BMC Medical Research Methodology

RESEARCH

Open Access

Calibration and XGBoost reweighting to reduce coverage and non-response biases in overlapping panel surveys: application to the Healthcare and Social Survey

Luis Castro[1], María del Mar Rueda[1,2], Carmen Sánchez-Cantalejo[3,4], Ramón Ferri[1] and Andrés Cabrera-León[3,4*]

Abstract

Background Surveys have been used worldwide to provide information on the COVID-19 pandemic impact so as to prepare and deliver an effective Public Health response. Overlapping panel surveys allow longitudinal estimates and more accurate cross-sectional estimates to be obtained thanks to the larger sample size. However, the problem of non-response is particularly aggravated in the case of panel surveys due to population fatigue with repeated surveys.

Objective To develop a new reweighting method for overlapping panel surveys affected by non-response.

Methods We chose the Healthcare and Social Survey which has an overlapping panel survey design with measurements throughout 2020 and 2021, and random samplings stratified by province and degree of urbanization. Each measurement comprises two samples: a longitudinal sample taken from previous measurements and a new sample taken at each measurement.

Results Our reweighting methodological approach is the result of a two-step process: the original sampling design weights are corrected by modelling non-response with respect to the longitudinal sample obtained in a previous measurement using machine learning techniques, followed by calibration using the auxiliary information available at the population level. It is applied to the estimation of totals, proportions, ratios, and differences between measurements, and to gender gaps in the variable of self-perceived general health.

Conclusion The proposed method produces suitable estimators for both cross-sectional and longitudinal samples. For addressing future health crises such as COVID-19, it is therefore necessary to reduce potential coverage and non-response biases in surveys by means of utilizing reweighting techniques as proposed in this study.

Keywords Public health, COVID-19, Panel surveys, Sampling, Machine learning, Non-response bias

*Correspondence:
Andrés Cabrera-León
andres.cabrera.easp@juntadeandalucia.es
Full list of author information is available at the end of the article

Background

Healthcare statistical services worldwide have used probability surveys to provide information on the social, economic and health impact of the disease, or on its seroprevalence [1] and evolution, or on the characteristics of the infected population, in particular members most vulnerable to the virus due to their age, risk of exclusion, health conditions or dependency [2]. These surveys allow valid inferences to be made about the population without having to incorporate hypotheses into the models, which is of great practical benefit [3]. Regarding the COVID-19 pandemic, most of the surveys created were based on non-probability sampling to provide a quick and efficient assessment of the situation based on predicting and quantifying the main parameters involved in this phenomenon [4].

The Healthcare and Social Survey (ESSA, Encuesta Sanitaria y Social de Andalucía) research project arises from the need to provide data on the evolution of the COVID-19 impact which can be considered when making decisions to prepare and deliver an effective Public Health response in the different populations concerned, particularly in the most vulnerable ones, including, for example, the elderly, the chronically ill, or persons at risk of exclusion [5]. The objective of this survey is to determine the magnitude, characteristics, and evolution of the impact of COVID-19 on overall health and its socioeconomic, psychosocial, behavioral, occupational, environmental, and clinical determinants in the general population and in the population at higher risk for socioeconomic deprivation. The study is based on a Real-World Data design integrating observational data extracted from multiple sources including information obtained from different surveys and clinical, population, and environmental registries. The ESSA has an overlapping panel design [6]. It consists of a series of measurements broken down into a new sample and a longitudinal sample for each measurement, except for the first measurement where the entire sample is new. Compared to rotating panel surveys [7], the ESSA sampling design is therefore non-rotational, i.e. the units included in each measurement remain in the following measurements until the final one.

This type of overlapping panel design is often used when the main objectives are to obtain cross-sectional estimates at time t and short-term longitudinal estimates of net and gross change between t and $t+1$, as is the case for ESSA. This way, the use of new samples at each measurement t permits whole population representativeness at time $t+1$, and therefore also permits cross-sectional estimation at this time. This feature means that one of the key aspects of overlapping panel surveys lies in cross-sectional estimation, i.e. how to combine the different samples selected at the same time. Another key aspect of panel surveys is the response obtained in each measurement of the longitudinal samples. The lack of response thus grows with the number of occasions or measurements, due, amongst other reasons, to the panelist fatigue with repeated interview. For this reason, partial replacement of units is common to guarantee a minimum number of units in the final sample. Estimation from data obtained with this structure is not easy, especially if the desire is to take into account the biases produced both by lack of response and lack of sample coverage and representativeness.

Some methods of handling wave non-response in panels are provided in [8–10]. Thus, the two main methods used to handle it in panels are based on weighting the effective sample according to the theoretical sample in the strata used, and reweighting by calibration in terms of population totals for sociodemographic stratification variables such as sex, age or territory (e.g. region, province or habitat level) [11]. Another set of studies focuses on modelling different types of response patterns in panels. Kern et al. [12] compares the usage of different Machine Learning (ML) methods for modeling non-response in the German Socio-Economic Panel Study (GSOEP) and recent study [13] proposes a general framework for building and evaluating non-response prediction models with panel data, although this study focuses on model building and evaluation without utilizing the predictions obtained to correct bias in the estimations.

Non-response in panel studies has traditionally been tackled by using non-response weights. Although reweighting methods do exist for addressing these types of biases, they have been proposed fundamentally for cross-sectional surveys and there are few studies that provide a formal methodology for their treatment in this type of panel survey. In [14], the authors discuss adjustments for non-response and how calibration can be carried out in panel studies in general and what effects it creates. They consider three possible calibration approaches: initial calibration (at the beginning of the panel, the weights of the units in the panel are calibrated), final calibration (at measurement t the weights of the individuals in the sample are adjusted by calibration) and initial and subsequent final calibration (both initial and final calibration are carried out). Several approaches are tested in [15] to produce calibration estimators which are suitable for survey data affected by non response where auxiliary information exists at both the panel and population level.

Longitudinal and cross-sectional weighting are considered in [7] for rotating samples in the context of the SILC survey in France. The sampling each year in this survey is formed by combining nine panel subsamples, and the longitudinal weights are allocated as an average of the weights in each time during which a unit belongs to the sample, using the weight-share method [16]. This method is also used in [17] for obtaining cross-sectional indicators for the SILC survey in Switzerland based on a four-panel rotation scheme. Verma et al. [18] develops longitudinal and cross-sectional weighting procedures in a rotational household panel with reference to the EU-SILC design (4-year rotational design) using a step-by-step procedure starting with design weights, followed by adjustments for non-response and calibration to external controls, and finally trimming and scaling as required to obtain the initial weights.

However, these authors do not consider the application of these adjustment methods in designs such as the study described herein, i.e. overlapping panel surveys where the units included in each measurement remain in the following measurements until the final one, and in which each measurement is completed with a new sample (except for the first one since the whole sample is new). This is the research gap addressed by this paper.

In this work, we therefore propose an empirical study of the associations between choice of research methodology and study outcomes. Accordingly, we combine suitable reweighting methods such as Propensity Score Adjustment (PSA), XGBoost and calibration to address the biases associated with dropout from overlapping panel survey data for estimating totals, proportions, ratios and differences in a study outcome. Other ML methods than XGBoost technique (such as logistic regression, decision trees, random forests and so on) could be used, but several papers [13, 19, 20] show that the set of predictor variables used in general mattered more than the type of ML technique. With regard to neural networks, they have been hugely successful for image, text or audio data due to the use of structures far more advanced than deep feedforward networks. However, for tabular data as in our case, the inefficacy and unreliability of neural networks is widely known. Arik and Pfister [21] further explains this issue in its introduction. Thus, those statistical techniques (PSA, XGBoost and calibration) are formulated on the outcome self-perceived general health from the ESSA survey and can be applied to any other variable and epidemiological research based on overlapping panel design.

Methods

The ESSA study framework

The Healthcare and Social Survey (ESSA, Encuesta Sanitaria y Social de Andalucía) provides a follow-up over time of the impact of the pandemic and its resulting lockdown on the population of Andalusia over the age of 16. Andalusia is a southern region of Spain with 8.4 million inhabitants. It is also the fifth most populated region in Europe, with a population similar in size to that of other European countries such as Austria or Switzerland.

As shown in Fig. 1, the ESSA study includes four measurements. The first one, M_1, coincided with the beginning of the Spanish State of Alarm in April 2020 (coinciding with the lockdown), while the second measurement M_2 was taken in June and July (a month after the first interview, coinciding with the de-escalation); the third measurement M_3 was taken in November and December (6 months after the first interview and coinciding with the second wave of the pandemic), and the fourth measurement M_4 in April and May 2021 (12 months after the first interview, coinciding with the relaxation of mobility restrictions and the end of the state of alarm). All the measurements had an effective size of around 3000 people, except for the second one which was 2500. They were obtained using an overlapping panel design, so the individuals from the previous measurement are sampled again. Each measurement thus had its own panel of people who were interviewed again in the following measurements $(P_1, ..., P_4)$. Non-response was offset in each measurement with another sample which included new individuals. The details of this non-response and the effective sample size for each measurement and panel can be consulted in Fig. 1. It also provides a description of the evolution of the SARS-COV-2 pandemic in Andalusia during 2020 and 2021 in terms of active infection diagnostic tests and deaths.

With respect to the sampling method, the new sample in each measurement was selected by stratified simple random sampling according to province and degree of urbanization: urban, semi-urban and rural, based on the methodology described by EUROSTAT for the allocation of territorial typologies in statistical grids of 1 km^2 where population resides; more information in [22]. This implies that within each stratum any person has the same probability of being selected, i.e. self-weighted samples are obtained in each stratum. The new sample was thus distributed across the 8 Andalusian provinces in proportion to province population size. Within each province, sample allocation was proportional to the population size of each degree of urbanization. Regarding for the longitudinal sample of a given measurement, this comprised the samples in the previous measurements, i.e. of the panel created in each measurement $(P_1, ..., P_4)$, with the exception of the first measurement, which did not have a longitudinal sample given that it was the first one. The population framework used for the extraction of population samples aged 16 years old and over residing in family dwellings in Andalucía, came from the Longitudinal Population Database of Andalusia (BDLPA) as of 1 January 2019. The BDLPA originates from integrating data obtained from the Civil Registries with respect to births,

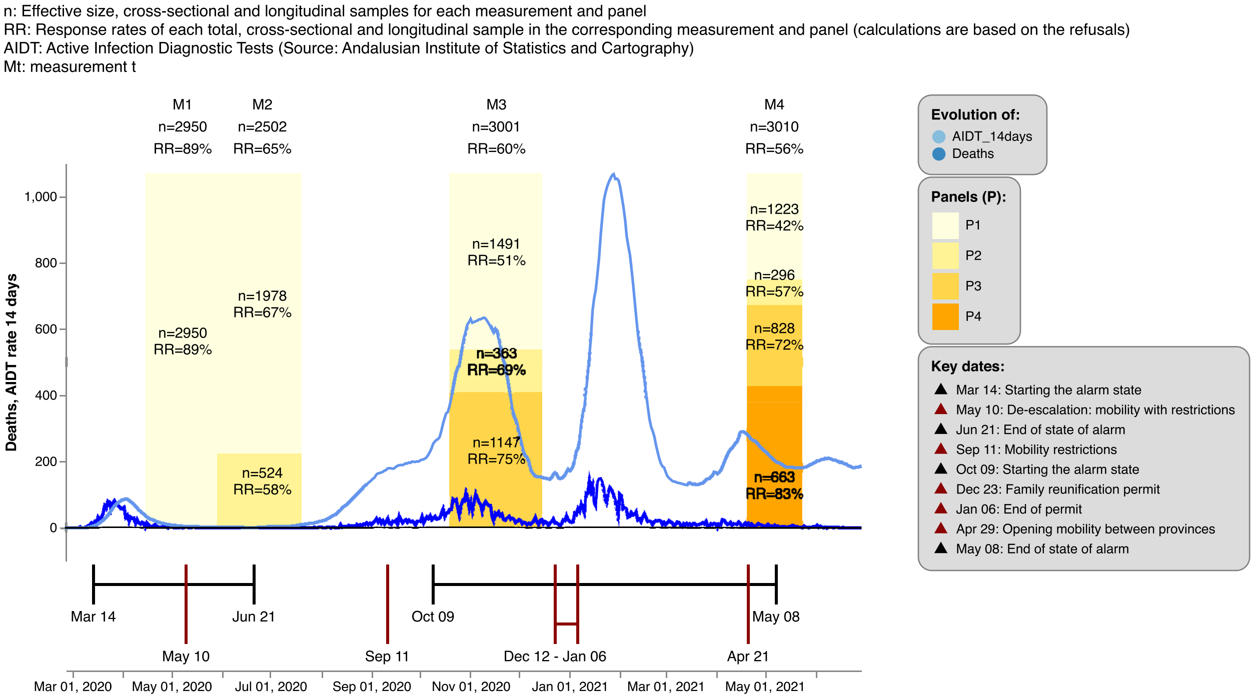


Fig. 1 Temporal scope, response rates (RR) and effective sample size for each measurement in ESSA

deaths, and marriages (i.e., vital statistics), as well as that reported in the population and housing censuses, give rise to an integrated longitudinal frame for population and territorial statistics in Andalusia [23]. The Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym) was responsible for population framework and extraction of the samples. A detailed description of the protocol followed for this survey can be seen in [5].

$$s_{rh}^{(1)} = \{k \in s^{(1)}/\text{respond in stratum h}\}$$
$$s_{fh}^{(1)} = \{k \in s^{(1)}/\text{missing in stratum h}\}.$$

Sampling setup in overlapping panels

Let U denote a finite population of size N, $U = \{1, \ldots, k, \ldots, N\}$. We want to estimate a population parameter of a variable of interest, y.

On the first measurement ($M1$) a theoretical sample $s^{(1)}$ of size $n^{(1)}$ is selected from the population U by stratified simple random sampling. Let h be the stratum to which unit i belongs, ($h = 1, \ldots L$) and $s_h^{(1)}$ be the sample corresponding to stratum h on measurement 1 as well as the first panel of the survey ($P1$).

There is a total lack of response in the sample $s^{(1)}$ which is divided into

Let $n_{rh}^{(1)}$ denote the number of the observations obtained from the $n_h^{(1)}$ sampled units, that is $\sum_h n_{rh}^{(1)} = n_r^{(1)}$ is the effective size of $s_r^{(1)}$ as well as of P_1. Thus, $s_r^{(1)}$ will be the theoretical sample of the second measurement M_2, $s_r^{(1,2)}$ the effective sample of M_2 (respondents in M_1 and M_2 of P_1), and $n_r^{(1,2)}$ the effective size of P_1 in M_2.

In each of the following measurements until t, $M_2, \ldots, M_t$, we denote by $s_r^{(1,t)}$ the effective sample in M_t of P_1, i.e. respondents in M_1, $M_2, \ldots, M_t$ of the first sample obtained $s^{(1)}$ as well as of the first panel created P_1; and by $n_r^{(1,t)}$ its effective size. In a similar way, we denote by $s_r^{(i,j)}$, where $i = 1, \ldots, j-1$, $j = 2, \ldots, t$, and $i < j$, the effective sample in M_j of P_i, i.e. respondents in M_i, $M_{i+1}, \ldots, M_j$ of the theoretical sample obtained in M_i, $s^{(i)}$, as well as of the i panel created, P_i; and by $n_r^{(i,j)}$ its effective size. By contrast, we denote by $s_f^{(i,j)}$ the missing sample in M_j of P_i, i.e. non-respondents in any M_i, $M_{i+1}, \ldots, M_j$ of the i theoretical sample obtained in M_i, $s^{(i)}$, which created the i panel, P_i. Thus, due to this non-response sample and in order to achieve the required sample size, we complete the sample of M_j with a new theoretical sample $s^{(j)}$ from the same population U by the same sampling design but independently of the new samples extracted in previous measurements. Therefore, for the extraction of the new theoretical samples, $s^{(1)}, \ldots, s^{(t)}$, in each new measurement M_j, IECA verified that M_j and M_{j-1} had an empty intersection. Therefore, the theoretical sample of M_j comprises by the effective samples of M_{j-1} and the new theoretical sample of M_j, i.e. $s^{M_j} = s_r^{M_{j-1}} \cup s^{(j)}$; while the effective sample of M_j comprises the effective samples of the panels created until M_{j-1}, $P_1, ..., P_{j-1}$, and the new effective sample of M_j, i.e. $s_r^{M_j} = \cup_{i=1}^{j-1} s_r^{(i,j)} \cup s_r^{(j)}$.

With respect to sample size, let $n^{(j)}$ be the theoretical size of the new theoretical sample $s^{(j)}$ in measurement j, and denote by $n_r^{(j)}$ the effective size of $s_r^{(j)}$, i.e. the respondents of the new sample in that measurement. Thus, the theoretical size of M_j comprises the effective sizes of M_{j-1} and the new theoretical size of M_j, i.e. $n^{M_j} = n_r^{M_{j-1}} + n^{(j)}$; while the effective size of M_j is composed by the effective sizes of the panels created until M_{j-1}, $P_1, ..., P_{j-1}$, and by the new effective size of M_j, i.e. $n_r^{M_j} = \sum_{i=1}^{j-1} n_r^{(i,j)} + n_r^{(j)}$.

Let $y_i^{(1)}$ be the value of the target variable associated to the i-th unit in M_1, and let d_i be the design weight associated to the i-th unit equal to the inverse of the inclusion probability in the theoretical sample, an estimation of the total of Y in the first measurement is given by [24]:

$$\hat{Y}_{s^{(1)}} = \sum_{i \in s_r^{(1)}} d_i y_i^{(1)} \,. \tag{1}$$

This estimator is a naive estimator. In the case of simple random sampling design for unit i is $d_i = \frac{N}{n^{(1)}}$.

Design weights should be adjusted to consider non-response in order to reduce the possible bias of resulting estimates, which may arise when there is a different propensity in answering for different groups. In the first measurement a response rate is determined in each class and a new weight is defined as the product of the design weight and the inverse of the response rate. The response rate is evaluated as $r^{(1)} = \frac{n_r^{(1)}}{n^{(1)}}$. Then the initial weight of unit i is replaced with the new weight $d_i^{(1)} = \frac{d_i}{r^{(1)}}$ and the estimator is given by

$$\hat{Y}_{s_r^{(1)}} = \sum_{i \in s_r^{(1)}} d_i^{(1)} y_i^{(1)} \,. \tag{2}$$

For the following measurements until t, different estimators can be obtained from the different effective samples of measurements and panels. Thus, to fix the notation, we will term cross-sectional estimator of a parameter θ at time j as being those estimators that are obtained from the effective sample of M_j, i.e. $s_r^{M_j}$, where $j = 1, ..., t$; while we will call longitudinal estimators of a parameter θ at times j and $j-1$, to those obtained from the effective samples of those panels which belong to two consecutive measurements M_j and M_{j-1}, i.e. $s_r^{(M_j,M_{j-1})} = \cup_{i=1}^{j-1} s_r^{(i,j)}$, where $j = 2, ..., t$; being the same sample as $s_{cross}^{(j)}$ but without the new effective sample of M_j. The process for obtaining them is shown in the following sections.

Cross-sectional estimation

The objective of most cross-sectional surveys is to produce unbiased estimates of totals or means at a given point in time, and, in the case of repeated surveys, to produce estimates of the net change that occurred in the population between two time points [25].

Cross-sectional estimates can be derived from longitudinal survey data to improve the cost-effectiveness of surveys, assuming that the survey design takes this possibility into account, and that estimation procedures are developed to satisfy cross-sectional as well as longitudinal requirements [26]. For this we will use both the longitudinal samples from the panels and the fresh or new samples obtained in each measurement. This way, the sample we work with always has the maximum sample size possible and we reduce the final estimator variance.

Point estimation of parameters of the cross-sectional population based on data from longitudinal surveys has been studied by [27] among others and the problem of formal comparison of the estimates from two years, which requires variance estimation for the difference of the estimates, is considered in [28]. We will follow community-agreed standards appropriate for the survey methodology used in those works, but implementing a new approach to achieve more suitable estimators in overlapping panel surveys. We will thus devise a cross-sectional weighting scheme that includes a non-response adjustment, optimal combination of the samples from the panels involved, and calibration for completing the representativeness of the population at a given measurement. This proposal is described below.

Weight adjustment based on propensities

A simple adjusted estimator accounting for initial non-response and attrition can be obtained by adjusting the basic weights of the Horvitz-Thompson estimator by the fraction of non-response. This adjustment based on weighting within classes assumes that unit non-response may be modelled by response homogeneity groups, and that these response homogeneity groups are given by the strata. This may be a reasonable assumption at baseline but it seems unlikely that non-response at any point in time will be suitably explained by the strata defined at baseline.

Therefore, although weighting within classes is a commonly used procedure for non-response cross-sectional and longitudinal weighting in panels, a more pragmatic alternative is to use a regression-based approach, all the more so when numerous auxiliary variables are available [18]. For this we are going to use the popular Propensity Score Adjustment (PSA) method [20, 29, 30] to model the probability that a unit k of the new theoretical sample $s^{(j)}$ responds to M_j, where $j = 1, ..., t$, or that another unit k of the effective sample $s_r^{(i)}$ responds to M_j, where $i = 1, ..., j-1$, $j = 2, ..., t$, and $i < j$.

For each sample unit k in $s^{(j)}$ let be $\delta_k^{(j)} = 1$ if $k \in s_r^{(j)}$ and $\delta_k^{(j)} = 0$ if $k \in s^{(j)} - s_r^{(j)}$, and regarding each sample unit k in $s_r^{(i)}$ let be $\delta_k^{(i,j)} = 1$ if $k \in s_r^{(i,j)}$ and $\delta_k^{(i,j)} = 0$ if $k \in s_r^{(i)} - s_r^{(i,j)}$. We assume that the selection mechanism of response is ignorable, this is:

$$\pi_k^{(j)} = P(\delta_k^{(j)} = 1|y_k, \mathbf{x}_k) = P(\delta_k^{(j)} = 1|\mathbf{x}_k);\, k \in s_r^{(j)} \tag{3}$$

where $j = 1, ..., t$, and for $\delta_k^{(i,j)}$:

$$\pi_k^{(i,j)} = P(\delta_k^{(i,j)} = 1|y_k, \mathbf{x}_k) = P(\delta_k^{(i,j)} = 1|\mathbf{x}_k);\, k \in s_r^{(i,j)}\,. \tag{4}$$

where $i = 1, ..., j-1$, $j = 2, ..., t$, and $i < j$.

We also assume that the mechanism follows a parametric model:

$$P(\delta_k^{(j)} = 1|y_k, \mathbf{x}_k) = m_{(j)}(\mathbf{x}_k, \lambda_{(j)}) \tag{5}$$

and

$$P(\delta_k^{(i,j)} = 1|y_k, \mathbf{x}_k) = m_{(i,j)}(\mathbf{x}_k, \lambda_{(i,j)})\,. \tag{6}$$

for some known functions $m_{(j)}(\cdot)$ and $m_{(i,j)}(\cdot)$ with second continuous derivatives with respect to unknown parameters $\lambda_{(j)}$ and $\lambda_{(i,j)}$, respectively. A commonly adopted parametric model is the logistic regression model [31, 32].

We use a state-of-the-art machine learning method: XGBoost [33] for estimating $\pi_k^{(j)}$ and $\pi_k^{(i,j)}$. This technique builds decision trees ensembles which optimize an objective function via Gradient Tree Boosting [34]. More details can be found in Annex 2. Kern et al. [12] has shown the effectiveness of this technique when studying non-response in the GSOEP panel. Ferri-García and Rueda [20] showed that Gradient Tree Boosting can lead to selection bias reductions in situations of high dimensionality or where the selection mechanism is Missing at Random (MAR). Lee et al. [35], Lee et al. [36], McCaffrey et al. [37], McCaffrey et al. [38], Tu [39], Zhu et al. [40], and Rueda et al. [41] have applied boosting algorithms in propensity score weighting showing better results than conventional parametric models.

In order to obtain the estimated propensities $\hat{\pi}_k^{(j)}$, we train a model with $s^{(j)}$ where $\mathbf{x}_k$ includes every available variable observed in the BDLPA population framework, while to obtain the estimated propensities $\hat{\pi}_k^{(i,j)}$, we train a model with $s_r^{(i)}$ where $\mathbf{x}_k$ includes every available variable observed in $s_r^{(i,j)}$. This model minimizes the weighted logistic loss for $\delta_k^{(j)}$; $k \in s^{(j)}$ and for $\delta_k^{(i,j)}$; $k \in s_r^{(i)}$.

Since the values we are interested in, $\hat{\pi}_k^{(j)}$ and $\hat{\pi}_k^{(i,j)}$ for $k \in s^{(j)}$ and $k \in s_r^{(i,j)}$, respectively, are a subset of the values used for training, $\delta_k^{(j)}$ and $\delta_k^{(i,j)}$ for $k \in s^{(j)}$ and $k \in s_r^{(i)}$, respectively, overfitting is likely to happen. This means that we will obtain values extremely close to 1 instead of real propensities. Hyperparameter optimization is essential in order to avoid this problem. This optimization can be applied as described in the Results section. Another important technique to consider is class balancing [42]. Classification models learn best when every class is equally represented in the training dataset. In practice, response rates are rarely close to 0.5 and therefore our model would often be biased. Class balancing ensures valid estimates, even when the response rate is high or low, by assigning $(1-p)\delta_k + p(1-\delta_k)$ as instance weight for training, where p represents the observed response rate. However, this method also distorts output probabilities. Consequently, they should be corrected as described by [43]:

$$\hat{\pi}_{corrected} = \frac{\hat{\pi} p}{\hat{\pi} p + (1-\hat{\pi})(1-p)\,.}$$

Then we applied this correction to the inverse of the estimated response propensity $\hat{\pi}_k^{(j)}$, which is ultimatley used as weight for constructing the estimator based on the new effective sample in M_j, $s_r^{(j)}$:

$$\hat{Y}_{s_r^{(j)}}^{\mathrm{PSA}} = \sum_{k \in s_r^{(j)}} \frac{N}{n^{(j)}} \frac{n^{(j)}}{n_r^{(j)}} \frac{1}{\hat{\pi}_k^{(j)}} y_k^{(j)} = \sum_{k \in s_r^{(j)}} d_k^{(j)} \frac{1}{\hat{\pi}_k^{(j)}} y_k^{(j)} = \sum_{k \in s_r^{(j)}} d_k^{(j)PSA} y_k^{(j)}\,, \tag{7}$$

where $j = 1, ..., t$; and we use the inverse of $\hat{\pi}_k^{(i,j)}$ as weight for constructing the estimator based on each effective sample of its corresponding panel P_i created in the previous measurements until M_j, $s_r^{(i,j)}$

$$\hat{Y}_{s_r^{(i,j)}}^{\mathrm{PSA}} = \sum_{k \in s_r^{(i,j)}} \frac{N}{n_r^{(i,j-1)}} \frac{n_r^{(i,j-1)}}{n_r^{(i,j)}} \frac{1}{\hat{\pi}_k^{(i,j)}} y_k^{(i,j)} = \sum_{k \in s_r^{(i,j)}} d_k^{(i,j)} \frac{1}{\hat{\pi}_k^{(i,j)}} y_k^{(i,j)} = \sum_{k \in s_r^{(i,j)}} d_k^{(i,j)PSA} y_k^{(i,j)}\,. \tag{8}$$

where $i = 1, ..., j-1$, $j = 2, ..., t$, and $i < j$.

Combining these estimators, we can consider the following cross-sectional estimator for the total:

$$\hat{Y}_{M_j}^{\mathrm{PSA}} = \sum_{i=1}^{j-1} \alpha_i \hat{Y}_{s_r^{(i,j)}}^{PSA} + \alpha_j \hat{Y}_{s_r^{(j)}}^{PSA}\,, \tag{9}$$

where $j = 1, ..., t$ and α_i are nonnegative constants such that $\alpha_1 + \alpha_2 + ... + \alpha_j = 1$.

There are several ways to assign these constants. A simple solution is to weight each estimator by the weight that sample has in the total effective sample available at the time j. This permits the procedure not to depend on the variable to be estimated and also to calculate only a few α values, making the process of estimating the variables simpler and more systematic. This is the procedure we followed.

Calibration on population totals

In addition to modification of weights for handling non-response, it may also be carried out to take auxiliary information into account. Calibration [11] is the technique most used for weights adjustment and can aim to ensure consistency among estimates of different sample surveys, reduce biases in the sample due to non-response, non-coverage and other distortions, and also reduce variances [44–47].

Let $\mathbf{x}^{*(j)}$ be a set of auxiliary variables related to y such that their population totals at the stratum level are known at measurement j, $\mathbf{X}_h^{*(j)} = \sum_{\mathcal{U}_h} \mathbf{x}_{kh}^{*(j)}$.

We denote by

$$\hat{Y}_{M_j} = \sum_{k \in s_r^{M_j}} D_k^{(j)} y_k^{(j)}$$

any of the cross-sectional estimators obtained using the previous adjustment method, where $s_r^{M_j} = \cup_{i=1}^{j-1} s_r^{(i,j)} \cup s_r^{(j)}$, as we defined in the Sampling setup in overlapping panels section.

The calibration total estimator is obtained as:

$$\hat{Y}_{M_j}^{\mathrm{CAL}} = \sum_{k \in s_r^{M_j}} w_k^{(j)} y_k^{(j)}, \tag{10}$$

where the weights $w_k^{(j)}$, are as close as possible, with respect to a given distance G, to the weights $D_k^{(j)}$ obtained in the phase of reweighting and combination of samples:

$$\min_{\omega_k} \sum_{k \in s_r^{M_j}} G\left(w_k^{(j)}, D_k^{(j)}\right) \tag{11}$$

fulfilling the calibration condition

$$\sum_{k \in s_{rh}^{M_j}} w_{kh}^{(j)} \mathbf{x}_{jh}^{*(j)} = \sum_{\mathcal{U}_h} \mathbf{x}_{kh}^{*(j)} \tag{12}$$

for all stratum h given by the calibration variables considered.

Estimating changes compared to the first measurement

A parameter of interest is the absolute change of a variable between one measurement and the first measurement. We denote by $\theta_{M_j}^{\mathrm{ABS}} = Y_{M_j} - Y_{M_1}$ this parameter, where $j = 1, ..., t$. Variations over time are measured more accurately with overlapping samples with respect to the case where samples on different occasions do not overlap (see [48]). An estimator of this parameter for measurement j based on the previous calibration total estimators can be obtained as follows:

$$\hat{\theta}_{M_j}^{\mathrm{ABS}} = \hat{Y}_{M_j}^{\mathrm{CAL}} - \hat{Y}_{M_1}^{\mathrm{CAL}}. \tag{13}$$

Another parameter of interest in panel surveys is the relative change $\theta_{Mj}^{\mathrm{REL}} = \frac{Y_{M_j} - Y_{M_1}}{Y_{M_1}}$ between measurement 1 and measurement j, which is estimated as:

$$\hat{\theta}_{Mj}^{\mathrm{REL}} = \frac{\hat{\theta}_{M_j}^{\mathrm{ABS}}}{\hat{Y}_{M_1}^{\mathrm{CAL}}}. \tag{14}$$

The estimator is a quotient of two estimators of the total based on two different samples, meaning that its properties are not equivalent to those of the ratio estimator commonly used in survey sampling, but its theoretical properties can be derived by using Taylor linear approximation.

Estimating gender gaps in each measurement

The impact of COVID-19 on the social determinants of health may have differed significantly between women and men as shown in recent studies [49]. It is therefore of great interest to define the estimators of the gender gap observed in each measurement, and also in absolute and relative terms, in order to observe their evolution.

Let $Gen = \{M, W\}$ be the variable measured in $s^{(j)}, j = 1, ..., t$ which reflects whether a respondent is a man (M) or a woman (W). We define the two indicator variables: $I_{kh}^M = 1$ if the unit k in stratum h is a man and 0 elsewhere, and I_{kh}^W in a similar way.

We start by defining the absolute gender gap estimator as follows:

$$\widehat{GG}_{M_j}^{\mathrm{ABS}} = \hat{Y}_{M_j}^{\mathrm{CAL}W} - \hat{Y}_{M_j}^{\mathrm{CAL}M} =$$

$$= \sum_h \sum_{k \in s_{rh}^{M_j}} w_{kh}^{(j)} y_{kh}^{(j)} I_{kh}^W - \sum_h \sum_{k \in s_{rh}^{M_j}} w_{kh}^{(j)} y_{kh}^{(j)} I_{kh}^M. \tag{15}$$

This estimator is defined as the linear combination of two estimators in certain domains, hence its theoretical properties can be easily derived [48]. This estimator is the most simple one that can be built on the gender gap and can differentiate between men and women in measurement j. However, this estimator is subject to the base rate of each variable. For this reason, we define the relative gender gap estimator as follows:

$$\widehat{GG}_{M_j}^{\mathrm{REL}} = \frac{\widehat{GG}_{M_j}^{\mathrm{ABS}}}{\hat{Y}_{M_j}^{\mathrm{CAL}M}} = \frac{\hat{Y}_{M_j}^{\mathrm{CAL}W} - \hat{Y}_{M_j}^{\mathrm{CAL}M}}{\hat{Y}_{M_j}^{\mathrm{CAL}M}}. \tag{16}$$

This estimator allows us to observe the gender gap in measurement j taking into account the base rate of the given target variable.

Thus, to obtain the cross-sectional estimator for the study variables of each ESSA measurement, we start from the H-T estimator (1) adjusted for non-response (7), combined from the panel and new samples (9) and finally calibrate to increase the representativeness of the sample (10). This estimator serves as the basis for calculating the absolute (13) and relative (14) change estimators between measurement j and 1, and for obtaining the different estimators to measure the absolute and relative gender gap in a given measurement (15 and 16).

Longitudinal estimation

The primary objective of panel surveys is the production of longitudinal data series which are appropriate for studying the gross change in the population between collection dates, and for research on causal relationships among variables. To study these changes and understand their relationships, it is more convenient to use longitudinal samples than cross-sectional ones, since they reflect the variations of the variable in each individual and enable additional parameters to be estimated, such as the number of population individuals whose value of y increases, decreases or remains the same between a measurement and the previous one. The drawback of working with the longitudinal sample is that its size is smaller at each time and therefore the variance of the estimates can be large.

In this section, the previous estimated propensities for each unit k of sample $s_r^{(i,j)}$, $\hat{\pi}_k^{(i,j)}$, are used to reweight for non-response when estimating the absolute difference from M_j to M_{j-1} as:

$$\hat{Y}^{\text{PSA}}_{M_j-M_{j-1}} = \sum_{k \in s_r^{(M_j,M_{j-1})}} D_k^{(M_j,M_{j-1})} (y_k^{(j)} - y_k^{(j-1)}) \,. \tag{17}$$

where $s_r^{(M_j,M_{j-1})} = \cup_{i=1}^{j-1} s^{(i,j)_r}$, and $j = 2, ..., t$. In this situation, the estimator is calculated by modelling the non-response of each panel P_i created until M_{j-1}, that is, we estimate the propensities given by 6.

Thus, the estimated propensities for each unit k of the samples $s_r^{(i,j)}$, $\hat{\pi}_k^{(i,j)}$, are used in the first stage to reweight for adjusting non-response, obtaining the total estimator given by 17; and, in the second stage, calibration is applied to reweight these weights and obtain new ones, $v_k^{(j,j-1)}$, so as to obtain better population representativeness. The longitudinal estimator of the absolute difference can be defined as follows:

$$\hat{Y}^{\text{CAL}}_{M_j-M_{j-1}} = \sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} (y_k^{(j)} - y_k^{(j-1)}) \,. \tag{18}$$

The longitudinal nature of the estimator allows us to define new estimators on the number of population individuals whose value of y increases, decreases or remains the same between M_j and $Mj - 1$. Let A be a subset of interest ($\mathbb{R}^+$, $\mathbb{R}^-$ or 0 if we are interested in the units whose value of y increases, decreases or remains the same, respectively); the estimator of the number of population individuals for which $y^{(j)} - y^{(j-1)} \in A$ can be estimated as follows:

$$\hat{\theta}^A_{M_j-M_{j-1}} = \sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_A, I_A = \begin{cases} 1 \;\; y_k^{(j)} - y_k^{(j-1)} \in A \\ 0 \;\; y_k^{(j)} - y_k^{(j-1)} \notin A \end{cases} \cdot \tag{19}$$

We can also obtain the estimator of the rate of people whose value in y has decreased between $t-1$ and t, in reference to the people whose value in y has increased between $t-1$ and t. For example, if the variable y measures health status, this rate can be considered a deterioration/improvement rate, $\hat{\theta}^{\text{RATE}}_{M_j-M_{j-1}}$. The formula can be defined as follows:

$$\begin{aligned} \hat{\theta}^{\text{RATE}}_{M_j-M_{j-1}} &= \frac{\hat{\theta}^{A_{R-}}_{M_j-M_{j-1}} - \hat{\theta}^{A_{R+}}_{M_j-M_{j-1}}}{\hat{\theta}^{A_{R+}}_{M_j-M_{j-1}}} = \\ &= \frac{\sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_{A_{R-}} - \sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_{A_{R+}}}{\sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_{A_{R+}}}. \end{aligned} \tag{20}$$

where

$$I_{A_{R+}} = \begin{cases} 1 \;\; y_k^{(j)} - y_k^{(j-1)} > 0 \\ 0 \;\; y_k^{(j)} - y_k^{(j-1)} \leq 0 \end{cases}$$

and

$$I_{A_{R-}} = \begin{cases} 1 \;\; y_k^{(j)} - y_k^{(j-1)} < 0 \\ 0 \;\; y_k^{(j)} - y_k^{(j-1)} \geq 0 \end{cases}$$

Based on both previous estimators, those based on the absolute and relative gender gap of the absolute difference between j and $j-1$ are defined as follows, respectively:

$$\begin{aligned} \widehat{GG}^{\text{ABS}A}_{M_j-M_{j-1}} &= \hat{\theta}^{A_W}_{M_j-M_{j-1}} - \hat{\theta}^{A_M}_{M_j-M_{j-1}} = \\ &= \sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_A I_k^W - \sum_{k \in s_r^{(M_j,M_{j-1})}} v_k^{(j,j-1)} I_A I_k^M \,, \end{aligned} \tag{21}$$

$$\widehat{GG}^{\text{REL}A}_{M_j-M_{j-1}} = \frac{\widehat{GG}^{\text{ABS}A}_{M_j-M_{j-1}}}{\hat{\theta}^{A_M}_{M_j-M_{j-1}}} \,. \tag{22}$$

A positive value of these estimators would indicate, in absolute (percentage points) or relative terms (percentages), that the percentage of women who improved/increased, remained the same, or deteriorated/decreased their outcome in the target difference variable was higher than the corresponding percentage in the male population, while a negative value would indicate that the percentage was lower in women. Estimator 21 is defined as the linear combination of two estimators in certain domains, while estimator 22 is a quotient of two estimators based on the same samples, hence their theoretical properties can be easily derived.

Variance estimation

It is no simple task to develop suitable variance estimators for these proposed estimators taking into account the panel design used. The variance estimation problem in longitudinal surveys is addressed in several papers. For example, [28] considers variance estimation for Canada's Survey of Labor and Income Dynamics within a Taylor linearization approach and a bootstrap method.

Some other works are developed for rotation panels: [17] considers the estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland based on a four-panel rotation scheme where the non-response is modeled using a Poisson design. [50] considers variance estimation for weighting in the SILC survey in France with a rotation scheme consisting of four panels. Ardilly and Osier [31] considers the case of a panel survey in which solely the units in the original sample are followed over time, without reentry or late entry units at subsequent times to represent possible newborns. They assume a non-response model where the response probability at time t can be explained by the variables observed at times $0, t-1$, including the variables of interest. Zhou and Kim [51] also consider the estimation of a mean for a panel survey, in case of monotone non-response.

On the other hand, there is little work about variance estimation for machine learning methods. Some work about variance estimation for tree-based methods is the infinitesimal jackknife [52].

In this study, the formulas used for estimating the variance of indicators must take into account the structure and complexity of the ESSA survey. The main factors to consider for estimating the variance of the proposed estimators are the non-linearity of the estimators, total non-response at different survey stages and the use of machine learning models in conjunction with calibration. Therefore, we consider the application of bias-corrected and accelerated bootstrap [53]. It is well suited for a wide variety of scenarios, including ours, and it is easy to efficiently implement via Scipy [54], a standard Python scientific library.

Results

Observed non-response biases

To illustrate the observed biases produced mainly by non-response in the ESSA survey, Figs. 2 and 3 show the differences between the sample and the study population at measurement 4. These differences are according to the intersection of the sex variable with age, province, degree of urbanization and nationality. Thus, with respect to age, the largest differences between the values observed from the sample and those from the population are found in the youngest men (under 30 years old), in middle-aged women (between 35 and 54 years old) and in the oldest

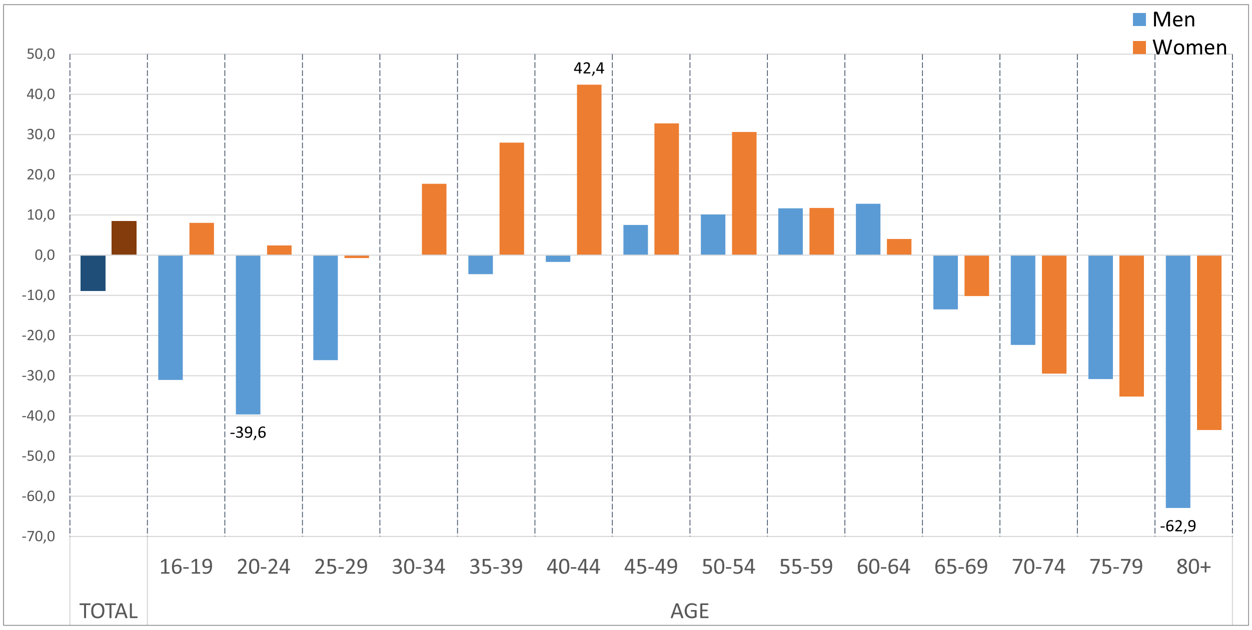


Fig. 2 Observed biases for the calibration variables in measurement 4 (age and sex)

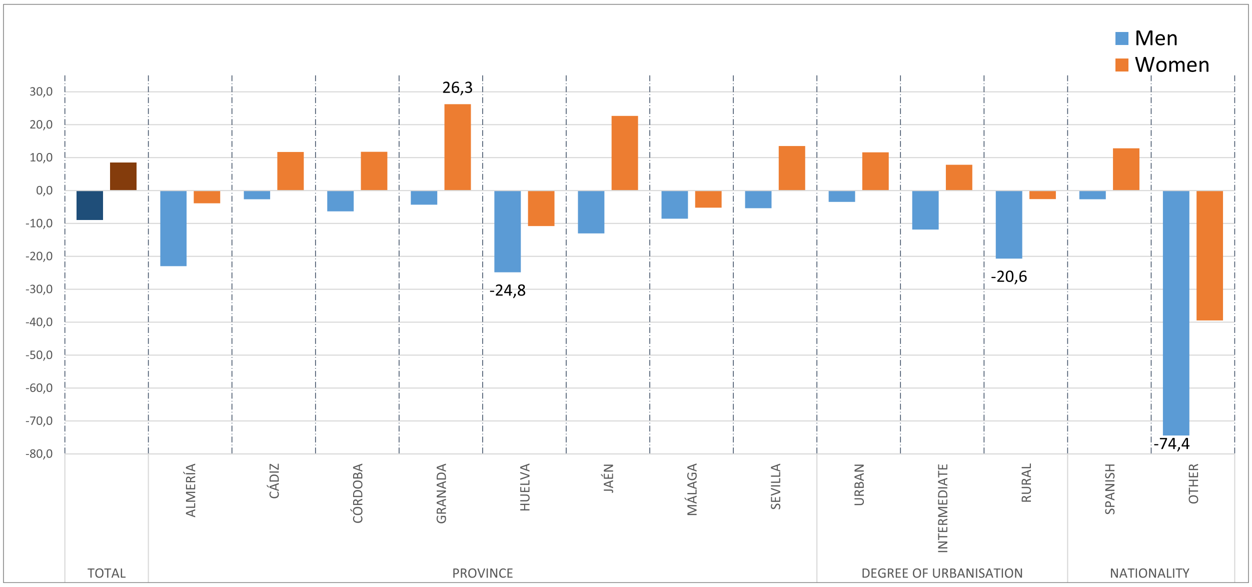


Fig. 3 Observed biases for the calibration variables in measurement 4 (sex-province, sex-urbanization and sex-nationality)

women and men (over 70 years old), these differences increasing with age. With regard to the other segmentation variables, the largest differences were found among people with a nationality other than Spanish, especially among men. These results are also observed although to a lesser extent in the previous measurements, showing a lower participation of these population groups in the ESSA, and therefore justify the need to adjust the sample weights.

Modelling non-response

ESSA thus has a non-monotone missing pattern and shows a lower participation of some population groups. This non-response, given in the new theoretical samples created in each measurement and in the panels involved in each measurement, is modelled with PSA as explained in the Methods section. In order to ensure that the XGBoost model is learning properly, we considered the following hyper-parameters:

- Number of estimators $\in [10, 1000]$: the number of trees forming the ensemble.
- Learning rate $\in [0.001, 0.9]$: the weight shrinkage applied after each boosting step.
- Maximum depth $\in [1, 30]$: the maximum number of splits that each tree can contain.
- Minimum child weight $\in [0, 10]$: the minimum total of instance weights needed to consider a new partition.
- Subsample $\in [0.6, 1]$: proportion of training data which is randomly sampled for each iteration.

The accuracy of the algorithm was tested with cross-validation. Therefore, training data is partitioned into 5 complementary subsets so that each one has the same proportion of $\delta_k^{(t)} = 1$ and $\delta_k^{(t)} = 0$ as the total. Then 5 models are trained leaving each one of the subsets out of the training data. For each model, the logistic loss was calculated for its corresponding remaining subset. The mean logistic loss is the estimated error.

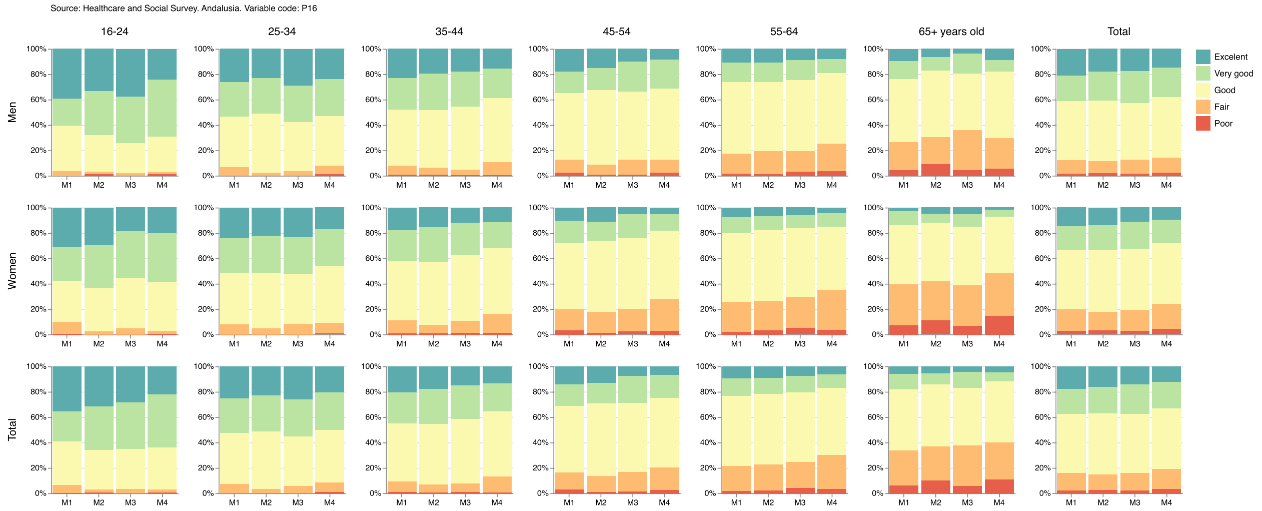
The values for the hyperparameters minimizing this estimated error were obtained using the Tree-structured Parzen Estimator (TPE) algorithm [55, 56]. TPE is implemented as default method in Optuna [57], an optimization library for Python.

The cross-sectional and longitudinal estimators were calculated by using these PSA weights.

Calibrating sample representativeness

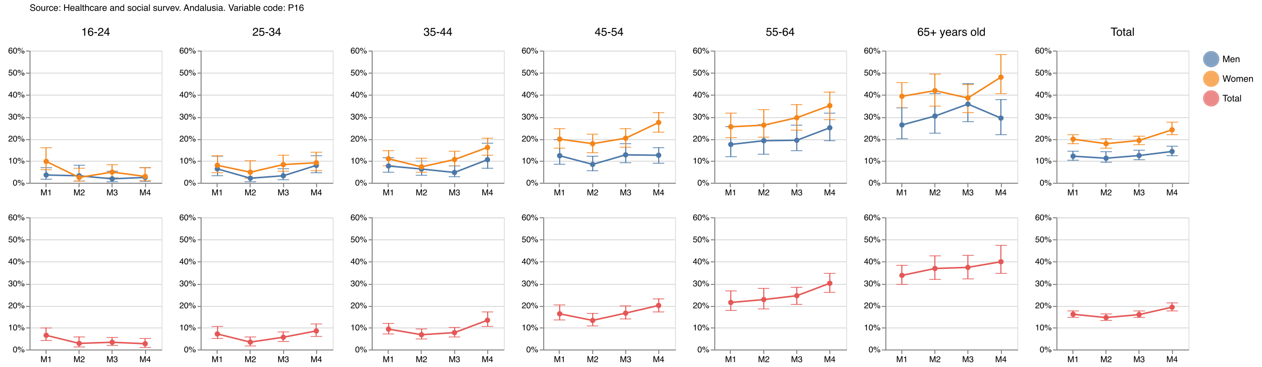
As explained in the Methods section, the weights obtained to adjust non-response are reweighted by calibration to achieve better representativeness of the population and reduce biases in the cross-sectional and longitudinal estimators.

The first ESSA measurement was carried out by IECA as another edition of the Social Household Survey that they have been conducting since 2007. Similarly, to deal with the observed biases, we had to apply the same


Source: Healthcare and Social Survey. Andalusia. Variable code: P16

16-24 25-34 35-44 45-54 55-64 65+ years old Total

Fig. 4 Estimations grouped by sex and age for the original categories of self-perceived general health


Source: Healthcare and social survev. Andalusia. Variable code: P16

16-24 25-34 35-44 45-54 55-64 65+ years old Total

Fig. 5 Percentages and confidence intervals at 95% level of people with fair or poor self-perceived general health

adjustment as IECA for the sample weights of the new samples and panels of each ESSA measurement, i.e. truncated raking calibration and the total population size for the intersection of the sex variable with province, age, urbanization grades and nationality as auxiliary information. The data for these totals were obtained from the 2019 Municipal Register of Inhabitants [58].

Cross-sectional estimators

Supplementary Table 1 shows for measurement 4 the percentages with corresponding 95% confidence intervals in addition to the sample size for each original category of the self-perceived general health variable grouped by sex and age. It may be observed from the chart that the percentages for the 'excellent' or 'very good' categories do not follow a clear pattern throughout measurements for the population between 16 and 34 years old, and 65+, either for men or for women. However, 'excellent' or 'very good' self-perceived health decreases for the population between 35 and 64 years old as the pandemic advances. This can be observed more as age increases, especially in women. This reduction results in an increment for the 'fair' and 'bad' categories. However, the 'good' general health category remains stable throughout the pandemic for each sex and age group. Figure 4 shows the percentages and confidence intervals given in Supplementary Table 1 not only for measurement 4, but also for all other ESSA measurements.

Based on these results, we dichotomized this variable with the categories 'excellent, very good and good' and 'fair and poor'. For each ESSA measurement, Supplementary Table 2 shows the percentages and 95% confidence intervals of this dichotomized self-perceived general health variable. These results can be seen in Fig. 5, which shows an increase in 'fair and poor' self-perceived health in measurements 3 and 4, this increase being slightly larger among women. Regarding age groups, evolution remained stable throughout the pandemic from lockdown onwards for the population aged between 16 and 24 for men and women alike. However, for the population aged over 25, the evolution worsens as age increases

Source: Healthcare and social survev. Andalusia. Variable code: P16_R

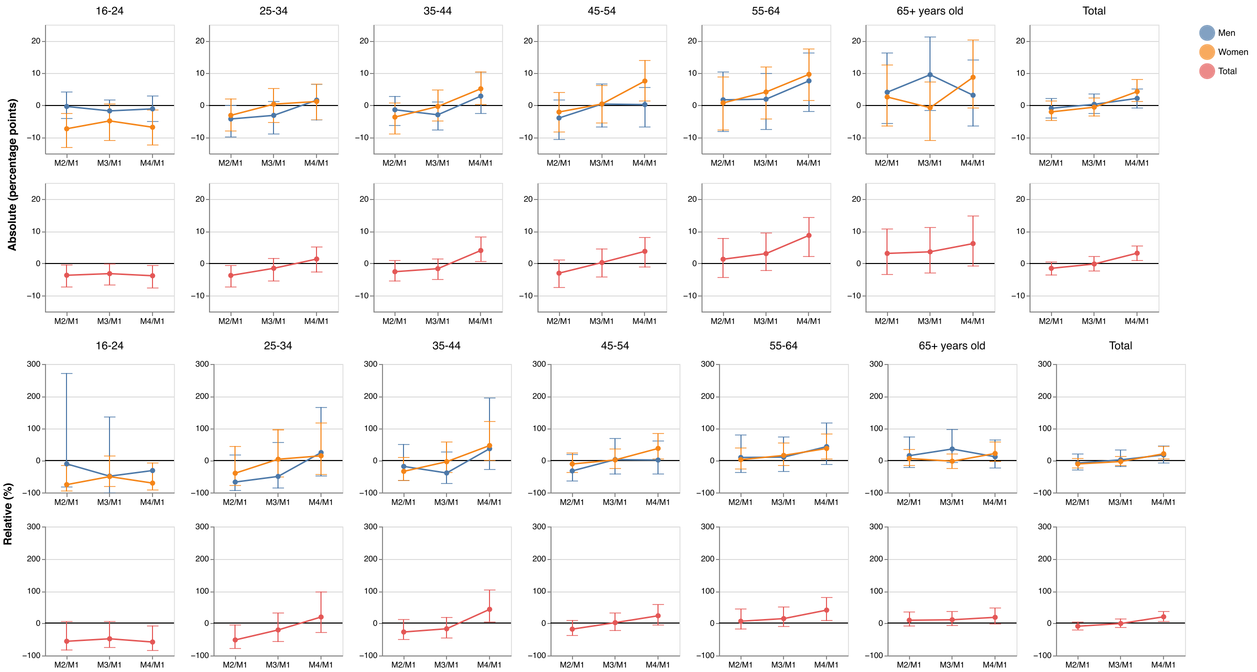


Fig. 6 Absolute percentage changes and 95% confidence intervals for people with fair or poor self-perceived general health

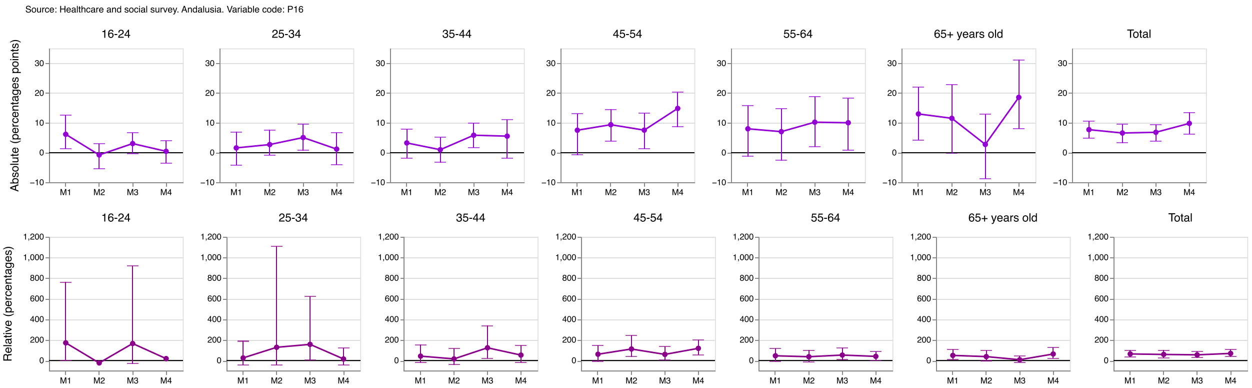
and the pandemic advances, especially in women. Therefore, this subpopulation got the highest 'fair or poor' general health values from the beginning of the lockdown for every age group above 25 years old.

Supplementary Table 3 shows the relative percentage changes and 95% confidence intervals for each measurement compared to measurement 1 for the 'fair or poor' self-perceived general health variable, while Fig. 6 shows the absolute percentage changes. It can be seen that fair or poor perception of general health increased in the general population by a 20% (CI95%=[5.2; 35.7] in measurement 4 compared to measurement 1. This increase was observed in all age groups, except for people under the age of 24 and over 65 years old, with no differences between women and men.

Supplementary Table 2 also includes absolute and relative gender gaps, i.e. the absolute difference (in percentage points) and the relative difference (in percentages) between women and men of a given measurement compared to the first measurement in which the target variable was gathered. This can be interpretated as a positive value indicating that women showed a positive difference (absolute or relative) in comparison to men in their 'fair or poor' self-perceived general health. This result could therefore be seen as a negative gender gap in the corresponding measurement (i.e. worse result or unfavorable to women as the reference category is 'fair or poor'). By contrast, a negative value would indicate that women showed a negative difference (absolute or relative) in comparison to men in their 'fair or poor' self-perceived general health, which could be seen as a positive gender gap (i.e. better result or favorable to women). These results are shown in Fig. 7; we can see, for example that both the absolute and relative gender gaps were positive throughout the pandemic, confirming an increasingly negative impact on women compared to men in terms of fair or poor self-perceived general health. Results by age reveal that the largest positive gender gaps were observed in people over 45 years old.

Longitudinal estimators

Supplementary Table 4 shows estimates of better, equal or worse self-perception of health in the population for a given measurement compared to the same population in the previous measurement. Thus, 21.7% of the study population improved their self-perceived general health in measurement 2 compared to measurement 1, but this percentage was slightly smaller in subsequent measurements. By contrast, 23.8% of this population group presented worse self-perceived general health in measurement 2 compared to measurement 1, with this percentage being slightly higher in subsequent measurements. When we analyze these results by sex and age, it can be observed that it is women between 25 and 54 years old who experience the decreases in the improvement of general health over the course of the pandemic and, conversely, women between 45-54 years old who experience


Source: Healthcare and social survey. Andalusia. Variable code: P16

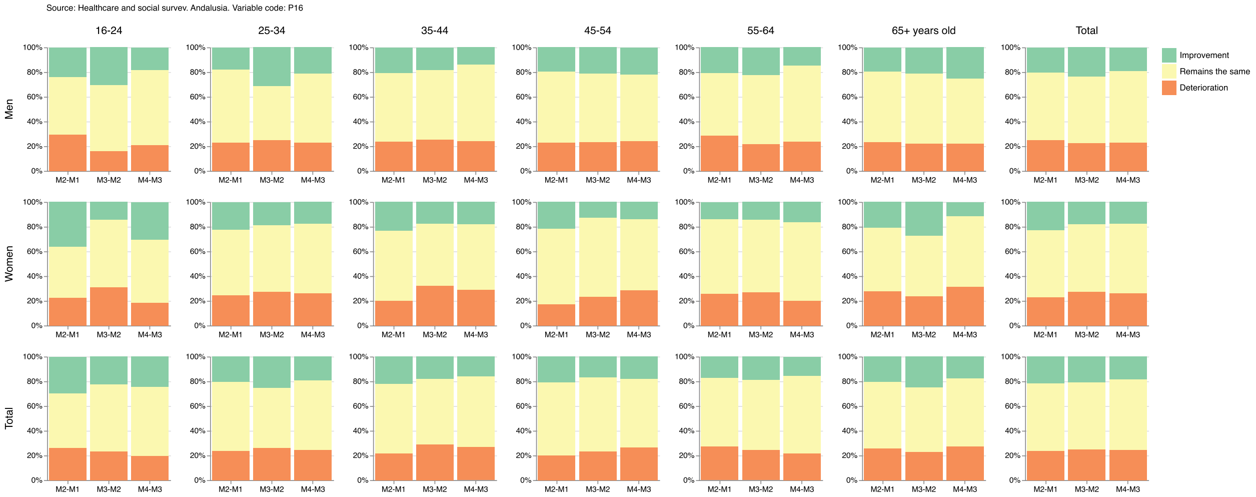
16-24 25-34 35-44 45-54 55-64 65+ years old Total

Fig. 7 Absolute and relative gender gap for the change in the fair or poor self-perceived health in each measurement (M)

an increase in the deterioration of self-perceived general health. On the other hand, the percentage of people who had remained the same self-perceived general health status in a given measurement compared to the previous one did not vary over the course of the pandemic, except for the population below 24 years old which did experience increases in the aforementioned percentage, going from 44% in measurement 2 to 55.9% in measurement 4. These results are shown in Fig. 8.

If we calculate the ratio of the population with worsening self-perceived general health (in a given measurement compared to the previous one) and the population where it improves, a positive value means that there are more people whose self-perceived general health has deteriorated than people whose health has improved, as seen in Fig. 9. In relative terms it could be observed that, in measurement 2 compared to measurement 1, 10% more of the population had worse self-perceived health than better health; this percentage increased to 19.5% and 34.2% in measurements 3 and 4 compared to measurements 2 and 3, respectively. These differences are larger in women, reaching values of 51.9% and 49.8% in measurements 3 and 4, respectively. If the ratio is analyzed according to the age of individuals regarding measurement 3 compared to measurement 2, deterioration of health was more frequently observed in women of any age.

Supplementary Table 4 also shows absolute and relative gender gaps in improvement self-perceived general health, staying the same or deteriorating in a measurement compared to the previous one in the same population. On the one hand, absolute gender gap is the absolute difference (in percentage points) between women and men with better, equal or worse self-perceived health in a measurement compared to the previous one, and on the other hand relative gender gap is the relative difference (in percentage) between women and men with better, equal or worse self-perceived


Source: Healthcare and social survev. Andalusia. Variable code: P16

16-24 25-34 35-44 45-54 55-64 65+ years old Total

Fig. 8 Percentage of population whose self-perceived general health improves, deteriorates or remains the same

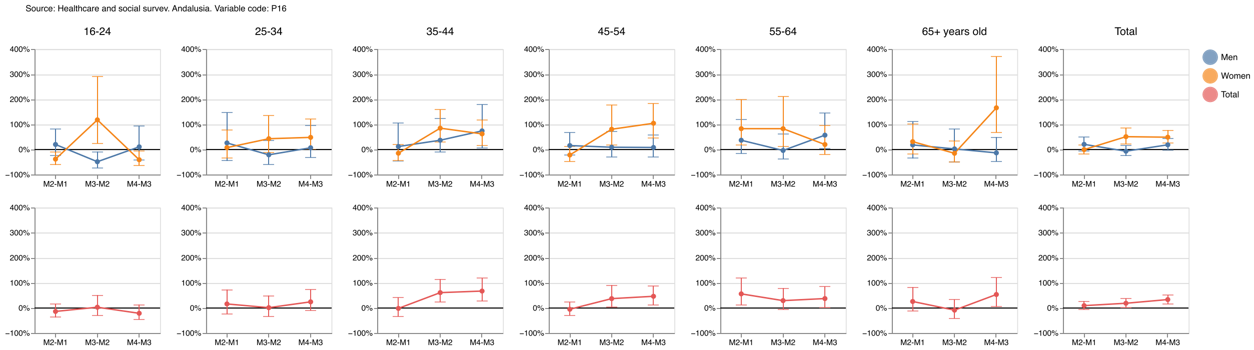


Fig. 9 Population whose health worsens between one measurement and the previous one compared to the population whose health improves

health. This means that a positive value in the gap (absolute or relative) indicates that the percentage of the female population with improved, equal or worsened self-perceived general health was greater than the corresponding percentage in the male population. A negative value would indicate that the percentage was smaller in women. Regarding deterioration of health, we observe in Fig. 10 that the percentage of the female population whose self-perceived health was worse in measurement 2 than in measurement 1 was 8.2% lower than among their male counterpart. However, this relative gender gap in health deterioration became positive in subsequent measurements, increasing to 20.9% and 13.8%, i.e. the deteriorating percentages were greater among women in measurement 3 and in measurement 4. This result was observed across all age groups, except for the population younger than 24 years old and the population between 55 and 64 years old.

Tables 1 and 2 summarize the name, table, figure, formula and interpretation relating to the estimators developed throughout this paper for cross-sectional and longitudinal samples, respectively.

Discussion

The rapid evolution of the COVID-19 pandemic has forced researchers to provide timely estimates on the disease's impact on the population. This has often led to the creation of survey studies which did not meet the criteria for being considered probabilistic, entailing many sources of error that may affect the final estimates obtained from them. In this sense, a recent scoping review on the methodological characteristics of the health surveys conducted in Spain early on in the COVID-19 pandemic included 55 studies (among over 3000 initially identified) [4]. An outcome of this review worth noting is the low proportion of longitudinal surveys identified (12.7%) and the implementation of some type of sampling adjustment (30.9%), even though most of the surveys were based on non-probability sampling (92.7%). Moreover, none of them considered the ESSA design or the reweighting approach described in this paper. Therefore, the ESSA survey is particularly valuable in the sense that its overlapping probability panel design offers the opportunity to obtain reliable estimates, both cross-sectional and longitudinal, on the impact of COVID-19 on health and its determinants. However, the analysis of the survey has not been

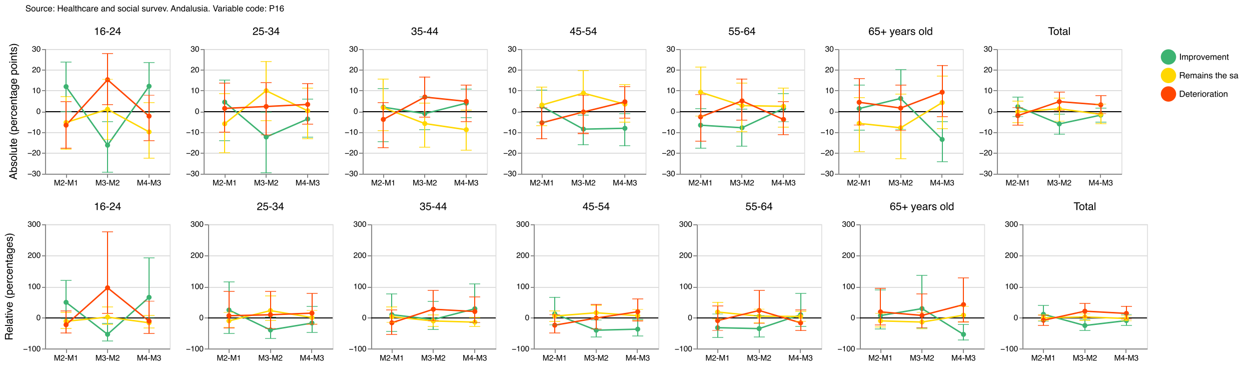


Fig. 10 Absolute and relative gender gaps in the improved, equal or worse self-perceived general health

Table 1 Name, table, figure, formula and interpretation of each estimator developed for the cross-sectional samples

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Original variables	1	4	(10)	Percentages, confidence intervals at 95%, sample size and population estimations at measurement 4, grouped by sex and age, for the original categories of self-perceived general health.
Dichotomized variables	2	5	(10)	Evolution of percentages and confidence intervals at 95%, grouped by sex and age, of people with fair or poor self-perceived general health. If the confidence intervals for the same measurement do not overlap, it can be said that there are statistically significant differences between women and men. Similarly, if the confidence intervals of two different measurements do not overlap, it can be said that there are statistically significant differences between them.
Absolute/Relative change	No/3	6	(13)/(14)	Evolution of absolute/relative changes and confidence intervals at 95%, grouped by sex and age, of people with fair or poor self-perceived general health in each measurement compared to measurement 1. A positive value indicates an increase, in percentage points/terms, in the fair or poor self-perception of overall health of the corresponding measure compared to the first measure. Conversely, a negative value indicates a decrease, in percentage points/terms, in the fair or poor self-perception of overall health of the corresponding measure compared to the first one. If the confidence interval does not include the value 0, this increase or decrease can be said to be statistically significant. Similarly, if the confidence intervals for the same measurement do not overlap, it can be said that there are statistically significant differences between women and men.
Absolute/Relative gender gap	2	7	(15)/(16)	Evolution in each measurement (M) of absolute/relative gender gaps (women versus men) and confidence intervals at 95%, grouped by age, of people with fair or poor self-perceived general health. A positive value indicates that women show, in percentage points/terms, a larger value in comparison to men in their 'fair or poor' self-perceived general health of the corresponding measurement. Therefore, this result could be seen as a negative gender gap (i.e., worse result or unfavorable to women) in the corresponding measurement. Conversely, a negative value indicates that women showed, in percentage points/terms, a smaller value in comparison to men in their 'fair or poor' self-perceived general health. It could be seen as a positive gender gap (i.e., better result or favorable to women) in the corresponding measurement. If the confidence interval does not include the value 0, the corresponding gender gap can be said to be statistically significant.

Table 2 Name, table, figure, formula and interpretation of each estimator developed for the longitudinal samples

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Longitudinal difference	4	8	(19)	Percentage of population and confidence intervals at 95% whose self-perceived general health increases/improves, decreases/deteriorates or remains the same between a measurement and the previous one
Decrease Increase Rate	No	9	(20)	Percentage of the population and confidence intervals at 95% that worsens their general health (in a given measurement compared to the previous one) and the population that improves it. A positive value means that there are more people whose self-perceived general health has deteriorated than people whose health has improved.
Absolute/Relative gender gap in the absolute difference	4	10	(21)/(22)	Absolute/Relative difference (in percentage points/terms) and confidence intervals at 95% between women and men with better, equal or worse self-perceived health in a measurement compared to the previous one. A positive value indicates that the percentage of the female population with improved, equal or worsened self-perceived general health was greater than the corresponding percentage in the male population. A negative value would indicate that the percentage was smaller in women.

exempt from statistical adjustments to correct for attrition and survey non-response.

The two-step adjustment procedure has been established in this study to remove the two main sources of error in the sampling design: population non-response, understood as people who did not take part in the survey despite having been selected in the sample, which was treated in the calibration step, and panel non-response, understood as people who participated in some of the measurements but did not follow up in subsequent ones. Panel non-response has been treated using PSA, which is a technique often used for addressing selection bias in online surveys [59] but which can also be used for non-response; in fact, it was originally adapted from [29] for this matter [60].

In our study, the XGBoost technique has been used to model lack of response from one measurement to another. The authors in [21] propose a novel neural network structure and they compare it with advanced gradient boosting methods such as XGBoost, using a justification that these are still state-of-the-art in spite of their age. It would be of great interest to consider their recent proposal in order to model non-response. However, it is still very experimental, as evidenced by the scarcity of papers and implementations. For such an important application as the ESSA, we therefore prefer an established method. The application developed in this work is one example where techniques in the machine learning field need to be combined with other important techniques in survey research, such as calibration and PSA, when studying non-response in a panel setting.

The propensities obtained could also be further analyzed by interpretable models [61] in order to determine the factors associated with non-response. This would help establish strategies for obtaining higher response rates, or at least for offsetting low response rates by targeting specifically hard-to-reach groups.

The results observed in the different estimates obtained from the self-perceived general health variable show that the impact of the pandemic has affected age age groups and genders differently. More precisely, self-perceived general health seems to have decreased more notably in older age groups and among women, according to the evolution of cross-sectional estimates and longitudinal estimates alike. The gender gap in both absolute and relative terms generally increased as the pandemic advanced, meaning that the differences (mostly decreases in self-perceived general health) have been larger and worse in women than in men. We chose this important health outcome because of the enormous amount of research it invests in studying risk factors and policy interventions in Public Health. This is due to its ability to summarize more objective measures such as morbidity, mortality, and clinical assessments of health conditions [62].

Some limitations must be noted in this study. Firstly, it is a well-known fact that subjective variables usually entail measurement errors, as the response given in such questions by the interviewee may depend on numerous unmeasurable factors unrelated to the subject being studied, but which distance the final response from the objective value that should be given. Further studies should consider the measurement of such variables using validated instruments for a more objective understanding of the subject. In any case, the methodology developed in this research can be extended to any variables and scales.

Secondly, we assume a covariate-dependent missingness pattern, as is usual in propensity score adjustment [63–65]. In a panel survey, it may be more realistic to assume Missing at Random which allows for dependence on the observed y-values in the previous years [31, 51], but has the drawback of the adjustment weights varying for each variable, which is not useful for multipurpose surveys such as the ESSA. This survey has more than 400 variables and is used by health researchers from different specialties, the objective being to give adjusted weights to each unit of the sample so that each researcher can use them to carry out their specific studies related to the variables that interest them. It would also be interesting to see the differences between the estimates with these two different patterns and whether this difference in accuracy offsets the complexity of having to build a different response model for each variable.

Another limitation in this work is that we have considered a situation in which the study population does not vary over time. This is justified because the new measurements are made with little difference compared to the first measurement (at one month, 6 months and 12 months) and all the samples are obtained from the same sampling frame [23], so we have assumed that the sample designs refer to the same population. In fact, the difference in population between the 2019 and 2020 population frameworks is 0.6% in relative terms, or, in absolute terms, about 42,000 people out of almost 7.2M people over 16 years of age residing in Andalusia [23].

These methods would therefore not be well suited to overlapping panel surveys where samples are drawn from very different frames in different years and therefore from different populations. In such cases, the proposed methodology would have to be adapted.

Conclusion

For addressing future health crises such as COVID-19, potential coverage and non-response biases in surveys must be reduced by means of utilizing reweighting techniques.

In this respect, we propose a new reweighting approach to produce suitable estimators for both cross-sectional and longitudinal samples in overlapping panel surveys. To achieve this, first the original sampling design weights are corrected by modelling non-response in respect of the longitudinal sample obtained in a previous measurement using machine learning techniques, and then, they are calibrated using the auxiliary information available at the population level.

We apply this methodology to estimate totals, proportions, ratios, and differences between measurements as well as gender gaps in the variable of self-perceived general health. The descriptive results for this variable are an example applied to this paper to show the different estimators, tables and figures developed which can be replicated with other variables and scales from other overlapping panel surveys. In fact, they are all extended to the 400+ ESSA variables through the web platform at www.easp.es/info/ESSA. On this website, after selecting the set of variables to be described, the estimators to be shown and the segmentation variables to be considered (sex and age or sex and degree of urbanization), the user obtains the corresponding interactive figures to help interpret the selected variables. This will allow the scientific epidemiological research community not only to access the descriptive results for all the ESSA variables, but also to carry out their own analyses by downloading the ESSA database and code developed, used as the basis for the conclusions of this paper.

Abbreviation

ESSA	Spanish acronym for Encuesta Sanitaria y Social de Andalucía (Andalusia Healthcare and Social Survey)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02171-z.

Additional file 1: Table 1. Estimations grouped by sex and age for the original categories of self-perceived general health at measurement 4 (Excel file).

Additional file 2: Table 2. Percentages, gender gaps and confidence intervals at 95% of people with fair or poor self-perceived general health (Excel file).

Additional file 3: Table 3. Relative percentage changes and 95% confidence intervals for people with fair or poor self-perceived general health (Excel file).

Additional file 4: Table 4. Percentage of people whose self-perceived general health improves, deteriorates or remains the same, and absolute and relative gender gap (Excel file).

Acknowledgements

We thank Iria Enrique and her team at the Institute of Statistics and Cartography of Andalusia (IECA, Instituto de Estadística y Cartografía de Andalucía) for carrying out the first ESSA measurement in the framework of their Social Household Survey (ESOC, Spanish acronym), and for providing the population and sample data to adjust the sampling weights of the ESSA measurements.

Authors' contributions

Conceptualisation, A.C., M.R.; Methodology, A.C., M.R., R.F. and L.C.; Resources, A.C., C.S.; Data curation, C.S.; Writing-original draft preparation, A.C., M.R., R.F. and L.C.; Writing-review and editing, A.C., M.R., R.F., C.S. and L.C.; Project administration, A.C. and C.S.; Funding acquisition, A.C. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding

This work is funded by the Spanish Ministerio de Ciencia and Innovation MCIN/AEI/10.13039/501100011033. It is also funded by the Spanish Ministerio de Ciencia and Innovation PI20/00855 and the Ministerio de Economía y Competitividad [grant PID2019-106861RB-I00]. ESSA funding was obtained from the Andalusian Institute of Statistics and Cartography (IECA), the Andalusian School of Public Health (EASP), and the competitive tenders of the Santander Universities SUPERA COVID-19 Fund (SAUN), the Conference of Rectors of Spanish Universities (CRUE), and the Higher Council for Scientific Research (CSIC), in addition to the COVID Competitive Aid Program -19 from Pfizer Global Medical Grants. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The dataset and code supporting the conclusions of this article is available in the ESSA repository at www.easp.es/info/ESSA.

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations. An informed consent was obtained from all subjects involved in the ESSA survey on which this study is based. It was also approved by the Research Ethics Committee of the Department of Health and Families of the Andalusian Regional Government (protocol code 10/20, dated 07 December 2020). The ESSA is an activity included in the processing activities registry of the Department of Health and Families of the Andalusian Regional Government and is linked to the Andalusian Health Survey (EAS, Encuesta Andaluza de Salud), an official statistical operation included in the Andalusian Statistical and Cartographic Plan 2023-2029 (Plan Estadístico y Cartográfico de Andalucía), code 04.02.16.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Statistics and Operational Research, University of Granada, Granada, Spain. [2]Institute of Mathematics, University of Granada, Granada, Spain. [3]Department of Public Health, Andalusian School of Public Health, Granada, Spain. [4]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain.

Received: 16 June 2023 Accepted: 5 February 2024
Published: 15 February 2024

References

1. Pastor-Barriuso R, Pérez-Gómez B, Oteo-Iglesias J, Hernán MA, Pérez-Olmeda M, Fernández-de Larrea N, et al. Design and Implementation of a Nationwide Population-Based Longitudinal Survey of SARS-CoV-2 Infection in Spain: The ENE-COVID Study. Am J Publ Health. 2023;113(5):525–32. https://doi.org/10.2105/AJPH.2022.307167.
2. Lazarus JV, Romero D, Kopka CJ, et al., The COVID-19 Consensus Statement Panel. A multinational Delphi consensus to end the COVID-19 public health threat. Nature. 2022;611:332–345. https://doi.org/10.1038/s41586-022-05398-2.
3. Lohr SL. Sampling: Design and Analysis. 3rd ed. Boca Raton: CRC Press; 2022.
4. Sánchez-Cantalejo Garrido C, Yucumá Conde D, Rueda García M, Martín Ruiz E, Olry de Labry Lima A, Higueras C, et al. Scoping Review of the methodology of large health surveys conducted in Spain early on the COVID-19 pandemic. Front Public Health. 2023;1–11. https://doi.org/10.3389/fpubh.2023.1217519.
5. Sánchez-Cantalejo C, Rueda MdM, Saez M, Enrique I, Ferri R, Fuente MdL, et al. Impact of COVID-19 on the Health of the General and More Vulnerable Population and Its Determinants: Health Care and Social Survey-ESSOC, Study Protocol. Int J Environ Res Public Health. 2021;18(15):8120. https://doi.org/10.3390/ijerph18158120.
6. Kalton G, Citro CF. Panel surveys: Adding the fourth dimension. Innov Eur J Soc Sci Res. 1995;8(1):25–39. https://doi.org/10.1080/13511610.1995.9968429.
7. Ardilly P, Lavallée P. Weighting in rotating samples: The SILC survey in France. Surv Methodol. 2007;33(2):131–7.
8. Kalton G, Lepkowski J, Lin TK. Compensating for wave nonresponse in the 1979 ISDP research panel. In: Proceedings of the Survey Research

Methods Section. vol. 372. 732 North Washington Street Alexandria, VA 22314-1943 USA: American Statistical Association; 1985. p. 377.

9. Lepkowski JM. Treatment of wave nonresponse in panel surveys. In: Kalton G, Lepkowski J, Heeringa S, Lin TK and Miller ME. The treatment of person-wave nonrespose in longitudinal surveys. No 26. U.S. Suitland, MD 20746 United States: Department of Commerce Bureau of the Census; 1987. p. 90–130.
10. Kalton G, Brick JM. Weighting schemes for household panel surveys. Surv Methodol. 1995;21(1):33–44.
11. Deville JC, Särndal CE. Calibration Estimators in Survey Sampling. J Am Stat Assoc. 1992;87(418):376–82. https://doi.org/10.1080/01621459.1992.10475217.
12. Kern C, Klausch T, Kreuter F. Tree-based Machine Learning Methods for Survey Research. Surv Res Methods. 2019;13(1):73–93.
13. Kern C, Weiß B, Kolb JP. Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning. J Surv Stat Methodol. 2021;smab009. https://doi.org/10.1093/jssam/smab009.
14. Rendtel U, Harms T. Weighting and Calibration for Household Panels. In: Lynn P, editor. Methodology of Longitudinal Surveys. Chichester: Wiley; 2009. p. 265–86. https://doi.org/10.1002/9780470743874.ch15.
15. Arcos A, Rueda MdM, Pasadas-del Amo S. Treating Nonresponse in Probability-Based Online Panels through Calibration: Empirical Evidence from a Survey of Political Decision-Making Procedures. Mathematics. 2020;8(3):423. https://doi.org/10.3390/math8030423.
16. Lavallée P, Deville J. Theoretical Foundations of the Generalised Weight Share Method. In: Proceedings of the International Conference on Recent Advances in Survey Sampling. International Conference on Recent Advances in Survey Sampling. 2002. p. 127–36.
17. Massiani A. Estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland. Surv Methodol. 2013;39(1):121–49.
18. Verma V, Betti G, Ghellini G. Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC. Piazza San Francesco, 7, Siena, Italy: Università di Siena, Dipartimento di metodi quantitativi; 2006.
19. Castro-Martín L, Rueda MdM, Ferri-García R. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment. Mathematics. 2020;8(11):2096. https://doi.org/10.3390/math8112096.
20. Ferri-García R, Rueda MdM. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PLoS ONE. 2020;15(4):e0231500. https://doi.org/10.1371/journal.pone.0231500.
21. Arik SÖ, Pfister T. Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35. 2275 East Bayshore Road, Suite 160 Palo Alto, CA 94303 USA: AAAI Press; 2021. p. 6679–87.
22. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Degree of urbanization. 2020. https://www.juntadeandalucia.es/institutodeestadisticaycartografia/gradourbanizacion/. Accessed 11 Feb 2024.
23. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Longevity. 2020. https://www.juntadeandalucia.es/institutodeestadisticaycartografia/longevidad/. Accessed 11 Feb 2024.
24. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663–85.
25. Roberts G, Kovacevic M, Mantel H, Phillip O. Cross-sectional inference based on longitudinal surveys: some experiences with statistics canada surveys. Stat Canada. 2001;1–10.
26. Kim S. In: Gu D, Dupre ME, editors. Cross-Sectional and Longitudinal Studies. Cham: Springer International Publishing; 2021. pp. 1251–5.
27. Lavallee P. Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. Surv Methodol. 1995;21(1):25–32.
28. Kovacevic MS. Cross-sectional inference based on longitudinal surveys: Some experiences with statistics Canada surveys. In: Federal Committee on Statistical Methodology Conference. Federal Committee on Statistical Methodology; 2001.
29. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. https://doi.org/10.1093/biomet/70.1.41.
30. Ferri-García R, Rueda MdM. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat Oper Res Trans. 2018;42(2):159–62.
31. Juillard H, Chauvet G. Variance estimation under monotone nonresponse for a panel survey. Surv Methodol. 2018.
32. Chen Y, Li P, Wu C. Doubly robust inference with nonprobability survey samples. J Am Stat Assoc. 2020;115(532):2011–21.
33. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
34. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat. 2000;28(2). https://doi.org/10.1214/aos/1016218223.
35. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29(3):337–46. https://doi.org/10.1002/sim.3782.
36. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLoS ONE. 2011;6(3):e18174. https://doi.org/10.1371/journal.pone.0018174.
37. McCaffrey DF, Ridgeway G, Morral AR. Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. Psychol Methods. 2004;9(4):403–25. https://doi.org/10.1037/1082-989X.9.4.403.
38. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013;32(19):3388–414. https://doi.org/10.1002/sim.5753.
39. Tu C. Comparison of various machine learning algorithms for estimating generalized propensity score. J Stat Comput Simul. 2019;89(4):708–19. https://doi.org/10.1080/00949655.2019.1571059.
40. Zhu Y, Coffman DL, Ghosh D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. J Causal Infer. 2015;3(1):25–40. https://doi.org/10.1515/jci-2014-0022.
41. Rueda MdM, Pasadas-del Amo S, Rodríguez BC, Castro-Martín L, Ferri-García R. Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. Biom J. 2022. https://doi.org/10.1002/bimj.202200035.
42. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8. 15-17, 69121 Heidelberg, Germany: Springer, Tiergartenstr; 2001. p. 63–6.
43. Saerens M, Latinne P, Decaestecker C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Comput. 2002;14(1):21–41. https://doi.org/10.1162/089976602753284446.
44. Rueda M, Martínez S, Martínez H, Arcos A. Mean estimation with calibration techniques in presence of missing data. Comput Stat Data Anal. 2006;50(11):3263–77. https://doi.org/10.1016/j.csda.2005.06.003.
45. Kott PS, Liao D. One step or two? Calibration weighting from a complete list frame with nonresponse. Surv Methodol. 2015;41(1):165–82.
46. Cabrera-León A, Lopez-Villaverde V, Rueda M, Moya-Garrido MN. Calibrated prevalence of infertility in 30- to 49-year-old women according to different approaches: a cross-sectional population-based study. Hum Reprod. 2015;30(11):2677–85. https://doi.org/10.1093/humrep/dev226.
47. Devaud D, Tillé Y. Rejoinder on: Deville and Särndal's calibration: revisiting a 25-year-old successful optimization problem. TEST. 2019;28(4):1087–91. https://doi.org/10.1007/s11749-019-00685-z.
48. Särndal CE, Swensson B, Wretman JH. Model assisted survey sampling. 1st ed. Springer series in statistics. New York Berlin Heidelberg: Springer; 2003.
49. Salas Quijada C, López-Contreras N, López-Jiménez T, Medina-Perucha L, León-Gómez BB, Peralta A, et al. Social Inequalities in Mental Health and Self-Perceived Health in the First Wave of COVID-19 Lockdown in Latin America and Spain: Results of an Online Observational Study. Int J Environ Res Public Health. 2023;20(9). https://doi.org/10.3390/ijerph20095722.
50. Ardilly P, Osier G. Cross-sectional variance estimation for the French "Labor Force Survey". In: Survey Research Methods. European Survey Research Association (ESRA). vol. 1. Unter Sachsenhausen 6-8 D-50667

Cologne Germany: Leibniz Institute for the Social Sciences; 2007. p. 75–83.
51. Zhou M, Kim JK. An efficient method of estimation for longitudinal surveys with monotone missing data. Biometrika. 2012;99(3):631–48.
52. Wager S, Hastie T, Efron B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. J Mach Learn Res. 2014;15(1):1625–51.
53. Efron B. Better bootstrap confidence intervals. J Am Stat Assoc. 1987;82(397):171–85.
54. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
55. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc.; 2011. https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html. Accessed 23 Nov 2021.
56. Bergstra J, Yamins D, Cox D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: Proceedings of the 30th International Conference on Machine Learning. PMLR; 2013. pp. 115–23. https://proceedings.mlr.press/v28/bergstra13.html. Accessed 23 Nov 2021.
57. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage: ACM; 2019. pp. 2623–31. https://doi.org/10.1145/3292500.3330701.
58. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Population and Housing Census. 2020. https://www.juntadeandalucia.es/institutodeestadisticaycartografia/padron/. Accessed 11 Feb 2024.
59. Lee S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J Off Stat. 2006;22(2):329.
60. Little RJ. Survey nonresponse adjustments for estimates of means. Int Stat Rev/Rev Int Stat. 1986;54:139–57.
61. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Commun ACM. 2019;63(1):68–77.
62. Jylha M. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. Soc Sci Med. 2009;307–16. https://doi.org/10.1016/j.socscimed.2009.05.013.
63. Castro-Martín L, Rueda MdM, Ferri-García R. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys. J Comput Appl Math. 2022;404. https://doi.org/10.1016/j.cam.2021.113414.
64. Castro-Martín L, Rueda MdM, Ferri-García R, Hernando-Tamayo C. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. Mathematics. 2021;9(23):2991. https://doi.org/10.3390/math9232991.
65. Ferri-García R, Rueda MdM, Cabrera-León A. Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment. Mathematics. 2021;9(7):791.

Publisher's Note