EMPIRIA. Revista de Metodología de las Ciencias Sociales
ISSN: 1139-5737
ISSN: 2174-0682
empiria@poli.uned.es
Universidad Nacional de Educación a Distancia
España

Cabrera-Álvarez, Pablo
Datos agregados para corregir los sesgos de no respuesta y de cobertura en encuestas
EMPIRIA. Revista de Metodología de las Ciencias Sociales, núm. 49, 2021, -, pp. 39-64
Universidad Nacional de Educación a Distancia
España

DOI: https://doi.org/10.5944/empiria.49.2021.29231

Datos agregados para corregir los sesgos de no respuesta y de cobertura en encuestas[1]

Aggregate data to correct nonresponse and coverage bias in surveys

Pablo Cabrera-Álvarez

Universidad de Salamanca
pablocal@usal.es (ESPAÑA)

Recibido: 04.06 2019
Aceptado: : 16.09.2020

RESUMEN

En las últimas décadas la incidencia creciente de los sesgos de no respuesta y cobertura en las encuestas han puesto en entredicho la capacidad de inferir los resultados a la población. Una forma extendida de corregir los sesgos de no respuesta y cobertura en las encuestas es el uso de ponderaciones que equilibran la muestra final de entrevistados. La construcción de ponderaciones requiere información auxiliar, totales poblacionales que estén disponibles para los que responden y para los que no cooperan. En este trabajo, a partir de simulaciones estadísticas, se comprueba la capacidad de la información agregada para corregir el sesgo de no respuesta. Para ello se comparan el ajuste con datos individuales y el sistema de datos agregados, dando como resultado que el uso de datos agregados puede ser útil si se cumplen tres requisitos: 1) la variable estimada está agrupada, 2) la variable estimada y la auxiliar están correlacionadas y 3) la probabilidad de completar la encuesta está relacionada con la variable auxiliar.

PALABRAS CLAVE

Metodología de encuestas, No respuesta, Ponderaciones, Datos agregados, Simulaciones estadísticas.

[1] El proyecto que ha generado estos resultados ha contado con el apoyo de una beca de la Fundación Bancaria "la Caixa" (ID 100010434), cuyo código es LCF/BQ/ES16/11570005.

ABSTRACT

In the last decades the effect of nonresponse and coverage bias in surveys have questioned the ability of inferring the results to the population. An extended procedure used to correct nonresponse and coverage problems is the use of weights to balance the sample of respondents. However auxiliary information available for respondents and nonrespondents is required to compute weights. In this paper statistical simulations are used to test the potential of aggregate data to correct nonresponse bias. This research compares individual data adjustments to the use of auxiliary aggregate data. The results show the use of aggregate data can improve survey representativity if three requirements are met: 1) the dependent variable is grouped, 2) the dependent and auxiliary variables are correlated and 3) the auxiliary variable is correlated with response propensities.

KEY WORDS

Survey methodology, Nonresponse, Weighting, Aggregate data, Statistical simulations.

1. INTRODUCCIÓN

Fue en 1788 cuando John Sinclair coordinó una de las primeras encuestas documentadas, un cuestionario con más de 100 preguntas dirigido a los pastores de todas las parroquias de la Iglesia de Escocia. Tras 23 recordatorios, el último de ellos escrito en rojo sangre, consiguió una tasa de respuesta del 100% (de Leeuw y Hox 2011). Mucho ha cambiado la investigación con encuestas desde que John Sinclair pusiera en marcha su censo de parroquias. Ahora cualquier experto daría por imposible alcanzar una tasa de respuesta cercana al 100%, incluso contando con un volumen de recursos suficiente como para poner en marcha la más sofisticada estrategia de recogida de datos.

En las últimas décadas, la extensión de la investigación por internet con el uso de paneles no probabilísticos unida a una caída sostenida de las tasas de respuestas ha dado lugar a un panorama de incertidumbre. Tanto en encuestas telefónicas como presenciales, cada vez menos personas están dispuestas a responder a las preguntas de los encuestadores. Por ejemplo, la tasa de respuesta en encuestas telefónicas en Estados Unidos ha caído del 36% al 6% entre 1997 y 2018 (Kennedy y Hartig 2019). Estos fenómenos —la caída en la tasa de respuesta y la extensión e la investigación online— arrojan dudas sobre el proceso de inferencia en el que descansa la encuesta, por el cual es posible extrapolar la información de la muestra a la población (Valliant, Dever y Kreuter 2017).

Para corregir los sesgos de la muestra que puedan comprometer el proceso de inferencia se puede recurrir, una vez que ha concluido el trabajo de campo, a la generación de ponderaciones basadas en coeficientes que modifican el peso

original de cada caso. Para calcular esas ponderaciones se utiliza información auxiliar, es decir, variables que están disponibles para todos los elementos de la población, tanto los que responden como los que deciden no cooperar. La teoría estadística establece que en la medida en que esas variables auxiliares estén correlacionadas con la probabilidad de responder y con la variable de interés, el sesgo de la estimación será corregido (Bethlehem, Cobben y Schouten 2011).

Algunos trabajos han demostrado que la clave para ajustar una encuesta reside, más que en el método empleado para computar las ponderaciones, en el conjunto de variables auxiliares que se tienen en cuenta (Mercer *et al.* 2018). Sin embargo, las restricciones de acceso a los microdatos poblacionales condicionan la capacidad de implementar los ajustes. Una alternativa a los microdatos consiste en recurrir a los totales poblacionales de fuentes como el censo, que pueden ser utilizados para detectar desviaciones en la distribución de la muestra y posteriormente ajustarla. Además, esos totales poblacionales pueden ser tratados como variables contextuales, es decir, como información del lugar, ya sea una sección censal, un municipio o una empresa, en la que se encuadra el elemento poblacional seleccionado en la muestra.

Esta investigación pretende, a partir de simulaciones estadísticas, determinar la idoneidad de usar totales poblacionales como variables contextuales frente a las variables individuales para ajustar los sesgos presentes en las encuestas. Los resultados apuntan a que el nivel de agrupación de la variable a estimar es el factor más determinante a la hora de que un ajuste con variables contextuales sea efectivo, aunque también deben concurrir dos elementos más, la correlación entre la variable auxiliar y la variable a estimar y la correlación de la variable auxiliar y la probabilidad de responder a la encuesta.

En el primer apartado de este trabajo se presenta el marco teórico y los precedentes de esta investigación. En el segundo se presentan una serie de hipótesis, y posteriormente se exponen los detalles sobre la simulación de los datos y su análisis. En la cuarta sección se trasladan los resultados de las simulaciones. Por último, se discuten los resultados y se presentan las conclusiones.

2. MARCO TEÓRICO

El análisis de la realidad social con encuestas descansa en la posibilidad de inferir las características de la población a partir de una muestra elegida de forma aleatoria. Para ello, la muestra debe ser elegida empleando métodos probabilísticos, y además no deben existir sesgos derivados de la falta de cooperación o de la imposibilidad de entrevistar a algunos elementos de la población, un escenario cada vez más improbable. Dos fenómenos han contribuido a acrecentar los problemas de cobertura y no respuesta en los últimos años. El primero es la caída sostenida de las tasas de respuesta (de Leeuw, Hox y Luiten 2018) y el segundo es la expansión de la investigación por internet basada en muestras no probabilísticas (Blom *et al.* 2016; ESOMAR 2017).

La caída generalizada de las tasas de respuesta arroja dudas sobre si el uso de muestras probabilísticas es, de por sí, suficiente para garantizar el proceso de inferencia a la población. El problema de la no respuesta radica en que los subconjuntos de la población tienen probabilidades diferentes de participar en las encuestas, y la existencia de esa diferencia sistemática provoca que las estimaciones estén sesgadas (Groves y Couper 1998; Dillman *et al.* 2002). La caída en la tasa de respuestas afecta tanto a encuestas presenciales (Beullens *et al.* 2018; de Leeuw, Hox y Luiten 2018) como a las telefónicas (Kennedy y Hartig 2019).

Otro fenómeno que afecta a la calidad de los datos de una encuesta es el sesgo de cobertura que se produce cuando parte de la población objetivo no puede ser contactada. Esta incidencia puede ocurrir porque los elementos poblacionales son inaccesibles, como por ejemplo las personas que residen en centros de internamiento, porque el modo de administración hace imposible que sean entrevistadas, como en el caso de los hogares que no tienen acceso a internet en las encuestas web a población general, o porque los elementos poblacionales no están incluidos en el marco muestral (Weiseberg 2005).

Existen diferentes métodos para corregir el sesgo de cobertura y no respuesta antes (Hansen 2007; Manfreda *et al.* 2008; Mohorko, Leeuw y Hox 2011; Ryu, Couper y Marans 2006; Singer, Groves y Corning 1999), durante (Groves y Heeringa 2006; Lepkowski *et al.* 2013; Olson y Peytchev 2007) y después de la recogida de los datos (Levy y Lemeshow 2013; Little y Vartivarian 2005; Sakshaug y Eckman 2017). Esta investigación se centra en los ajustes que se realizan una vez que ha concluido el trabajo de campo, es decir, las ponderaciones que tienen como objetivo equilibrar la composición de la muestra con respecto a la población. El ejercicio de ponderación en su versión más sencilla consiste en generar un peso w para cada subgrupo j que en su conjunto fuercen a la muestra a reflejar la distribución de la población con respecto a los grupos de la variable auxiliar (z):

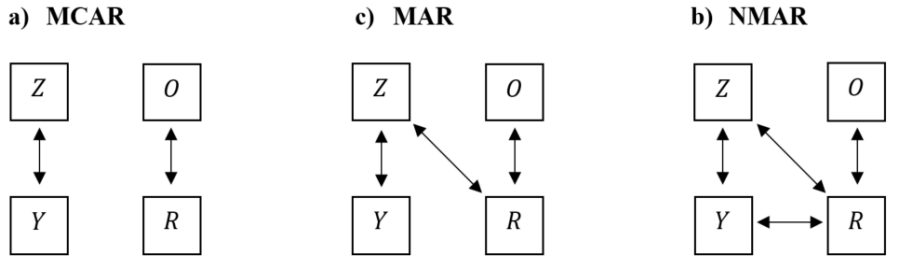
$$w_{zj} = \frac{N_{zj}}{n_{zj}}$$

donde N_{zj} es el total poblacional para los subgrupos de la variable z, y n_{zj} se refiere a los mismos totales, pero para la muestra. Esta es la forma más sencilla de computar una ponderación, el ajuste por celdas, que consiste en crear un cociente entre el total poblacional y el total muestral para las diferentes categorías de una variable. Otros métodos para generar ajustes son la calibración y la postestratificación (Dever, Rafferty y Valliant 2008; Särndal 2007; Tsung, Valliant y Elliott 2018; Zhang 2000). Mediante el primero, la muestra es forzada a replicar la distribución marginal de las variables auxiliares en la población usando para ello los totales poblacionales de cada subgrupo, mientras que, en la postestratificación, además de los marginales también se tienen en cuenta las frecuencias conjuntas. Por su parte, los pesos basados en modelos de respuesta

parten de la probabilidad estimada de que un elemento muestral responda a la encuesta (Bethlehem, Cobben y Schouten 2011; Elliott y Valliant 2017). Para estimar las probabilidades de respuesta se emplean modelos que requieren de un marco muestral con información auxiliar para los que responden y los que no (Bethlehem *et al.* 2011). Por último, cuando se trata de corregir el sesgo de autoselección, se pueden utilizar muestras probabilísticas de referencia (de Pedraza *et al.* 2010; Gummer y Roßmann 2018; Lee y Valliant 2009; Pasek 2016), o técnicas de propensity score matching (Elliott y Valliant 2017; Mercer *et al.* 2018) para determinar cuál es la probabilidad de que un caso dado decida tomar parte en la encuesta, y así poder ajustar la composición de la muestra.

La efectividad de las ponderaciones está determinada por el mecanismo que subyace a los datos perdidos. En la literatura se diferencian tres mecanismos de datos perdidos: MCAR (*missing completely at random* por sus siglas en inglés), MAR (*missing at random*) y NMAR (*not missing at random*) (Bethlehem et al. 2011; Little y Rubin 1987). La Figura 1 adaptada de Bethlehem *et al.* (2011) presenta un resumen de cómo operan los diferentes mecanismos. Bajo el mecanismo MCAR la ponderación es innecesaria ya que la estimación no está sesgada, y bajo el mecanismo NMAR es fútil ya que la participación en la encuesta depende directamente de la variable a estimar. Solo en el caso de MAR se dan las condiciones para que la ponderación, basada en las variables auxiliares (Z), corrija el sesgo en la variable a estimar (Y).

Figura 1. Mecanismos de datos perdidos.



Y : variable a estimar; Z: variable auxiliar; O: variables no observadas; R: participación en la encuesta.

2.1 Variables auxiliares y datos poblacionales agregados

Los estudios sobre el efecto de la ponderación establecen que para disminuir el sesgo de las estimaciones debe darse una doble condición. Por un lado, las

variables auxiliares deben estar relacionadas con la probabilidad de respuesta de los elementos muestrales y, por el otro, las variables auxiliares también deben estar relacionadas con la variable de interés que se pretende estimar (Bethlehem, Cobben y Schouten 2011). El proceso de búsqueda de variables auxiliares que cumplan esta doble condición presenta ciertas limitaciones, en ocasiones teóricas, ya que puede no existir un desarrollo teórico que oriente sobre cuáles son las variables relevantes, y en la mayoría de los casos prácticas, debido a que suele ser reducido el número de variables que contienen información de los que no responden a la encuesta.

Esas variables auxiliares son utilizadas en los ajustes según la forma en la que esté disponible la información poblacional. Cuando la información poblacional existe en forma de microdatos, los datos de encuestas se pueden unir con el marco muestral con el fin, por ejemplo, de construir un modelo para calcular las probabilidades de respuesta (*p. ej.* Park *et al.* 2013). En ese caso estaríamos ante un ajuste individual, porque la información poblacional está disponible de forma desagregada. Un caso diferente es cuando en la muestra existen las variables auxiliares, pero la información poblacional está en forma agregada. En ese escenario las técnicas como la calibración o la postestratificación funcionarían, ya que solo requieren los totales subpoblacionales de las variables auxiliares (Särndal y Lundström 2005). También existe una posibilidad adicional, que es utilizar la información agregada como variables contextuales, es decir cada elemento de la muestra contaría con datos sobre el entorno en el que se encuadra. Por ejemplo, el registro de una persona entrevistada de la que se conoce su municipio puede ser enriquecido con datos como la proporción de personas de más de 65 años o la proporción de coches de lujo en el municipio. Posteriormente, las ponderaciones pueden ser generadas en función de esa información poblacional agregada. Este trabajo se centra en esta última alternativa, que apenas ha sido tratada en la investigación sobre ponderaciones y en su capacidad de ajustar las muestras.

Para ajustar el sesgo presente en las estimaciones realizadas a partir de encuestas se utiliza de forma recurrente la información poblacional agregada. Un ejemplo son los datos del censo, que se utilizan para ajustar la muestra en términos de sexo, edad y distribución territorial (*p. ej.* Park *et al.* 2013). En los últimos años, con la aparición de nuevas fuentes de datos, existe un interés renovado en utilizarlos para corregir el sesgo de las encuestas (Burrows y Savage 2014; Couper 2013). De hecho, ha habido intentos de sistematizar la recogida y uso de información auxiliar como es el caso de la estrategia de datos multinivel integrados (MIDA en inglés), en la que diferentes fuentes de datos auxiliares son combinadas con los datos originales de la encuesta o el marco muestral con el fin de ampliar las posibilidades de ajustar la muestra (Smith 2011; Smith y Kim 2013).

Sin embargo, son pocos los trabajos en los que se han utilizados datos agregados como variables auxiliares para ajustar una encuesta. En uno de ellos, Biemer y Peytchev (2012; 2013) utilizaron datos censales agregados con el fin de corregir el sesgo en las estimaciones realizadas a partir de una encuesta telefónica en los Estados Unidos. A la luz de los resultados los autores concluyeron

que el uso de datos agregados del censo solo es efectivo para ajustar encuestas si los individuos con una determinada característica están agrupados y esta característica está correlacionada con la variable de interés. Más recientemente, en Reino Unido, se comprobó la eficacia de los datos administrativos agregados en el marco de la Encuesta Social Europea (Butt y Lahtinen 2016). Para ello utilizaron diversas fuentes de datos como los registros de criminalidad, el censo, los índices de exclusión social, los datos del Ministerio de Educación o del de Transporte y Medio Ambiente. Los datos, que estaban agregados a nivel municipal o inferior, no resultaron efectivos para corregir el posible sesgo de no respuesta en las estimaciones.

2.2. Modalidades de los datos agregados

A pesar de que existen algunos trabajos empíricos sobre el efecto de los datos agregados utilizados como variables contextuales para reducir el nivel de sesgo en las encuestas, no se ha realizado un análisis teórico sobre en qué casos pueden resultar efectivos a la hora de reducir el sesgo. El trabajo de Biemer y Peytchev (2013) emplea un marco derivado de las características del ajuste estadístico con datos individuales. Ese marco se basa en la demostración de que, si la variable auxiliar está fuertemente correlacionada con la probabilidad de participar en el estudio y con la variable estimada, el sesgo de la estimación será corregido (Bethlehem *et al.* 2011). Según estos autores, para aplicar este marco a los datos agregados existe un requisito adicional: la variable auxiliar debe estar conglomerada para ser un buen proxy de la característica individual. Por ejemplo, si en una sección censal hay un 99% de mujeres, esa información agregada es un buen indicador del sexo de la persona entrevistada en caso de que no haya desvelado esa información. Sin embargo, los autores pasan por alto el papel del nivel de conglomeración de la variable estimada.

Existen varias modalidades que explican cómo los datos auxiliares agregados están relacionados con las variables de interés de la encuesta. En estos modelos básicos intervienen tres elementos, la variable auxiliar (Z), los conglomerados a los que pertenecen los casos (K) y la variable objetivo (Y). Las variables auxiliares recogen información de toda la población objetivo de la encuesta, por lo que pueden ser utilizadas para corregir los sesgos de la muestra. Ejemplos de variables auxiliares agregadas son el nivel de renta o los resultados electorales a nivel de sección censal, el número de coches de lujo matriculados a nivel de municipio o el porcentaje de alumnos de un colegio que tienen asignada una beca de comedor. Por su parte, los conglomerados a los que pertenecen los casos (K) cambian según la población de la encuesta. Así, si la encuesta es a la población general, el municipio o la sección censal son variables de agrupación, mientras que, si la encuesta aborda a la población de estudiantes, el centro escolar o la clase son otras posibles variables de agrupación.

Las relaciones entre el nivel de agrupación (*K*) y la variable auxiliar (*Z*) y objetivo (*Y*) están representadas por la correlación intraclase ρ, que se define como:

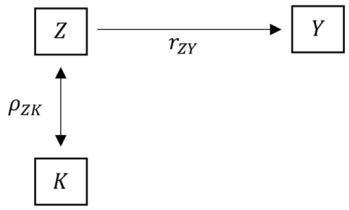
$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

en la que σ_b^2 se refiere a la varianza entre los grupos, que están definidos por la variable de agrupación, y σ_w^2 a la varianza dentro de los grupos (Liljequist, Elfving y Roaldsen 2019). Por lo tanto, ρ toma valores entre 0 y 1, en el que 0 implica que no existe relación entre las variables y 1 que existe una relación perfecta.

Así, el nivel de agrupación de la variable *Z* viene determinado por la correlación intraclase ρ_{ZK}; el nivel de agrupación de *Y* está determinado por ρ_{YK} y la relación entre *Z* e *Y* se expresa con el coeficiente de correlación de Pearson r_{zy}. Aquí se presentan tres posibles escenarios de generación de los datos, en el primero el nivel de agregación de *Y* es dependiente de la relación entre *Z* y *K*, en el segundo la agregación de *Z* depende de la relación entre *Y* y *K*, y en el tercero los niveles de agregación de *Z* e *Y* son independientes.

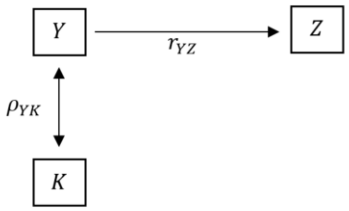
En el primer escenario la variable auxiliar (*Z*) juega un papel determinante al establecer el nivel de agrupación de la variable estimada (*Y*), como se observa en la Figura 2. Para clarificarlo, pensemos que queremos estimar la distribución de la afiliación religiosa de la población (*Y*) a partir de una encuesta. Con el fin de ajustar la encuesta para corregir las desviaciones introducidas por la no respuesta o la falta de cobertura se puede utilizar una variable auxiliar como el país de procedencia. Algunos estudios has mostrado que las personas procedentes de otros países presentan una distribución de la afiliación religiosa diferente que la población autóctona (Santiago y Pérez-Agote 2013) y, además, son más propensos a responder a las encuestas (Morales y Ros 2013). Asimismo, la información sobre el país de procedencia se puede obtener del Instituto Nacional de Estadística agregada a nivel de sección censal o municipio (*K*). En este escenario, a la hora de generarse los datos, la relación entre la variable religión y la variable de conglomeración depende de dos factores, el primero es la medida en que las personas tienden a agruparse en el territorio según su país de procedencia (ρ_{ZK}) y el segundo es la correlación entre la procedencia y la afiliación religiosa (r_{zy}).

Figura 2. El nivel de agrupación de Y es determinado por r_{ZY}.



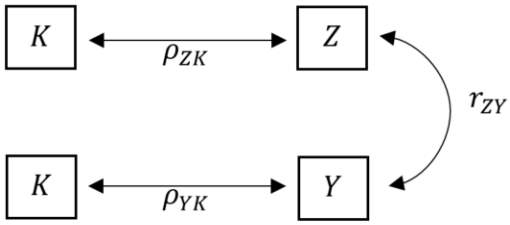
En el segundo escenario la variable *Y* está agrupada de manera independiente, mientras que el nivel de agrupación de *Z* viene dado por la correlación r_{ZY} (Figura 3). Este escenario es idéntico al primer caso planteado, pero en este sería *Y* la variable que determinaría el nivel de agrupación de *Z*.

Figura 3. La agregación de Z ***depende de la relación de*** Y ***con*** K.



Por último, en el tercer escenario las variables *Z* e *Y* son generadas de forma independiente a partir de su relación con la variable de agrupación (*K*), que viene determinada por ρ_{ZK} y ρ_{YK} respectivamente (Figura 4). Para ejemplificar este escenario pensemos en una encuesta a trabajadores que están agrupados en empresas (*K*) en la que se pretende estimar el porcentaje de ellos que disfrutan de la jornada intensiva (*Y*). Para ajustar la encuesta se recurre a la variable contextual de porcentaje de trabajadores en la empresa según rama de conocimiento en la que se formaron (*Z*). El hecho de disfrutar de la jornada intensiva está relacionado directamente con la empresa en la que se trabaja, ya que es la propia organización la que establece la regulación del horario. También es posible encontrar a más trabajadores con similar formación en la misma empresa. Sin embargo, estas dos variables no tienen por qué estar relacionadas entre ellas.

Figura 4. Los niveles de agregación de Z e Y son independientes.



3. HIPÓTESIS

En este apartado se presentan las principales hipótesis del trabajo basadas en la teoría expuesta en la sección anterior.

H1. *Los datos agregados usados como variables contextuales pueden tener una capacidad de ajuste equiparable o incluso mayor que los datos agregados utilizados como totales poblacionales*.

En la mayoría de las encuestas los datos agregados no se utilizan como variables contextuales por diversos motivos, como pueden ser su baja efectividad (Butt y Lahtinen 2015), la falta de correlación con las características individuales de los entrevistados (Biemer y Peytchev 2013), o por el posible contra efecto de la falacia ecológica (Robinson 2011). Esta investigación trata de determinar si el uso de variables contextuales puede llegar a presentar mejores resultados que cuando los datos agregados se usan como totales poblacionales a la hora de tratar el sesgo de las estimaciones.

H2. *La capacidad de ajuste de los datos agregados usados como variables contextuales depende del grado de agrupación de la variable auxiliar medido con la correlación intraclase*.

Biemer y Peytchev (2013) plantean que uno de los requisitos para que los datos agregados puedan ser efectivos a la hora de reducir el sesgo de no respuesta es que la variable auxiliar esté agrupada. La lógica que siguen los autores es que las características contextuales deben ser predictoras de las características individuales, por ejemplo, se espera que la media de ingresos de una sección censal sea un buen indicador de los ingresos del individuo incluido en la muestra.

H3. *La correlación entre la variable auxiliar y la variable estimada es relevante tanto en el ajuste con datos individuales como en el ajuste con datos agregados*.

En línea con la H3, Biemer y Peytchev (2013) plantean un segundo requisito que consiste en que la variable auxiliar agregada esté correlacionada con la va-

riable de interés. En definitiva, lo que hipotetizan estos autores es que el marco de reducción del sesgo que se aplica a las variables individuales es igualmente válido cuando se emplean variables contextuales. En el marco de los datos individuales, para que una ponderación funcione, la variable auxiliar debe estar correlacionada con la propensión a responder y con la variable estimada.

H4. *El efecto de la correlación entre la variable auxiliar y la dependiente y el nivel de agregación de la variable auxiliar sobre la capacidad de ajuste de los datos agregados depende de la modalidad de generación de estos*.

La modalidad de los datos, es decir, como son generados, determina en qué medida la correlación entre la variable auxiliar y la dependiente o el nivel de conglomeración de las variables afecta a la capacidad de reducir los sesgos. Por ejemplo, se espera que en el caso de que la variable auxiliar esté relacionada directamente con los conglomerados, la correlación entre la información auxiliar y la variable estimada juegue un rol importante a la hora de reducir el sesgo.

H5. *El tamaño de los conglomerados o nivel de agregación de los datos no está relacionado con la capacidad de ajuste de los datos agregados*.

En línea con lo descubierto por Butt y Lahtinen (2016) en su investigación con datos de la Encuesta Social Europea en Reino Unido, es de esperar que una vez que los datos están conglomerados, el nivel al que han sido agrupados no esté relacionado con la capacidad de ajuste. En la práctica sería indiferente que los datos utilizados estén agrupados a nivel de sección censal o municipio porque el efecto sería muy similar.

H6. *La magnitud del sesgo de las estimaciones no está relacionada con la capacidad de ajuste de los datos agregados*.

Se podría hipotetizar que en escenarios en los que la magnitud del sesgo es mayor, la capacidad de ajuste de los datos también *puede* serlo. Sin embargo, esta posibilidad solo se materializa si la variable auxiliar está relacionada con la probabilidad de responder y con la variable estimada. Por lo tanto, lo relevante no es la magnitud del sesgo, sino la capacidad de corrección de las variables auxiliares utilizadas.

4. METODOLOGÍA

En esta sección, en primer lugar, se expone el proceso de generación de los datos simulados. Posteriormente, se explica el procedimiento de ajuste seguido y, por último, se presenta la metodología empleada para evaluar la eficacia de las ponderaciones.

4.1. Generación de datos simulados

Las simulaciones tienen como fin determinar cuál es el potencial de las variables agregadas para reducir el impacto del sesgo de cobertura o no respuesta,

y qué condiciones se requieren para que ese potencial se despliegue. En este caso se ha llevado a cabo una simulación por escenarios en el que se han combinado posibles valores de los parámetros poblacionales. Para ello se han simulado 500.000 poblaciones (N = 100.000) y tres variables, el conglomerado al que pertenecen los casos (K), una variable a estimar binaria (Y) y otra variable auxiliar binaria (Z). Al generar las poblaciones con tres variables relacionadas se plantea el problema de cuál es la modalidad de los datos, es decir, si las variables son generadas secuencialmente, qué orden debe seguir el proceso. Por ello los datos se han generado siguiendo los tres esquemas propuestos en el marco teórico con el fin de estudiar cómo la modalidad de los datos agrupados puede afectar a la efectividad de los ajustes a la hora de reducir el sesgo de las estimaciones.

En el primer método de simulación (*congZ*) se genera Z dado un nivel de relación con K, que se establece a través de la correlación intraclase ρ_{ZK}. Posteriormente, Y es generada a partir de su relación con Z, determinada por r_{ZY}. En este caso el nivel de agrupación de la variable Z determina la agrupación de Y. En el segundo método (*congY*) la variable Y es generada teniendo en cuenta la correlación intraclase ρ_{YK}, para posteriormente computar Z con un nivel de correlación determinado por r_{zy}. Este método es idéntico a *congZ*, aunque aquí la variable estimada es el referente para establecer el nivel de agrupación. Por último, en el tercer método (*congInd*) las variables Z e Y son generadas de forma independiente a partir de su relación con la variable de agrupación (K). Dentro de cada conglomerado, la variable Y es reordenada para alcanzar un nivel de correlación con la variable auxiliar (Z) determinado por r_{zy}.

Las poblaciones fueron generadas utilizando el paquete fabricatr de R (Blair *et al.* 2018). Al conformar las poblaciones se han utilizado dos procesos, uno para generar los datos agregados y otro para generar las variables correlacionadas. En primer lugar, para crear la variable agregada se ha generado la variable auxiliar (Z) o estimada (Y) a partir del nivel de la correlación intraclase ρ. Para simplificar el siguiente desarrollo, se asume el escenario *congZ*, en el que el valor z de cada elemento i en cada conglomerado k viene definido por:

$$t_i \sim Bern(p_i)$$

$$u_{ik} \sim Bern\left(\sqrt{\rho}\right)$$

$$z_{ik} = \begin{cases} z_{ik} \sim Bern(p_i), & u_{ik} = 1 \\ t_k, & u_{ik} = 0 \end{cases}$$

en la que p_i es la probabilidad de que un elemento presente la característica de interés.

En segundo término, para simular una variable fijando el nivel de correlación se sigue un proceso de cinco pasos. Asumiendo que la variable Z ya ha sido

simulada, se trata de simular la variable Y a partir de un nivel de correlación r_{zy} predeterminado. En el primer paso se calculan los cuantiles de la variable Z:

$$Z_q = F^{-1}(Z)$$

en la que F representa la distribución empírica de la variable (Z). En el segundo paso se extraen los cuantiles a partir de una distribución normal estándar:

$$Z_{std} = \Phi(Z_q).$$

En tercer lugar, se genera una distribución normal estándar de la variable (Y_{std}) a partir del nivel preestablecido de r_{zy} de la siguiente forma:

$$Y_{std} \sim N\left(r_{ZY}\, Z_{std}, (1 - r_{ZY}^2)\right),$$

para posteriormente generar los cuantiles de la variable Y a partir de la distribución normal:

$$Y_q = \Phi^{-1}(Y_{std})$$

Finalmente, la variable Y se genera a partir de la distribución objetivo (G) y los valores de Y_q:

$$Y = G(Y_q)$$

Una vez generadas las poblaciones, se han extraído diferentes muestras, forzando un determinado nivel de sesgo (0,05; 0,1; 0,15; 0,20; 0,25) en la media de la variable estimada. Al seguir este esquema se garantiza que el sistema de datos perdidos oscile entre MAR y NMAR, en el primero (MAR) la probabilidad de responder es explicada por la variable auxiliar, lo que permite corregir el sesgo de la estimación. En el segundo (NMAR), la probabilidad de responder está determinada por una serie de variables no observadas y afecta directamente a la variable estimada. En los datos simulados, la muestra presenta un sesgo inducido en Y, por lo que el caso de MAR ocurre cuando la variable auxiliar (Z) está relacionada en cierta medida con la variable estimada (Y), mientras que NMAR se produce cuando el valor de esa correlación es de cero.

Los parámetros tenidos en cuenta para generar las poblaciones y extraer las muestras se presentan en la Tabla 1. Dado el número de condicionantes incluidos al generar las poblaciones con el diseño factorial, hubo que elegir una muestra de 500.000 poblaciones que fueron simuladas utilizando un sistema de computación en la nube de Microsoft Azure.

Tabla 1. Parámetros tenidos en cuenta al simular las poblaciones y extraer las muestras.

Parámetro	Descripción	Valores
Población		
K	Número de conglomerados	De 50 a 1000 en grupos de 100
$\overline{Y}$	Probabilidad media poblacional de la variable estimada (Y)	Valores de 0,05 a 0,5 en pasos de 0,05
$\overline{Z}$	Probabilidad media poblacional de la variable auxiliar (Z)	Valores de 0,05 a 0,5 en pasos de 0,05
ρ_{YK}	Correlación intraclase entre la distribución en conglomerados y la variable a estimar	De 0,05 a 0,95 en pasos de 0,1
ρ_{ZK}	Correlación intraclase entre la distribución en conglomerados y la variable auxiliar	De 0,05 a 0,95 en pasos de 0,1
r_{zy}	Correlación entre la variable auxiliar y la variable estimada	De 0,05 a 0,95 en pasos de 0,1
Modalidad de los datos	Modalidad usada para generar los datos	*congZ*, *congY y congInd*
Muestra		
$B_{(\bar{y})}$	Sesgo introducido en *y* al seleccionar la muestra	0,05; 0,1; 0,15; 0,20; 0,25
n	Tamaño de la muestra	200; 500; 1000; 2000

4.2. Ajuste de los datos

Una vez generadas las poblaciones simuladas y seleccionadas las muestras se procedió a generar dos ponderaciones, una utilizando los datos individuales (DI), que es el procedimiento habitual y sirve en esta investigación como punto de referencia, y otra empleando los datos agregados (DA), es decir, una variable de tipo contextual.

La primera ponderación se realizó utilizando la variable auxiliar de nivel individual y el total poblacional (DI). Una vez que la muestra había sido seleccionada, tomando como referencia el total poblacional de la variable auxiliar, se procedió a generar la ponderación utilizando el método de calibración lineal. En la calibración lineal, partiendo de una muestra que cuenta con unos pesos de diseño, en este caso iguales a uno, la ponderación final es el resultado de minimizar la distancia entre los pesos de diseño y los pesos finales bajo la condición

de que la distribución de las variables auxiliares sea igual a la de los totales poblacionales de esas variables (Lundstrom y Sarndal 2001).

La segunda ponderación se basó en la variable auxiliar agregada (DA). En este caso se empleó el mismo sistema para calcular la ponderación, pero aquí la información utilizada fue la variable agregada, es decir, un resumen de la variable auxiliar en el conglomerado al que pertenecía el caso. Concretamente, para generar la variable agregada, primero, se procedió a calcular la media de la variable auxiliar en cada conglomerado, y posteriormente esta variable contextual fue dividida en cuartiles para facilitar su uso en la calibración. Finalmente, las variables auxiliares fueron añadidas a los datos muestrales utilizando para ello el conglomerado de pertenencia como clave.

4.3. Evaluación del efecto de los ajustes

Por último, las estimaciones ponderadas por ambos sistemas fueron comparadas con la media poblacional para establecer en qué medida el sesgo presente en la estimación sin ponderar se había reducido. Para ello se calculó una medida de cambio relativo del sesgo (*CRS*) para cada ponderación:

$$CRS = \frac{\left|\boldsymbol{B}_{(\bar{y}_w)}\right| - \left|\boldsymbol{B}_{(\bar{y})}\right|}{\left|\boldsymbol{B}_{(\bar{y})}\right|}$$

en la que $|\boldsymbol{B}_{(\bar{y}_w)}| = |\overline{Y} - \bar{y}_w|$ representa el valor absoluto del sesgo de la estimación ponderada y $|\boldsymbol{B}_{(\bar{y})}| = |\overline{Y} - \bar{y}|$ se refiere al sesgo absoluto de la estimación sin ponderar. Estas medidas de cambio en el sesgo de las estimaciones fueron modeladas por separado, la ponderación individual (DI) y la agregada (DA), con el fin de determinar el impacto de los diferentes factores incluidos en la simulación. Para ello se utilizaron modelos de regresión lineal ajustados con mínimos cuadrados ordinarios. En la Tabla 2 se presentan los estadísticos descriptivos de las variables incluidas en los modelos de regresión. Las interacciones y los términos cuadráticos fueron omitidos para facilitar la interpretación de la tabla.

Tabla 2. Estadísticos descriptivos de las variables incluidas en el modelo de regresión.

Variable	Casos	Media	Desv. Típ.	Min	Max
Variables dependientes					
CRS (DA)	499.500	-0,07	0,16	-1	0,80
CRS (DI)	499.500	-0,08	0,15	-1	0,81
Variables independientes					
k=50 (ref.)	499.500	0,20	0,40	0	1
k=150	499.500	0,20	0,40	0	1
k=250	499.500	0,20	0,40	0	1
k=350	499.500	0,20	0,40	0	1
k=450	499.500	0,20	0,40	0	1
Media Y (p_y)	499.500	0,27	0,14	0,01	0,57
Media Z (p_z)	499.500	0,27	0,14	0,01	0,52
Corr. intraclase *Y* (ρ_{YK})	499.500	0,23	0,29	0,00	0,99
Corr. intraclase *Z* (ρ_{ZK})	499.500	0,22	0,28	0,00	0,87
Corr. XY (r_{zy})	499.500	0,17	0,20	-0,16	0,96
c*ongZ* (ref.)	499.500	0,33	0,47	0	1
congY	499.500	0,33	0,47	0	1
congInd	499.500	0,33	0,47	0	1
Nivel de sesgo	499.500	0,15	0,07	0,05	0,25
n–200 (ref.)	499.500	0,25	0,43	0	1
n–500	499.500	0,25	0,43	0	1
n–1000	499.500	0,25	0,43	0	1
n–2000	499.500	0,25	0,43	0	1

5. RESULTADOS

En esta sección se presentan los resultados de los dos modelos de regresión[2], uno en el que la variable dependiente es el cambio relativo en el sesgo (*CRS*) cuando se utilizan datos agregados (DA) para ajustar la muestra y otro en el que la variable dependiente es el CRS cuando se emplean datos individuales (DI). Las variables independientes son los diferentes parámetros incluidos en las simulaciones (Tabla 2). Los modelos pueden ser consultados en el Anexo I.

[2] El código utilizado para generar y analizar los datos se encuentra disponible en https://github.com/pablocal/pub_empiria_simulations

La Figura 5 presenta el efecto que tiene cada factor incluido en la simulación sobre la capacidad de la ponderación de corregir el sesgo de las estimaciones. Cada gráfico representa, en el eje horizontal, una de las características relevantes incluidas en las simulaciones, mientras que el eje vertical representa, en todos los casos, la proporción en la que varía el sesgo de las estimaciones al aplicar la ponderación (CRS). Cada gráfico, a su vez, contiene cuatro líneas, tres correspondientes a las ponderaciones hechas a partir de datos agregados (DA) y una correspondiente a la ponderación individual (DI). Las tres líneas que representan a las ponderaciones hechas a partir de datos agregados (DA) simbolizan los diferentes mecanismos utilizados para generar las poblaciones: *congZ*, *congY* y *congInd*.

Los resultados se pueden resumir en tres puntos: 1) el nivel de agrupación de la variable estimada tiene un impacto destacado en la reducción del sesgo cuando se utilizan variables auxiliares agregadas (DA); 2) el nivel de correlación entre la variable auxiliar y la variable a estimar también es un factor relevante y 3) el nivel de impacto de estos dos factores depende de la modalidad usada para generar de los datos (*congZ*, *congY* y *congInd)*.

En primer lugar, sobre la relevancia de la agrupación de la variable a estimar, el gráfico a) de la Figura 5 muestra la relación entre la correlación intraclase de la variable estimada y el cambio en el sesgo de la estimación al aplicar la ponderación. Cuando se utilizan los datos individuales (DI) para ponderar, el nivel de agrupación de la variable estimada no afecta a la capacidad de reducir el sesgo. Distinto es el caso de las ponderaciones hechas con datos agregados (DA), en las que cuanto mayor es el nivel de agregación de la variable estimada, mayor es la capacidad de la ponderación de reducir el sesgo. Sin embargo, esta tendencia no es uniforme, se observan diferencias según el sistema utilizado para generar los datos.

Tabla 3. Predicción del cambio relativo del sesgo (CRS) para diferentes valores de la correlación intraclase de Y.

		DA: *congZ*	DA: *congY*	DA: *congInd*	DI
Corr. intraclase Y	**0,0**	0,01	0,01	0,01	-0,08
	0,1	-0,08	-0,04	0,00	-0,08
	0,2	-0,17	-0,09	-0,01	-0,09
	0,3	-0,26	-0,14	-0,03	-0,09
	0,4	-0,35	-0,20	-0,04	-0,10
	0,5	-0,45	-0,25	-0,05	-0,10

La Tabla 3 es una ampliación del gráfico a) en la que se observa con más detalle que el impacto de la agrupación de la variable a estimar varía según sea la modalidad de los datos (*congZ, congY, congInd*). El caso más favorable se da con el sistema *congZ*, en el que el nivel de agrupación de la variable estimada es determinado por su correlación con la variable auxiliar. En ese escenario el

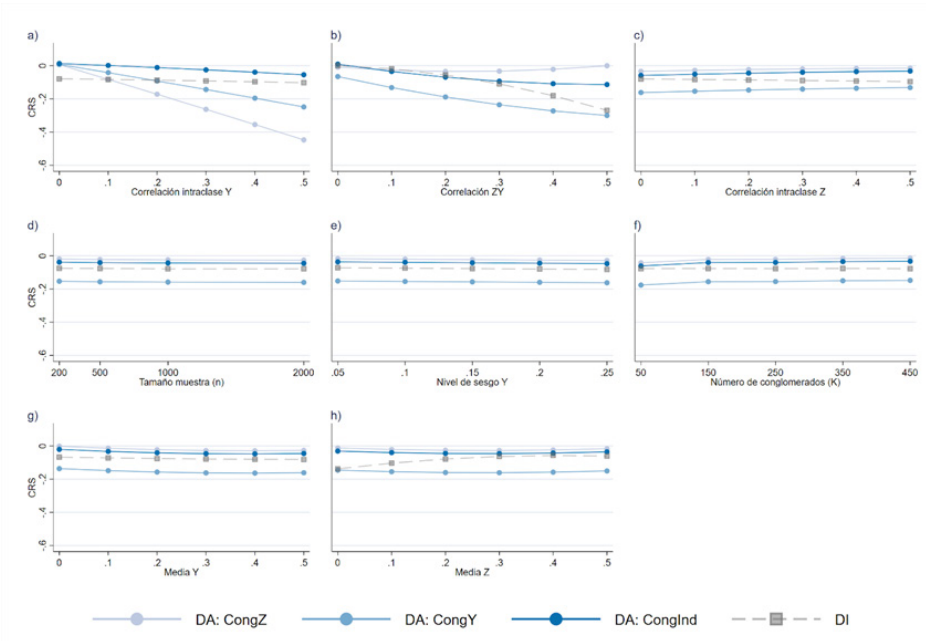
aumento del nivel de la correlación intraclase en una décima supone una reducción media del sesgo de 9 puntos porcentuales, que contrasta con la reducción en las modalidades *congY* (5 puntos) y *congInd* (1 punto). Esta diferencia se explica por las condiciones bajo las que se da la agrupación de la variable estimada. En el caso de *congZ*, para que se de un nivel alto de agrupación deben concurrir dos supuestos: 1) la variable auxiliar debe presentar un nivel alto de agrupación y 2) la variable auxiliar debe estar correlacionada con la variable a estimar. En cambio, bajo el sistema *congY*, que la variable a estimar esté agrupada en no conlleva que la correlación con la variable auxiliar sea alta, por ello el impacto de la agrupación es menor.

En segundo lugar, acerca del nivel de correlación entre la variable auxiliar y la variable a estimar, el gráfico b) representa la relación entre esa correlación y la reducción relativa del sesgo. Si se utiliza la ponderación individual (DI), un aumento en la magnitud de la correlación implica una mayor reducción del sesgo de las estimaciones. La correlación también es importante en el caso de la ponderación computada con datos agregados (DA) generados con el método *congY*, en el que los niveles de reducción del sesgo son incluso mayores que en el caso de los datos individuales (DI).

La Tabla 4 es una ampliación del gráfico b) en la que se observa el comportamiento de la reducción del sesgo según el nivel de la correlación y la modalidad de los datos. El caso más destacado se da bajo el sistema *congY*, en el que, por ejemplo, cuando el nivel de la correlación aumenta en una décima hasta $r_{zy} = 0{,}1$, el *CRS* se reduce de -0,07 a -0,13, mientras que en el caso de los datos individuales (DI), ese mismo cambio en r_{zy} solo implica una variación mínima del *CRS* (de 0,0 a -0,02).

Tabla 4. Predicción del cambio relativo del sesgo (CRS) para diferentes valores de la correlación entre Z e Y

		DA: *congZ*	DA: *congY*	DA: *congInd*	DI
Corr. ZY	**0,0**	-0,01	-0,07	0,01	0,00
	0,1	-0,03	-0,13	-0,04	-0,02
	0,2	-0,03	-0,19	-0,07	-0,06
	0,3	-0,03	-0,24	-0,09	-0,11
	0,4	-0,02	-0,27	-0,11	-0,18
	0,5	0,00	-0,30	-0,11	-0,27



DA: Estimación ponderada con datos agregados; **DI**: Estimación ponderada con datos individuales.

6. DISCUSIÓN

La primera hipótesis (H1) que plantea este trabajo establece la posibilidad de que los datos agregados puedan ser útiles para ajustar desviaciones producidas por la falta de cobertura o la no respuesta. Frente a los resultados de trabajos anteriores (Biemer y Peytchev 2013; Butt y Lahtinen 2016), las simulaciones muestran que bajo determinadas circunstancias el uso de datos agregados puede funcionar e incluso mejorar los resultados que se obtienen al utilizar datos individuales. Sin embargo, esas circunstancias, la agrupación de los elementos por la variable de interés y la correlación de esta con la variable auxiliar, son difíciles de encontrar en los datos que generalmente se usan en Ciencias Sociales. En cuanto a la conglomeración de los elementos según la variable de interés, existen análisis que, teniendo en cuenta variables factuales y actitudinales, indican que la correlación intraclase suele estar por debajo de 0,1 (Kish, Groves y Krotki 1976). Por otra parte, una potencial ventaja de utilizar datos agregados es que son más accesibles y existe una mayor variedad de fuentes, por lo que podría ser más fácil encontrar variables auxiliares correlacionadas con la propensión a responder y las variables de interés. No obstante, en la mayoría de los estudios es difícil encontrar variables auxiliares que presenten niveles altos de correlación con la

variable de interés y la probabilidad de responder. Por lo tanto, con respecto a la H1, el uso de información agregada para corregir desviaciones puede ser efectiva si los sujetos están agrupados según la variable de interés y esa variable está correlacionada con la variable auxiliar.

La segunda hipótesis (H2) parte de las conclusiones del trabajo empírico de Biemer y Peytchev (2013), en el que se establece que la agrupación de la variable auxiliar es necesaria para que los ajustes con variables contextuales tengan éxito. Sin embargo, los datos de las simulaciones apuntan en otra dirección, lo relevante no es el nivel de agregación de la variable auxiliar, sino el de la variable estimada. De todos los factores incluidos en las simulaciones, la correlación intraclase es el más relevante, aunque, como se ha planteado en el párrafo anterior, no es realista asumir en Ciencias Sociales niveles de la correlación intraclase por encima de 0,1, lo que limita el alcance de este hallazgo. Esta hipótesis se complementa con la H3, que se refiere al efecto de la correlación entre la variable estimada y la auxiliar. Esta correlación ya se sabía determinante en el caso de los ajustes con datos individuales (Groves y Couper 1998), pero también es relevante cuando se utilizan datos agregados, hasta el punto de que con el sistema congY, con niveles de correlación entre 0,1 y 0,5, la capacidad de reducir el sesgo está sustancialmente por encima del caso de los datos individuales (DI). Este último escenario abre la puerta a ajustes con datos agregados siempre que se cumplan las condiciones mencionadas anteriormente: 1) que la variable estimada tenga un nivel de agregación por encima de $\rho = 0,1$, 2) que la variable auxiliar esté correlacionada con la variable estimada y 3) que la variable auxiliar y la probabilidad de responder estén correlacionadas.

La forma en que los datos agregados son generados (H4) es fundamental para entender cómo funcionan los ajustes posteriormente. En este trabajo se comparan tres mecanismos de generación de los datos, congZ, en el que la variable auxiliar es la que está conglomerada, congY, en el que es la variable estimada la que está agrupada y congInd, en el que ambas variables son agrupadas de forma independiente. En el párrafo anterior se ha expuesto que cuando se emplean datos agregados para ajustar la muestra, tanto la correlación intraclase de la variable estimada, como la correlación entre la variable auxiliar y la dependiente son elementos clave para determinar el éxito del ajuste. Pero hay que apuntar que estas dos características se ven afectadas por la forma en que los datos han sido generados. En el caso de congZ, cuando la variable estimada está más agrupada, la capacidad de reducir el sesgo es mayor que en cualquiera de los otros sistemas. Esto ocurre porque, para que en este sistema la variable estimada presente un nivel de agrupación alto deben concurrir otros dos elementos, y es que la variable auxiliar esté agrupada y además exista una correlación alta con la variable estimada. Bajo el sistema congY, por su lado, es importante que concurran la dos circunstancias, la correlación de la variable estimada y la auxiliar, así como un nivel alto de agrupación de la variable estimada. En el sistema congInd, al ser los niveles de agrupación independientes, lo más relevante es que exista un nivel alto de correlación entre la variable auxiliar y la dependiente.

Uno de los hallazgos del trabajo de Butt y Lahtinen (2016) tiene que ver con el nivel de agregación de los datos, que no es determinante, ya que una vez que los datos han sido agregados es indiferente al nivel que se realice. Para comprobar este extremo (H5), en las simulaciones, una de las variables manipuladas ha sido el número de conglomerados. En los resultados queda claro que, en consonancia con el trabajo citado, el número de conglomerados no está relacionado con la capacidad de reducir el sesgo cuando se utilizan datos agregados. Lo realmente relevante es que, en esos conglomerados, independientemente de su tamaño, la variable estimada esté agrupada. Sin embargo, hay que señalar que en este trabajo no se han utilizado conglomerados de diferente tamaño, o se han reproducido diferentes sistemas de conglomeración sobre las mismas poblaciones, por lo que no se puede realizar una comprobación definitiva de esta hipótesis.

Otro aspecto para comprobar en esta investigación era si la magnitud del sesgo de la muestra estaba relacionada con la capacidad de corregir las estimaciones (H6). Se podría argumentar que cuanto mayor es el sesgo, mayor capacidad de corregir pueden alcanzar los ajustes estadísticos. Sin embargo, a luz de los resultados, ni en el caso de los datos agregados, ni en el de los individuales, se confirma esta hipótesis. Es cierto que cuanto mayor es el sesgo de la estimación mayor debe ser la corrección de la desviación, pero el tamaño del sesgo no está relacionado con la capacidad de las variables auxiliares de reducirlo.

7. CONCLUSIONES

Para concluir este trabajo se responde a dos cuestiones clave, la primera es sobre la conveniencia de utilizar datos agregados para ajustar los sesgos de no respuesta y cobertura en las encuestas. La segunda trata sobre las limitaciones y el futuro de la presente investigación.

Los datos agregados presentan dos ventajas, existe una gran variedad de fuentes y son más accesible que los microdatos al presentar menos problemas de privacidad. La cuestión es, cómo son más útiles, porque existen diferentes formas de usar los datos agregados: como totales poblacionales en calibraciones individuales o como variables contextuales. Los totales poblacionales usados en calibraciones individuales tienen la limitación de que la información de cada elemento muestral que responda debe ser conocida para realizar el ajuste, algo que puede ser costoso y que no siempre está en los planes de los investigadores en la fase de diseño del estudio. Por el contrario, el uso de variables contextuales permite mucha más flexibilidad, ya que una amplia variedad de predictores puede ser usados una vez que ha concluido el trabajo de campo sin necesidad de recoger ninguna información extra aparte de la unidad geográfica a la que pertenece el elemento muestral. Sin embargo, esta ventaja se ve eclipsada por los supuestos adicionales que deben cumplirse para que las variables agregadas puedan reducir el nivel de sesgo: 1) la variable auxiliar agregada debe estar correlacionada con la probabilidad de responder y la variable estimada y 2) la variable estimada debe estar agrupada. Sobre todo, el segundo es un supuesto improbable, por lo

que sugerimos que los investigadores hagan esta comprobación antes de plantear el uso de variables agregadas en la construcción de ponderaciones.

Esta investigación presenta varias limitaciones que abren nuevas líneas de trabajo para el futuro. En primer lugar, el resultado de las simulaciones, en las que se observa el potencial de los datos agregados bajo determinadas circunstancias, contrasta con las evidencias empíricas que existen hasta el momento. Los casos expuestos en esta investigación en los que se ha intentado utilizar este tipo de datos (Biemer y Peytchev 2013; Butt y Lahtinen 2015; 2016) comparten una característica en común, se trata de encuestas probabilística de alta calidad. Cabe la posibilidad de que la no respuesta o el sesgo de cobertura no sean inconvenientes en estos estudios, mientras que, en otras investigaciones en las que la incidencia de estos fenómenos sea mayor, el uso de variables contextuales pueda ser de ayuda para corregir los sesgos. Más investigación es necesaria en este frente para acercar los resultados de las simulaciones al contexto en el que se mueven los datos reales. Otro interrogante que queda abierto tiene que ver con la influencia del tamaño de los conglomerados y el nivel de agregación. Como ya se ha comentado anteriormente, es necesario seguir trabajando en el efecto que puede tener el tamaño diferencial de los conglomerados, y en el impacto del nivel de agregación de los datos en el plano empírico. Además, un aspecto que esta investigación no ha tratado es el efecto de los ajustes con datos agregados sobre los errores de las estimaciones. Por último, queda abierta la necesidad de desarrollar medidas empíricas que ayuden a decidir a los investigadores sobre la conveniencia de utilizar datos agregados en los ajustes.

8. BIBLIOGRAFÍA

BETHLEHEM, J., COBBEN, F., y SCHOUTEN, B. (2011): Handbook of Nonresponse in Household Surveys, Nueva Jersey, Wiley and Sons.

BEULLENS, K., LOOSVELDT, G., VANDENPLAS C., y STOOP I. (2018): "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?", Survey Methods: Insights from the Field. https://surveyinsights.org/?p=9673

BIEMER, P., y PEYTCHEV, A. (2012): "Census geocoding for nonresponse bias evaluation in telephone surveys", Public Opinion Quarterly, 76(3), 432-452. https://doi.org/10.1093/poq/nfs035

BIEMER, P., y PEYTCHEV, A. (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

BLAIR, G., COOPER, J., HUMPHREYS, A. C. M., Rudkin, A., y Fultz, N. (2018): fabricatr: Imagine Your Data Before You Collect It.

BLOM, A. G., BOSNJAK, M., CORNILLEAU, A., COUSTEAUX, A. S., Das, M., DOUHOU, S., y KRIEGER, U. (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

BURROWS, R., y SAVAGE, M. (2014): "After the crisis? Big Data and the methodological challenges of empirical sociology". Big Data y Society, 1(1), 205395171454028. https://doi.org/10.1177/2053951714540280

BUTT, S., y LAHTINEN, K. (2015): Using auxiliary data to model nonresponse bias The challenge of knowing too much about nonrespondents rather than too little?, presentado en el International Workshop on Household Nonresponse 2015, 02 Sep 2015 - 04 Sep 2015, Leuven, Bélgica.

BUTT, S., y LAHTINEN, K. (2016): ADDResponse : auxiliary data driven non response bias analysis technical report on appending geocoded auxiliary data to Round 6 of European Social Survey (UK), Londres, City University.

COUPER, M. P. (2013): "Is the sky falling? New technology, changing media, and the future of surveys", Survey Research Methods, 7(3), 145-156.

de LEEUW, E. D., y HOX, J. J. (2011): "Internet surveys as part of a mixed-mode design", Social and Behavioral Research and the Internet, 45-76.

de LEEUW, E., HOX, J., y LUITEN, A. (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. https://doi.org/10.13094/SMIF-2018-00008

de PEDRAZA, P., TIJDENS, K., de BUSTILLO, R. M., y STEINMETZ, S. (2010): "A Spanish Continuous Volunteer Web Survey: Sample Bias, Weighting and Efficiency", Revista Española de Investigaciones Sociológicas, 131(1), 109-130.

DEVER, J., RAFFERTY, A., y VALLIANT, R. (2008): "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" Survey Research Methods, 2(2), 47-60. https://doi.org/10.18148/srm/2008.v2i2.128

DILLMAN, D., ELTINGE, J., GROVES, R. M., y LITTLE, R. (2002): "Survey nonresponse in design, data collection and analysis", en Survey nonresponse, Nueva York, Wiley & Sons, 3–26.

ELLIOTT, M. R., y VALLIANT, R. (2017): "Inference for Nonprobability Samples", Statistical Science, 32(2), 249-264. https://doi.org/10.1214/16-STS598

ESOMAR. (2017): Global Market Research 2017. Amsterdam.

GROVES, R.M., y COUPER, M. (1998): Nonresponse in household interview surveys, Nueva York, Wiley and Sons.

GROVES, R.M., y HEERINGA, S. G. (2006): "Responsive design for household surveys: tools for actively controlling survey errors and costs", Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(3), 439-457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

GUMMER, T., y ROßMANN, J. (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

HANSEN, K. (2007): "The effects of incentives, interview length, and interviewer characteristics on response rates in a CATI-study", International Journal of Public Opinion Research, 19(1).

KENNEDY C., y HARTIG, H. (2019): Response rates in telephone surveys have resumed their decline, disponible en https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/ [consultado: 7-09-2020].

KISH, L., GROVES, R. M., KROTKI, K. P. (1976): Sampling errors for fertility surveys. Voorburg, Netherlands: International Statistical Institute.

LEE, S., y VALLIANT, R. (2009): " Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment", Sociological Methods y Research, 37(3), 319-343.

LEPKOWSKI, J. M., MOSHER, W. D., GROVES, R. M., WEST, B. T., WAGNER, J., y GU, H. (2013): "Responsive Design, Weighting, andVariance Estimation in the 2006-2010 National Survey of Family Growth", Vital and Health Statistics. Series 2, Data Evaluation and Methods Research, (158), 1-52.

LEVY, P. S., y LEMESHOW, S. (2013): Sampling of Populations: Methods and Applications, Nueva Jersey, Wiley and Sons.

LILJEQUIST, D., ELFVING, B., ROALDSEN, K. S. (2019): "Intraclass correlation – A discussion and demonstration of basic features", PLoS ONE 14(7), e0219854. https://doi.org/10.1371/journal.pone.0219854

LITTLE, R.J. y RUBIN, D. (1987): Statistical Analysis with Missing Data, Wiley, New York., 381. https://doi.org/10.1002/9781119013563

LITTLE, R. J. A., y VARTIVARIAN, S. (2005): Does Weighting for Nonresponse Increase the Variance of Survey Means? Survey Methodology, 31(2), 161-168.

LUNDSTROM, S., y SARNDAL, C. E. (2001): Estimation in the Presence of Nonresponse and Frame Imperfection, Estocolmo, Statistics Sweden.

MANFREDA, K. L., BERZELAK, J., VEHOVAR, V., BOSNJAK, M., y HAAS, I. (2008): "Web Surveys versus other Survey Modes: A Meta-Analysis Comparing Response Rates", International Journal of Market Research, 50(1), 79-104. https://doi.org/10.1177/147078530805000107

MERCER, A., LAU, A., y KENNEDY, C. (2018): For Weighting Online Opt-In Samples, What Matters Most?, Washington, Pew Research.

MOHORKO, A., LEEUW, E. De, y HOX, J. (2011): "Internet Coverage and Coverage Bias Trends across Countries in Europe and over Time", Background, Methods, Question Wording and Bias Tables, 29(4), 1-28.

MORALES, L., y ROS, V. (2013): "Comparing the response rates of autochthonous and migrant populations in nominal sampling surveys: The LOCALMULTIDEM study in Madrid", en Surveying Ethnic Minorities and Immigrant Populations, Amsterdam, Amsterdam University Press, 147-166.

OLSON, K., y PEYTCHEV, A. (2007): "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes", Public Opinion Quarterly, 71(2), 273-286. https://doi.org/10.1093/poq/nfm007

PARK, A., BRYSON, C., CIERY, E., CURTICE, J., y PHILLIPS, M. (2013): British Social Attitudes 30th Report, Londres, NatCen Social Research.

PASEK, J. (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), 269-291. https://doi.org/10.1093/ijpor/edv016

RYU, E., COUPER, M. P., y MARANS, R. W. (2006): "Survey incentives: Cash vs. in-kind; Face-to-face vs. mail; Response rate vs. nonresponse error", International Journal of Public Opinion Research. https://doi.org/10.1093/ijpor/edh089

SAKSHAUG, J. W., y ECKMAN, S. (2017): "Are survey nonrespondents willing to provide consent to use administrative records? Evidence from a nonresponse follow-up survey in Germany", Public Opinion Quarterly, 81(2), 495-522. https://doi.org/10.1093/poq/nfw053

SANTIAGO, J., y PÉREZ-AGOTE, A. (2013): La nueva pluralidad religiosa, Madrid, Ministerio de Justicia.

SÄRNDAL, C., y LUNDSTRÖM, S. (2005): Estimation in Surveys with Nonresponse.
SÄRNDAL, C. (2007): "The calibration approach in survey theory and practice", Survey Methodology, 33(2), 99-119.
SINGER, E., GROVES, R.M., y CORNING, A.D. (1999): "Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation", Public Opinion Quarterly, 63(2), 251-260. https://doi.org/10.1086/297714
SMITH, T. 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, T. W., y KIM, J. (2013): "An Assessment of the Multi-level Integrated Database Approach", Annals of the American Academy of Political and Social Science (Vol. 645). https://doi.org/10.1177/0002716212463340
TSUNG, K., VALLIANT, R. L., y ELLIOTT, M. R. (2018): "Model-assisted calibration of non-probability sample survey data using adaptive LASSO", (12).
VALLIANT, R., DEVER, J. A., y KREUTER, F. (2018): Practical tools for designing and weighting survey samples, Cham, Springer.
WEISEBERG, H. (2005): The total survey error approach, Chicago, The University of Chicago Press.
ZHANG, L.C. (2000): "Post-Stratification and Calibration-A Synthesis", The American Statistician, 54(3), 178. https://doi.org/10.2307/2685587

ANEXO I: MODELOS DE REGRESIÓN

Tabla 5. Modelos MCO para determinar la reducción del sesgo en el escenario de datos agregados (DA) y datos individuales

	DA	DI
Media Y	-0,119***	0,390***
	(0,004)	(0,003)
Media Y^2	0,206***	-0,472***
	(0,007)	(0,005)
Media Z	-0,147***	-0,053***
	(0,004)	(0,003)
Media Z^2	0,181***	0,049***
	(0,007)	(0,005)
Rho Y	-0,705***	-0,176***
	(0,007)	(0,005)
Rho Y^2	-0,047***	-0,020***
	(0,002)	(0,002)
Rho Z	-0,004*	-0,004**
	(0,002)	(0,001)
Rho Z^2	-0,049***	-0,008***
	(0,002)	(0,002)
Corr. ZY	-0,314***	-0,160***
	(0,002)	(0,002)
Corr. ZY^2	0,487***	-0,862***
	(0,003)	(0,002)
Nivel de sesgo	-0,053***	-0,050***
	(0,002)	(0,001)
CongInd	0,016***	-0,002***
	(0,001)	(0,000)
CongY	-0,008***	-0,005***
	(0,001)	(0,001)
n = 500	-0,003***	-0,001***
	(0,000)	(0,000)
n = 1.000	-0,005***	-0,003***
	(0,000)	(0,000)
n = 2.000	-0,007***	-0,003***
	(0,000)	(0,000)
k = 150	0,020***	0,001
	(0,000)	(0,000)
k = 250	0,021***	0,000
	(0,000)	(0,000)
k = 350	0,026***	0,001***
	(0,000)	(0,000)
k = 450	0,028***	-0,000
	(0,000)	(0,000)
CongInd*Rho Y	0,418***	0,188***
	(0,007)	(0,005)
CongY*Rho Y	0,739***	0,200***
	(0,007)	(0,005)
CongInd*Rho Z		-0,145***
		(0,005)
CongY*Rho Z		0,017***
		(0,001)
CongInd*Corr. ZY	-0,052***	0,021***
	(0,002)	(0,002)
CongY*Corr. ZY	0,106***	0,020***
	(0,002)	(0,002)
Rho Y*Rho Z	0,076***	0,001
	(0,002)	(0,002)
Rho Y*Corr. ZY	-1,113***	0,058***
	(0,003)	(0,003)
Rho Z*Corr. ZY	0,356***	0,109***
	(0,004)	(0,003)
Media Y*Media Z	0,022***	0,002
	(0,006)	(0,005)
Constante	0,044***	-0,038***
	(0,001)	(0,001)
F	39869,39	67714,39
Grados de libertad	28	30
P-valor	0,000	0,000
R cuadrado	0,69	0,80
Casos	499500	499500

DA: Datos agregados; DI: Datos individuales.
Las interacciones entre método y Rho Z fueron omitidas para el primer modelo debido a la falta de observaciones y su efecto adverso en las predicciones.
* p<0.05, ** p<0.01, *** p<0.001