# Aggregate administrative data to adjust selection bias in estimates from nonprobability samples


Pablo Cabrera Álvarez

Department of Sociology and Communication

A thesis submitted for the degree of

*Doctor of Philosophy*

Supervisor

Modesto Escobar Mercado

July 2021

This PhD thesis entitled "*Aggregate administrative data to adjust selection bias in estimates from nonprobability samples*" was funded by "La Caixa" Foundation (LCF/BQ/DE16/11570005).

La presente tesis titulada "*Aggregate administrative data to adjust selection bias in estimates from nonprobability samples*" ha sido financiada por la Fundación de "La Caixa" (LCF/BQ/DE16/11570005).

*A mi padrino y a mi abuela*
*que siempre me acompañan.*

La gloria fulminante de un único minuto,
tu explosión contenida, su apariencia exultante
con ser todo no es nada, con no ser nada es todo:
La mar, el mar.

— Gabriel Celaya

Data-driven prediction can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves.

— Nate Silver

La presente tesis doctoral corresponde a un compendio de trabajos previamente publicados, que se especifican a continuación:

Cabrera-Álvarez, Pablo. 2021. "Datos Agregados Para Corregir los Sesgos de No Respuesta y de Cobertura en Encuestas." *Empiria. Revista de Metodología de Ciencias Sociales* (49):39. doi: 10.5944/empiria.49.2021.29231.

Cabrera-Álvarez, Pablo, and Modesto Escobar. 2019. "The Effect of Weighting and Multiple Imputation on Bias in Spanish Election Polls." *Revista Española de Investigaciones Sociológicas* 165:45–64. doi: 10.5477/cis/reis.165.45.

Cabrera-Álvarez, Pablo. 2021. "Datos Administrativos Agregados y Estimación a Partir de Muestras No Probabilísticas." *Revista Internacional de Sociología* 79(1):e180. doi: 10.3989/ris.2021.79.1.19.350.

D. Rafael Modesto Escobar Mercado, Catedrático de Sociología de la Universidad de Salamanca.

CERTIFICA: Que el trabajo doctoral por compendio de artículos o publicaciones realizado bajo su dirección por D. Pablo Cabrera Álvarez, titulado "*Aggregate administrative data to adjust selection bias in estimates from nonprobability samples*", reúne las condiciones de originalidad requeridas para optar al grado de Doctor en Ciencias Sociales por la Universidad de Salamanca.

Y para que así conste, firma la presente certificación en Salamanca, a 19 de julio de 2021

Fdo. Rafael Modesto Escobar Mercado

# Agradecimientos

Ha llegado el momento de cerrar una etapa. En estas páginas quiero detenerme a pensar en todas aquellas personas que me han acompañado en este trayecto, a los que anduvieron a mi lado y a los que permanecen; a los que iluminaron la senda y a los que me dieron palabras de aliento cuando más lo necesité; a los que apoyaron con los medios y a los que me hicieron reflexionar.

Este viaje no habría sido posible sin Modesto, en él encontré a un guía, pero también a un compañero leal. He tenido la oportunidad de compartir estos años con alguien a quien admiro por su excelencia y su ética del trabajo, pero, sobre todo, por su calidad humana. A Modesto le debo las enseñanzas más valiosas de esta etapa. Él me enseñó que en la academia hay cabida para la generosidad, que la principal virtud de un tutor no es hablar, sino saber escuchar y que la excelencia no es más que la humildad de saber reconocer nuestros horizontes y la voluntad para caminar hacia ellos. Gracias por ser, más allá de mi director de tesis, el maestro paciente y la mano amiga que me ha acompañado durante estos años. En mí quedan para siempre tu entusiasmo por aprender y tu afán por ayudar a los demás. También pienso en quien posibilitó esta aventura. Siempre estaré agradecido a Paca por alentarme desde el principio a mirar más allá de Somosaguas, animarme a hacer las maletas para estudiar en el Reino Unido y, después, hacer de puente con la que se convertiría en mi casa.

También quiero mostrar mi agradecimiento a quienes han hecho posible esta tesis. IMOP Insights, y en especial su directora, Isabel Peleteiro, Sara Varela y todo el equipo. Ellas me enseñaron, hace años, el arte de entrevistar, la magia de convertir una conversación en el principio de un descubrimiento, y durante esta etapa han continuado abriéndome las puertas y apoyándome en la investigación. Los artículos de esta tesis no serían posibles sin su colaboración. También quiero reconocer el papel de la Fundación de “La Caixa”, al permitirme dedicar varios años íntegramente a desarrollar esta investigación, y a su equipo, principalmente Eli, que con su paciencia y empatía siempre se ha mostrado dispuesta a ayudarnos a sortear los recovecos de la burocracia.

También quiero mostrar mi gratitud con las instituciones que me han permitido desarrollar esta etapa. La Universidad de Salamanca y la Facultad de Ciencias Sociales, por brindarme un entorno privilegiado para desarrollarme como investigador y docente. Aquí tuve la oportunidad de enseñar por primera vez, de devolver una parte de lo que durante años me ha sido entregado a través de la educación pública, y por lo que siempre estaré agradecido. En la tarea de enseñar ha sido fundamental el apoyo de la dirección del departamento de Sociología y Comunicación y del resto de mis compañeros, ellos confiaron en mi y me permitieron desarrollar el programa docente de la asignatura de Técnicas Cuantitativas del que me llevo un excelente recuerdo. También tengo palabras de agradecimiento para el seminario de doctorado y su director, Alberto del Rey. El seminario es un espacio abierto y amable en el que muchos hemos tenido la oportunidad de presentar por primera vez nuestra investigación. Por último, quiero agradecer a la *University College of London* (UCL) y, en particular, a Lisa Calderwood por acogerme durante la estancia que tuve la oportunidad de disfrutar en 2019.

Pienso, cómo no, en mis compañeras, ahora amigas, que hicieron de Salamanca un lugar para volver. Tania, Eva y Elena, gracias por aquel primer año en el que volví a sentirme en casa y todo lo que ha seguido hasta el día de hoy. También a los que vinieron después, a Nacho, Juan y Jesús. Con todos ellos he compartido el 101, después 326, lo que es sinónimo de tés, cruzar párrafos, comidas y hasta ratos de ping-pong, los pequeños momentos que hacen especial a la etapa predoctoral. Luis, mi compañero de Madrid con el que, a pesar de la distancia, he compartido en buena medida este camino y que es en parte el culpable de mi inmersión en la ciencia de datos. A todos los compañeros del departamento con los que he compartido momentos. David, por recordarme con una sonrisa que el sur siempre va contigo; Jaime, con quien compartí asignatura, por todo el apoyo brindado durante estos años. Los colegas de la "La Caixa", todos tan distintos, pero con ese afán de escuchar al otro con infinita curiosidad, ya fuera acerca de biología evolutiva, yacimientos arqueológicos o las primeras burbujas económicas. Tampoco me olvido de los compañeros de la UCL —Nancy, Ke, Fivi, Stella y Dawid—, que me ayudaron a reconciliarme con Londres.

Por último, quiero dedicar unas palabras a todas las personas que me han acompañado en este viaje que traspasa las fronteras de lo profesional y lo académico. Mis padres y mi hermano, por todo el amor, la paciencia y el aliento durante estos años. Gracias por intentar entenderme, por el apoyo infinito, sin vosotros nada de esto hubiera sido posible. A mis padrinos, que desde pequeño me animaron a mirar más lejos sin olvidar a los que están a mi lado. A mis abuelas, por el abrigo que me dais todos los días de mi vida. A mis tías, a mis tíos, primas y primos, por el cariño incondicional. A mis politólogos, que me acompañan desde que soñábamos en los pasillos de Somosaguas, por estar siempre, a pesar de los años y los kilómetros. A Cristina y Nelia por ser mi vínculo con el sur y siempre volver a encontrarnos. A Klaudia, Caro y Edu por compartir las desapacibles tardes de Londres en las que buscábamos luz. A Frie por encontrar siempre un rato para reírnos de nuestras aventuras en Colchester. A los que habéis llegado a mi vida en esta última etapa para compartir tantos buenos momentos. A todos, gracias.

**Aggregate administrative data to adjust selection bias in estimates from nonprobability samples**


## Abstract

In recent years, the concurrence of two phenomena has revitalised the methodological debate about inference from nonprobability samples. On the one hand, probability samples increasingly suffer from nonresponse and noncoverage errors, increasing survey costs and leading to biased estimates. On the other hand, the emergence and expansion of the Internet have led to an exponential growth in the use of web surveys with samples recruited using nonprobability methods. Inference from nonprobability samples requires an explicit or implicit model that explains the selection mechanism with respect to the target variable.

This thesis explores an intersection between the need to reduce selection bias in the estimates from nonprobability samples and the opportunity to explain the selection mechanism emerging from newly available aggregate administrative data. To this end, this thesis encompasses three papers that present statistical simulations and two methodological applications using a set of face-to-face and two web surveys conducted in Spain. The first paper uses statistical simulations to explore the conditions under which aggregated data as contextual variables and population totals can reduce or remove selection bias from the estimates. The second paper explores adding sociodemographic and past vote auxiliary variables to the weighting as well as using multiple imputation to improve the quality of the estimates using the pre and post-election surveys of the *Centro de Investigaciones Sociológicas* (CIS) that combine probability selection methods with quotas. The third article tests the effect of including aggregate administrative data at the municipality level to tackle selection bias and improve the quality of the survey estimates using two surveys from an experimental panel of internet users sponsored by the Association for Media Research (AIMC).

The results show that aggregate administrative data is insufficient to correct selection bias in survey estimates, especially when used as contextual variables. The results also

suggest that the aggregate nature of the data is the main impediment to control for selection bias in the estimates.

**Keywords:** inference, nonprobability sampling, administrative data, auxiliary variables, model-based inference.

**Datos administrativos agregados para corregir el sesgo de selección en estimaciones de muestras no probabilísticas**


## Resumen

En los últimos años, la concurrencia de dos fenómenos ha revitalizado el debate metodológico sobre la inferencia a partir de muestras no probabilísticas. Por un lado, las muestras probabilísticas adolecen cada vez más de errores derivados de la no respuesta y la falta de cobertura, lo que aumenta los costes de las encuestas y da lugar a estimaciones sesgadas. Por otro lado, la aparición y la expansión de internet han provocado un crecimiento exponencial del uso de encuestas web con muestras reclutadas mediante métodos no probabilísticos. La inferencia a partir de muestras no probabilísticas requiere un modelo explícito o implícito que explique el mecanismo de selección con respecto a la variable objetivo.

Esta tesis explora una intersección entre la necesidad de reducir el sesgo de selección en las estimaciones realizadas a partir de muestras no probabilísticas y la oportunidad de explicar el mecanismo de selección que surge de los nuevos datos administrativos agregados disponibles. Para ello, esta tesis engloba tres trabajos que presentan una serie de simulaciones estadísticas y dos aplicaciones metodológicas utilizando un conjunto de encuestas presenciales y dos encuestas web realizadas en España. En primer lugar, las simulaciones estadísticas exploran las condiciones bajo las cuales los datos agregados como variables contextuales y totales poblacionales pueden reducir o eliminar el sesgo de selección de las estimaciones. En segundo lugar, utilizando las encuestas pre y postelectorales del Centro de Investigaciones Sociológicas (CIS) que combinan métodos de selección probabilística con cuotas, se explora la adición de variables auxiliares sociodemográficas y recuerdo de voto a la ponderación, así como el uso de técnicas de imputación múltiple para mejorar la calidad de las estimaciones. En tercer lugar, utilizando dos encuestas de un panel experimental de internautas patrocinado por la Asociación para la Investigación de los Medios de Comunicación (AIMC), se comprueba el efecto de incluir datos administrativos

agregados a nivel municipal para atajar el sesgo de selección y mejorar la calidad de las estimaciones de la encuesta.

Los resultados muestran que los datos administrativos agregados son insuficientes para corregir el sesgo de selección en las estimaciones de la encuesta, especialmente cuando se utilizan como variables contextuales. Los resultados también sugieren que la naturaleza agregada de los datos es el principal impedimento para controlar el sesgo de selección en las estimaciones.

**Palabras clave:** inferencia, muestreo no probabilístico, datos administrativos, variables auxiliares, inferencia basada en modelos.

# Table of contents

# 1. Introduction

The most defining feature of surveys is their ability to infer the characteristics of the sample to the population. The sampling method used to select the elements of the population is crucial to enable inference. Statistical theory states that selecting a sample using probability methods is sufficient to support the inference process, given the absence of noncoverage and nonresponse errors (Kish 1965). However, since the beginning of modern sampling, different nonprobability selection methodologies have been developed and applied despite the doubts about their ability to infer (Baker *et al.* 2013). In recent years, the concurrence of two phenomena has revitalised the methodological debate about inference from nonprobability samples. On the one hand, probability samples increasingly suffer from nonresponse and noncoverage errors, increasing survey costs and leading to biased estimates. On the other hand, the emergence and expansion of the Internet have led to an exponential growth in the use of web surveys with samples recruited using nonprobability methods.

The advent of the Internet has prompted new methodological opportunities. The popularity of web surveys lies in their higher speed of data collection at a relatively lower cost compared to other survey modes (Schonlau and Couper 2017). However, web surveys are often used with samples recruited using nonprobability methods, which puts the inference to the population at risk. Moreover, a part of the population does not have internet access, adding another source of error to the estimation process. These issues—uncontrolled selection and noncoverage of the offline population—underly selection bias that occurs if those included in the sample are different from the rest of the population regarding the target variable, making the inference unfeasible (Valliant, Dever, and Kreuter 2018:571). Model-based estimation can deal with selection bias and other sources of error and enable inference (Baker *et al.* 2013:93). This approach to inference relies on implicit or explicit assumptions about the structure of the population, a model, ignoring the sample selection method. An issue, however, is that such models generally require auxiliary variables available for those selected and those not selected in the sample.

Another consequence of technological development is the availability of new data sources or data that previously existed but could not be stored and processed (Baker 2017;

Woollard 2014). These changes have also affected administrative data originally generated to manage and monitor public services, which can also be used for research purposes. The main advantage of this data source is that it often covers the entire population, suggesting that it can help detect and correct biases in survey estimates (Smith and Kim 2013). Public administrations usually release administrative records in the form of totals or means aggregated at certain geographical levels. Such geographically aggregated data provide extensive and varied information that can be used to fit models that aim to improve the quality of the survey estimates.

This research explores an intersection between the need to reduce selection bias in the estimates from nonprobability samples and the opportunity to explain the selection mechanism emerging from newly available aggregate administrative data. To this end, this thesis encompasses three papers that present a set of statistical simulations and two methodological applications using face-to-face and web surveys conducted in Spain. The first paper uses statistical simulations to explore the conditions under which aggregated data as contextual variables[1] and population totals can reduce or remove selection bias from the estimates. The second paper explores adding sociodemographic and past vote auxiliary variables to the weighting and using multiple imputation to improve the quality of the estimates from the pre and post-election surveys of the *Centro de Investigaciones Sociológicas* (CIS) that combine probability selection methods with quotas. The third article tests the effect of including aggregate administrative data at the municipality level to tackle selection bias and improve the quality of the survey estimates using two surveys from an experimental panel of internet users sponsored by the Association for Media Research (AIMC).

In addition to the three articles that present the research results, this PhD thesis contains an introductory chapter that covers the research problem and the relevance of the thesis, the research objectives and provides a common theoretical ground to the articles. First, the introductory chapter presents the research problem, discusses the relevance of the

[1] In this thesis, contextual variables refer to aggregate information at a geographical level, generally administrative data, that informs about the context of the sample member. For instance, it could be the unemployment rate, the election results or the average personal income tax of their census tract, postcode or municipality. This information is geocoded and can be matched to sample members.

thesis, and outlines the research objectives. Second, it reviews theoretical and empirical developments on inference from probability and nonprobability samples. Third, it summarises the adjustment methods developed to reduce the bias of the estimates and enable inference to the population. Fourth, it covers the auxiliary variables, with particular attention to administrative data. Chapters 2, 3, 4 correspond to the three articles that compose the PhD thesis. Finally, chapter 5 presents some conclusions and reflections as a closure to the research.

## 1.1 Research problem and relevance

Survey research has always been aware of technological changes and social dynamics. In survey research, methodological changes stem from the intersection of new information needs and scientific and technical developments. In his work on the evolution of survey research, Groves (2011) points out that the increasing demand for information in the private and public sectors alongside the technical developments in statistics and psychology explains the expansion of survey research in the United States during 1930-1960. At that time, the government needed to understand social dynamics and monitor the impact of public policies in a period of social unrest (Converse 2009). In the private sector, journalists began to support the news with public opinion data, while the expanding service sector demanded information about consumers to develop and sell new products. The increase in information needs coincided with the development of probability sampling theory (Neyman 1934) and the emergence of new methodologies in psychology for measuring attitudes (Likert 1932). This intersection between information needs and technical developments has marked the development of survey methodology up to the present day.

After the era of survey expansion (1960-1990), we now find ourselves in a stage of transformation marked technologically and socially by the irruption of the Internet and methodologically by the sustained fall in response rates in probability surveys (Groves 2011). On the one hand, response rates have followed a declining trend in recent decades (Brick and Williams 2013; de Leeuw *et al.* 2018), which casts doubt on the quality of survey estimates. On the other hand, the spread of the Internet[2] among the population has led to the emergence of a new survey mode, web surveys (Couper 2000; Tourangeau, Conrad, and Couper 2013), while opening the door to the generation, processing and storage of new data sources (Baker 2017). The last decade has seen multiple reflections about the present and future of survey research that agree on two challenges: the role of nonprobability samples and the integration of surveys and new data sources (Couper 2013; Groves 2011; Kalton 2019; Lohr 2017; Miller 2017; Rao and Fuller 2017; Singer 2016). The first challenge is related to the issues arising from the increasing use of nonprobability samples,

[2] According to Eurostat in the EU-27, 90% of households had internet access in 2019, 10% more than in 2014. This percentage reached 91% in Spain and 96% in the United Kingdom (Eurostat 2019).

such as online opt-in panels, and the effects of falling response rates in probability samples. In Groves words, the second challenge is "to discover how to combine designed data with organic data[3], to produce resources with the most efficient information-to-data ratio" (2011:869).

The expansion of the Internet has enabled data collection using web surveys, a relatively new survey mode. This mode allows collecting data in less time and at a lower cost compared to other modes, which explains its popularity in the survey world (Díaz de Rada, Domínguez, and Pasadas 2019). Evidence of the growing popularity of the web survey is the proliferation of web panels mainly recruited using nonprobability methods (Callegaro *et al.* 2014). In 2017, the European Society for Opinion and Marketing Research (ESOMAR), the association that brings together the main market and opinion research agencies, published that the volume of business derived from online quantitative research (27%) almost double that from telephone or personal surveys (15%) (ESOMAR 2017). In the area of methodological research, in addition to a growing body of literature and evidence, the American Association for Public Opinion Research (AAPOR) commissioned two reports to a group of experts on the use of online panel surveys and nonprobability samples (Baker *et al.* 2010, 2013).

Despite the advantages of web surveys, using this mode with nonprobability samples cast doubts on the inference process, the foundation of survey research. From the inference point of view, web surveys present two issues: the noncoverage of the offline population and the lack of a web sampling frame for most populations. A common approach to address these problems in the field of market and public opinion research is the use of opt-in panels made up of volunteers recruited through various methods such as the use of pop-ups, banners or advertisements on a selection of websites (Callegaro *et al.* 2014). Moreover, apart from recruiting volunteers using nonprobability selection methods, these panels usually ignore the offline population. Using these recruitment methods and the systematic exclusion of a part of the population may introduce selection bias in the survey

[3] Groves's definition of organic data refers to data generated automatically to track transactions of all sorts (2011:868). He includes data from social media and the Internet of Things such as radio frequency identification or traffic cameras. This concepts covers most new data sources covered in some definitions of big data.

estimates, preventing inference to the population (Elliott and Valliant 2017:571). Those internet users who voluntarily join the panel may differ from the part of the population without internet access or who did not choose to join the panel. In this respect, the reports of the AAPOR about internet panels and nonprobability samples reach similar conclusions: to estimate population values accurately, the use of nonprobability samples or online panels should be avoided (Baker *et al.* 2010).

Another threat to inference, both with probability and nonprobability samples, is the declining trend in response rates. This trend affects probability surveys such as the Labour Force Survey or the European Social Survey, which have experienced declining response rates over the last decades (de Leeuw *et al.* 2018). In public opinion research, Pew Research has recorded a drop in response rates in telephone surveys from 36% to 9% in the period between 1997 and 2016 (Keeter *et al.* 2017). In Spain, although response rates are typically below 70% (Díaz de Rada 2013), the trend over time is not clear. For example, in the framework of the European Social Survey, a series of interventions aimed at reducing nonresponse resulted in an increase in the response rate from 53% to 66% between 2004 and 2010 (Riba, Torcal, and Morales 2010). The problem with low response rates is that the estimates will be biased to the extent that those who respond are different from those who decline to participate in the study with respect to the variable of interest (Groves and Couper 1998). Although there is no clear relationship between response rates and the magnitude of bias (Groves 2006), a lower response rate opens the door to such biases, which are difficult to control in surveys with many variables that aim to answer different research questions.

Another challenge facing survey methodology, also related to the emergence of the Internet, is integrating surveys with new data sources that have recently appeared or other sources that, although they already existed, could not be accessed or processed before. A few years ago, some voices announced a paradigm shift in which surveys could be replaced by other data sources (Kitchin 2014). Savage and Burrows (2007), in their diagnosis of the incipient crisis of empirical sociology, pointed out that institutions and decision-makers had more valuable data at their disposal to achieve their goals. These new data sources, sometimes referred to as big data, have emerged due to the increasing technological

capacity to generate, process and store information (Japec *et al.* 2015). Depending on the definition adopted[4], the concept of big data encompasses a broad spectrum of data sources, such as data from transactions, administrative data, data generated in social networks, or data produced by sensors and other devices capable of tracking individuals' behaviour (Baker 2017).

However, the initial optimism generated around the increase in technological capacity and the emergence of new data sources has been tempered after confirming the limitations of most of them (Couper 2013). Some of the main limitations are the reduced number of variables contained in some datasets, which makes it challenging to develop explanatory models (Miller 2017; Prewitt 2013); the selection bias arising from the unobserved part of the population (Hsieh and Murphy 2017; Schober *et al.* 2016); the restricted access imposed on proprietary data (Couper 2013), or the measurement bias that occurs when the definition of research concepts differs from the definitions used to generate the data (Connelly *et al.* 2016; Hand 2018).

The limitations of the new data sources do not allow to envisage the replacement of surveys in the near future. However, surveys face challenges such as the inference from nonprobability samples or the drop in response rates in probability surveys. In this sense, there are multiple pronouncements in favour of synergies between survey methodology and new data sources in order to improve the quality of the estimates and expand data coverage (Forsyth and Boucher 2015; Kalton 2019; Miller 2017). Conversely, some researchers also advocate using surveys to improve the quality of new data sources (Kim and Tam 2020; Rafei, Flannagan, and Elliott 2020). In recent years there has been an intense activity to find synergies between the emergence of new data sources and surveys (Hill *et al.* 2019).

[4] Although the notion of a large volume of data is common to most definitions, scholars have not reached a consensus about the ground characteristics of big data (Ward and Barker 2013). One of the first and most extended definitions focuses on three features: velocity, volume, and variety (Laney 2001). Another early, however less known, definition of big data covers all interactions among individuals, institutions, and things are recorded and stored digitally (Negroponte 1995).

This thesis explores a possible synergy between data that are now accessible, aggregated administrative data, and inference from nonprobability samples. Although there are issues that hinder the use of nonprobability samples to infer to the population, unbiased estimation is possible using a model. These models ignore the sample design and only rely on the population structure to estimate (Elliott and Valliant 2017; Särndal 2010). However, models that adjust the sample for selection bias require auxiliary information, variables available for those who participate in the study and those who are not part of the sample. This requirement has historically limited survey adjustments since only a few variables were available for all the population members, typically from the census. However, the increasingly available data sources open up new opportunities for specifying adjustment models. Administrative data consists of information originally collected to organise, manage, monitor or deliver services and can be used for research purposes (Playford *et al.* 2016). One of the main advantages of this data source is that it usually covers the entire population. Although individual-level data are often unavailable for use due to data security and privacy reasons, they are released aggregated at different geographical levels. This research explores the use of aggregated administrative data to correct selection bias when inferring from nonprobability samples.

Building from this context, the thesis seeks to make contributions in three aspects. First, this thesis explores the use of aggregate administrative data to adjust estimates for selection bias. The amount of aggregated administrative data at different geographical levels accessible for research has multiplied in recent years. From a methodological point of view, this data source has some significant advantages, such as the fact that it often covers the entire population and is varied in its domains (Künn 2015). A challenge in using aggregate data is whether to do so as contextual variables, which inform about the environment of the sample element—its census tract or municipality—or as population totals to, through the responses of the sample elements, balance the sample. A series of statistical simulations explore the circumstances that make administrative data in the form of contextual variables or population totals suitable to reduce the bias of the survey estimates.

Second, this thesis investigates using different auxiliary variables as population totals to adjust voting intention estimates. Although this is not a new research question, some aspects, such as the impact of the quality of the auxiliary variables on the estimates, remain unexplored (Bethlehem, Cobben, and Schouten 2011:9). This is the case of the past vote variable, which sometimes is excluded from the adjustments of voting intention estimates due to the possible effect of measurement and item nonresponse errors (Crespi 1988). This research attempts to use imputation models to preserve the quality of the auxiliary variables and enhance the estimation from nonprobability samples.

Third, this thesis examines the feasibility of using geographically aggregated administrative data as contextual variables to reduce the bias of the estimates from two surveys from an online panel. Given that such data are accessible and available for the whole population, it is necessary to test whether these auxiliary variables effectively reduce the bias of the estimates (Smith and Kim 2013). Two open questions are whether the use of administrative aggregate data as contextual variables can help adjust survey estimates and whether the type of model used to adjust the estimates is more relevant than the auxiliary variables that explain selection bias (Mercer, Lau, and Kennedy 2018). The third article adds new evidence to these open questions about auxiliary variables and inference from nonprobability samples.

## 1.2 Research objectives

This research aims to add new evidence on the feasibility of using aggregate administrative data to adjust nonprobability surveys. This section presents the objectives covered in the three research articles. The research hypotheses are presented in each article. The first article, "*Aggregate data to correct for nonresponse and coverage biases in surveys*", includes a series of statistical simulations that test the conditions under which the use of aggregate data as contextual variables is effective in reducing estimation biases:

- To establish the effectiveness of using aggregated auxiliary information as contextual variables as opposed to population totals to adjust survey estimates for coverage and nonresponse biases.

- To determine the data characteristics that enable the use of aggregated auxiliary in adjustment models data to minimise the bias of survey estimates.

The second article, *"The effect of weighting and multiple imputation on bias in Spanish election polls"*, presents the use of different combinations of aggregate auxiliary variables to adjust the estimates of voting intention for the general elections held in Spain between 1982 and 2016 using calibration weighting. In addition, it covers the use of imputation techniques to improve data quality. For this purpose, the pre and post-election surveys of the *Centro de Investigaciones Sociológicas* (CIS) are used. These are the primary objectives:

- To explore the effect on the bias of the voting intention estimates of adding sociodemographic and past vote auxiliary variables to the calibration model.

- To determine the effect of multiple imputation to remove the item nonresponse bias from the auxiliary variable, past vote, and the target variable, voting intention.

The last article of the thesis, *"Aggregate administrative data and estimation from nonprobability samples"*, addresses the use of aggregate administrative data as contextual variables to adjust two surveys of a panel of internet users. Machine learning techniques are applied to deal with a large number of auxiliary variables:

- To examine the potential of aggregate administrative data used as contextual variables to correct the estimates and enable inference from a nonprobability sample.

- To compare the effectiveness of the model specification with the modelling approach—quasi-randomisation, superpopulation, and doubly robust—in adjusting the survey estimates.

- To assess the change in the variance of the estimates after adding the aggregate administrative data as contextual variables to the adjustment model.

Table 1.1 presents a summary of the main features of the three articles of the thesis.

Table 1.1. Summary of the research questions and methods of the articles

| Article | Objectives | Survey data | Adjustment methods | Auxiliary variables |
|---|---|---|---|---|
| *Aggregate data to correct for nonresponse and coverage bias in surveys* | (1) To establish the effectiveness of using aggregated auxiliary information as contextual variables as opposed to population totals to adjust survey estimates for coverage and nonresponse biases.<br><br>(2) To determine the data characteristics that enable the use of aggregated auxiliary in adjustment models data to minimise the bias of survey estimates. | Simulation of populations given different levels of correlation among the auxiliary, target and clustering variables.<br><br>Samples selected assuming selection bias (noncoverage and nonresponse). | Calibration weighting. | Simulated auxiliary variable used as aggregate data contextual variable and population totals. |
| *The effect of weighting and multiple imputation on bias in Spanish election polls* | (1) To explore the effect on the bias of the voting intention estimates of adding sociodemographic and past vote auxiliary variables to the calibration model.<br><br>(2) To determine the effect of multiple imputation to remove the item nonresponse bias from the auxiliary variable, past vote, and the target variable, voting intention. | Face-to-face pre and post-election surveys of the Spanish general elections from 1982 to 2016.<br><br>Multistage sampling. Probability methods to select households and quotas to select household members. | Calibration weighting and multiple imputation for item nonresponse of past vote and voting intention. | Weighting. Auxiliary variables used as population totals.<br>(1) Raw past vote and imputed past vote for missing values.<br>(2) Sociodemographics: age-sex, region, education and status.<br><br>Imputation.<br>(1) Sociodemographics: sex, age, education.<br>(2) Attitudinal: past vote, ideology, evaluation political and economic situation. |
| *Aggregate administrative data and estimation from nonprobability samples* | (1) To examine the potential of aggregate administrative data used as contextual variables to correct the estimates and enable inference from a nonprobability sample.<br><br>(2) To compare the effectiveness of the model specification with the modelling approach—quasi-randomisation, superpopulation, and doubly robust—in adjusting the survey estimates.<br><br>(3) To assess the change in the variance of the estimates after adding the aggregate administrative data as contextual variables to the adjustment model. | Two surveys from an online panel of internet users (AIMC-Q panel) in Spain.<br><br>Panel recruited back to the General Media Study (EGM) with no coverage of the offline population. | Quasi-randomisation: propensity score weighting.<br><br>Superpopulation: model-assisted calibration.<br><br>Doubly robust combining quasi-randomisation and superpopulation. | Sociodemographics used as population totals and 1,099 auxiliary variables derived from municipality-level aggregate administrative data used as contextual variables. |

## 1.3 Sampling strategies, selection methods and inference

A double inference process characterises survey research. On the one hand, the characteristics of the individuals selected in the sample are inferred from their answers. On the other hand, the characteristics of the individuals that form the sample are the base to describe the target population (Groves *et al.* 2013). This research focuses on the second inference process, which allows us to know the population from observing a part—sample—of it.

**Probability sampling and design-based inference**

Kish's definition of survey sampling refers to "methods for selecting and observing a part (sample) of the population in order to make inferences about the whole population" (1965:18). This definition does not limit sampling to the sample selection method; it includes the link between the selection method and inference. In the same vein, Cochran states that "[Sampling theory] attempts to develop methods of sample selection and methods of estimation which will provide, at the lowest possible cost, estimators that are sufficiently accurate" (1971:30). Both definitions of sampling converge in relating the selection method to inference. In other words, how the sample is selected is critical to establish the link to the population.

Since the ultimate goal of sample selection is to make inferences to the population, the sampling strategy must integrate the method for selecting units and the estimation strategy (Hansen, Hurwitz, and Madow 1953). These two elements are connected in such a way that, for example, a random selection of population units allows inferring to the population, assuming a degree of uncertainty. In this case, the use of a given selection method—simple random sampling—makes it possible to infer to the population based on mathematical principles that establish how estimators behave when the selection procedure is repeatedly implemented. In contrast, nonprobability selection methods, such as snowball or quota sampling, require models to enable inference from the sample to the population.

Randomisation as a basis for inferring from the sample to the population has been the dominant idea in survey sampling for the last 90 years. In a probability sample, the selection of each element of the population is made based on randomness and according to a known probability of selection (Cochran 1971; Hansen *et al.* 1953; Kish 1965). Although

there are different variants within probability sampling, such as stratified or cluster sampling, all of them have in common four characteristics: 1) the possibility of defining a set of samples through the application of the sampling procedure; 2) each possible sample of the population has a known probability of being selected; 3) each element of the population has a known non-zero probability of selection and 4) the method for computing the estimate must lead to a unique estimate for any specific (Cochran 1971:31).

Randomness plays a dual role in the sampling strategy: on the one hand, it guarantees the representativeness[5] of the sample with respect to any variable measured in the survey, and, on the other hand, it allows measuring the uncertainty derived from observing only a part of the population. Statistical theory states that the selection of successive random samples from the same population allows the frequency distribution of the estimators to be calculated, meaning that the variability of the distribution can be taken into account for the inference process. The selection of a probability sample, assuming that there are no other sources of error, such as nonresponse or the absence of an adequate sampling frame, allows inferring the characteristics of the sample to the population. This approach to inference is called design-based inference (Valliant *et al.* 2018:323–25).

**Nonprobability sampling and model-based inference**

There are selection methods in which randomisation does not play a role, and the population elements are selected according to some arbitrary rules. Nonprobability sampling encompasses a series of selection methods that, unlike probability methods, do not have a theoretical basis that enables the inference to the population. These methods have in common that generally exclude a large number of population members, rely on self-selected volunteers and present high levels of nonresponse (Baker et al. 2013:94). Under the umbrella of nonprobability selection methods, a wide range of alternatives offer

[5] The concept of representativeness, although widely used, is problematic in the field of survey sampling. Kish warned that this term was used equally to describe the outcome of probability and nonprobability selection methods and recommended avoiding it (Kish 1965:26). Likewise, Kruskal and Mosteller (1979a; 1979b; 1979c, 1980) made an in-depth analysis of the term representativeness in different fields, such as science or survey sampling, identifying up to nine different meanings of the term. Bethlehem, Cobben and Schouten (2011:180–81) propose that a sample is representative of the population with respect to a variable, when the probability of response is constant for the different levels of that variable. In this thesis, the concept of representativeness is also restricted to the variable to estimate, omitting broader definitions that evoke the idea of a representation of the population.

different opportunities and risks for estimation. There are different typologies of nonprobability selection methods (e.g. Baker *et al.*, 2013; Elliott and Valliant, 2017; Kish, 1965). This paper uses the AAPOR proposal (Table 1.2) used in Elliott and Valliant (2017), which classifies these selection methods into three groups: convenience sampling, sample matching and network sampling.

Table 1.2. Nonprobability sampling methods

| Convenience sampling | Sample matching | Network sampling |
|---|---|---|
| • Mall intercept survey/intercept survey<br>• River sampling<br>• Opt-in web panels | • Quota sampling<br>• Sample matching | • Snowball sampling<br>• Respondent driven sampling |

Source: classification based on the AAPOR report on nonprobability samples (2013).

*Convenience sampling*

Convenience sampling consists in selecting a sample of elements without the intervention of any sampling plan, prioritising the ease of locating and recruiting participants. In all these methods, the selection only requires that the individual belongs to the target population. A method that has been largely used in market research, the mall intercept survey, recruits volunteers in public places or malls (Bush and Hair 1985). A modern version of this method is river sampling, a technique to recruit an online sample that relies on pop-ups, banners or advertisements on a series of web pages (Callegaro, Manfreda, and Vehovar 2015:48–51).

Opt-in panels are research infrastructures that have become popular among research agencies (Baker *et al.* 2010). Online panels consist of a database of potential participants who have declared their intention to cooperate if selected for a survey during the recruitment stage. These panels can be recruited using probability and nonprobability methods (Callegaro *et al.* 2014), which affects the estimation strategy. Opt-in panels rely on volunteers recruited using nonprobability methods such as advertisements, pop-ups, banners or registration agreements from other websites (Callegaro *et al.* 2015:207). Moreover, in these panels, the offline population is usually excluded. In addition to the aforementioned sources of bias, a recruitment system largely dependent on individuals has led to a

significant proportion of professionalised participants that affects data quality (Tourangeau *et al.* 2013). Samples within the panel are usually selected using quotas or other sample matching methods.

*Sample matching*

In sample matching, the sample selection is based on a set of relevant population characteristics related to the selection mechanism. The most widespread method of sample matching is quota sampling. In this method, a set of variables are used to control the selection process so that the resulting sample distribution matches the population with respect to those variables (Kish 1965:562–63; Smith 1983). To the extent that the variables used to generate the quotas control the selection mechanism into the sample —with respect to the target variable—, the estimates would be unbiased, enabling inference (Gittelman *et al.* 2015). Another form of sample matching is to use a statistical model to guide the selection of population units using a reference probability sample (Rivers 2007; Schonlau *et al.* 2009). In this method, elements are selected according to their affinity with cases in the probability sample using a set of auxiliary variables present in both samples.

*Network sampling*

Network sampling requires an initial sample of participants who, in turn, identify other members of the population with whom they are connected. Snowball sampling starts with a group of participants selected by the researchers, who invite their contacts to participate as long as they meet the selection requirements. This method has been used to recruit samples from rare populations where connections among the members exist (Kalton and Anderson 1986). Another method of network sampling is respondent-driven sampling, a more sophisticated version of snowball sampling in which each respondent identifies all their contacts who belong to the reference population (Heckathorn 1997; Salganik and Heckathorn 2004). If the recruitment of the initial sample allows for the calculation of selection probabilities, RDS could be a case of probability sampling. However, the conditions surrounding network sampling rarely allow for the calculation of initial selection probabilities.

*Selection bias and model-based inference*

The use of nonprobability samples to study populations dates back to the beginnings of modern sampling when Kiaer (1897) proposed his representative method for selecting a sample so that the distribution of some indicators matched the population. Despite the consensus on the merits of probability sampling, nonprobability samples continued to be the rule in some areas such as market research. In addition, in recent years, the emergence of the Internet and new data collection techniques has boosted the expansion of these selection methods (Callegaro *et al.* 2015). In 2013, AAPOR published a report on the opportunities and limitations of nonprobability samples (Baker *et al.* 2013). The report acknowledges the ability of these selection methods to accelerate data collection and reduce costs. However, the major drawback of such methods relates to the second element of the sampling strategy, the estimation method. The lack of control over sample selection makes inference difficult if not impossible because, unlike in probability sampling, there is no theoretical framework to support it.

The main drawback of nonprobability methods is that the lack of control over the selection process can introduce selection bias in the estimates (Elliott and Valliant 2017). Selection bias occurs when the units included in the sample differ from those excluded with respect to the variable to be estimated. Coverage error, for example, might be at the origin of selection bias (Valliant and Dever 2011). This error refers to the fact that some population elements cannot participate in the survey either because they are not included in the sampling frame or because they do not have access to the technology that enables responding to the survey (Weiseberg 2005). The recruitment process for a general population survey from a web opt-in panel can illustrate how this bias works. First, although the target population is all residents in the country, the recruitment is dependent on having internet access and visiting the web pages where the panel is advertised. The fact that a part of the population has no chance of joining the panel is coverage error, one of the sources of selection bias. Secondly, only a part of those who can enter the panel will do so, increasing the magnitude of the selection bias if they are systematically different from those who did not join. Finally, when selected for a specific survey, the panel members can refuse to cooperate with the study, adding another source of bias to the selection process. Non response or self-selection can also be the source of selection bias in selection processes.

However, an issue with nonprobability sampling is that most of the time there is not enough information to dissentagle the contribution of each part of the selection process to the selection bias.

Despite the disadvantages that the use of nonprobability selection methods may entail, survey inference is not an exclusive capability of probability samples. In this respect, Kish stated that "probability sampling is not a dogma, but a strategy, especially for large numbers" (1965:29). Under certain assumptions, a probability sample is not necessary for inference. In a recent paper, Cornesse and her colleagues (2020:8) point to four types of justifications for inference from nonprobability samples. First, some research questions deal with phenomena widespread throughout the population so that inference would be appropriate regardless of the type of sampling employed. Second, the nonprobability sample design manages to control for possible biases. For example, quotas that control the selection process can minimise or eliminate bias in the estimates. Third, biases produced by selection using nonprobability methods are controlled using analytical methods after data collection. Fourth, the combination of the sampling method and the production of statistical adjustments will produce accurate population estimates. These four assumptions under which inference can be made from nonprobability samples have in common that they require a model that rules out the impact of selection bias. The sampling design is ignored in model-based inference, which only considers the population structure (Little 2004; Valliant *et al.* 2018).

**Model-assisted inference**

A third hybrid type of inference, model-assisted inference, has emerged between the design-based and model-based approaches (Särndal, Swensson, and Wretman 1992). This inference method requires a probability sample but deviates from the design-based inference in that some of the basic assumptions of the latter are not met. Design-based inference assumes a perfect sampling frame that covers all units in the population and that all selected units respond to the survey. However, in the last decades, response rates have been steadily falling in addition to increasing coverage problems (de Leeuw, Hox, and Luiten 2018). In model-assisted inference, apart from starting from a probability sample, models are used to avoid biases caused by nonresponse and noncoverage. Such models

take, in most cases, the form of nonresponse weights or calibration, which rebalance the sample using auxiliary variables related to the probability of response and the target variable (Bethlehem *et al.* 2011; Särndal 2010).

**Survey inference through five discussions**

Certain events, discussions and publications can explain the development of sampling theory and practice. The following pages describe some of the major discussions and events that have shaped sampling strategies and inference evolution. In addition to providing a historical context for the thesis, these discussions show some common threads. First, from the beginning of the development of modern sampling, there has been a tension between design-based inference and model-based inference. Second, the development of sampling strategies has occurred in parallel in the area of official statistics, academia, and public opinion research. Despite the interconnections between these fields, there have been significant differences marked by the demands of information and the resources available, which, for example, have led pollsters to explore inference from nonprobability samples. Finally, in line with other methodological advances, the development of sampling strategies has occurred as a consequence of changes in the information needs at each historical moment and technological development.

*The beginnings of modern sampling: representativeness and probability*

The idea of describing population characteristics from a sample has been around for centuries. As early as 1662, John Graunt attempted to estimate the population of London from what we would call a sample. Later, in 1802, Laplace estimated the population of France from a selection of administrative districts (Brewer 2013). However, the origin of modern sample surveys is attributed to Kiaer's (1897) *representative method*, which he developed from his position as director of Statistics Norway. Kiaer advocated the possibility of knowing a population from what he called a "partial investigation" in which some population elements were selected in such a way that they mirrored the population. To this end, he proposed arbitrarily selecting population units to achieve the desired representativeness with respect to certain variables. Kiaer applied his method to several surveys in Norway during the second half of the nineteenth century.

The *representative method* was widely contested by his colleagues, mainly for two reasons. The first was a reactionary argument. During the nineteenth century, the dominating idea was that the census, an enumeration of the entire population, was the only suitable method to know the population. In this scenario, the idea that the study of a sample could lead to knowing the population characteristics was not well understood. The second argument was more elaborated and alluded to the lack of rigour of a purposive selection method in which the judgement of the researcher played a vital role (Lie 2002). This second criticism, the lack of a theoretical underpinning to the inference process, would be wielded to this day to dismiss or at least question the use of nonprobability sampling methods.

The representative method, although criticised, opened the door to the development of the incipient idea of inference from a sample. At the International Statistical Institute meeting in Berlin in 1903, a new discussion about the feasibility of the representative method and partial investigations allowed Lucien March to present his idea about the use of randomness for sample selection in order to produce population estimates (Kruskal and Mosteller 1980). After this discussion, Bowley (1906, 1913) developed the idea of random selection and implemented it to select a sample in the English town of Reading. In 1925 the International Statistical Institute accepted sampling as a valid method for the study of populations. However, there was no recommendation on the best method—probability or purposive sampling—for making inference at that early stage, although in the report, there was a reference to the desirability of using randomisation (Brewer 2013). This decision marked the beginning of a dispute over the most appropriate method for selecting samples and inferring to the population that continued for the next decade. Despite having its centre in official statistics and academia (Bowley 1926), this debate also developed in the emerging polling industry, especially in the United States. The debate about sampling methods and inference experienced a turning point in the 1930s, beginning the hegemony of probability sampling and design-based inference.

*The triumph of design-based inference: Neyman versus Gini and Galvani*

The early 1930s saw a growing acceptance of using samples to study a population, although the debate persisted about the suitability of purposive—nonprobability—and probability methods for selection. At this time, two events marked the beginning of the

hegemony of probability sampling. The first episode, which took place in the realm of official statistics and academia, was Neyman's critique (Neyman 1934) of Gini and Galvani's attempt in 1929 to select a sample of responses from the Italian census using a purposive selection method. The second episode occurred in the field of public opinion research when the Literary Digest Poll failed to estimate the vote shares in the 1936 presidential election in the United States.

At the end of the 1920s, the Italian census office needed to free up space to store the forms for the next census edition scheduled for 1931. Gini and Galvani decided to select a sample of the forms using a purposive selection method to preserve some of the information from the previous census for future research. They selected the forms corresponding to 29 of the 214 administrative districts of the country in such a way that they reflected the profile of the Italian population with respect to seven variables (Neyman 1952). The researchers found that, despite the coincidence between the sample and the population with respect to the variables used for the selection, there were considerable deviations in some of the variables not used in the sampling design. At the time, the authors' conclusion, far from embracing the probability method, was to question the possibility of using samples to study the population (Gini 1928; Gini and Galvani 1929).

The milestone that marks the beginning of the hegemony of probability sampling and design-based inference is the publication of Neyman's (1934) article, in which he analyses the sampling procedure and conclusions from Gini and Galvani sampling. In his review, Neyman argues that the problem does not lie in using samples for inference but in the inadequacy of using purposive selection methods. In addition, he criticises the model—the seven variables—used to select districts as inappropriate and unrealistic, which biased the estimates. In contrast to purposive sampling, the probability method allows using the framework of probability theory to interpret the estimates without using a model while provides an alternative way to quantify the precision of the estimates by calculating standard errors and confidence intervals. In his words, "the only method which can be advised for general use is the method of stratified random sampling" (Neyman 1934:588). This demonstration was a turning point in the use of purposive sampling, especially in official statistics (Kruskal and Mosteller 1980). Proof of this is that the first generation of sampling

handbooks focuses exclusively on probability selection methods and design-based inference (Cochram, 1953; Deming, 1950; Yates, 1949).

*Sample representativeness in public opinion research*

The 1930s also saw the emergence of the polling industry in the United States. Two names stood out, Gallup and Roper, who founded two companies that pioneered surveys and data analysis to respond to emerging information demands. The demand for information mainly came from the service sector, which wanted to understand consumers behaviour, and the journalism, which systematically started to use data to support their news (Converse 2009). The relationship between journalism and data had already begun to develop in the preceding decades with the attempt of using some primitive techniques to collect public opinion data (Groves 2011). For instance, straw polls, an open consultation to gather insight about the popular opinion on a specific matter, were used to predict the election outcome. This technique was used by the Literary Digest, a weekly magazine, to predict the winner of the 1936 US presidential election. This prediction turned out wrong and helped to consolidate the nascent polling industry, which also relied on nonprobability selection methods at that time.

The Literary Digest straw polls, though unscientific, were backed by the success of its predictions of presidential elections during the 1920s (Squire 1988). In the months leading up to the 1936 election that pitted Democratic President Roosevelt against Republican Landon, the Literary Digest mailed 10 million postcards to a sample selected from automobile registration and telephone listings. The magazine received more than 2.3 million ballots that were used to predict the victory for the Republican candidate, with 55% of the support to President Roosevelt's 41%. Roosevelt won 61% of the vote on the election day, more than 20 percentage points ahead of his opponent. Although different hypotheses have been proposed, an incomplete sample frame and a higher propensity of Republican voters to return the ballot are the main causes of the mismatch between the prediction and the outcome of the election (Lohr and Brick 2017; Lusinchi 2012). In this election, new polling firms, such as Gallup, which claimed to use a scientific method for selecting the sample, were much more accurate in estimating the outcome.

Although the failure of the Literary Digest has been a landmark in invalidating inference from nonprobability samples, the Gallup and Roper polls at the time also used selection methods that, although more sophisticated, were not probability-based[6] such as quota sampling. In the spirit of Kiaer's representative method, they intended to create a microcosm of the United States with respect to the variables sex, age, and occupation (Berinsky, 2006). Each interviewer had a list of profiles to track in a given geographic location to complete the quotas, having a great degree of intervention in sample selection. However, at the time, the use of quotas was advocated by the country most influential pollsters to ensure the quality of the sample and reduce data collection time (Gallup 1944).

It took another polling failure in 1948 for pollsters to consider abandoning quota sampling and replacing it with probability methods. In that presidential election, the three main organisations dedicated to analysing the public opinion—Gallup, Roper and Crossley—estimated a victory for the Republican Dewey over the Democrat Truman, which the election results refuted. This time, as the polls had gained popularity, the Social Science Research Council undertook an independent investigation of the causes of this mismatch. The report revealed that one of the main issues was the use of quota sampling[7] that, excluding some profiles, affected the sample composition (Mosteller *et al.* 1949). Twelve years later, the unrepresentativeness of the sample that had led to the failure of the Literary Digest Poll was also partly responsible for the failure of polls that claimed to employ a more sophisticated selection methodology[8]. The 1948 polling failure led most pollsters to revise their selection methods during the 1950s (Berinsky 2006).

After 1948 the use of probability sampling was hardly contested, although nonprobability methods have continued to be used in areas such as market research. Also, in the

[6] Gallup, Roper and Crossley used the Literary Digest failure to construct a rhetoric that presented their methods as scientific, establishing a non-existent link to probability sampling (Lusinchi 2017).

[7] In addition, the report notes that the fieldwork was not able to detect last-minute movements in the vote; information on respondents' willingness to vote was not taken into account, and the omission of those who did not reveal their electoral preferences in the estimation (Moon 1999).

[8] The lack of representativeness of samples has been a recurrent cause of pre-election polling failures. This lack of representativeness has been related to the use of quotas in sampling (Callegaro and Gasperoni 2008; Jowell *et al.* 1993). Recently, the investigation following the failure of polls in the 2015 United Kingdom general election showed that the most plausible cause of the widespread deviation of the voting estimates was the use of quota samples that were not representative of the population (Sturgis *et al.* 2016).

1950s and '60s, in academia and official statistics, some advances related to probability samples anticipated the later development of model-based inference. In the 1950s, some statisticians developed estimators based on auxiliary information[9] to connect the sample and the population. These are the cases of the ratio estimator and the regression estimator (Cochran 1971). Both estimators[10] were based on a simple model that did not fit into the framework of design-based inference framework (Särndal 1978). In addition, another notable development was Sudman's proposal to combine probability area sampling with random walks and quotas, a system adopting aspects of probability and nonprobability sampling, which had some impact in Europe (Sudman 1966). For instance, the Spanish *Centro de Investigaciones Sociológicas* (CIS), which carries out most surveys face-to-face, has been using a version of this method to design its samples for decades. Some evidence has shown that the combination of probability selection methods and quotas can lead to more accurate estimates of some non-quota variables such as education level compared to probability samples (Díaz de Rada and Martínez 2020).

*Model-based inference and superpopulation models*

During the 1970s, new research contributed to the strength of the model-based inference approach along two lines. First, some work questioned inference from probability samples (Valliant *et al.* 2018:326). The underpinning theory to probability sampling and design-based inference states that repeated samples from the same population form a distribution that allows uncertainty to be estimated, providing a framework for inference. However, some work has explored scenarios where such a deviation between samples exists that the design-based inference framework would not allow to produce accurate population estimates (Smith 1976, 1979, 1983). Second, some new techniques extended the inference framework for nonprobability samples, as is the case of superpopulation models[11] (Royall 1970). Superpopulation models, which are further explained in the next

[9] The term auxiliary information or auxiliary variable refers to information available for the whole population, those who are part of the sample or respond to the survey and those who are not. A more detailed description of the role of auxiliary information in estimation can be found in the last section of this chapter.

[10] The ratio estimator is defined as $t_R = \bar{x}\bar{y}/\bar{x}_s$ where $\bar{y}_s$ is the mean of the variable to be estimated in the sample $s$, $\bar{x}$ is the mean of the auxiliary variable in the population, and $\bar{x}_s$ in sample. In the regression estimator $t_{REG} = \bar{y}_s + \beta(\bar{x} - \bar{x}_s)$ (Cochran 1971).

[11] Superpopulation models assume the existence of a set of auxiliary variables, available for those selected in the sample and the rest of the population, that are related to the probability of being part of the

section, are based on the idea of modelling the population structure regardless of the method used to select the sample. In addition, Särndal (1978) compared model-based and design-based inference approaches, concluding that the use of models can produce similar results compared to the design-based inference and that, under certain circumstances, they could offer advantages over classical inference.

In parallel to these two developments, the decline in response rates started, and concerns about design-based inference emerged (Koop 1974). Nonresponse could bias the estimates even though the sample design was based on a random selection of cases. Therefore, despite the decisive weight of the design, inference began to rely partially on models to adjust for possible deviations. In this context, Rubin and Little (Little 1982; Little and Rubin 1987; Rubin 1976) presented their work on inference in the presence of selection bias and nonresponse. This research outlines three mechanisms to explain the effect of missing data: MCAR (missing completely at random), MAR (missing at random) and NMAR (not missing at random). Under MCAR, the estimate of the target variable is not biased despite nonresponse, as the selection mechanism acts entirely at random. In NMAR, the opposite is true; participation in the survey depends directly on the target variable. In between these two mechanisms is MAR, where the selection is independent of the target variable and dependent on a set of auxiliary variables. In a MAR scenario, the auxiliary variables can be used to adjust the sample and eliminate the bias from the estimates.

*Web surveys and inference*

The emergence of web surveys has marked the most recent episode in the debate on inference from nonprobability samples. Although the first cases of surveys distributed over the web date back to the 1980s, the spread of the Internet and the development of Web 2.0 helped to expand the use of web surveys (Díaz de Rada *et al.* 2019). This expansion was due to the development of technology, the growing need for information by

final sample and to the target variable. These auxiliary variables are used, first, to fit a model in the sample that predicts the variable of interest. Second, to predict the variable of interest in the rest of the population. The final estimator is composed of the information collected in the survey and the predictions made for the rest of the population (Valliant *et al.* 2000).

organisations[12], and the exponential increase in internet coverage. Despite the advantages offered by web surveys, especially in terms of costs and time needed for data collection, the lack of coverage affecting a part of the population and the absence of adequate sampling frames difficult the selection of probability samples, casting doubts on the inference process (Tourangeau *et al.* 2013).

In the early 2000s, with the expansion of web surveys, several research streams emerged based on the previous work on inference from nonprobability samples and model-assisted inference. Most of the early work was intended to document the experiences of organisations, mainly polling companies, that had begun to generate a systematic methodology for conducting web surveys. Later, some research began to focus on two areas related to the inference from nonprobability samples. First, given the high volume of nonprobability surveys, there have been some efforts to develop a theoretical framework to underpin the process of inference from these samples (e.g. Elliott and Valliant 2017; Mercer 2018). Second, there has been a proliferation of research on statistical methods to produce unbiased survey estimates based on the idea of model-based inference (e.g. Chen, Valliant, and Elliott 2018; Ferri-García and Mar Rueda 2020; Rafei et al. 2020; Wiśniowski et al. 2020).

**Developments in the theory of inference from nonprobability samples**

The widespread use of nonprobability selection methods for conducting web surveys has stimulated the theoretical discussion about inference from nonprobability samples. One of the most recurrent criticisms of inference from nonprobability samples is that it lacks a theoretical framework to support the process and that each target variable requires the specification of a model (Baker *et al.* 2013:103). In response to this criticism, there have been some attempts to create a consistent framework for model-based inference. The following paragraphs present the contributions of Elliott and Valliant (2017), Mercer (2018; Mercer et al. 2017) and the concept of *fit for purpose*.

Elliott and Valliant (2017) provide a conceptual framework for inference from nonprobability samples that organises estimation methods to produce unbiased estimates

[12] According to ESOMAR (2017) the market and public opinion research sector turnover from conducting research using online methodologies accounted for 26% of the total in 2011 and grew to 44% in 2016.

with a measure of error. They propose a general framework for inference from nonprobability samples that relates the work of Smith (1983) on the inference from nonprobability samples to the contributions of Rubin (1976) and Little (1982) about the selection mechanisms and the effects of nonresponse on the estimates. Estimation of the variable of interest is possible whenever the selection process is MCAR—there is no selection—or MAR—the selection mechanism can be controlled using auxiliary variables.

Two inference strategies, quasi-randomisation and superpopulation, are derived from this conceptual framework. Although the common basis is that a model is needed to control selection bias, each strategy approaches the problem differently. In quasi-randomisation, the pseudo-probability of selection is modelled, assuming that it is related to a set of auxiliary variables observed for all cases (Elliott and Valliant 2017:255–57). This strategy includes the propensity score weighting and propensity score matching techniques discussed in the next section. The result of this strategy is the calculation of a pseudo-selection weight used to estimate all survey variables. The superpopulation strategy, instead of modelling the pseudo-probability of selection, focuses on predicting the target variable by using the auxiliary information available for all elements of the population (Elliott and Valliant 2017:257–61). In this approach, the estimation of each survey variable requires a superpopulation model. Another relevant contribution of this framework is that it enables the calculation of the variance of the estimators. A criticism of inference from nonprobability samples is that there is no theory to support the calculation of a measure of uncertainty that informs about the precision of the estimate. In the quasi-randomisation framework, they propose to use resampling methods[13] such as bootstrap or jackknife to incorporate sample variability and the variability due to the computation of the pseudo-weights (Elliott and Valliant 2017:257). For superpopulation models, the authors recommend using the jackknife method or the variance sandwich estimator (Valliant, Dorfman, and Royall 2000).

Mercer makes a twofold contribution to the theory of inference from nonprobability samples. First, he reflects on why the framework used in survey methodology to analyse the quality of estimates, the Total Survey Error (TSE), is not suitable for dealing with

[13] Resampling techniques such as Bootstrap or jackknife consist of selecting sub-samples and calculating the estimates in order to find measures of variability of the estimators (Efron 1982).

inference from nonprobability samples. Second, he proposes using concepts from causal inference analysis to enable inference from nonprobability samples (Mercer *et al.* 2017). The TSE framework is a systematic way of addressing the quality of survey estimates (Biemer 2010; Biemer and Lyberg 2003). It divides the sources of error into two groups, measurement and representativeness. The representativeness side of TSE covers coverage error, sampling error, nonresponse error and errors arising from adjustments made after data collection. Each of these errors has, in turn, two components, one random, the variance, and one systematic, the bias. In his PhD thesis, which focuses on inference from online opt-in panels, Mercer presents three arguments to rule out the use of this theoretical framework when dealing with nonprobability samples. First, there is no concern or aspiration to fully cover the population in the nonprobability sample framework, as per the TSE standard (2018:8). Instead, in the context of opt-in panels, the goal is to have a representation of all relevant profiles of the population so the estimation can account for the differences across these profiles. Second, the linear process drawn by the TSE does not apply to the selection methods of many nonprobability samples (2018:9). The clearest example is that the TSE does not consider selection bias as such but merely includes coverage and nonresponse errors. Finally, TSE does not consider the possibility that inference can be model-based; it was designed for design-based inference (2018:10). In recent years other researchers have supported this proposal which rules out the use of the TSE as a framework to analyse inference from nonprobability samples (Cornesse *et al.* 2020).

In addition, Mercer and his colleagues (2017) presented an alternative theoretical framework based on the causal inference analysis (Rubin 1974). The authors draw a parallel between the role of randomisation in estimating causal effects and sample selection in surveys. They establish that inference from nonprobability samples, as the estimation of treatment effects with observational data, relies on three assumptions: exchangeability, positivity and comparability of the composition of the treatment and control groups. First, exchangeability refers to that all covariates involved in the selection process are measured for all sample elements (Rosenbaum and Rubin 1983). Second, positivity requires that all relevant groups in the population are represented in the sample. The set of variables that explain the selection mechanism determine the relevant groups that must be present in the sample. Finally, the sample composition must be aligned with the population with respect

to the variables used to control for selection. Otherwise, it would be necessary to balance the sample through survey weights.

Finally, it is essential to highlight the development of the idea of *fit for purpose* applied to the inference from nonprobability samples. Based on the TSE scheme, the quality of survey estimates has been measured in terms of bias and variance. However, some voices have called for a data quality paradigm that considers cost, and more importantly, the purpose for which the data was collected (Baker *et al.* 2013:98). Several national statistical offices have been working on a flexible framework for survey design that considers the purpose of the information rather than obtaining the highest quality in any case (Statistics Canada 2017). The expansion of this framework could also positively impact the costs and time required to conduct the research (Kohler, Kreuter, and Stuart 2019). Nonprobability surveys fit into this framework as long as the quality of the estimates allows reaching a conclusion that fulfils the need for information. In this respect, some works have shown that estimates from bivariate (Pasek 2016) and multivariate (Dassonneville *et al.* 2020) analyses are similar in nonprobability surveys and probability surveys.

## 1.4 Adjustments for selection bias

The adjustment methods for survey estimates are not new developments. They existed already to correct coverage and nonresponse bias in probability surveys and were implemented in nonprobability surveys such as political polls. Before 2000, Vehovar and his colleagues (1999) already wondered whether the use of weights would enable inference from a nonprobability web survey. Table 1.3 presents adjustment methods according to whether they model the selection into the sample or the target variable. The idea behind using these techniques is to model the selection mechanism using auxiliary variables to minimise the selection bias of the survey estimates. These models need to be correctly specified with auxiliary variables correlated with the probability of selection and the target variable. There are two types of adjustments, the global adjustments that model the selection in the sample and outcome-specific adjustments focused on the target variable (Cornesse *et al.* 2020).

Table 1.3. Methods for adjusting estimates from nonprobability samples

| Global adjustments | Outcome-specific adjustments |
|---|---|
| • Calibration and poststratification | • Superpopulation models |
| • Propensity score weighting | • Model-assisted calibration |
| • Sample matching | • Multilevel regression and poststratification |

Source: adapted from Cornesse *et al.* (2020).

### Global adjustments

Calibration and poststratification were originally designed to adjust the estimates from probability samples for coverage and nonresponse bias (Deville and Särndal 1992; Särndal and Lundström 2005). Calibration weighting consists of a model that aims to produce a set of weights that forces the sample to match the population totals with respect to the auxiliary variables (Kott 2006). An advanced form of calibration is poststratification, in which the sample is split into mutually exclusive groups formed by the interaction of all auxiliary variables (Zhang 2000). These models are able to remove the bias from the estimates if the groups formed by the variables used to explain the selection mechanism are representative of the population regarding the target variable. This type of adjustment has been widely used to correct estimates from nonprobability samples (e.g. Dever, Rafferty,

and Valliant 2008; Loosveldt and Sonck 2008; Taylor 2000; Terhanian *et al.* 2000). Finally, another application of this technique consists in combining a probability and a nonprobability sample and using calibration to adjust the joint sample (Disogra *et al.* 2011; Fahimi *et al.* 2015).

Propensity score weighting involves using a probability sample reference survey with the nonprobability sample to fit a model that predicts the selection into the nonprobability sample (Valliant and Dever 2011). The regression model used to predict the probability of selection uses a set of auxiliary variables measured identically in both surveys. Finally, the weight that adjusts the nonprobability sample is the inverse of the predicted probability of selection. Propensity score models were developed in the framework of causal inference theory to control for the probability of being assigned to an experimental group in observational studies (Rosenbaum and Rubin 1983) and have also been used in probability surveys to adjust for nonresponse bias (Bethlehem *et al.* 2011:8). Such adjustments have been widely used since the beginning of the web survey to improve the quality of the estimates (e.g. Börsch-Supan *et al.* 2004; Duffy *et al.* 2005; Lee 2006; Lee and Valliant 2009; Pedraza *et al.* 2010; Schonlau, van Soest, and Kapteyn 2007; Terhanian *et al.* 2000). For this weighting method to effectively minimise bias, the auxiliary variables must be related to the probability of selection and the target variable. A similar approach is to use the probability and nonprobability samples together for estimation after computing the weight for the nonprobability sample (Elliott 2009).

Sample matching relies on a statistical model to select a sample using a reference probability sample (Bethlehem 2016; Rivers 2007; Vavreck and Rivers 2008). The objective of this method is that the resulting sample mirrors the probability sample with respect to a set of auxiliary variables that explain the selection into the sample. To this end, the nonprobability sample is selected from a large set of cases recruited using nonprobability methods, such as an opt-in web panel, so each case in the reference sample is matched to one case in the nonprobability sample. The probability sample only acts as a reference survey and is discarded after the selection.

**Outcome-specific adjustments**

The second group of adjustments focuses on modelling the selection mechanism with respect to a specific target variable. A drawback of these methods is that each estimate requires fitting a model (Elliott and Valliant 2017). The first method is the superpopulation model. The idea underlying the superpopulation model is that a common model explains the data-generating process in the nonprobability sample and the rest of the population and that this model can be specified using auxiliary information. A statistical model is fitted in the nonprobability sample to predict the target variable for the nonsampled part of the population. At the estimation stage, a joint estimator combines the observed nonprobability sample and the predicted outcome for the rest of the population (Valliant *et al.* 2000). This technique has been used to adjust nonprobability surveys in different domains, such as election polling (Pavía 2005; Pavía and Larraz 2012).

Model-assisted calibration can also be used in the framework of superpopulation models. This method involves generating specific weights for a variable of interest using a model fitted with a set of auxiliary variables for which population totals are known (Wu and Sitter 2001). This method was recently employed by Chen *et al.* (2018; Chen, Valliant, and Elliott 2019) using a LASSO model, a machine learning analysis technique useful in scenarios with a large number of predictors. This method is adapted to adjust a nonprobability sample using aggregate administrative data as auxiliary variables in the fourth chapter of the thesis.

A method used in political polling to adjust survey estimates is multilevel regression with poststratification (MRP) (Gelman 2007; Park, Gelman, and Bafumi 2004). This method uses a multilevel model to predict the outcome variable in a nonprobability sample using a set of auxiliary variables for which the population totals are known. Predictions are made for the cell resulting from combining the categories of the auxiliary variables, the poststrata. Finally, these predictions are weighted using poststratification to calculate the estimator. An advantage of these models is that they allow subgroup estimates, even if there are relatively few units of the group in the sample. An example of the use of this methodology was the estimation of the voting intention in the US 2012 presidential election using a large sample of Xbox users (Wang *et al.* 2015). Despite the significant departures

of the sample from the population with respect to the target variable—voting intention—and a set of sociodemographics, the MRP model successfully removed the bias from the estimates. The final estimate from the model was closer to the actual election result than the polling average. Again, this technique to be effective requires the correct specification of the selection mechanism with respect to the target variable in the multilevel model (Buttice and Highton 2013).

After presenting the different adjustment options for nonprobability samples, two questions arise. The first is whether, after implementing the adjustments, the estimates from nonprobability samples are comparable to those from probability samples. In this respect, Cornesse *et al.* (2020), after reviewing studies that have compared estimates from probability and nonprobability samples to population benchmarks, concluded that probability samples produce less biased estimates. In addition, they also concluded that, while adjustments do reduce bias to some extent, they are not effective in eliminating it. For example, Yeager and his colleagues (2011) compare telephone and web surveys estimates based on probability and nonprobability samples against population benchmarks. The result showed that estimates from probability samples were less biased and that the effect of poststratification on estimates from the nonprobability sample did not always reduce the bias from the estimates.

The second question concerns which adjustment method is more effective in minimising bias. To answer this question, some researchers have used statistical simulations to compare adjustment methods. For example, Valliant (2019) compared estimates from propensity score weighting, superpopulation modelling, a doubly robust approach—the combination of propensity score and superpopulation—and MRP. Although the result was inconclusive, combining the propensity score weighting model with the calibration, the doubly robust approach, achieved the most significant reduction in bias. Also, Ferri-García and Rueda (2018) compared calibration and propensity score weighting effectiveness in reducing selection bias in different missing data scenarios. The result shows the superiority of propensity score weighting or the combination of this technique and calibration over the use of calibration alone, especially when the missing data mechanism is NMAR.

## 1.5 Auxiliary data in model-based and model-assisted estimation

All the methods listed in the previous section that seek to adjust the estimates for selection bias have in common that require auxiliary variables capable of explaining both the selection mechanism and the outcome of the analysis. Along with the selected adjustment methods, the model specification plays a fundamental role in minimising the incidence of selection bias. In this respect, Mercer, Lau and Kennedy (2018) compared MRP and propensity score weighting to adjust the estimates from a nonprobability panel using different sets of auxiliary variables. They concluded that the model specification is more relevant than the method used in removing selection bias from the estimates.

**Use of auxiliary variables to adjust surveys**

The selection of auxiliary variables has received little attention compared to estimation and adjustment methods. These variables have historically been limited to the census and other administrative data, imposing an important constraint to the specification of the nonresponse models in probability surveys. The limited availability of auxiliary variables is also behind the lack of empirical evidence about the role and selection of these variables in the weighting models. The main contributions in this field come from the area of probability sample adjustments and official statistics (Bethlehem *et al.* 2011; Kreuter and Olson 2011; Särndal and Lundström 2005, 2008; Schouten 2007).

Auxiliary variables are available for all population members, those who participate in the survey, and those not selected. These variables are required to fit any model that aims to address biases arising from sample selection and nonresponse. Auxiliary variables can be available for all population units, available only for the sample—respondent and nonrespondent—members, or available on an aggregated population level (Bethlehem *et al.* 2011:247–48). This classification arises from the traditional data sources used to adjust probability surveys for nonresponse and coverage errors. The first type of auxiliary information, population level, is available in the sample frame. The second type, sample level, mainly covers paradata[14] such as interviewer observations in face-to-face surveys. The third type, aggregated population level information, refers to population totals or means,

[14] Paradata are by-products generated during the survey fieldwork. For example, the number of calls before getting the interview or the time used to complete the questionnaire (Kreuter 2013).

such as the number of men and women in the population. This thesis explores a fourth type of auxiliary data, geographically aggregate data used as contextual variables. This type of auxiliary data differentiates from aggregated population level in that the former is aggregated at a lower level, for example, census tract or municipality, giving information about where the sample elements live. Table 1.4 shows a summary of the four types of auxiliary variables.

Table 1.4. Comparison of auxiliary data for survey adjustments

| | **Population level** | **Sample level** | **Aggregated population level[15]** | **Aggregated at a lower level/ contextual variables** |
|---|---|---|---|---|
| **Definition** | Information available for all population members at the individual level. | Information available only for members selected in the sample, usually collected during the fieldwork. | Population totals or means for the overall population/subgroups. | Population totals or means aggregated at a lower geographical level (i.e. census tract or municipality in a general population survey). |
| **Examples** | Information on income tax at the individual level; population register data at individual level. | Interviewer observations collected during fieldwork. | Population totals for age and sex groups from the census; election results. | Unemployment rate in the municipality where respondent lives; election results in the census tract. |
| **Adjustment method** | Propensity score and other individual-level models. Also possible to use calibration/poststratification if aggregate the information to obtain population totals. | Propensity score and other individual-level models. | Calibration and poststratification. | Propensity score and other individual-level models. Calibration and poststratification. |
| **Use in probability surveys** | Yes. Especially in official statistics, where they can match sample/respondent data to administrative records. | Usually face-to-face that allow to collect data for respondents and nonrespondents. | Yes. | An application using the National Comorbidity Survey Replication (Biemer and Peytchev 2013). |
| **Use in nonprobability surveys** | No. This data is seldom accessed from outside official statistics for privacy and data security reasons (general population records)—instead uses probability reference survey. | No. Instead use of probability reference surveys. | Yes. Used given the availability of population totals in contrast to individual-level records. | No application reported. |

The type of auxiliary data available—population level, sample level or aggregated population level—condition the adjustment methods that can be used. For instance, in the framework of probability surveys, the use of propensity score weighting requires population or sample level information (Bethlehem *et al.* 2011:8). In contrast, aggregated

[15] Throughout this thesis aggregated population level auxiliary data is also referred as individual-level data. This is because the use of this type of auxiliary information requires to measure the variable for the respondent sample. For instance, if the population total for age groups is intended to be used in a calibration, the research team need to know the age of each respondent—at the individual level.

population level data allows for calibration or poststratification adjustments (Särndal 2007). Regardless of the adjustment method, auxiliary variables reduce the bias of the estimates if correlated with the probability of being in the sample and the outcome variable. Moreover, there is a third condition to consider: the auxiliary variable must be correlated with the domains of the survey, which are the subpopulations used for estimation (Särndal and Lundström 2005:110). An incorrect model specification would be useless to adjust the estimates, and it may negatively affect the magnitude of the standard errors (Little and Vartivarian 2005). Therefore, the search for auxiliary variables relies on a theoretical analysis of the factors capable of explaining the selection mechanism regarding the target variable.

This logic, developed in the framework of probability surveys, is transferable to the inference from nonprobability surveys for the most part. Most of the adjustment methods used to correct nonprobability samples were first used to adjust probability samples. Nevertheless, there are some notable differences. First, in nonprobability surveys, the use of reference probability surveys instead of population or sample level data (e.g. propensity score weighting or propensity score matching) is widespread. Since the variables in a survey are more varied, covering a wider range of topics than those from population registers or other official sources, the model specification can improve the effectiveness of the adjustment. In the case of aggregate variables, it has also been proposed to use probability surveys as benchmarks to calculate population totals and thus extend the range of auxiliary variables available (Ferri-García and Rueda 2018).

The second difference is that the nature of the selection mechanism to be modelled usually differs between probability surveys—noncoverage and nonresponse—and nonprobability surveys—selection bias. This difference is evident when analysing the response process in an opt-in panel survey. In this case, the selection mechanism must consider the population that does not have internet access and account for the individual's decision to volunteer in the panel and respond to the survey invitation. This selection mechanism has different features than those covered by nonresponse and noncoverage in probability surveys.

The emergence of web surveys and the more frequent use of nonprobability samples has also contributed to fostering the research about the auxiliary variables that can best adjust the survey estimates. The developers of one of the first online volunteer panels in the United States, Harris Polls, coined the term webographics to refer to a series of attitudinal variables included in their adjustment models (Terhanian *et al.* 2000). Although some work has shown the potential of these variables to discriminate between the online and offline populations (Schonlau *et al.* 2004, 2007), their limited use in practice suggests that their ability to reduce bias is often insignificant (Schonlau and Couper 2017).

Another relevant aspect is the method used to select auxiliary variables included in the adjustment model. The selection of auxiliary variables has not been a significant issue in the probability sample world due to the few available variables, mainly from the census and other population registers. These variables have traditionally been selected following experience or best practices (Bethlehem *et al.* 2011:248). However, there have been some proposals to systematise this selection. Särndal and Lundström (2005:117–22) propose calculating two indicators, IND1 and IND2, to measure the contribution of each variable to explain the response mechanism and the relationship of the auxiliary variables with the target variable. Both indicators should be assessed jointly for each possible combination of auxiliary variables. Schouten (2007) developed a methodology for selecting auxiliary variables based on a measure of the absolute bias of the estimate. More recently, machine learning techniques, such as LASSO, have been used to select variables in a scenario of a high number of auxiliary variables (Chen *et al.* 2018). In the fourth chapter of this thesis, this methodology is used to select auxiliary variables among a large number of predictors.

**New auxiliary variables**

The expansion of the Internet and the possibility of generating, processing and storing large volumes of data, the so-called big data phenomenon, opens up new opportunities to improve the specification of the adjustment models (Baker 2017; Couper 2013). In recent years, various data sources have been tested as auxiliary variables to reduce coverage and nonresponse biases in probability samples. Among the new data sources that have been considered to adjust estimates are individual-level commercial data (Disogra *et al.* 2011; Pasek *et al.* 2014; West *et al.* 2015), paradata (Kreuter *et al.* 2010; Wagner *et al.* 2014;

West 2013) or georeferenced data (Butt, Lahtinen, and Fitzgerald 2015). Administrative data is another source of data, which has also been tested to adjust estimates.

*Aggregated administrative data*

Administrative data are products or by-products generated from the interaction of the public administration with citizens, businesses or other organisations. This data aims to organise, manage, or monitor services and public policies (Playford *et al.* 2016; Woollard 2014) and can also be used for research purposes. For survey adjustments, administrative data has the advantage that they often cover the entire population. Despite this advantage, access to administrative data at the population level, available for all population members, is limited to public bodies mainly for privacy and data security reasons. An alternative is to use geographically aggregated data published by public administrations.

The use of aggregated data has been recurrent to adjust survey estimates. As discussed in the previous section, some adjustment methods such as calibration or poststratification require aggregated data to fit the models (Särndal and Lundström 2005:49–65). Moreover, in both calibration and poststratification, auxiliary variables must be measured in the sample as part of the survey. Another approach consists in using aggregated data to a lower geographical level as contextual variables, which provide information about the environment—census tract or municipality—of the population members. Only two research have addressed this use of administrative data to tackle estimates bias (Biemer and Peytchev 2012, 2013; Lahtinen and Butt 2015). In both cases, they used probability samples such as the National Comorbidity Survey Replication or the European Social Survey in the UK. Also, in both cases, the aggregate variables used from the census and other government sources showed no ability to adjust for nonresponse bias.

# 2. Article I: Datos agregados para corregir los sesgos de no respuesta y de cobertura en encuestas

Cabrera-Álvarez, Pablo. 2021. "Datos Agregados Para Corregir los Sesgos de No Respuesta y de Cobertura en Encuestas." *Empiria. Revista de Metodología de Ciencias Sociales* (49):39–64. doi: 10.5944/empiria.49.2021.29231.

**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.

An English version of this paper can be found in appendix B.

# *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 c omo p ara p oner e n 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 de 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 ($z$):

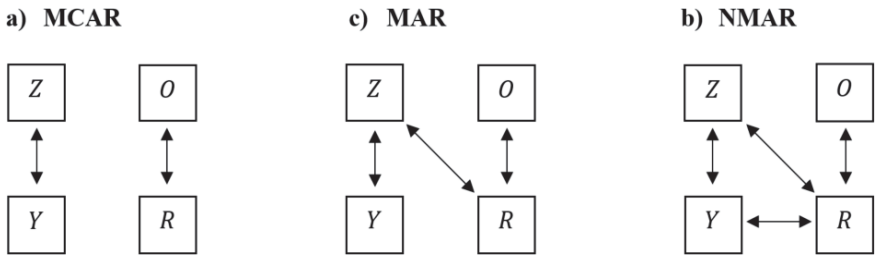
$$w_{zj} = \frac{N_{zj}}{n_{zj}}$$

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 prácticos, 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 $\rho$, que se define como:

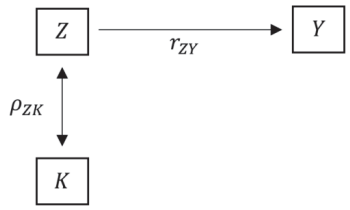
$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

en la que $\sigma_b^2$ se refiere a la varianza entre los grupos, que están definidos por la variable de agrupación, y $\sigma_w^2$ a la varianza dentro de los grupos (Liljequist, Elfving y Roaldsen 2019). Por lo tanto, $\rho$ 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 $\rho_{ZK}$; el nivel de agrupación de $Y$ está determinado por $\rho_{YK}$ 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.

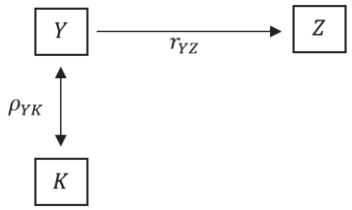
En el primer escenario la variable auxiliar (*Z*) juega un papel determinante al establecer el nivel de agrupación de la variable estimada (*Y*), como se observa en la Figura 2. Para clarificarlo, pensemos que queremos estimar la distribución de la afiliación religiosa de la población (*Y*) a partir 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 ($\rho_{ZK}$) 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*** $r_{ZY}$.



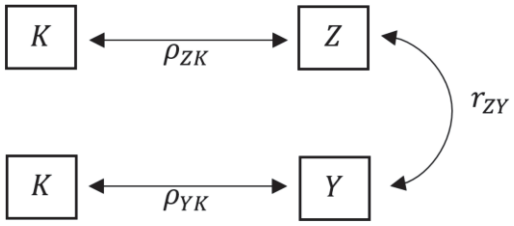

En el segundo escenario la variable $Y$ está agrupada de manera independiente, mientras que el nivel de agrupación de $Z$ viene dado por la correlación $r_{zy}$ (Figura 3). Este escenario es idéntico al primer caso planteado, pero en este sería $Y$ la variable que determinaría el nivel de agrupación 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 $\rho_{ZK}$ y $\rho_{YK}$ 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 $\rho_{ZK}$. 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 $\rho_{YK}$ 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 $\rho$. 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 $Y_q$:

$$Y = G\left(Y_q\right)$$

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 |
| $\rho_{YK}$ | Correlación intraclase entre la distribución en conglomerados y la variable a estimar | De 0,05 a 0,95 en pasos de 0,1 |
| $\rho_{ZK}$ | 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:

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

en la que $|\boldsymbol{B}_{(\bar{y}w)}| = |\bar{Y} - \bar{y}w|$ representa el valor absoluto del sesgo de la estimación ponderada y $|\boldsymbol{B}_{(\bar{y})}| = |\bar{Y} - \bar{y}|$ 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* ($\rho_{YK}$) | 499.500 | 0,23 | 0,29 | 0,00 | 0,99 |
| Corr. intraclase *Z* ($\rho_{ZK}$) | 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

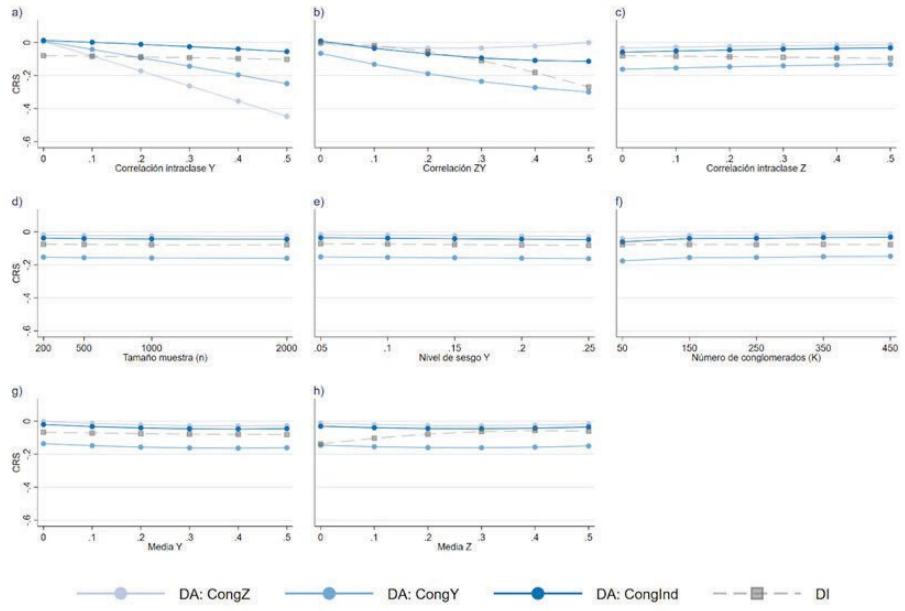
aumento del nivel de la correlación intraclase en una décima supone una reducción media del sesgo de 9 puntos porcentuales, que contrasta con la reducción en las modalidades *congY* (5 puntos) y *congInd* (1 punto). Esta diferencia se explica por las condiciones bajo las que se da la agrupación de la 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 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 p = 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 PEREZ-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

# 3. Article II: The effect of weighting and multiple imputation on bias in Spanish election polls

Cabrera-Álvarez, Pablo, and Modesto Escobar. 2019. "The Effect of Weighting and Multiple Imputation on Bias in Spanish Election Polls." *Revista Española de Investigaciones Sociológicas* 165:45–64. doi: 10.5477/cis/reis.165.45.

**Resumen**

Este artículo tiene como objetivo evaluar la eficacia de las correcciones realizadas en encuestas electorales para ajustar las estimaciones por la incidencia de los sesgos de selección y no respuesta. Para ello se ponen a prueba distintos métodos de ponderación e imputación múltiple en todos los estudios preelectorales y postelectorales de elecciones generales al Congreso de los Diputados español llevados a cabo por el Centro de Investigaciones Sociológicas desde 1982. Para ello se utilizaron seis métodos de ponderación, según incluyeran variables sociodemográficas y diferentes versiones de la variable recuerdo de voto—con y sin imputación—y cuatro variantes de la imputación para la variable intención de voto según el tipo de modelo y las variables utilizadas en su especificación. Los resultados muestran la ventaja de utilizar la variable recuerdo de voto en la ponderación cuando hay estabilidad en las preferencias de los electores. De modo complementario, para tratar la no respuesta, el uso de técnicas de imputación tiene un efecto limitado y condicionado por las variables incluidas en el modelo.

**Palabras clave:** estimación de voto, imputación múltiple, recuerdo de voto, ponderación, Total Survey Error.

doi:10.5477/cis/reis.165.45

# The Effect of Weighting and Multiple Imputation on Bias in Spanish Election Polls

## *El efecto de la ponderación y la imputación en el sesgo de los estudios electorales en España*

**Pablo Cabrera-Álvarez and Modesto Escobar**


### Key words

Polls
- Voting Estimates
- Multiple Imputation
- Past Vote
- Weighting
- Total Survey Error



### Abstract

The purpose of this article is to assess the effectiveness of post-survey adjustments made to electoral polls in order to correct for non-response bias. To do so we have used different weighting and multiple imputation methods using pre-election and post-election polls conducted by Spain's Centre for Sociological Research for all Spanish general elections since 1982. The results show the benefit of weighting by past vote when voters' preferences remain stable. However, the use of multiple imputation techniques to address missing values has a limited effect and is influenced by the variables included in the model.



### Palabras clave

Encuestas
- Estimación de voto
- Imputación múltiple
- Recuerdo de voto
- Ponderación
- *Total Survey Error*



### Resumen

Este artículo tiene como objetivo evaluar la eficacia de las correcciones realizadas en encuestas electorales para ajustar el efecto de la no respuesta. Para ello se ponen a prueba distintos métodos de ponderación e imputación múltiple en todos los estudios preelectorales y postelectorales de elecciones generales al Congreso de los Diputados español llevados a cabo por el Centro de Investigaciones Sociológicas desde 1982. Los resultados muestran la ventaja de utilizar la variable recuerdo de voto en la ponderación cuando hay estabilidad en las preferencias de los electores. De modo complementario, para tratar la no respuesta, el uso de técnicas de imputación tiene un efecto limitado y condicionado por las variables incluidas en el modelo.



### Citation

Cabrera-Álvarez, Pablo and Escobar, Modesto (2019). "The Effect of Weighting and Multiple Imputation on Bias in Spanish Election Polls". *Revista Española de Investigaciones Sociológicas,* 165: 45-64. (http://dx.doi.org/10.5477/cis/reis.165.45)



**Pablo Cabrera-Álvarez:** Universidad de Salamanca | pablocal@usal.es
**Modesto Escobar:** Universidad de Salamanca | modesto@usal.es

## INTRODUCTION[1]

Estimates of voting based on pre-election polls tend to approximate final election results the closer they are carried out to the actual election. Although in recent years studies suggest that the accuracy of pre-election polls continues to meet satisfactory limits of what is expected (Jennings and Wlezien, 2018), we have also seen cases, such as the general elections in the United Kingdom in 2015 and in Spain in 2016, that have caused widespread debate about the usefulness and need for pre-election polling.

Among the reasons for the lack of accuracy in pre-election polls is the existence of non-response bias, that is, a systematic difference between the voting intentions of those who participate in the study and those who do not. To mitigate this bias, there are statistical techniques that correct possible deviations in the sample profile using complementary information. One of the commonly used auxiliary variables in this procedure is past vote, based on asking survey participants about their past voting behaviour. However, debate continues over the use of this variable, as there is no evidence that it always has a positive effect on the accuracy of voting estimates. With the same aim, multiple imputation (MI) techniques can be used to assign valid values to those who say they do not know their voting intention or do not answer the question.

This article, focused on the Spanish case, seeks to determine the effect of the use of multiple imputation and weighting on the accuracy of voting estimations, using a perspective that compares pre and post-electoral polls for Spanish parliamentary elections in the period from 1982 to 2016.

Compared to previous research (Escobar *et al.*, 2014; Pavía and Larraz, 2012; Rivas *et al.,* 2010), this study represents a new and important contribution for three reasons. The first is the use of different transformations of the past vote variable in the weighting, along with a set of socio-demographic variables that have not been previously used in Spain. Second is the extensive use of multiple imputation to address both voting intention and past voting. The third is that both the weights and imputations are tested using pre and post-electoral studies by Spain's Centre for Sociological Research (CIS), covering general elections in Spain since 1982. This perspective over time is necessary to clarify if the success of these techniques changes from one election to another and if there is a trend that explains such variability.

The article is divided into four sections. In the first, we look at how accuracy is defined and propose the *Total Survey Error* (TSE) paradigm as the theoretical framework for studying election polls. Next, we present our working hypotheses and the data and methodology used. This is followed by the results of our analysis and discussion. Lastly, we present our conclusions.

## ACCURACY AND SOURCES OF ERROR IN PRE-ELECTION POLLS

The concept of accuracy in election polls involves two issues. On the one hand, the notion of variability resulting from sample size and variance in estimation, represented by the margin of sampling error, and on the other, the notion of fit, as the difference between the estimation and population data, in this case, the results of the election. In studies such as this one, which analyse poll performance, the concept of fit is used taking into account that this comparison between polls and election results can be altered by the effects of the electoral campaign (Crespi, 1988; Sturgis *et al*., 2016).

[1] This research has been supported by the predoctoral scholarship programme of the Obra Social "La Caixa".

The lack of accuracy in pre-election polls is a recurring theme in the literature on public opinion and elections (Caballé *et al.*, 2013; Callegaro and Gasperoni, 2008; Durand *et al.*, 2004; Sanders, 2003; Traugott, 2005). Among the causes of the discrepancy between opinion polls and actual election results, we find changes in the preferences of voters between the time of the survey and the elections (Abrams, 1970; Shlapentokh, 1994), the methods used to determine the likelihood of voting (Durand *et al.*, 2004; McEwen, 2004; Sturgis *et al.*, 2016), the sample used (Abramson, 2007; Curtice, 1997; Lynn and Jowell, 1996; Worcester, 1996), problems related to population coverage (Callegaro and Gasperoni, 2008; Durand *et al.*, 2001; Sauger, 2008), and how cases that do not respond to the survey or to the question regarding voting intention are treated (Anderson, 1992; Jowell *et al.*, 1993; Katz, 1941).

The TSE framework allows us to systematically analyse the sources of error that exist in the design of the study, as well as in the gathering, processing and analysis of data. Knowing and controlling these sources of error is essential for maintaining the quality of estimates (Biemer, 2010; Biemer and Lyberg, 2003). The TSE divides sources of error into two groups: those related to measurement and those related to representativeness. Regarding measurement, we find issues of validity, measurement errors and errors in data processing. Regarding representativeness, we find errors in coverage, sampling error, non-response error and errors derived from the adjustments made after the gathering of the data (Groves *et al.*, 2013).

## Non-response bias in pre-election surveys

This study focuses on the non-response error. Non-response refers to a lack of information due to a sample element not being reached or not collaborating in the survey or a part of the survey (Lynn, 2008). In the context of this study, we say that there is a non-response bias when those that respond in a poll, or specifically to a question regarding voting intention, have voting preferences that are different from those that do not respond.

Unit non-response (that is, when the sample element rejects participation in the survey or is not contacted) has been identified as one of the causes of problems in the accuracy of pre-election polls in different countries, among them Spain (Durand *et al.*, 2004; Jowell *et al.*, 1993; Smith, 1996; Pavía *et al.*, 2016). In this regard, some studies have shown that the propensity to respond to socio-political surveys is related to the level of interest in politics the citizens in the sample have (Voogt and Saris, 2003). More recently, it has been shown that those who are going to vote tend to be over-represented in post-election studies, contributing to estimates of participation based on polls exceeding real figures (Ansolabehere and Hersh, 2012; Sciarini and Goldberg, 2016).

The other factor associated with non-response bias is the refusal of some participants to reveal their voting intention. This has also been identified as a possible cause of the lack of accuracy in pre-election surveys (Curtice, 1997; McEwen, 2004; Sauger, 2008). Regarding Spain, Urquizu (2005) showed that in the 1980s conservative voters were less likely to reveal their voting preferences, while this tendency was reversed in the 1990s.

## Weighting and imputation as methods for adjusting non-response bias in studies of voting

Once the data have been gathered, it is possible to apply adjustments to reduce the impact of biases caused by both unit and item non-response. In the case of total

non-response, weighting techniques are used to rebalance the final sample. In the same way, imputation techniques can be used to attribute valid response categories to those that avoid answering a specific question.

The use of weighting and similar techniques to calibrate results has been common in combination with quota sampling (Särndal, 2007). However, there is a debate in the literature over the usefulness of including past vote as an auxiliary variable in the weighting. For example, in the United States the majority of polling firms have avoided using this variable (Voss *et al.*, 1995). In a classic study on pre-election polling methodology in the U.S., Crespi (1988: 40-41) states that the main reason cited for not using this variable is measurement problems associated with it, including the over-representation of voters in comparison to abstentionists, and the over-representation of those who choose the winning candidate or party versus those that vote for the losers. Worcester (1996), in regard to the United Kingdom, argued that the use of past vote does not help and could actually lead to less accurate estimates. Along the same lines, a recent study on pre-election polls in Canada and France by Durand *et al.* (2015) shows that past vote can both improve and worsen the accuracy of voting estimates.

Despite this debate, the use of the past vote variable in weighting is common. For example, in the United Kingdom and in France this variable has been used to correct for under-representation of conservative voters (Crewe, 2001) and National Front voters (Durand, 2008) respectively. In Spain, only a few academic studies have attempted to shed light on this phenomenon. Escobar *et al.* (2014) compared different methodologies used to carry out estimates of voting behaviour based on polls using past vote as a weighting variable. In that study, they found that in elections that produced a change in the ruling party, the use of the past vote variable weakened the accuracy of estimates in the period 1979 to 2011. Pavia and Larraz (2012) also experimented with different forms of weighting by past vote, reaching the conclusion that post-stratification was not the best method in response to non-response bias.

Regarding item non-response, once the data have been gathered, researchers have to decide how to proceed with cases that are likely to vote but that do not reveal their voting intention. This problem has been addressed using *ad hoc* techniques defined by each research organisation (Crespi, 1988: Lynn and Jowell, 1996; Sturgis *et al.*, 2016). In Spain, we find the study of Varela *et al.* (1998), which describes different methods that can be used for imputing a valid response to those who do not reveal their voting preferences, as well as that of Pavía and Larraz (2012), who employed imputation by expert criteria to address non-responses in voting intention and past vote. In addition, Rivas *et al.* (2010) study discusses the relevance of the use of imputation to address partial non-response in voting intention, using the 2000 elections as a case study. Their conclusion is that this technique is only effective when the predictors allow us to differentiate all the possible categories of voter intention.

Multiple imputation is a technique that is used to assign valid values to cases that have missing values, but its application in pre-election studies has been limited. King *et al.* (2001) analysed its potential use in political science, suggesting it could be used to study the preferences of non-voters. Bernhagen and Marsh (2007) used this technique to assign valid values to individuals that did not declare their voting intention, and Liu (2014) used multiple imputation techniques to assign preferences to individuals that did not reveal their voting intention in a pre-election study in Taiwan, but without success. For Spain, Escobar and Jaime (2013) were also not able to achieve greater

accuracy in their estimates using different methods of multiple imputation in pre and post-election studies from the Centre for Sociological Research in the 2011 general elections.

## Research hypotheses

To meet our objective of determining the effect of the use of weighting and multiple imputation on the accuracy of estimates of voting behaviour in Spain, we have formulated the following hypotheses:

Hypothesis 1: *Weighting the sample using socio-demographic variables positively affects the accuracy of vote estimates.*

However, we expect the impact to be limited because quotas are already used and because in general these types of variables have little relationship to voting intention. In addition, weighting techniques reduce the total non-response bias in estimates when the information employed is correlated with the propensity to respond to the variable of interest, in this case, voting intention (Särndal, 2007).

Hypothesis 2: *Weighting the sample by past vote increases the accuracy of vote estimates*.

Using the variable past voting behaviour, quite widespread in the polling industry, has a different impact, given that it is correlated with intention to vote (Crespi, 1988; Crewe, 1997). Some studies in other countries, however, have shown that the effect of using this variable has been minimal (Duran, Deslauriers and Valois, 2015).

Hypothesis 3: *The use of multiple imputation techniques in treating the past vote variable reduces the level of bias present in this variable, and as a consequence, the use of this imputed variable in the weighting increases the accuracy of the vote estimates*.

To the extent that the recall of past vote in previous elections could be impacted by memory problems or by item non-response bias (Crespi, 1988; Worcester, 1996), we propose a procedure to correct deviations by the use of multiple imputation techniques.

Hypothesis 4: *The use of weighting that combines socio-demographic variables and past vote will be the most effective in reducing the error level of vote estimates (PV+SD).*

This hypothesis is a corollary of the preceding ones. If weighting with socio-demographic variables and past vote separately affects vote estimates positively, it would be expected that combining these variables would improve the results.

Hypothesis 5: *The use of the past vote variable to weight results has a positive effect on the accuracy of estimates in elections in which there is political continuity.*

Escobar *et al.* (2014) reveal a trend in the use of past vote in weighting in the Spanish case (1979-2011): when in certain elections the ruling government fails to maintain its hegemony, the use of past vote in weighting has a negative effect on accuracy. In the 2014 elections, Spain's party system changed with the emergence of two new parties and the decline in the percentage of support given to the two main political parties (Orriols and Cordero, 2016; Rama, 2016). It is necessary to see whether this generalisation is confirmed in the subsequent 2015 and 2016 elections.

Hypothesis 6: *The use of multiple imputation techniques to assign a voting intention or behaviour to those that do not know or do not answer increases the accuracy of vote estimates*.

In this regard, we advocate for the use of multiple imputation to assign valid values to those that do not reveal their voting behaviour in elections being held close to the date of the survey (King, 2001).

Hypothesis 7: *In the same way that occurs with weighting by past vote, the effect of the use of multiple imputation techniques depends on whether election results return the ruling party to government or not*.

Escobar and Jaime (2013) showed the positive effect that imputation had on the accuracy of estimates in the 2011 elections. However, to date, no other study on the effects of this procedure on other elections has been published.

Sample

In order to examine the implications of using different types of weightings and imputations, we have worked with pre-election and post-election polls carried out by the CIS between 1982[2] and 2016 for Spain's general elections. All of the CIS samples used multi-stage stratified sampling by province and population size, with the selection of households by random routes and subjects by sex and age quotas. The studies used are shown in the following table:

**TABLE 1.** *Year, number and size of CIS studies used*

| Year | CIS study number | Pre-electoral sample size | Post-electoral sample size |
|---|---|---|---|
| 1982 | 1.326 y 1.327 | 24,832 | 2,394 |
| 1986 | 1.526 y 1.542 | 25,304 | 6,842 |
| 1989 | 1.821/37 y 1.842 | 27,122 | 2,508 |
| 1993 | 2.060 y 2.061 | 2,462 | 4,225 |
| 1996 | 2.207 y 2.210 | 6,544 | 4,610 |
| 2000 | 2.382 y 2.384 | 24,040 | 4,386 |
| 2004* | 2.555 | 24,109 | |
| 2008 | 2.750 y 2.757 | 18,221 | 5,247 |
| 2011 | 2.915 y 2.920 | 17,201 | 6,056 |
| 2015 | 3.117 y 3.126 | 17,403 | 5,457 |
| 2016 | 3.141 y 3.146 | 17,458 | 5,136 |

* The 2004 post-electoral study was excluded from the analysis as it did not include the variable, past vote.

*Note:* Starting in 2000, with the exception of 2004 and 2016, the pre and post-electoral studies were panel type studies.

## METHODOLOGY

In this part, we present the methodology in four sections. In the first one we describe our data sources. In the next section, we address the different criteria used for weighting the data, and in the third section, the imputation procedures used. In the last section we present the criteria used to determine the accuracy of voting estimates.

The inclusion of post-electoral studies stems from the limitation resulting from the dates of the fieldwork for the CIS pre-election polls, around one month before the elections. Post-election studies seek to mitigate the

[2] The 1977 and 1979 elections are not included because no post-electoral study is available. In addition, in the case of the 1977 elections, no past vote variable, essential for this study, is available.

possible bias introduced by the effects of the electoral campaign, which are not detected in the pre-electoral polls. However, it is also necessary to point out limitations presented by post-electoral studies: 1) voters in elections are over-represented (Sciarini and Goldberg, 2016); 2) in the case of panel-type studies, these may have conditioning effects on participants (Sturgis *et al.*, 2009), and 3) there is usually an over-representation of the winning party in the recent elections (Crespi, 1988).

### Weighting criteria

In pre-electoral studies, the CIS designs a stratified sample by province with non-proportionate allocation. In these cases, a selection weight has to be applied. These weights ($w_k$ ) are equal for interviewees in the same electoral district, and their formula is the following:

$$w_k = e_k \, In_k$$

with $e_k$ being the size of the electoral census and $n_k$ being the number of interviews carried out in each strata, province or electoral district.

The remaining weights were calculated using the logistic calibration method. There are other methods for generating weighting coefficients, such as the use of non-response models to determine the probability of responding in a survey, or methods based on *propensity score matching* techniques. Comparison of these methods shows that the key lies in the predictors used (Mercer *et al.*, 2018), rather than in the statistical technique used to generate the weights. In this case, given population data can only be obtained in aggregate form, the technique used was calibration in its logistic version; its advantage over the linear version is that it avoids the generation of negative weighting coefficients.

After obtaining the sample, logistic calibration requires a comparison of the distribution of one or more of its variables to see if they coincide with the parameters of the population, in order to calculate weights that ensure that the sample results coincide with

**TABLE 2.** *Weightings by province, past vote (PV) and sociodemographic variables used in the design of the research*

| Abbreviation | Weighted variables | Imputed PV | PV sphere * |
|---|---|---|---|
| BE | Province | | |
| SD | Province and sociodemographics** | | |
| PV0 | Province and past vote (PV) | No | NR, NV and NVM excluded |
| PV1 | Province and past vote | No | NR, NV and NVM included |
| PV2 | Province and past vote | Yes (NR imputed) | NR, NV and NVM included |
| PV3 | Province and past vote | Yes (NR and NVM imputed) | NR, NV and NVM included |
| SD+PV | Province and sociodemographics** and PV | Yes (NR and NVM imputed) | NR, NV and NVM included |

NR: non-response (DK (don't know) and NA (no answer)); NV: did not vote in prior elections; NVM: did not vote, minor.

* Non imputed categories, whether for no response (NR) or for not voting due to being a minor (NVM), only the weight corresponding to the rest of the weighting criteria are assigned.

** Age by sex, autonomous region, size of habitat, education level and employment status.

the population totals in the selected variables[3]. The first calibration criterion used includes only socio-demographic variables, while the rest of the weights use past vote[4]. A summary of the weights used is provided in Table 2.

### Imputation criteria

The method proposed by Rubin (1987) was used for the analysis and imputation of incomplete data, which consists of reconstructing new data sets, as many as the researcher establishes, with randomly simulated values based on other variables from the study that contain more complete information[5]. In contrast to a single imputation, which consists of estimating the data only once, multiple imputation makes a series of estimates – by simulating a number of complete data sets – from which a single estimate can be constructed, supplemented by the variation from the diverse estimates carried out. Consequently, the variances of the parameters can be obtained more accurately than with a single imputation.

There are different imputation procedures based on Bayesian and frequentist principles. Essentially, we can distinguish between univariate imputations (one variable at a time), based on the posterior predictive distribution of missing data, and chained imputations, which involve feedback of the imputed variables. Table 3 summarises the types of imputations taken into account in this study (note the correspondence between the versions of imputed past vote and the weights of the same name).

The choice of the predictors included in the imputation models was made based on theoretical criteria (Escobar and Jaime, 2013) and considering the limitations resulting from the research design. First, studies of electoral behaviour were used to determine which socio-demographic and political predictors are related to voting intention (for example, Bosch and Riba, 2005; Jaime and Saéz, 2001; Lago and Lago, 2005). Secondly, the list was limited to be able to apply the same model to all the elections studied.

Given that the variable of interest for the imputation is always the vote ($x_i$), we used a multinomial model in which the probabilities for the $k$ categories of the variable respond to the following formula, where $\mathbf{z}_i$ is the vector of the variables used in the imputation:

$$\Pr(x_i = k \mid \mathbf{z}_i) = \frac{1}{1 + \sum_{l=2}^{K} \exp(\mathbf{z}_i' \ \boldsymbol{\beta}_i)}, \ si \ k = 1$$

$$\Pr(x_i = k \mid \mathbf{z}_i) = \frac{\exp(\mathbf{z}_i' \ \boldsymbol{\beta}_i)}{1 + \sum_{l=2}^{K} \exp(\mathbf{z}_i' \ \boldsymbol{\beta}_i)}, \ si \ k > 1$$

### Evaluation of the accuracy of results

To assess the results of the estimates and imputations we use the weighted mean absolute error (WMAE), used in the literature on

[3] For more information regarding the calculation of weighting coefficients by calibration, see the studies of Särndal (2005) and Lundström and Särndal (2001). The calibration was carried out in Stata, using the calibrate package designed by D'Souza (2011).

[4] To carry out the calibration it is necessary to have the population distributions for the auxiliary variables. The data to carry out the weighting by past vote and population size come from Spain's Interior Ministry. The information on population distributions for the variables sex, age and autonomous regions come from Spain's National Statistics Institute (INE). Historical data on education level and employment status come from studies by Fuente and Domenech (2015) and Fuente (2015), respectively. In the case of the last two elections, 2015 and 2016, for which no data exists in the just mentioned studies, values for the populations were obtained from the INE's Labour Force Survey for employment status and were interpolated for education level.

[5] A basic introduction along with the way to obtain these models with Stata can be found in the book dedicated to multiple imputation (Stata, 2015). In addition, a theoretical and applied presentation in Spanish is found in the already cited book by Rivero (2011).

**TABLE 3.** *Imputations included in the design for past vote and voting intention*

| A) Past vote (weighting for this variable) | | |
|---|---|---|
| Name* | Objective variable (model) | Predictive variables |
| PV1 not imputed | | |
| PV2 imputed (NR) | Past vote (multinomial) | Mixed set ** Mixed set |
| PV3 imputed (NR and NVM) | Past vote (multinomial) | Mixed set ** Mixed set |

| B) Voting intention (to estimate value) | | |
|---|---|---|
| Name | Objective variable (model) | Predictive variables |
| 1. Not imputed | | |
| 2. Basic univariate | Voting intention (multinomial) | Basic set *** |
| 3. Enhanced univariate | Voting intention (multinomial) | Enhanced set **** |
| 4. Basic chained | Voting intention (multinomial)<br>Past vote (multinomial)<br>Ideology (ordinal) | Basic set *** |
| 5. Enhanced chained | Voting intention (multinomial)<br>Past vote (multinomial)<br>Ideology (ordinall) | Enhanced set **** |

* In past vote, the non-response (NR) or those who did not vote in the previous elections because they were minors (NVM) were imputed (univariate method).
** Sex, age, habitat size, autonomous community, voting intention and ideology.
*** Sex, age, education level, past vote and ideology.
**** Sex, age, education level, past vote, ideology, evaluation of the economic situation and evaluation of the political situation.

forecasting time series[6]. The formula is the following:

$$\sum_{k=1}^{K} \quad k \quad k \quad k$$

where $p_k$ is the electoral result for each $k$ party and $p^{\wedge}_k$ their corresponding estimates.

[6] There are three measures used for these ends: absolute mean error, quadratic mean error and standardised quadratic mean error. For their visibility, we have used the first, adding the weighting of the average errors. The latter was done because we are forecasting non-bipartisan elections, so it is logical to give greater importance to the errors committed in regard to the most voted parties. See Lewis (2005) and Hyndman and Koehler (2005).

## Models

To discover which weighting and imputations methods are the best for election forecasting, we have considered two results:

a) The WMAE(S) independently obtained for each survey based on voting intention (or past vote in the case of post-election studies) for the elections considered (models 1 to 4 in table 1 of the appendix).

The predictors included in these models were:

1) Year of the election (models 1 and 2).
2) Classification of elections as elections of change or continuity (models 3 and 4).
3) Type of survey (pre or post-election).

These two variables serve to control for the effect of the electoral campaign and the political climate.

4) Weighting modalities (the seven presented in table 2), with the aim of verifying *hypotheses 1, 2* and *4.*
5) Methods of imputing voting intentions (the five presented in table 3.b[7]), to verify *hypothesis 6.*
6) Interaction between the form of weighting and the year (models 1 and 2), or classification of the elections based on whether they are elections of change or continuity (models 3 and 4), considered in *hypothesis 5.*
7) Interaction of the imputation method and year (models 1 and 2), or classification of the elections based on whether they are elections of change or continuity (models 3 and 4), considered in *hypothesis 7*.

b) The WMAE(P) of past voting in elections prior to the poll in question (model 5 of table 1A in the appendix) for testing *hypothesis 3.* In this case, the predictors were year, type of survey and method of imputation of past vote (the three included in table 3.a).

With the different treatments of imputation and estimation by election and survey, we obtained 595 different estimates[8] of the WMAE(S) and 63 of the WMAE(P).

The hypotheses mentioned were compared within regression models, adjusted with least squares, through specific contrasts of the estimated means of the weighted mean absolute errors. Once the F values of these comparisons were calculated, the Bonferroni correction was applied to avoid Type I errors (Rosenthal *et al.*, 2000).

## Results and discussion

The first hypothesis proposed in this study refers to the need to use socio-demographic weighting to improve the accuracy of voting estimates. We stated that the improvement in accuracy after balancing the sample by sex, age, education and economic activity would depend on whether these variables are related to voting intention and the probability of responding to the survey (Särndal, 2007). Graph 1.a shows that during the 1980s, the use of this weighting contributes slightly to improving estimates, although if the complete period is analysed there are no differences between the estimates without weighting and adjusted estimates by socio-demographic profiles ($F_{(1,\ 341)} = 2.59$; $p = 0.650$). The fact that socio-demographic variables do not have a clear relationship to voting intention or the likelihood of responding in a survey is not surprising, as other studies in the American and British contexts also point in this direction (Crespi, 1988; Sturgis *et al.,* 2016).

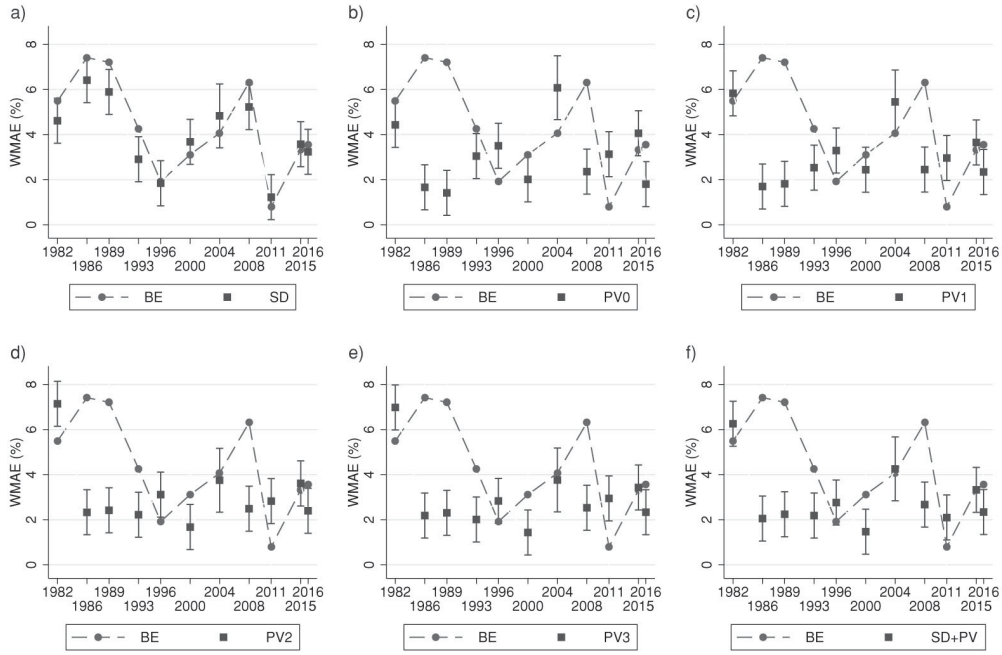
Faced with the ineffectiveness of using socio-demographic factors, an alternative is to use past voting, which is related to voting intention, and with the likelihood of accepting to participate in a survey (Voogt and Saris, 2003). The results show that in general the use of this variable in weighting helps reduce the error level in voting estimates (graphs 1.b to 1.f).

As described in the methodology section, we have worked with several versions of weighting by past vote in order to understand

[7] But only three imputation modalities are possible before 2000, as the surveys did not include questions on evaluation of the government and the economy. As a result, models 1 and 3 are split (in which we only consider three imputation modalities for the whole period analysed) in models 2 and 4 in which there are more imputation modalities, but over less time: only the last six elections.

[8] Among them, 441 do not contain broader imputations and, therefore, they are available in the 11 elections addressed, while 385 correspond to the surveys in which participants were asked to evaluate the government and the economy and, therefore, exclude predictions prior to the year 2000.

**GRAPH 1.** *Graph 1 Comparison of the Weighted Mean Absolute Error (WMAE) for estimating the weighted vote in its different versions in comparison with the base estimate (BE) without weighting*




The reference for comparison is the base estimate (**BE**).
**SD**: Sociodemographic (sex, age, region, employment status and qualifications); **PV0**: Past vote excluding those DK/NA or did not vote; **PV1**: Past vote; **PV2**: Imputed past vote (DK and NA); **PV3**: Imputed past vote (DK, NA and underage previous elections); **SD+PV**: Sociodemographic and PV3.

if the treatment of this variable through imputation has a positive impact on the quality of estimates (hypotheses 2, 3 and 4). Two of these alternatives (PV0 and PV1) do not use imputation techniques to correct for possible bias resulting from item non-response, while the other three versions do use this technique (PV2, PV3 and SD+PV).

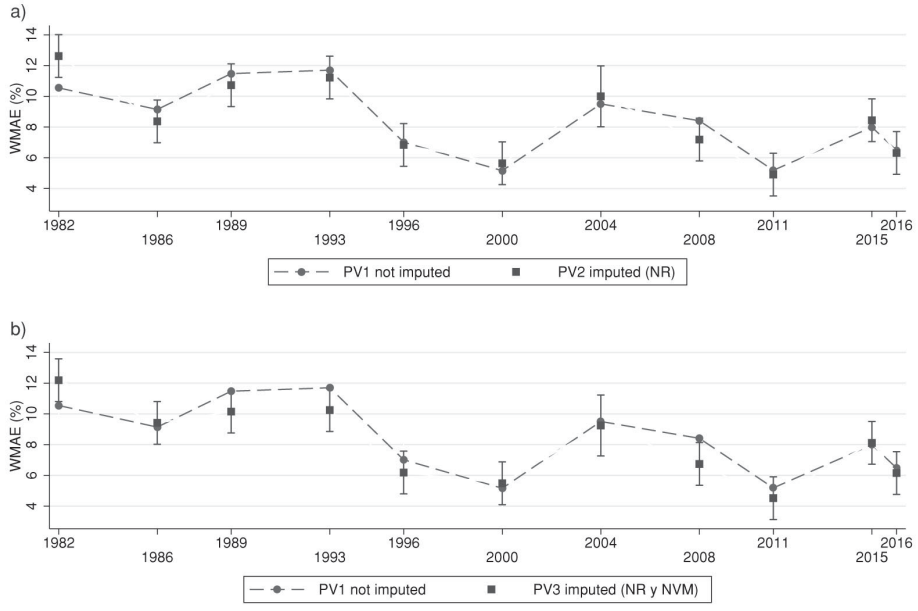
In general, although public information regarding this is limited, we know that polling firms carry out a minimum transformation of the past vote variable before including it in any weighting. In this study, we have tried to replicate this strategy in two weightings. In the case of PV0, we excluded from the process those who do not remember or do not respond to the question regarding past vote, as well as those who said they did not vote in previous elections. In the majority of the elections studied, the use of this weighting improves the accuracy of vote estimations, in comparison to the use of socio-demographic weighting ($F_{(1, 341)} = 15.98$; $p = 0.001$) or the absence of weighting ($F_{(1, 341)} = 31.44$; $p < 0.001$). The same behaviour is found using the PV1 weighting, which in contrast to the case of PV0, includes those respondents who do not reveal their votes in prior elections or who did not vote (without weighting: $F_{(1, 341)} = 27.22$; $p < 0.001$; socio-demographic: $F_{(1, 341)} = 13.01$; $p = 0.002$). This overall positive effect contrasts with the reticence shown by Worcester (1996) or the findings presented by Durand *et al.* (2015) for French and Canadian elections. Despite the deficiencies that this variable may present, overall its use is positive in the Spanish case for the period studied (hypothesis 2).

A possible improvement in the weighting of past vote in prior elections would consist of correcting for deviations in this variable, imputing valid values to those who did not

answer. The PV2, PV3 (hypothesis 3) and SD+PV (hypothesis 4) versions are variations in the weighting by past vote, in which this variable was previously treated with multiple imputation techniques. In the case of PV2, a valid value was assigned to those who did not respond, and with PV3, a valid value was also imputed to those who were not old enough to vote in prior elections. In both cases (graphs 1.c and 1.d), the use of the past vote variable increases the accuracy of estimates in comparison to the absence of weighting (PV2: $F_{(1,\ 341)} = 29.35$; $p < 0.001$; PV3: $F_{(1,\ 341)} = 34.95$; $p < 0.001$) or the use of socio-demographic weighting (PV2: $F_{(1,\ 341)} = 14.50$; $p = 0.001$; PV3: $F_{(1,\ 341)} = 18.51$; $p < 0.001$). However, there is no significant improvement if we compare the results with versions of the past vote without imputation (PV0 and PV1). In the case of weighting SD+PV, which combines socio-demographic variables with imputed past vote (PV3), we do not obtain better results than those found with the rest of the weightings of past vote, presumably due to the null effect of socio-demographic variables in the majority of elections.

The use of imputation to reduce the bias caused by item non-response in the past vote variable does not improve the accuracy of estimates weighted by this variable. This can be seen in graph 2, as the corrections applied to assign valid values to those who did not respond are small (PV2 imputed (NR)), or to those who did not answer and to those who could not vote in prior elections (PV3 imputed (NR and ME)), barely reduce the error level of the variable with the exception of the 1993 and 2008 elections, where we find improvements of more than one percentage point. This may be due to two factors: the predictors included in the imputation and the magnitude of the bias due to partial non-response. On the one hand, the predictors included in the imputation models may not effectively discriminate among the

**GRAPH 2.** *Comparison of the Weighted Mean Absolute Error (WMAE) of the variable past vote without imputation (PV1) and the variables imputed past vote (PV2 and PV3)*



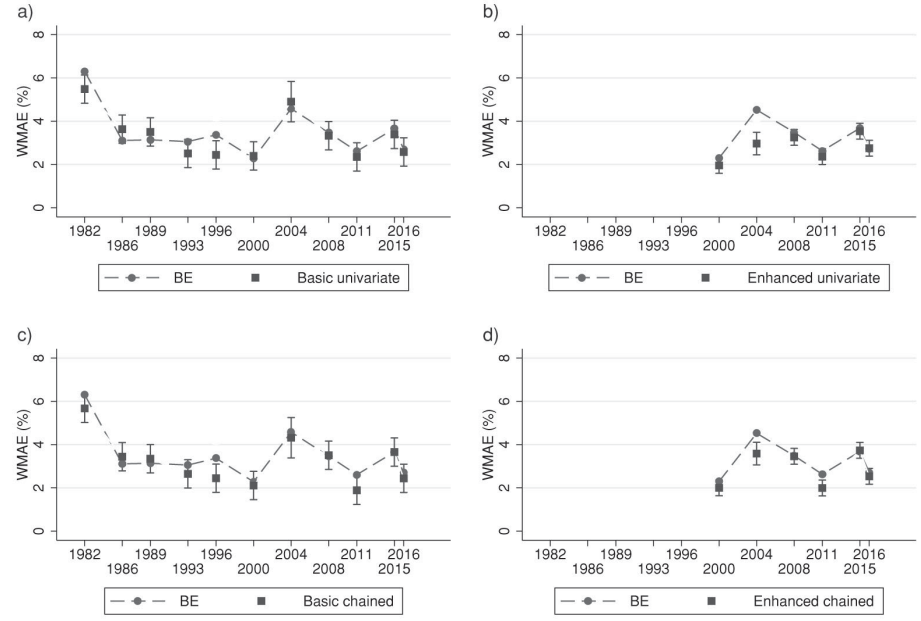

The reference for comparison is the base estimate (**PV1 not imputed**).
**PV2 imputed (NR)**: Imputed non–response (DK and NA); **PV3 imputed (NR and NVM)**: Imputed non–response and underage in previous election (NVM).

different options regarding past vote. On the other hand, the bias generated by partial non-response is likely to be minimal and despite its correction, there may still be a high level of error in past vote recall due to other factors, such as unit non-response. For example, the latter would occur if all those who respond to the survey correctly recall their voting behaviour in past elections, but their voting intentions are different from those who decided not to participate in the survey.

This positive use of past vote in weighting is however not uniform in the period studied (hypothesis 5). In line with the findings of Escobar *et al.* (2014), in the elections in which a change of government took place (1982, 1996, 2004 and 2011), the use of past vote has no positive effect on estimates (for example, PV3: $F_{(1, 338)} = 5.39$, $p = 0.251$). This trend continued in the new political cycle that began in 2014, characterised by the emergence of Podemos and Ciudadanos in the party system (Orriols and Cordero, 2016; Rama, 2016), so that in the 2015 elections, the use of past vote does not contribute to improving the estimate (for example, PV3: $F_{(1, 404)} = 0.01$, $p = 0.999$), while it does seem to do so in surveys of the 2016 elections, although the comparison test does not provide robust results (for example, PV3: $F_{(1, 404)} = 1.66$, $p = 0.198$).

In estimating the vote, it is also important to look at the possible deviations from voting intention, due to the fact that those who reveal their electoral preferences plan to vote differently from those who do not express a preference. The use of multiple imputation (hypothesis 6) can be a way of correcting for these possible deviations (King, 2001; Liu, 2014). As a result, two different imputation techniques have been proposed, univariate and chained, and two sets of variables, one basic (sex, age, education, past vote and ideology) and the other enhanced (also including evaluation of the government and the economic situation).

**GRAPH 3.** *Comparison of the Weighted Mean Absolute Error (WMAE) for estimating the imputed vote in its different versions in comparison with the base estimate (BE) without imputation*




The reference for comparison is the base estimate (**BE**).

The use of univariate imputation (graph 3.a) or chained imputation (graph 3.c) employing a basic set of variables does not improve the estimates. Only in the case of expanded versions (graphs 3.b and .c) – which include survey participants' assessment of government action and the economic situation and which can only be computed starting with the 2000 elections –, do we find a slight improvement in the accuracy of estimates (simple: $F_{(1, 318)} = 12.97$; $p = 0.002$; chained: $F_{(1, 318)} = 8.51$; $p = 0.015$). Liu (2014), along the same lines, using a similar methodology to correct for voting intention in elections in Taiwan, shows the null capacity of this technique to correct for possible bias in voting intention. However, with the data available from the CIS, we can conclude that the low effectiveness of the imputation model is caused by the inadequate choice of predictors.

Thus, in the case of expanded imputation, we find a positive effect, which seems to be reflected in the 2004 and 2011 elections. However, it cannot be shown that there is a clear tendency related to whether the elections are ones in which there is a change in government (hypothesis 7), as occurs in the case of the use of weightings. In part this is because the magnitude of the improvement in the accuracy of the estimates using expanded imputation is minimal.

## Conclusions

The first hypothesis raised in this study addresses the effectiveness of the use of socio-demographic weighting. The results show that despite the fact that the use of weighting produces a limited improvement in estimates until 1993, the contribution has been null since the mid-1990s. This is consistent with what has been found in other case studies (Crespi, 1988; Durand *et al.*, 2015). Either because socio-demographic variables are increasingly less and less related to voting or because the samples are relatively balanced in this regard, the effect of this type of weighting on voting estimates is very limited. As suggested in Sturgis *et al.* (2016), one of the keys is to use population characteristics that are capable of correcting deviations resulting from non-response.

This study, in line with others previously cited, has shown the positive effect of weighting by past vote in the Spanish case. Thus, the second hypothesis is confirmed. However, even if the sample is balanced with respect to the preferences of voters in the most recent election, for this weighting to be successful it is necessary that the voters for a specific party that respond to the survey be representative of those not participating. This requirement is never fully met, and therefore the weighting by recall of a past vote is not sufficient to completely eliminate the bias present in the vote estimate.

In addition, the past vote variable itself may have deficiencies, due to respondents' failures of memory or to the social desirability of the response in certain contexts (Crespi, 1988). To correct these possible problems, two hypotheses were put forward in this paper: The first is that the combination of socio-demographic variables with past vote should be the most effective way of weighting the estimate; the second is that the correction of possible defects in the past vote variable could help to improve the quality of the adjustment. Neither of these hypotheses has been confirmed. What has been determined is that multiple imputation, in the way it has been used, has only lightly corrected the bias present in the past vote variable. This result, rather than raising doubts about imputation as a technique, reveals the importance of the predictors that are selected (Mercer *et al.*, 2018).

In addition, the use of imputation, as stated in the fourth hypothesis, has not resulted in a substantial improvement in the accuracy of the estimates. However, one significant contribution is that in the expanded imputation models, where predictors are added re-

garding respondents' assessment of the government and the economic situation, the ability to reduce the error in the estimate is greater. In line with the findings of Rivas *et al.* (2010), this points to the fact that the choice of predictors is the key, which also applies to the imputation of past vote.

The positive effect of weighting, and more tentatively, of enhanced imputation, is not consistent over time. This trend had already been detected by Escobar *et al.* (2014): in elections that bring about political change, the effect of weighting by past vote is nil, and in some cases, counter productive. This trend is also reflected in this study with the inclusion of data for 2015 and 2016. In the 2015 elections, which are considered elections of political change, the use of the weighting by past vote has a null effect, a situation that was partially reversed in 2016, elections in which the results of 2015 were largely reproduced. In the case of enhanced imputation, no conclusive trend is observed, although a positive effect is concentrated in the 2004 and 2011 studies, elections of political change in which weighting by past vote does not work correctly.

It would make sense in the future to abandon pre-election polls due to the limited number of questions they include, and it would be advisable to study which variables included in post-election studies could yield better results in improving estimates of the electoral behaviour of citizens through surveys.

## Bibliography


Abrams, Mark (1970). "The Opinion Polls and the 1970 British General Election". *Public Opinion Quarterly,* 34(3): 317-324.

Abramson, Paul R. (2007). "The French Presidential Election of 2007: Was Sarkozy the Condorcet Winner?". *French Politics* 5(3):287-291.

Anderson, Leslie (1992). "Surprises and Secrets: Lessons from the 1990 Nicaraguan Election". *Studies In Comparative International Development*, 27(3):93-119.

Ansolabehere, Stephen and Hersh, Eitan (2012). "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate". *Political Analysis,* 20(4):437-459.

Bernhagen, Patrick and Marsh, Michael (2007). "The Partisan Effects of Low Turnout: Analyzing Vote Abstention as a Missing Data Problem". *Electoral Studies,* 26(3):548-560.

Bethlehem, Jelke; Cobben, Fannie and Schouten, Barry (2011). *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Biemer, Paul (2010). "Total Survey Error: Design, Implementation, and Evaluation". *Public Opinion Quarterly*, 74(5):817-848.

Biemer, Paul and Lyberg, Lars E. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons.

Bosch, Agustí, and Riba, Clara (2005). "Coyuntura económica y voto en España". *Revista de Sociología,* 75: 117-140.

Caballé, Adriá; Grima, Pere and Marco-Almagro, Lluís (2013). "Are Election Polls Right? Analysis of the Accuracy of Election Polls Predictions Published in the Press". *Revista Española de Investigaciones Sociológicas,* 143:25-46.

Callegaro, Mario and Gasperoni, Giancarlo (2008). "Accuracy of Pre-Election Polls for the 2006 Italian Parliamentary Election: Too Close to Call". *International Journal of Public Opinion Research* 20(2):148-170.

Crespi, Irving (1988). *Pre-Election Polling: Sources of Accuracy and Error*. New York: Russell Sage Foundation.

Crewe, Ivor (1997). "The Polls: Confidence Restored?" *Parliamentary Affairs,* 50: 569-585.

Crewe, Ivor (2001). "The Opinion Polls: Still Biased to Labour". *Parliamentary Affairs,* 54(4):650-665.

Curtice, John (1997). *So How Well Did They Do? The Polls in the 1997 Election*. Centre for Research into Elecctions and Social Trends. London: Centre for Research into Elecctions and Social Trends.

D'Souza, John (2010). *Calibrate: a Stata Program for Calibration Weighting*. London: Stata User Group.

D'Souza, John (2011). *Calibrate: Stata module to calibrate survey datasets to population totals. Statistical Software Components S457240.* Boston College Department of Economics.

Durand, Claire (2008). "The Polls of the 2007 French Presidential Campaign: Were Lessons Learned from the 2002 Catastrophe?". *International Journal of Public Opinion Research,* 20(3):275-298.

Durand, Claire; Blais, André and Larochelle, Mylène (2004). "The Polls - Review. The Polls in the 2002 French Presidential Election: An Autopsy". *Public Opinion Quarterly* 68(4):602-622.

Durand, Claire; Blais, André and Vachon, Sébastien (2001). "A Late Campaign Swing or a Failure of the Polls? The Case of the 1998 Quebec Election". *Public Opinion Quarterly,* 65(1):108-123.

Durand, Claire; Deslauriers, Melanie and Vallois, Isabelle (2015). "Should Recall of Previous Votes Be Used to Adjust Estimates of Voting Intention?". *Survey Methods: Insights from the Field* 1-14. Available at: http://surveyinsights.org/?p=3543, access September 19, 2017

Escobar, Modesto and Jaime, Antonio M. (2013). "Métodos de Imputación Múltiple para Predecir Resultados Electorales". In: Mendoza Velázquez, A. (ed.). *Aplicaciones en Economía y Ciencias Sociales con Stata*. Texas: Stata Press.

Escobar, Modesto; Rivière Gómez, Jaime and Cilleros Conde, Roberto (2014). *Los Pronósticos Electorales con Encuestas: Elecciones Generales en España (1979-2011)*. Madrid: Centro de Investigaciones Sociológicas.

Fuente, Angel de la and Domenech Vilarino, Rafael (2015). *El Nivel Educativo de la Población en España y sus Regiones: 1960-2011*. BBVA Bank, Economic Research Department. Madrid: BBVA Research. Available at: https://www.bbvaresearch.com/wp-content/uploads/2015/02/WP_15-07_Educacion.pdf, access September 19, 2017.

Fuente, Angel de la (2015). *Series Enlazadas de los Principales Agregados Nacionales de la EPA, 1964-2014*. Instituto de Análisis Económico (CSIC). Madrid: FEDEA Research. Available at: http://documentos.fedea.net/pubs/eee/eee2015-07.pdf, access September 19, 2017.

Groves, Robert M. *et al.* (2013). *Survey Methodology.* New York: Wiley.

Hyndman, Rob J. and Koehler, Anne B. (2005). "Another Look at Measures of Forecast Accuracy," *Monash Econometrics and Business Statistics Working Papers 13/05*. Monash University, Department of Econometrics and Business Statistics.

Jaime Castillo, Antonio M., and Sáez Lozano, José L. (2001). *El comportamiento electoral en la democracia española*. Madrid: Centro de Estudios Políticos y Constitucionales.

Jennings, Will, and Wlezien, Christopher (2018). "Election polling errors across time and space". *Nature Human Behaviour* 1.

Jowell, Roger *et al.* (1993). "Review: The 1992 British Election: The Failure of the Polls". *Public Opinion Quarterly,* 57(2):238-263.

Katz, Daniel (1941). "The Public Opinion Polls and the 1940 Election". *Public Opinion Quarterly,* 5(1):52-78.

King, Gary *et al.* (2001). "Analyzing Incomplete Political Science Data". *American Political Science Review,* 85(1269):49-69.

Lago Peñas, Ignacio, and Lago Peñas, Santiago (2005). "Does the economy matter? An empirical analysis of the causal chain connecting the economy and the vote in Galicia". *Economics and Politics,* 17: 215-243.

Lewis-Beck, M. S. (2005), "Election Forecasting: Principles and Practice". *The British Journal of Politics & International Relations*, 7: 145-164.

Liu, Frank C. S. (2014). "Using Multiple Imputation for Vote Choice Data: A Comparison across Multiple Imputation Tools". *Open Journal of Political Science,* 4:39-46.

Lundström, Sixten and Särndal, Carl E. (2001). *Estimation in the Presence of Nonresponse and Frame Imperfection*. Suecia: Statistics Sweden.

Lynn, Peter (2008). "The Problem of Nonresponse". In: European Association of Methodology (ed.). *The International Handbook of Survey Methodology*. New York: Lawrence Erlbaum Associates.

Lynn, Peter and Jowell, Roger (1996). "How Might Opinion Polls Be Improved? The Case for Probability Sampling". *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 159(1):21-28.

McEwen, Nicola (2004). "Opinion Polling in Scotland: An Analysis of the 2003 Scottish Parliament Election". *Journal of Elections, Public Opinion and Parties,* 14(1):171-190.

Mercer, Andrew; Lau, Arnold and Courtney, Kennedy (2018). *For Weighting Online Opt-In Samples, What Matters Most?*. Washington: Pew Research Centre.

Orriols, Lluis and Cordero, Guillermo (2016). "The Breakdown of the Spanish Two-Party System:

The Upsurge of Podemos and Ciudadanos in the 2015 General Election". *South European Society and Politics*, 21:4, 469-492.

Pavía, Jose M.; Badal, Elena and García-Cárceles, Belén (2016). "Spanish Exit Polls. Sampling Error or Nonresponse Bias?". *Revista Internacional de Sociología,* 74(3):e043

Pavía, José M. and Larraz, Beatriz (2012). "Sesgo de no-respuesta y modelos de superpoblación en encuestas electorales". *Revista Española de Investigaciones Sociológicas,* 137:121-150.

Rama, José (2016). *Crisis económica y sistema de partidos: síntomas de cambio político en España.* Barcelona: Institut de Ciéncies Politiques i Socials.

Rivas, Cristina; Martínez Rosón, María del Mar and Galindo, Purificación (2010). "La Imputación Múltiple como Alternativa al Análisis de la No Respuesta en la Variable Intención de Voto". *Revista Española de Ciencia Política,* 22, 99-118.

Rivero Rodríguez, Gonzalo (2011). *Análisis de datos incompletos en Ciencias Sociales*. Madrid: CIS.

Rosenthal, Robert; Rosnow, Ralph L. and Donald B. Rubin (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge: Cambridge University Press.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Sanders, David (2003). "Pre-Election Polling in Britain, 1950-1997." *Electoral Studies,* 22(1):1-20.

Särndal, Carl E. (2007). "The Calibration Approach in Survey Theory and Practice". *Survey Methodology,* 33(2):99-119.

Särndal, Carl E. and Lundström, Sixten (2005). *Estimation in Surveys with Nonresponse*. England: John Wiley & Sons.

Sauger, Nicolas (2008). "Assessing the Accuracy of Polls for the French Presidential Election: The 2007 Experience". *French Politics,* 6(2):116-136.

Sciarini, Pascal and Goldberg, Andreas C. (2016). "Turnout Bias in Postelection Surveys: Political Involvement, Survey Participation, and Vote Overreporting". *Journal of Survey Statistics and Methodology,* 4(1):110-137.

Shlapentokh, Vladimir (1994). "The Polls - a Review - the 1993 Russian Election Polls". *Public Opinion Quarterly,* 58(46302):579-602.

Smith, Fred T. M. (1996). "Public Opinion Polls: The UK General Election, 1992". *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 159(3):535-545.

Stata (2015). *Stata 14 Base Reference Manual.* College Station. Texas: Stata Press.

StataCorp (2017). *Stata Multiple Imputation Reference Manual*. Texas: StataCorp.

Sturgis, Patrick; Allum, Nick and Brunton-Smith, Ian (2009). "Attitudes Over Time: The Psychology of Panel Conditioning". In: Groves, R. M.; Kalton, G.; Rao, J. N.; Schwarz, N.; Skinner C. and Lynn, P. (eds.). *Methodology of Longitudinal Surveys*. Wiley: Nueva York.

Sturgis, Patrick *et al.* (2016). *Report of the Inquiry into the 2015 British General Election Opinion Polls*. British Polling Council: London.

Traugott, Michael W. (2005). "The Accuracy of the National Preelection Polls in the 2004 Presidential Election". *Public Opinion Quarterly,* 69(5 SPEC. ISS.):642-654.

Urquizu Sancho, Ignacio (2005). "El Voto Oculto en España". *Revista Española de Ciencia Política,* 13:119-156.

Varela Mallou, Jesús *et al.* (1998). "Estimación de la Respuesta de los 'No Sabe/No Contesta' en los Estudios de Intención de Voto". *Revista Española de Investigaciones Sociológicas,* 83:269-287.

Voogt, Robert J. J. and Saris, William E. (2003). "To Participate or Not to Participate: The Link Between Survey Participation, Electoral Participation, and Political Interest". *Political Analysis* 11(2):164-179.

Voss, Stephen; Gelman, Andrew and King, Gary (1995). "Preelection Survey Methodology: Details from Eight Polling Organizations, 1988 and 1992". *Public Opinion Quarterly,* 59:98-132.

Worcester, Robert (1996). "Political Polling: 95% Expertise and 5% Luck." *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 159(1):5.

**RECEPTION:** September 25, 2017
**REVIEW:** February 14, 2018
**ACCEPTANCE:** May 18, 2018

# Appendix 1. Regression models

**TABLE 1A.** *Linear regression models*

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| Elections | | | | | |
| 1982 | | | | | |
| 1986 | 1.16 | | | | -1.41 |
| | (0.81) | | | | (0.96) |
| 1989 | 1.04 | | | | 0.93 |
| | (0.81) | | | | (0.96) |
| 1993 | -1.40 | | | | 1.15 |
| | (0.81) | | | | (0.96) |
| 1996 | -3.44*** | | | | -3.53*** |
| | (0.81) | | | | (0.96) |
| 2000 | -2.84*** | | | | -5.38*** |
| | (0.81) | | | | (0.96) |
| 2004 | -1.93 | 1.04* | | | -1.04 |
| | (1.00) | (0.48) | | | (1.18) |
| 2008 | 0.37 | 3.18*** | | | -2.14* |
| | (0.81) | (0.39) | | | (0.96) |
| 2011 | -4.85*** | -1.78*** | | | -5.36*** |
| | (0.81) | (0.39) | | | (0.96) |
| 2015 | -2.56** | 0.53 | | | -2.55* |
| | (0.81) | (0.39) | | | (0.96) |
| 2016 | -2.29** | 0.73 | | | -4.06*** |
| | (0.81) | (0.39) | | | (0.96) |
| Political change | | | | | |
| Elections of continuity | | | | | |
| Elections of change | | | -1.97*** | -1.62*** | |
| | | | (0.47) | (0.36) | |
| Weighting | | | | | |
| Base estimate (BE) | | | | | |
| SD | -0.87 | 0.56 | -0.75 | -0.27 | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |
| PV0 | -1.06 | -0.87** | -3.26*** | -2.09*** | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |
| PV1 | 0.34 | -0.51 | -3.10*** | -1.74*** | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |
| PV2 | 1.65* | -1.14*** | -3.05*** | -1.93*** | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |
| PV3 | 1.49* | -1.33*** | -3.18*** | -2.00*** | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |
| SD + PV | 0.77 | -1.31*** | -3.15*** | -1.94*** | |
| | (0.72) | (0.31) | (0.39) | (0.27) | |

**TABLE 1A.** *Linear regression models* (continuation)

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| Imputation methods | | | | | |
| Base estimate (BE) | | | | | |
| Basic univariate | -0.82 | 0.11 | 0.03 | -0.05 | |
| | (0.47) | (0.26) | (0.25) | (0.23) | |
| Enhanced univariate | | -0.34 | | -0.18 | |
| | | (0.26) | | (0.23) | |
| Basic chained | -0.63 | -0.17 | | -0.13 | |
| | (0.47) | (0.26) | | (0.23) | |
| Enhanced chained | | -0.30 | | -0.17 | |
| | | (0.26) | | (0.23) | |
| Type of study | | | | | |
| Pre-electoral | -0.53*** | -0.67 | -0.62*** | -0.84*** | 0.44 |
| Post-electoral | (0.12) | (0.07) | (0.16) | (0.11) | (0.25) |
| Treatment of PV | | | | | |
| PV1 without imputation | | | | | |
| PV2 imputed (NR) | | | | | 2.07* |
| | | | | | (0.96) |
| PV3 imputed (NR y NVJ) | | | | | 1.64 |
| | | | | | (0.96) |
| Constant | 6.23*** | 3.26*** | 5.61*** | 4.65*** | 10.34*** |
| | (0.58) | (0.28) | (0.32) | (0.25) | (0.69) |
| F | 7.66 | 16.66 | 11.71 | 11.35 | 11.24 |
| Degrees of freedom | 99 | 66 | 18 | 22 | 33 |
| P-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| R squared | 0.60 | 0.73 | 0.30 | 0.37 | 0.85 |
| Casos | 441 | 385 | 441 | 385 | 63 |

M1: Weighted Mean Absolute Error (WMAE) for estimated vote in the period 1982-2016 only includes imputation methods with basic variables.

M2: Weighted Mean Absolute Error (WMAE) for estimated vote in the period 2000-2016 includes all of the imputation methods.

M3: M1 with the variable Elections substituted by an indicator of political change (Elections of change: 1982, 1996, 2004, 2011 and 2015).

M4: M2 with the variable Elections substitued by an indicator of political change.

M5: WMAE of the past vote variable for the period 1982-2016.

Weightings. SD: Sociodemographic variables; PV0: past vote filtering NR/NA and did not vote;

PV1: past vote without filtering;

PV2; by past vote with imputed NR and NA;

PV3 by past vote NR, NA and without imputed age; PV+SD: for sociodemographic variables and PV3.

Treatment of PV (past vote): PV1 without imputation; PV2 imputed (NR): past vote imputed to those who did not respond;

PV3 imputed ( NR and NVM): past vote imputed to those that did not respond and that were not of voting age.

* p<0.05, ** p<0.01, *** p<0.01.

# 4. Article III: Datos administrativos agregados y estimación a partir de muestras no probabilísticas

Cabrera-Álvarez, Pablo. 2021. "Datos Administrativos Agregados y Estimación a Partir de Muestras No Probabilísticas." Revista Internacional de Sociología 79(1):e180. doi: 10.3989/ris.2021.79.1.19.350.

**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.

An English version of this paper can be found in appendix C.

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 fexibles 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 sociodemográicas, 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 sociodemográicas 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** | ficina del atastro | 2016 (anual) | uperficie 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. in 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 ( D+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 $\boldsymbol{A}$, que puede variar entre 0 y $\boldsymbol{+oo}$, y un parámetro $\boldsymbol{a}$ 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:

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

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 $\boldsymbol{\alpha}$ y $\boldsymbol{\lambda}$, necesarios para ajustar el modelo, se procedió a calcular diez validaciones cruzadas para seis valores diferentes de $\boldsymbol{\alpha}$. Por defecto, *glmnet* calcula el modelo para un conjunto de 100 valores de $\boldsymbol{\lambda}$. Los valores de $\boldsymbol{\alpha}$ y $\boldsymbol{\lambda}$ 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 $\boldsymbol{w}^{SP}$ 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 interés en 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 predi-cha en la población.

En la calibración asistida por modelo (Särndal y Lundström 2005; Wu y Sitter 2001), las distancias entre los pesos de diseño $\boldsymbol{d}_i$ y las ponderaciones finales $\boldsymbol{w}_i$ se obtienen minimizando la función $\boldsymbol{g}$ en la que $\boldsymbol{q}_i$ 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 $\boldsymbol{q}_i=1$ y que $\boldsymbol{g}$ corresponde a la distancia chi-cuadrado $g(w_i^{SP}, d_i) = \left(w_i^{SP} - d_i\right)^2/d_i$:

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

donde $\boldsymbol{d}$ son los pesos de diseño de la muestra, $\boldsymbol{D}$ 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 $\mathbf{T}^{\mathbf{M}} = (N, \sum_i^N \hat{y}_i)$.

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 $\boldsymbol{y}$ y $\hat{\boldsymbol{N}}$ 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 *J* 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 incon-veniente 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) pro-pone 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 $y_s$ 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 1 (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 ( D+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 gráico 1 presenta el promedio del sesgo relativo y los gráicos 2-14 representan las estimaciones en porcentaje sin ponderar y ponderadas para cada variable, junto con el valor de la estimación del EGM.*

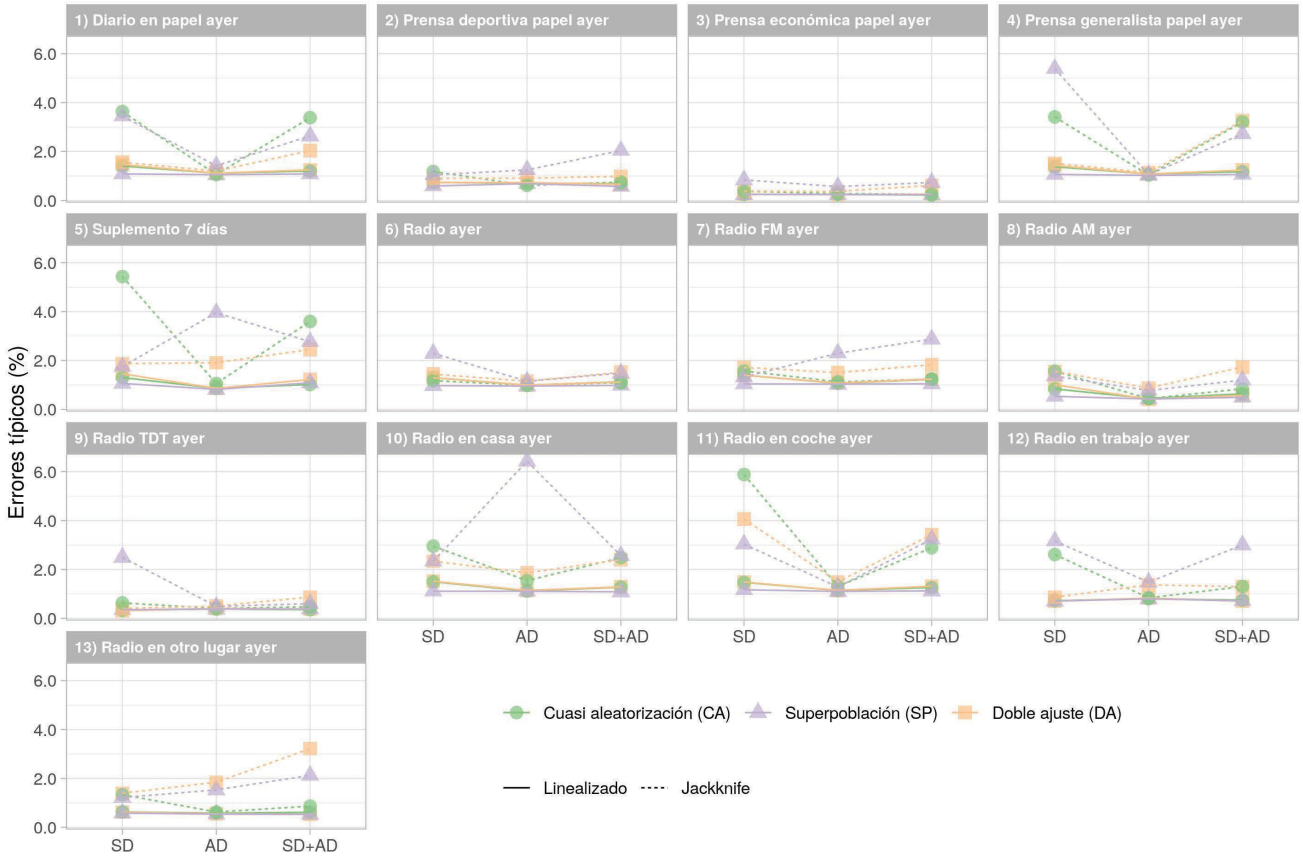

a la hora de reducir el sesgo de las estimaciones. Además, este efecto se acrecienta cuando se aplica el método de cuasi aleatorización, aunque solo supone una mejora de 2,2 puntos porcentuales sobre la estimación sin ponderar. Con respecto a la estimación, el efecto de emplear las variables auxiliares 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 refeja 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 . 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 beneits: 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.

# 5. Conclusions

This section presents the conclusions of the doctoral thesis alongside some reflections that emerged during the research. This thesis fits in a research line that seeks to understand the conditions that make inference possible and develop estimation methods. The thesis comprises three articles that seek to add new evidence to the cumulative process that has been developing for years on the inference from nonprobability samples and the correction of selection bias.

**Aggregate auxiliary data must correlate with the outcome and the grouping variables to effectively adjust for selection bias.**

The first research objective was to establish whether using aggregate data as contextual variables could effectively remove the bias from the estimates. The second was to determine under what circumstances aggregate administrative data effectively adjust the bias of the estimates. To this end, a series of statistical simulations were developed. The aggregate data was used in two ways: population totals and contextual variables that give information about the cluster where the sample unit belongs. The results show that for aggregate data as contextual variables to work, the target variable must be 1) grouped, 2) correlated with the auxiliary variable, and 3) the auxiliary variable must correlate with the selection probability. Only in such a scenario, aggregated data used as contextual variables effectively removes bias from the estimates. This finding partly confirms Biemer and Peytchev (2013) work using georeferenced census data to adjust a telephone survey based on a probability sample. They alluded to the need for auxiliary variables to be clustered to act as proxies for individual characteristics, to which these simulations add the need for the variable of interest to be clustered as well.

This finding questions the applicability of aggregate data used as contextual variables to adjust estimates in social sciences, mainly because the correlation between the target and grouping variables required is unusual. In this regard, Kish, Groves and Krotki (1976) analysed the level of clustering—the relationship between the target and the grouping variable—of several attitudinal and factual variables, concluding that the level of this

relationship is usually below the minimum necessary for estimates to be adjusted by using aggregate auxiliary variables. The simulations also show that the size of the clusters is irrelevant in determining the effectiveness of aggregate auxiliary variables, in line with the findings of Butt and Lahtinen (2016) in their work on the use of contextual variables to identify nonresponse bias in the UK European Social Survey. Yet this conclusion is only valid to the extent that the level of clustering of the variables does not change by, for example, reducing the size of the clusters (i.e. using a lower geographical level to aggregate the data).

In contrast to the problematic assumptions that aggregate data used as contextual variables need to meet, the simulations show that the correlation between the auxiliary variable and the target variable is sufficient if using aggregate information as population totals in a calibration model. However, in the simulations, this data cannot completely remove the bias from the estimates. The results of the simulations are subsequently corroborated in the applied cases discussed in the other two articles of the thesis.

**Aggregate data used as population totals fail to fully adjust the estimates even though they correlate with the probability of responding and the variable of interest.**

The essential requirement for an auxiliary variable to effectively adjust the bias of the estimates is that it can explain the selection mechanism with respect to the target variable. In other words, the auxiliary variable must be correlated with the probability of being part of the final sample and with the outcome variable (Särndal and Lundström 2005:110). Keeping this objective in mind, the second article of the thesis proposes using a set of model specifications to estimate voting intention from a series of pre and post-election surveys carried out by the Spanish *Centro de Investigaciones Sociológicas* (CIS). The CIS employs a combination of probability and nonprobability sampling for the selection of individuals. A sample of census tracts is drawn using multistage sampling with probability selection methods; households are selected using random paths and individuals within the household using sex and age quotas (Díaz de Rada 2005:70–78). The adjustments covered in this article include sociodemographic variables—sex and age, region, municipality size, education and employment—and different modalities of the past vote variable. The

population totals derived from administrative data were used to calibrate the sample with respect to these variables. In vote estimations, the use of past vote as an auxiliary variable has the advantage that it usually fulfils the twofold requirement of being correlated with the probability of being part of the sample and the target variable —voting intention.

The results show that the use of sociodemographic variables has a minimal effect in correcting the bias of the vote estimate, in line with other findings in the same field (Crespi 1988; Durand, Deslauriers, and Valois 2015). In contrast, using the past vote as an auxiliary variable in the calibration model improves the quality of the estimates, but this effect is not observed in all the years studied and in no case does it completely eliminate the bias in the estimates. This result is a further example that adjusting methods can help improve estimates, but they are far from guaranteeing an unbiased estimate (Cornesse *et al.* 2020). Also, in this paper, multiple imputation techniques were used to mitigate the effect of item nonresponse on the past vote variable, which generally presents some measurement and nonresponse errors (Crespi 1988). The multiple imputation failed to improve the quality of the auxiliary variable and the voting estimates.

These results also suggest that biases evolve and that the specification of selection models must evolve with them. The fact that the variable past vote, which correlates with voting intention, is sometimes ineffective in reducing the bias of the estimates indicates that the nature of the bias may change from one election to the next, and the auxiliary variables in the estimation model must take this into account. Jowell and colleagues (1993), in their discussion of the role of the variables used in opinion poll quotas in the UK, point to the lack of effort made to specify a theoretical model to support the use of quotas in each case. The quotas are generally limited to groups of sex, age and social status. Similarly, the selection of auxiliary variables in the adjustments has been guided by experience and good practices (Bethlehem *et al.* 2011:248). In the near future, where the amount of auxiliary information will increase significantly, beyond developing methods that allow the selection of auxiliary variables, it seems advisable to rethink the place of theories that explain the causes of the phenomena to be estimated and their relationship with the selection bias.

**The effectiveness of combining weighting with multiple imputation to deal with selection bias depends on the specification of the models.**

The imputation models carried out to tackle item nonresponse at the auxiliary variable past vote and voting intention must be correctly specified. Although the imputation had little effect on the quality of the voting intention estimates, the only exception—2004 elections—is in the model that includes additional attitudinal predictors, compared to the basic set of sociodemographic variables. This result reinforces the idea of the relevance of the auxiliary variables selected for the models, even over and above the method used. In this research, the choice of auxiliary information was limited by using all pre and post-election surveys that cover the period from 1982 to 2016, which have only a few common variables. In another scenario, in which a larger number of attitudinal variables could be added to the models, the result of combining weighting with imputation models could be improved.

**Aggregate administrative data at the municipality level and used as auxiliary—contextual—variables do not perform well in the adjustment models to remove selection bias from the estimates.**

The third article used aggregated administrative data to correct two surveys from a panel of internet users, the AIMC-Q panel, devoted to measuring aspects related to media consumption. The panel surveys generally employ an adjustment using sociodemographic variables—age, sex, municipality size and region—collected in the survey and the population totals of those variables in a calibration. In addition to this approach, this research tested the feasibility of using administrative variables aggregated at the municipality level to adjust radio and press consumption survey estimates. The administrative variables covered different domains such as population registers, unemployment, income data, tax records, car registration, and election results.

The results of the analysis show, as the simulations of chapter 2, that using aggregate administrative variables has little effect in reducing the bias of the estimates. These results align with other research that addressed the utility of aggregate census and

administrative data to tackle nonresponse in high-quality probability surveys (Biemer and Peytchev 2013; Butt and Lahtinen 2016). One possible explanation for the lack of effect of the aggregate variables is that they are not correlated with the target variables, a set of press and radio consumption measures. However, taking into account the results of the simulations, we can conclude that the aggregate nature of the variables prevents them from effectively reducing the bias of the estimates.

**The different estimation methods achieve similar results using the same vector of auxiliary variables, which is also the case for the standard errors.**

Another issue addressed in this thesis is the performance of the different models against the set of auxiliary variables used to fit them to reduce the level of selection bias in the estimates. In the third article of the thesis, three methods were used to estimate from two nonprobability samples: quasi-randomisation, superpopulation models using a model-assisted calibration and a combination of both. As described in the previous paragraph, the adjustments included sociodemographic variables and a set of administrative variables aggregated at the municipality level.

Although the auxiliary variables did little to reduce the bias of the estimates, the variables in the model were more relevant than the method employed for most of the estimates. This finding aligns with Mercer, Lau and Kennedy (2018), who compared a set of methods and variables to adjust the estimates from a nonprobability panel, concluding that the auxiliary variables had a more prominent role. This observation is essential because the focus in research has been on estimation models versus the potential of auxiliary variables. Moreover, this is not only about the auxiliary variables but the theoretical models that inform their selection.

Another research question was about the impact of the administrative aggregate variables used as contextual variables on the standard errors of the estimates. Sometimes, when the auxiliary variables are not related to the probability of selection and the target variable, survey weights have a negative effect on the variance of the estimates (Little and Vartivarian 2005). In this case, the use of the aggregate administrative variables reduces

the variance of the estimates. However, this effect stems from the little variation of the weights based on contextual variables derived from aggregate data.

**The Total Survey Error (TSE) framework is not suitable for analysing inference from nonprobability samples.**

Although it is not a primary objective of the thesis, it is appropriate to address the role played by the TSE and the conclusions reached during the research. The TSE is the predominant framework for analysing the quality of survey estimates (Biemer 2010; Biemer and Lyberg 2003). This framework systematically captures all possible sources of error that can bias an estimate, including, on its representativeness side, sampling, coverage, nonresponse and adjustment errors. Using the TSE to identify the source of bias affecting estimates from nonprobability samples may be attractive. Not surprisingly, the article presented in chapter 3 had the secondary objective of adapting the TSE to explain the sources of error in the vote estimates made from the CIS surveys. Following the TSE approach, the bias in the estimates was due to noncoverage and nonresponse errors.

Still, the use of quotas to select the respondents and the high degree of interviewers discretion in this last step (Díaz de Rada 2005) make the TSE framework insufficient to describe the selection process precisely in the CIS surveys. Mercer (2018) provides a critical review of the TSE to address the quality of estimates from nonprobability samples—opt-in panels—, which in the course of this research served to rethink the theoretical framework. Applied to the quota sampling of the CIS surveys, two aspects emerge that disconnect the TSE from the analysis of the quality of the estimates. First, the coverage and nonresponse bias of the TSE omits the process of arbitrary selection by the interviewer that occurs when applying quotas. For instance, it does not account for the situation where two individuals in the household fit the quota. Similarly, this framework does not take into account the process of household substitution. Secondly, the TSE does not focus on model-based estimation but on eliminating errors during data collection. This strategy for minimising estimation error does not apply to nonprobability samples, which always need a model that explains the process of data generation in the population to enable estimation.

It can be concluded that the TSE is a valuable framework for the analysis of the quality of estimates in probability surveys, and some of the errors that it contemplates, for example coverage error, constitute valid categories for analysing estimates made from nonprobability samples. Yet the TSE is insufficient to explain the complexity of the selection and response processes in nonprobability surveys.

**All theoretical developments start and end at the model specification.**

One of the main criticisms of inference from nonprobability samples stems from the lack of a coherent theoretical framework to back the generalisation of the survey estimates (Baker *et al.* 2013). Recent years have seen developments that seek to address this lack of a theoretical underpinning to the inference process. On the one hand, Valliant and Elliott (2017) have proposed using the previous work on selection mechanisms and how they affect survey estimates to generate a consistent framework for estimation from nonprobability samples. On the other hand, Mercer and his colleagues (2017) have proposed transferring the framework used for the analysis of causal inference to inference from nonprobability samples. Both contributions share the need, through a model, to control the selection process with respect to the outcome variable.

Although these proposals aim to create a specific framework, different from the TSE, for the use with nonprobability samples, they fail to dispel the main criticism of inference from nonprobability samples, the need to rely on a specific model for each estimate. In essence, these theoretical developments do not lead to new conclusions, but rather synthesise the conditions and form that models must take to be effective. The theoretical framework that defines the possibility of inference from nonprobability samples is model-based inference, and in any case, requires some assumptions about the process of data generation in the population. The need for implicit or explicit models is common to both proposed frameworks.

Another criticism of nonprobability samples is that it is impossible to know in which situations the models work —manage to control selection bias. This criticism is not dispelled when using the new theoretical frameworks, improved adjustment methods or

having a wider variety of data sources to specify the models. In each case, a model is needed, and most of the time, it will not be possible to be confident whether it allows for producing accurate population estimates. In other words, the possibility of inferring from models relies on assumptions about the process of data generation in the population that, in most cases, are not testable.

**From the optimism of the big data to the return to theory.**

The beginning of this research was marked by a moment of optimism justified by technological expansion and the possibility of accessing and processing large amounts of information. The initial idea of this research was born out of an interest in finding synergies that could provide new solutions to old needs. This research sought precisely to use a part of the information that is now accessible—aggregate administrative data—to improve the quality of estimates made from nonprobability samples.

This optimism led me to exhaustively track public data sources containing information at the municipality or lower levels and systematise the collection and curation of data to adjust estimates from nonprobability samples. The conclusion with which I close this thesis is far from that initial optimism. There is no doubt that the greater availability of data sources will help improve the quality of survey estimates and other aspects of survey research. Proofs of this are the numerous synergies being developed in many areas of survey research, including hybrid estimation (Hill *et al.* 2019). But it is also true that the abundance of data does not replace the need to understand the phenomena under study in order to be effective in, for example, making adjustments to surveys.

In recent years, some voices have demanded, in an environment dominated by an abundance of data, to look again to theory, also to specify adjustment models. Peytchev, Presser and Zhang (2018), using the US General Social Survey—a probability survey—proposed to get back to theory to choose auxiliary variables for adjustments. Their paper shows how using past vote and volunteering improves the nonresponse bias of some survey estimates. There are many considerations to make about this more traditional approach to weighting; we need to keep experimenting with new data sources, but the look for and use

of new auxiliary variables needs to be companied by theory. Probably, giving theory a more central role in thinking about the specification of estimation models can help to select better predictors. Probably, the future is not in analysing large volumes of data but in using the wider variety of available data sources to identify the variables that require the adjustment model. Undoubtedly, the future lies in keeping investigating the mechanisms that enable inference from nonprobability samples. The key is to find the right balance between the role of theory and the opportunities arising from the abundance of data and data-driven approaches.

## 5.1 Implications and future research

The most relevant methodological implication of this thesis is to discourage the use of aggregate administrative data as contextual variables in the adjustments of the selection bias in nonprobability samples. Even though the case studied in this thesis corresponds to a specific domain—surveys measuring media audience—the results coincide with the conclusions drawn from the statistical simulations, aggregate administrative data as contextual variables shows poor performance in correcting selection bias. For aggregate data as contextual variables to work in survey adjustments, the level of clustering of the target variable must reach a magnitude not expected in the framework of most social surveys. Moreover, this conclusion reinforces existing evidence from high-quality probability surveys where the census and administrative data were used to tackle nonresponse bias, such as the National Comorbidity Survey Replication (Biemer and Peytchev 2013) and the European Social Survey (Butt and Lahtinen 2016). This evidence altogether is sufficient to conclude that the low or null effect of using aggregate administrative data to correct survey estimates stems from the aggregate nature of the variables. It may be possible in the future to test the usefulness of such auxiliary variables in the framework of surveys where some target variables exhibit higher levels of clustering, such as school surveys. However, such surveys are a minority, and administrative aggregate data availability is not guaranteed at all levels (e.g. classroom or school).

The second implication of this work concerns the potential of administrative aggregate data as population totals to reduce the bias of the estimates. The results of both the

simulations and the article analysing the use of different vectors of auxiliary variables to correct for bias in vote estimates show the potential of this type of auxiliary data. This is not a finding attributable to this research since the use of aggregate data as population totals is widely implemented in both industry and official statistics (Bethlehem *et al.* 2011:9). However, the auxiliary data to be effective needs to rely on an accurate model that explains the selection into the sample with regards to the target variable, which connects with the initiative to use theoretical models in order to implement more specific adjustments (Peytchev et al. 2018). This improvement in model specification is possible thanks to the greater availability of auxiliary information.

Finally, although the central question of this thesis has been answered, there are still many open questions to explore related to the inference from nonprobability samples. On the theoretical level, the fit for purpose framework needs further elaboration. Although there is some consensus on the need to rule out nonprobability samples when accurate population estimates are needed (Baker et al. 2013), some research questions can benefit from data collected from nonprobability samples. More research is needed to address the circumstances that allow for the use of nonprobability samples for inference. Also, at the theoretical level, the uniqueness of the selection methods that fall under the umbrella of nonprobability sampling needs further exploration. The differences between these methods recommend a careful and more precise treatment of each of them with respect to the inference process. At the empirical level, some evidence has shown that selection bias affects differently univariate, bivariate and multivariate estimators (Pasek 2016). The effect of survey adjustments in bivariate and multivariate analyses also calls for further research. Another empirical issue that requires more research effort concerns selecting the adjustment method depending on the characteristics of the sampling. Although some comparisons between adjustment methods (e.g. Cornesse et al. 2020; Ferri-García and Rueda 2018; Valliant 2020), these simulations are far from exhaustive. More adjustment methods and estimation contexts need to be considered.

# 6. References


Baker, Reginald P. 2017. “Big Data.” Pp. 47–69 in *Total Survey Error in Practice*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Baker, Reginald P., Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, and Paul J. Lavrakas. 2010. “AAPOR Report on Online Panels.” *Public Opinion Quarterly* 74(4):711–81. doi: 10.1093/poq/nfq048.

Baker, Reginald P., J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau. 2013. “Summary Report of the Aapor Task Force on Non-Probability Sampling.” *Journal of Survey Statistics and Methodology* 1(2):90–105. doi: 10.1093/jssam/smt008.

Berinsky, Adam J. 2006. “American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data.” *Public Opinion Quarterly* 70(4):499–529.

Bethlehem, Jelke. 2016. “Solving the Nonresponse Problem With Sample Matching?” *Social Science Computer Review* 34(1):59–77. doi: 10.1177/0894439315573926.

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Biemer, Paul P. 2010. “Total Survey Error: Design, Implementation, and Evaluation.” *Public Opinion Quarterly* 74(5):817–48. doi: 10.1093/poq/nfq058.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley & Sons.

Biemer, Paul P., and Andy Peytchev. 2012. “Census Geocoding for Nonresponse Bias Evaluation in Telephone Surveys.” *Public Opinion Quarterly* 76(3):432–52. doi: 10.1093/poq/nfs035.

Biemer, Paul P., and Andy Peytchev. 2013. “Using Geocoded Census Data for Nonresponse Bias Correction: An Assessment.” *Journal of Survey Statistics and Methodology* 1(1):24–44. doi: 10.1093/jssam/smt003.

Börsch-Supan, Axel, Detlev Elsner, Heino Faßbender, Rainer Kiefer, Daniel Mcfadden, and Joachim Winter. 2004. “How to Make Internet Surveys Representative: A Case Study of a Two-Step Weighting Procedure.” *Publication Series of the MPI for Social Policy* 067–04.

Bowley, Arthur L. 1906. “Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science.” *Journal of the Royal Statistical Society* 69(3):540–58.

Bowley, Arthut L. 1913. “Working-Class Households in Reading.” *Journal of the Royal Statistical Society* 76(7):672–701.

Brewer, Ken. 2013. “Three Controversies in the History of Survey Sampling.” *Survey*

*Methodology* 39(2):249–62.

Brick, J. Michael, and Douglas Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *Annals of the American Academy of Political and Social Science* 645(1):36–59. doi: 10.1177/0002716212456834.

Bush, Alan J., and Joseph F. Hair. 1985. "An Assessment of the Mall Intercept as a Data Collection Method." *Journal of Marketing Research* 22(2):158–67. doi: 10.1177/002224378502200205.

Butt, Sarah, and Kaisa Lahtinen. 2016. *ADDResponse : Auxiliary Data Driven NonResponse Bias Analysis Technical Report on Appending Geocoded Auxiliary Data to Round 6 of European Social Survey (UK)*. London.

Butt, Sarah, Kaisa Lahtinen, and Rory Fitzgerald. 2015. *Using Geocoded Auxiliary Data to Predict Nonresponse in Address-Based Samples: Are Household- Level Commercial Data Any Better than Aggregate-Level Census Data?* London.

Buttice, Matthew K., and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21(4):449–67. doi: 10.1093/pan/mpt017.

Callegaro, Mario, Reg P. Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas. 2014. *Online Panel Research: A Data Quality Perspective*. New York: Wiley.

Callegaro, Mario, Katja L. Manfreda, and Vasja Vehovar. 2015. *Web Survey Methodology*. SAGE Publications.

Callegaro, Mario, and Giancarlo Gasperoni. 2008. "Accuracy of Pre-Election Polls for the 2006 Italian Parliamentary Election: Too Close to Call." *International Journal of Public Opinion Research* 20(2):148–70. doi: 10.1093/ijpor/edn015.

Chen, Jack Kuang Richard L. Valliant, and Michael R. Elliott. 2019. "Calibrating Non-Probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling." *Journal of the Royal Statistical Society. Series C: Applied Statistics* 68(3). doi: 10.1111/rssc.12327.

Chen, Jack Kuang, Richard L. Valliant, and Michael R. Elliott. 2018. "Model-Assisted Calibration of Non-Probability Sample Survey Data Using Adaptive LASSO." *Survey Methodology* 44(1): 117–144.

Cochran, William G. 1971. Técnicas de Muestreo. Mexico: Continental.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research* 59:1–12. doi: 10.1016/j.ssresearch.2016.04.015.

Converse, Jean M. 2009. Survey Research in the United States: Roots and Emergence 1890-1960. *Transaction Publishers*.

Cornesse, Carina, Annelies G. Blom, David Dutwin, Jon A. Krosnick, Edith D. De Leeuw,

Stéphane Legleye, Josh Pasek, Darren Pennay, Benjamin Phillips, Joseph W. Sakshaug, Bella Struminskaya, and Alexander Wenz. 2020. "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." *Journal of Survey Statistics and Methodology* 8(1):4–36. doi: 10.1093/jssam/smz041.

Couper, Mick P. 2000. "Web Surveys." *Public Opinion Quarterly* 64(4):464–94. doi: 10.1086/318641.

Couper, Mick P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3):145–56.

Crespi, Irving. 1988. *Pre-Election Polling: Sources of Accuracy and Error*. Russell Sage Foundation.

Dassonneville, Ruth, André Blais, Marc Hooghe, and Kris Deschouwer. 2020. "The Effects of Survey Mode and Sampling in Belgian Election Studies: A Comparison of a National Probability Face-to-Face Survey and a Nonprobability Internet Survey." *Acta Politica* 55(2):175–98. doi: 10.1057/s41269-018-0110-4.

Dever, Jill, Ann Rafferty, and Richard L. Valliant. 2008. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods* 2(2):47–60. doi: 10.18148/srm/2008.v2i2.128.

Deville, Jean Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–82. doi: 10.1080/01621459.1992.10475217.

Díaz de Rada, Vida, Juan Antonio Domínguez, and Sara Pasadas. 2019. *Internet Como Modo de Administración de Encuestas*. Madrid: Centro de Investigaciones Sociológicas.

Díaz de Rada, Vidal. 2005. *Manual de Trabajo de Campo de La Encuesta*. Madrid: Centro de Investigaciones Sociológicas.

Díaz de Rada, Vidal. 2013. "La No Respuesta En Encuestas Presenciales Realizadas En España." *Revista Internacional de Sociología* 71(2):357–81. doi: 10.3989/ris.2012.02.07.

Díaz de Rada, Vidal, and Valentín Martínez. 2020. "Household Sampling Designs: Differences and Similarities between Probability Sampling and Route and Quota Sampling." *Revista Española de Investigaciones Sociológicas* 171:23–42. doi: 10.5477/cis/reis.171.23.

Disogra, Charles, Curtiss Cobb, Elisa Chan, and J. Michael Dennis. 2011. "Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics." *JSM proceedings* 4501–15.

Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47(6):615–30. doi: 10.1177/147078530504700602.

Durand, Claire, Melanie Deslauriers, and Isabelle Valois. 2015. "Should Recall of Previous Votes Be Used to Adjust Estimates of Voting Intention?" *Survey Insights: Methods from the Field* 1–14. doi: 10.13094/SMIF-2015-00002.

Efron, Bradley. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for industrial and applied mathematics.

Elliott, Michael R. 2009. "Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights." *Survey Practice* 2(6):1–7. doi: 10.29115/sp-2009-0025.

Elliott, Michael R., and Richard L. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32(2):249–64. doi: 10.1214/16-STS598.

ESOMAR. 2017. *Global Market Research 2017*. Amsterdam.

Eurostat. 2019. "Digital Economy and Society Statistics - Households and Individuals - Statistics Explained." Retrieved July 5, 2021 (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals).

Fahimi, Mansour, Frances M. Barlas, Randall K. Thomas, and Nicole Buttermore. 2015. "Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse." *Survey Practice* 8(6):1–11. doi: 10.29115/sp-2015-0031.

Ferri-García, Ramón, and María del Mar Rueda. 2018. "Efficiency of Propensity Score Adjustment and Calibration on the Estimation from Non-Probabilistic Online Surveys." *SORT* 42(November):159–82. doi: 10.2436/20.8080.02.73.

Ferri-García, Ramón, and María del Mar Rueda. 2020. "Propensity Score Adjustment Using Machine Learning Classification Algorithms to Control Selection Bias in Online Surveys." *PLoS ONE* 15(4):1–19. doi: 10.1371/journal.pone.0231500.

Forsyth, John, and Leah Boucher. 2015. "Why Big Data Is Not Enough." *Research World* 2015(50):26–27. doi: 10.1002/rwm3.20187.

Gallup, George. 1944. *A Guide to Public Opinion Polls*. Princeton, NJ, US: Princeton University Press.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22(2):153–64. doi: 10.1214/088342306000000691.

Gini, Corrado. 1928. "Une Application de La Méthode Représentative Aux Matériaux Du Dernier Recensement de La Population Italiene." *Bulletin of the International Statistical Institute* 23(2):198–215.

Gini, Corrado, and Luigi Galvani. 1929. "Di Una Applicazione Del Metodo Rappresentativo All'ultimo Censimento Italiano Della Popolazione." *Annli Di Statistica Series* 6(4):1–107.

Gittelman, Steven H., Randall K. Thomas, Paul J. Lavrakas, and Victor Lange. 2015. "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples." *Journal of Advertising Research* 55(4):368–79. doi:

10.2501/JAR-2015-020.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5):646–75. doi: 10.1093/poq/nfl033.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75(5 SPEC. ISSUE):861–71. doi: 10.1093/poq/nfr057.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2013. *Survey Methodology* (Google EBook).

Hand, David J. 2018. "Statistical Challenges of Administrative and Transaction Data." Journal of the Royal Statistical Society. *Series A: Statistics in Society* 181(3):555–605. doi: 10.1111/rssa.12315.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1953. *Sample Survey Methods and Theory. Vol. I. Methods and Applications*. Oxford, England: John Wiley.

Heckathorn, Douglas D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44(2):174–99. doi: 10.2307/3096941.

Hill, Craig A., Paul Biemer, Trent Buskirk, Mario Callegaro, Ana Lucía Córdova Cazar, Adam Eck, Lilli Japec, Antje Kirchner, Stas Kolenikov, Lars Lyberg, and Patrick Sturgis. 2019. "Exploring New Statistical Frontiers at the Intersection of Survey Science and Big Data: Convergence at 'BigSurv18.'" *Survey Research Methods* 13(1):123–34. doi: 10.18148/srm/2019.v13i1.7467.

Hsieh, Yuli Patrick, and Joe Murphy. 2017. "Total Twitter Error." *Total Survey Error in Practice* 23–46. doi: 10.1002/9781119041702.ch2.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. 2015. "Big data in survey research." *Public Opinion Quarterly* 79(4):839-880.

Jowell, Roger, Barry Hedges, Peter Lynn, Graham Farrant, and Anthony Heath. 1993. "Review: The 1992 British Election: The Failure of the Polls." *Public Opinion Quarterly* 57(2):238. doi: 10.1086/269369.

Kalton, Graham. 2019. "Developments in Survey Research over the Past 60 Years: A Personal Perspective." *International Statistical Review* 87(S1):S10–30. doi: 10.1111/insr.12287.

Kalton, Graham, and Dallas W. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society. Series A (General)* 149(1):65. doi: 10.2307/2981886.

Keeter, Scott, Nick Hatley, Courtney Kennedy, and Arnold Lau. 2017. "What Low Response Rates Mean for Telephone Surveys." *Pew Research Center* 1–39.

Kiaer, Anders N. 1897. *The Representative Method of Statistical Surveys*. Translation.

Oslo: Norwegian Central Bureau of Statistics.

Kim, Jae-kwang, and Siu-Ming Tam. 2020. “Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference.” *International Statistical Review* 1–30.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Kish, Leslie, Robert M. Groves, K. P. Krotki 1976. “Sampling Errors for Fertility Surveys.” *Occasional Papers* 17:1–61.

Kitchin, Rob. 2014. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data and Society* 1(1):1–12. doi: 10.1177/2053951714528481.

Kohler, Ulrich, Frauke Kreuter, and Elizabeth A. Stuart. 2019. “Nonprobability Sampling and Causal Analysis.” *Annual Review of Statistics and Its Application* 6:149–72. doi: 10.1146/annurev-statistics-030718-104951.

Koop, J. C. 1974. “Notes for a Unified Theory of Estimation for Sample Surveys Taking into Account Response Erros.” *Metrika* 21(1):19–39. doi: 10.1007/BF01893890.

Kott, Phillip S. 2006. “Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors.” *Survey Methodology* 32(2):133–42.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivellore E. Raghunathan. 2010. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 173(2):389–407. doi: 10.1111/j.1467-985X.2009.00621.x.

Kreuter, Frauke. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley & Sons.

Kreuter, Frauke, and Kristen Olson. 2011. “Multiple Auxiliary Variables in Nonresponse Adjustment.” *Sociological Methods and Research* 40(2):311–32. doi: 10.1177/0049124111400042.

Kruskal, William, and Frederick Mosteller. 1979. “Representative Sampling, I: Non-Scientific Literature.” *International Statistical Review* 47(2):13–24.

Kruskal, William, Frederick Mosteller. 1979. “Representative Sampling, II: Scientific Literature, Excluding Statistics Excluding Statistics.” *International Statistical Review* 47(2):111–27.

Kruskal, William, Frederick Mosteller. 1979. “Representative Sampling, III: The Current Statistical Literature.” *International Statistical Review* 47(3):245–65.

Kruskal, William, and Frederick Mosteller. 1980. “Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939.” *International Statistical Review* 48(2):169–95.

Künn, Steffen. 2015. “The Challenges of Linking Survey and Administrative Data*.” IZA*

*World of Labor* 1–10. doi: 10.15185/izawol.214.

Lahtinen, Kaisa, and Sarah Butt. 2015. *Using Auxiliary Data to Model Nonresponse Bias The Challenge of Knowing Too Much about Nonrespondents Rather than Too Little?* London.

Laney, Doug. 2001. "META Delta." *Application Delivery Strategies* 949(February 2001):4. doi: 10.1016/j.infsof.2008.09.005.

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2):329–49.

Lee, Sunghee, and Richard L. Valliant. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research* 37(3):319–43.

de Leeuw, Edith D., Joop J. Hox, and Annemieke Luiten. 2018. "International Nonresponse Trends across Countries and Years: An Analysis of 36 Years of Labour Force Survey Data." *Survey Methods: Insights from the Field* 1–11. doi: 10.13094/SMIF-2018-00008.

Lie, Einar. 2002. "The Rise and Fall of Sampling Surveys in Norway, 1875-1906." *Science in Context* 15(3):385–409.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22 140:55.

Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. Vol. Second.

Little, Roderick J. A., and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2):161–68.

Little, Roderick J. A. 2004. "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99(466):546–56. doi: 10.1198/016214504000000467.

Little, Roderick J. A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77(378):237–50.

Lohr, Sharon L. 2017. "Comments on the Rao and Fuller (2017) Paper." *Survey Methodology* 43(2):173–78.

Lohr, Sharon L., and J. Michael Brick. 2017. "Roosevelt Predicted to Win: Revisiting the 1936 Literary Digest Poll." *Statistics, Politics and Policy* 8(1). doi: 10.1515/spp-2016-0006.

Loosveldt, Geert, and Nathalie Sonck. 2008. "An Evaluation of the Weighting Procedures for an Online Access Panel Survey." *Survey Research Methods* 2(2):93–105. doi: 10.18148/srm/2008.v2i2.82.

Lusinchi, Dominic. 2012. "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" *Social Science History* 36(1):23–54.

doi: 10.1215/01455532-1461650.

Lusinchi, Dominic. 2017. "The Rhetorical Use of Random Sampling: Crafting and Communicating the Public Image of Polls As a Science (1935–1948)." *Journal of the History of the Behavioral Sciences* 53(2):113–32. doi: 10.1002/jhbs.21836.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. *For Weighting Online Opt-In Samples, What Matters Most?* Washington DC: Pew Research Center.

Mercer, Andrew W., Frauke Kreuter, Scott Keeter, and Elizabeth A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys." *Public Opinion Quarterly* 81:250–79. doi: 10.1093/poq/nfw060.

Mercer, Andrew William. 2018. *Selection Bias in Nonprobability Surveys: A Causal Inference Approach*. University of Maryland.

Miller, Peter V. 2017. "Is There a Future for Surveys?" *Public Opinion Quarterly* 81:205–12. doi: 10.1093/poq/nfx008.

Moon, Nick. 1999. *Opinion Polls. History, Theory and Practice*. Manchester: Manchester University Press.

Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, and David B. Truman. 1949. *The Pre-Election Polls of 1948*. New York: Social Sciences Research Council.

Negroponte, Nicholas. 1995. *Being Digital*.

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. doi: 10.2307/2342192.

Neyman, Jerzy. 1952. *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, D.C.: Graduate School, U.S. Department of Agriculture.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–85. doi: 10.1093/pan/mph024.

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. doi: 10.1093/ijpor/edv016.

Pasek, Josh, S. Mo Jang, Curtiss L. Cobb, J. Michael Dennis, and Charles Disogra. 2014. "Can Marketing Data Aid Survey Research? Examining Accuracy and Completeness in Consumer-File Data." *Public Opinion Quarterly* 78(4):889–916. doi: 10.1093/poq/nfu043.

Pavía, José M., and Beatriz Larraz. 2012. "Nonresponse Bias and Superpopulation Models in Electoral Polls." *Revista Española de Investigaciones Sociológicas* 237–64. doi: 10.5477/cis/reis.137.237.

Pavía, José M. 2005. "Forecasts from Nonrandom Samples: The Election Night Case." *Journal of the American Statistical Association* 100(472):1113–22. doi: 10.1198/016214504000001835.

Pedraza, Pablo, Kea Tijdens, Rafael Muñoz de Bustillo, and Stephanie Steinmetz. 2010. "A Spanish Continuous Volunteer Web Survey: Sample Bias, Weighting and Efficiency." *Revista Española de Investigaciones Sociológicas* 131(1):109–30.

Peytchev, Andy, Stanley Presser, and Mengmeng Zhang. 2018. "Improving Traditional Nonresponse Bias Adjustments: Combining Statistical Propertieswith Social Theory." *Journal of Survey Statistics and Methodology* 6(4):491–515. doi: 10.1093/jssam/smx035.

Playford, Christopher J., Vernon Gayle, Roxanne Connelly, and Alasdair JG J. G. Gray. 2016. "Administrative Social Science Data: The Challenge of Reproducible Research." *Big Data and Society* 3(2):1–13. doi: 10.1177/2053951716684143.

Prewitt, Kenneth. 2013. "The 2012 Morris Hansen Lecture: Thank You Morris, et Al., for Westat, et Al." *Journal of Official Statistics* 29(2):223–31. doi: 10.2478/jos-2013-0018.

Rafei, Ali, Carol A. C. Flannagan, and Michael R. Elliott. 2020. "Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees." *Journal of Survey Statistics and Methodology* 8(1):148–80. doi: 10.1093/jssam/smz060.

Rao, J. N. K., and Wayne A. Fuller. 2017. "Sample Survey Theory and Methods: Past, Present, and Future Directions." *Survey Methodology* 43(2):145–60.

Riba, Clara, Mariano Torcal, and Laura Morales. 2010. "Estrategias Para Aumentar La Tasa de Respuesta y Los Resultados de La Encuesta Social Europea En España." *Revista Internacional de Sociología* 68(3):603–35. doi: 10.3989/ris.2008.12.17.

Rivers, Douglas. 2007. "Sampling for Web Surveys." in *Joint Statistical Meeting*.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Royall, Richard M. 1970. "On Finite Population Sampling Theory under Certain Linear Regression Models." *Biometrika* 57(2):377–88.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3):581–92.

Salganik, Matthew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34(1):193–240. doi: 10.1111/j.0081-1750.2004.00152.x.

Särndal, Carl-Erik. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33(2):99–119.

Särndal, Carl-Erik. 1978. “Design-Based and Model-Based Inference in Survey Sampling.” *Scandinavian Journal of Statistics* 5(1):27–52.

Särndal, Carl-Erik. 2010. “Models in Survey Sampling.” S*tatistics in Transition New Series* 11(3):112–27.

Särndal, Carl-Erik, and Sixten Lundström. 2008. “Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator.” *Journal of Official Statistics* 24(2):167–91.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Särndal, Carl-Erik, and Sixten Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley.

Savage, Mike, and Roger Burrows. 2007. “The Coming Crisis of Empirical Sociology.” *Sociology* 41(5):885–99. doi: 10.1177/0038038507080443.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. 2016. “Social Media Analyses for Social Measurement.” *Public Opinion Quarterly* 80(1):180–211. doi: 10.1093/poq/nfv048.

Schonlau, Matthias, and Mick P. Couper. 2017. “Options for Conducting Web Surveys.” *Statistical Science* 32(2):279–92. doi: 10.1214/16-STS597.

Schonlau, Matthias, Arthur van Soest, and Arie Kapteyn. 2007. “Are ‘Webographic’ or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?” *Survey Research Methods* 1(3):155–63. doi: 10.2139/ssrn.1006108.

Schonlau, Matthias, Arthur Van Soest, Arie Kapteyn, and Mick Couper. 2009. “Selection Bias in Web Surveys and the Use of Propensity Scores.” *Sociological Methods and Research* 37(3):291–318. doi: 10.1177/0049124108327128.

Schonlau, Matthias, Kinga Zapert, Lisa Payne Simon, Katherine Haynes Sanstad, Sue M. Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra H. Berry. 2004. “A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey.” *Social Science Computer Review* 22(1):128–38. doi: 10.1177/0894439303256551.

Schouten, Barry. 2007. “A Selection Strategy for Weighting Variables under a Not-Missing-at-Random Assumption.” *Journal of Official Statistics* 23(1):51–68.

Singer, Eleanor. 2016. “Reflections on Surveys’ Past and Future.” *Journal of Survey Statistics and Methodology* 4(4):463–175. doi: 10.1093/jssam/smw026.

Smith, T. M. F. 1976. “The Foundations of Survey Sampling: A Review.” *Journal of the Royal Statistical Society. Series A (General)* 139(2):183. doi: 10.2307/2345174.

Smith, T. M. F. 1979. “Statistical Sampling in Auditing : A Statistician ’ s Viewpoint.” *Journal of the Royal Statistical Society . Series D (The Statistician)* 28(4):267–80.

Smith, T. M. F. 1983. “On the Validity of Inferences from Non-Random Sample.” *Journal*

*of the Royal Statistical Society. Series A (General)* 146(4):394. doi: 10.2307/2981454.

Smith, Tom W. and Jibum Kim. 2013. "An Assessment of the Multi-Level Integrated Database Approach." *The Annals of the American Academy of Political and Social Science* 645:185–221.

Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed." *Public Opinion Quarterly* 52(1):125–33. doi: 10.1086/269085.

Statistics Canada. 2017. *Statistics Canada's Quality Assurance Framework*. Ottawa.

Sturgis, Patrick, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Jouni Kuha, Ben Lauderdale, and Patten Smith. 2016. *Report of the Inquiry into the 2015 British General Election Opinion Polls*. London: National Center for Research Methods.

Sudman, Seymour. 1966. "Probability Sampling with Quotas." *Journal of the American Statistical Association* 61(315):749–71. doi: 10.1080/01621459.1966.10480903.

Taylor, Humphrey. 2000. "Does Internet Research Work?" *International Journal of Market Research* 42(1):1–11. doi: 10.1177/147078530004200104.

Terhanian, George, John Bremer, Renee Smith, and Randy Thomas. 2000. "Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment." *Harris Interactive White Paper* 1–13.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. "The Science of Web Surveys." *The Science of Web Surveys*. doi: 10.1093/acprof:oso/9780199747047.001.0001.

Valliant, Richard L., Alan H. Dorfman, and Richard M. Royall. 2000. *Finite Population Sampling And Inference: A Prediction Approach*. New York: Wiley.

Valliant, Richard L. 2020. "Comparing Alternatives for Estimation from Nonprobability Samples." *Journal of Survey Statistics and Methodology* 8:231–63. doi: 10.1093/jssam/smz003.

Valliant, Richard L., and Jill A. Dever. 2011. Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research* 40(1):105–37.

Valliant, Richard L., Jill A. Dever, and Frauke Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. Cham: Springer International Publishing.

Vavreck, Lynn, and Douglas Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion and Parties* 18(4):355–66. doi: 10.1080/17457280802305177.

Vehovar, Vasja, Zenel Batagelj, and Katja Manfreda. 1999. "Web Surveys: Can the Weighting Solve the Problem?" *Proceedings of the Section on Survey Research Methods* 962–67.

Wagner, James, Richard L. Valliant, Frost Hubbard, and Li Charley Jiang. 2014. "Level-of-Effort Paradata and Nonresponse Adjustment Models for a National Face-to-Face

Survey." *Journal of Survey Statistics and Methodology* 2(4):410–32. doi: 10.1093/jssam/smu012.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31(3):980–91. doi: 10.1016/j.ijforecast.2014.06.001.

Ward, Jonathan Stuart, and Adam Barker. 2013. "Undefined By Data: A Survey of Big Data Definitions." *arXiv preprint.* doi: 10.1145/2699414.

Weiseberg, Herbert. 2005. *The Total Survey Error Approach*. Chicago: The University of Chicago Press.

West, Brady T. 2013. "An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 176(1):211–25. doi: 10.1111/j.1467-985X.2012.01038.x.

West, Brady T., James Wagner, Frost Hubbard, and 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. doi: 10.1093/jssam/smv004.

Wiśniowski, Arkadiusz, Joseph W. Sakshaug, Diego Andres Perez Ruiz, and Annelies G. Blom. 2020. "Integrating Probability and Nonprobability Samples for Survey Inference." *Journal of Survey Statistics and Methodology* 8(1):120–47. doi: 10.1093/jssam/smz051.

Woollard, Matthew. 2014. "Administrative Data: Problems and Benefits: A Perspective from the United Kingdom." *Facing the future: European research infrastructures for the humanities and social sciences*, 49.

Wu, Changbao, and Randy R. Sitter. 2001. "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data." *Journal of the American Statistical Association* 96(453):185–93. doi: 10.1198/016214501750333054.

Zhang, Li-Chun. 2000. "Post-Stratification and Calibration-A Synthesis." *The American Statistician* 54(3):178. doi: 10.2307/2685587.

# Appendix A: Resumen en español

Esta sección contiene un resumen de la introducción y las conclusiones de la tesis, con el fin de cumplir con lo dispuesto en el reglamento de doctorado de la Universidad de Salamanca.

## Introducción

La posibilidad de inferir las características de la muestra a la población es el rasgo más definitorio de la encuesta. Para ello, es clave el método empleado para la selección. La teoría estadística respalda que la selección de una muestra empleando métodos probabilísticos basta para sustentar el proceso de inferencia (Neyman 1934). Sin embargo, desde el comienzo del muestreo moderno se han desarrollado diferentes metodologías de selección que, bajo la etiqueta de no probabilísticas, han sido utilizadas a pesar de las dudas que existen sobre su capacidad de inferir (Baker *et al.* 2013).

Esta tesis estudia una posible intersección entre la necesidad de reducir los sesgos de las estimaciones provocados por el sesgo de selección en muestras no probabilísticas y la oportunidad que ofrecen los datos administrativos agregados de explicar esa selección. Para ello, se presentan una simulación estadística y dos aplicaciones metodológicas a una serie de encuestas, presencial y web, realizadas en España. Primero, las simulaciones estadísticas analizan las condiciones en las que el uso de datos agregados geográficamente como variables contextuales pueden ser útiles para reducir el sesgo de selección de las estimaciones. Segundo, utilizando los estudios preelectorales y postelectorales del Centro de Investigaciones Sociológicas (CIS), que combinan métodos probabilísticos de selección con el uso de cuotas, se explora el potencial de añadir nuevas variables auxiliares para explicar el mecanismo de selección y mejorar la calidad de las estimaciones, además de utilizar técnicas de imputación. Tercero, utilizando dos encuestas provenientes de un panel experimental de internautas reclutado por la Asociación para la Investigación de los Medios de Comunicación (AIMC), se emplean datos administrativos agregados para mejorar el modelo de ajuste tradicional basado en variables sociodemográficas.

**Problema de investigación y relevancia de la tesis**

En la última década se han sucedido múltiples intervenciones acerca del presente y futuro de la investigación con encuestas que han identificado dos retos principales (Couper 2013; Groves 2011; Kalton 2019; Lohr 2017; Miller 2017; Rao y Fuller 2017; Singer 2016). El primero consiste en los problemas derivados de la inferencia a partir de muestras no probabilísticas, como los *opt-in panels*, así como los efectos de la caída de las tasas de respuesta en las muestras probabilísticas. El segundo, es integrar las nuevas fuentes de datos que han aparecido gracias a la expansión tecnológica con las encuestas.

La irrupción y expansión de internet entre la población está detrás de la aparición de un nuevo modo de administración, la encuesta web. Este modo permite recoger información en menos tiempo y con un menor coste, de ahí la expansión de su uso en las últimas décadas (Díaz de Rada *et al.* 2019). Esta popularidad se ha hecho notar entre los profesionales de la investigación y en el mundo académico. Una prueba de la creciente popularidad de la encuesta web es la proliferación de paneles de internautas, tanto los reclutados con métodos probabilísticos (Schonlau y Couper 2017) como los reclutados a partir de métodos no probabilísticos (Callegaro *et al.* 2014). A pesar de las ventajas de la encuesta web sobre otros modos de administración y su amplia aceptación, su uso en conjunción con muestras no probabilísticas pone en cuestión el proceso de inferencia, la base sobre la que se sustenta la encuesta.

Otra amenaza a la inferencia, tanto con el uso de muestras probabilísticas como no probabilísticas, es la tendencia declinante de las tasas de respuesta. Esta tendencia se ha mostrado en encuestas probabilísticas como la *Labour Force Survey* o la Encuesta Social Europea, en las que se ha experimentado un descenso en las tasas de respuesta durante las últimas décadas (de Leeuw *et al.* 2018). En la investigación de la opinión pública, *Pew Research* ha registrado una caída de las tasas de respuestas en encuestas telefónicas del 36% al 9% en el período entre 1997 y 2016 (Keeter *et al.* 2017). Aunque no existe una relación clara entre la tasa de respuesta y la magnitud del sesgo (Groves 2006), una menor tasa de respuesta abre la puerta a que se den estas desviaciones, que son difíciles de controlar en el marco de encuestas que cuentan con un elevado número de variables y son utilizadas para responder a diferentes preguntas de investigación.

El otro reto que afronta la metodología de encuestas, también relacionado con la aparición de internet, es la integración de las encuestas con las nuevas fuentes de datos que han surgido recientemente u otras fuentes que, aunque ya existían, no era posible acceder a ellas o procesar los datos.

Esta tesis explora una posible sinergia entre una fuente de datos que ahora son accesibles y procesables, los datos administrativos agregados, y la inferencia a partir de muestras no probabilísticas. A pesar de que las muestras no probabilísticas presentan problemas a la hora de inferir a la población, existen algunos supuestos en los que la estimación sin sesgos sería posible. Uno de esos supuestos consiste en que, a partir de un modelo, se controle el sesgo de selección de la muestra con respecto a la variable a estimar (Elliott y Valliant 2017; Särndal 2010). Todo modelo de ajuste precisa de información auxiliar que esté disponible para los que participan en el estudio y para los que no forman parte de la muestra. El aumento del número de fuentes de datos disponibles abre nuevas oportunidades para especificar los modelos de ajuste. Los datos administrativos son agregados, dado su accesibilidad y variedad, abren nuevas oportunidades para ajustar las muestras.

Partiendo de este contexto, esta tesis contribuye a incrementar el conocimiento metodológico en tres aspectos. Primero, se busca responder a la pregunta de en qué circunstancias los datos agregados pueden ser útiles en los ajustes realizados en presencia de sesgo de cobertura y no respuesta. Segundo, se aborda la cuestión de la calidad de las variables auxiliares en la estimación, enfocado en el caso del recuerdo de voto en estimaciones electorales. Por último, se explora la utilidad de usar datos agregados administrativos en el ajuste de una encuesta realizada a partir de una muestra no probabilística.

**Objetivos de investigación**

Los objetivos de investigación de la tesis doctoral son:

- Determinar la idoneidad de usar información auxiliar agregada como variables contextuales frente a los totales poblacionales para ajustar los sesgos presentes en las encuestas.
- Determinar el efecto de la ponderación sobre la estimación al añadir variables auxiliares sociodemográficas a la calibración para estimar la intención de votar.

- Determinar el efecto del uso de la imputación múltiple y la ponderación sobre la precisión de la estimación de voto.
- 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.
- Comparar la efectividad de los modelos de cuasi-aleatorización, superpoblación y doble ajuste para reducir el sesgo de las estimaciones provenientes de muestras no probabilísticas en comparación con la capacidad de ajuste de las especificaciones de los modelos.
- Analizar el cambio en la varianza de los estimadores al usar variables provenientes de datos administrativos agregados en la especificación del modelo.

**Estrategias de muestreo, método de selección e inferencia**

Dado que el objetivo final de la selección de una muestra es inferir, la estrategia de muestreo debe integrar el método para elegir a las unidades y la estrategia de estimación (Hansen, Hurwitz, y Madow 1953). El uso de la aleatorización como base para inferir desde la muestra a la población ha sido la idea predominante en los últimos 90 años en el ámbito de la investigación con encuestas. En una muestra probabilística la selección de cada elemento de la población se hace en base a la aleatoriedad y con arreglo a una probabilidad conocida de selección (Cochran 1971; Hansen *et al.* 1953; Kish 1965). La aleatoriedad juega un doble papel en la estrategia de muestreo, por un lado, garantiza la representatividad de la muestra con respecto a cualquier variable que sea medida en la encuesta y, por el otro, permite medir la incertidumbre derivada de observar solo a una parte de la población. Este método de inferencia se denomina inferencia basada en el diseño (Valliant, Dever, y Kreuter 2018:323–25).

Existen métodos de selección en los que la aleatorización carece de protagonismo y en el que los elementos de la población son elegidos con arreglo al criterio del equipo de investigación o de los propios elementos poblacionales. El muestreo no probabilístico engloba a una serie de métodos de selección de diversa naturaleza que, al contrario que los métodos probabilísticos, no cuentan con una base teórica común que posibilite el proceso

de inferencia. El principal inconveniente de los métodos no probabilísticos es que la falta de control sobre el proceso puede introducir un sesgo de selección en las estimaciones (Elliott y Valliant 2017). El sesgo de selección ocurre cuando las unidades incluidas en la muestra difieren de aquellas que no lo son con respecto a la variable a estimar. Inferir a partir de muestras no probabilísticas requiere un modelo que controle el impacto del sesgo de selección. En la inferencia basada en modelos se ignora el diseño de la muestra para considerar la estructura de la población (Little 2004; Valliant *et al.* 2018).

En los últimos años, el uso extendido de métodos de selección no probabilísticos para realizar encuestas web ha estimulado la discusión teórica acerca de la inferencia a partir de muestras no probabilísticas. Una de las críticas más recurrentes a la inferencia a partir de muestras no probabilísticas es que carece de un marco teórico que respalde el proceso, sino que en cada caso se precisa la especificación de un modelo concreto (Baker *et al.* 2013:103). Elliott y Valliant (2017), por un lado, ofrecen un marco conceptual para la inferencia a partir de muestras no probabilísticas, y por el otro, a partir de dicho marco articulan diferentes métodos de estimación que tienen como objetivo reducir los sesgos de las estimaciones y aportan alternativas para calcular la varianza de los estimadores.

Mercer, por su parte, realiza una doble aportación a la teoría de la inferencia a partir de muestras no probabilísticas. En primer lugar, hace una reflexión sobre porqué el marco utilizado en la metodología de encuestas para analizar la calidad de las estimaciones, el Error Total de Encuestas (TSE), no es adecuado para abordar el uso de muestras probabilísticas para inferir. En segundo término, propone utilizar los conceptos del análisis de la inferencia causal para enmarcar la inferencia desde muestras no probabilísticas (Mercer *et al.* 2017).

La idea de *fit for purpose* aplicada a la inferencia a partir de muestras no probabilísticas apuesta por un esquema de evaluación de la calidad de las estimaciones que tenga en consideración el coste, y de forma más importante, el objetivo para el que los datos se han recogido (Baker *et al.* 2013:98).

**Ajustes en muestras no probabilísticas**

El desarrollo de una teoría consistente que de cobertura a la inferencia a partir de muestras no probabilísticas ha ido en paralelo a los avances que se producían en las técnicas

de ajuste que permiten reducir el sesgo de las estimaciones y posibilitar la inferencia. La Tabla A.1 presenta las técnicas según su objetivo sea modelar la probabilidad de selección o la variable de interés.

Tabla A.1. Métodos de ajustes de estimaciones obtenidas a partir de muestras no probabilísticas

| Ajustes globales | Ajustes basados en la variable de interés |
|---|---|
| • Calibración y postestratificación<br>• Ponderación basada en la pobabilidad de formar parte de la muestra<br>• Muestreo por correspondencia | • Superpoblación<br>• Calibración asistida por modelo<br>• Regresión multinivel con posestratificación |

Fuente: Adaptado de Cornesse *et al.* (2020).

La idea que subyace al uso de estas técnicas es intentar controlar los sesgos de las estimaciones a partir de modelar el mecanismo de selección utilizando variables auxiliares que están disponibles tanto para los individuos seleccionados como para el resto de la población. En el caso de que el sesgo exista, la especificación del modelo es la clave para poder realizar la inferencia. Lo ajustes se pueden clasificar en dos grupos, aquellos que tienen como objetivo modelar la probabilidad de selección y aquellos cuyo objetivo es modelar la variable de interés.

- La calibración consiste en un modelo que tiene como objetivo producir una ponderación que hace coincidir la distribución de una serie de variables auxiliares en la muestra con los totales poblacionales (Kott 2006). Una forma avanzada de calibración es la postestratificación, en la que la muestra es dividida en grupos formados por la interacción de todas las variables auxiliares (Zhang 2000).
- La ponderación basada en la probabilidad de formar parte de la muestra consiste en utilizar una encuesta de referencia con una muestra probabilística junto con la muestra no probabilística con el fin de ajustar un modelo que prediga la probabilidad de selección en la muestra no probabilística (Valliant y Dever 2011). La ponderación que ajusta la muestra no probabilística es el inverso de la probabilidad de selección predicha.
- Las técnicas de muestreo por correspondencia parten de un modelo estadístico para seleccionar la muestra no probabilística utilizando para ello una muestra probabilística de referencia (Bethlehem 2016; Rivers 2007; Vavreck y Rivers 2008). El

objetivo de este método es que la muestra no probabilística sea un reflejo de la muestra probabilística con respecto a una serie de variables auxiliares que explican el modelo de selección.

El segundo grupo de ajustes incluye aquellos que modelan el mecanismo de selección a partir de la variable a estimar. Un inconveniente de estos métodos es que la estimación de cada variable requiere el ajuste de un modelo (Elliott y Valliant 2017):

- Los modelos de superpoblación se basan en la idea de que existe un modelo común que explica la generación de los datos en la muestra observada como en el resto de la población. Este método consiste en ajustar un modelo para la variable de interés en la muestra observada, con el fin de utilizar ese modelo para predecir la variable entre los casos de la población no incluidos en la muestra. El estimador final combina tanto la muestra observada como las predicciones realizadas con los datos auxiliares en el resto de la población (Valliant, Dorfman, y Royall 2000).
- La calibración asistida por modelo consiste en la generación de ponderaciones específicas para una variable de interés a partir de un modelo de predicción especificado con una serie de variables auxiliares para las que se conocen los totales poblacionales (Wu y Sitter 2001).
- La regresión multinivel con postestratificación consiste en utilizar un modelo multinivel para predecir la variable de interés en la muestra no probabilística a partir de una serie de variables auxiliares. Las predicciones se realizan para las celdas resultantes de combinar las categorías de las variables utilizadas en el modelo, postestratos. Por último, esas predicciones son ponderadas de forma que los postestratos de la muestra tengan el peso que les corresponde en la población (Gelman 2007; Park, Gelman, y Bafumi 2004). La efectividad de esta técnica está determinada por la correcta especificación del mecanismo de selección en el modelo multinivel (Buttice y Highton 2013).

**Datos auxiliares en la inferencia asistida y basada en modelos**

Todas las técnicas enumeradas en la sección anterior que buscan ajustar las estimaciones ante la presencia del sesgo de selección tienen en común que precisan de variables auxiliares. Las variables auxiliares están disponibles tanto para los que participan en el

estudio como para el resto de los elementos de la población, lo que permite que sean utilizadas para, por ejemplo, predecir la probabilidad de responder a la encuesta. Las variables auxiliares, tradicionalmente, pueden presentarse en tres formatos: a nivel individual para todos los elementos de la población, a nivel individual para todos los elementos de la muestra o a nivel agregado (Bethlehem, Cobben, y Schouten 2011:247–48).

El tipo de datos auxiliares disponibles —nivel poblacional, nivel muestral o agregados— condiciona los métodos de ajuste que se pueden emplear. Así, en el marco de la encuesta probabilística, el uso de modelos para predecir la probabilidad de responder sólo es posible ajustarlos si la información existe a nivel poblacional o muestral (Bethlehem *et al.* 2011:8). Por su parte, los datos agregados permiten los ajustes por calibración o poststratificación (Särndal 2007). Independientemente del método de ajuste empleado, una variable auxiliar cumple su función de ajustar el sesgo de las estimaciones si está correlacionada con la probabilidad de responder y la variable de interés (Särndal y Lundström 2005).

Esta lógica, desarrollada en el marco de las encuestas probabilísticas, es trasladable a la inferencia a partir de encuestas no probabilísticas en su mayor parte. Sin embargo, hay algunas diferencias destacables. La primera es que, en las encuestas no probabilísticas, está extendido el uso de encuestas probabilísticas como referentes, por lo que no es necesario que las variables auxiliares estén disponibles para todos los elementos de la población. La segunda diferencia, también relacionada con las variables auxiliares, es que la naturaleza del mecanismo de selección a modelar suele diferir entre encuestas probabilísticas —cobertura y no respuesta— y las no probabilísticas.

Con la aparición de las encuestas web y el uso más frecuente de muestras no probabilísticas también han visto la luz algunas investigaciones acerca de las variables auxiliares que pueden ser efectivas para eliminar el sesgo de las estimaciones. En esta línea, los promotores de uno de los primeros paneles de voluntarios de los Estados Unidos, Harris Polls, promovieron el término *webographics* para referirse a una serie de variables actitudinales que eran incluidas en sus modelos de ajuste (Terhanian *et al.* 2000). Aunque algunos trabajos han mostrado el potencial de estas variables para discriminar entre la población online y offline (Schonlau *et al.* 2004; Schonlau, van Soest, y Kapteyn 2007), su escaso

uso en la práctica hace pensar que la capacidad de reducir los sesgos es reducida en muchas ocasiones (Schonlau y Couper 2017).

La expansión de internet y la posibilidad de generar, procesar y almacenar grandes volúmenes de datos, el fenómeno denominado *big data*, abre nuevas oportunidades para mejorar la especificación de los modelos de ajuste de las estimaciones (Baker 2017; Couper 2013). Los datos administrativos son productos o subproductos generados en la interacción de la administración pública con los ciudadanos, empresas u otras organizaciones con el fin de organizar, gestionar o monitorizar servicios (Playford *et al.* 2016; Woollard 2014). Estos datos, de cara a ser utilizados como variables auxiliares, tienen la ventaja de que suelen cubrir a la totalidad de la población. A pesar de esta ventaja, el acceso a los datos administrativos a nivel poblacional está limitado a los organismos públicos principalmente por razones relacionadas con la privacidad y la seguridad. Una alternativa es utilizar los datos agregados a un nivel geográfico que publican las administraciones públicas como parte de su labor informativa.

El uso de los datos agregados ha sido recurrente para ajustar las estimaciones de las encuestas. Como se exponía en la sección anterior, algunos métodos de ajuste como la calibración o las postestratificación precisan de esos datos agregados para ajustar los modelos (Särndal y Lundström 2005:49–65). Además, tanto en la calibración como la postsestratificación, las variables auxiliares deben ser medidas en la muestra como parte de la encuesta. Frente a este uso de los totales poblacionales producidos por las variables administrativas, cabe utilizar los datos agregados a un nivel geográfico menor como variables contextuales, que informan acerca del entorno —sección censal o municipio— en el que reside el individuo seleccionado en la muestra.

## Conclusiones

Esta sección presenta un resumen de las conclusiones incluidas en la tesis.

**Los datos auxiliares, para ser efectivos a la hora de ajustar el sesgo de selección, deben estar correlacionados con la variable de interés y la de agrupamiento.**

El primer objetivo de esta investigación buscaba determinar bajo qué circunstancias los datos administrativos agregados son efectivos a la hora de ajustar el sesgo de las

estimaciones. Para ello, a partir de una serie de simulaciones estadísticas se ha comprobado la efectividad de estos datos incluidos de dos formas diferentes, como totales poblacionales y como variables agregadas a un nivel inferior que informan acerca del contexto del elemento muestral. Los resultados muestran que, para que las variables auxiliares que contienen información agregada funcionen debe 1) estar agrupada la variable objetivo, 2) correlacionada con la variable auxiliar y, a su vez, 3) la variable auxiliar correlacionar con la probabilidad de selección en la muestra. Sólo en ese caso los datos auxiliares agregados como variables contextuales son efectivos a la hora de eliminar el sesgo de las estimaciones. Este hallazgo pone en cuestión la aplicabilidad de los datos agregados utilizados como variables contextuales para ajustar las estimaciones en el ámbito de las Ciencias Sociales, principalmente porque los niveles de correlación entre la variable objetivo y la de agrupación necesarios no son habituales.

**Las variables auxiliares no consiguen ajustar totalmente las estimaciones a pesar de que estén correlacionadas con la probabilidad de responder y la variable de interés.**

El requisito básico para que una variable auxiliar sea efectiva a la hora de ajustar el sesgo de las estimaciones es que sea capaz de explicar el mecanismo de selección con respecto a la variable objetivo (Särndal y Lundström 2005:110). Con este objetivo, en el segundo artículo de la tesis se propone el uso de diferentes especificaciones para estimar la intención de voto a partir de una serie de encuestas pre y postelectorales llevadas cabo por el Centro de Investigaciones Sociológicas (CIS). Los ajustes empleados incluyen variables sociodemográficas—sexo y edad, región, tamaño de municipio, educación y empleo—y diferentes modalidades del recuerdo de voto. En las estimaciones de voto, el uso del recuerdo de voto como variable auxiliar presenta la ventaja de que suele cumplir el doble requisito para eliminar el sesgo de las estimaciones, estar correlacionada con la probabilidad de formar parte de la muestra final y la variable de interés, la intención de voto.

Los resultados muestran que el uso de las variables sociodemográficos tiene un efecto mínimo a la hora de corregir el sesgo de la estimación de voto, en línea con otros hallazgos en el mismo campo (Crespi 1988; Durand, Deslauriers, y Valois 2015). El uso de la variable auxiliar recuerdo de voto consigue mejorar la calidad de las estimaciones,

pero este efecto no se observa en todos los casos estudiados y en ningún caso se consigue eliminar completamente el sesgo de las estimaciones. Los resultados de este análisis también sugieren que los sesgos evolucionan y que la especificación de los modelos de selección debe evolucionar con ellos. El hecho de que la variable recuerdo de voto, claramente correlacionada con la intención de voto, no sea efectiva en algunas ocasiones para reducir el sesgo de las estimaciones, señala que la naturaleza del sesgo es diferente y las variables auxiliares del modelo de estimación deben de tenerlo en cuenta.

**Los datos agregados a nivel geográfico y utilizados como variables auxiliares contextuales no tienen una gran capacidad de incidir sobre sesgo de las estimaciones.**

El tercer artículo es una aplicación del uso de datos administrativos agregados para corregir dos encuestas provenientes de un panel de internautas dedicado a la medición de aspectos relacionados con el consumo de medios de comunicación. Los resultados del análisis muestran, como ya anunciaban las simulaciones del primer artículo (capítulo 2) de la tesis, que el uso de las variables administrativas agregadas no tiene apenas efecto a la hora de reducir el sesgo de las estimaciones. Estos resultados coinciden con los observados en otras investigaciones (Biemer y Peytchev 2013; Butt y Lahtinen 2016), a pesar de que en aquellos trabajos se utilizaran muestras probabilísticas que probablemente contaban con un menor nivel de desajuste.

**Los diferentes métodos de estimación alcanzan resultados similares si emplean el mismo vector de variables auxiliares.**

Otra cuestión que se aborda en esta tesis es la capacidad de los diferentes modelos frente al conjunto de las variables auxiliares empleadas para ajustarlos. En el tercer artículo que compone la tesis se utilizaron tres métodos para estimar a partir de una muestra no probabilística: la cuasi-aleatorización, los modelos de superpoblación a partir de una calibración asistida y una combinación de ambos. Como se relata en párrafo anterior, los ajustes contaron con una serie de variables sociodemográficas y un conjunto de variables administrativas agregadas al nivel de municipio. A pesar de que las variables auxiliares apenas consiguieron reducir el sesgo de las estimaciones, en la mayoría de los casos se cumple lo observado por Mercer, Lau y Kennedy (2018): las variables son más relevantes que el modelo utilizado para realizar el ajuste.

**Los problemas del *Total Survey Error* (TSE) para analizar la inferencia a partir de muestras no probabilísticas.**

Aunque no corresponde con un objetivo principal de la tesis, es preciso comentar el papel que ha jugado el TSE y las conclusiones que se han alcanzado durante la investigación. El TSE es el marco predominante para analizar la calidad de las estimaciones realizadas a partir de encuestas (Biemer 2010; Biemer y Lyberg 2003). Sin embargo, en las encuestas no probabilísticas, como las del CIS empleadas en este trabajo, el uso de cuotas para seleccionar a la persona entrevistada y el alto grado de discreción con el que cuenta el entrevistador (Díaz de Rada 2005) hace que el marco del TSE sea insuficiente para abordar el proceso.

Se puede concluir que el TSE es un marco válido para el análisis de la calidad de las estimaciones en las encuestas probabilísticas y algunos de los errores que contempla constituyen categorías válidas para analizar las estimaciones realizadas a partir de muestras no probabilísticas. Sin embargo, el TSE no es suficiente para explicar la complejidad de los procesos de selección y respuesta que se dan en este tipo de muestras.

**Todos los desarrollos teóricos conducen a la especificación teórica de los modelos.**

Una de las principales críticas que recibe la inferencia a partir de muestras no probabilísticas es la falta de un marco teórico coherente (Baker *et al.* 2013). En los últimos años han visto la luz desarrollos teóricos que pretenden dar respuesta a esa falta de un respaldo teórico al proceso de inferencia (Elliott y Valliant 2017; Mercer *et al.* 2017). A pesar de que estas propuestas pretenden crear un marco propio diferenciado del TSE para el uso de muestras no probabilísticas, no consigue disipar la principal crítica que recibe la inferencia a partir de muestras no probabilísticas y que es inherente a su naturaleza, la necesidad de utilizar un modelo de estimación para cada variable. Otra crítica al uso de muestras no probabilística señala que no es posible saber en qué situaciones funcionan —consiguen controlar el sesgo— los modelos en los que se basa la inferencia. La posibilidad de inferir a partir de modelos descansa en unos supuestos sobre el proceso de generación de los datos en la población que en la mayoría de los casos no son comprobables.

### Del optimismo del *big data* a la vuelta a la teoría.

El comienzo de esta investigación estuvo marcado por un momento cierto optimismo justificado por la expansión tecnológica y la posibilidad de acceder y procesar ingentes cantidades de información. Esta investigación trataba precisamente de utilizar una parte de la información que ahora es accesible —los datos administrativos agregados— para mejorar la calidad de las estimaciones realizadas a partir de muestras no probabilísticas.

La conclusión con la que cierro este trabajo dista de aquel primer optimismo. No cabe duda de que la mayor disponibilidad de fuentes de datos va a ayudar a mejorar la calidad de las estimaciones realizadas a partir de encuesta, es por ello que son múltiples las áreas de la metodología de encuestas en las que se están desarrollando sinergias (Hill *et al.* 2019). Pero también es cierto que la abundancia de datos no va a sustituir la necesidad de entender los fenómenos para ser efectivos a la hora de, por ejemplo, realizar ajustes en las estimaciones realizadas a partir de encuestas. Probablemente, dar a la teoría un papel más central a la hora de pensar la especificación de los modelos de estimación puede ayudar a seleccionar mejores predictores. Probablemente, el futuro no está en hacer un análisis con grandes volúmenes de datos, sino en utilizar la mayor variedad de fuentes de datos disponibles para identificar las variables que precisan los ajustes. Probablemente, el futuro pasa por seguir trabajando para entender los mecanismos que posibilitan la inferencia a partir de muestras no probabilísticas.

## Implicaciones y futuras líneas de investigación

La implicación metodológica más relevante de esta tesis es desaconsejar el uso de datos administrativos agregados como variables contextuales en los ajustes del sesgo de selección de las estimaciones realizadas a partir de muestras no probabilísticas. Además, la evidencia que existe es suficiente para subrayar que el problema no radica en la relación de las variables auxiliares con las variables objetivo o la probabilidad de selección en la muestra, sino la naturaleza agregada de las variables. Podría caber, en el futuro, comprobar la utilidad de este tipo de variables auxiliares en el marco de encuestas, como pueden ser las realizadas en el ámbito educativo, en el que algunas variables objetivo y auxiliares presentan niveles más altos de conglomeración. Sin embargo, este tipo de encuestas son un

una minoría y la disposición de datos agregados administrativos no está garantizada en todos los niveles.

La segunda implicación de este trabajo tiene que ver el potencial de los datos agregados administrativos como totales poblacionales para reducir el sesgo de las estimaciones. Los resultados presentes en esta tesis muestran que las variables auxiliares utilizadas en los ajustes deben ser adaptadas a las características que definen el proceso de selección y su relación con la variable a estimar en cada momento. Esto se une a la iniciativa de recuperar la reflexión teórica con el fin de implementar modelos de ajuste cada vez más específicos (Peytchev *et al.* 2018), en parte, gracias también a la mayor disponibilidad de variables auxiliares que permiten materializar esos modelos teóricos.

Por último, a pesar de que la principal pregunta de esta tesis ha encontrado respuesta, permanecen abiertas infinidad de cuestiones que explorar, tanto en el plano teórico como empírico, dentro del ámbito de la inferencia a partir de muestras no probabilísticas. En el plano teórico, es preciso profundizar en el marco del fit for purpose. También en el plano teórico, es preciso profundizar en la singularidad de los métodos de selección que se inscriben en el muestreo no probabilístico. En el plano empírico, en lo tocante a las variables auxiliares de los modelos de ajuste, es preciso investigar el impacto del uso de ponderaciones en análisis de tipo bivariado y multivariado, donde parece que el impacto del sesgo de selección puede ser diferente (Pasek 2016). Otro aspecto empírico que está por tratar es la conveniencia de emplear los diferentes métodos de ajuste disponibles según las características del método de selección y el contexto de la estimación.

# Appendix B: Aggregate data to correct for nonresponse and coverage bias in surveys

Cabrera-Álvarez, Pablo. 2021. "Datos Agregados Para Corregir los Sesgos de No Respuesta y de Cobertura en Encuestas." *Empiria. Revista de Metodología de Ciencias Sociales* (49):39–64. doi: 10.5944/empiria.49.2021.29231.

**Abstract**

In the last decades, the effect of nonresponse and coverage biases in surveys has questioned the ability to infer 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 that using 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.

**keywords:** survey methodology, nonresponse, weighting, aggregate data, statistical simulations.

## Introduction

In 1788 John Sinclair carried out one of the first documented surveys. It was a questionnaire with more than 100 questions sent to each parish of the Church of Scotland. After 23 reminders, including the last one written in blood red, he achieved a 100% response rate (de Leeuw and Hox, 2011). Many things have changed in survey research since John Sinclair carried out his census of parishes. Nowadays, any survey practitioner would consider it impossible to achieve a response rate close to 100%, even in a scenario with enough resources to implement the most sophisticated data collection strategy.

In the last decades, the spread of internet research using nonprobability panels and a steady decline in response rates has led to a picture of uncertainty. In both telephone and face-to-face surveys, fewer and fewer people are willing to respond. For example, the response rate in telephone surveys in the United States has fallen from 36% to 6% between 1997 and 2018 (Kennedy and Hartig 2019). These phenomena—the drop in response rates and the spread of online research—cast doubt on the inference process that allows extrapolating information from the sample to the population (Valliant, Dever and Kreuter 2017).

To correct for sample biases that may compromise the inference process, once the fieldwork has been completed, researchers can generate weights that modify the original weight of each case. These weights are calculated using auxiliary information, i.e. variables available for all population elements, respondents and nonrespondents. Statistical theory states that to the extent that these auxiliary variables are correlated with the probability of responding and the variable of interest, the bias in the estimate will be corrected (Bethlehem, Cobben and Schouten 2011).

Some work has shown that the key to adjusting a survey lies in the set of auxiliary variables taken into account rather than the method used to compute the weights (Mercer *et al.* 2018). However, restrictions on access to population-based microdata constrain the ability to implement adjustments. An alternative to microdata is to rely on population totals from sources such as the census, which can be used to detect deviations in the sample distribution and subsequently adjust it. These population totals can be treated as contextual variables that collect information from the place, like a census ward, a local authority, or a company, where the sample element belongs.

Based on statistical simulations, this research aims to determine the suitability of using population totals as contextual variables versus individual variables to adjust the surveys. The results suggest that the clustering level of the target variable is the most relevant factor for a weight computed with context variables to be effective. Furthermore, two other elements must be present, the correlation between the auxiliary variable and the target variable and the correlation between the auxiliary variable and the probability of response.

The first section of this paper presents the theoretical framework and background of this research. The second section presents a series of hypotheses, followed by details on the simulation of the data and its analysis. The fourth section contains the results of the simulations. Finally, the results are discussed, and the conclusions presented.

## Background

Survey research relies on the ability to infer population characteristics from a random sample. The inference process requires selecting a probability sample and the full cooperation of those included in the sample, an implausible scenario in social research. In the last years, two phenomena have enlarged the issues caused by noncoverage and nonresponse. The first is the continuous drop in response rates (de Leeuw, Hox y Luiten 2018), and the second is the expansion of online research using nonprobability samples (Blom *et al.* 2016; ESOMAR 2017).

The widespread drop in response rates casts doubt on whether the use of probability samples is sufficient to guarantee the inference process. The problem of nonresponse is that subsets of the population have different probabilities of participating in the survey. These differences can lead to biased estimates (Groves and Couper 1998; Dillman *et al.* 2002). The drop in response rate affects both face-to-face surveys (Beullens *et al.* 2018; de Leeuw, Hox and Luiten 2018) and telephone surveys (Kennedy and Hartig 2019).

Another issue that affects the quality of survey data is the undercoverage bias: the inability to reach a part of the target population. There are various reasons for undercoverage. First, in some instances, a part of the population is not reachable because they live in institutions or other places where the survey organisation cannot access them. Second, the survey mode can also have an impact on the coverage. For example, some households

cannot participate in a web survey because they do not have a device or an internet connection. Finally, sometimes the sample frame fails to contain all members of the population (Weiseberg 2005).

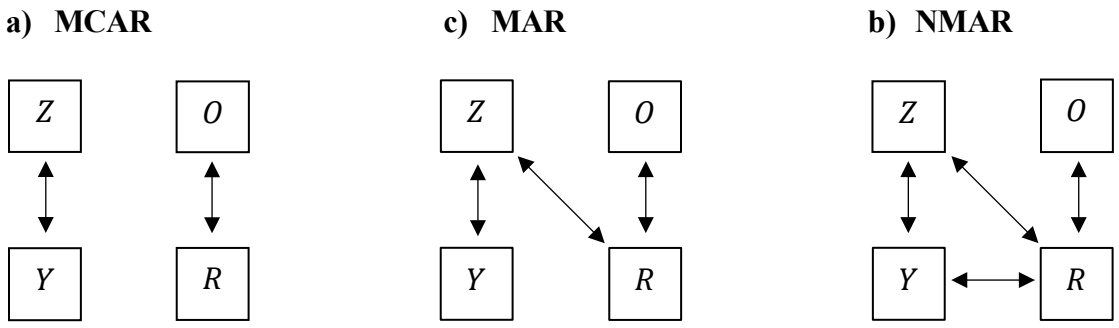
Different methods exist to correct for undercoverage and nonresponse bias before (Hansen 2007; Manfreda *et al.* 2008; Mohorko, Leeuw and Hox 2011; Ryu, Couper and Marans 2006; Singer, Groves and Corning 1999), during (Groves and Heeringa 2006; Lepkowski *et al.* 2013; Olson and Peytchev 2007) and after data collection (Levy and Lemeshow 2013; Little and Vartivarian 2005; Sakshaug and Eckman 2017). This paper focuses on the survey adjustments made once the fieldwork has concluded. These adjustments or weights seek to balance the final sample, matching the distribution population in some key characteristics. The simplest case of weighting consists in calculating a coefficient *w* for each subgroup *j* defined by the auxiliary variable *z*:

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

[B.1]

where $N_{zj}$ is the population total for the subgroups of the variable $z$, and $n_{zj}$ refers to the same total, but for the sample. This is the simplest way to compute a weight, the cell adjustment, which consists in creating a ratio between the population total and the sample total for the different categories of a variable. Other methods for generating adjustments are calibration and post-stratification (Dever, Rafferty and Valliant 2008; Särndal 2007; Tsung, Valliant and Elliott 2018; Zhang 2000). The former forces the final sample to match the marginal distribution of the population totals of the auxiliary variables. The latter adjusts the final sample to match the distribution of the auxiliary variables and their interactions. Another method consists in computing the response propensity of the sample elements (Bethlehem, Cobben and Schouten 2011; Elliott and Valliant 2017). The response propensities are computed using statistical models that require a sampling frame with auxiliary information for respondents and nonrespondents (Bethlehem *et al.* 2011). Finally, reference probability samples (Pedraza *et al.* 2010; Gummer and Roßmann 2018; Lee and Valliant 2009; Pasek 2016) or propensity score matching techniques (Elliott and Valliant 2017; Mercer *et al.* 2018) can also calculate the response propensities.

The missing data mechanism underlying the sample determines the success of survey weights tackling nonresponse and undercoverage bias. Scholars have differentiated three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Bethlehem *et al.* 2011; Little and Rubin 1987). Figure B.1 adapted from Bethlehem *et al.* (2011) presents a summary of how the different mechanisms operate. Under the MCAR mechanism, weighting is unnecessary as the estimate is unbiased, and under the NMAR mechanism, it is futile as participation in the survey depends directly on the target variable. Only under the MAR assumption can the use of weights reduce the bias of the estimate. In that case, correcting the sample imbalance of $Z$ entails reducing the bias of the target variable $Y$.




$Y$: target variable; $Z$: auxiliary variable; $O$: unobserved variables; $R$: participation in the survey.

Figure B.1. Unit nonresponse mechanisms (Bethlehem et al. 2011)

**Auxiliary variables and aggregate population data**

Survey weights need to meet two requirements to reduce the bias of the estimates. On the one hand, the auxiliary variables must be related to the response probability of the sample elements, and, on the other hand, the auxiliary variables must also be related to the target variable (Bethlehem, Cobben and Schouten 2011). The search for auxiliary variables that meet these requirements presents some theoretical and practical limitations. The former refers to the lack of theory that guides the auxiliary variable selection process. The

latter is the most usual and points to the reduced number of variables populated for those who do not participate in the survey or are not part of the sample frame.

Auxiliary variables contain information at the individual level or present a summary statistic, like a total or average, aggregated at a geographical level. This circumstance determines the type of weight that can be computed. For instance, if the auxiliary information is available at the individual level, the sample frame can be merged with the weighting variables to model the response propensity and compute the weights (*p. ej.* Park *et al.* 2013). In this case, each individual in the survey sample is matched to a record of the auxiliary variables. Most of the time, the auxiliary information is unavailable at the individual level, but statistical summaries exist for the whole population. In such a scenario, calibration and poststratification techniques can be used to generate a weight since they only require population totals (Särndal and Lundström 2005). There is a further possibility: using aggregate information like contextual variables that, for example, bring information about the sample member's neighbourhood in a general population survey. Other examples of contextual variables are the proportion of elderlies who live in a census tract or the proportion of luxury cars registered in a municipality. This information can also be used in the computation of survey weights. This paper focuses on the possibility of using contextual variables to improve survey weights, a topic that barely has been addressed in the statistical literature.

Using aggregate data to adjust survey estimates is a common approach in survey research. The clearest example is the census data. General population surveys generally include weights that adjust the sample distribution to match the census figures of sex, age, and region (*e.g.* Park *et al.* 2013). In recent years, with the emergence of new data sources, there is renewed interest in using them to correct for survey bias (Burrows and Savage 2014; Couper 2013). In this line, Smith and Kim (2013) developed a system to systematise the collection, storage, and use of additional auxiliary information, the Multilevel Integrated Data-Base Approach (MIDA). They remark that one of the main objectives of this system is to identify nonresponse bias and adjust the surveys.

However, few papers have used aggregate data as auxiliary variables to adjust a survey. In one paper, Biemer and Peytchev (2012; 2013) used aggregate census data to

correct for bias in estimates from a telephone survey in the United States. In light of the results, the authors concluded that aggregate census data is only effective for adjusting surveys if individuals with a certain characteristic are clustered together, and this characteristic correlates with the variable of interest. More recently, in 2016, a study using the European Social Survey in the United Kingdom addressed the effectiveness of administrative aggregate data (Butt and Lahtinen 2016). They used a wide range of administrative data such as crime statistics, the census, area deprivation measures, or statistics from the departments of transport, education, or environment. The auxiliary variables derived from this data, which were aggregated at the local authority or census tract levels, did not help identify nonresponse and correct the bias in the survey estimates.

**Aggregate data mechanisms**

Although some empirical contributions focused on how aggregate contextual variables can help reduce the impact of selection and nonresponse bias in surveys, there is no theoretical analysis of the requirements needed for this approach to work. Biemer and Peytchev (2013) use a framework of survey adjustments with individual data. Under this framework, if the auxiliary variables are strongly correlated with the response propensity and the target variable, the bias of the survey estimate will be adjusted (Bethlehem *et al.* 2011). In addition to this, Biemer and Peythev (2013) point out that the aggregate auxiliary variables must be a good proxy of the individual characteristics. For instance, if 99% in a census tract are women, this aggregate information is a good indicator of the sex of the sample member in case they do not provide this information. However, these authors do not address the role of the clustering of the target variable.

Several mechanisms explain how aggregate auxiliary variables may be related to survey target variables. Three elements intervene in these basics models, the auxiliary variable ($Z$), the clusters ($K$), and the target variable ($Y$). Auxiliary variables cover the whole population and can be used to correct the bias derived from nonresponse and selection. Examples of aggregate auxiliary variables are the average income in a census tract, the electoral results in the ward, the number of luxury cars registered in a municipality, or the percentage of students that receive free school meals. Besides, some populations are

grouped in clusters. For example, the general population are grouped in municipalities or census tracts as students are grouped in schools and classes.

The relationship between the clustering variable ($K$) and the auxiliary ($Z$) and target ($Y$) variables can be calculated through the intraclass correlation (ICC) $\rho$, which is:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \quad [B.2]$$

where $\sigma_b^2$ is the variance between groups, and $\sigma_w^2$ is the variance within groups (Liljequist, Elfving and Roaldsen 2019). Hence, $\rho$ takes 0 if the clustering is not related to the target variable and 1 if the clustering variable fully explains the variance of the target variable.

Thus, the level of clustering of the variable $Z$ is determined by the intraclass correlation $\rho_{ZK}$ and the level of clustering of $Y$ is determined by $\rho_{YK}$. The relationship between $Z$ and $Y$ is expressed by the correlation coefficient $r_{ZY}$. These three elements can be arranged in three different scenarios. In the first scenario, the aggregation level of $Y$ depends on the correlation between $Z$ and $K$. In the second scenario, the level of clustering of $Z$ is explained by the relationship between $Y$ and $K$, and, under the third mechanism, the clustering of $Z$ and $Y$ are independent.

In the first scenario, shown in Figure B.2, the auxiliary variable ($Z$) plays a crucial role in determining the level of aggregation of the target variable ($Y$). To better explain this mechanism, let us assume that we want to estimate the religious affiliation ($Y$) of the general population from a survey. The auxiliary variable, country of birth, can be used to adjust nonresponse and undercoverage bias in the survey. Previous research has shown that individuals born in other countries present a different distribution of religious affiliation than the local population (Santiago and Pérez-Agote 2013). Furthermore, people born in other countries are more likely to participate in surveys (Morales and Ros 2013). The information about the country of birth can be obtained from the National Office of Statistics, aggregated at the census tract or municipality levels ($K$). In this scenario, the relationship between religion and the clustering variable depends on whether people who were born in other

countries tend to live together ($\rho_{ZK}$) and the correlation between the country of birth and the religious affiliation ($r_{ZY}$).

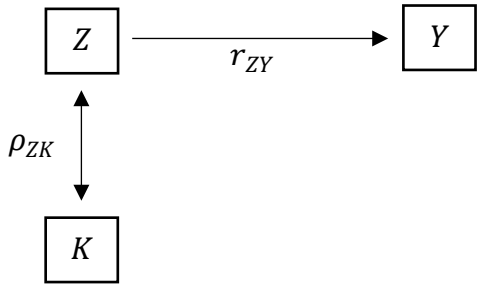



Figure B.2. Clustering of variable Y determined by $r_{zy}$

In the second scenario, variable $Y$ is grouped independently, while the level of clustering of $Z$ is given by the correlation $r_{ZY}$ (Figure B.3). This case is identical to the first scenario, but here variable $Y$ determines the level of clustering of $Z$.

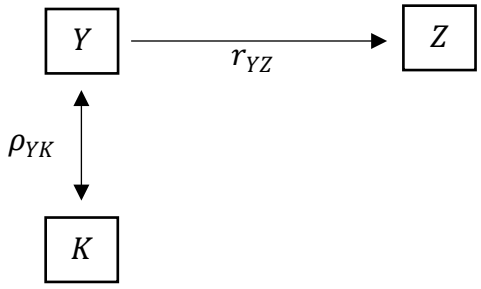



Figure B.3. Clustering of variable Z determined by the relationship of Y and K

Finally, in the third scenario, variables $Z$ and $Y$ are generated independently based on their relationship with the clustering factor ($K$), which are denoted by $\rho_{ZK}$ and $\rho_{YK}$, respectively (Figure B.4). An example of this scenario could be a survey of employees, grouped in companies ($K$), aiming to estimate the percentage of full-time personnel ($Y$). A posible auxiliary variable that can be used to diagnose and adjust deviations caused by nonresponse and undercoverage is the employees' area of studies ($Z$). This variable is an indicator of the percentage of workers who studied science, health, engineering, social sciences, or humanities. Being a full-time employee has to do with the company's policies,

and, at the same time, it is possible to find employees from the same area of knowledge in the same company. However, these two variables can be uncorrelated.

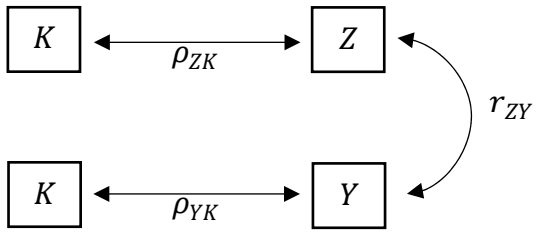



Figure B.4. Clustering of Y and Z are independent

## Research hypotheses

This section presents the main hypotheses of the work based on the theory presented in the previous section.

**H1**. *Aggregate data used as contextual variables may have comparable or even greater ability to tackle bias than aggregate data used as population totals*.

In most surveys, aggregate data is not used as contextual variables for various reasons, such as low effectiveness (Butt and Lahtinen 2015), lack of correlation with individual respondent characteristics (Biemer and Peytchev 2013), or the possible counter-effect of an ecological fallacy (Robinson 2011). This paper aims to discover whether contextual variables can improve the efficiency of the survey adjustments reducing nonresponse and undercoverage bias.

**H2**. *The power of aggregate data used as contextual variables to adjust survey estimates depends on the extent to which the auxiliary variables are clustered, measured by the intraclass correlation.*

Biemer and Peytchev (2013) argue that one of the requirements for aggregate data to be effective in reducing nonresponse bias is that the auxiliary variable is clustered. The authors suggest that contextual variables need to be good predictors of the individuals' characteristics to be effective. For example, the average income in a census tract is expected to be an accurate indicator of individuals' wealth.

**H3**. *The correlation between the auxiliary and the target variables is relevant in the adjustment with individual-level data and the adjustment with aggregate data*.

Biemer and Peytchev (2013) also found that the auxiliary variable needs to be correlated with the target variable. In summary, they point out that the same framework employed to understand individual-level weighting can be applied to the use of contextual variables. An effective weight requires the auxiliary variable to be correlated with the response propensity and the target variable.

**H4**. *The effect of the correlation between the auxiliary and the target variables and the level of clustering of the auxiliary variable on the effectiveness of the weight depends on the aggregate data mechanism.*

The aggregate data mechanism or how the aggregate data is generated affects the correlation between the auxiliary and the dependent variables and their levels of clustering. At the same time, these affect the effectiveness of the survey weight to reduce the bias of the estimates. For instance, if the auxiliary variable is strongly clustered, the correlation between the auxiliary and target variables will play a key role in reducing the level of bias.

**H5**. *The size of the clusters is not related to the effectiveness of the contextual variables in the weighting.*

As showed by Butt and Lahtienen (2016) in their research using the European Social Survey in the United Kingdom, the geographical level of aggregation of the data is not related to the ability of the weights to adjust the bias. This implies that for a general population survey it is not relevant whether the information of the contextual data is grouped at the municipality or census tract level.

**H6**. *The size of the bias is not related to the potential of the aggregate data to adjust them.*

It would be possible that, as the magnitude of the bias increases, the potential of the aggregate data to reduce the bias also increases. However, this possibility only occurs if the auxiliary variable correlates with the response propensity and the target variable. Hence

rather than the magnitude of the bias, it is the ability of the auxiliary variables to adjust the survey estimates that is relevant.

## Methodology

First, this section outlines the data simulation method employed in the paper. It then provides a detailed explanation of the methodology followed to compute the survey weight. Finally, it presents the measures used to assess the effect of the survey weights.

### Generation of simulated data

The purpose of the simulations is to determine the potential of aggregate variables to reduce the impact of coverage or nonresponse bias and the requirements for this potential to unfold. The simulation of scenarios tested various combinations of the parameters across 500,000 populations each with a size of 100,000. Each population contained three variables: the clustering indicator ($K$), the binary target variable ($Y$) and the binary auxiliary variable ($Z$). The simulation of the data accounted for the aggregate data generation mechanism. This is important to establish how the random variables are generated. The data generation mechanism presented in the previous section was implemented to assess the efficiency of the auxiliary contextual variables in the survey weights.

The first simulation method (*clusterZ*) generates $Z$ from its relationship with $K$, measured as the intraclass correlation $\rho_{ZK}$. Then, $Y$ is computed from its relationship with $Z$, represented as $r_{ZY}$. In this case, the clustering level of $Y$ is determined by $Z$. The second simulation method (*clusterY*), first, generates $Y$ using the intraclass correlation $\rho_{YK}$, followed by $Z$ based on the correlation $r_{ZY}$. This method is equivalent to *clusterZ*, but here the target variable ($Y$) is used to establish the level of clustering. Finally, the third method (*cluserInd)* generates clustered $Z$ and $Y$ independently, based on their relationship with the clustering variable ($K$). Within each cluster, $Y$ and $Z$ variables are arranged to achieve a given correlation level ($r_{ZY}$).

The simulated data was generated using the R package fabricatr (Blair *et al.* 2018). The simulations followed two steps. First, the aggregate variable was computed, so the auxiliary $Z$ or the target $Y$ variables were generated alongside the clustering indicator ($K$)

given the intraclass correlation $\rho$. For ease of explanation, assume the scenario *clusterZ*, in which the value *z* for each population element *i* in a cluster *k* is defined as:

$$t_i \sim Bern(p_i)$$

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

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

[B.3]

where $p_i$ is the probability that an item has the characteristic of interest.

Second, a five-step process is followed to simulate the other variable, given a correlation coefficient. Let assume that variable $Z$ was already computed in the previous step, so the next step is to simulate $Y$ given a correlation $r_{ZY}$. The first step consists in computing the quantiles of $Z$:

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

[B.4]

where $F$ represents the empirical distribution of the variable ($Z$). In the second step, the quantiles are drawn from a normal distribution:

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

[B.5]

Third, a normal standardised distribution of the variable is generated ($Y_{std}$), given the correlation $r_{ZY}$, as follows:

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

[B.6]

Later, the quantiles of $Y$ are generated from the normal distribution:

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

[B.7]

Finally, the variable $Y$ is generated from the target distribution ($G$) and the values of $Y_q$:

$$Y = G(Y_q)$$

[B.8].

After simulating the data, samples were drawn given a level of bias (0.05; 0.1; 0.15; 0.20; 0.25) in the man of the target variable $Y$. Following this method guarantees that the missing data mechanism is either MAR or NMAR. Under MAR, the response probability—being selected in the sample—is totally or partially explained by $Z$. The NMAR mechanism occurs when the response probability is related to the outcome variable and explained by a set of unobserved variables. In the simulated data, the mean estimate of the variable $Y$ was biased. In that scenario, the case of MAR occurs when the auxiliary variable ($Z$) is related to the target variable ($Y$), while NMAR occurs when the value of that correlation is zero.

The values of the parameters used in the simulations are presented in Table B.1. Given the large number of factors included in the design, a sample of 500,000 populations was selected. The simulations were done using cloud computing in Microsoft Azure.

Table B.1. Parameters used in the simulation of the populations

| Parameter | Description | Values |
|---|---|---|
| | **Population** | |
| $K$ | Number of clusters | From 50 to 1000 in steps of 100 |
| $\bar{Y}$ | Population probability of the target variable ($Y$) | From 0.05 to 0.5 in steps of 0.05 |
| $\bar{Z}$ | Population probability of the auxiliary variable ($Z$) | From 0.05 to 0.5 in steps of 0.05 |
| $\rho_{YK}$ | Intraclass correlation of $Y$ and $K$ | From 0.05 to 0.95 in steps of 0.1 |
| $\rho_{ZK}$ | Intraclass correlation of $Z$ and $K$ | From 0.05 to 0.95 in steps of 0.1 |
| $r_{ZY}$ | Correlation between auxiliary and target variables | From 0.05 to 0.95 in steps of 0.1 |
| **Aggregate mechanism** | Aggregate data mechanism | clusterZ, clusterY and clusterInd |
| | **Sample** | |
| $\lvert B_{(\bar{y})} \rvert$ | Bias of $\bar{y}$ | 0.05; 0.1; 0.15; 0.20; 0.25 |
| $n$ | Sample size | 200; 500; 1000; 2000 |

## Survey weights

Two survey weights were computed to correct the bias of the survey estimates in the simulations. The first weight employed the auxiliary variable at the individual level (ID), which is the standard procedure and was used as a reference point. The second weight utilised the auxiliary variable aggregated (AD) by the clustering indicator $k$.

The first weight was computed using the auxiliary variable at the individual level (ID) and the population totals. The technique used to compute this weight was linear calibration. Starting from a design weight, this technique minimises the distance between the starting and the output weight while forcing the sample distribution to match the population totals on key variables (Lundstrom and Särndal 2001).

The second weighting was based on the aggregate auxiliary variable (AD). In this case, the same system was used to calculate the weight, but the information was the aggregate variable, a summary of the auxiliary variable in the cluster to which the case belonged. Specifically, to generate the aggregate variable, the mean of the auxiliary variable in each cluster was calculated, and then this contextual variable was divided into quartiles to facilitate its use in the calibration. Finally, the auxiliary variables were added to the sample data using the cluster of membership as a key.

**Evaluation of the effect of the weights**

Finally, the weighted estimates of the target variables using both adjustments were compared to the population mean. This evaluation consisted of comparing each weighted estimate to the unweighted one and measuring the relative reduction in bias. The measure of the relative change in bias (RCB) is:

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

[B.9]

where $|\boldsymbol{B}_{(\bar{y}_{\boldsymbol{w}})}| = |\bar{Y} - \bar{y}_w|$ represents the absolute values of the bias in the weighted estimate and $|\boldsymbol{B}_{(\bar{y})}| = |\bar{Y} - \bar{y}|$ is the absolute bias of the unweighted estimate. These measures of relative bias were modelled using OLS linear regression. The objective of this analysis was to assess the impact of the different factors manipulated at the simulation stage on the effectiveness of the weights. Table B.2 presents the descriptive statistics of the dependent variables and predictors included in the regression model. The interaction and quadratic terms were omitted to ease the interpretation of the table.

Table B.2. Descriptive statistics of the variables included in the regression models

| Variable | Cases | Mean | Std. dev. | Min. | Max. |
|---|---|---|---|---|---|
| **Dependent variables** | | | | | |
| **RCB (AD)** | 499,500 | -0.07 | 0.16 | -1 | 0.80 |
| **RCB (ID)** | 499,500 | -0.08 | 0.15 | -1 | 0.81 |
| **Predictors** | | | | | |
| $k = 50$ **(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 |
| **Mean Y (**$\bar{y}$**)** | 499,500 | 0.27 | 0.14 | 0.01 | 0.57 |
| **Mean Z (**$\bar{z}$**)** | 499,500 | 0.27 | 0.14 | 0.01 | 0.52 |
| **ICC Y (**$\rho_{YK}$**)** | 499,500 | 0.23 | 0.29 | 0.00 | 0.99 |
| **ICC Z (**$\rho_{ZK}$**)** | 499,500 | 0.22 | 0.28 | 0.00 | 0.87 |
| **Corr. XY (**$r_{ZY}$**)** | 499,500 | 0.17 | 0.20 | -0.16 | 0.96 |
| ***clusterZ*** **(ref.)** | 499,500 | 0.33 | 0.47 | 0 | 1 |
| ***clusterY*** | 499,500 | 0.33 | 0.47 | 0 | 1 |
| ***clusterInd*** | 499,500 | 0.33 | 0.47 | 0 | 1 |
| **Bias** | 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 |

## Results

This section focuses on the two regression models that assess the impact of the factor manipulated on the effectiveness of the weights. The first model considers the relative change in bias (RCB) when using the aggregate data (AD) to compute the weights. The dependent variable of the second model is the relative change in bias when using individual-level data to build the adjustments. The predictors included in the model are presented in Table B.2.

Figure B.5 presents the effect of each factor included in the simulation on the ability of the weights to correct the bias of the estimates. Each graph represents, on the horizontal axis, one of the relevant characteristics included in the simulations, while the vertical axis represents, in all cases, the relative change in bias (CRS). Each graph contains four lines, three corresponding to the weights made from aggregate data (AD) and one corresponding

to the individual weight (ID). The three lines of the weights computed with aggregate data represent the three data generation mechanisms used for the simulations: *clusterZ, clusterY and clusterInd.*

The results can be summarised in three findings: 1) the level of clustering of the target variable has a significant impact on the reduction of bias when aggregate auxiliary variables (AD) are used; 2) the level of correlation between the auxiliary and the target variables is also a relevant factor; and 3) the impact of these two factors is moderated by the data mechanism (i.e., *clusterZ*, *clusterY*, and *cluterInd*).

First, regarding the level of clustering of the target variable, Graph a) in Figure B.5 shows the relationship between the intraclass correlation of the target variable and the relative change in bias. The clustering of the target variable is irrelevant if the auxiliary variable employed to compute the weight is at the individual-level (ID). Hence, it has no impact on the reduction of bias. In contrast, the weights based on aggregate data (AD) show better performance as the level of clustering of the target variable increases. Still, this relationship varies depending on data generation mechanisms.

Table B.3. Predicted relative change in bias (RCB) of the estimates for different values of the ICC of Y

| | | **AD: *clusterZ*** | **AD: *clusterY*** | **AD: *clusterInd*** | **ID** |
|---|---|---|---|---|---|
| **ICC 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 |

Table B.3 expands the information of Graph a) and allows us to observe the influence of the clustering of the target variable conditional on the data generation mechanism. Under the *clusterZ* system, where the clustering of the target variable is fixed by the clustering of the auxiliary variable, the effect of the aggregate data weight is maximised. In this scenario, the increase by one decimal point of the ICC results in a reduction of 9 percentage points of the bias, which is significantly better than the *clusterY* (5 p.p.) and *clusterInd* (1 p.p.) scenarios. The mechanism used to generate the target variable explains these

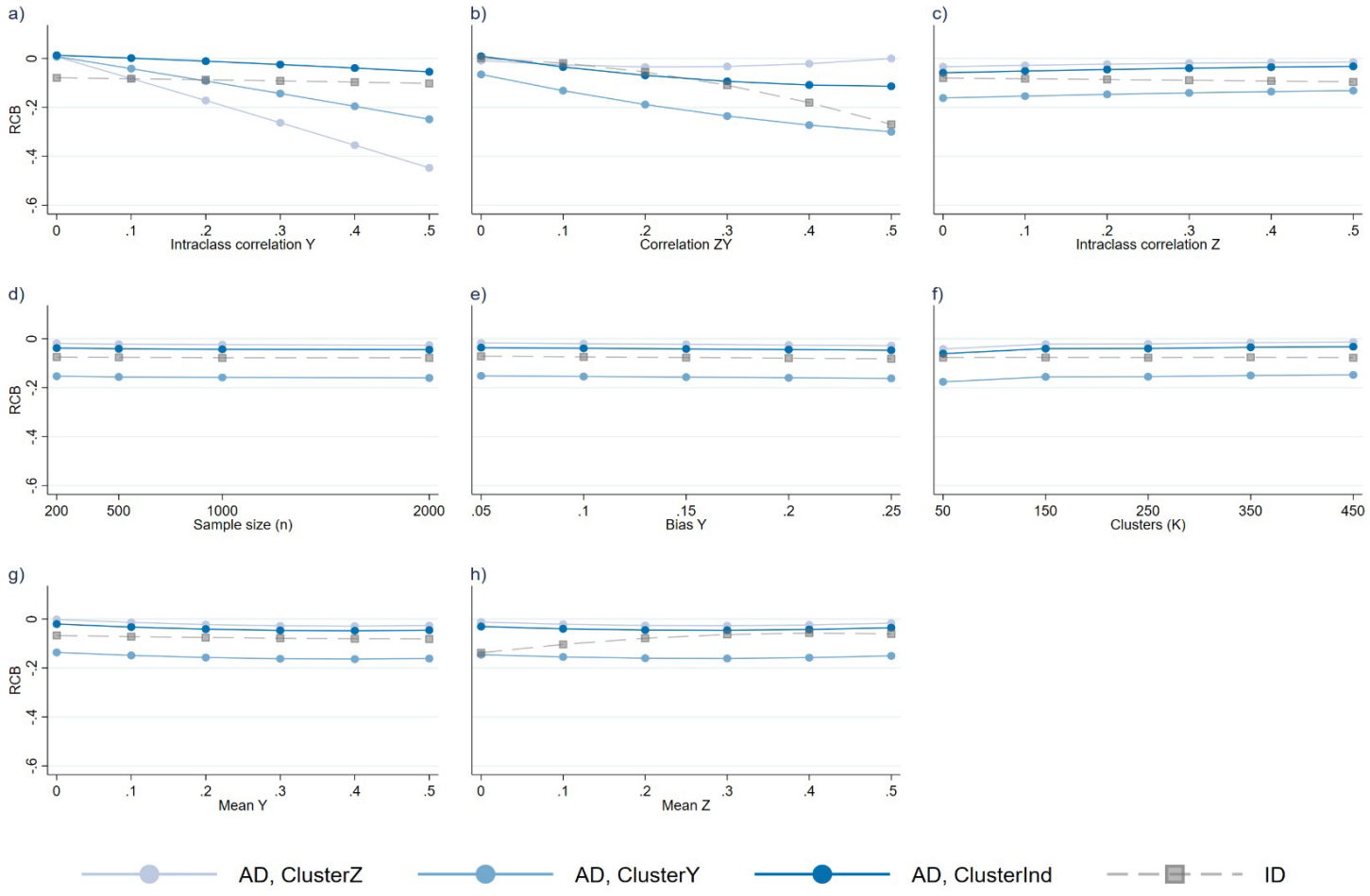
differences. Under the *clusterZ* mechanism, the target variable will be clustered if the auxiliary variable is clustered, and there is a high correlation between the auxiliary and target variables. In contrast, under *clusterY* the clustering of the target variable does not imply a high correlation with the auxiliary variable, which directly affects the effectiveness of the weight.

Second, another relevant factor is the correlation between the auxiliary and target variables. Graph b) represents how this correlation explains the effectiveness of the weights with aggregate or individual-level data. If the individual-level data (ID) is used for the weight, the higher the correlation, the more significant the reduction in bias. This factor is also relevant if aggregate data (AD) is used, especially under the *clusterY* data generation mechanism, where the impact is greater than using individual-level data (ID). Table B.4 reproduces the data underlying Graph b). For *clusterY*, the increase of the correlation of one decimal point to $r_{ZY} = 0.1$ produces a reduction in the RCB from -0.07 to -.013, while the same increase in the correlation has almost a null effect for the individual-level data (ID), a change from 0.0 to -0.02.

Table B.4. Prediction of the relative change in bias (RCB) for different values of the correlation between Z and Y

| | | **AD: *clusterZ*** | **AD: *clusterY*** | **AD: *clusterInd*** | **ID** |
|---|---|---|---|---|---|
| **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 |

Other factors included in the regression models are the intraclass correlation (ICC) of the auxiliary variable *Z* (Graph c), the sample size (d), the level of bias in *Y* (e), the number of clusters (f), the population mean of *Y* (g) and the population mean of *Z* (h). However, none of these has a significant effect on the performance of the weights to reduce bias.




**AD**: Weighted estimate with aggregate data; **ID**: Weighted estimate with individual level data.

Figure B.5. Predictions from the regression models

## Discussion

The first hypothesis (H1) of this paper posits that aggregate data will be useful to adjust the sample deviations derived from the lack of coverage or response. Contrary to the results of previous work (Biemer and Peytchev 2013; Butt and Lahtinen 2016), simulations show that under certain circumstances, the use of aggregate data can work and even improve the results obtained by using individual data. However, those circumstances, the clustering of the target variable and its correlation with the auxiliary variable, are unlikely to be present in the data generally used by social scientists. Regarding the clustering of the target variable, empirical work on factual and attitudinal variables has shown intraclass correlations lower than 0.1 (Kish, Groves and Krotki 1976). On the other hand, a potential advantage of using aggregate data is that they are more accessible, and there is a wider variety of sources, which makes it easier to find auxiliary variables correlated with propensity to respond and the variables of interest. However, in most studies it is difficult to find auxiliary variables highly correlated with the variable of interest and the propensity to respond. Therefore, with respect to H1, the use of aggregate information to correct for bias can be effective if subjects are grouped according to the variable of interest and that variable is correlated with the auxiliary variable.

The second hypothesis (H2) is based on the findings of the empirical work of Biemer and Peytchev (2013), which shows that the clustering of the auxiliary variable is needed for the aggregate data adjustments to work. However, the results of the data simulations slightly depart from this idea. The target variable also needs to be clustered for the adjustments to remove the bias from the estimate successfully. From all the factors included in the simulations, the intraclass correlation of the target variable has emerged as the most relevant. However, as discussed above, it is unrealistic to expect a high level of geographical clustering in most of the variables used in social sciences. This hypothesis is complemented by H3, which focuses on the correlation between the target and auxiliary variables. This correlation has been found to be crucial to explain the effect of survey weights using individual-level data (Groves and Couper 1998). Yet, it is also relevant in cases where aggregate data is employed to compute the weights.

The data mechanism used to generate the data (H4) explains how the survey adjustments work. This paper compares three data generation mechanisms, *clusterZ*, where the auxiliary variable is clustered; *clusterY*, where the target variable is aggregated; and *clusterInd*, where the target and auxiliary variables are aggregated independently. Above

I have discussed the importance of the clustering of the target variable and the correlation between the auxiliary and target variables to adjust the sample using a survey weight. But these two features are affected by the mechanism underlying the data generation. Under *clusterZ* the performance of the weights is better compared to the other two scenarios—*clusterY* and *clusterInd*. This is because the clustering of the target variable requires both the clustering of the auxiliary variable and the correlation between the target and the auxiliary variable.

Butt and Lahtinen (2016) found that the level at which the data was aggregated was irrelevant. The critical point was that the data was aggregated regardless of the level. In the simulations, one of the manipulated variables was the number of clusters (H5). The results show that the number and size of the groups are not related to the performance of the weights. The relevant feature is the clustering of the target variable within the groups rather than their size and number. However, in each simulation, all clusters have the same size, and only a clustering strategy was implemented in each simulated population. These characteristics should be manipulated to fully assess this hypothesis.

This research aimed to assess the role of the magnitude of the bias on the performance of the weights (H6). The extent of the bias would theoretically allow for a more significant correction. However, the results confirm that the relevant characteristics are the correlations between the auxiliary and the target variable, as well as the level of clustering of the target variable.

## Conclusions

The first conclusion is about the appropriateness of using aggregate data to adjust for nonresponse and coverage biases in surveys. The second is about the limitations and future of this research.

Aggregate data has two advantages: there is a wide variety of sources, and they are more accessible than microdata because they present fewer privacy issues. The question is how aggregate data is most valuable: population totals in individual calibrations or contextual variables. Population totals used in individual-level calibrations have the limitation that the information for each sample element must be known to make the adjustment. This can be costly and is not always in the research plan at the design stage. In contrast, the use of contextual variables allows for more flexibility, as a wide variety of predictors can be used once the fieldwork has been completed, without the need to collect

any extra information beyond the geographical unit to which the sample element belongs. However, this advantage is overshadowed by the additional assumptions that must be met for aggregate variables to reduce the level of bias: 1) the aggregate auxiliary variable must be correlated with the probability of response and the target variable, and 2) the estimated variable must be clustered. In particular, the second is an unlikely assumption, so we suggest that researchers check this requirement before considering the use of aggregate variables in the construction of weights.

This research has several limitations. First, the results, which show the potential of aggregate data under certain circumstances, contrasts with the empirical evidence so far. The previous works where contextual variables were used to assess or correct nonresponse were high-quality probability surveys (Biemer and Peytchev 2013; Butt and Lahtinen 2015; 2016). Nonresponse or coverage bias may not be an issue in these studies. Nevertheless, in other research where the incidence of these phenomena is higher, contextual variables may help correct for bias. More research is needed on this front to bring the results of simulations closer to the context of social research. Another open question concerns the influence of the cluster size and the level of aggregation. Further work is needed on the effect that the differential size of clusters may have. Moreover, this research has not addressed the effect of adjustments with aggregate data on estimation errors. Finally, there is a need to develop empirical measures to help researchers decide on the appropriateness of using aggregate data in adjustments.

## Bibliography

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

BEULLENS, K., LOOSVELDT, G., VANDENPLAS C. , and 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., and 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., and 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., and 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., and 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., and SAVAGE, M. (2014): "After the crisis? Big Data and the methodological challenges of empirical sociology". Big Data and Society, 1(1), 205395171454028. https://doi.org/10.1177/2053951714540280

BUTT, S. , and LAHTINEN, K. (2015): Using auxiliary data to model nonresponse bias The challenge of knowing too much about nonrespondents rather than too little? presented at the International Workshop on Household Nonresponse 2015, 02 Sep 2015 - 04 Sep 2015, Leuven, Belgium.

BUTT, S., and 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 ), London, 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., and 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., and 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., and 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., and 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., and LITTLE, R. (2002): "Survey nonresponse in design, data collection and analysis", in Survey nonresponse, New York, Wiley & Sons, 3-26.

ELLIOTT, M. R., and 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., and COUPER, M. (1998): Nonresponse in household interview surveys, New York, Wiley and Sons.

GROVES, R. M., and 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., and ROßMANN, J. (2018): 'The effects of propensity score weighting on attrition biases in attitudinal, behavioural, 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., and HARTIG, H. (2019): Response rates in telephone surveys have resumed their decline, available at https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/ [accessed: 7-09-2020].

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

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

LEPKOWSKI, J. M., MOSHER, W. D., GROVES, R. M., WEST, B. T., WAGNER, J., and 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., and LEMESHOW, S. (2013): Sampling of Populations: Methods and Applications, New 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. and RUBIN, D. (1987): Statistical Analysis with Missing Data, Wiley, New York. 381. https://doi.org/10.1002/9781119013563

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

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

MANFREDA, K. L., BERZELAK, J., VEHOVAR, V., BOSNJAK, M., and 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., and KENNEDY, C. (2018): For Weighting Online Opt-In Samples, What Matters Most? , Washington, Pew Research.

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

OLSON, K., and 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., and PHILLIPS, M. (2013): British Social Attitudes 30th Report, London, 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., and 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., and 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., and PÉREZ-AGOTE, A. (2013): La nueva pluralidad religiosa, Madrid, Ministerio de Justicia.

SÄRNDAL, C., and LUNDSTRÖM, S. (2005): Estimation in Surveys with Nonresponse.

SÄRNDAL, C. (2007): "The calibration approach in survey theory and practice", Survey Methododogy, 33(2), 99-119.

SINGER, E., GROVES, R.M., and 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., and 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., and ELLIOTT, M. R. (2018): 'Model-assisted calibration of non-probability sample survey data using adaptive LASSO', (12).

VALLIANT, R. , DEVER, J. A., and 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

# Appendix I: Regression models

Table B.5. OLS models to determine the bias reduction in the aggregated data (AD) and individual (ID) data scenario

| | AD | ID |
|---|---|---|
| Mean Y | -0.119*** | 0.390*** |
| | (0.004) | (0.003) |
| Mean Y^2 | 0.206*** | -0.472*** |
| | (0.007) | (0.005) |
| Mean Z | -0.147*** | -0.053*** |
| | (0.004) | (0.003) |
| Mean 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) |
| Bias | -0.053*** | -0.050*** |
| | (0.002) | (0.001) |
| ClusterInd | 0.016*** | -0.002*** |
| | (0.001) | (0.000) |
| ClusterY | -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) |
| ClusterInd*Rho Y | 0.418*** | 0.188*** |
| | (0.007) | (0.005) |
| ClusterY*Rho Y | 0.739*** | 0.200*** |
| | (0.007) | (0.005) |
| ClusterInd*Rho Z | | -0.145*** |
| | | (0.005) |
| ClusterY*Rho Z | | 0.017*** |
| | | (0.001) |
| ClusterInd*Corr. ZY | -0.052*** | 0.021*** |
| | (0.002) | (0.002) |
| ClusterY*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) |
| Mean Y*Mean Z | 0.022*** | 0.002 |
| | (0.006) | (0.005) |
| Intercept | 0.044*** | -0.038*** |
| | (0.001) | (0.001) |
| F | 39869.39 | 67714.39 |
| Degrees of freedom | 28 | 30 |
| P-value | 0.000 | 0.000 |
| R-square | 0.69 | 0.80 |
| N | 499500 | 499500 |

AD: Aggregate data; ID: Individual level data.
Interaction between methods ClusterY, clusterInd and rho Z were omitted from the first model due to the lack
of cases and the negative effect on the estimates.
* p<0.05, ** p<0.01, *** p<0.001

# Appendix C: Aggregate administrative data and estimation from nonprobability samples

Cabrera-Álvarez, Pablo. 2021. "Datos Administrativos Agregados y Estimación a Partir de Muestras No Probabilísticas." Revista Internacional de Sociología 79(1):e180. doi: 10.3989/ris.2021.79.1.19.350.

**Abstract**

In recent years survey research has been marked by the spread of internet research and the use of nonprobability samples, and the steady decline in response rates. Adjustments are needed to infer the survey estimates to the population. However, these adjustments are statistical models that require a set of auxiliary variables, information about the whole population. This paper tests the potential of administrative data aggregated at the municipality level to adjust two surveys from a panel of internet users, the AIMC-Q panel, promoted by the Spanish Association for Media Research (AIMC). The results show that the ability of aggregate administrative variables to reduce the bias of the estimates is minimal.

**Keywords:** survey methodology, nonprobability samples, machine learning, selection bias, administrative data.

## Introduction

In recent years two phenomena have affected the drift of survey research: the expansion of the Internet and the more frequent use of nonprobability samples, and the sustained fall in response rates. Both phenomena affect the inference process—the possibility of knowing the characteristics of the population from the sample—underlying survey research. However, just as inference can be made from selecting a probability sample, it is also possible to use statistical models to eliminate or reduce the bias present in the estimates after data collection. These statistical models require auxiliary variables available for the population as a whole. Moreover, to be effective, these variables have to be related to the survey variables of interest and the response probability. This paper focuses on one source of auxiliary variables: aggregate administrative data.

Aggregate administrative data, as opposed to other types of data, such as administrative microdata or commercial data, are more numerous, accessible and varied. Indeed, some administrative data sources, such as the census, have been used for decades to obtain population totals for weighting surveys. However, little research has been done on the use of contextual variables, such as the characteristics of the respondent's neighbourhood or municipality, to adjust sample deviations.

The main objective of this research is to explore the potential of using aggregate administrative data as contextual variables to correct for bias in estimates derived from a nonprobability survey. To this end, three sets of auxiliary variables and three estimation methods are compared to adjust two web surveys conducted from a panel of Internet users promoted by the Association for Media Research in Spain (AIMC).

This paper has five sections. The first section discusses the theoretical and empirical background. The second presents a series of hypotheses and the third, the data and methodology employed. The fourth contains the results. Finally, the results are discussed, and conclusions are presented.

## Background

In recent years we have witnessed an exponential growth in the number of online surveys in social and market research (Hays, Liu and Kapteyn 2015; Blom *et al.*

2016). A significant proportion of these surveys are based on samples drawn from panels of internet users recruited using nonprobability methods (Callegaro, Manfreda and Vehovar 2015). The use of these procedures may cause the emergence of selection bias, which is the existence of systematic differences between those who are part of the panel and those who are not. Selection bias is composed of two distinct issues: coverage bias and self-selection bias. Coverage bias occurs when part of the population is not eligible to participate in the study (Weiseberg 2005), such as households without internet access in a general population web survey. Self-selection bias refers to the differential probability of population elements voluntarily joining, for example, a panel of internet users (Bethlehem and Biffignandi 2011; Blom, Gathmann and Krieger 2015).

The drop in response rates also casts doubt on whether the use of probability samples is sufficient to guarantee the inference process (de Leeuw, Hox and Luiten 2018; Elliott and Valliant 2017). The problem with nonresponse lies in the fact that some groups have a higher probability of participation, and such systematic differences lead to biased estimates (Groves and Couper 1998).

**From design inference to model inference**

These two elements, the expansion of internet research and the deteriorating quality of probability samples, challenge the inference process based on random samples (Pasek 2015). Consequently, increasingly complex adjustments are needed to ensure data quality. Brick (2011), in his work on the future of survey sampling, distinguishes between two types of inference. The first is based on the use of probability samples and called *design-based inference*. The second relies on statistical models fitted after data collection and receives the name *model-based inference*.

*Design inference* is based on the probability mechanism underlying the selection of a random sample (Kish 1965; Neyman 1934). Since the bulk of sampling theory was developed in the mid-20th century, most surveys have relied on probability principles to select representative samples of the population (Baker *et al.* 2013). In a probability sample, all members of the population have a known probability of being selected that is non-zero (Levy and Lemeshow 2013). If all sample elements respond to the survey, estimates can be inferred to the population with a certain degree of precision.

Increasingly, selecting a probability sample does not guarantee the inference process because of self-selection and nonresponse bias. An example is surveys conducted from panels of Internet users recruited using nonprobability methods. In these cases, one can choose to rely on statistical models to support the inference process—inference *from models*—where the statistical apparatus controls the selection bias (Valliant, Dorfman and Royall 2000). There are different methods to enable inference from nonprobability samples: quasi-randomisation models, superpopulation models, and a combination of both, the doubly robust estimator (Elliott and Valliant 2017; Valliant 2019).

In the quasi-randomisation method, a statistical model is used to derive the pseudo selection probabilities of the sample elements. This statistical model generally uses data from a reference probability survey (Gummer and Roßmann 2018; Lee and Valliant 2009; Pasek 2016; de Pedraza *et al.* 2010; Valliant and Dever 2011). Another method to derive the pseudo selection probabilities is matching the nonprobability survey cases with those in a probability sample using *propensity score* (Elliott and Valliant 2017; Mercer, Lau and Kennedy 2018; Ferri-García and Rueda 2018). Calibration or post-stratification methods are also valid to calculate pseudo inclusion probabilities using aggregate auxiliary information such as population totals (Dever, Rafferty and Valliant 2008; Pasek 2016; Peytchev, Presser and Zhang 2018).

The superpopulation method involves fitting a model to predict the variable of interest in the nonprobability sample and projecting it to the population (Buelens, Burger and Brakel 2018; Dorfman and Valliant 2005; Wang *et al.* 2015). For this method to be effective, the sample and the population must follow a common model that can be discovered from the survey. Superpopulation models are less flexible than quasi-randomisation since, in theory, it is necessary to generate a weight for each variable of interest. In recent years, machine learning techniques have been used to develop superpopulation models (e.g., Chen, Valliant and Elliot 2018).

It is also possible to combine the two strategies above and perform a double adjustment. This involves calculating pseudo selection probabilities which, in turn, are used to fit the superpopulation model (Kang and Schafer 2007; Brick 2015). The bias of the estimates will be reduced to the extent that one or both models are correctly specified. In recent years, several studies have compared some of these adjustment

strategies. Ferri-García and Rueda (2018) found that the combination of *propensity score* and calibration methods generated more efficient adjustments. Valliant (2019), based on simulations, compared the effectiveness of several estimation strategies in nonprobability surveys. He compared quasi-randomisation, superpopulation models and multilevel regression with poststratification, finding that a combination of quasi-randomisation with superpopulation models was the best option to reduce the level of bias in the estimates.

All modelling strategies have one thing in common, they require auxiliary variables to be correlated with both the variables of interest and the probability of participating in the survey (West and Little 2013). A recent study using surveys of a panel of internet users in the United States shows that the specification of the models is more relevant than the technique used to fit them (Mercer, Lau and Kennedy 2018).

**Auxiliary variables to correct for self-selection and nonresponse biases**

Traditionally, the information to construct survey adjustments came from official statistics and surveys such as the census (Barboza and Williams 2005). However, in recent years multiple data sources have emerged: commercial data (Peytchev and Raghunathan 2013; West *et al.* 2015), paradata (Kreuter 2013), georeferenced data (Lahtinen, Kaisa and Butt 2015) and administrative data (Couper 2013). This paper focuses on one of these sources: aggregate administrative data.

Administrative data are products or by-products generated in the interaction of public administration with citizens, businesses or other organisations (Playford *et al.* 2016). Woollard (2014) states that administrative data are collected to organise, manage or monitor services. However, this information can also be useful for answering research questions in social sciences. These data have some characteristics that make them good candidates for auxiliary variables. First, they tend to be subject to less error than survey data, although the definition of the concepts and the instruments used to collect the information may differ (Connelly *et al.* 2016). Second, administrative data have a wide coverage, in many cases reaching the entire population (Künn 2015). As a counterpoint, access to the data depends on the will of the administration and may be restricted to ensure the privacy of citizens or organisations (Dibben *et al.* 2015; Stevens

and Laurie 2014). But this disadvantage affects aggregate administrative data to a lesser extent.

Aggregate administrative data have played an important role in correcting for selection and nonresponse bias for decades. For example, census data are often used to adjust the sample distribution for sex and age (*e.g.* Morris *et al.* 2016; Park *et al.* 2013). In order to make these adjustments, the researcher needs to know in advance which variables will be used to calibrate the final sample and include them in the questionnaire.

A greater variety of administrative data has recently opened the door to exploring how to combine them with survey data (Smith 2011; Smith and Kim 2013; Lohr and Raughnathan 2017). One possibility is to use aggregate administrative data as contextual variables, summary information about the environment of the sampled items, such as the neighbourhood or municipality. This could be, for example, the percentage of luxury cars, the prevalence of voting for a particular party, or the value of buildings in the area where the sample unit resides. In addition to being very varied in its subject and easily accessible, this contextual information can help fit the models that correct bias in the survey estimates.

However, few studies have used aggregate administrative data as contextual variables to address selection or nonresponse bias. Biemer and Peytchev (2012, 2013) used aggregate administrative data at the census tract, municipality and county levels to detect and correct for the effect of nonresponse in the National Comorbidity Survey, a telephone survey conducted in the United States. They concluded that the use of administrative data was not effective in improving estimates. More recently, a study has assessed the effectiveness of administrative data aggregated at the census tract or municipality level as contextual variables to adjust for the nonresponse bias present in the European Social Survey in the United Kingdom (Lahtinen, Kaisa and Butt 2015). The research results concluded that aggregate data were not related to the probability of response in this survey. Despite the poor results of these investigations, it is worth noting that both cases are based on probability surveys where the presence of bias is usually attenuated by the use of carefully designed fieldwork procedures.

## Research hypotheses

The theory and research mentioned in the previous section leads to a set of hypotheses presented in this section.

**H1**. *The use of administrative data aggregated at the municipality level as auxiliary variables results in a greater reduction in the level of bias compared to the use of individual-level sociodemographic variables*.

The advantage of using aggregate administrative data lies in their availability and variety. In this paper, as detailed in the following section, three sets of auxiliary variables are compared to generate the survey weights: sociodemographic, administrative and the combination of both. The sociodemographic variables are usually used to weight the survey used in this research (sex, age, region of residence and size of municipality). The administrative variables are more numerous and cover a wide range of topics, from income to electoral behaviour. This greater variety suggests that some of the variables will effectively reduce the level of bias in the estimates.

**H2**. *The combination of administrative auxiliary variables and sociodemographic variables is the most effective alternative to reduce the bias of the estimates*.

**H3**. *The effectiveness of auxiliary variables in reducing bias in the estimates is independent of the strategy used to generate the weighting.*

A set of weights has been developed to analyse the effectiveness of aggregate administrative variables using three strategies: quasi-randomisation, superpopulation models, and the doubly robust method. Despite the differences between the estimation strategies, similar patterns are expected to be observed: administrative aggregate data, being more varied, lead to more effective weights.

**H4**. *The standard errors of the estimates will be smaller when administrative variables are used to calculate the weights*.

The error of the estimates was calculated using a linearised method employed in sampling with replacement and a *jackknife* replication method. The weights can reduce the error of the estimates if the auxiliary variables correlate with the probability

of response and the variable of interest. In this case, I hypothesise that the use of administrative variables to calculate the weights will make the estimates more efficient.

## Data and methodology

This section presents the data and methodology of the analysis carried out using two AIMC-Q panel surveys, the General Media Study (EGM) and aggregate administrative data.

### Sources of data

This analysis uses three sources of data: aggregate administrative data used as contextual variables, data from the General Media Study used as a population reference to calculate the bias of the estimates, and two surveys from the AIMC-Q panel.

The collection of administrative data was designed using the national list of statistical operations from the National Statistics Institute (INE). This list groups all the data collected and produced by the central government and their levels of aggregation. This paper uses data aggregated at the municipality level for two reasons. First, data sources available below the municipality level, such as the census, are scarce. Second, and more importantly, the surveys used in this paper only contained identifiers of the respondent's municipality, making it unfeasible to use aggregate information at a lower level. The data aggregated at the municipality level were included in a database with a total of 1,099 variables. These variables came from various sources, including the 2011 census, the population registers, income tax statistics, electoral results, unemployment data, or information on the make of vehicles registered in the municipality, as shown in Table C.1.

The EGM is a study carried out in three waves each year whose main objective is to collect data on the media consumption of residents of Spain aged 14 or older. This survey is formed by a multimedia sample, that looks at the consumption of radio, press, and television, and three other single-media samples to collect information from each medium (radio, press or television). The multimedia sample contains 30,000 responses collected over the year, while the specialised samples range from 13,000 (television) to 45,000 (press) responses. In the multimedia study, data collection consists of

interviewing informants face-to-face in a random selection of households. In the single-media surveys, data is collected by telephone using a combination of landline and mobile phones. Data from the different phases are adjusted to preserve the representativeness of the sample. This paper uses the estimates from the first (January-March) and second (April-June) waves of 2017 as benchmarks to assess the estimates from the AIMC-Q panel. The fact that the benchmark comes from a survey is a limitation of this study. Despite the large sample size of the EGM and the use of probability sampling, the estimates might be biased. However, in the absence of population benchmarks, it is common to assume that estimates from a survey such as the EGM will be less biased than those from an internet panel (Schonlau *et al.* 2009; Yeager *et al.* 2011).

Table C.1. Administrative variables included in the research

| Source of data | Institution | Year (period) | Variables | Number of variables |
|---|---|---|---|---|
| **Register of municipalities** | National Statistics Institute | 2018 (twice per year) | Municipalities, total population, population by sex. | 2 |
| **Population and housing census** | National Statistics Institute | 2011 (every ten years) | Sex, age, marital status, educational level, country of birth, nationality. Dwellings by size, type of property, number of rooms and persons residing. | 145 |
| **Population register** | National Statistics Institute | 2018 (twice per year) | Sex, age, nationality, country of birth, relationship to place of birth and residence. | 303 |
| **Cadastre** | Cadastral Office | 2016 (annual) | Surface area according to use, average land value, land typology | 20 |
| **Local taxes** | Ministry of Finance | 2016 (annual) | Local taxes. | 21 |
| **Personal income tax** | Ministry of Finance | 2016 (annual) | Tax base, filers, holders, deductions, average gross and average disposable income. | 32 |
| **Local budget** | Ministry of Finance | 2016 (annual) | Local budget statistics. | 32 |
| **Unemployment and registered contracts** | Ministry of Labour and Social Security | 2018 (monthly) | Unemployed and contracts concluded. | 23 |
| **Driver census** | Department of Traffic | 2017 (annual) | Drivers, sex. | 2 |
| **Vehicle census** | Department of Traffic | 2017 (annual) | Passenger cars, mopeds, motorbikes, make of passenger car. | 58 |
| **Vehicle enrolments** | Department of Traffic | 2017 (annual) | Enrolments. | 1 |
| **Accidents** | Department of Traffic | 2017 (annual) | Accident victims. | 1 |
| **Elections** | Home Office | 2016 (annual) | Local, European and general election results (1977-2016). | 459 |

To assess the impact of aggregate data on the survey estimates, we use two surveys from a probability panel of internet users managed by AIMC. This experimental panel of internet users, which started in 2013, is comprised of EGM respondents with internet access who agreed to participate. In 2017 the panel had 4,514 members, who are invited to complete surveys periodically. This work focuses on two surveys, the first is regarding press consumption (n = 2,013), whose fieldwork took place in March 2017, and the second is on radio consumption (n = 2,058), carried out in June 2017. It should be noted that this panel is composed of a probability subsample of the EGM designed to study the population of internet users residing in Spain. However, in this research, it is assumed that the same panel can cover the general Spanish population using model-based adjustments. The objective here was to emulate other panels of internet users whose members are recruited using nonprobability methods.

Table C.2 presents the sample profiles of both surveys and the distribution of the same variables in the population. In both surveys, the group of over-64s is underrepresented, while individuals aged 35-54 are overrepresented. Those living in provincial capitals are also overrepresented—41% of the sample in both surveys, but only 32% of the population.

Table C.2. Sample profile of the AIMC-Q panel press and radio surveys and population distribution

| | | Population | Press | Radio |
|---|---|---|---|---|
| **Sex** | Man | 48.6 | 55.0 | 55.9 |
| | Woman | 51.4 | 45.0 | 44.1 |
| | | | | |
| **Age** | 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 or more | 21.9 | 6.7 | 5.9 |
| | | | | |
| **Size of municipality (inhabitants)** | Less than 2000 | 6.1 | 4.6 | 3.7 |
| | 2001 to 5000 | 6.6 | 5.0 | 4.8 |
| | 5001 to 10000 | 8.3 | 6.2 | 6.4 |
| | 10001 to 50000 | 26.5 | 22.4 | 23.3 |
| | 50001 to 200000 | 15.3 | 15.1 | 14.5 |
| | 200001 to 500000 | 5.0 | 5.4 | 5.9 |
| | Provincial capital | 32.2 | 41.3 | 41.5 |

## Methodology

The effectiveness of nine weights was tested to assess their potential to reduce the bias in the estimates. These weights were the product of using three sets of auxiliary variables and three estimation methods. The three sets of auxiliary variables are a selection of sociodemographic variables (SD), the aggregate administrative data (AD) and a combination of both (SD+AD). Each set has been used to estimate a quasi-randomisation model (QR), a superpopulation model (SP) and a doubly robust adjustment model (DR). In the superpopulation and doubly robust adjustment models, a weight was generated for each of the 13 target variables.

*Auxiliary variables*

The first set of auxiliary data corresponds to the sociodemographic (SD) variables used to adjust the panel surveys. This is a basic set that includes sex, age groups, seven categories of habitat size and the region of residence. These variables are commonly used because population totals are available to researchers, and the information is usually complete for all elements of the sample. However, the effectiveness of this set of variables in reducing bias is not guaranteed as there may be a weak relationship between the sociodemographic variables, the probability of participating and the variables of interest. This is a baseline scenario to compare the effectiveness of aggregate administrative variables.

The second set is made up of 1,099 aggregated administrative variables (AD) collected from the national list of operations held by the National Statistical Office. This is a large set of easily accessible contextual variables that could improve the estimates of other studies if they prove useful in adjusting the survey estimates. Before being used in the models, these variables were treated in three steps: 1) in some cases, population information was not available for all municipalities, so missing values were imputed using the nearest neighbour method; 2) the variables, which were totals for each municipality, were converted into percentages; and 3) to fit the regularised regression models they were scaled and standardised. Finally, the research design also includes a combination of both sets of administrative and sociodemographic variables (SD+AD).

*Estimation techniques*

The three estimation techniques used in this paper—computation of pseudo selection probabilities, superpopulation models and doubly robust estimator—usually rely on linear models such as logistic regression. Here, however, the large number of auxiliary variables has led to replacing the generalised linear models with a machine learning technique: regularised regression.

Regularised regression models

In recent years different machine learning techniques have been used to adjust survey estimates: *random forest* (Valliant, Dever and Kreuter 2018), support vector machines and neural networks (Buelens, Burger and Brakel 2018; Buskirk *et al.* 2018), and regularised regression (Chen, Valliant and Elliott, 2018). The main reason to adopt the regularised regression model in this research is because of its effectiveness for the automatic selection of predictors in the presence of multicollinearity.

Generalised linear models are problematic in the presence of a large number of predictors; assumptions such as the absence of multicollinearity are likely to be broken. Regularised regression is based on the idea that a selection of independent variables contains the most relevant effects of the model (Hastie, Tibshirani and Wainwright 2015). A penalty term in the objective function allows the identification of the most relevant variables. This term limits the magnitude of the coefficients so that the coefficients can only increase if there is a comparable decrease in the objective function.

The most widespread penalties are *ridge*, *lasso* and *elastic net,* the final one containing a combination of the other two. Here we describe the *elastic net* penalty applied to logistic regression, which is the model used in this paper (Friedman, Hastie and Tibshirani 2010). To model the variable $y$, which takes values 0 and 1, from a vector of predictors $\boldsymbol{x_i}$:

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

In the model using the *elastic net* penalty, the objective function used to fit it includes the penalty $\lambda$ which varies between 0 and $+\infty$ and a parameter $\alpha$ taking values

between 0 and 1. The parameter $\alpha$ determines the extent to which the *ridge* $\frac{\|\beta\|_2^2}{2}$ or *lasso* $\|\beta\|_1$ penalties, or a combination of both, is applied:

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

[C.2]

The models using the *elastic net* penalty have been calculated using the R package *glmnet.*

Quasi-randomisation model (QR)

In the quasi-randomisation method, to calculate the pseudo selection probabilities, a sample selected from the population data was used as the reference survey (Valliant, Dever and Kreuter 2018). The baseline survey (n = 10,000) was combined with the samples of the press and radio surveys. The combined dataset had a variable indicating whether each case came from the panel survey, in which case it took the value of 1, or whether it came from the benchmark survey, which took the value of 0.

The variable indicating the origin of the sample was modelled using a regularised logistic regression. Three models were run for each survey, one with each set of auxiliary variables (SD, AD, SD+AD). To determine the values of $\alpha$ and $\lambda$ needed to fit the model, ten cross-validations were calculated for six different values of $\alpha$. By default, *glmnet* calculates the model for a set of 100 values of $\lambda$. The values of $\alpha$ and $\lambda$ that resulted in the lowest classification error were used to compute the final model that predicted the probability of being part of the panel sample. The final weight was calculated as the inverse of that probability:

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

[C.3]

in which $\hat{\pi}(\boldsymbol{x}_i)$ represents the estimated probability of being part of the panel survey using a vector of auxiliary variables $\boldsymbol{x}$.

Superpopulation models (SP)

To calculate the weights with superpopulation models $\mathbf{w}^{\mathrm{SP}}$ the calibration method was adapted from an adaptive *lasso* model proposed by Chen, Valliant and Elliott (2018). The method used involves 1) fitting a regularised regression model with an *elastic net* penalty to predict the variable of interest in the sample; 2) this model is projected into the population to predict the values of the variable of interest; 3) weights are generated from a calibration model using the total of the predicted variable in the population as a reference.

In the model-assisted calibration (Deville and Särndal 2005; Wu and Sitter 2001), the distances between the design weights $d_i$ and the final weights $w_i$ are minimised by the function $g$ where $q_i$ is a constant independent of the design weight:

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

[C.4]

with the conditions that $\sum_{i\in s} w_i^{SP} = N$ y $\sum_{i\in s} w_i^{SP}\hat{y} = \sum^{N}\hat{y}_i$. Assuming $q_i = 1$ and that $g$ corresponds to the chi-squared distance $g(w_i^{SP}, d_i) = (w_i^{SP} - d_i)^2/d_i$:

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

[C.5]

where $\mathbf{d}$ are the design weights of the sample, $\mathbf{D}$ corresponds to the matrix on whose diagonal are the design weights, $\mathbf{M} = [\mathbf{d}, \sum_{i\in A}\hat{y}]$ and $\mathbf{T}^{\mathrm{M}} = (N, \sum^{N}\hat{y})$.

In this case, for each dependent variable and set of auxiliary variables -SD, AD and SD+AD- a weight was generated. The procedure described in the previous section was followed to construct the regularised logistic regression models.

Doubly robust (DR) estimator

The third estimation strategy combines the two previous ones, the quasi-randomisation model and the superpopulation model. Two steps were followed to apply this strategy. First, the weight from the quasi-randomisation model for each set of auxiliary

variables was used as the starting point. Subsequently, the final weight was derived from a weighted superpopulation model, following the steps explained in the previous section.

Standard error of estimates

The standard errors of the estimates were calculated following two methods: a linearised procedure designed for random sampling with replacement and an abbreviated *jackknife* replication procedure. The linearised method has been proposed as an alternative to approximate the error when the estimates are derived from superpopulation or quasi-randomisation models (Valliant, Dever and Kreuter 2018). The advantage of using this estimator is that it does not require excessive computational resources and is implemented in most statistical software. The standard error of the estimator is:

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

[C.6]

in which $\hat{z}_i = w_i z_i$ and $z_i$ is a measure of deviation associated with $\hat{y}$. The main drawback of this method is that it does not consider that the weights have been constructed using a random sample. For example, in the case of quasi-randomisation, the pseudo selection probabilities are estimates derived from using a reference sample.

The *jackknife* procedure consists in replicating the estimation $n$ times, excluding one case at a time, in order to calculate the deviation from the average estimates. Some empirical evidence supports this replication method to calculate the variance of estimates made from nonprobability samples (Wolter 2007; Valliant 2019). However, this method is computationally intensive—each regularised regression model has to be fitted *n* times to estimate the variables of interest. To overcome this issue, Valliant (2019) proposed a shortened version in which cases are randomly grouped into sets, and one is excluded at a time. The *jackknife* standard error is defined as:

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

[C.7]

where $\hat{\bar{y}}_{(j)}$ is the estimate of the mean of the variable of interest excluding the units of the group $j$. The number of groups $J$ was set to 50. Hence, for each estimate, all weights were computed 50 times to calculate $se_j$.

*Evaluation of the weights*

A set of 13 factual variables present in both the EGM and the AIMC-Q panel surveys were used to evaluate the impact of the auxiliary variables and the weights. The evaluation consisted in calculating a weighted measure of relative bias, which compares the EGM estimate with the panel survey estimate:

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

[C.8]

where $\hat{\bar{y}}_{EGM}$ is the estimate of the mean or proportion of the variable in the EGM and $B_r$ is a measure of the relative bias of each estimate:

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

[C.9]

where $\hat{\bar{y}}$ is the estimate of the mean of the target variable.

## Results

The two AIMC-Q panel surveys—radio and press—were used to test the potential of the aggregate administrative data to adjust the bias in the survey estimates. The summary table of descriptive statistics for the weights can be found in the Annex I. Figure C.1 presents, for each target variable, the EGM estimate, the unweighted AIMC-Q panel survey estimate, and nine weighted estimates. The nine weights correspond to the three sets of auxiliary variables—SD, AD, and SD+AD—and the three estimation methods—quasi-randomisation method (QR), the superpopulation models (SP), and the doubly robust estimator (DR). In addition, the first graph presents an average of the relative level of bias in the estimates for each combination of estimation method and set of auxiliary variables.

Administrative auxiliary variables (AD) showed a minimal ability to reduce the bias in the estimates. On average, the improvement in the relative bias of the estimates (graph 1 in Figure C.1) barely reaches one percentage point when unweighted estimates are taken as a reference. At best, when administrative data are used with the superpopulation (SP) method, the reduction in relative bias is one percentage point. Most estimates (graphs 2 to 14) adjusted using administrative data were similar to the unweighted ones. Only five variables experienced changes in the bias after using the administrative variables. However, in three of them, the effect of the weight resulted in a slight increase in the bias. These are the cases of *sports press yesterday*, *AM radio yesterday* and *radio at work yesterday*. For example, all three estimates of the proportion of *radio* consumption *at work yesterday* show increases in the level of bias, reaching 0.9 percentage points (SP and DR models). Also, in cases where weighting with administrative data has a bias-reducing effect, the magnitude of the bias is minimal. The most notable case is in the variable measuring *radio* consumption *in the home yesterday*, where the bias is reduced by 2.8 percentage points (SP) from the unweighted estimate. However, there is still a difference of 7.1 percentage points with respect to the EGM estimate.

Sociodemographic variables (SD) were slightly more effective than administrative variables when the quasi-randomisation method is used. However, the opposite occurred under the doubly robust method (DR) and superpopulation models (SP). Moreover, in both cases, the average relative bias of the estimates increases by one percentage point with respect to the unweighted estimates. The use of sociodemographic variables in the weighting compared to administrative variables produces larger variations in the estimates. In five out of 13 variables, the use of sociomodegraphics resulted in an increase in bias. For example, reading a *paper newspaper yesterday* or reading *supplements in the last 7 days*. Positive effects were mainly observed in the variables of the radio survey. For example, the variable listening to the *radio in the car yesterday* weighted by the superpopulation model (SP) reduced the bias of the estimate by 4.6 percentage points. Also, the doubly robust estimator (DR) reduced the bias of listening to the *radio at work yesterday* by 2.3 points.


Figure C.1. Bias of estimates by adjustment method and set of auxiliary variables. Graph 1 presents the average relative bias and graphs 2-14 represent the unweighted and weighted estimates for each variable together with the value of the EGM estimate.

The combination of sociodemographic and administrative variables (SD+AD) was, on average, the most effective set in reducing the bias of the estimates. Moreover, this effect was enhanced under the quasi-randomisation method, although it only represents an improvement of 2.2 percentage points compared to the unweighted estimate. In fact, the impact of using the SD+AD auxiliary variables in the estimation process is, in most cases, very similar to the effect produced by the sociodemographic variables; only the estimates of *radio on TV yesterday* and *radio at home yesterday* presented patterns more similar to the administrative data adjustments.

The estimation models used in this research—QR, SP, and DR—added some variability to the estimates, especially when the set of sociodemographic variables is taken into account. However, this is blurred when looking at the estimates separately. For most variables, using the different estimation methods yields similar results when the same set of auxiliary information is used.

Figure C.2 presents the standard errors of the estimates calculated using a linearised method and sample replications. It is noteworthy that, in general, the use of AD auxiliary variables had a smaller impact on the variance of the estimates, which was in line with the results observed in the bias analysis. The set of administrative variables was not related to most of the target variables and the probability of participating in the study, which prevents the bias from being reduced and keeps the standard error of the estimates at lower levels. In contrast, sociodemographic and SD+AD adjustments tended to increase the variance of most estimates. Moreover, the errors calculated with the *jackknife* method were larger than the linearised ones. This is because the linearised errors do not consider the variability derived from using estimates to compute the weights.

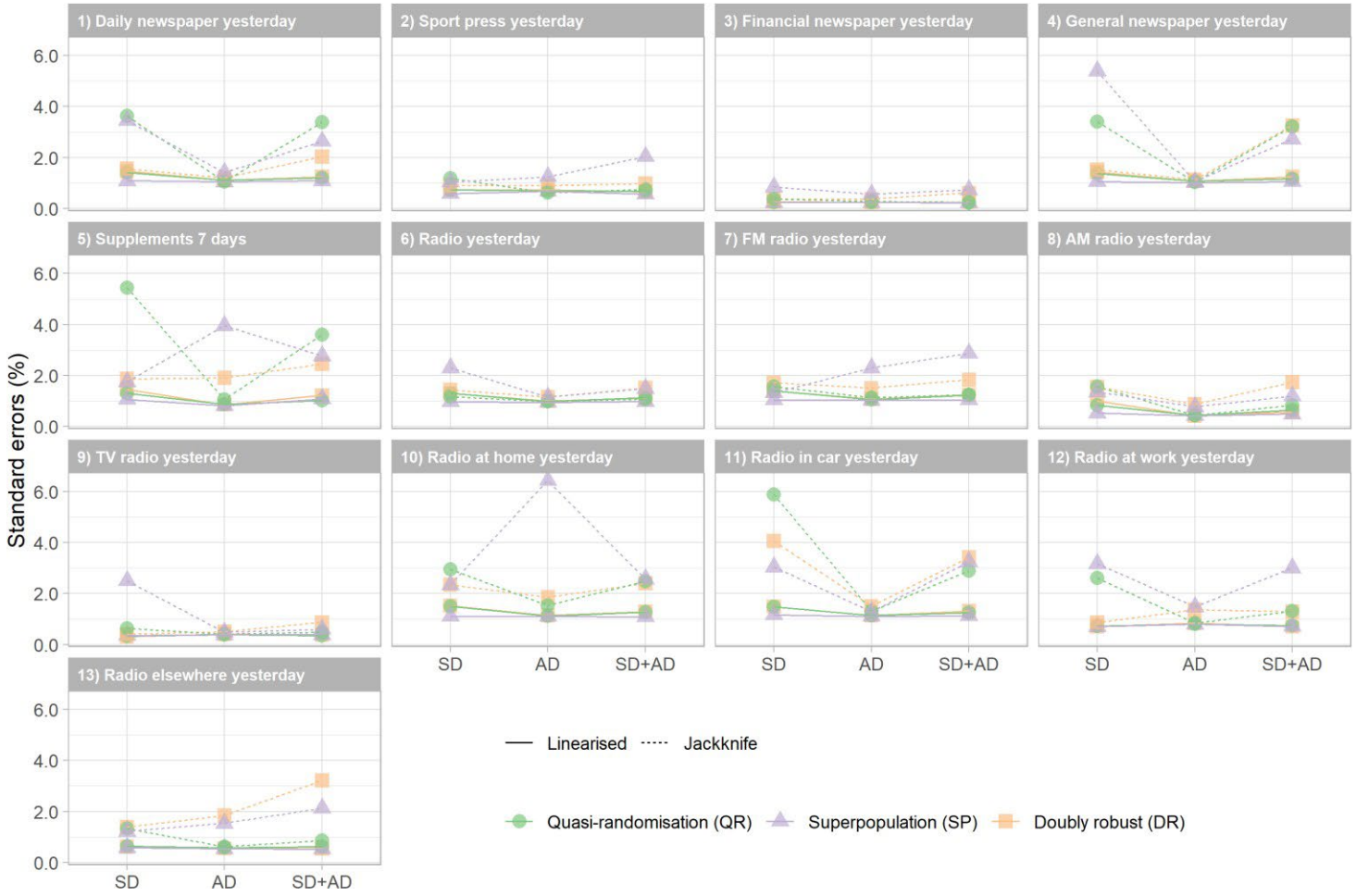



Figure C.2. Standard errors (%) of the estimates of each variable of interest by set of auxiliary variables and variance estimation method

## Discussion and conclusions

The first two hypotheses raised in this paper dealt with the most effective set of variables to reduce the bias of the estimates. This research tests the effect of using three sets of auxiliary variables—sociodemographic, administrative aggregated at the municipality level and the combination of both—to reduce the bias in the estimates from two surveys from an internet panel. Due to their large number and variety, aggregated administrative data could be used as auxiliary variables to adjust the estimates. However, the results of this research do not support this possibility; weights based on administrative data only slightly reduce the level of bias in the estimates. This finding is in line with Biemer and Peytchev (2013) and Lahtinen, Kaisa and Butt (2015), who did not find aggregate administrative variables useful for correcting nonresponse bias. In contrast to previous work, which used probability samples such as the European Social Survey, this research simulates a nonprobability sample. However, even in this scenario, aggregate administrative data did not help reduce the bias of the estimates.

The most effective approach to adjusting the sample was, on average, the combination of sociodemographic and aggregate administrative data and the quasi-randomisation method. However, this result should be taken with caution for two reasons. First, the difference between the average bias of the weighted and unweighted estimates was only 1.6 percentage points. And second, the effectiveness of this weight is mainly due to the sociodemographic variables. The analysis of the estimates using the weights constructed from the administrative and sociodemographic variables separately reinforces this idea.

Aggregate administrative variables do not reduce the bias of the estimates because they are not correlated with the probability of being part of the sample or with the variables of interest. However, it remains to be discerned whether the problem lies in the variables used—from voting behaviour to the proportion of luxury cars—or in the aggregate nature of the auxiliary variables. In this regard, Biemer and Peytchev (2013) argue that aggregate data must be correlated with the individual characteristics of the sample elements in order to be effective. This approach, however, needs to be tested. More research is required in order to find out whether there is a context in which aggregate data, given their nature, can be used to adjust survey estimates.

On the other hand, Peytchev, Presser and Zhang (2018) attempted to rethink the selection of auxiliary variables over time in *big data*. According to the authors, it is necessary to have data that theoretically match the variables to be estimated and the probability of responding, rather than using a large number of variables that may not be related to the object of study. The results of this research reinforce this approach: theory is necessary when selecting auxiliary variables.

The third hypothesis concerned the interaction between the data used and the estimation technique. In this research, the auxiliary variables have been used in three estimation models: quasi-randomisation, superpopulation models and doubly robust estimator. The quasi-randomisation and doubly robust estimators managed to optimise the auxiliary information to a certain extent. However, the superpopulation models in conjunction with the administrative variables resulted in a slight increase in the average bias. In general, the variability in the estimates is due more to the auxiliary variables used than to the estimation method.

Finally, it was expected that the use of administrative variables would have a positive impact on reducing the variance of the estimates. The estimated errors indicate that the weights generated from the administrative variables produce smaller standard errors. However, this effect has to do with the minimal variability of the weights themselves and not with the ability of the weights to adjust the estimates.

The first limitation of this research has to do with the comparability of individual calibrations using aggregate data as contextual variables. First, the ideal design for this research would have been to use the same set of variables at the individual level and as contextual variables. However, limiting the variables to sex and age meant leaving out the potential advantage of using aggregate data, which is more accessible. Another limitation concerns the generalizability of the conclusions drawn from the analysis of the two AIMC-Q panel surveys to other situations where administrative data can be used. Although using two surveys from an internet panel do not allow the conclusions to be extrapolated, they do provide new evidence that contributes to building knowledge on the use of aggregate data in the treatment of nonresponse and coverage bias.

## Bibliography


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

Barboza, Iffigenia and Rohan Williams. 2005. "Post-stratification and response bias in survey data with applications in political science. " Artículo presentado en Annual Meeting of the Midwest Political Science Association.

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

Biemer, Paul and 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 and 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 and 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 and 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 and Jan A. van den Brakel. 2018. "Comparing inference methods for nonprobability samples. " *International Statistical Review* 86(2):322-43. https://doi.org/10.1111/insr.12253

Buskirk, T. D., A. Kirchner, A. Eck and 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, and V. Vehovar. 2015. *Web survey methodology*. London: SAGE Publications.

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

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle and 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 and Richard Valliant. 2008. "Internet surveys: can statistical adjustments eliminate coverage bias? " *Survey Research Methods* 2(2):47-60.

Dibben, Chris, Mark Elliot, Heather Gowans and 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. and 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. and Richard Valliant. 2017. "Inference for nonprobability samples. " *Statistical Science* 32(2):249-64. https://doi.org/10.1214/16-STS598

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

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

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

Gummer, Tobias and Joss Roßmann. 2018. "The effects of propensity score weighting on attrition biases in attitudinal, behavioral, and sociodemographic 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 and M. Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalisations*. CRC press.

Hays, Ron D., Honghu Liu and 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. and 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*. New 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 and Sarah Butt. 2015. "Using auxiliary data to model nonresponse bias The challenge of knowing too much about nonrespondents rather than too little? " Londres.

Lee, Sunghee and 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 and 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. and Stanley Lemeshow. 2013. *Sampling of Populations: Methods and Applications*. Nueva York: John Wiley & Sons.

Lohr, Sharon L. and 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 and Courtney Kennedy. 2018. *For Weighting Online Opt-In Samples, What Matters Most?* Washington: Pew Research. Consulta 11 de Marzo 2019 (http://www. pewresearch. org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most)

Morris, Sarah, Alun Humphrey, Pablo Cabrera Álvarez and Olivia D'Lima. 2016. *The UK Time Diary Study 2014-2015. Technical Report.* Londres: NatCen Social Research. Consulta 11 de Marzo 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 and M. Phillips. 2013. *British Social Attitudes 30th Report*. Londres: NatCen Social Research. Consulta 11 de Marzo 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

Pedraza, Pablo, Kea Tijdens, Rafael Muñoz de Bustillo and 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 and 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 and 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 and 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 and Sixten Lundström. 2005. *Estimation in surveys with nonresponse*. New York: John Wiley & Sons. https://doi.org/10.1002/0470011351

Schonlau, M., A. Van Soest, A. Kapteyn and 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 multilevel 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. and Jibum Kim. 2013. "An assessment of the multilevel 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. and Graeme Laurie. 2014. "The administrative data research centre scotland: a scoping report on the legal & ethical issues arising from access & linkage of administrative data. " *SSRN Electronic Journal* (August).

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

Valliant, Richard and 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 and 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 and 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. and Roderick J. A. Little. 2013. "Nonresponse 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 and 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 and G. Wagner. Berlin: SCIVERO.

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

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

# Annex I: Tables

Table C3. Descriptive statistics of the weights

| | | Media | Standard deviation | Min. | Max. | DEFF |
|---|---|---|---|---|---|---|
| **Press study** | | | | | | |
| **Quasi-randomisation** | | | | | | |
| | 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 |
| **Superpopulation** | | | | | | |
| **Daily newspaper yesterday** | 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 |
| **Sport press yesterday** | 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 |
| **Economic newspaper yesterday** | 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 |
| **General newspaper yesterday** | 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 |
| **Supplements 7 days** | 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 |
| **Doubly robust** | | | | | | |
| **Daily newspaper yesterday** | 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 |
| **Sport press yesterday** | 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 |
| **Economic newspaper yesterday** | 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 |
| **General newspaper yesterday** | 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 |
| **Supplements 7 days** | 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 |
| **Radio study** | | | | | | |
| **Quasi-randomisation** | | | | | | |
| | 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 |
| **Superpopulation** | | | | | | |
| | SD | 20103.6 | 1636.6 | 16782.9 | 26952 | 1.01 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Radio yesterday** | AD | 20103.6 | 535.3 | 18060 | 22798.3 | 1.00 |
| | SD+AD | 20103.6 | 1623.3 | 15625 | 27274.8 | 1.01 |
| **FM radio yesterday** | 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 |
| **AM radio yesterday** | 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 |
| **TDT radio yesterday** | 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 at home yesterday** | 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 in car yesterday** | 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 at work yesterday** | 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 elsewhere yesterday** | 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 |
| **Doubly robust** | | | | | | |
| **Radio yesterday** | 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 |
| **FM radio yesterday** | 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 | | 129349 | 1.36 |
| **AM radio yesterday** | 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 |
| **TDT radio yesterday** | 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 at home yesterday** | 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 in car yesterday** | 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 at work yesterday** | 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 elsewhere yesterday** | 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 |