1	1,95
	AD	20103,6	6577,6	344,1	57220,4	1,11
	SD+AD	20103,6	12503,9	2,2	136433,1	1,39
Radio en otro lugar ayer	SD	20103,6	19021,4	5278,4	380383,3	1,89
	AD	20103,6	5610,4	315,4	58569,9	1,08
	SD+AD	20103,6	12113,2	1,2	133481,1	1,36

Pablo Cabrera-Álvarez es investigador en formación del Departamento de Sociología de la Universidad de Salamanca. Cursó la licenciatura en Ciencias Políticas en la Universidad Complutense de Madrid (2013), la licenciatura en Sociología (2013) y el máster en *Survey Methods for Social Research* en la *University of Essex* en Reino Unido. Posteriormente, trabajó como estadístico de encuestas en *NatCen Social Research*. Su tesis doctoral trata sobre los problemas de representatividad de las encuestas y cómo estos pueden ser tratados empleando nuevas fuentes de datos.