REWEIGHTING WITH CALIBRATION AND XGBOOST IN PANEL SURVEYS: APPLICATION TO THE HEALTH CARE AND SOCIAL SURVEY.

L. Castro-Martín

department of quantitative methods in economics.university of granada, granada, spain.

https://orcid.org/0000-0002-0934-4219

M. Rueda

department of statistics and operational research, math institute of the university of granada, university of granada, granada, spain.

https://orcid.org/0000-0002-2903-8745

C. Sánchez-Cantalejo

andalusian school of public health, granada, spain.

https://orcid.org/0000-0002-3600-7007

R. Ferri-García

department of statistics and operational research,

university of granada, granada, spain.

https://orcid.org/0000-0002-9655-933x

J. Hidalgo Calderón

department of geometry and topology, math institute of

the university of granada, granada, spain.

A. Cabrera

andalusian school of public health, granada,

spain.https://orcid.org/0000-0002-4812-1026

Acknowledgments

We thank Iria Enrique and her team at the Instituto de Estadística y Cartografía de Andalucía (IECA) for carrying out the first ESSOC measurement in the framework of their Social Household Survey, as well as for providing the population and sample data

to adjust the sampling weights of the ESSOC measurements.

Corresponding author: M. Rueda, Department of Statistics and O.R. Faculty of Sciences. University of Granada. 18071 Granada, Spain. Email: mrueda@ugr.es

REWEIGHTING WITH CALIBRATION AND XGBOOST IN PANEL SURVEYS: APPLICATION TO THE HEALTH CARE AND SOCIAL SURVEY

Abstract

Healthcare statistical services worldwide have used probability surveys to respond to such information needs. The Health Care and Social Survey (ESSOC) research project arises from the need to provide data on the evolution of the COVID-19 impact that can be considered when making decisions to prepare and provide an effective Public Health response in the different affected populations. This survey has an overlapping panel design with 4 measurements throughout 1 year with random samples stratified by province and degree of urbanization. Thus, each ESSOC measurement is composed of two samples: a longitudinal sample from previous measurements and a new sample at each measurement. The advantage of this design is that, in addition to being able to obtain longitudinal estimates, cross-sectional estimates are more accurate because of the larger sample size. However, the problem of non-response is particularly aggravated in the case of panel surveys, due to the fatigue of the population to be repeatedly surveyed. Taking into account the design, timing and objectives of this survey, in this work, we test a new reweighting method that produces suitable estimators for overlapping panel surveys affected by non-response. In each measurement, missing units are substituted by new surveyed units, allowing the obtention of cross-sectional and longitudinal estimates. The weights are the result of a two-step process: the original sampling design weights are corrected during a 1st phase by modeling the non-response with respect to the longitudinal sample obtained in the first measurement using XGBoost. Then, during a 2nd phase, they are calibrated using the auxiliary information available at the population level. The proposed method is applied to the estimation of totals, proportions, differences between measurements as well as gender gaps in the ESSOC.

Key words: Public health, COVID-19, panel surveys, sampling, machine learning, non-response

1. Introduction

The urgent need to control the expansion rate of COVID-19 requires a quick and efficient assessment of the situation, based on predicting and quantifying the main parameters involved in this phenomenon. Healthcare statistical services worldwide have used probability surveys to respond to information needs concerning the social, economic and health impact of the disease, or on its seroprevalence and evolution or on the characteristics of the infected population, especially those most vulnerable to the virus due to their age, risk of exclusion, health conditions or dependency. 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.

The Health Care and Social Survey (ESSOC, Encuesta Sanitaria y SOCial) research project arises from the need to provide data on the evolution of the COVID-19 impact that can be considered when making decisions to prepare and provide an effective Public Health response in the different affected populations, especially in the most vulnerable ones, such as, among others, the elderly, the chronically ill, or persons at risk of exclusion[1]. 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 that with greater socioeconomic vulnerability. 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 ESSOC has an overlapping panel design[2]. It consists of a series of measurements broken down into a new sample and a longitudinal sample for each measurement, with the exception of the first measurement where the entire sample is new. Thus, compared to rotating panel surveys[3], the ESSOC sampling design is non-rotational, i.e. the units included in each measurement remain in the following measurements until the last 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 ESSOC. In this way, the use of new samples at each measurement t enables to be representative of the whole population at time $t+1$, and therefore enables cross-sectional estimation at this time as well. 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. Thus, the lack of response grows with the number of occasions or measurements, due, among others, to the fatigue of the panelist to be repeatedly interviewed. 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 one wants to take into account the biases produced both by the lack of response, as well as by the lack of coverage and representativeness of the sample.

Some methods of handling wave nonresponse in panels are provided in[4],[5] and[6]. However, the methods used in those studies are based solely on weighting the effective sample according to the theoretical sample in the strata used and, at most, calibrating the sample weights in terms of population totals for socio-demographic stratification variables such as sex, age or territory (e.g. region, province or habitat level). Another set of studies focuses on modeling different types of response patterns in panels.[7] compare the usage of different Machine Learning (ML) methods for modeling nonresponse in the German Socio-Economic Panel Study (GSOEP) and recently[8] propose a general framework for building and evaluating nonresponse prediction models with panel data, but this study is focused on model building and evaluation without utilizing the obtained predictions to correct the bias in the estimations.

Nonresponse in panel studies has traditionally been tackled by using nonresponse weights. Although there are reweighting methods to deal with these types of biases, they have been proposed fundamentally for the case of cross-sectional surveys and there are few studies that provide a formal methodology for their treatment in this type of panel as the following. In[9], the authors discuss adjustments for nonresponse and how calibration can be carried out in panel studies in general and what effects it creates. They consider three possible ways of calibration: 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 as well as final calibration, are carried out). Several approaches are tested in[10] to produce calibration estimators that are suitable for survey data affected by non response where auxiliary information exists at both the panel level and the population level.

Longitudinal and cross-sectional weighting are considered in[3] for rotating samples in the context of the SILC survey in France. In this survey the sampling each year is formed by combining nine panel subsamples and the longitudinal weights are assigned as an average of the weights in each time in which a unit belongs to the sample following the weight-share method[11]. This method is also used in[12] for obtaining cross-sectional indicators for the SILC survey in Switzerland based on a four-panel rotation scheme.[13] develop longitudinal and cross-sectional weighting procedures in rotational household panel with reference to the EU-SILC design (4-year rotational design) by a step-by-step procedure starting from 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, those authors do not consider the application of these adjustment methods in designs such as the present study, i.e. in overlapping panel surveys where the units included in each measurement remain in the following measurements until the last one, and in which each measurement is completed with a new sample, with the exception of the first one as the whole sample is new, which is the research gap that this paper addresses.

Therefore, in this work we propose weighting methods to address the bias associated with the dropout from overlapping panel survey data for estimating totals, proportions and change or differences of a population characteristic using various combined methods such as Propensity Score Matching, machine learning and calibration. The reweighting methods are formulated based on the ESSOC survey so they can be adapted to any other type of overlapping panel design.

The paper is organized as follows. First, we introduce and describe a real survey about COVID-19, the Health Care and Social Survey, in Section 2. Then, in Section 3 we review the estimation in overlapping panels to set the framework and the notation. We present cross-sectional and longitudinal estimators in Sections 4 and 5 and we show how to use machine learning methods to reweighting for non-response based on the data of previous occasions. In Section 6, we apply some of the estimators developed and proposed methods to a specific variable (self-perceived general health) from the Health Care and Social Survey. Finally, we highlight the most relevant findings and conclusions in Section 7.

2. The ESSOC Study Framework

The Health Care and Social Survey (ESSOC, Encuesta Sanitaria y SOCial) provides a follow-up over time on the impact of the pandemic, and its resulting lockdown, on the population of Andalusia (Spain) over the age of 16.

As shown in Figure 1, the ESSOC study includes four measurements. The first one, $s^{(1)}$, coincides with the beginning of the Spanish State of Alarm on April 2020, while the 2nd measurement $s^{(2)}$ was carried out in June and July (a month after the 1st interview, coinciding with the de-escalation), the 3rd measurement $s^{(3)}$ in November and December (6 months after the 1st interview and coinciding with the 2nd wave of the pandemic), and the 4th measurement $s^{(4)}$ in April and May 2021 (12 months after the 1st interview, coinciding with the opening of mobility and the end of the state of alarm). All the theoretical samples have a size of 3000 people. They were obtained using an overlapping panel design so the individuals from the previous measurement are sampled again. However, the non-response is compensated with another sample including new individuals. The details of this non-response and the effective sample size for each measurement can be consulted at Figure 1. That figure 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 selection of the new sample in each measurement is stratified simple random sampling according to province and degree of urbanization (urban, semi-urban and rural, according to the methodology described by EUROSTAT for the allocation of territorial typologies in statistical grids of 1 km2 where population resides, more information at[14]. 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. Thus, the new sample is distributed among the 8 Andalusian provinces in proportion to the population size of the province. Within each province, the sample allocation is proportional to the population size of each degree of urbanization. As for the longitudinal sample of a given measurement, this is composed of the sample of the previous measurement, with the exception of the first measurement, which would not have a longitudinal sample as it is the first one. The population framework used for the extraction of the samples of the population aged 16 years old and over residing in family dwellings in Andalucía, comes from the Longitudinal Population Database of Andalucía (BDLPA) as of 1 January 2019 (more information on this register is available at[15]. A detailed description of the protocol followed for this survey can be seen in[1].

To visualize the observed biases produced mainly by non-response in the ESSOC surveys, Figures 2 and 3 depict the differences between the sample and the population at measurement 4. Those differences are according to the cross 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 youngest men (under 30 years old), in middle aged women (between 35 and 54 years old) and in oldest women and men (over 70 years old), with higher differences as age increases. As for the other segmentation variables, the largest differences were found among people with a nationality other than Spanish, especially among men. These results, although to a lesser extent, are also observed in the previous

FIGURE 1.

Temporal scope, response rates (RR) and effective sample size for each measurement in ESSOC

Description of the ESSOCgeneral measurements (overlapping panel design) and evolution of the SARS-COV-2 pandemic in Andalusia during 2020 and 2021

m: Effective total, cross-sectional and longitudinal samples for each measurement
RR: Response rates of each total, cross-sectional and longitudinal sample in the corresponding measurement (calculations are based on the refusals)
AIDT: Active Infection Diagnostic Tests (Source: Andalusian Institute of Statistics and Cartography)
s(t): sample of the measurement t

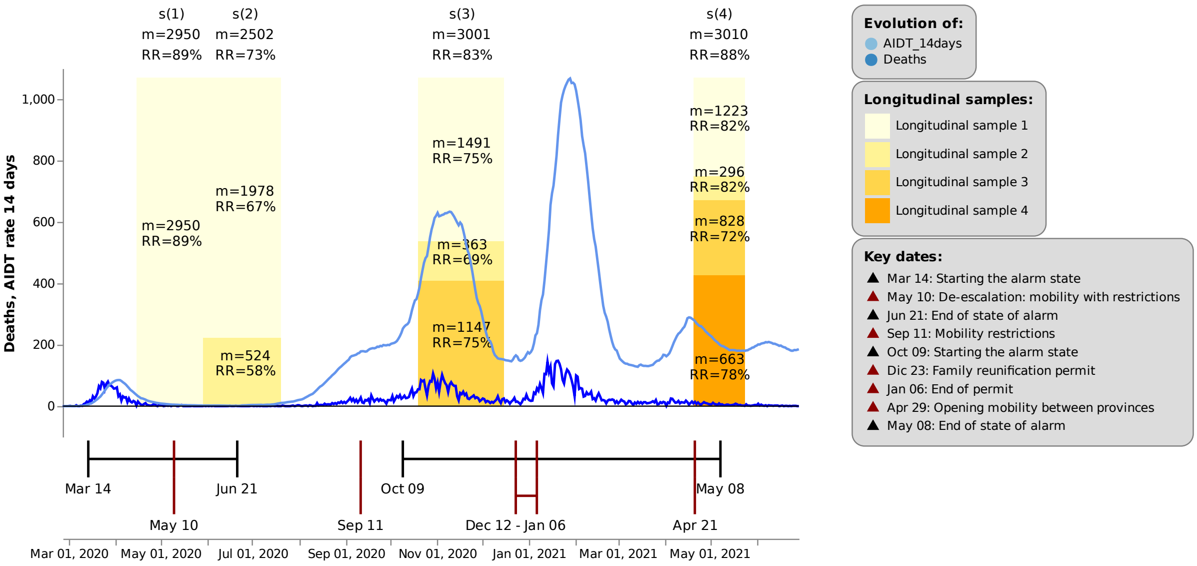
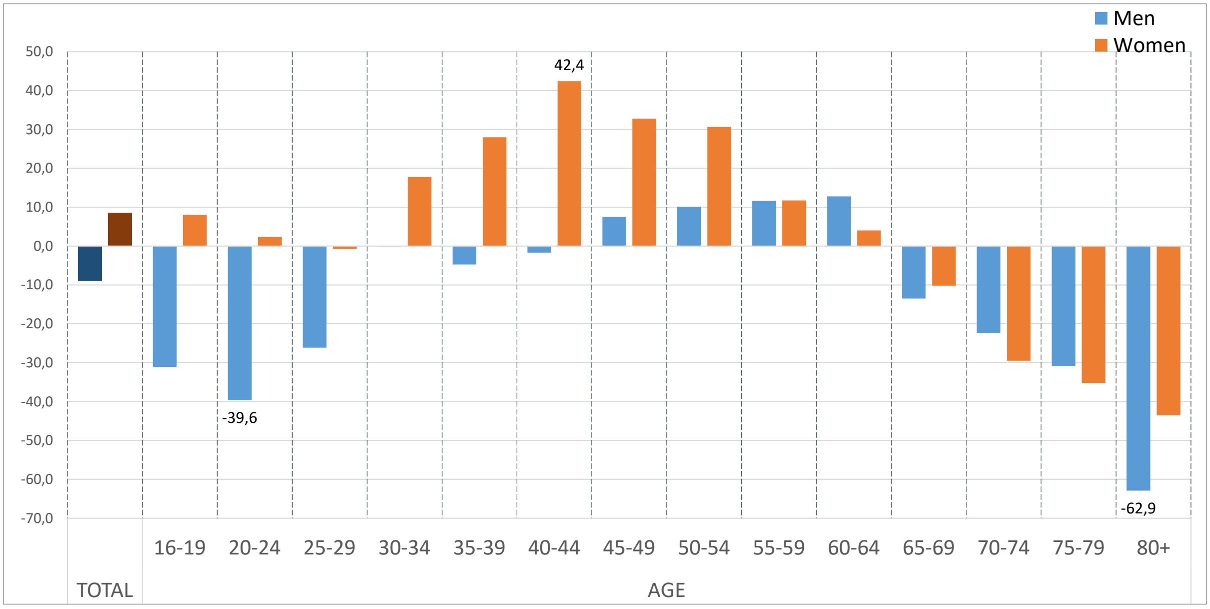


Figure 2.

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



measurements, showing a lower participation of those population groups in the ESSOC and therefore justify the need to adjust the sample weights.

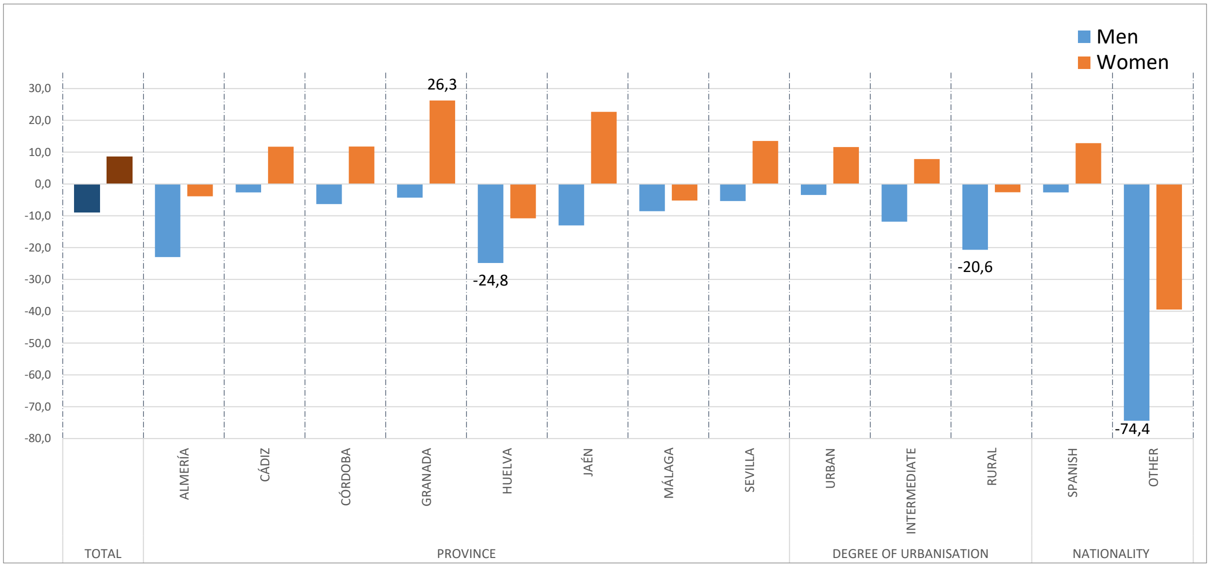
3. Sampling setup in overlapping panels.

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

On the first measurement $(t = 1)$ a sample $s^{(1)}$ of size $n^{(1)}$ is selected from the population U by stratified random sampling. Let h be the stratum to which unit i belongs, $(h = 1, ...L)$ and $s_h^{(1)}$ be the sample corresponding to stratum h on occasion 1. There is a total lack of response in the sample $s^{(1)}$ which is divided into

Figure 3.

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



$$s_{rh}^{(1)} = \{i \;\in\; s^{(1)}/\text{respond in stratum h }\}$$

$$s_{fh}^{(1)} = \{i \;\in\; s^{(1)}/\text{missing in stratum h }\}.$$

Let $m_h^{(1)}$ denote the number of the observations obtained form the $n_h^{(1)}$ sampled units, that is $\sum_h m_h^{(1)}$ is the size of $s_r^{(1)}$.

In each of the following measurements $t = 2, 3, ..., k$ we denote by $s_{rh}^{(t)}$ the sample of respondents in measurement t in stratum h of the original sample $s^{(1)}$. The size of which we denote by $m_h^{(t)}$. To complete the sample, a new sample $s_{new}^{(t)}$ is selected from the population U by stratified sampling independently of the sample $s^{(1)}$. The Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym) was responsible for the population framework and the extraction of the samples according to the ESSOC design described in section 2. For each extraction, the IECA verified that the samples $s_{new}^{(t)}$ and the sample $s^{(t)_r}$ had an empty intersection. Let $n_{hnew}^{(t)}$ be the size of the sample $s_{new}^{(t)}$ in stratum h and denote by $m_{hnew}^{(t)}$ the size of the sample of respondents in this stratum, $s_{rhnew}^{(t)}$. Thus, the total sample of respondents in each stratum and measurement would be $m_{htotal}^{(t)} = m_h^{(t)} + m_{hnew}^{(t)}$.

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

$$\hat{Y}_{ht}^{(1)} = \sum_h \sum_{i \in s_{rh}^{(1)}} d_{ih} y_{ih}^{(1)} . \tag{1}$$

This estimator is a naive estimator. In the case of stratified simple random sampling design for unit i belonging to stratum h is $d_{ih} = \frac{N_h}{n_h^{(1)}}$.

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 occasion 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 in stratum h is evaluated as $r_h = \frac{m_h^{(1)}}{n_h^{(1)}}$.

Then the initial weight of unit i in stratum h d_{ih} is replaced with the new weight $d_{ih}^{(1)} = \frac{d_{ih}}{r_h}$ and the estimator is given by

$$\hat{Y}^{(1)} = \sum_{h} \sum_{i \in s_{rh}^{(1)}} d_{ih}^{(1)} y_{ih}^{(1)} \,. \tag{2}$$

For the following measurements, different estimators can be obtained from the new sample obtained in each measurement and from the longitudinal samples of the previous measurements. The process to obtain them is shown in the next sections.

To fix the notation, we will call cross-sectional estimator of a parameter θ at time t to those estimators that are obtained from the cross-sectional sample $s_{cross}^{(t)} = s_r^{(t)} \cup s_{rnew}^{(t)}$, while we will call longitudinal estimators those obtained from the longitudinal sample $s_r^{(t)}$, where $s_r^{(t)} = \cup_h s_{rh}^{(t)}$ and $s_{rnew}^{(t)} = \cup_h s_{rhnew}^{(t)}$.

4. Cross-sectional estimation

The objective of most cross-sectional surveys is to produce unbiased estimates of totals or means at a given time point, and, in the case of repeated surveys, to produce estimates of the net change that occurred in the population between two time points. In order to improve the cost-effectiveness of surveys, one can derive cross-sectional estimates from longitudinal survey data 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. For this we will use both the longitudinal sample and the refreshing sample. In this way, the sample with which we work always has a sample size close to 3000 and we manage to reduce the variance of the final estimator.

Point estimation of parameters of the cross-sectional population based on data from longitudinal surveys has been studied by[16] 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[17]. We will follow a methodology similar to that used in these works. We will elaborate a cross-sectional weighting scheme that includes a non-response adjustment, an optimal combination of the two samples, and a calibration for completing representativeness of the population at a given time. This proposal is described below.

4.1. A first adjustment based on homogeneous groups

A simple adjusted estimator accounting for initial non-response and attrition can be obtained by adjusting the basic weights of the H-T estimator by the fraction of non-response $\frac{n_{hnew}^{(t)}}{m_{hnew}^{(t)}}$ obtaining the total estimator for the new sample in measurement t:

$$\hat{Y}_n^{(t)} = \sum_h \sum_{s_{rhnew}^{(t)}} \frac{N_h}{n_{hnew}^{(t)}} \frac{n_{hnew}^{(t)}}{m_{hnew}^{(t)}} y_{ih}^{(t)} = \sum_h \sum_{s_{rhnew}^{(t)}} d_{ihn}^{(t)} y_{ih}^{(t)} \,. \tag{3}$$

Weighting within classes is a commonly used procedure for non-response cross-sectional and longitudinal weighting in panels[13].

In a similar way, from the sample $s_r^{(t)}$ we can estimate the total as:

$$\hat{Y}_r^{(t)} = \sum_h \sum_{i\in s_{rh}^{(t)}} \frac{N_h}{n_h^{(1)}} \frac{n_h^{(1)}}{m_h^{(t)}} y_{ih}^{(t)} = \sum_h \sum_{i\in s_{rh}^{(t)}} d_{ihr}^{(t)} y_{ih}^{(t)} \,. \tag{4}$$

Combining these estimators we considered the following estimator

$$\hat{Y}^{(t)} = \alpha_1 \hat{Y}_r^{(t)} + \alpha_2 \hat{Y}_n^{(t)} \,, \tag{5}$$

where α_1 and α_2 are nonnegative constants such that $\alpha_1 + \alpha_2 = 1$.

Next, we consider the problem of selection of the best coefficients.

We denote the values $V(\hat{Y}_r^{(t)})$, $V(\hat{Y}_n^{(t)})$ by V_1, V_2 respectively. Thus, the variance of $\hat{Y}^{(t)}$ is:

$$V(\hat{Y}^{(t)}) = \alpha_1^2 V_1 + (1-\alpha_1)^2 V_2 \tag{6}$$

and its minimum value is obtained for

$$\alpha_1 = 1 - \alpha_2 = \frac{V_2}{V_1 + V_2}\,.$$

However, the values V_1 and V_2 are unknown. One possibility is to estimate them from the sample and substitute them in the previous expression to calculate the coefficients α but that does not ensure their optimality. A simple solution is to weight each estimator by the weight that sample has in the total sample available at the time t. In this way we consider the self-weighted total estimator

$$\hat{Y}_{sw}^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} \frac{m_h^{(t)}}{m_h^{(t)} + m_{hnew}^{(t)}} \frac{N_h}{m_h^{(t)}} y_{ih}^{(t)} + \sum_h \sum_{i \in s_{rhnew}^{(t)}} \frac{m_{hnew}^{(t)}}{m_h^{(t)} + m_{hnew}^{(t)}} \frac{N_h}{m_{hnew}^{(t)}} y_{ih}^{(t)}$$

$$= \sum_h \frac{N_h}{m_h^{(t)} + m_{hnew}^{(t)}} (\sum_{i \in s_{rh}^{(t)}} y_{ih}^{(t)} + \sum_{i \in s_{rhnew}^{(t)}} y_{ih}^{(t)}) = \sum_h \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} d_{ich}^{(t)} y_{ih}^{(t)}\,. \tag{7}$$

The weights $d_{ich}^{(t)}$ are the same for all units included in stratum h, so the sample within each stratum is self-weighted. Self-weighted estimators have the advantage of being easy to calculate and allow the application of more advanced inference techniques that are usually designed for simple random samples where the units have the same probability of selection.

4.2. Weight adjustment based on propensities

The adjustment based on weighting within classes assumes that unit non-response may be modeled 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 may be well explained by the strata defined at baseline.

An alternative is to use a regression-based approach. When many auxiliary variables are available, this approach is preferable to the previous one[13]. For this we are going to use the popular Propensity Score Adjustment (PSA) method[18,19,20] to model the probability that a unit of the sample $s_r^{(1)}$ responds on occasion t.

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

$$\pi_k^{(t)} = P(\delta_k^{(t)} = 1|y_k, \mathbf{x}_k) = P(\delta_k^{(t)} = 1|\mathbf{x}_k); k \in s_r^{(t)} . \tag{8}$$

We also assume that the mechanism follows a parametric model:

$$P(\delta_k^{(t)} = 1|y_k, \mathbf{x}_k) = m_t(\mathbf{x}_k, \lambda_t) \tag{9}$$

for some known function $m_t(\cdot)$ with second continuous derivatives with respect to an unknown parameter λ_t. A commonly adopted parametric model is the logistic regression model[21,22].

We use a state-of-the-art machine learning method: XGBoost[23] for estimating $\pi_k^{(t)}$. This technique builds decision trees ensembles which optimize an objective function via Gradient Tree Bosting[24]. More details can be found in Annex 2.[7] has shown the effectiveness of this technique when studying nonresponse in the GSOEP panel.[20] showed that Gradient Tree Bosting can lead to selection bias reductions in situations of high dimensionality or where the selection mechanism is Missing At Random (MAR).[25,26,27,28,29,30,31] have applied boosting algorithms in propensity score weighting showing better results than conventional parametric models.

In order to obtain the estimated propensities $\hat{\pi}_k^{(t)}$, we train a model with $s_r^{(1)}$ where $\mathbf{x}_k$ includes every available variable observed in $s_r^{(1)}$. Said model minimizes the weighted logistic loss for $\delta_k^{(t)}; k \in s_r^{(1)}$.

Since the values we are interested in, $\hat{\pi}_k^{(t)}$ for $k \in s_r^{(t)}$, are a subset of the values used for training, $\delta_k^{(t)}$ for $k \in s_r^{(1)}$, overfitting is likely to happen. This means we will obtain values extremely close to 1 instead of real propensities. Hyperparameter optimization is essential in order to avoid this issue.

Then we use the inverse of the estimated response propensity $\hat{\pi}_k^{(t)}$ as weight for constructing the estimator based on the sample $s_r^{(t)}$:

$$\hat{Y}_P^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ih}^{(1)} \frac{1}{\hat{\pi}_{ih}^{(t)}} y_{ih}^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ihPSA}^{(t)} y_{ih}^{(t)} \,. \tag{10}$$

Once the estimator $\hat{Y}_P^{(t)}$ is defined, we can considered the following cross-sectional estimator for the total:

$$\hat{Y}^{(t)}_{PSA} = \alpha_1 \hat{Y}^{(t)}_{P} + \alpha_2 \hat{Y}^{(t)}_{n} \,, \tag{11}$$

in a similar way as in the previous subsection.

4.3. Calibration on population totals

Further improvement in the representativeness of the sample can be made through its calibration against more reliable external information. Such calibration can reduce biases in the sample due to non-response, non-coverage and other distortions, and also reduce variances. Besides the modification of weights for handling non-response, weights adjustment may also be carried out to take into account auxiliary information. Calibration[32] is the most used technique for weights adjustment and can have the aim to insure consistency among estimates of different sample surveys, can reduce biases in the sample due to non-response, non-coverage and other distortions, and also reduce variances[33,34,35,36].

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

We denote by

$$\hat{Y}^{(t)}_{cross} = \sum_h \sum_{i \in s^{(t)}_{rh} \cup s^{(t)}_{rhnew}} D^{(t)}_{ih} y^{(t)}_{ih}$$

any of the cross-sectional estimators obtained using any of the previous adjustment methods.

The calibration total estimator is obtained as:

$$\hat{Y}_{\mathrm{CAL}}^{(t)} = \sum_{h} \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} y_{ih}^{(t)}, \tag{12}$$

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

$$\min_{\omega_k} \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} G\left(w_{ih}^{(t)}, D_{ih}^{(t)}\right) \tag{13}$$

fulfilling the calibration condition

$$\sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} \mathbf{x}_{ih}^{*(t)} = \sum_{\mathcal{U}_h} \mathbf{x}_{ih}^{*(t)} \tag{14}$$

for all stratum h.

4.4. Estimating changes and gender gaps

A parameter of interest is the absolute change from one measurement to the first measurement of the variable and we denote by $\theta^{(t)} = Y^{(t)} - Y^{(1)}$ this parameter. Variations over time are measured more accurately with overlapping samples with respect to the case where samples on different occasions do not overlap (see[37]). An estimator of this parameter for measurement t based on the previous calibration total estimators can be obtained as follows:

$$\hat{\theta}_{abs}^{(t)} = \hat{Y}_{\mathrm{CAL}}^{(t)} - \hat{Y}_{\mathrm{CAL}}^{(1)}. \tag{15}$$

Other parameter of interest in panel surveys is the relative change $\theta_{rel}^{(t)} = \frac{Y^{(t)} - Y^{(1)}}{Y^{(1)}}$ between measurement 1 and measurement t, which is estimated as:

$$\hat{\theta}^{(t)}_{rel} = \frac{\hat{\theta}^{(t)}_{abs}}{\hat{Y}^{(1)}_{\text{CAL}}} \,. \tag{16}$$

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.

The impact of the COVID-19 in the social determinants of health might have been widely different between genders. For this reason, it is of great interest to define the estimators of the gender gaps observed in the absolute and relative changes defined in previous sections, both in absolute and relative terms as well, in order to observe if the changes were significantly larger among people of a given gender in comparison to their counterpart.

Let $Gen = \{M, W\}$ be the variable measured in $s^{(t)}, t = 1, 2, 3, ..., k$ which reflects whether a respondent is a man (M) or a woman (W). We define the two indicator variables: $I^{M}_{ih} = 1$ if the unit i in stratum h is a man and 0 elsewhere, and I^{W}_{ih} in a similar way.

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

$$\begin{gathered} G\hat{G}abs^{(t)}_{abs} = \hat{\theta}^{(t)}_{W} - \hat{\theta}^{(t)}_{M} = \\ = (\textstyle\sum_h \sum_{i \in s^{(t)}_{rh} \cup s^{(t)}_{rhnew}} w^{(t)}_{ih} y^{(t)}_{ih} I^{W}_{ih} - \sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^{W}_{ih}) - \\ (\textstyle\sum_h \sum_{i \in s^{(t)}_{rh} \cup s^{(t)}_{rhnew}} w^{(t)}_{ih} y^{(t)}_{ih} I^{M}_{ih} - \sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^{M}_{ih}) \,. \end{gathered} \tag{17}$$

The estimator $G\hat{G}abs_{abs}^{(t)}$ is defined as the linear combination of two estimators in certain domains, hence its theoretical properties can be easily derived (see section 5.4 in[37]). This estimator is the most simple one can build on the gender gap and can tell the difference in the absolute change between men and women between measurement t and measurement 1. However, this estimator is subject to the base rate on each variable. For this reason, we define the relative gender gap estimator in the absolute change as follows:

$$G\hat{G}abs_{rel}^{(t)} = \frac{G\hat{G}abs_{abs}^{(t)}}{\hat{\theta}_M^{(t)}} = \frac{\hat{\theta}_W^{(t)} - \hat{\theta}_M^{(t)}}{\hat{\theta}_M^{(t)}}\,. \tag{18}$$

The estimator $G\hat{G}abs_{rel}^{(t)}$ allows us to observe the gender gap in the growth between measurement 1 and measurement t taking into account the base rate of the given target variable, but its theoretical properties are more difficult to develop as it is a nonlinear combination of two estimators from non-overlapping samples.

We define the absolute gender gap in the relative change as follows:

$$G\hat{G}rel_{abs}^{(t)} = \hat{\theta}_{relW}^{(t)} - \hat{\theta}_{relM}^{(t)} = \frac{\hat{\theta}_W^{(t)}}{\sum_h \sum_{i\in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^W} - \frac{\hat{\theta}_M^{(t)}}{\sum_h \sum_{i\in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^M}\,. \tag{19}$$

The estimator $G\hat{G}rel_{abs}^{(t)}$ allows us to observe the difference in percentage points in the relative growth of a given variable between women and men. We define the relative gender gap in the relative change as follows:

$$G\hat{G}rel_{rel}^{(t)} = \frac{G\hat{G}rel_{abs}^{(t)}}{\hat{\theta}_{relM}^{(t)}}\,. \tag{20}$$

Thus, for the study variables of each ESSOC measurement, we start from the H-T estimator (1) that is adjusted for non-response (10), combined from the cross-sectional and longitudinal samples (11) and, finally, calibrated to increase the representativeness of the sample (12). This estimator serves as the basis for calculating the absolute (15) and relative (16) change estimators between measurement t and 1, which are also used to obtain the different estimators to measure the absolute and relative gender gap in the absolute and relative changes of a measurement with respect to the first (17 and 18, and 19 and 20, respectively).

5. Longitudinal estimation

The primary objective of panel surveys is the production of longitudinal data series that 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 the longitudinal sample, since they reflect the variations of the variable in each individual and allow estimating additional parameters such as the number of population individuals whose value of y increases, decreases or remains the same between $t-1$ and t. 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 estimated propensities for each unit i of sample $s_{rh}^{(t)}$, $\hat{\pi}_{ih}^{(t)}$, are used to reweighting for nonresponse, and we define an estimator for $\theta^{(t)} = \sum_U (y_i^{(t)} - y_i^{(1)})$ from the longitudinal sample of respondents on occasion t by:

$$\hat{\theta}_l^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ih}^{(1)} \frac{1}{\hat{\pi}_{ih}^{(t)}} (y_{ih}^{(t)} - y_{ih}^{(1)}) = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ihPSA}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \,. \tag{21}$$

If updated totals are available then calibration to new totals can reduce presence of bias. Thus, in the next phase, calibration is applied to change the weights. So we get some weights $v_{ih}^{(t)}$, minimizing:

$$\sum_{i \in s_{rh}^{(t)}} G\left(v_{ih}^{(t)}, d_{ihPSA}^{(t)}\right) \tag{22}$$

subject to

$$\sum_{i \in s_{rh}^{(t)}} v_{ih}^{t} \mathbf{x}_{ih}^{*(t)} = \sum_{\mathcal{U}_h} \mathbf{x}_{ih}^{*(t)} \tag{23}$$

for all stratum h. The final calibrate estimator is given by

$$\hat{\theta}_c^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \,. \tag{24}$$

The researcher may also be interested in estimating the change from occasion t to the occasion $t-1$, $\theta^{(t,t-1)} = \sum_U y_i^{(t)} - y_i^{(t-1)}$. In this situation the estimator is calculated in the same way but modeling the non-response with respect to the sample obtained in the previous occasion, that is, we estimate the new propensities:

$$P(\delta_k^{(t,t-1)} = 1 | y_k, \mathbf{x}_k) = g_t(\mathbf{x}_k) \,, \tag{25}$$

being $\delta_k^{(t,t-1)} = 1$ if $k \in s_r^{(t)}$ and $\delta_k^{(t)} = 0$ if $k \in s_r^{(t-1)} - s_r^{(t)}$.

The estimated propensities for each unit i of sample $s_{rh}^{(t)}$, $\hat{\pi}_{ih}^{(t,t-1)}$, are used in the first stage to reweighting for adjusting the nonresponse, and in the second stage, calibration is applied to reweight these weights and obtain new ones, $v_{ih}^{(t,t-1)}$, so as to obtain better representativeness of the population. The longitudinal estimator of $\theta^{(t,t-1)}$ can be defined as follows:

$$\hat{\theta}_c^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)})\,. \tag{26}$$

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 $t-1$ and t. 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 $y^{(t)} - y^{(t-1)} \in A$ can be estimated as follows:

$$\hat{\theta}_{cA}^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A, I_A = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \in A \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \notin A \end{cases}. \tag{27}$$

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. If the variable y measures health status, this rate can be considered a deterioration/improvement rate, $DIRate$. The formula can be defined as follows:

$$\widehat{DIRate}_c^{(t,t-1)} = \frac{\hat{\theta}_{cA_{R-}}^{(t,t-1)} - \hat{\theta}_{cA_{R+}}^{(t,t-1)}}{\hat{\theta}_{cA_{R+}}^{(t,t-1)}} = \frac{\sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R-}} - \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R+}}}{\sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R+}}}, \tag{28}$$

where

$$I_{A_{R+}} = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} > 0 \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \leq 0 \end{cases}$$

and

$$I_{A_{R-}} = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} < 0 \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \geq 0 \end{cases}$$

Based on previous estimators, estimators of the gender gap of the change between $t-1$ and t can be defined as follows:

$$GG\hat{l}ong_{abs}^{(t)} = \hat{\theta}_{cW}^{(t,t-1)} - \hat{\theta}_{cM}^{(t,t-1)} = \sum_h \sum_{\imath \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) I_{ih}^W - \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) I_{ih}^M\,, \tag{29}$$

$$GG\hat{l}ong_{absA}^{(t)} = \hat{\theta}_{cAW}^{(t,t-1)} - \hat{\theta}_{cAM}^{(t,t-1)} = \sum_h \sum_{\imath \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A I_{ih}^W - \sum_h \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A I_{ih}^M\,, \tag{30}$$

$$GG\hat{l}ong_{rel}^{(t)} = \frac{GG\hat{l}ong_{abs}^{(t)}}{\hat{\theta}_{cM}^{(t,t-1)}}\,, \tag{31}$$

$$GG\hat{l}ong_{relA}^{(t)} = \frac{GG\hat{l}ong_{absA}^{(t)}}{\hat{\theta}_{cAM}^{(t,t-1)}}\,. \tag{32}$$

5.1. Variance estimation

The development of suitable variance estimators for these proposed estimators taking into account the panel design used is not a simple task. The variance estimation problem in longitudinal surveys is addressed in several papers. For example,[17] considers variance estimation for Canada's Survey of Labour and Income Dynamics within a Taylor linearization approach and a bootstrap method.

Some other works are developed for rotation panels:[12] 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.[38] consider variance estimation for weighting in the SILC survey in France with a rotation scheme consisting of four panels.[21] consider the case of a panel survey in which the sole 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.[39] also consider the estimation of a mean for a panel survey, in case of monotone nonresponse.

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[40].

In this study, the formulas used for estimating the variance of indicators take account the structure and complexity of the ESSOC survey. The main factors taken into account in estimating the variance of the proposed estimators are the non-linearity of the estimators, total non-response at different survey stages and calibration. We introduce how to get the variance estimators accordingly in the Annex 1.

6. Application to the Health Care and Social Survey (ESSOC)

6.1. Calibrating the representativeness of the sample

As shown in section 2 and in Figures 2 and 3, the ESSOC has a non-monotone missing pattern and shows a lower participation of some population groups. The first measurement was carried out by the 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 adjustment as the IECA for the sample weights of the new samples of each ESSOC measurement, i.e. truncated linear calibration with 0.1 to 10 limits and the total population size for the cross of the sex variable with province, age, urbanization grades and nationality as auxiliary information. The data for said totals are obtained from the 2019 Municipal Register of Inhabitants[41].

6.2. Modeling the non-response

Starting from the data and calibrated weights provided by the IECA, we proceeded to perform a readjustment using propensities.

The non-response at $s_r^{(t)}$, $\pi_k^{(t)}$ for $k \in s_r^{(t)}$, is modeled with PSA considering every variable of $s_r^{(1)}$. In order to ensure that the XGBoost model is learning properly, we have considered the following hyperparameters:

- Number of estimators $\in [10, 400]$: The number of trees forming the ensemble.

- Learning rate $\in [0.01, 1]$: the weight shrinkage applied after each boosting step.
- Maximum depth $\in [1, 60]$: The maximum number of splits that each tree can contain.
- Minimum child weight $\in [1, 6]$: The minimum total of instance weights needed to consider a new partition.

The accuracy of the algorithm is 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 is calculated for its corresponding remaining subset. The mean logistic loss is the estimated error.

The values for the hyperparameters minimizing this estimated error are obtained using the Tree-structured Parzen Estimator (TPE) algorithm[42,43]. TPE is implemented in Optuna[44], an optimization library for Python, as its default method.

The cross-sectional and longitudinal estimated are calculated by using these PSA weights.

6.3. Cross-sectional estimators

Table 1 shows, for measurement 4, the percentages with corresponding confidence intervals at 95% as well as 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

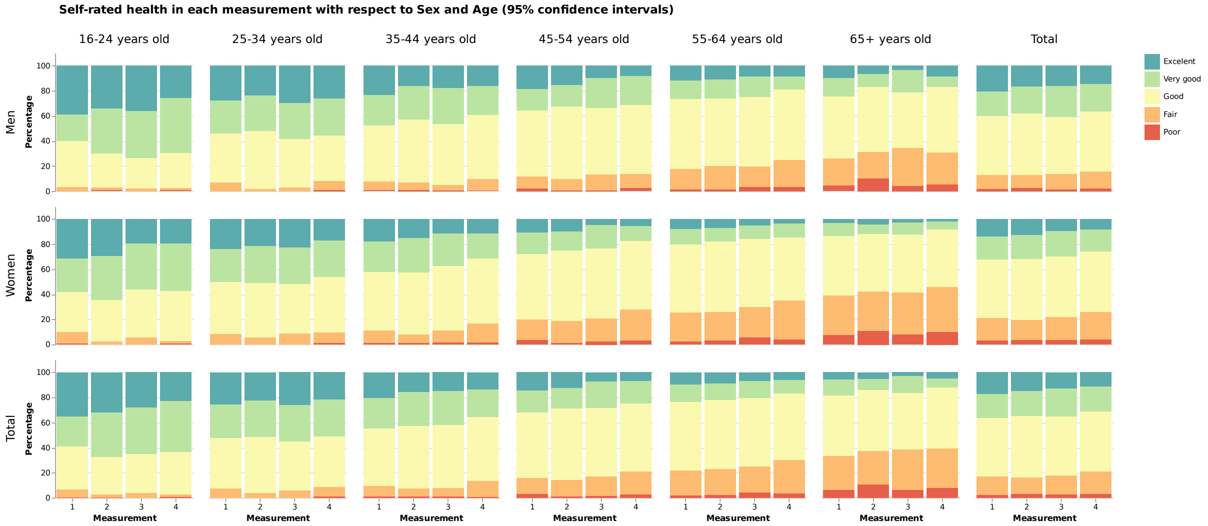
TABLE 1.

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

Age group (years)	Self-perceived general health	Total					Men					Women				
		sample			Confidence Interval 95%		sample			Confidence Interval 95%		sample			Confidence Interval 95%	
		size	Population	Percentage	lower limit	upper limit	size	Population	Percentage	lower limit	upper limit	size	Population	Percentage	lower limit	upper limit
Total	Excelent	339	808765	11.5%	10.2%	12.8%	184	502606	14.6%	12.4%	16.7%	155	306160	8.5%	7.2%	9.9%
	Very good	586	1378519	19.6%	18.0%	21.1%	287	771297	22.3%	19.8%	24.9%	299	607222	16.9%	15.0%	18.7%
	Good	1499	3388935	48.1%	46.1%	50.0%	671	1636596	47.4%	44.4%	50.4%	828	1752339	48.7%	46.1%	51.3%
	Fair	501	1259015	17.9%	16.2%	19.5%	166	461496	13.4%	11.0%	15.7%	335	797519	22.2%	19.9%	24.4%
	Bad	80	214926	3.1%	2.3%	3.8%	31	79900	2.3%	1.4%	3.2%	49	135026	3.8%	2.6%	4.9%
	Total	3005	7050160	100%			1339	3451896	100%			1666	3598265	100%		
16-24	Excelent	65	179749	22.6%	17.5%	27.7%	29	103885	25.2%	17.1%	33.3%	36	75864	19.7%	13.7%	25.7%
	Very good	117	326469	41.0%	35.0%	47.0%	51	183790	44.6%	35.3%	53.8%	66	142679	37.1%	29.7%	44.5%
	Good	99	268791	33.7%	28.0%	39.5%	31	112763	27.3%	19.0%	35.7%	68	156028	40.6%	33.0%	48.2%
	Fair	6	17120	2.2%	0.3%	4.0%	2	9326	2.3%	-0.8%	5.4%	4	7794	2.0%	0.0%	4.1%
	Bad	2	4773	0.6%	-0.2%	1.4%	1	2680	0.7%	-0.6%	1.9%	1	2094	0.5%	-0.5%	1.6%
	Total	289	796903	100%			114	412444	100%			175	384459	100%		
25-34	Excelent	91	223549	21.9%	17.4%	26.3%	49	138750	26.0%	18.9%	33.2%	42	84799	17.4%	12.4%	22.3%
	Very good	116	298896	29.3%	24.4%	34.1%	53	156852	29.4%	22.1%	36.7%	63	142043	29.1%	22.8%	35.4%
	Good	177	413081	40.5%	35.3%	45.6%	71	196425	36.9%	29.2%	44.6%	106	216656	44.4%	37.7%	51.0%
	Fair	33	77329	7.6%	5.0%	10.2%	15	36091	6.8%	3.3%	10.3%	18	41238	8.5%	4.6%	12.3%
	Bad	3	8319	0.8%	-0.1%	1.8%	2	4853	0.9%	-0.4%	2.2%	1	3465	0.7%	-0.7%	2.1%
	Total	420	1021173	100%			190	532972	100%			230	488201	100%		
35-44	Excelent	85	182116	13.6%	10.5%	16.7%	42	107783	16.0%	10.8%	21.2%	43	74333	11.2%	7.9%	14.5%
	Very good	145	292297	21.8%	18.4%	25.3%	66	155939	23.1%	17.6%	28.6%	79	136359	20.6%	16.4%	24.7%
	Good	338	687717	51.4%	47.1%	55.6%	142	347035	51.4%	44.7%	58.1%	196	340682	51.4%	46.2%	56.6%
	Fair	79	166859	12.5%	9.7%	15.3%	24	62907	9.3%	5.3%	13.3%	55	103952	15.7%	11.8%	19.6%
	Bad	6	9684	0.7%	0.1%	1.3%	1	2112	0.3%	-0.3%	0.9%	5	7572	1.1%	0.1%	2.2%
	Total	653	1338673	100%			275	675775	100%			378	662898	100%		
45-54	Excelent	48	91484	6.8%	4.9%	8.8%	28	57292	8.5%	5.4%	11.6%	20	34193	5.1%	2.9%	7.4%
	Very good	117	237169	17.7%	14.5%	20.9%	70	159215	23.6%	18.2%	29.0%	47	77955	11.7%	8.5%	14.9%
	Good	377	730633	54.5%	50.5%	58.5%	170	362979	53.8%	47.7%	59.9%	207	367654	55.2%	50.1%	60.4%
	Fair	126	245229	18.3%	15.2%	21.4%	35	79231	11.7%	7.8%	15.7%	91	165998	24.9%	20.4%	29.5%
	Bad	19	36081	2.7%	1.5%	3.9%	8	15904	2.4%	0.7%	4.0%	11	20177	3.0%	1.2%	4.8%
	Total	687	1340596	100%			311	674620	100%			376	665976	100%		
55-64	Excelent	30	66397	6.1%	3.9%	8.4%	21	45756	8.7%	4.9%	12.6%	9	20641	3.7%	1.3%	6.2%
	Very good	58	117423	10.9%	8.2%	13.6%	29	57481	10.9%	7.1%	14.8%	29	59942	10.8%	7.0%	14.6%
	Good	281	569480	52.7%	48.3%	57.1%	149	289687	55.1%	48.8%	61.5%	132	279793	50.4%	44.2%	56.6%
	Fair	134	288216	26.7%	22.7%	30.7%	52	114366	21.8%	16.2%	27.3%	82	173850	31.3%	25.6%	37.1%
	Bad	19	39277	3.6%	2.0%	5.3%	9	18377	3.5%	1.2%	5.8%	10	20900	3.8%	1.4%	6.1%
	Total	522	1080793	100%			260	525667	100%			262	555126	100%		
>=65	Excelent	20	65470	4.5%	2.3%	6.6%	15	49140	7.8%	3.5%	12.1%	5	16330	1.9%	0.3%	3.6%
	Very good	33	106265	7.2%	4.7%	9.7%	18	58021	9.2%	4.8%	13.6%	15	48244	5.7%	2.8%	8.6%
	Good	227	719234	48.9%	43.7%	54.0%	108	327708	52.0%	43.5%	60.4%	119	391527	46.5%	40.1%	53.0%
	Fair	123	464262	31.5%	26.4%	36.7%	38	159575	25.3%	16.6%	34.0%	85	304687	36.2%	29.9%	42.5%
	Bad	31	116792	7.9%	5.0%	10.8%	10	35974	5.7%	1.9%	9.6%	21	80818	9.6%	5.5%	13.7%
	Total	434	1472024	100%			189	630418	100%			245	841605	100%		

Figure 4.

Estimated percentages grouped by sex and age for the original categories of self-perceived general health at measurement 4

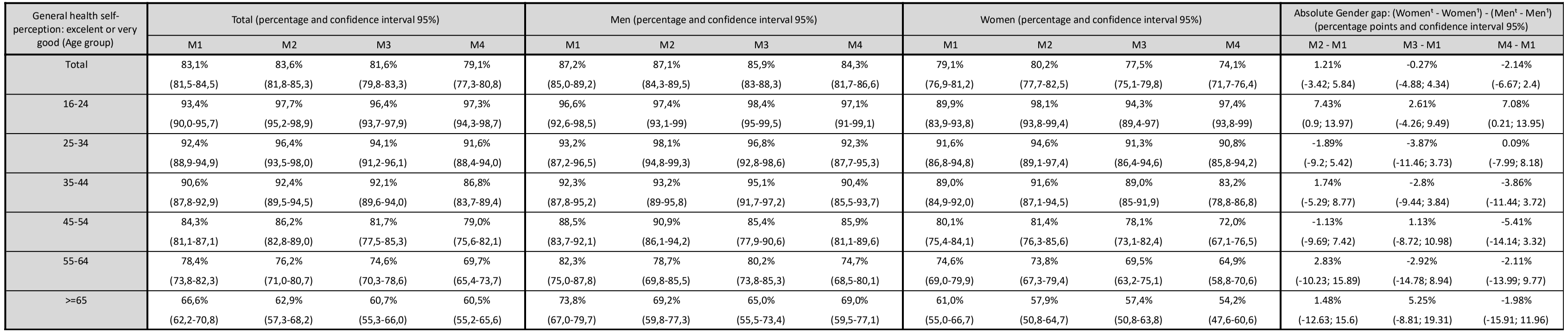


old, neither for men nor for women. However, the excellent or very good self-perceived health decreases for the population older than 35 years as the pandemic advances. This can be observed more as age increases, especially in women. This reduction results in an increment for the 'fair' and 'bad' categories. However, the 'good' general health category stays stable throughout the pandemic for each sex and age group.

Based on these results, we dichotomized that variable with the categories 'excellent and very good' and 'good, fair and bad'. Figure 4 shows the percentages and confidence intervals given in table 1 not only for measurement 4, but also for all other ESSOC measurements. Table 2 shows, for each measurement of the ESSOC, the percentages and confidence intervals at 95% of the dichotomized self-perceived general health variable as described in the previous paragraph. These results can be seen at figure 5,

Table 2.

Evolution in each measurement (M) of percentages, gender gaps and confidence intervals at 95% of people with excellent or very good self-perceived general health.



General health self-perception: excelent or very good (Age group)	Total (percentage and confidence interval 95%)				Men (percentage and confidence interval 95%)				Women (percentage and confidence interval 95%)				Absolute Gender gap: (Women^t - Women^1) - (Men^t - Men^1) (percentage points and confidence interval 95%)		
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M2 - M1	M3 - M1	M4 - M1
Total	83,1% (81,5-84,5)	83,6% (81,8-85,3)	81,6% (79,8-83,3)	79,1% (77,3-80,8)	87,2% (85,0-89,2)	87,1% (84,3-89,5)	85,9% (83-88,3)	84,3% (81,7-86,6)	79,1% (76,9-81,2)	80,2% (77,7-82,5)	77,5% (75,1-79,8)	74,1% (71,7-76,4)	1.21% (-3.42; 5.84)	-0.27% (-4.88; 4.34)	-2.14% (-6.67; 2.4)
16-24	93,4% (90,0-95,7)	97,7% (95,2-98,9)	96,4% (93,7-97,9)	97,3% (94,3-98,7)	96,6% (92,6-98,5)	97,4% (93,1-99)	98,4% (95-99,5)	97,1% (91-99,1)	89,9% (83,9-93,8)	98,1% (93,8-99,4)	94,3% (89,4-97)	97,4% (93,8-99)	7.43% (0.9; 13.97)	2.61% (-4.26; 9.49)	7.08% (0.21; 13.95)
25-34	92,4% (88,9-94,9)	96,4% (93,5-98,0)	94,1% (91,2-96,1)	91,6% (88,4-94,0)	93,2% (87,2-96,5)	98,1% (94,8-99,3)	96,8% (92,8-98,6)	92,3% (87,7-95,3)	91,6% (86,8-94,8)	94,6% (89,1-97,4)	91,3% (86,4-94,6)	90,8% (85,8-94,2)	-1.89% (-9.2; 5.42)	-3.87% (-11.46; 3.73)	0.09% (-7.99; 8.18)
35-44	90,6% (87,8-92,9)	92,4% (89,5-94,5)	92,1% (89,6-94,0)	86,8% (83,7-89,4)	92,3% (87,8-95,2)	93,2% (89-95,8)	95,1% (91,7-97,2)	90,4% (85,5-93,7)	89,0% (84,9-92,0)	91,6% (87,1-94,5)	89,0% (85-91,9)	83,2% (78,8-86,8)	1.74% (-5.29; 8.77)	-2.8% (-9.44; 3.84)	-3.86% (-11.44; 3.72)
45-54	84,3% (81,1-87,1)	86,2% (82,8-89,0)	81,7% (77,5-85,3)	79,0% (75,6-82,1)	88,5% (83,7-92,1)	90,9% (86,1-94,2)	85,4% (77,9-90,6)	85,9% (81,1-89,6)	80,1% (75,4-84,1)	81,4% (76,3-85,6)	78,1% (73,1-82,4)	72,0% (67,1-76,5)	-1.13% (-9.69; 7.42)	1.13% (-8.72; 10.98)	-5.41% (-14.14; 3.32)
55-64	78,4% (73,8-82,3)	76,2% (71,0-80,7)	74,6% (70,3-78,6)	69,7% (65,4-73,7)	82,3% (75,0-87,8)	78,7% (69,8-85,5)	80,2% (73,8-85,3)	74,7% (68,5-80,1)	74,6% (69,0-79,9)	73,8% (67,3-79,4)	69,5% (63,2-75,1)	64,9% (58,8-70,6)	2.83% (-10.23; 15.89)	-2.92% (-14.78; 8.94)	-2.11% (-13.99; 9.77)
>=65	66,6% (62,2-70,8)	62,9% (57,3-68,2)	60,7% (55,3-66,0)	60,5% (55,2-65,6)	73,8% (67,0-79,7)	69,2% (59,8-77,3)	65,0% (55,5-73,4)	69,0% (59,5-77,1)	61,0% (55,0-66,7)	57,9% (50,8-64,7)	57,4% (50,8-63,8)	54,2% (47,6-60,6)	1.48% (-12.63; 15.6)	5.25% (-8.81; 19.31)	-1.98% (-15.91; 11.96)

where it can be observed that the excellent and very good self-perceived health decreased in measurements 3 and 4, with the decrease being slightly larger among women than among men. Regarding age groups, the evolution has been stable throughout the pandemic since the lockdown for the population between 16 and 34 years old, for men as well as for women. However, for the population above 35 years old, the evolution worsens as the age increases and the pandemic advances, especially in women. Therefore, this subpopulation got the lowest 'excellent or very good' general health values at the beginning of the lockdown for every age group above 35 years old and, also, it was when the difference with respect to men was bigger.

Figure 5.

Evolution of percentages and confidence intervals at 95% level of people with excellent or very good self-perceived general health with respect to age and sex

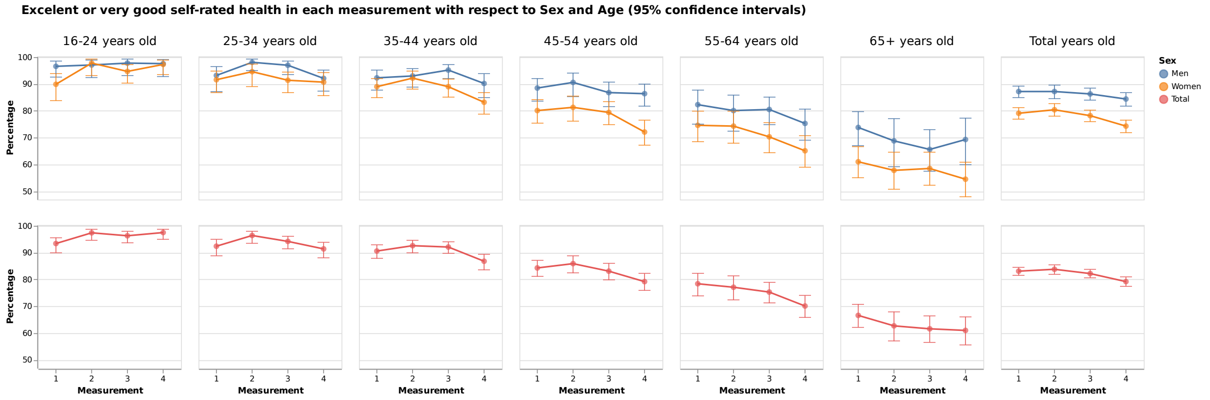
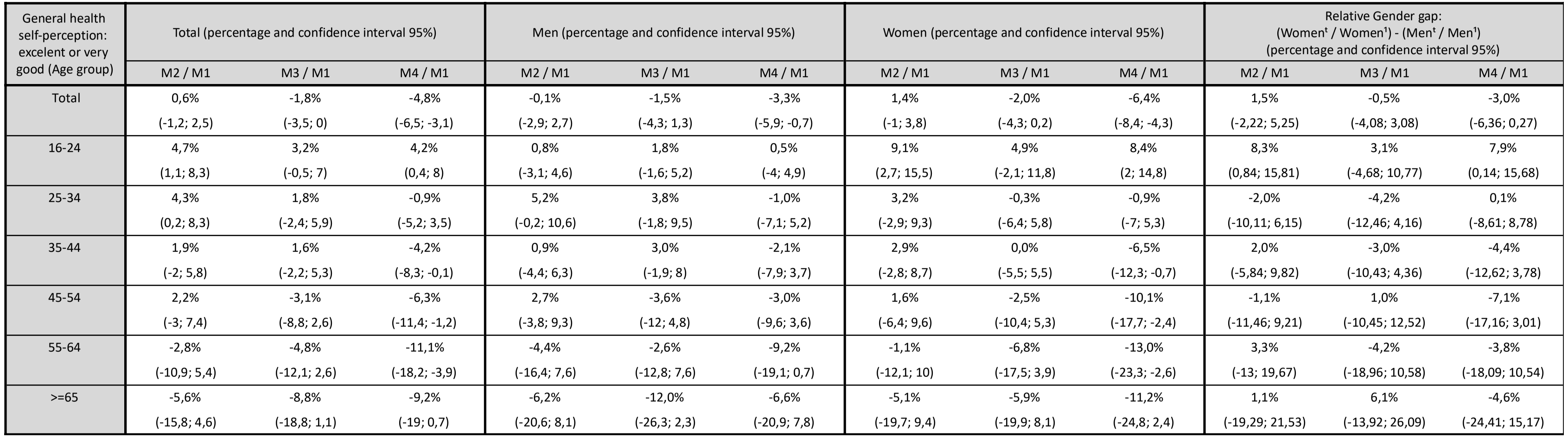


Table 3 and Figure 6 shows the relative percentage changes and 95% confidence intervals for each measurement with respect to measurement 1 for the 'excellent or very good' self-perceived general health variable. It could be observed that excellent or very good general health decreased in the general population by a 4.8% (CI95%=[-6.5;-3.1] in

Table 3.

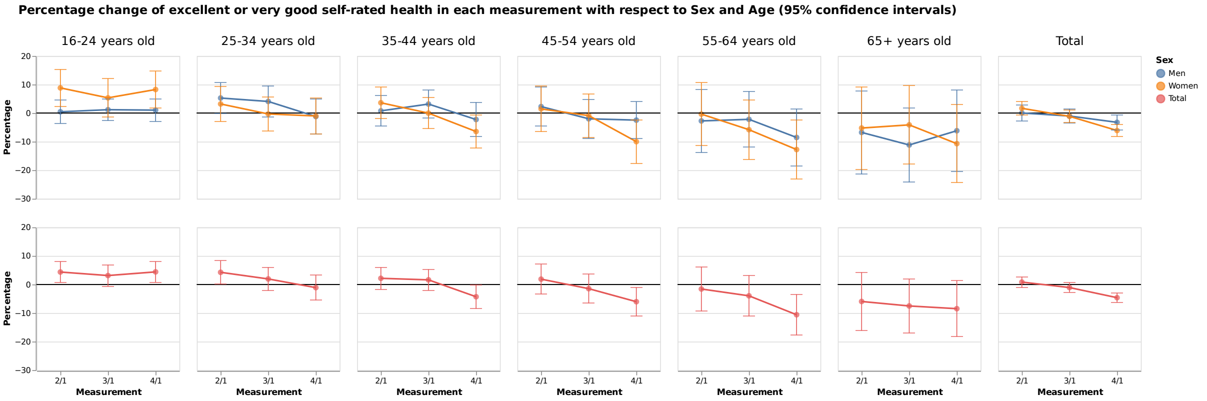
Evolution in each measurement (M) with respect to measurement 1 of relative percentage changes, gender gaps and 95% confidence intervals for people with excellent or very good self-perceived general health.



General health self-perception: excelent or very good (Age group)	Total (percentage and confidence interval 95%)			Men (percentage and confidence interval 95%)			Women (percentage and confidence interval 95%)			Relative Gender gap: (Women^t / Women^1) - (Men^t / Men^1) (percentage and confidence interval 95%)		
	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1
Total	0,6% (-1,2; 2,5)	-1,8% (-3,5; 0)	-4,8% (-6,5; -3,1)	-0,1% (-2,9; 2,7)	-1,5% (-4,3; 1,3)	-3,3% (-5,9; -0,7)	1,4% (-1; 3,8)	-2,0% (-4,3; 0,2)	-6,4% (-8,4; -4,3)	1,5% (-2,22; 5,25)	-0,5% (-4,08; 3,08)	-3,0% (-6,36; 0,27)
16-24	4,7% (1,1; 8,3)	3,2% (-0,5; 7)	4,2% (0,4; 8)	0,8% (-3,1; 4,6)	1,8% (-1,6; 5,2)	0,5% (-4; 4,9)	9,1% (2,7; 15,5)	4,9% (-2,1; 11,8)	8,4% (2; 14,8)	8,3% (0,84; 15,81)	3,1% (-4,68; 10,77)	7,9% (0,14; 15,68)
25-34	4,3% (0,2; 8,3)	1,8% (-2,4; 5,9)	-0,9% (-5,2; 3,5)	5,2% (-0,2; 10,6)	3,8% (-1,8; 9,5)	-1,0% (-7,1; 5,2)	3,2% (-2,9; 9,3)	-0,3% (-6,4; 5,8)	-0,9% (-7; 5,3)	-2,0% (-10,11; 6,15)	-4,2% (-12,46; 4,16)	0,1% (-8,61; 8,78)
35-44	1,9% (-2; 5,8)	1,6% (-2,2; 5,3)	-4,2% (-8,3; -0,1)	0,9% (-4,4; 6,3)	3,0% (-1,9; 8)	-2,1% (-7,9; 3,7)	2,9% (-2,8; 8,7)	0,0% (-5,5; 5,5)	-6,5% (-12,3; -0,7)	2,0% (-5,84; 9,82)	-3,0% (-10,43; 4,36)	-4,4% (-12,62; 3,78)
45-54	2,2% (-3; 7,4)	-3,1% (-8,8; 2,6)	-6,3% (-11,4; -1,2)	2,7% (-3,8; 9,3)	-3,6% (-12; 4,8)	-3,0% (-9,6; 3,6)	1,6% (-6,4; 9,6)	-2,5% (-10,4; 5,3)	-10,1% (-17,7; -2,4)	-1,1% (-11,46; 9,21)	1,0% (-10,45; 12,52)	-7,1% (-17,16; 3,01)
55-64	-2,8% (-10,9; 5,4)	-4,8% (-12,1; 2,6)	-11,1% (-18,2; -3,9)	-4,4% (-16,4; 7,6)	-2,6% (-12,8; 7,6)	-9,2% (-19,1; 0,7)	-1,1% (-12,1; 10)	-6,8% (-17,5; 3,9)	-13,0% (-23,3; -2,6)	3,3% (-13; 19,67)	-4,2% (-18,96; 10,58)	-3,8% (-18,09; 10,54)
>=65	-5,6% (-15,8; 4,6)	-8,8% (-18,8; 1,1)	-9,2% (-19; 0,7)	-6,2% (-20,6; 8,1)	-12,0% (-26,3; 2,3)	-6,6% (-20,9; 7,8)	-5,1% (-19,7; 9,4)	-5,9% (-19,9; 8,1)	-11,2% (-24,8; 2,4)	1,1% (-19,29; 21,53)	6,1% (-13,92; 26,09)	-4,6% (-24,41; 15,17)

Figure 6.

Evolution of relative percentage changes and 95% confidence intervals for people with excellent or very good self-perceived general health in each measurement (M) with respect to measurement 1.



measurement 4 with respect to measurement 1. In addition, the decrease among women was of double the decrease among men (-6.4 and -3.3% respectively) and this gap was larger in older people, specially for women, except for women between 16 and 24 years old, where an increase was observed in all measurements with respect to measurement 1. Tables 2 and 3 incorporate absolute gender gaps on absolute and relative changes respectively, i.e. the absolute difference (in percentage points) between men and women of absolute and relative changes of a given measurement with respect to measurement 1. Their interpretation would be that a positive value indicates that women showed a larger change (absolute or relative) in comparison to men in their 'excellent or very good' self-perceived general health. Therefore, this result could be seen as a positive gender gap (i.e., better result or favorable to women) in the corresponding measurement compared to the first one. On the contrary, a negative value indicates that women showed a smaller change (absolute or relative) in comparison to men in their 'excellent or very good' self-perceived general health, which could be seen as a negative (or unfavourable to women) gender gap. These results are visualized in Figure 7; it can be observed that, for example, the gender gap was negative along the pandemic, confirming an increasingly negative impact on women with respect to men in relation to self-perceived general health. The gender gap went from positive in measurement 2 with respect to measurement 1, by 1.5 percentage points, to negative and statistically significant in measurement 4 by -3 percentage points. Results by age reveal that negative gender gaps were observed in people over 35 years old in measurement 4 (with respect to measurement 1), with the group between 45 and 54 years old showing the largest gender gap, of 7 percentage points (relative change).

FIGURE 7.

Gender gap for the change on the excellent or very good self-perceived health from one measurement with respect to the first measurement

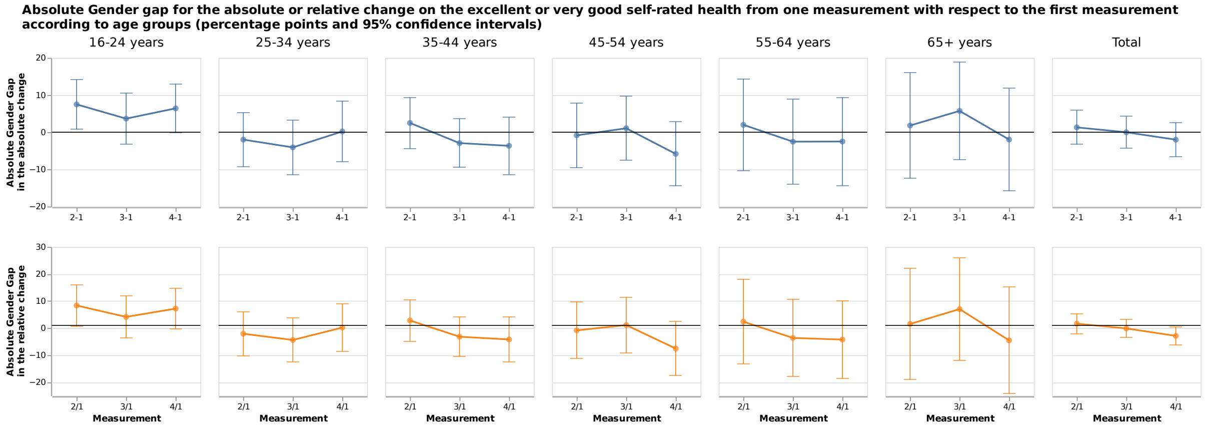


Table 4.

Percentage of people whose self-perceived general health improves, deteriorates or remains the same between measurements, and absolute and relative gender gap in the percentage

Age group (years)	Self-perceived general health	Total			Women			Men			Absolute gender gap (change in percentage points: Women - Men)			Relative gender gap (percentages change rate: Women / Men)		
		M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)
Total	Health improves	20,75% (18,75-22,76)	20,58% (18,36-22,8)	18,32% (16,56-20,07)	19,47% (16,44-22,51)	24,21% (20,59-27,82)	19,92% (17,1-22,73)	21,98% (19,35-24,61)	17,12% (14,54-19,7)	16,78% (14,66-18,9)	2.51 (-1.51; 6.53)	-7.09 (-11.53; -2.65)	-3.14 (-6.66; 0.38)	12.73% (-12.15; 37.61)	-29.39% (-46.52; -12.27)	-16.19% (-33.81; 1.44)
	Remains the same	54,96% (52,5-57,42)	54,04% (51,46-56,62)	56,97% (54,74-59,2)	55,21% (51,44-58,98)	53,24% (49,24-57,25)	56,31% (52,86-59,75)	54,72% (51,54-57,89)	54,80% (51,51-58,08)	57,61% (54,75-60,47)	-0.49 (-5.43; 4.44)	1.55 (-3.63; 6.73)	1.3 (-3.17; 5.78)	-1.03% (-14.19; 12.13)	2.75% (-11.26; 16.77)	1.79% (-10.12; 13.7)
	Health deteriorates	24,29% (22,16-26,42)	25,38% (23,15-27,61)	24,71% (22,74-26,68)	25,32% (22,03-28,61)	22,55% (19,22-25,88)	23,78% (20,75-26,81)	23,30% (20,57-26,03)	28,09% (25,12-31,06)	25,61% (23,07-28,15)	-2.02 (-6.29; 2.26)	5.54 (1.07; 10)	1.83 (-2.12; 5.79)	-8.08% (-26.6; 10.43)	24.37% (-1.67; 50.42)	7.16% (-12.97; 27.29)
16-24	Health improves	29,90% (23,88-35,93)	21,48% (15,6-27,37)	24,10% (18,41-29,8)	25,30% (17,35-33,24)	28,37% (18,86-37,88)	20,92% (12,43-29,41)	34,80% (25,83-43,77)	14,16% (7,76-20,55)	27,49% (20-34,98)	9.5 (-2.48; 21.49)	-14.22 (-25.67; -2.76)	6.58 (-4.75; 17.9)	37.57% (-29.22; 104.35)	-50.1% (-81.14; -19.06)	31.44% (-41.35; 104.22)
	Remains the same	44,26% (37,76-50,76)	53,92% (46,73-61,12)	57,08% (50,39-63,77)	45,83% (36,62-55,03)	53,79% (43,12-64,46)	59,93% (49,77-70,08)	42,60% (33,45-51,75)	54,06% (44,49-63,63)	54,05% (45,47-62,62)	-3.23 (-16.21; 9.75)	0.27 (-14.07; 14.6)	-5.88 (-19.17; 7.41)	-7.04% (-43.39; 29.3)	0.5% (-38.85; 39.85)	-9.81% (-42.8; 23.17)
	Health deteriorates	25,84% (20,09-31,59)	24,59% (18,34-30,85)	18,82% (13,64-24)	28,88% (20,44-37,32)	17,84% (9,55-26,12)	19,16% (11,24-27,07)	22,60% (14,92-30,28)	31,78% (22,66-40,91)	18,46% (11,86-25,06)	-6.27 (-17.69; 5.14)	13.95 (1.62; 26.28)	-0.7 (-11; 9.61)	-21.73% (-62.34; 18.88)	78.18% (-33.05; 189.42)	-3.64% (-61.56; 54.29)
25-34	Health improves	20,39% (14,01-26,77)	24,09% (17,85-30,34)	20,28% (15,3-25,26)	17,93% (7,67-28,2)	29,62% (19,52-39,71)	23,06% (14,85-31,26)	22,92% (15,32-30,52)	18,39% (11,46-25,33)	17,42% (12,01-22,83)	4.99 (-7.79; 17.76)	-11.22 (-23.47; 1.02)	-5.64 (-15.47; 4.19)	27.8% (-69.74; 125.35)	-37.9% (-74.51; -1.29)	-24.45% (-64.79; 15.9)
	Remains the same	55,39% (48,35-62,42)	48,07% (41,14-55)	54,24% (48,29-60,2)	58,48% (47,55-69,41)	43,25% (33,01-53,49)	52,19% (42,97-61,41)	52,19% (43,44-60,95)	53,03% (44,12-61,94)	56,36% (48,92-63,81)	-6.29 (-20.29; 7.72)	9.78 (-3.8; 23.36)	4.18 (-7.67; 16.03)	-10.75% (-43.25; 21.75)	22.61% (-22.76; 67.99)	8% (-26.98; 42.99)
	Health deteriorates	24,23% (18,62-29,84)	27,84% (21,27-34,41)	25,47% (20,36-30,58)	23,58% (15,43-31,74)	27,13% (16,81-37,45)	24,75% (17,04-32,47)	24,89% (17,24-32,54)	28,57% (20,48-36,67)	26,21% (19,56-32,87)	1.3 (-9.88; 12.49)	1.44 (-11.67; 14.56)	1.46 (-8.73; 11.65)	5.53% (-48.55; 59.6)	5.32% (-55.13; 65.77)	5.9% (-42.95; 54.75)
35-44	Health improves	20,39% (15,85-24,93)	20,19% (15,29-25,09)	16,42% (12,85-19,98)	17,78% (10,74-24,83)	22,66% (14,44-30,88)	15% (9,51-20,49)	23,04% (17,41-28,67)	17,68% (12,47-22,9)	17,87% (13,35-22,38)	5.26 (-3.77; 14.28)	-4.98 (-14.71; 4.76)	2.86 (-4.25; 9.98)	29.55% (-36.93; 96.04)	-21.96% (-63.49; 19.56)	18.77% (-38.83; 76.37)
	Remains the same	56,87% (51,07-62,68)	51,37% (45,6-57,14)	56,09% (51,27-60,91)	57,66% (48,2-67,13)	53,14% (43,8-62,48)	59,28% (51,71-66,84)	56,08% (49,43-62,72)	49,57% (42,87-56,27)	52,85% (46,93-58,76)	-1.59 (-13.15; 9.98)	-3.56 (-15.06; 7.93)	-6.43 (-16.03; 3.17)	-2.75% (-33.51; 28)	-6.71% (-36.49; 23.08)	-11.1% (-34.09; 11.9)
	Health deteriorates	22,73% (17,78-27,68)	28,44% (23,44-33,44)	27,49% (23,17-31,81)	24,55% (16,33-32,77)	24,20% (16,6-31,81)	25,72% (18,99-32,45)	20,89% (15,43-26,34)	32,74% (26,44-39,05)	29,29% (23,9-34,68)	-3.67 (-13.54; 6.2)	8.54 (-1.34; 18.42)	3.57 (-5.06; 12.19)	-14.94% (-56.03; 26.15)	35.28% (-21.88; 92.45)	13.54% (-28.92; 56.01)
45-54	Health improves	20,83% (16,97-24,69)	17,21% (13,52-20,91)	17,64% (13,8-21,49)	19,26% (13,75-24,78)	21,34% (15,37-27,31)	21,80% (15,31-28,3)	22,38% (17-27,76)	13,13% (8,63-17,63)	13,49% (9,59-17,39)	3.12 (-4.59; 10.82)	-8.21 (-15.68; -0.73)	-8.32 (-15.89; -0.74)	16.78% (-31.81; 65.36)	-38.08% (-67.22; -8.93)	-38.33% (-67.16; -9.49)
	Remains the same	58,96% (54,22-63,7)	60,44% (55,2-65,68)	55,65% (51,1-60,2)	58% (50,92-65,08)	57,10% (48,76-65,45)	53,57% (46,54-60,6)	59,92% (53,58-66,26)	63,74% (57,38-70,1)	57,73% (52,01-63,46)	1.92 (-7.58; 11.42)	6.64 (-3.86; 17.13)	4.17 (-4.9; 13.23)	3.83% (-23.15; 30.81)	12.32% (-19.48; 44.12)	7.46% (-18.02; 32.94)
	Health deteriorates	20,21% (16,39-24,03)	22,35% (17,66-27,03)	26,70% (22,77-30,64)	22,74% (16,94-28,54)	21,56% (13,99-29,13)	24,63% (18,79-30,46)	17,70% (12,7-22,71)	23,13% (17,56-28,7)	28,78% (23,53-34,02)	-5.03 (-12.69; 2.62)	1.57 (-7.83; 10.97)	4.15 (-3.7; 12)	-21.75% (-54.76; 11.27)	7.95% (-45.11; 61.01)	16.52% (-23.63; 56.67)
55-64	Health improves	17,21% (12,7-21,72)	18,76% (14,3-23,22)	16,31% (12,69-19,93)	20,02% (12,74-27,3)	22,26% (15,71-28,8)	15,78% (10,76-20,81)	14,54% (9,17-19,91)	15,43% (9,36-21,51)	16,82% (11,61-22,02)	-5.48 (-14.53; 3.56)	-6.82 (-15.75; 2.11)	1.03 (-6.2; 8.27)	-26.99% (-68.66; 14.69)	-30.28% (-68.06; 7.5)	7.03% (-44.98; 59.05)
	Remains the same	56,22% (50,32-62,13)	56,01% (50,29-61,73)	61,52% (56,77-66,27)	52,50% (43,49-61,5)	55,93% (47,91-63,95)	60,37% (53,63-67,11)	59,77% (52,18-67,36)	56,09% (47,93-64,24)	62,62% (55,93-69,3)	7.28 (-4.5; 19.05)	0.16 (-11.28; 11.59)	2.24 (-7.25; 11.74)	14.5% (-20.37; 49.37)	0.83% (-30.22; 31.88)	4.2% (-21.6; 30)
	Health deteriorates	26,57% (21,14-32)	25,23% (20,19-30,27)	22,17% (18,11-26,22)	27,48% (19,02-35,94)	21,82% (15,27-28,36)	23,84% (17,96-29,72)	25,69% (18,82-32,56)	28,48% (20,94-36,02)	20,57% (14,98-26,15)	-1.79 (-12.69; 9.11)	6.66 (-3.32; 16.65)	-3.28 (-11.38; 4.83)	-5.99% (-51.82; 39.83)	31.26% (-29.42; 91.94)	-13.34% (-49.38; 22.7)
>=65	Health improves	18,90% (14,3-23,5)	22,50% (16,06-28,94)	17,61% (13,21-22,01)	18,43% (11,5-25,35)	23,43% (12,04-34,81)	23,38% (15,81-30,96)	19,27% (13,11-25,43)	21,77% (14,6-28,94)	13,05% (7,95-18,14)	0.85 (-8.42; 10.11)	-1.66 (-15.11; 11.8)	-10.34 (-19.47; -1.21)	3.13% (-52.75; 59.01)	-8.52% (-73.11; 56.06)	-45.38% (-75.57; -15.2)
	Remains the same	54,21% (48,16-60,25)	53,26% (46,53-59,98)	57,34% (51,36-63,32)	55,54% (46,22-64,86)	54,68% (43,61-65,76)	53,58% (43,97-63,19)	53,16% (45,22-61,1)	52,13% (43,81-60,46)	60,31% (52,9-67,73)	-2.38 (-14.62; 9.87)	-2.55 (-16.41; 11.31)	6.73 (-5.4; 18.87)	-5.62% (-36.55; 25.32)	-6.15% (-38.36; 26.06)	10.21% (-23.1; 43.51)
	Health deteriorates	26,89% (21,39-32,4)	24,24% (18,79-29,7)	25,05% (19,42-30,68)	26,03% (17,3-34,77)	21,89% (13,77-30,01)	23,04% (13,53-32,55)	27,57% (20,51-34,62)	26,10% (18,8-33,39)	26,64% (19,86-33,42)	1.53 (-9.7; 12.76)	4.21 (-6.7; 15.12)	3.6 (-8.07; 15.28)	4.4% (-48; 56.8)	17.36% (-43.15; 77.88)	13.22% (-52.72; 79.16)

6.4. Longitudinal estimators

Table 4 shows estimates of better, equal or worse self-perception of health in the population for a given measurement with respect to the same population in the previous measurement. 20.75% of the andalusian population below 16 years old improved their self-perceived general health in measurement 2 with respect to measurement 1, but this percentage was slightly smaller in the following measurements, specially in measurement 4 with respect to measurement 3. On the contrary, 24.29% of this population group presented a worse self-perceived general health in measurement 2 with respect to measurement 1, with this percentage being slightly greater in next measurements. When we analyze these results by sex, it can be observed that it is the women that experiment that decrease in the improvement of general health, as well as the increase in the deterioration of self-perceived general health. Regarding the age, the decreases in the improvement of general health along the pandemic are observed among women between 25 and 54 years old, and the increase in deterioration percentages are observed in those women between 45 and 54 years old. On the other hand, the percentage of people that remained with the same self-perceived general health status, in a given measurement with respect to the previous one, did not vary along the pandemic, except for the population below 24 years old that did experiment increases in the aforementioned percentage, going from 44.26% in measurement 2 to 57.08% in measurement 4. These results are visualized in Figure 8.

If we calculate the ratio of the population that worsens their general health (in a given measurement with respect to the previous one) and the population that improves it, a

Figure 8.

Percentage of population whose self-perceived general health improves, deteriorates or remains the same between a measurement and the previous one

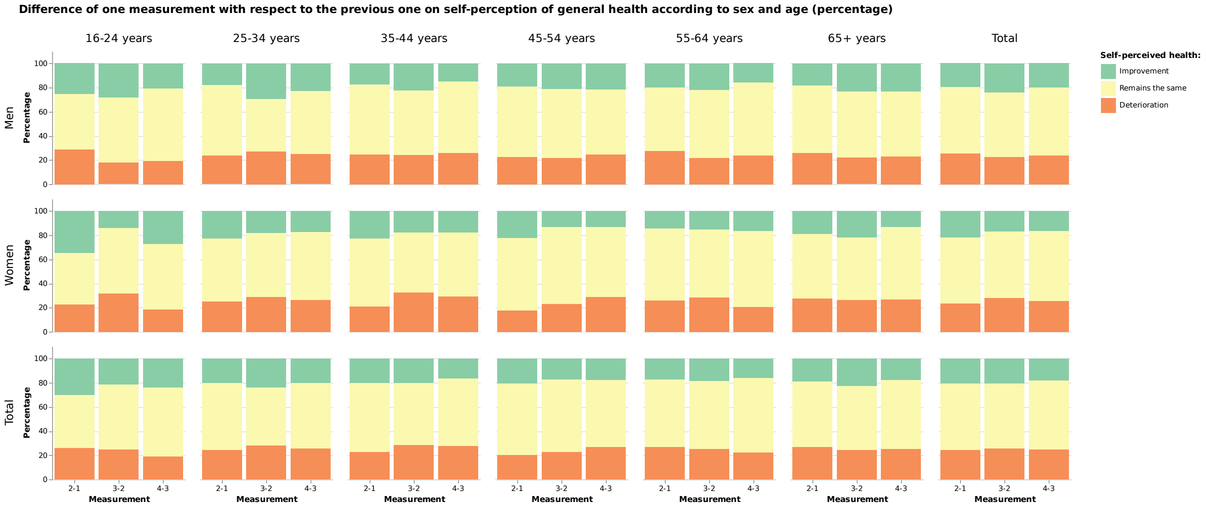
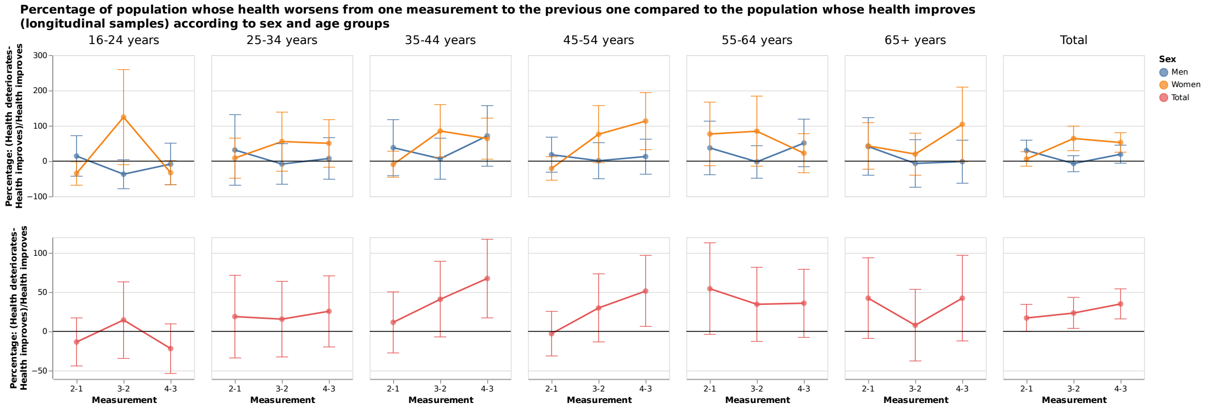


Figure 9.

Percentage of population whose health worsens from one measurement to the previous one compared to the population whose health improves

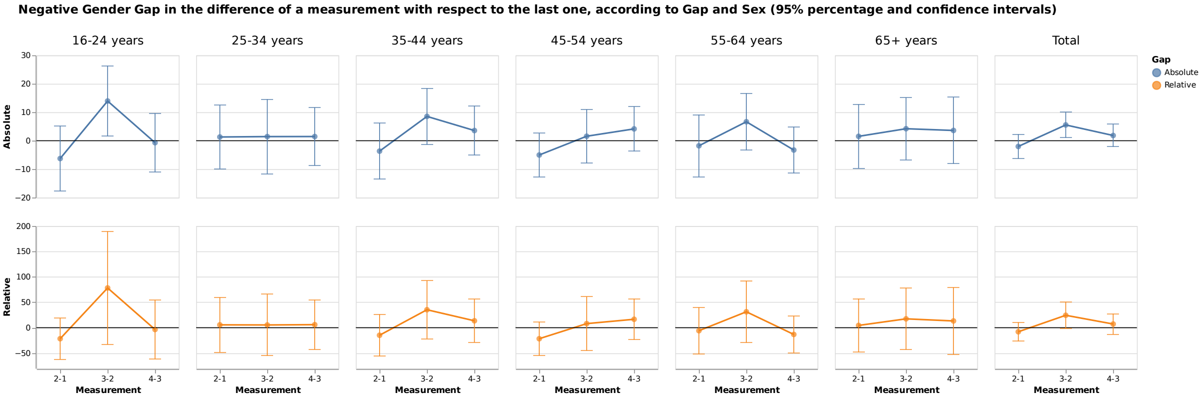


positive value means that there are more people whose self-perceived general health has deteriorated than people whose health has improved, as seen at Figure 9. In relative terms, it could be observed that, in measurement 2 with respect to measurement 1, there was 17.06% more population with worse health than with better health; this percentage increased to 23.32% and 34.88% in measurements 3 and 4 with respect to measurements 2 and 3 respectively. These differences are greater in women, reaching values of 64.08% and 52.32% in measurements 3 and 4 respectively. If the ratio is analyzed according to the age of the individuals, we can spot different patterns in men and women. In this sense, more men population under 55 years old perceived a deterioration in their health in measurement 2 with respect to measurement 1, while this was only observed in women above that age in the women population. Regarding measurement 3 with respect to measurement 2, deterioration of health was more frequently observed in women of any age, with almost no changes between age groups in men. Finally, deterioration of health in measurement 4 with respect to measurement 3 was more frequently observed in men population between 25 and 64 years old, something that was also observed in women population above 25 years old.

Table 4 also shows absolute and relative gender gaps in the improvement, in staying the same or in the deterioration of self-perceived general health in a measurement with respect 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 a better, equal or worse perceived health in a measurement with respect to the previous one, and on the other hand relative gender gap is the relative difference (in percentage) between women and men with a better, equal or worse perceived health. This means that a positive value in the gap (absolute or relative) indicates that the population percentage of women that improved, stayed the same or worsened their self-perceived general health was greater than the corresponding percentage in men population. A negative value would indicate that the percentage was smaller in women. Regarding deterioration of health , we observe at Figure 10 that the percentage of women population whose self-perceived health was worse in measurement 2 than in measurement 1 was 8% lower than its men population counterpart. However, this relative gender gap on health deterioration became positive in next measurements, 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.

Figure 10.

Negative Gender Gap in the difference of a measurement with respect to the last one



7. Conclusions

The rapid evolution of the COVID-19 pandemic has forced researchers to provide timely estimates on the impact of the disease in the population. This has often lead to the establishment of survey studies which did not meet the criteria to be considered probabilistic, entailing many sources of error that may have affect the final estimates obtained from them. For this reason, the ESSOC survey is particularly valuable in the sense that its overlapping probability panel design offers the possibility of obtaining 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 exempt from statistical adjustments to correct for attrition and survey nonresponse.

The two-step adjustment procedure has been established in this study to remove the two main sources of error in the sampling design: the population nonresponse, 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 nonresponse, understood as people who participated in some of the measurements but did not follow up in further ones. Panel nonresponse has been treated using PSA, which is a technique often used for addressing selection bias in online surveys[45] but which can also be used for nonresponse; in fact, it was originally adapted from[18] for that matter[46].

In our study, the XGBoost technique has been used to model the lack of response from one measurement to another. Other ML methods (as logistic regression, decision trees, random forests, ...) could be used, but several papers[8,47,20] show that the set of predictor variables used in general mattered more than the type of ML technique. Regarding neural networks, even though they have had great success for image, text or audio data, this success is due to the use of structures much more advanced than deep feedforward networks. However, for tabular data as it is our case, the inefficacy and unreliability of neural networks is widely known.[48] further explain this issue in their introduction. They propose a novel neural network structure, and they compare it against advanced gradient boosting methods, such as XGBoost, since as they justify those are the current state-of-the-art despite their longevity. It would be of great interest to consider their recent proposal in order to model non-response. However, it is yet very experimental as the scarcity of papers and implementations shows. Therefore, for such an important application as the ESSOC, we prefer an established method.

The application developed in this work is one example where techniques of the machine learning field have to be combined with other important techniques in survey research as calibration and PSA, when studying nonresponse in a panel setting.

Tables 5 and 6 summarise the name, table, figure, formula and interpretation related to the estimators developed throughout this paper for the cross-sectional and longitudinal samples, respectively.

The results observed in the different estimates obtained from the survey show that the impact of the pandemic has hit differently across age groups and genders. More precisely, the self-perceived general health seems to have decreased more notably in older age groups and women, both according to the evolution of cross-sectional estimates and longitudinal estimates. The gender gap, both in absolute and relative terms, has mostly grown as the pandemic advanced, meaning that the changes (mostly decreases in self-perceived general health) have been larger and worse in women in comparison to men. The variable of interest has been the self-perceived general health. It is a well-known fact that subjective variables usually entail measurement errors, as

Table 5.

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

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Original variables	1	4	(12)	Percentages estimations with confidence intervals at 95% and sample size at measurement 4, grouped by sex and age, for the original categories of self-perceived general health.
Dichotomized variables	2	5	(12)	Evolution of percentages and confidence intervals at 95%, grouped by sex and age, of people with excellent or very good 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	No/6	(15)/(16)	Evolution in each measurement with respect to measurement 1 of absolute/relative changes and confidence intervals at 95%, grouped by sex and age, of people with excellent or very good self-perceived general health. A positive value indicates an increase, in percentage points/terms, in the excellent or very good 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 excellent or very good 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 in the absolute change	2/No	7/No	(17)/(18)	Evolution in each measurement (M) with respect to measurement 1 of absolute/relative gender gaps (women versus men) in the absolute/relative change and confidence intervals at 95%, grouped by age, of people with excellent or very good self-perceived general health. A positive value indicates that women show, in percentage points/terms, a larger absolute/relative change in comparison to men in their 'excellent or very good' self-perceived general health of the corresponding measure compared to the first one. Therefore, this result could be seen as a positive gender gap (i.e., better result or favorable to women) in the corresponding measurement compared to the first one. On the contrary, a negative value indicates that women showed, in percentage points/terms, a smaller absolute/relative change in comparison to men in their 'excellent or very good' self-perceived general health. It could be seen as a negative gender gap (i.e., worse result or unfavorable to women) in the corresponding measurement compared to the first one. If the confidence interval does not include the value 0, the corresponding gender gap can be said to be statistically significant.
Absolute/Relative gender gap in the relative change	3/No	7/No	(19)/(20)	

Table 6.

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

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Longitudinal difference	4	8	(26)	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	(28)	Percentage of the population and confidence intervals at 95% that worsens their general health (in a given measurement with respect 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	(29)/(30)	Absolute/Relative difference (in percentage points/terms) and confidence intervals at 95% between women and men with a better, equal, or worse self-perceived health in a measurement with respect to the previous one. A positive value indicates that the percentage of women that improved, stayed the same or worsened their self-perceived general health was greater than the corresponding percentage in men population. A negative value would indicate that the percentage was smaller in women.

the response given in such questions by the interviewee may depend on many unmeasurable factors unrelated to the matter of study but that move the final response away 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 matter.

The descriptive results for the general health self-perception variable are an example applied to this paper in order to show the different estimators, tables and figures developed. All these are extended to the more than 400 ESSOC variables through the web platform developed at `www.researchprojects.com/ESSOC`. 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 the interpretations for the selected variables. This will allow the scientific community not only to access the descriptive results for all the variables of the ESSOC, but also to carry out their own analyses.

Some limitations have to be noted in this study. Firstly, we assume a covariate-dependent missingness pattern, as is usual in propensity score adjustment[49,50,51]. As a referee has revealed, in a panel survey a more realistic assumption can be a missing at random assumption which allows for dependence on the observed y-values in the previous years[39,21] but it has the drawback that the adjustment weights will vary for each variable, which is not useful for multipurpose surveys such as the ESSOC. 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 if this difference in accuracy compensates for the complexity of having to build a different response model for each variable.

On the other hand, in this work we have considered a situation in which the population under study does not vary in time. This is justified because the new measurements are made with little difference compared to the first measurement (at one month, at 6 months and at 12 months) and all the samples are obtained from the same sampling frame[15] so we have assumed that the sample designs refer to the same population. Thus these methods would not be well suited to rotating panel surveys where samples are drawn from different frames on different years and therefore from different populations.

Funding

Funding has been obtained from the competitive calls of the SUPERA COVID-19 Fund of Santander Universities (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. This work is also partially supported by Ministerio de Economía y Competitividad and Ministerio de Ciencia y Educación of Spain [grant PID2019-106861RB-I00 and PI20/00855, respectively], IMAG-Maria de Maeztu CEX2020-001105-M/AEI/

10.13039/501100011033 and from FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (FQM170-UGR20, A-SEJ-154-UGR20).

References

1. Sánchez-Cantalejo C, Rueda MdM, Saez M 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. *International Journal of Environmental Research and Public Health* 2021; 18(15): 8120. DOI:10.3390/ijerph18158120. URL `https://www.mdpi.com/1660-4601/18/15/8120`.

2. Kalton G and Citro CF. Panel surveys: Adding the fourth dimension. *Innovation: The European Journal of Social Science Research* 1995; 8(1): 25–39. DOI:10.1080/13511610.1995.9968429. URL `http://www.tandfonline.com/doi/abs/10.1080/13511610.1995.9968429`.

3. Ardilly P and Lavallée P. Weighting in rotating samples: The silc survey in france. *Survey Methodology* 2007; 33(2): 131–137.

4. Kalton G, Lepkowski J and Lin TK. Compensating for wave nonresponse in the 1979 ISDP research panel. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, volume 372. p. 377.

5. Lepkowski JM. Treatment of wave nonresponse in panel surveys. *Panel surveys* 1989; .

6. Kalton G and Brick JM. Weighting schemes for household panel surveys. *Survey Methodology* 1995; 21(1): 33–44.

7. Kern C, Klausch T and Kreuter F. Tree-based Machine Learning Methods for Survey Research. *Survey research methods* 2019; 13(1): 73–93. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7425836/`.

8. Kern C, Weiß B and Kolb JP. Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning. *Journal of Survey Statistics and Methodology* 2021; : smab009DOI:10.1093/jssam/smab009. URL `https://academic.oup.com/jssam/advance-article/doi/10.1093/jssam/smab009/6364780`.

9. Rendtel U and Harms T. Weighting and Calibration for Household Panels. In Lynn P (ed.) *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley & Sons, Ltd. ISBN 9780470743874 9780470018712, 2009. pp. 265–286. DOI:10.1002/9780470743874.ch15. URL `https://onlinelibrary.wiley.com/doi/10.1002/9780470743874.ch15`.

10. Arcos A, Rueda MdM and 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. DOI:10.3390/math8030423. URL `https://www.mdpi.com/2227-7390/8/3/423`.

11. Lavallée P and Deville J. Theoretical foundations of the generalised weight share method. In *Proceedings of the International Conference on Recent Advances in Survey Sampling*. pp. 127–136.

12. Massiani A. Estimation of the variance of cross-sectional indicators for the silc survey in switzerland. *Survey Methodology* 2013; 39(1): 121–149.

13. Verma V, Betti G and Ghellini G. *Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC*. Università di Siena, Dipartimento di metodi quantitativi, 2006.

14. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Degree of urbanization, 2020. URL `https://www.juntadeandalucia.es/institutodeestadisticaycartografia/gradourbanizacion/`.

15. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Longevity, 2020. URL `https://www.juntadeandalucia.es/institutodeestadisticaycartografia/longevidad/`.

16. Lavallee P. Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology* 1995; 21(1): 25–32.

17. Kovacevic MS. Cross-sectional inference based on longitudinal surveys: Some experiences with statistics Canada surveys. In *Federal Committee on Statistical Methodology Conference*.

18. Rosenbaum PR and Rubin DB. The central role of the propensity score in

observational studies for causal effects. *Biometrika* 1983; 70(1): 41–55. DOI:10.1093/biomet/70.1.41. URL `https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41`.

19. Ferri-García R and Rueda MdM. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Statistics and Operations Research Transactions* 2018; : 159–162URL `https://raco.cat/index.php/SORT/article/view/347847`.

20. Ferri-García R and Rueda MdM. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLOS ONE* 2020; 15(4): e0231500. DOI:10.1371/journal.pone.0231500. URL `https://dx.plos.org/10.1371/journal.pone.0231500`.

21. Juillard H and Chauvet G. Variance estimation under monotone non-response for a panel survey. *Survey Methodology* 2018; .

22. Chen Y, Li P and Wu C. Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* 2020; 115(532): 2011–2021.

23. Chen T and 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 California USA: ACM. ISBN 9781450342322, pp. 785–794. DOI:10.1145/2939672.2939785. URL `https://dl.acm.org/doi/10.1145/2939672.2939785`.

24. Friedman J, Hastie T and Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics* 2000; 28(2). DOI:10.1214/aos/1016218223. URL https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-2/Additive-logistic-regression--a-statistical-view-of-boosting-With/10.1214/aos/1016218223.full.

25. Lee BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; 29(3): 337–346. DOI:10.1002/sim.3782. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.3782.

26. Lee BK, Lessler J and Stuart EA. Weight Trimming and Propensity Score Weighting. *PLoS ONE* 2011; 6(3): e18174. DOI:10.1371/journal.pone.0018174. URL https://dx.plos.org/10.1371/journal.pone.0018174.

27. McCaffrey DF, Ridgeway G and Morral AR. Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 2004; 9(4): 403–425. DOI:10.1037/1082-989X.9.4.403. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.9.4.403.

28. McCaffrey DF, Griffin BA, Almirall D et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 2013; 32(19): 3388–3414. DOI:10.1002/sim.5753. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.5753.

29. Tu C. Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation* 2019; 89(4): 708–719. DOI:10.1080/00949655.2019.1571059. URL `https://www.tandfonline.com/doi/full/10.1080/00949655.2019.1571059`.

30. Zhu Y, Coffman DL and Ghosh D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. *Journal of Causal Inference* 2015; 3(1): 25–40. DOI:10.1515/jci-2014-0022. URL `https://www.degruyter.com/document/doi/10.1515/jci-2014-0022/html`.

31. Rueda MdM, Pasadas-del Amo S, Rodríguez BC et al. 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. *Biometrical Journal* 2022; DOI:https://doi.org/10.1002/bimj.202200035. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202200035`. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.202200035`.

32. Deville JC and Särndal CE. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* 1992; 87(418): 376–382. DOI:10.1080/01621459.1992.10475217. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217`.

33. Rueda M, Martínez S, Martínez H et al. Mean estimation with calibration techniques in presence of missing data. *Computational Statistics & Data Analysis*

2006; 50(11): 3263–3277. DOI:10.1016/j.csda.2005.06.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167947305001349`.

34. Kott PS and Liao D. One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology* 2015; 41(1): 165–182.

35. Cabrera-León A, Lopez-Villaverde V, Rueda M et al. Calibrated prevalence of infertility in 30- to 49-year-old women according to different approaches: a cross-sectional population-based study. *Human Reproduction* 2015; 30(11): 2677–2685. DOI:10.1093/humrep/dev226. URL `https://academic.oup.com/humrep/article-lookup/doi/10.1093/humrep/dev226`.

36. Devaud D and Tillé Y. Rejoinder on: Deville and Särndal's calibration: revisiting a 25-year-old successful optimization problem. *TEST* 2019; 28(4): 1087–1091. DOI:10.1007/s11749-019-00685-z. URL `http://link.springer.com/10.1007/s11749-019-00685-z`.

37. Särndal CE, Swensson B and Wretman JH. *Model assisted survey sampling.* 1. softcover print ed. Springer series in statistics, New York Berlin Heidelberg: Springer, 2003. ISBN 9780387406206.

38. Ardilly P and Osier G. Cross-sectional variance estimation for the french" labor force survey". In *Survey Research Methods*, volume 1. pp. 75–83.

39. Zhou M and Kim JK. An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika* 2012; 99(3): 631–648.

40. Wager S, Hastie T and Efron B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* 2014; 15(1): 1625–1651.

41. Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym). Population and housing census, 2020. URL `https://www.juntadeandalucia.es/institutodeestadisticaycartografia/padron/`.

42. Bergstra J, Bardenet R, Bengio Y et al. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html`.

43. Bergstra J, Yamins D and 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, pp. 115–123. URL `https://proceedings.mlr.press/v28/bergstra13.html`.

44. Akiba T, Sano S, Yanase T et al. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM. ISBN 9781450362016, pp. 2623–2631. DOI:10.1145/3292500.3330701. URL `https://dl.acm.org/doi/10.1145/3292500.3330701`.

45. Lee S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics* 2006; 22(2): 329.

46. Little RJ. Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique* 1986; : 139–157.

47. Castro-Martín L, Rueda MdM and Ferri-García R. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment. *Mathematics* 2020; 8(11): 2096. DOI:10.3390/math8112096. URL `https://www.mdpi.com/2227-7390/8/11/2096`.

48. Arik SÖ and Pfister T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. pp. 6679–6687.

49. Castro-Martín L, Rueda MdM and Ferri-García R. Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *Journal of Computational and Applied Mathematics* 2022; 404. DOI:10.1016/j.cam.2021.113414. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100251813&doi=10.1016%2fj.cam.2021.113414&partnerID=40&md5=10675d46a5fb2ea604fcfc99155317ab`.

50. Castro-Martín L, Rueda MdM, Ferri-García R et al. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics* 2021; 9(23): 2991. DOI:10.3390/math9232991. URL `https://www.mdpi.com/2227-7390/9/23/2991`.

51. Ferri-García R, Rueda MdM and 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.

52. Wu C and Thompson ME. *Sampling theory and practice.* Springer, 2020.

53. Biau G and Cadre B. Optimization by gradient boosting. In *Advances in Contemporary Statistics and Econometrics.* Springer, 2021. pp. 23–44.

54. Wolter KM. *Introduction to variance estimation*, volume 53. Springer, 2007.

55. Devroye L, Györfi L and Lugosi G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Annex 1: Variance estimation.

Given the independence of the sampling designs used to select the samples $s^{(1)}$ and $s^{(t)}$ in the case of the combined estimator of the type 5, we have considered the estimator of the variance:

$$\widehat{V} = \alpha_1^2 \hat{V}_1 + (1-\alpha_1)^2 \hat{V}_2 \,,$$

where $\hat{V}_1$ and $\hat{V}_2$ are the usual Horvitz-Thompson estimators of the variances $V(\hat{Y}_r^{(t)})$ and $V(\hat{Y}_n^{(t)})$ respectively.

The properties of the inverse propensity weighting estimators as a method for handle missing data is developed, among others, in[52] section 9.6 and in[47] for treating the bias of non-probability sampling. Under certain regularity conditions for the response model and the sampling design model, the IPW estimator

$$\hat{Y}_P^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ih}^1 \frac{1}{\hat{\pi}_{ih}^{(t)}} y_{ih}^{(t)} = \sum_{i \in s_r^{(t)}} d_i^1 \frac{1}{\hat{\pi}_i^{(t)}} y_i^{(t)}$$

is asymptotically unbiased for the population total $Y^{(t)}$ ($\hat{Y}_P^{(t)} - Y^{(t)} = O_p(n^{-1/2})$) and the asymptotic expression for its variance is given by

$$V(\hat{Y}_P^{(t)}) = \sum_U (y_i^{(t)}/\hat{\pi}_i^{(t)} - \mathbf{b}_{1t}^T \mathbf{x}_i)^2 (1 - \hat{\pi}_i^{(t)}) \hat{\pi}_i^{(t)} + \mathbf{b}_{1t}^T D \mathbf{b}_{1t} \,, \tag{33}$$

where $\mathbf{b}_{1t}^T = \sum_U (1 - \hat{\pi}_i^{(t)}) y_i \mathbf{x}_i^T / \sum_U \hat{\pi}_i^{(t)} (1 - \hat{\pi}_i^{(t)}) \mathbf{x}_i \mathbf{x}_i^T$, and $D_t = V_p(\sum_{i \in s_r^{(t)}} d_i \hat{\pi}_i^{(t)} \mathbf{x}_i)$ where V_p denotes the design-based variance under the sampling design p.

Thus, there are two sources of variation for the PSA estimator: the probability sampling design and the missing mechanism described by the propensity score mode. The above asymptotic variance provides a plug-in method for variance estimation. Thus we consider the variance estimator given by

$$\hat{V}(\hat{Y}_P^{(t)}) = \sum_{s_r^{(t)}} (y_i/\hat{\pi}_i^{(t)} - \tilde{\mathbf{b}}_1^T \mathbf{x}_i)^2 (1 - \hat{\pi}_i^{(t)}) + \tilde{\mathbf{b}}_1^T \tilde{D} \tilde{\mathbf{b}}_1 \,, \tag{34}$$

where

$$\tilde{\mathbf{b}}_1^T = \sum_{s_r^{(t)}} \frac{(1 - \hat{\pi}_i^{(t)})}{\hat{\pi}_i^{(t)}} y_i \mathbf{x}_i^T / \sum_{s_r^{(t)}} (1 - \hat{\pi}_i^{(t)}) \mathbf{x}_i \mathbf{x}_i^T$$

and

$$\tilde{D} = \sum_{i,j \in s_r^{(t)}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{\pi}_i^{(t)} \hat{\pi}_j^{(t)}}{\pi_i \pi_j} \mathbf{x}_i \mathbf{x}_i^T$$

where

$\pi_i = \frac{n_h^{(1)}}{N_h}$, if $i \in U_h$,

$$\pi_{ij} = \begin{cases} \frac{n_h^{(1)}(n_h^{(1)}-1)}{N_h(N_h-1)} \text{ if } i,j \in U_h \\ \\ \\ \pi_i\pi_j \text{ if } i \in U_h \text{ and } j \in U_k,\ k \neq h. \end{cases}$$

Note. Regarding the sample design used (stratified random sampling) and the assumed non-response model (9) it is clear that they verify the regularity conditions given in[52]. In this work we also have to assume that the $\widehat{m_t}$ estimator provided by the XGBoost method gives consistent estimates for the propensities. The large sample properties (including consistency) of the gradient boosting algorithms are shown in[53].

The absolute gender gap estimators in the absolute change, $G\hat{G}abs_{abs}^{(t)}$ is defined as the linear combination of two estimators in certain domains, and thus for its variance estimation we have used Taylor Linearization Approach in the same form as in[17] section 3.1.

The estimator for the variance of the relative gender gap estimator in the absolute change, $G\hat{G}abs_{rel}^{(t)}$ is obtained using the usual variance estimation method for a ratio estimator[54].

The variance estimator for the estimator 21 of the change $\theta^{(t)}$ is obtained in a similar way by changing the values y_i by $y_i^{(t)} - y_i^{(1)}$. The variance estimator for 26 is obtained by changing the values y_i by $y_i^{(t)} - y_i^{(t-1)}$ and the estimated propensities $\hat{\pi}_i^{(t)}$ by $\hat{\pi}_i^{(t,t-1)}$.

Finally, variance estimation for calibrated estimators is developed following the methodology proposed in[32] using estimated residual for the weighted regression.

Annex 2: XGBoost

As stated, the XGBoost[23] algorithm is chosen as the machine learning method for estimating the propensities. Each model is expressed as $F : \mathbb{R}^m \rightarrow \mathbb{R}$ which, for the given values of x_k, returns the associated propensity of the individual k participating in the survey. The algorithm minimizes a risk functional

$$C(F) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}(F(x_k), \delta_k)\,,$$

where $\mathcal{L}$ is a loss function measuring the error and $\delta_k = 1$ if the individual k participated in the survey or $\delta_k = 0$ otherwise.

For this purpose, it considers linear combinations of a class $\mathcal{F}$ of functions $f : \mathbb{R}^m \rightarrow \mathbb{R}$ which are expressed as binary decision trees. Each decision tree $f \in \mathcal{F}$ takes the form $f = \sum_{j=1}^{T} \beta_j \chi_{A_j}$ where T determines the number of final nodes, $(\beta_1, ..., \beta_T) \in \mathbb{R}^T$ and $A_1, ..., A_T$ is a partition of $\mathbb{R}^m$ as described in[55] Chapter 20.

In our case, $C(F)$ is regularized in order to minimize the loss while avoiding complex trees (which would cause overfitting):

$$C_{reg}(F) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}(F(x_k), \delta_k) + \sum_{i=1}^{K} \Omega(f_i)\,,$$

where the chosen $\mathcal{L}$ is the logistic loss as described in Section 4.2, K is the number of trees forming the linear combination and $\Omega : \mathcal{F} \rightarrow \mathbb{R}^+$ penalizes complex trees:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\beta\|^2\,,$$

with γ and λ being hyperparameters.

XGBoost achieves this minimization via Gradient Tree Boosting[24], an iterative process. In each iteration $i = 1, ..., K$ with $F^{(0)} = 0$, a new tree f_i is added to the previous estimator $F^{(i-1)}$ such that

$$C_{reg}^{(i)} = \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}(F^{(i-1)}(x_k) + f_i(x_k), \delta_k) + \Omega(f_i)$$

is minimized. Since finding the optimal value for $f_i \in \mathcal{F}$ is not computationally feasible, a greedy approach is considered by iteratively adding branches which reduce the error. Finally, $F^{(i)} = F^{(i-1)} + \eta f_i$ where η is another hyperparameter which assures convergence[53].

As seen, the method depends on several hyperparameters for a proper functioning. The hyperparameter optimization process considered in this application is described in Section 6.2.