WID.world WORKING PAPER SERIES N° 2018/12

The Weight of the Rich:
Improving Surveys Using Tax Data

Thomas Blanchet
Ignacio Flores
Marc Morgan

First version: October 2018
This version: October 2019

The Weight of the Rich: Improving Surveys Using Tax Data[*]

Thomas Blanchet[†] Ignacio Flores[‡] Marc Morgan[†]

This version: October 3, 2019

Abstract

Household surveys fail to capture the top tail of income and wealth distributions, as evidenced by studies based on tax data. Yet to date there is no consensus on how to best reconcile both sources of information. This paper presents a novel method, rooted in calibration theory, which helps to solve the problem under reasonable assumptions. It has the advantage of endogenously determining a "merging point" between the datasets before modifying weights along the entire distribution and replacing new observations beyond the survey's original support. We provide simulations of the method and applications to real data. The former demonstrate that our method improves the accuracy and precision of distributional estimates, even under extreme assumptions, and in comparison to other survey correction methods using external data. The empirical applications provide useful and coherent illustrations in a wide variety of contexts. Results show that not only can income inequality levels change, but also trends. Given that our method preserves the multivariate distributions of survey variables, it provides a more representative framework for researchers to explore the socio-economic dimensions of inequality, as well as to study other related topics, such as fiscal incidence.

Keywords: Survey Representativeness, Tax Data, Reweighting, Calibration, Inequality.
Classification: C83, D31, N30.

[*] We gratefully acknowledge funding from the *Fundación Ramón Areces*, ERC (Grant 340831), Ford Foundation, INET (Grant INO14-00023) and from other partners of the World Inequality Lab. We thank Facundo Alvaredo, Yonatan Berman, François Bourguignon, Lucas Chancel, Mauricio De Rosa, Francisco Ferreira, Emmanuel Flachaire, Pablo Gutiérrez, Amory Gethin, Thanasak Jenmana, Nora Lustig, Brian Nolan, Thomas Piketty, Li Yang and Gabriel Zucman for helpful discussions of earlier versions of this paper, as well as participants at the 2018 INET seminar series at the University of Oxford, the May 2018 Workshop on harmonizing surveys and tax data at the Paris School of Economics, the ECINEQ 2019 conference, the inequality seminar at ECLAC Santiago (2019), the STEP Seminar at the Université Paris 1 Panthéon–Sorbonne and the Stone Center of Inequality Research seminar at INSEAD.

[†] Paris School of Economics and World Inequality Lab

[‡] INSEAD and World Inequality Lab

Introduction

For a long time, most of what we knew about the distribution of income, wealth and their covariates came from surveys, in which randomly chosen households are asked to fill a questionnaire. Household surveys have been an invaluable tool for tracking the evolution of society. But in recent years, the research community has grown increasingly concerned with their limitations. In particular, surveys have struggled to keep track of the evolution of the top tail of the distribution, due mainly to heterogeneous response rates, misreporting and small sample bias, which distort all sorts of distributional estimates. These biases end up affecting the way public policy is designed and evaluated.

For this reason, researchers have increasingly been turning to a different source to study inequality: tax data. The idea is not new; we can trace it back to the seminal work of Kuznets (1953), or even Pareto (1896). More recently, Piketty and Saez (2003) and Piketty (2003) applied their method to more recent data for France and the United States. This work was extended to more countries by many researchers whose contributions were collected in two volumes by Atkinson and Piketty (2007, 2010) and served as the basis for the World Inequality Database (`http://wid.world`).

But tax data have their own limitations. They usually only cover the top of the distribution and include at best a limited set of covariates. They do not capture well informal and tax-exempt income. They are often not available as microdata but rather as tabulations summarizing the distribution, which limits their use. The statistical unit that they use (individuals or households) depends on the local legislation and may not be comparable from one country to the next. This is why many indicators, such as poverty rates or gender gaps, have to be calculated from surveys. The use of different – and sometimes contradictory – sources to compute statistics can make it hard to build consistent and accurate narratives on distributional matters. This explains the ongoing effort to combine the different data sources at our disposal in a way that exploits their strengths and corrects their weaknesses.

The Distributional National Accounts (DINA) project is a prominent example of this effort. Its guidelines (Alvaredo et al., 2017) emphasize the need to look at the entire distribution, harmonize concepts, and where possible decompose the distribution according to socio-demographic characteristics. Piketty, Saez, and Zucman (2018) for the United States, and Garbinti, Goupille-Lebret, and Piketty (2018) for France have used both survey and tax data to construct distributional statistics that account for all of the income recorded in national accounts. The resulting dataset not only allowed them to reassess the evolution of income concentration statistics, but also to study subjects such as: gender gaps, growth incidence curves or the distributive impact of fiscal policy. But these examples depend in large part on the existence of reliable administrative microdata accessible to researchers, to which information from surveys can be added to account for the limited sources of income not covered in the tax data.

In many countries, both developed and less developed, such direct access is quite rare. Instead, tabulations of fiscal income, containing information on the number and declared income of individuals by income bracket, are more commonly available. The population coverage in the tabulations is often substantially less than the total adult population, and the difference varies with the country studied. Furthermore, in contexts of high informality, which is the case for many developing countries, even if tax declarations had full population coverage, they could not be assumed to be reliable across the whole distribution. In such cases it is better to proceed the other way round: rather than incorporating survey information into the tax data, we need to incorporate tax information into the survey data.

There has been a number of suggested approaches to deal with the problem of merging tax and survey data, yet the literature has largely failed to converge towards a standard. Crucially, most of the existing approaches directly adjust the income or wealth distributions, overlooking the goal of preserving the survey's representativeness in terms of covariates, while relying on arbitrary assumptions in the process. In this paper, we develop a methodology that has significant advantages over previous ones, and which should cover most practical cases within a single, united framework. Our method avoids relying, to the extent possible, on *ad hoc* assumptions and parameters. We present a data-driven way to determine the point in the survey data where the under-coverage of income starts. This is our "merging point" — the point in the distribution were survey data and tax data are merged. We perform necessary adjustments in a way that minimize distortions from the original survey, and preserve desirable properties, such as the continuity of the density function. Rather than directly making assumptions on complex summary statistics such as quantiles or bracket averages, our method makes assumptions that are easily interpretable at the level of observations. The algorithm acknowledges the presence of covariates, so that we ensure the representativeness of the survey in terms of income while maintaining — and possibly improving — its representativeness in terms of age, gender, or any other dimension along the distribution. As a result, we can preserve the richness of information in surveys, both in terms of covariates and household structure.

Our method proceeds in three steps, the first aimed at selecting the merging point between the datasets, and the other two aimed at correcting for the two main types of error in surveys: non-sampling error and sampling error. Non-sampling error refers to issues that cannot easily be solved with a larger sample size, and typically arise from unobserved heterogeneous response rates. In the second step, we correct for these issues using a reweighting procedure rooted in survey calibration theory (Deville and Särndal, 1992). In doing so, we address a longstanding inconsistency between the empirical literature on top incomes in surveys, and the established practice of most survey producers. Indeed, since Deming and Stephan (1940) introduced their raking algorithm, statistical institutes have regularly reweighted their surveys to match known demographic totals from census data. Yet the literature on income has mostly relied on adjusting the value of observations, rather than their weight, to enforce consistency between tax and survey data. We argue that the theoretical foundations of such approaches are less explicit and harder to justify.

This initial correction step addresses non-sampling error, but it is limited in its ability to correct for sampling error, meaning a lack of precision due to limited sample size.[1] A clear example is the maximum income, which is almost always lower in surveys than in tax data, something no amount of reweighting can do anything about. Top income shares of small income groups are also strongly downward biased in small samples (Taleb and Douady, 2015), so inequality will be underestimated even if all the non-sampling error has been corrected. To overcome this problem, we supplement the survey calibration with a further step, in which we replace observations at the top by a distribution generated from the tax data, and match the survey covariates to it. The algorithm for doing so preserves the distribution of covariates in the original survey, their correlation with income, and the household structure, regardless of the statistical unit in the tax data. The result is a dataset where sampling variability of income at the top has been mostly eliminated, and whose covariates have the same statistical properties as the reweighted survey. Because we preserve the nature of the original microdata, we can use the output to experiment with different statistical units, equivalence scales, calculate complex indicators, and perform decompositions along any dimension included in the survey.

In order to illustrate how the method operates in practice, we run two different types of applications. First, since the true distribution of income and wealth is always unknown, we simulate artificial populations that are drawn from parametric distributions. These include behavioral assumptions that define two main sources of bias, namely heterogeneous response rates and misreporting. Using these biases, we simulate a large number of consecutive surveys and then apply our correction method using synthetic tax tabulations. We use these experiments to assess the accuracy and precision of the resulting estimates and to compare them to those derived from both the raw sample and the most common alternative methods using external data — namely methods that directly replace survey incomes with tax incomes for the same quantiles in the distribution. We demonstrate that our method is superior to available options, not only because it relies on reasonable assumptions that enable the use of resulting micro-datasets — unlike the "replacing" alternative — but also because it produces estimates that are consistently closer to true values with lower variance.

In our second application, we apply our method to real data from five countries: France, U.K., Norway, Brazil and Chile. Our case studies are chosen to showcase the

[1] Calibration methods can, to some extent, correct for sampling error. But their ability to do so only holds asymptotically (Deville and Särndal, 1992), so it does not apply to narrow income groups at the top of the distribution.

wide applicability of the method to both developed countries and less-developed countries. The method makes upward revisions to inequality estimates in all cases, with varying degrees of magnitude, depending on the quality of the underlying data and the level of inequality in each country. It can also produce differing inequality trends. Moreover, our empirical results support the findings of our simulations concerning the difference between our method and the replacing alternative.

For practical use, we have developed a Stata command that applies our method. The program works with several input types, income concepts and statistical units, ensuring flexibility for users. Our method may therefore easily be used by researchers interested in analyzing the different dimensions of inequality.[2] The main goal of this paper is to describe the theoretical and practical details behind this readily usable method, as well as its advantages with respect to existing approaches.

The remainder of the paper is structured as follows. In section 1 we relate our paper to the existing literature. In section 2 we lay out the theoretical framework of our method. This is followed by applications to simulated distributions and practical applications to specific countries in section 3, before concluding.

1 Literature on Survey Correction Methods

Numerous studies have sought to adjust survey data primarily to improve the latter's representativeness and/or produce a more accurate distribution of income. In some instances this has been achieved with the aid of external administrative data. We identify three distinguishable methodological strands present in this literature. The first strand opts to reweight survey observations. The second strand replaces the income value of observations with a value typically drawn from a parametric distribution or an external data source. Finally, a third strand identifies the need to employ a hybrid procedure by combining reweighting and replacing.

1.1 Reweighting Observations

The studies that focus on reweighting usually formalize the bias as nonresponse. Many papers in this literature estimate a parametric model of nonresponse to adjust survey weights, but do not use direct data on the distribution of income. Korinek, Mistiaen, and Ravallion (2006) make this type of adjustment using nonresponse rates across geographic areas and the characteristics of respondents within regions. This type of approach can be sensitive to the degree of geographic aggregation used calculating response rates. This is an issue explored in more detail by Hlasny and Verme (2017; 2018) for the US and

[2] The package to download is `bfmcorr` for the correction method, which includes two sub-commands: `postbfm` for the post-estimation output and `bfmtoy` for parametric simulations. The command and its sub-commands come with a full set of user instructions.

European case respectively, using similar probabilistic models. Depending on the nature of the survey data, greater or less geographic dis-aggregation on nonresponse rates can be more appropriate to the adjustment at hand.

Crucially, these models do not use direct information on the income distribution — often due to lack of availability. Instead, they have to infer relationships between individual nonresponse and individual characteristics based on aggregate relationships between average nonresponse and an average of certain characteristics. This make these methods susceptible to the pitfalls of ecological inference. In particular, to the extent that nonresponse bias is a strongly nonlinear function of income (which we observe in practice), the relationship between income and nonresponse will look very different at the aggregate and the individual level. Our proposal instead makes use of direct administrative data to determine the relationship between income and nonresponse.

There are a few studies in this literature that combine surveys with external sources to measure inequality. An example of this is the case study of Argentina in Alvaredo (2011), in which the corrected Gini coefficient is estimated by assuming that the top of the survey distribution (top 1% or top 0.1%) completely misses the richest individuals that are represented in tax data. This accounts for the bias of nonresponse and corrects the distribution via an implicit reweighting procedure. The specific form of the nonresponse bias that is assumed tacitly is, nonetheless, a rather restrictive one. Indeed, the correction implies a deterministic nonresponse rate equal to 1 above a previously selected quantile and 0 under it. Furthermore, in both of the empirical applications (the US and Argentina) the threshold beyond which the tax data is used is chosen arbitrarily.[3] Our method on the other hand tries at best to avoid arbitrary choices on the portion of the survey distribution to be corrected or on the form of the bias implied by the correction.

To our knowledge the paper that comes closest to proposing an approach that resembles the one we propose here, in terms of criteria and methodology, is Medeiros, Castro Galvão, and Azevedo Nazareno (2018) applied to Brazilian data. That is, it is the only study that combines tabulated tax data with survey micro-data by explicitly reweighting survey observations. More specifically, the authors apply a Pareto distribution to incomes from the tax tabulation to correct the top of the income distribution calculated from the census. Their method involves re-weighting the census population by income intervals above a specified merging point, which is determined from the comparison of the median total income reported in each quantile of the tax data and in the Census (0.5% of the adult population sorted by income).

However, important differences remain. Contrary to our method, the choice of the merging point is not endogenous, but chosen by the authors as the most relevant point beyond which the tax data presents a more concentrated distribution. Thus, multiple

[3] In any case, the goal of the paper is not to tackle the nonresponse or misreporting biases directly, but to provide a simple estimation of a corrected Gini coefficient.

points can be used, and indeed the authors test two. Our method endogenously determines a single merging point based on a more comprehensive treatment of the form of the non-response bias. Importantly, our approach preserves the continuity of the density of income — something that only a specific choice of the merging point can ensure. To guide their choice of the merging point, Medeiros, Castro Galvão, and Azevedo Nazareno (2018) look at the rank at which income in the tax data exceed that of the survey. Yet from the perspective of correcting for non-response, such a point does not have any well-defined interpretation.

Moreover, while they increase the weight of observations above the merging point, they do not reduce the weight of individuals below this point, such that the corrected population ends up being larger than the original official population. The authors do not provide a way to ensure the representativeness of characteristics other than income after the adjustment either — their purpose is to remedy the underestimation of top incomes in surveys, without a unified calibration framework. Moreover, their method does not remedy the lack of precision at the top of the distribution arising from sampling limitations, resulting in downward biased income shares of small income groups, especially in small samples. In contrast, our method addresses all of these issues.

1.2 Replacing Incomes

The general feature of the "replacing" approach is that it involves the direct replacing of survey incomes with incomes from tax data. Although there is no unified theory or explicit justification behind the applications of this adjustment procedure, most of these methods share some defining characteristics. In practice, they generally adjust distributions by replacing cell-means in the survey distribution of income with those from the tax distribution for the same sized cells (i.e. fractiles) of equivalent rank in the population. The size of the cells varies by study (Burkhauser et al., 2016; Piketty, Yang, and Zucman, 2017; Chancel and Piketty, 2017; Czajka, 2017). Furthermore, the overall size of the population group whose income is to be adjusted is sometimes chosen arbitrarily, such as the top 20% in the distribution (Piketty, Yang, and Zucman, 2017), the top 10% (Burkhauser et al., 2016; Chancel and Piketty, 2017), the top 1% (Burkhauser, Hahn, and Wilkins, 2016; Alvaredo, 2011), or the top 0.5% of survey observations (DWP, 2015).

This decision can be made less arbitrary from the comparison of threshold or average incomes by fractile in the two distributions. The size of the group is then chosen as the point in the distribution where the two quantile functions cross (e.g. Czajka (2017)). As we have noted earlier, this point is not really meaningful from a statistical viewpoint. In fact, under the most natural assumption (increasing nonresponse profile) it should not even exist, i.e. the quantile functions do not intersect.

Other non-arbitrary choices take the minimum income level that requires mandatory tax filing (Diaz-Bazan, 2015). While these try to keep the use of survey data to measure the top of the income distribution to a strict minimum, they assume that the entire tax distribution is reliable. We argue however that not all the income in tax data should be considered reliable given the difference between declarable income thresholds and taxable income thresholds. The quality of tax data generally increases with income in a manner that is often not well defined, and given this uncertainty it makes sense to limit their use to the portion that is absolutely necessary.

In certain cases, the survey distribution stops being reliable before the tax data can be trusted. This happens in particular in countries where only a small part of the population file a tax return. The fact that quantiles from both sources may not cross is evidence of this problem. In such cases, from the point at which we stop trusting the survey to the point at which we start trusting the tax data, one option is to rescale upwards the income values from the survey distribution. This can be done using various profiles of rescaling coefficients (usually linear) (Chancel and Piketty, 2017; Piketty, Yang, and Zucman, 2017; Novokmet, Piketty, and Zucman, 2018). This procedure ensures at least that the quantile function is continuous. These rescaling methods can be seen as an extension of the general replacing methods.

Replacing survey-respondents' declared income has been viewed as adjusting for the misreporting bias in surveys (Burkhauser et al., 2018; Jenkins, 2017). In Appendix A we formalise the existence of this bias both when it operates alone and when it operates in the presence of non-response. We compare existing replacing methods to our own method, and we explain why they only correct for misreporting under very strong and unrealistic assumptions — namely that the income rank in the survey distribution and in the benchmark distribution are the same, and that underreporting is a deterministic function of this rank.

1.3 Combined Reweighting and Replacing

Some voices stress the need to combine the aforementioned correction approaches. Bourguignon (2018), while reviewing the typical adjustment methods employed, correctly highlights that any method must dwell on three important parameters: the amount of income to be assigned to the top, the size of this top group, and the share of the population added to the top in the survey. The definition of these three parameters implies a correction procedure combining reweighting and replacing methods. His analysis goes on to study the ways in which these choices impact the adjustments made to the original distribution. However, this analysis does not shed light on *how* to make these choices. Moreover, in reviewing multiple correction methods and applying them to Mexican survey data (including the combined case, where all three parameters mentioned take non-zero values), he only considers the situation “where nothing is known about the distribution of the missing income, unlike when tax records or tabulations are available” (Bourguignon, 2018). This is in contrast to our approach for correcting survey microdata, which combines the two previous methods, but which explicitly merges tax data with surveys to produce more realistic distributions of income.

In summary, contrary to existing methods, our method uses external tax data, endogenously finds a non-arbitrary merging point, and preserves the multivariate distribution of covariates and population totals. Moreover, it is grounded on a solid theoretical framework, which we now turn to explain in the following section.

2 Theory and Methodology

To describe our method and the theory behind it, we part from the simple univariate setting, where we adjust the weight of observations in the survey at different income levels. The second section explains how to use the theory of survey calibration to handle more complex multivariate settings. Finally, the third section explains how we address the problem of sampling error, which reweighting has only a limited ability to address.

2.1 Univariate Setting

In this section we first explain the intuition behind the correction before presenting how we choose the merging point between the two distributions.

2.1.1 Intuition

Let X and Y be two real random variables. We will use Y to represent the true income distribution, part of which we assume is recorded in the tax data.[4] And we will use X to represent the income distribution recorded in the survey. Each random variable has a probability density function (PDF) f_Y and f_X, a cumulative probability function (CDF) F_Y and F_X, and a quantile function Q_Y and Q_X.

Let $\theta(y) = f_X(y)/f_Y(y)$ be the ratio of the survey density to the true density at the income level y. This represents the number of people within an infinitesimal bracket $[y, y + \mathrm{d}y]$ according to the the survey, relative to the actual number of people in the bracket. If $\theta(y) < 1$, then people with income y are underrepresented in the survey. Conversely, if $\theta > 1$, then they are overrepresented.

The value of $\theta(y)$ may be interpreted as a relative probability. Indeed, let D be a binary random variable that denotes participation to the survey: if an observation is

[4] In reality, part of the true income may also be missing from the tax data due to non-taxable income not reported on the declaration and tax evasion. The extent of these omissions vary by country, and their treatment are beyond the scope of this paper.

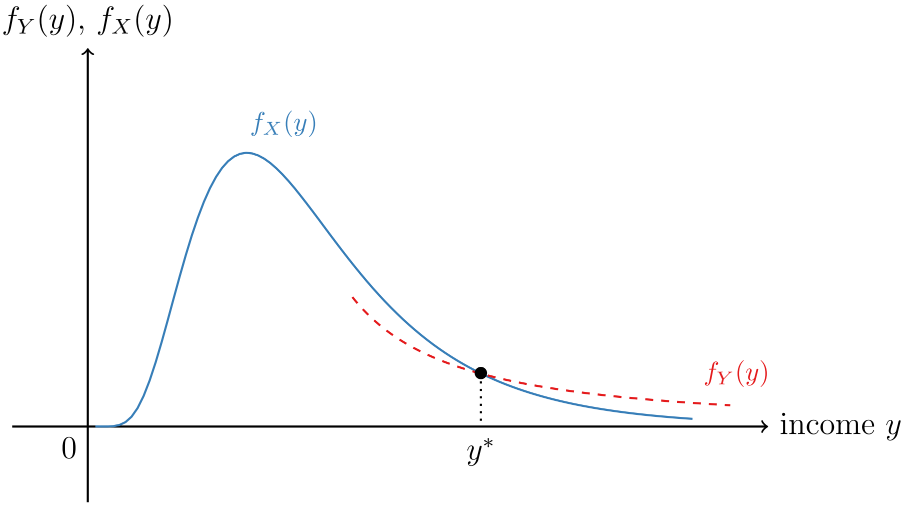
included in the sample, then $D = 1$, otherwise $D = 0$. Then Bayes' formula implies:

$$\theta(y) = \frac{f_X(y)}{f_Y(y)} = \frac{1}{f_Y(y)} \times f_Y(y) \frac{\mathbb{P}\{D = 1|Y = y\}}{\mathbb{P}\{D = 1\}} = \frac{\mathbb{P}\{D = 1|Y = y\}}{\mathbb{P}\{D = 1\}} \tag{1}$$

If everyone has the same probability of response, then $\mathbb{P}\{D = 1|Y = y\} = \mathbb{P}\{D = 1\}$, and $\theta(y) = 1$. Hence $f_X(y) = f_Y(y)$ and the survey is unbiased. What matters for the bias is the probability of response at a given income level relative to the average response rate, which is why we have the constraint $\mathbb{E}[\theta(Y)] = 1$. Intuitively, if some people are underrepresented in the survey, then mechanically others have to be overrepresented, since the sum of weights must ultimately sum to the population size.

This basic constraint has important consequences for how we think about the adjustment of distributions. Any modification of one part of the distribution is bound to have repercussions on the rest. In particular, it makes little sense to assume that the survey is not representative of the rich, and at the same time that it is representative of the non-rich.

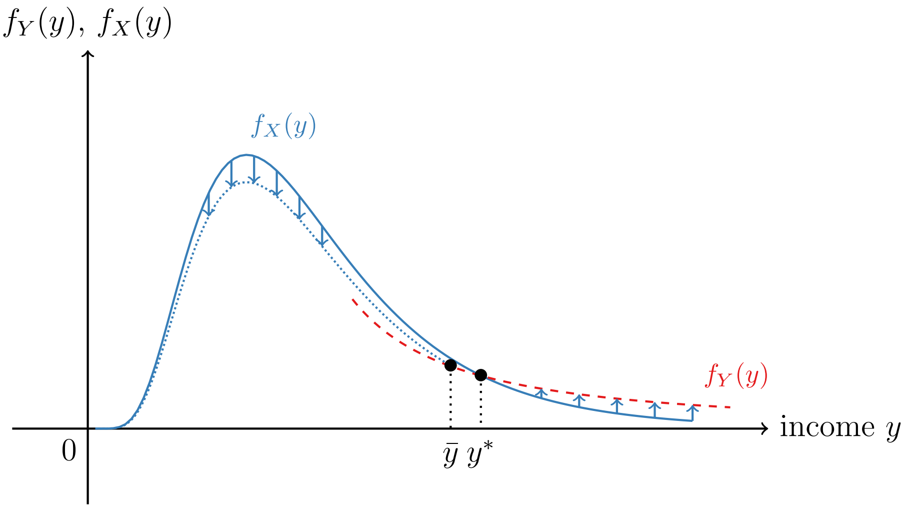
Figure 1: A "true" and biased income distribution



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y, which is only observed at the top. For high incomes, the survey density is lower than the tax data density, which means that high incomes are underrepresented. If some individuals are underrepresented, then other have to be overrepresented: they correspond to people below the point y^*.

Figure 1 represents the situation graphically, in the more common case where $\theta(y)$ is lower for top incomes. We show a truncated version of f_Y since tax data often only cover a limited part of the whole distribution. The fact that the dashed red line $f_Y(y)$ is above the solid blue line $f_X(y)$ mean that top incomes are underrepresented. Therefore, lower incomes must be overrepresented, which is what we see below the point y^*. This pivotal value is unique assuming that θ is monotone. The appropriate correction procedure here would be to increase the value of the density above it, and decrease its value below it. The intuition behind reweighting is that we have to multiply the survey density f_X by a factor $1/\theta(y)$ to make it equal to the true density f_Y. In practice, this means multiplying the weight of any observation Y_i by $1/\theta(Y_i)$.

Figure 2: The intuition behind reweighting



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y. Above the merging point $\bar{y}$, the reweighted survey data have the same distribution as the tax data (dashed red line). Below the merging point, the density has been uniformly lowered so that it still integrates to one, creating the dotted blue line.

When we observe both f_Y and f_X, we can directly estimate θ nonparametrically. But because we do not observe the true density over the entire support, we have to make an assumption on the shape of θ for values not covered by the tax data. We will assume a constant value. Behind this assumption, there are both theoretical motivations that we develop in section 2.2, and empirical evidence that we present in section 3. Intuitively, it means that there is no problem of representativeness within the bottom of the distribution, so that the overrepresentation of the non-rich is only the counterpart of the underrepresentation of the rich. We can therefore write the complete profile of θ as:

$$\theta(y) = \begin{cases} \bar{\theta} & \text{if } y < \bar{y} \\ f_X(y)/f_Y(y) & \text{if } y \geq \bar{y} \end{cases} \tag{2}$$

We call $\bar{y}$ the *merging point*. It is the value at which we merge observations from the tax data into the survey. A naive choice would be to use the tax data as soon as they become available, but this will often lead to poor results. This is because the point from which the tax data become reliable is not necessarily sharp and well-defined, so in practice it will be better to start using the tax data only when it becomes clearly necessary. The proper choice of that point is an important aspect of the method on which we return to in section 2.1.2. For now we will take it as given, and only assume that it is below the pivotal point y^* of figure 1. Figure 2 shows how the reweighting using (2) operates.

Let $\tilde{f}_X$ be the reweighted survey, i.e. $\tilde{f}_X(y) = f_X(y)/\theta(y)$. By construction, we have $\tilde{f}_X(y) = f_Y(y)$ for $y \geq \bar{y}$. As indicated by upward arrows on the right of figure 2, the density has been increased for $y > y^*$. Since densities must integrate to one, values for $y < y^*$ have to be lowered. The uniform reweighting below $\bar{y}$ creates the dotted blue line.

2.1.2 Choice of the Merging Point

For many countries, tax data only covers the top of the distribution. We use the term *trustable span* to name the interval over which the tax data may be considered reliable. It takes the form $[y_{\text{trust}}, +\infty[$. This interval is determined by country specific tax legislation. It relies on the portion of the distribution covered in the data (declarations) or just on the portion of the tax population that pays income tax (taxpayers).

We do not usually wish to use the tax data over the entire trustable span. First, because the beginning of the trustable span is not always sharp — the reliability of the tax data increases with income in a way that is not well-defined, therefore it is more prudent to restrict their use to the minimum that is necessary. Second, once we are past the point where there is clear evidence of a bias, we prefer to avoid distorting the survey in unnecessary ways.

We suggest a simple, data-driven way for choosing the merging point point with desirable properties. In particular, we seek to approximately preserve the continuity of the underlying density function after reweighting. We start from the typical case where $\bar{y}$ is inside the trustable span $[y_{\text{trust}}, +\infty[$. In Appendix B we consider cases where the trustable span may be too small to observe an overlap between the densities.

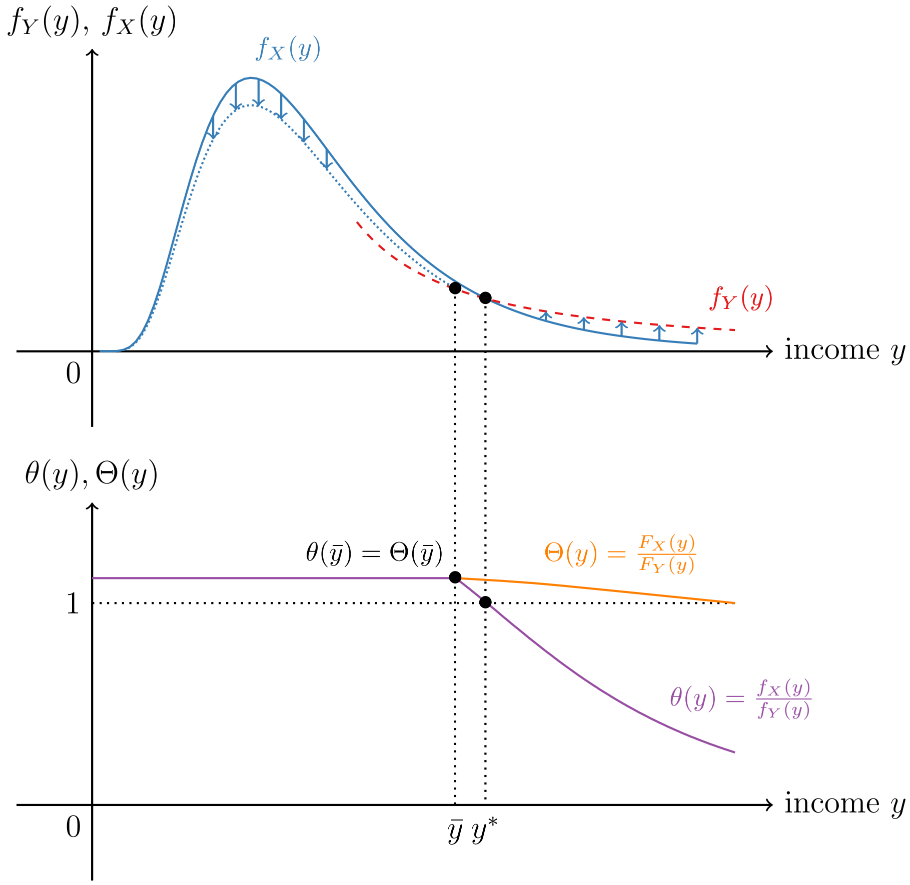
Merging Point in the Trustable Span Assume that the bias function $\theta(y)$ follows the form in (2). We introduce a second function, the cumulative bias, defined as:

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)} \tag{3}$$

In figure 3, we examine the shape of $\theta(y)$ and $\Theta(y)$ in relation to the density functions presented in figure 2. We have the relationship $\Theta(y)F_Y(y) = \int_{-\infty}^{y} \theta(t) f_Y(t)\,\mathrm{d}t$. Given (2), for $y < \bar{y}$, $\Theta(y) = \bar{\theta}$. As figure 3 shows, we should expect the merging point $\bar{y}$ to be the highest value y such that $\Theta(y) = \theta(y)$.

We can contrast this choice of merging point with the one implicitly chosen in at least some replacing approaches: the point at which the quantile functions of the survey and

Figure 3: Choice of Merging Point when $\bar{y} \geq y_{\textbf{trust}}$



the tax data cross.[5] This is equivalent to setting equal densities (i.e. $\theta(y) = 1$) until this merging point, which will in general be lower than ours. At thais point, there is a discontinuity in $\theta(y)$ which jumps above one, and then progressively decreases toward zero. As a result, the people just above the merging point are implicitly assumed to be overrepresented compared to those below, even though they are richer. This discontinuity and lack of monotonicity of θ is hard to justify, and our choice of merging point avoids it.

We can estimate both $\theta(y)$ and $\Theta(y)$ over the trustable span of the tax data. To determine the merging point in practice, we look for the moment when the empirical curves for $\Theta(y)$ and $\theta(y)$ cross, and discard the tax data below this point. This choice is the only one that can ensure that the profile of $\theta(y)$, and by extension the income density function, remains continuous.

The estimation of $\Theta(y)$ poses no difficulty as it suffices to replace the CDFs by their empirical counterpart in (3) to get the estimate $\hat{\Theta}_k$. For $\theta(y)$, however, we have to estimate densities. We define m bins using fractiles of the distribution (from 0% to 99%, then 99.1% to 99.9%, then 99.91% to 99.99% and 99.991% to 99.999%). We approximate the densities using histogram functions over these bins. This gives a first estimate for each bin that we call $(\tilde{\theta}_k)_{1 \leq k \leq m}$. The resulting estimate is fairly noisy, so we get a second,

[5] Appendix A.2 presents a theoretical comparison of both procedures.

more stable one named $(\hat{\theta}_k)_{1\leq k\leq m}$ using an antitonic (monotonically decreasing) regression (Brunk, 1955; Ayer et al., 1955; Eeden, 1958). That is, we solve:

$$\min_{\hat{\theta}_1,\dots,\hat{\theta}_m} \sum_{k=1}^{m} (\hat{\theta}_k - \tilde{\theta}_k)^2 \qquad \text{s.t.} \qquad \forall k \in \{2,\dots,m\} \quad \hat{\theta}_{k-1} \geq \hat{\theta}_k$$

We solve the problem above using the Pool Adjacent Violators Algorithm (Ayer et al., 1955). The main feature of this approach is that we force $(\hat{\theta}_k)_{1\leq k\leq m}$ to be decreasing. This turns out to be enough to smooth the estimate so that we can work with it, without the need introduce additional regularity requirements. We use as the merging point bracket the lowest value of k such that $\hat{\theta}_k < \hat{\Theta}_k$.

2.2 Multivariate Setting

The previous subsection presented the main idea of the method. But while this intuition works well in the univariate case, the introduction of other dimensions from the survey (gender, age, income composition, etc.) complicates the problem significantly. Indeed, it is not enough for the survey to be solely representative in terms of total income, we also need to preserve (or possibly enforce) representativeness in terms of these other variables. This subsection thus explains how we adapt our method to the survey-calibration framework, mainly to address two types of representational issues.[6] First, if the survey is already assumed to be representative at the aggregate level in terms of age or gender (i.e., because it has already been adjusted to fit census data), then we should aim to preserve such features. Second, when the adjustment is made using income alone (i.e. the univariate case), it corrects weights based on the observed probability of response conditional on income, ignoring interactions between total income and other characteristics, which are sometimes reported in tax data.[7] We start by presenting the theory in its general setting below, before explaining how to apply it to the problems at hand.

2.2.1 Calibration

Problem Survey calibration considers the following problem. We have a survey sample of size n. Each observation is a k-dimensional vector $\boldsymbol{x}_i = (x_{1i},\dots,x_{ki})'$. The sample can be written $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, and the corresponding survey weights are $(d_1,\dots,d_n)$. We know from a higher-quality external source the true population totals of the variables

[6] Survey calibration was introduced with the raking procedure of Deming and Stephan (1940). Deville and Särndal (1992) provided major improvements. While statistical institutes routinely use calibration methods with respect to age and gender variables, they are not yet traditionally used for income variables.

[7] For instance, if rich elderly persons are more likely to respond to surveys (say, because they have more free time) than younger rich people, then the univariate adjustment will produce an accurate income distribution without solving the over-representation of older people. A similar rationale can be applied to the issue of income composition.

$x_{1i}, \dots, x_{ki}$ as the vector $\boldsymbol{t}$. We seek a new set of weights, $(w_1, \dots, w_n)$, such that the totals in the survey match their true value, i.e. $\sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t}$.

This problem will in general have an infinity of solutions, therefore survey calibration introduces a regularization criterion to select the preferred solution out of all the different possibilities. The idea is to minimize distortions from the original survey data, so we consider:

$$\min_{w_1,\dots,w_n} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i} \qquad \text{s.t.} \qquad \sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t} \tag{4}$$

That is, we minimize the χ^2 distance between the original and the calibrated weights, under the constraint on population totals: this is called linear calibration. While alternative distances are sometimes used, linear calibration is advantageous in terms of analytical and computational tractability.

Solution Solving the problem (4) leads to:

$$\frac{w_i}{d_i} = 1 + \boldsymbol{\beta}\boldsymbol{x}_i \tag{5}$$

where $\boldsymbol{\beta}$ is a vector of Lagrange multipliers determined from the constraints as:

$$\boldsymbol{\beta} = \boldsymbol{T}^{-1}\left(\boldsymbol{t} - \sum_{i=1}^n d_i \boldsymbol{x}_i\right) \qquad \text{with} \qquad \boldsymbol{T} = \sum_{i=1}^n d_i \boldsymbol{x}_i \boldsymbol{x}_i'$$

where the matrix $\boldsymbol{T}$ is invertible as long as there are no collinear variables in the $\boldsymbol{x}_i$ (meaning neither redundancy nor incompatibility of the constraints).[8] One undesirable feature of linear calibration is that it may lead to weights below one or even negative, which prevents their interpretation as an inverse probability and is incompatible with several statistical procedures. Therefore, in practice, we enforce the constraints $w_i \geq 1$ for all i using an standard iterative method described in Singh and Mohl (1996, method 5). This is known as truncated linear calibration.

Interpretation This procedure can be interpreted in terms of a nonresponse model.[9] In this context, the survey weights are the inverse of the probability of inclusion in the survey sample. This probability of inclusion is the product of two components. The first one depends on whether a unit is selected for the survey, regardless of whether that unit accepts to answer or not. We note $D_i = 1$ if unit i is selected, and $D_i = 0$ otherwise. The value $\delta_i = 1/\mathbb{P}\{D_i = 1\}$ is called the design weight. The design weight in constructed by the survey producer and therefore known exactly. The second component depends on whether a unit contacted for the survey accepts to answer or not. We note $R_i = 1$ if unit

[8] In practice, we use the Moore–Penrose generalized inverse to circumvent the collinearity problem.

[9] For a geometric interpretation of linear calibration see Appendix C.

i accepts to participate in the survey, and $R_i = 0$ otherwise. The value $\rho_i = 1/\mathbb{P}\{R_i = 1\}$ is called the response weight. Since both D_i and R_i must be equal to 1 for a unit to be observed, the final weight is the product of these two components $\delta_i \rho_i$.

Nonresponse is unknown so it has to be estimated using certain assumptions. The simplest one is that ρ_i is the same for all units, therefore all weights are up-scaled by the same factor so that their sum matches the population of interest. More complex models use information usually available to the survey producer, that is, basic socio-demographic variables which we will write $\boldsymbol{U}_i$. The survey producer models nonresponse as a function of these variables: $\rho_i = \phi(\boldsymbol{U}_i)$. The survey producer provides weights equal to $\delta_i \phi(\boldsymbol{U}_i)$. If nonresponse is also a function of income, which is not observed by the survey producer, then the estimated nonresponse will fail to accurately reflect true nonresponse, leading to biased estimates of the income distribution. Using the tax data $\boldsymbol{Y}_i$, we can estimate a new model that takes income into account: $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)$. The final weight becomes:

$$\begin{aligned} w_i &= \frac{1}{\mathbb{P}\{D_i = 1\}} \frac{1}{\mathbb{P}\{R_i = 1\}} \\ &= \frac{1}{\mathbb{P}\{D_i = 1\}} \psi(\boldsymbol{U}_i, \boldsymbol{Y}_i) \\ &= \delta_i \phi(\boldsymbol{U}_i) \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \\ &= d_i \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \end{aligned} \tag{6}$$

Comparing equation (5) with (6), we see that the calibration problem suggests both a functional form and an estimation method for $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)/\phi(\boldsymbol{U}_i)$. This functional form assumes nonresponse profiles that are as uniform (thus non-distortive) as possible, and only modify the underlying distribution if it is necessary to do so. The preference for non-distortive functional forms can also help justify the use of a constant reweighting profile below the merging point in section 2.1.1.

Application to Income Data The calibration problem is presented so as to enforce the aggregate value of variables. In order to use it to enforce the distribution of a variable, we have to discretize this distribution. In the case of income tax data, the income distribution may be presented in various tabulated forms, and we use the generalized Pareto interpolation method of Blanchet, Fournier, and Piketty (2017) to turn it into a continuous distribution.[10] We output the distribution discretized over a narrow grid made up of all percentiles from 0% to 99%, 99.1% to 99.9%, 99.91% to 99.99% and 99.991% to 99.999%. We discard tax brackets below the merging point, whose choice is described in section 2.1.2. We then match the survey data to their corresponding tax bracket. In

[10] See `wid.world/gpinter` for an online interface and a R package to apply the method.

general, it is necessary to regroup certain tax brackets to make sure that we have at least one (and preferably more) observations in each bracket. Otherwise the calibration will not be possible. We automatically regroup brackets to have a partition of the income distribution at the top such that each bracket has at least 5 survey observations.

Assume that we eventually get m brackets, with the k-th bracket covering a fraction p_k of the population. We create dummy variables $b_1, \dots, b_m$ for each income bracket. If the total population is N and the sample-size is n, then the calibrated weights should satisfy:

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} = N p_k$$

Since these equations are expressed as totals of variables, they can directly enter the calibration problem (4). In practice, we are enforcing the income distribution through a histogram approximation of it.

The flexibility of the calibration procedure lets us put additional constraints in the calibration problem. In particular, if the survey is already assumed to be representative in terms of age or gender, then their distribution can be kept constant during the procedure. Hence we correct for the income distribution while maintaining the representativeness of the survey along the other dimensions. Additional constraints are also possible, if external information on other variables is available (see section 2.2.2).

For all the observations below the merging point, the dummy variables $b_1, \dots, b_m$ are all equal to zero, so the weight adjustment only depends on a constant and possibly other calibration variables such as age and gender, but not income. This matches the uniform adjustment profile (2) at the bottom of the distribution that we present in section 2.1.1. The calibration, by construction, avoids distorting the bottom of the distribution because it is not necessary to enforce the constraints of the calibration problem.

Our correction procedure also constrains the number of times the weights are expanded or reduced to avoid disproportionate adjustments to single observations already in the dataset. Consequently we introduce the condition that brackets with a $\theta(y)$ outside the boundary defined by $1/n \leq \theta(y) \leq n$ are automatically grouped into larger brackets. The default limit we choose is n = 5. Thus, in this case, no observation would have their weight multiplied by more than 5 times or less than 0.2 times.

2.2.2 Extensions

The calibration framework is generic enough to incorporate information into the survey in different forms. While the most standard problem is to directly correct the income distribution using the income concept of interest, more complicated settings can sometimes occur. The flexibility of the calibration framework makes it generally possible to deal with these settings without resorting to additional *ad hoc* assumptions. We discuss below three common cases.

Using Population Characteristics by Income Tax data sometimes provides information on the population characteristics by income level, typically, the gender composition. This can tell us how the interaction between income and other characteristics impacts the bias, so it can be useful to include this information in the survey.

Assume that we have m income tax brackets that contain a share $p_1, \dots, p_m$ of the overall population N. For each of them, we know the share $\boldsymbol{s} = (s_1, \dots, s_m)$ of people with a given characteristic, such as belonging to a certain gender or age group. Let v_i be the variable equal to 1 if unit i belongs to that group in the survey, and 0 otherwise. Let b_{ik} be the variable equal to 1 if unit i in the survey is in income bracket k, and 0 otherwise.

To make sure that the survey reproduces the information in the tax data, we add the following constraints to the calibration problem (4):

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} v_i = N s_k p_k$$

Using Income Composition Another source of information that is commonly available in tax data is the composition of income within brackets. Using that information is useful if we assume that the bias may be different for people that derive their income from, say, capital rather than labor.

Assume that we have m income brackets. For each of them, we know the share $\boldsymbol{s} = (s_1, \dots, s_m)$ of capital income. In the survey, total income is recorded as y_i and capital income as c_i. Let b_{ik} be a variable equal to 1 if unit i in the survey is in income bracket k. In order to enforce the constraint that the share of capital income within each bracket is the same as in the tax data, it suffice to enforce the constraints:

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} (c_i - s_k y_i) = 0$$

Indeed, the first part of the sum is $\sum_{i=1}^{n} w_i b_{ik} c_i$, which is the total capital income of the bracket. In the second part we have the total income of the bracket $\sum_{i=1}^{n} w_i b_{ik} y_i$, multiplied by the capital share s_k. This constraint can be expressed as a total of the variable $b_{ik}(c_i - s_k y_i)$. We can see that units will see their weight decrease or increase depending on whether their capital share is below or above the average of the bracket they belong to.

Using several income concepts Until now we have considered the case where the income recorded in tax data more or less matches the income concept of interest, which is the income likely to drive the bias. Yet sometimes only part of this income is recorded in the tax data. For example, in developing countries, only income from the formal sector may be recorded in the tax data, and there is a sizable informal sector only present in the survey data, which is widely spread across the distribution (as in Czajka (2017)).

In such cases, it would be problematic to directly apply the calibration method described previously. Indeed, since the adjustment factor of the weights would only depend on formal sector income, two people with the same income, one working in the formal sector and the other in the informal sector, would see their weight adjusted very differently. As a result, there would be almost no correction for the income distribution of the informal sector.

The solution to that problem is to use Deville's (2000) generalized calibration approach. The standard calibration approach formulated in (4) does not specify on what variable the weight adjustment factors should depend. In the solution of the problem, they depend directly on the variables used in the constraint. This is because the method always favors the least distorting adjustments, so it only uses the variables most directly related to the constraints.

If we have some prior knowledge of what the bias should depend on, then we can use generalized calibration to specify these variables *ex ante*. We still use $\boldsymbol{x}_i$ to denote the k calibration variables for which we know the true population totals $\boldsymbol{t}$. In the example, it would include formal sector income in addition to basic socio-demographic characteristics. We also define $\boldsymbol{z}_i$, a vector of instrumental calibration variables with the same size as $\boldsymbol{x}_i$. They may include variables in $\boldsymbol{x}_i$ (e.g. socio-demographic variables) but more importantly also some variables imperfectly correlated with the $\boldsymbol{x}_i$, in the example the sum of formal and informal sector income. We write the calibration problem as finding $w_1, \dots, w_n$ such that:

$$\sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{t} \qquad \text{and} \qquad \forall i \in \{1, \dots, n\} \qquad \frac{w_i}{d_i} = 1 + \boldsymbol{\beta} \boldsymbol{z}_i \tag{7}$$

When $\boldsymbol{x}_i = \boldsymbol{z}_i$, the problem (7) is equivalent to (4). The solution of (7) given by Deville (2000) is similar to that of (5):

$$\boldsymbol{\beta} = \boldsymbol{T}^{-1} \left(\boldsymbol{t} - \sum_{i=1}^{n} d_i \boldsymbol{x}_i \right) \qquad \text{with} \qquad \boldsymbol{T} = \sum_{i=1}^{n} d_i \boldsymbol{z}_i \boldsymbol{x}_i'$$

While we may view the standard calibration as performing a projection of the variable of interest y_i onto the calibration variables $\boldsymbol{x}_i$ using an OLS regression, the generalized calibration performs that same projection using an IV regression with $\boldsymbol{z}_i$ as a vector of instruments for $\boldsymbol{x}_i$. For this to work properly, we need $\boldsymbol{z}_i$ to be sufficiently correlated with $\boldsymbol{x}_i$, otherwise we face a weak instrument problem similar to that of traditional IV regressions (Lesage, Haziza, and D'Haultfoeuille, 2018). This is not a major concern in the example since the sum of formal and informal income is strongly correlated with formal income by construction.

2.3 Expanding the Support

After applying the methods of the previous sections, the survey should be statistically indistinguishable from the tax data. However, the precision that we get at the top of the income distribution may still be insufficient for some purposes. Indeed, the number of observations in the survey is still significantly lower than what we would get in theory from administrative microdata. The extent to which this represents a problem varies. If we use survey weights to, say, run regressions and get correct estimates of average partial effects in presence of unmodeled heterogeneity of effects (Solon, Haider, and Wooldridge, 2015), then the reweighting step is enough. But problems may arise if we wish to produce indicators of inequality, especially the ones that focus on the top of the distribution, like top income shares. The combination of a low number of observations with fat-tailed distributions can create small sample biases for the quantiles and top shares (Okolewski and Rychlik, 2001; Taleb and Douady, 2015), and skewed distributions of the sample mean (Fleming, 2007). In most cases, we would underestimate levels of inequality.

Unlike problems caused by, say, heterogeneous response rates, these biases are part of *sampling error*. They do not reflect fundamental issues with the validity of the survey, but arise purely out of its limited sample size. The calibration method (section 2.2) does, to some extent, reduce sampling error. Yet it only does so under asymptotic conditions (Deville and Särndal, 1992) that cannot hold for narrow groups at the top of the income distribution. For this reason, we prefer to consider that the role of survey calibration in our methodology is to deal with *non-sampling error*. We use a different approach to deal with sampling error.

In particular, we aim to solve the case where tax statistics include a positive number of income-declarations beyond the survey's support. That is, we need to account for individuals declaring higher income than the richest persons in the surveys, which cannot be solved by re-weighting observations. To do so, we start from the original tax tabulations, which were created from the entire population of taxpayers and should therefore be free of sampling error. We use it alongside a generalized Pareto interpolation to estimate a continuous income distribution (Blanchet, Fournier, and Piketty, 2017) that reproduces the features of the tax data with high precision. We then statistically match the information in the calibrated survey data with the tax data by preserving the rank of each observation.

More precisely: we inflate the number of data points in the survey by making k_i duplicates of each observation i. We attribute to each new observation the weight $q_i = w_i/k_i$, where w_i is the calibrated weight from the previous step. We choose $k_i = [\pi \times w_i]$ where $[x]$ is x rounded to the nearest integer. Therefore all new observations have an approximately equal weight close to $1/\pi$. The size of the new dataset, made out of the duplicated observations, can be made arbitrarily high by adjusting π, yet any linear weighted statistic will be the same over both datasets.

Let M be the number of observations in the new dataset. The weights are assumed to sum to the population size N. We will associate to each of them a small share $[0, q_{j_1}/N], [q_{j_1}/N, (q_{j_1} + q_{j_2})/N], \dots, [\sum_{k=1}^{M} q_{j_k}/N, 1]$ of the true population. If we attribute to each observation the average income of their population share in the tax data, then by construction the income distribution of the newly created survey will be the same as in the tax data. We rank observations in increasing order by income to preserve the joint distribution between income and the covariates in the survey.

From an intuitive perspective, this process can be described as replacing the income of observations beyond the merging point with the income of observations with equivalent weight and rank in the tax distribution. This step ensures that the we reproduce exactly the income distribution from tax data, preserve the survey's covariate distribution (including the household structure), and limit distortions in the relationship between income and covariates from survey data.

3 Applications

In section 3.1, we run controlled experiments with parametric distributions, using the Monte-Carlo approach, in order to assess the accuracy of estimates produced after applying both our adjustment method and the common replacing alternative found in the literature.[11] In section 3.2, we illustrate how the method operates with actual household surveys and tax statistics, applying it to data from five countries (France, U.K., Norway, Brazil and Chile). Our chosen case studies showcase the wide applicability of the method to both developed countries and less-developed ones — the latter's data tending to be more challenging.

3.1 Simulations

Our experiments start with the simulation of a 'true' distribution with several million individuals, which follow a parametric distribution. We emulate a typical tax-tabulation, which summarizes information on the richest fractiles of that same distribution in different intervals. We then draw a number of pseudo-random samples from the original distribution, simulating surveys to a given share of the population each time, which we adjust following both our method and the replacing method common in the literature.

All samples are biased by definition, including both the misreporting and non-response

[11] We choose the replacing alternative as it is the most prevalent one which utilises external data to correct surveys.

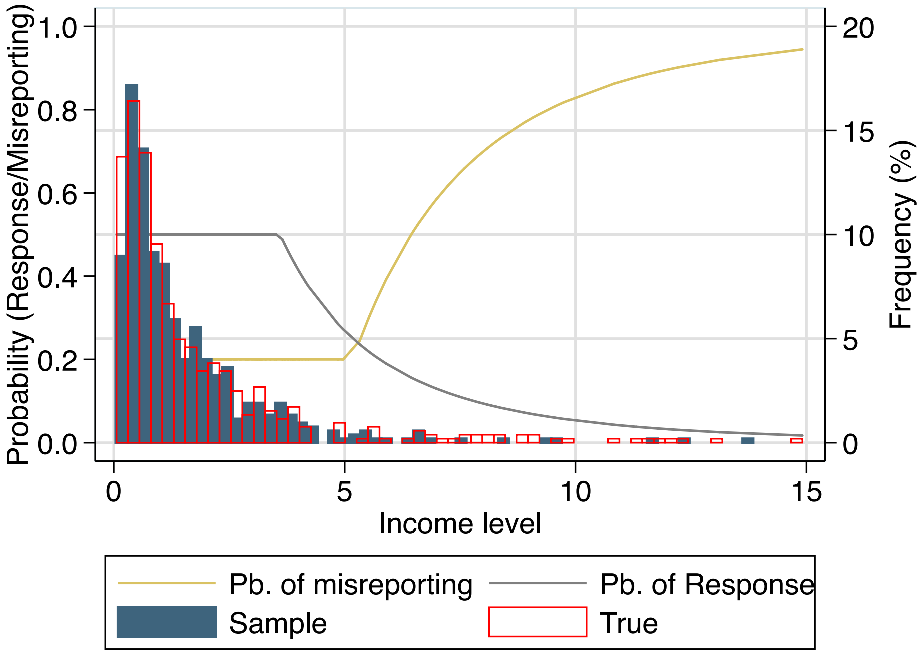
biases. The former is defined by a probability of misreporting that is assumed to be flat for most of the distribution and assumed to increase linearly with rank only at the top. The distribution of misreported income is also defined parametrically, in such a way to ensure a prevalence of under-reporting over over-reporting. Response rates are also assumed to be flat for most of the distribution and they only fall — linearly with rank — at the top. In what follows, we comment on what we consider to be our benchmark experiment. Yet, other experiments were conducted, using different sets of parameters and assumptions (Appendix D). These include alternative assumptions for each bias, variations in the replacing procedure, the size of the replaced population and the coverage of the simulated tax data. However, despite different — and sometimes extreme — assumptions, these experiments consistently demonstrate that our algorithm is adaptive and capable of implementing adjustments that push surveys closer to the true distribution when its right tail is biased.[12]

In our benchmark experiment, we study a population of 9 million individuals that are randomly drawn from a standard normal distribution. We use the exponential function of sampled values so that the distribution fits a lognormal distribution. We select a thousand random subsamples from it, whose size correspond to 1% of the total population. The expected response rate, conditional of being sampled, is 50% for most of the population; it then decreases from percentile 90 (P90) onward and tends to 0 for the richest individual, resulting in a general response rate of 47.5%. The probability of misreporting is 20% until percentile 95 (P95); it then increases, approaching 100% at the very top. The probability of misreporting is close to 22% on average and the distribution of misreported income is also a standard lognormal. In practice, all individuals in the simulated distribution have the same probability of being 'surveyed' (1 in 100), yet individuals have their own likelihood of answering the survey and if they do answer, their response can be either accurate or misreported. Hence, in such context, although the surveyed sample is 1% of the population, only close to 0.5% of the population effectively reports income. Figure 4 graphically depicts the set-up of our benchmark experiment, for one of the random samples. We apply both our adjustment method, described in the previous section, and the alternative replacing procedure. The latter corresponds to the most common form that is found in the literature, which consists in replacing the top 1% of the survey distribution with that from tax data.

Figure 5 compares the accuracy of distributional estimates that result from the raw simulated survey to those resulting from the application of both our method and replacing. It displays errors with respect to true values for a series of estimates. Kernel densities provide a visual appreciation of the set of measurements that are found for all the 1000

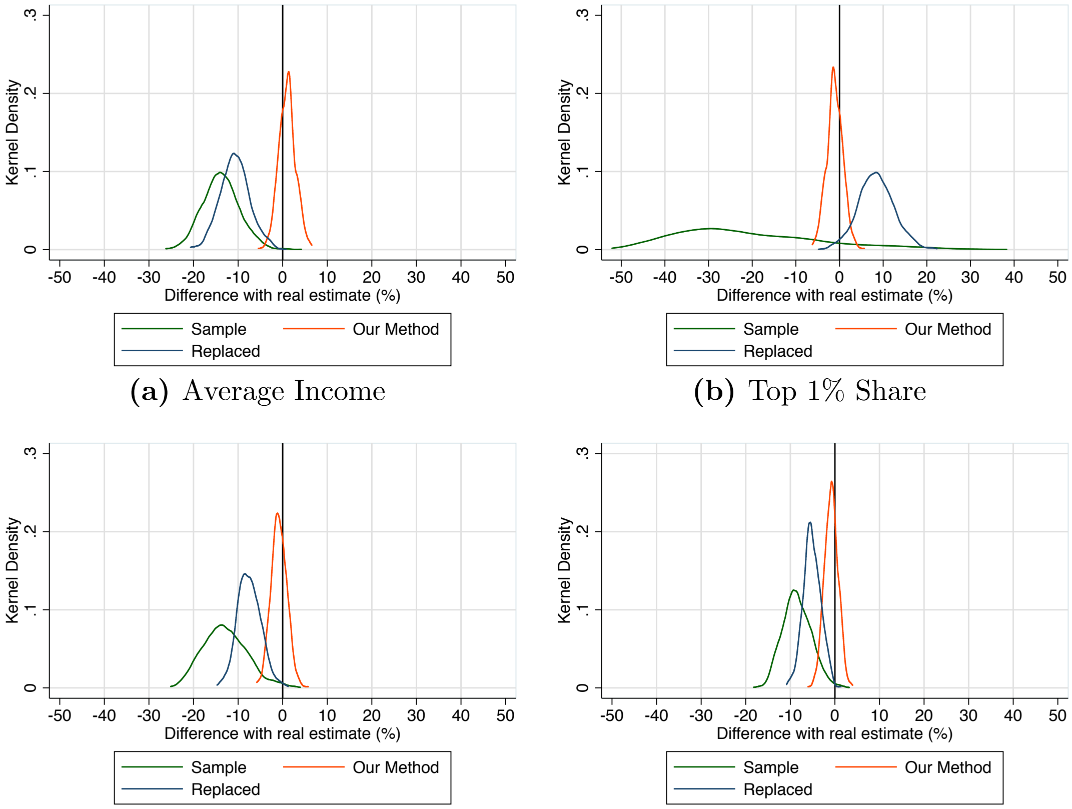
[12] All our experiments were conducted using the `bfmtoy` command that comes with the `bfmcorr` Stata package. Not only was it coded to be able to reproduce our experiments, but it also provides a tool for researchers to simulate artificial distributions and easily change all the parameters involved to test survey-adjustment methods.

Figure 4: Benchmark Experiment Set-Up



iterations. The true values are: an average close to 1.6, a Gini coefficient close to 0.52, a top 1% share close to 9.3% and a top 10% share close to 39%. It appears quite clearly that our method's estimates tend to be more accurate than others in all cases, as they are systematically closer to the true estimates and they are visibly less variant. Although both adjustment methods operate differently, in purely distributional terms they both reproduce the information of the top 1% that is found in tax data. That is, after applying the adjustment, the average income of the top percentile should be equivalent in both cases. However, the same is not necessarily true for the rest of the distribution, and thus not for average income either. Indeed, figure 5a shows that even if the average income gets closer to the true value with replacing, it still remains underestimated by a tenth of the true value on average, instead of 15% in the raw survey. The lower total income is thus what explains that in figure 5b, the top 1% shares seem to be systematically overestimated with replacing because the numerator of the top share is the same in both, but the denominator is underestimated in replacing. In the case of the top 10%, the error goes on the opposite direction (figure 5c). This is because an arbitrary correction of the top 1% is not enough to adjust for a distribution where the top decile is affected both by higher non-response and misreporting. When we focus on a synthetic indicator of inequality, such as the Gini coefficient, we find a similar hierarchy of estimates (figure 5d). It is also worth noticing that the raw sample estimate of the top 1% share is considerably less precise than any of the other estimates. This is due to a large extent to the small sample bias referred to by Taleb and Douady (2015), which is amplified by both nonresponse and misreporting.

Figure 5: Benchmark Experiment Results



(a) Average Income

(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

3.2 Real Data

Our method can be replicated for all countries with the requisite data, namely, survey micro-data covering the entire population and tax data covering at least a fraction of it.[13] We experiment with five real distributions, three European countries — making use of the common survey framework applied to them — and two less-developed Latin American countries, which can be imagined to present more of a challenge regarding data quality and scope.

3.2.1 Definitions and Data

A crucial preliminary step in the analysis is to reconcile both the definition of income and the unit of observation in national surveys with the ones that are used in tax declarations. Our algorithm functions under the supposition that these definitions have been made consistent in the two datasets by researchers. For France, Norway and the U.K., our analysis broadly covers the years 2004-2014. For Brazil, we cover 2007-2015 and for Chile we include the years 2009, 2011, 2013 and 2015. Consistent with the calibration procedure

[13] In the case where users only avail of tabulated survey data our method will still perform the correction, using percentile bracket-information from the synthetic micro-files produced by the *gpinter* program (see `wid.world/gpinter`).

explained in section 2.2 we preserve the representativeness — not only of income — but also of other variables for which the survey is assumed to be already representative, namely gender and age variables.[14]

Income Concept Given that we seek to approximate the benchmark distribution, our method is by definition anchored to the income concept that is used in the tax tabulations, which in all of our case studies is pre-tax income. However, countries differ in the income concept included in their respective surveys. Brazil's PNAD reports individuals' pre-tax income, while Chile's CASEN gives after-tax income. Thus, for Chile we require to impute taxes paid to arrive at gross income. Appendix E.1 explains how this imputation is done, as well as the construction of income units in surveys and their approximation with tax data in all countries. For European countries we work with gross incomes (pre-tax and employee contributions deducted at source) from the SILC database.[15] France is the exception since incomes reported in the tax files are net of employee contributions deducted at source. For this reason we use the concept of net income in SILC for France that deducts social contributions levied at source.

The tax data we use is presented in tabulated form, containing at the very least, the number of income recipients by given income intervals and the total or average income declared within each interval. For France, we use the tabulated tax statistics produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the ministry of finance's tax microdata. The data cover all tax units (*foyers fiscaux* – singles or married couples), with about 50% of these subject to positive income tax. For the U.K. we use tax tabulations from the Survey of Personal Incomes (SPI) available from the Office of National Statistics. The underlying data covers about 80-90% of tax units (individuals) aged 15+, with about 60% subject to positive income tax. For Norway, we use tax data from Statistics Norway, which covers 100% of tax units (individuals) aged 17 and over, of which roughly 90% have positive income tax payments. For Brazil we use tax data from the personal income tax declarations (DIPRF tables), which covers about 20% of the adult population, with about 14% subject to the personal income tax on taxable income. For Chile we exploit income tax data from the *Global Complementario* and *Impesto Único de Segunda Categoría* (IGC and IUSC tabulations), which covers 70% of the adult population, with about 15-20% subject to the personal income tax on taxable income. For all cases, we take the proportion of population with positive tax payments as the "trustable span" of the tax data. The intuition for this choice is that individuals subject to income tax are less likely to misreport their income compared individuals who declare but are under the tax-paying threshold.

[14] We do so using the command `holdmargins`. See the instructions to `bfmcorr` in Stata.

[15] In all countries, gross income is after employer social contributions.

Observational Unit Concerning the observational units, we anchor the definition to the official tax unit in each country. In all of our country cases declarations are made at the individual level, except in France and Brazil, where declarations are jointly filed by married couples (in the case of the latter, at their own discretion). However, for France we make use of the individually-declared fiscal income files produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the administratice microdata. Therefore for all countries, we define the unit of analysis across datasets as individual income, including for Brazil, where the joint income of couples is equally split between the component members (see Appendix E.1 and Morgan (2018) for further details).

3.2.2 Empirical Bias and Corrected Population

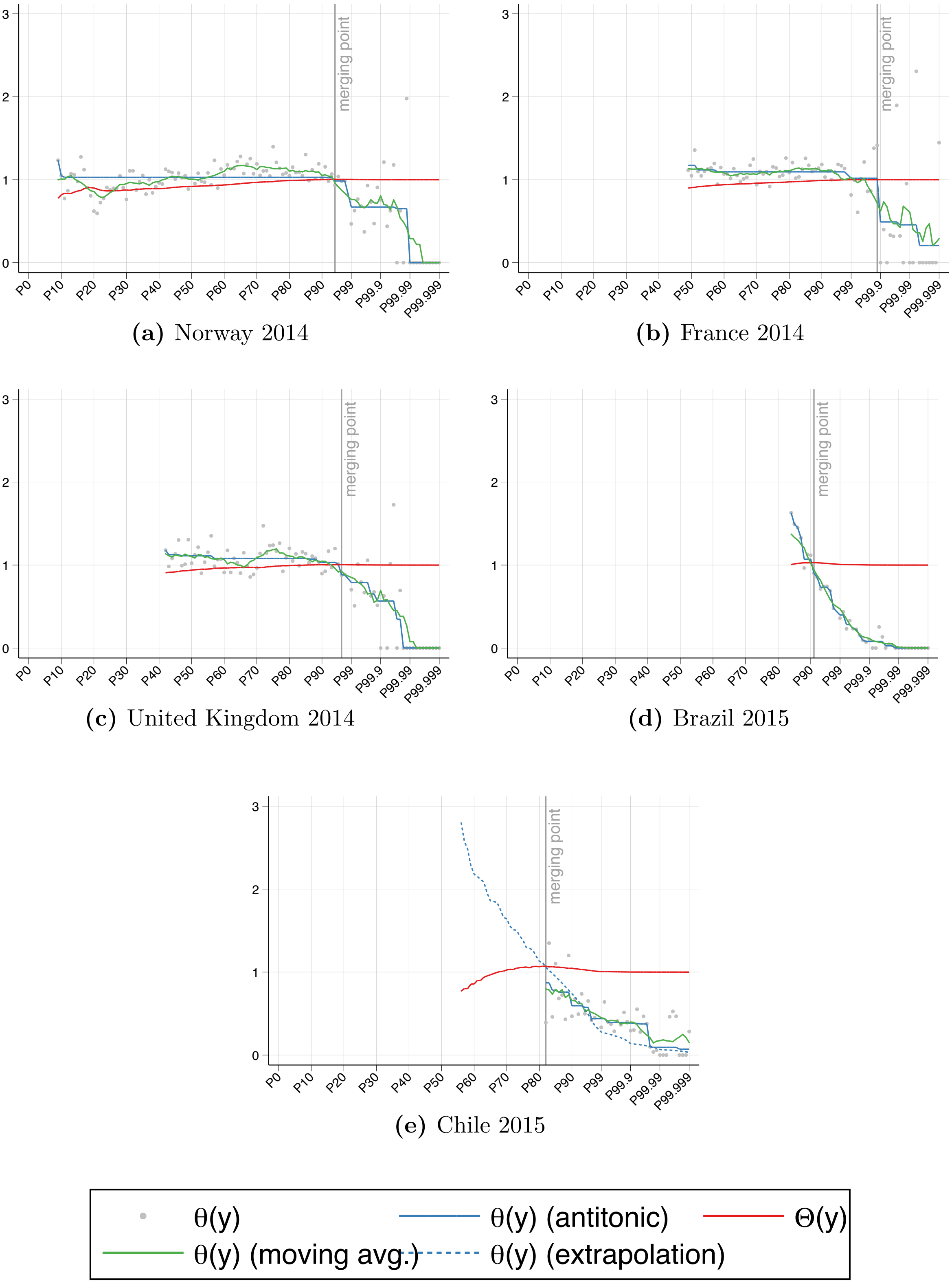
The Shape of the Bias Our method finds the merging point between surveys and tax data by comparing the population densities at specified income levels, as explained in section 2.1.2. To do so we first interpolate the fiscal incomes in the tabulation using the generalized Pareto interpolation (`https://wid.world/gpinter`) developed by Blanchet, Fournier, and Piketty (2017), which allows for the expansion of the tabulated income values into 127 intervals.[16] Using the thresholds of these intervals we can construct our key statistics: the frequency ($\theta(y)$) and cumulative frequency ($\Theta(y)$) of individuals along the income distribution.

Figure 6 presents depictions of the shape of the empirical bias within the tax data's "trustable span" for all countries for the latest available year. First of all, the shape of the bias we measure from the data is very similar to the one we present in the theoretical formalization, depicted in Figures 3 and 12. In particular, we always observe a convex shape in the top tail, to the right of the merging point. It thus appears that surveys tend to increasingly underestimate the frequency of incomes beyond a certain point in the distribution.

For the more developed countries (Norway, France and the United Kingdom), the shape of the empirical bias $\theta(y)$ can be observed for a more comprehensive share of the population, due to the greater population coverage in tax data. This enables us to empirically test our theoretical expectations on the specific behavior of the bias to the left of the merging point. We indeed observe on the left side of Figures 6a 6b 6c, a general stability in the relative rate of response, with averages trending above 1. The extent and quality of tax data below the merging point in less-developed countries is such that we

[16] These comprise of 100 percentiles from P0 to P100, where the top percentile (P99–100) is split into 10 deciles (P99.0, P99.1, . . . , P99.9-100), the top decile of the top percentile (P99.9–100) being split into ten deciles itself (P99.90, P99.91, . . . , P99.99-100), and so forth until P99.999. This interpolation technique, contrary to the standard Pareto interpolation, allows us to recover the income distribution without the need for parametric approximations. It estimates a full set of Pareto coefficients by using a given number of empirical thresholds provided by tabulated data. As such the Pareto distribution is given a flexible form, which overcomes the constancy condition of standard power laws, and produces smoother and more precise estimates of the distribution.

Figure 6: Merging Point in 5 Countries, Latest year



(a) Norway 2014

(b) France 2014

(c) United Kingdom 2014

(d) Brazil 2015

(e) Chile 2015

Notes: the figures depict the estimated bias in the survey relative to the tax data. Grey dots are, for each quantile of the fiscal income distribution, the ratio of income density in the survey over that of tax data. The green line is the centered average of $\theta(y)$ at each quantile and eight neighboring estimates. The blue line is the result of an *antitonic* regression applied to $\theta(y)$. It is constrained to be decreasing as it is used to find a single merging point. The blue dotted line, which only appears in figure 6e,is an extrapolation of the trend described by $\theta(y)$ based on a *ridge* regression (see Appendix B). The red line is the ratio of the cumulative densities. For details refer to section 2.1.2.

cannot observe the same trends.[17]

The merging points found by our algorithm vary by country and by year, again revealing differences in data quality and coverage between them. The Chilean case (Figure 6e) provides an example of our program needing to extrapolate the shape of the bias to find the merging point (see Appendix B for more details of this procedure). For this case we rely on parameters observed for Brazil (specifically, values for the elasticity of response to income) above its trustable span as inputs for the Chilean extrapolation.[18] The fit with the existing data seems to work quite well. The empirical bias that is observed in previous years for all countries is presented in Appendix E.2.1.

Corrected Population Our program then adjusts the individual weights of survey respondents in line with information from tax data, as described in section 2.1. We provide some summary statistics of the population we correct in Table 1, again using the last available year for each country as illustrations (see Appendix E.2.2 for other years). According to the comparison of surveys with tax records, a varying proportion of the total population is adjusted at the top of the survey distribution in each country (column [4] of Table 1), ranging from 5.9% in Chile to 0.05% in France for their most recent years.[19] This is derived from the comparison of the share of the population above the merging point in the two datasets. Since we use incomes in tax data as the benchmark for the top of the distribution, the share of the population above the merging point in tax data is directly related to the merging point. The share of the population above this point in surveys is always lower, indicating under-coverage of top incomes. But in both cases, the overwhelming majority of the adjustment (over 90%) can be seen to come from inside the survey support, rather than outside the survey's original support, suggesting that non-sampling issues related to heterogeneous response rates matter more than problems related to under-sampling for the size of the corrected population.

In general, this step of the algorithm is a useful guide to assess the income coverage of surveys across countries. For instance, it appears on the basis of our analysis that the Brazilian surveys do a better job at capturing gross income, given the lower share of the underrepresented population, than the Chilean household surveys. Moreover, comparing France and the UK, it seems that sampling error is greater in the UK surveys, given the higher share of the population beyond the survey's maximum income that needs to

[17] Tax enforcement issues affecting this portion of the distribution could be at play here, as well as the sharp difference in incomes between the top and the rest in these countries leading to higher inequality levels than developed countries.

[18] The value of the baseline elasticity of response to income, γ_1^*, extracted from the Brazilian data is -0.99.

[19] Across years there is less variation in this share, with Norway and particularly France being relative exceptions. In the French case, we believe the significant break in the series is due to the use of register data in SILC alongside the household survey from 2008. Despite the SILC survey making use of register data, the goal is not to over-sample the top of the distribution, but rather to improve the precision of responses.

Table 1: Structure of Corrected Population: Latest Year

Country	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
Chile	17.0%	11.1%	5.9%	99.99%	0.01%
Brazil	8.0%	5.3%	2.7%	99.0%	1.0%
UK	3.0%	2.5%	0.5%	93.6%	6.4%
Norway	5.0%	4.6%	0.4%	96.0%	4.0%
France	0.1%	0.05%	0.05%	99.0%	1.0%

Notes: The table orders countries by the size of the corrected population. Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum). Brazil and Chile refer to 2015, while all the European countries refer to 2014.

be added. Non-sampling error itself is greatest in Chile, derived from the share of the corrected population found inside the survey's support.

3.2.3 Results

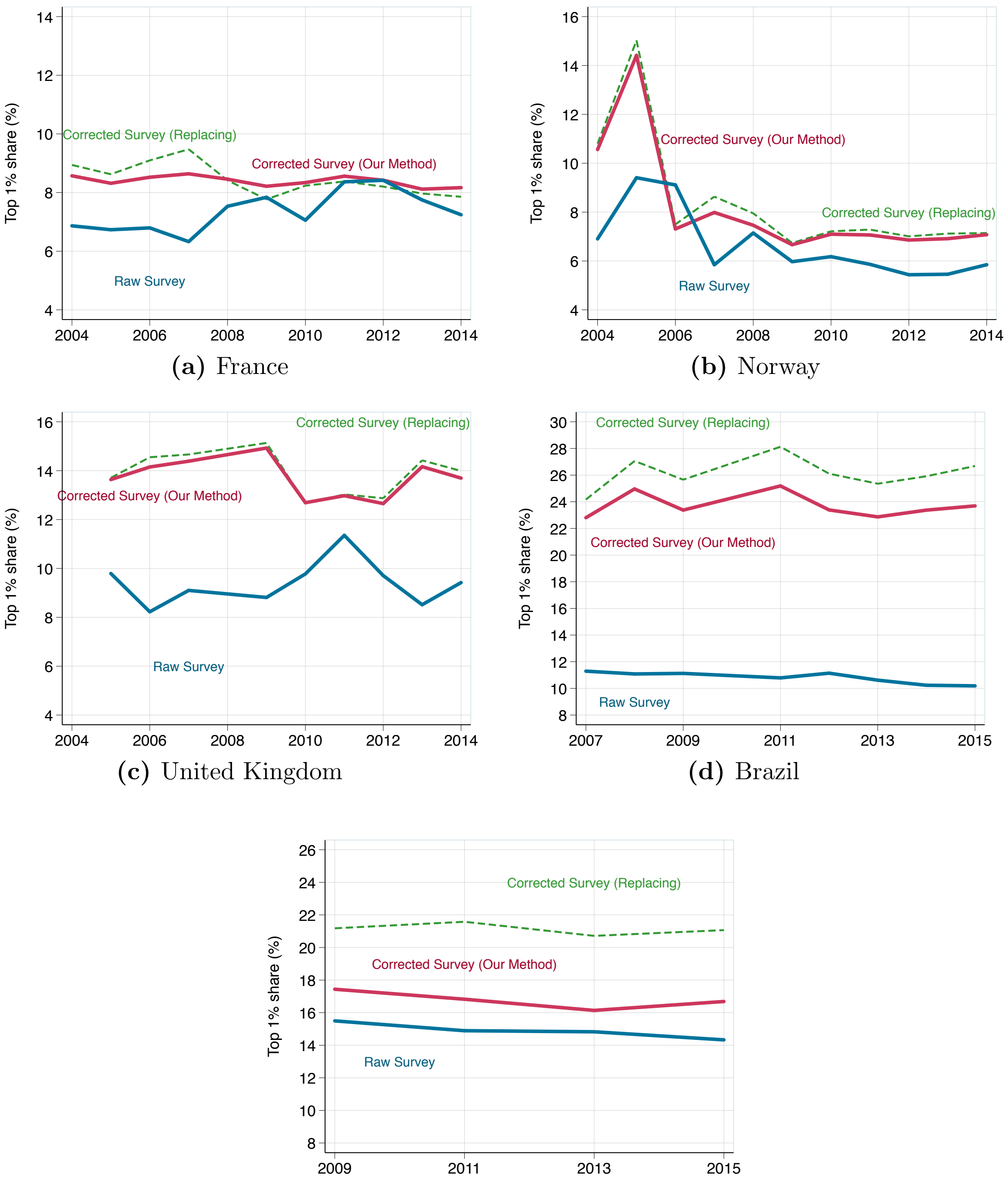
We now turn to unveil how different our merged distributions are with respect to the raw survey distributions and other corrected distributions based on the most common replacing method found in the literature that utilises external data. The latter corresponds to the procedure reproduced in the simulation in section 3.1, whereby the top 1% of the survey distribution is directly replaced by the top 1% of the tax distribution. We present results on top 1% income shares, Gini coefficients and average incomes.[20]

Top Income Shares In line with the improved income coverage that are method produces — by more accurately including upper incomes — estimates of the income concentrated at the top of the distribution are revised upwards in all countries. The size of the adjustment, however, varies by country. Figure 7 depicts this for the Top 1% share.[21] Brazil has the most extensive correction, with a top 1% share that increases by about 10 percentage points every year (Figure 7d). Conversely, France and Norway experience relatively smaller adjustments, starting from relatively lower levels of inequality. In addition, Brazil offers the clearest illustration of the distinct trends in inequality that can emerge after making a correction to the survey's representation of income. While

[20] Appendix E.3 presents results for other income groups in the distribution.

[21] The one exception to this upward correction is Norway in 2006 (see Figure 7b). However, this is likely due to a change in the local tax legislation affecting the distribution of business profits (Alstadsæter et al., 2016), as we explain in the text.

Figure 7: Top 1% Shares Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

the raw survey depicts falling top income shares, the corrected survey distribution shows slightly increasing top shares. Distinct trends are also visible, albeit for shorter periods of time, in the other countries.

The quality of both surveys and tax statistics may have a substantial impact on the size of the adjustment. For instance, in the case of France, several improvements were made to the survey's methodology starting in 2008. In particular, the matching of individuals across survey and register data allowed for the use of tax data as an external source to assess individual income without recourse to self-reporting. This testifies to the more accurate reporting of income in subsequent years, even though the gap in shares does not fully disappear in all years. Although this incorporation of register data remedies problems of misreporting and item non-response (failure to answer certain income questions), it cannot itself get around unit non-response (failure to answer the entire survey), or issues of under-sampling.

Moreover, when we compare the size of the adjustment in Chile and Brazil (Figures 7d and 7e respectively), two highly unequal Latin American countries, the latter has a considerably higher adjustment. One of the reasons that could be behind this phenomenon is the fact that capital income, especially dividends, is better recorded in Brazilian tax statistics. Indeed, the Brazilian tax agency has relatively good means to verify the accuracy of capital income declarations (Morgan, 2018), while Chilean tax authorities are generally constrained by bank secrecy (Fairfield and Jorratt De Luis, 2016). In this case, the limited quality of Chilean tax statistics explains the smaller correction.[22]

Following the same rationale, the inclusion or exclusion of some types of income in a given dataset can also affect the size of the correction. In the case of Norway, tax incentives started favoring the retention of corporate profits inside corporations after 2005, with the creation of a permanent dividends tax in 2006. This resulted in less dividend payments, and thus less income to be registered as personal income in tax data. The reform also gave strong incentives for higher-than-normal dividend payouts in 2005, which contributed to the sharp increase in top shares observed for this year (Aaberge and Atkinson, 2010; Alstadsæter et al., 2016). In Figure 7b, it can be clearly perceived that the size of the adjustment appears to drop durably after this year. Additionally, it should be noticed that the Norwegian survey appears to be rather insensitive to this change, implying that dividends where badly represented before 2005. Other potential explanations for the difference in the size of adjustments could have to do with behavioural differences between populations across countries related to response rates and reporting accuracy.

[22] There is also a considerable difference between these countries' tax systems and their respective incentives. In Chile most dividends received by individuals are taxed, while in Brazil they are not. This, in addition to the fact that Chilean realized capital gains are mostly un-taxed, provokes incentives towards the artificial retention of profits that are not as present in Brazil. This is why, in Chile, the imputation of undistributed profits to the distribution of personal income appears to be necessary when making international comparisons (Flores et al., forthcoming).

The extent of the adjustment, by definition, depends directly on the shape of the bias that is observed in Figure 6. Both the steepness of $\theta(y)$, when it is to the right side of the merging point, and the size of the corrected population (Column 4 in Table 1) are decisive factors for the size of such an increase. Another way to think about the size of the corrected population is to look at the size of the area between $\theta(y)$ and 1, to the right side of the merging point.

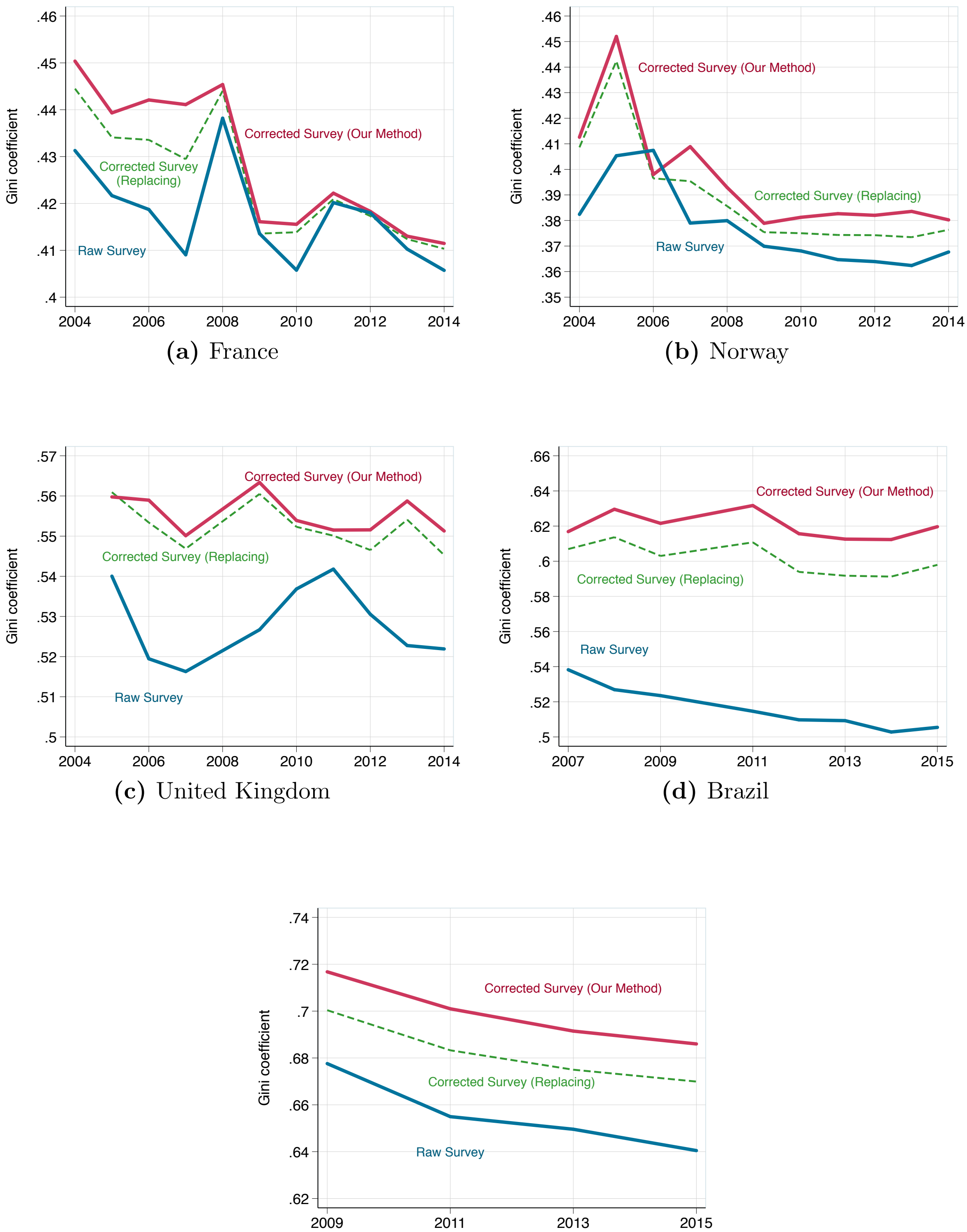
Finally, comparing between correction methods, we can observe — in line with our simulations — that the top 1% share is generally higher in the replacing scenario than in our method due to the fact that while the level of numerator incomes is equivalent in both settings, average incomes (the denominator) is underestimated in the former scenario, as we show further below.

Gini Coefficients Figure 8 shows the time series of the Gini coefficients before and after the correction for all available years. Overall, we find a similar hierarchy of estimates, mirroring our simulations in the previous section — inequality is corrected upwards, more so in countries whose raw survey is not already matched with any administrative source, and to different degrees depending on the year, thus producing distinct trends. This is further evidence that surveys need to be adjusted if they are to better represent the income distribution, in the same manner as they are currently calibrated to better represent the distribution of various demographic variables.

Again consistent with our simulations, the replacing procedure seems to undershoot inequality levels compared to our method, which more accurately accounts for higher non-response and misreporting at the top. An arbitrary correction of the top 1% is not enough to adjust the under-coverage of income coming from these errors. This is especially the case where the corrected population is larger than the arbitrarily chosen fractile, such as in Brazil and Chile (see Figure 6 and Table 1).

Average Incomes As alluded to before, the average income of the top percentile using both correction methods is the same, which is higher than the level observed in the raw surveys. However, the crucial difference between the two methods is that the average incomes for the other groups in the population are not equal. In our method, the weight of persons with lower incomes are reduced, while the replacing method keeps the same average income for the bottom income groups. This subsequently produces differences in the overall average income of the population in both cases. Figure 9 depicts that our method increases the average income in the surveys in all countries, although with highly varying degrees of magnitude. In the lower-income countries, which have the highest corrections — Brazil and Chile — average incomes increase broadly by 30-50%, with the gap increasing over time. The higher income countries in Europe experience lower corrections to their average incomes, with the orders of magnitude between them

Figure 8: Gini Coefficients, Before and After Correction



(a) France

(b) Norway

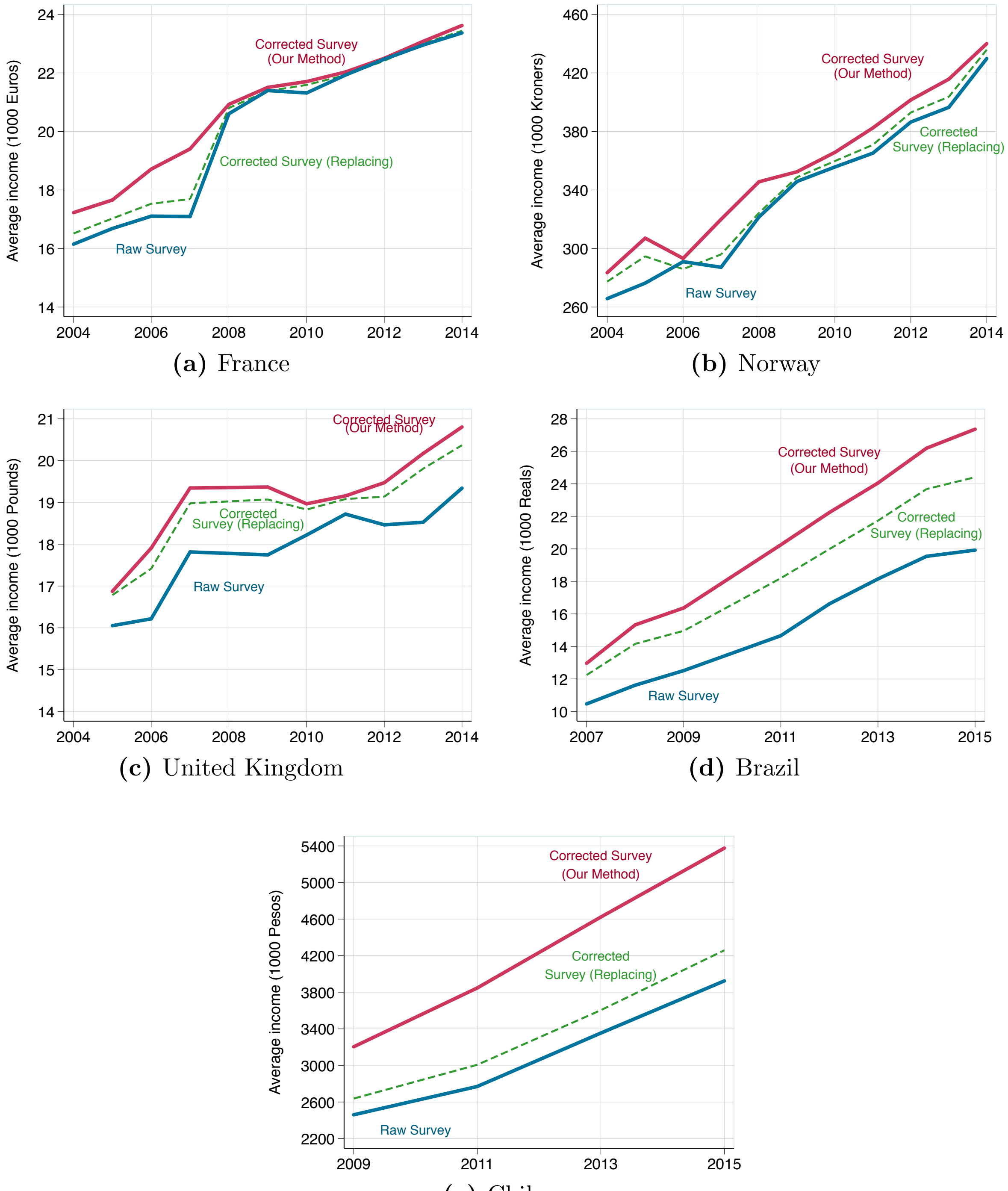
(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

Figure 9: Average Incomes Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution. Average incomes are rescaled accordingly.

reproducing the rank of countries by size of correction in Table 1 — the U.K. experiences a larger correction than Norway, which experiences a larger correction than France. Visibly, in Figure 9a the gap between the average in the raw data and corrected data is reduced from 2008 on-wards on account of the reduction in the size of the survey bias coming from the methodological novelties (see Table E.4 for further details).

The result for the replacing method goes in line with expectations. It is higher than the raw survey result, as more income is given to the top of the distribution, but it is also consistently lower than the average our method produces, since it does not reduce the weight of individuals with lower incomes. This is an inconsistency coming from its own rationale, as explained in Appendix A.2, which our method explicitly overcomes.

The relative underestimation of incomes is further evident in Figure E.6, which shows income coverage across datasets in the two countries with the largest corrected populations. The corrected survey income total from our method, which is already higher than the total from replacing as Figure 9 testifies, is closer to a broadly equivalent income total from national accounts in both Brazil and Chile.

Conclusion

The main objective of this paper is to provide a rigorous methodological tool that enables researchers to combine income or wealth surveys with administrative data in a simple and consistent manner. We present a new methodology on the combination of such sources, which incorporates a clearer formal understanding of the potential biases at play and a solution to remedy them. The result of our calibration-inspired approach, we argue, should be a more representative dataset that can serve as a basis to study the different dimensions of social inequality. Our algorithm is built in such way that it automatically generates, from raw surveys and tax data, an adjusted micro-dataset including new modified weights and new observations, while preserving the consistency of other pre-existing socio-demographic variables, at both the individual and aggregate level.

Our paper can thus be viewed as an attempt to improve survey representativeness by taking the income (or wealth) distribution into account. While it is common to adjust survey weights in accordance to external information on the distribution of basic socio-demographic variables, our paper motivates the use of auxiliary administrative data sources on the distribution of income to further improve the representativeness of the population.

Our procedure has several advantages compared to available options to correct surveys. First, it is based on a solid and intuitive theoretical framework. Second, our method avoids *a priori* assumptions on the size of the population to be corrected. Instead, it offers a clear procedure to find the merging point between datasets non-arbitrarily. Third, the algorithm can be applied to a wide variety of countries, both developed and less developed, since it accounts for different levels of data coverage. Fourth, our method respects original individual self-reported profiles and socio-demographic totals for variables other than income. We thus preserve the internal consistency of surveys, while better approximating the external consistency of its income distribution. Although we preserve socio-demographic totals for variables other than income, our method allows for their conditional distribution to vary upon the addition of new income information. However, our method also accommodates the input of distributional information of other variables (age, sex, income type, etc.) if they are available in the tax data. As such, one may also calibrate and correct the survey on covariates of income, in addition to income itself, if reliable statistics exist on their interaction. Finally, it should be clear that this method can serve multiple research objectives — from single-country and cross-country empirical analyses using income statistics as well as their covariates, to fiscal incidence analysis.

To the extent of harmonizing our correction procedure among different countries, we stress the importance of analyzing the underlying data in each case. For this, our method provides useful tools to researchers wishing to assess the population coverage of surveys conditional on income. Figure 6 and Table 1 are examples of the type of information directly computed by our algorithm, which is made available to users as a program on Stata. With standard survey and tax data at hand, researchers can perform our correction procedure with relative ease, as long as the income concepts are/can be made comparable across the datasets.

Our practical applications show the accuracy and scope of the method. The Monte Carlo simulations reveal that our method produces results — on average incomes and inequality indicators — that are closer to values from the true distribution with lower variance, compared to the drawn sample and the common "replacing" alternative employed in the literature. This is because the structure of our method's correction takes seriously the nature of the potential biases at play. Finally, when applied to real data, our approach is shown to be robust to different contexts, with the size of adjustments depending on data quality and inequality levels by country. The wider the gap between survey and administrative data and higher the level of inequality in the country, the greater the correction is likely to be. Our empirical results are consistent with experiments we run with simulated data. Overall, we claim that our method is accurate, robust and pragmatic in unifying the strengths of separate datasets on the distribution of income/wealth and their covariates into one source of information.

Appendix A Formal Biases and Adjustments

Our adjustment procedure is based on the interpretation of the whole difference between tax and survey densities as being due solely to nonresponse. However, the misreporting of income by survey respondents may also produce discrepancies. Misreporting tends to be negatively correlated with income.[23] That is, on average, the poor are more likely to overreport their true income while the rich tend to underreport. It is thus fair to ask: what would be the consequences of such behavior in our analytical framework? And how do replacing methods — which aim to adjust for underreporting at the top — compare to reweighting?

A.1 Double-Biased Density Functions

To define the misreporting bias, let $f_Y(y)$ be the true income distribution, $f_M(y)$ the distribution of misreported income, $p(y)$ the probability of misreporting for a given level of income, conditional on response, and $\bar{p}$ its average. Then we define f_Z as the income distribution of a sample that is drawn from $f_Y(y)$, including both the nonresponse and misreporting biases (the former is defined in equation (1)) :

$$f_Z(y) = f_Y(y)\theta(y)(1 - p(y)) + f_M(y)\bar{p} \tag{8}$$

The left side of the sum stands for those who report income correctly with a given (relative) probability of response that is defined in equation (1), i.e. $\theta(y)$. The right side of the sum accounts for those declaring misreported income equal to y, given that they respond. In this situation, the over- or under-estimation of f_Z with respect to the true distribution f_Y can be formulated as the ratio of the two distributions:

$$\frac{f_Z(y)}{f_Y(y)} = \theta(y)(1 - p(y)) + \frac{f_M(y)}{f_Y(y)}\bar{p} \tag{9}$$

If the ratio is higher than 1, the density is overestimated. If it is lower than 1, it is underestimated. Naturally, the shape of such bias depends on the characteristics of each of the variables at play. Following the empirical literature, it is reasonable to define the probability of misreporting as being higher in both ends of the distribution and relatively stable in the middle. However, explicit information on the shape of the misreported-income distribution is rare, since it relies on having individually-linked survey and register micro-data. In order to better understand the potential impact of assumptions on its shape, it can be useful to analyze a simplified situation where misreporting operates in

[23] See Bound and Krueger (1991), Bollinger (1998), Pedace and Bates (2000), Cristia and Schwabish (2009), and Abowd and Stinson (2013) for studies on the United States and Angel, Heuberger, and Lamei (2017) and Paulus (2015) for studies on Austria and Estonia respectively.

isolation. In that case we have:

$$\frac{f_Z(y)}{f_Y(y)} = 1 - p(y) + \frac{f_M(y)}{f_Y(y)}\bar{p} \tag{10}$$

If misreported income follows the same distribution as true income, that is $f_M(y) = f_Y(y)$, then densities are underestimated where the probability of misreporting is higher than its average ($p(y) > \bar{p}$). Symetrically, densities are overestimated where the same probability is lower than its average ($p(y) < \bar{p}$). Of course, it may seem odd to assume that misreported income is distributed exactly as true income. However, we consider this to be a useful simplification which helps to convey that both the nonresponse and misreporting biases can have a similar impact and that we are unable to tell them apart *ex-post*. Indeed, both biases, either working alone or together, can perfectly describe a profile as the one in Figure 3. If $f_M \neq f_Y$, we can still get a similar result under some circumstances. For instance, if both densities are of the same type but defined by different parameters (e.g. if both are log-normal with a different mean and standard error) — which does not seem to be a strong assumption — the ratio of the sample to true distribution would likely have a form similar to Figure 3 but with strong or slight perturbations near the mode of each distribution (assuming the true and misreported income-densities are unimodal).

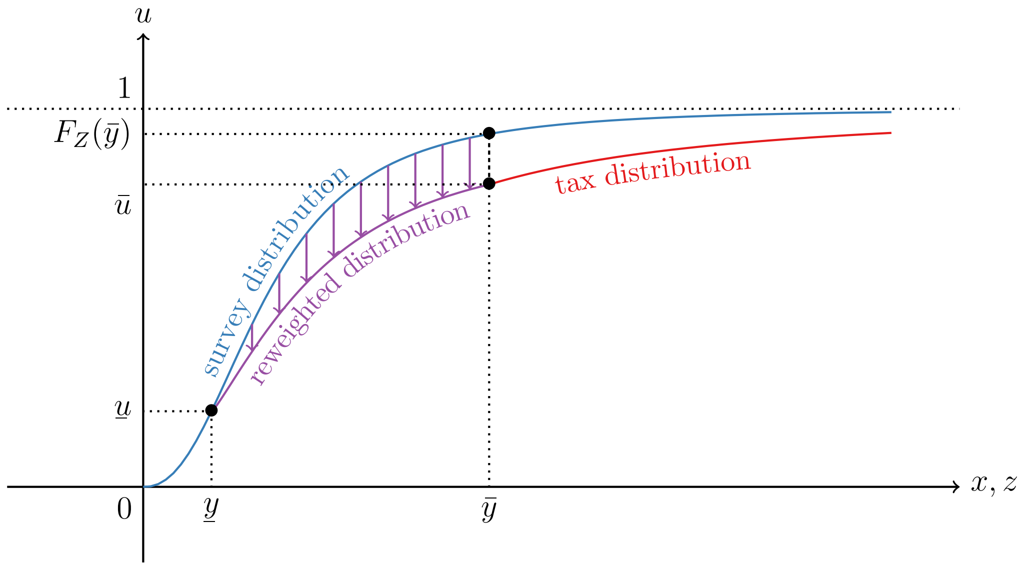
When we study the ratio of income distributions from actual tax and survey data — in Section 3.2.2 — the empirical estimate of the θ coefficient should be capturing the effect of both these biases. Figure 6 shows that estimates for countries with comprehensive tax coverage (e.g. Norway, France and the UK) depict rather flat shapes through most of the distribution and only fall closer to the right tail. Such a shape implies that, if the misreporting bias is present in the survey, the differences between f_M and f_Y are not big enough to cause perturbations that are easily distiguishable from noise in the θ coefficient. In any case, to our knowledge, it is not possible to measure the relative size of both the nonresponse and misreporting biases without access to individually-matched micro datasets.

A.2 Adjustment Methods: Reweighting vs. Replacing

In practice, researchers face the following problem while combining survey and tax data: on one side, survey data supposedly covers the whole population but fails to properly capture the top tail of the income distribution. On the other side, they have a tax data distribution which is assumed to be accurate, at least at the top.[24]

[24] The issues of tax avoidance and evasion are issues of underreporting, but are more difficult to remedy without access to third-party/offshore data. Therefore it is useful to think of tax data, at least above a certain top threshold, as being an accurate lower bound for incomes.

Figure 10: Correcting for nonresponse by reweighting



Reweighting The reweighting solution in this scenario can be represented as in figure 10, which displays the Cumulative Distribution Functions (CDF) of the survey, tax data and "reweighted" distributions. The tax data start at the value $\bar{y}$, which correspond to the population fractile $\bar{u}$. If nonresponse is higher at the top, the corresponding fractile in the survey — $F_Z(\bar{y})$ — will be higher as shown. We can also define a low income level $\underline{y}$ with corresponding fractile $\underline{u}$ below which we do not want to alter the survey (e.g. the national poverty line). If there is no such concern, then we can set $\underline{u} = 0$ and $\underline{y} = -\infty$.

Replacing While the reweighting method adjusts the weight of survey observations, replacing methods adjust their value. The usual rationale behind replacing methods is different. It accounts for the discrepancy between the survey and the tax data by assuming that people misreport their income, rather than by assuming that people refuse to answer the survey or its income-related questions.

Either problem may happen in reality, and mathematically it is not possible to disentangle them without linking tax and survey data directly (see Appendix A.1). But the case for reweighting relies in part on the fact that even if misreporting is the problem, it is unclear that pure replacing does a better job of solving it than reweighting. To convey this, let us start with a formulation of the misreporting problem. We have the following relationship between two random variables Y and Z, which represent true income and misreported income respectively:

$$Z = Y\Lambda$$

where Λ is a random variable that may depend on Y. We call $1 - \Lambda$ the rate of underreporting. In this setting, the PDF of Z will depend on the joint PDF of Y

and Λ:

$$f_Z(z) = \int_{-\infty}^{+\infty} \frac{1}{|\lambda|} f_{Y\Lambda}[z/\lambda, \lambda] \, \mathrm{d}\lambda$$

The expression above raises some major tractability issues. In particular, it is not possible to recover $f_{Y\Lambda}$ from the knowledge of f_Y and f_Z separately, so Λ may only be estimated when we can link misreported income and its covariates (i.e. Z and X) at the individual level, which is not common in practice. Otherwise, there will infinitely many Λ that satisfy the problem. For these reasons, previous researchers working with replacing methods have made some very strong (even if implicit) assumptions, which we make explicit below.

Assumption 1. The rate of underreporting is a deterministic function of the rank in Y: $\Lambda = \lambda(F_Y(Y))$.

Assumption 2. The rank is the same in the true income distribution and in the survey income distribution: $F_Y(Y) = F_Z(Z) = U$.

These assumptions on λ are very strong and unavoidable. It is not possible to interpret $\lambda(u)$ as an average underreporting given rank u (i.e. $\lambda(u) = \mathbb{E}[\Lambda | F_Y(Y) = u]))$, because f_Z depends on the entire joint distribution of (Y, Λ). Using these assumptions, estimating the underreporting function λ is very simple. Indeed, since misreporting leaves the rank unchanged, we have:

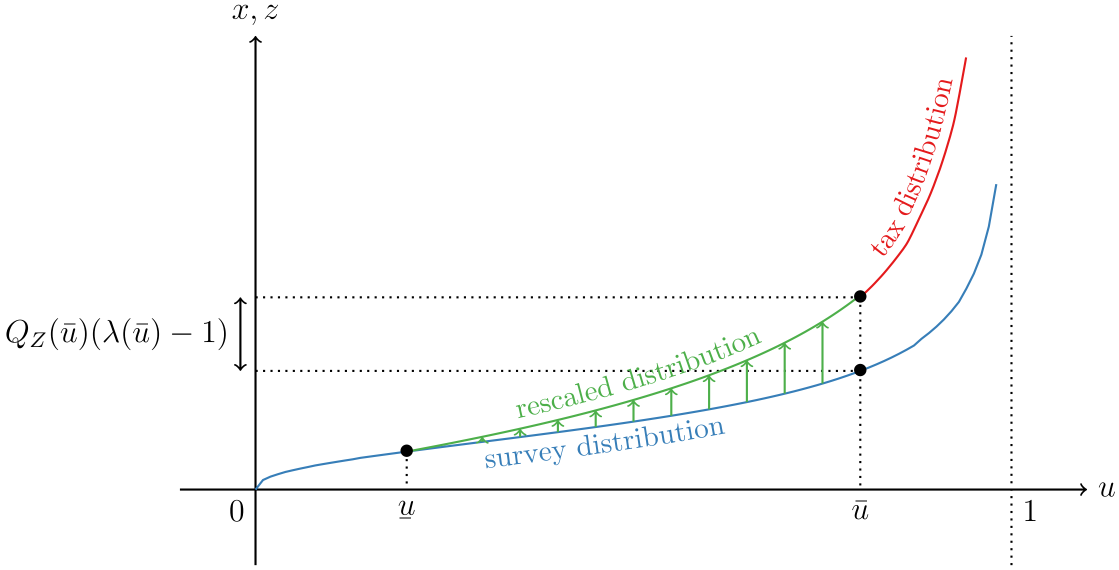
$$\lambda(u) = \frac{Q_Z(u)}{Q_Y(u)} \tag{11}$$

where Q_Y and Q_Z are the quantile functions of Y and Z. The replacing approach to correcting survey data proceeds as follows.[25] We assume a rank $\underline{u}$ below which we do not alter the survey data, assuming it is already accurate, so $\lambda(\underline{u}) = 1$. The tax data start at the rank $\bar{u}$, at which the rate of underreporting is observed directly: $\lambda(\bar{u}) = Q_Z(\bar{u})/Q_Y(\bar{u})$. The situation is pictured in figure 11. Between $\underline{u}$ and $\bar{u}$, we must assume a certain shape of the function λ. A simple and common choice is the linear rescaling profile $\lambda(u) = 1 + (\lambda(\bar{u}) - 1)\frac{u - \underline{u}}{\bar{u} - \underline{u}}$.

This procedure may make sense if we view it as a manipulation of the distribution in itself. But given the extremely strong and unrealistic assumptions stated above, any interpretation in terms of individual behaviour is slippery. And if we only understand the replacing approach as a manipulation of distribution at the aggregate level, then we should expect reweighting to perform similarly well. Indeed, reweighting simply involves adjusting the survey distribution in figure 11 horizontally rather than vertically. Therefore, we have the following equivalence between reweighting and replacing coefficients for income y and

[25] Here we present the most extreme class of replacing methods, which we label 'rescaling'. In this case there is a part of the survey distribution that is adjusted (rescaled) for which there are no tax data values to replace it. See section 1.2.

Figure 11: Correcting for misreporting by replacing



rank u, so that reweighting may be interpreted as a specific case of replacing with:

$$\Theta(y) = \frac{F_Z[Q_Z(u)/\lambda(u)]}{u}$$

In the end, unlike reweighting, it is unclear what problem exactly replacing methods end up solving. In any case, reweighting does, at least, an equally good job at solving it. Furthermore, reweighting has a clear interpretation, it is consistent with widely accepted calibration methods, it is easier to generalize to more complex settings and it always preserves the continuity of density functions, which is highly desirable and not the case in the replacing procedures, especially those adjusting arbitrary portions of the distribution (e.g. the top 1%).

Appendix B Merging Point Below the Trustable Span

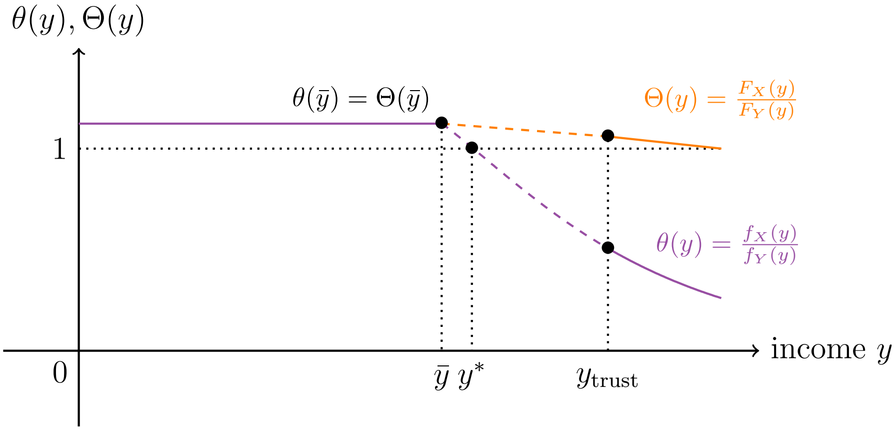
Sometimes the part of the distribution covered by the tax data is too limited to observe a merging point such that $\Theta(y) = \theta(y)$. This situation is represented in figure 12. Below y_{trust}, the value of $\theta(y)$ and $\Theta(y)$ have to be extrapolated until both curves cross, which is where we define the merging point.

We need to define a functional form for $\theta(y)$ in order perform the extrapolation (the value of $\Theta(y)$ follows from that of $\theta(y)$). We will assume the following:

$$\log \theta(y) = \gamma_0 - \gamma_1 \log y \tag{12}$$

which may also be written $\theta(y) = \mathrm{e}^{\gamma_0} y^{-\gamma_1}$. In addition to fitting the shape of the bias observed in practice, this form has the property of preserving Pareto distributions. Indeed,

Figure 12: Choice of Merging Point when $\bar{y} < y_{\text{trust}}$


$\theta(y), \Theta(y)$
$\theta(\bar{y}) = \Theta(\bar{y})$
$\Theta(y) = \frac{F_X(y)}{F_Y(y)}$
1
$\theta(y) = \frac{f_X(y)}{f_Y(y)}$
income y
0
$\bar{y}$ y^*
y_{trust}

if $f_Y(y) \propto x^{-\alpha-1}$, then $f_X(y) = \theta(y) f_Y(y) \propto x^{-\gamma_1-\alpha-1}$, which is also a Pareto density. The parameter γ_1 may be interpreted as an elasticity of nonresponse: when the income of people increases by 1%, how much less likely are they to be represented in the survey.

While the equation (12) can be estimated by OLS, we need to take into account situations where tax data covers such a small share of the distribution that the number of data points is insufficient to estimate the regression reliably. Since the frontier between having and not having enough data is blurry, our preferred approach is to deal with the two cases at once using a ridge regression. The idea is that we can know from experience a typical value for γ_1 called γ_1^*. In the absence of data, it represents our baseline estimate.[26] As we observe new data, we may be willing to deviate from that value, but only to the extent that there is enough evidence for doing so. The ridge regression formalizes this problem as:

$$\min_{\gamma_0,\gamma_1} \sum_{i=1}^{m} (\log \tilde{\theta}_k - \gamma_0 - \gamma_1 \log y_k)^2 + \lambda(\gamma_1 - \gamma_1^*)^2$$

The first term is the same sum of squares as the one minimized by standard OLS. The second term is a Tikhonov regularization parameter that penalizes deviations from γ_1^*. If $m = 1$, then $\gamma_1 = \gamma_1^*$ and the sum of squares only determines the intercept. As we get more data points, the sum of squares gets more weight and results get closer to OLS. The parameter λ determines the strength of the penalization. The problem has an explicit solution expressible in matrix form (e.g. Hoerl and Kennard, 2000). We can have a Bayesian interpretation of the method where our prior for γ_1 is a normal distribution centered around γ_1^* and λ determines its variance. The solution of the ridge regression gives the mean value of the posterior. Once we have the estimation of γ_0, γ_1 we can simulate a tax data distribution by reweighting the survey data: the point at which $\theta(y)$ crosses $\Theta(y)$ becomes the merging point $\bar{y}$, and the reweighted survey from $\bar{y}$ to y_{trust} can

[26] In practice, γ_1^* can be drawn from other "similar countries" that have sufficient data. For example, in our applications, we use the Brazilian γ_1^* to extrapolate the Chilean merging point (see section 3.2.2).

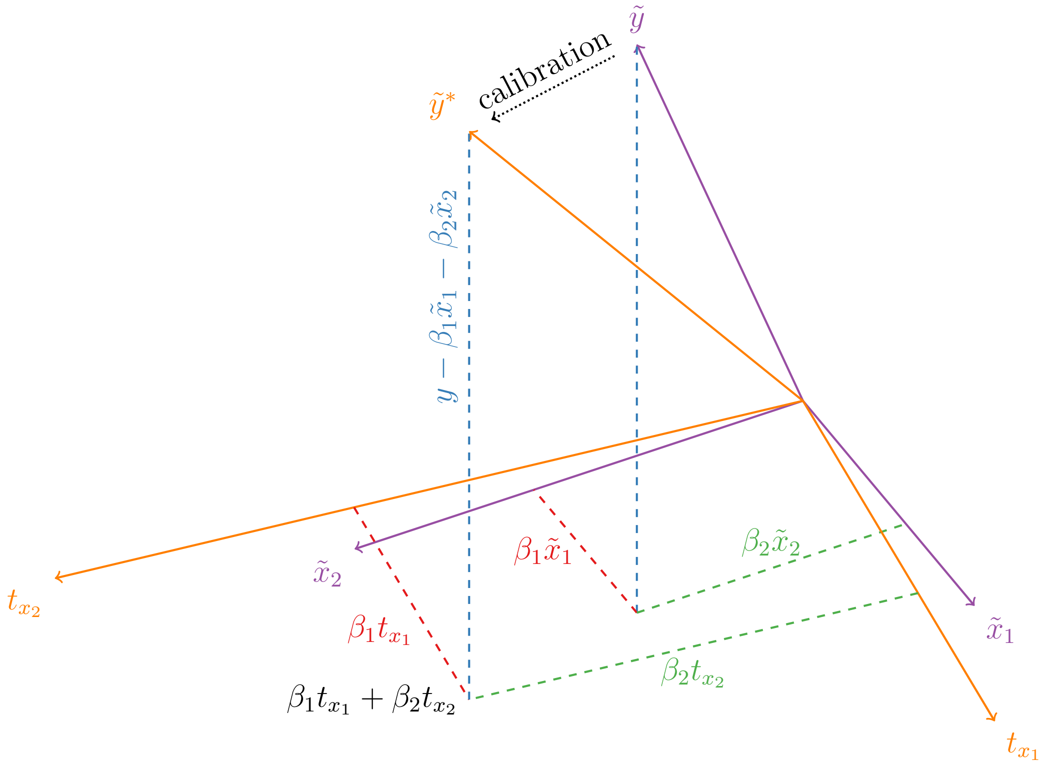
be used to complete the tax data.

Appendix C Geometrical Interpretation of Calibration

A further interpretation of the linear calibration presented in section 2.2 is geometrical. It comes from the relationship between (4) and the generalized regression estimator (GREG). Assume that we seek to estimate the total of a survey variable y. We can directly use the survey total, which we will write $\tilde{y}$. But if we wish to exploit the information on the true population totals of the auxiliary variables $x_1, \dots, x_k$, we can use the GREG estimator, whose logic is represented in figure 13. The idea is to first use the survey to project the variable of interest y onto the auxiliary variables $x_1, \dots, x_k$ using an ordinary least squares regression. Hence we get a linear prediction $\hat{y}_i = \boldsymbol{\beta x}_i$ of y_i, which corresponds to the part of y that can be explained by the auxiliary variables $x_1, \dots, x_k$. We can then substitute the survey totals by their true population counterpart in the linear prediction to get a new, corrected prediction of y. Adding back the unexplained part of y leads to the GREG estimator $\tilde{y}^* = \tilde{y} + \boldsymbol{\beta}(\boldsymbol{t} - \tilde{\boldsymbol{x}})$.

It can be shown algebraically that linear calibration is identical to the GREG procedure (Deville and Särndal, 1992). By using the calibrated weights, we systematically project the variable of interest on the calibration variables and perform the correction described above, without having to explicitly calculate the GREG estimator every time.

Figure 13: Geometrical Interpretation of Linear Calibration



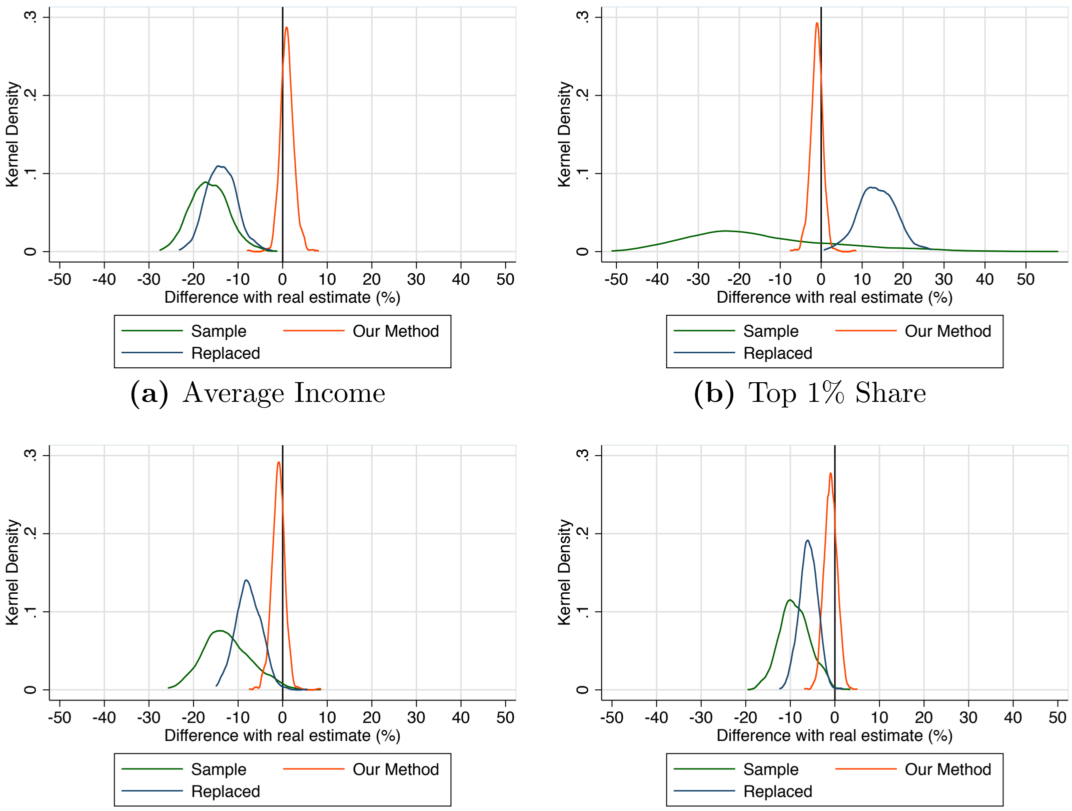
The survey totals $\tilde{y}$, $\tilde{x}_1$ and $\tilde{x}_2$ are shown in purple. The GREG estimator, which is equivalent to linear calibration, first projects $\tilde{y}$ onto $\tilde{x}_1$ and $\tilde{x}_2$ (dashed blue line). This projection is equal to $\beta_1\tilde{x}_1 + \beta_2\tilde{x}_2$. The true population totals t_{x_1} and t_{x_2} are in orange. We substitute them for $\tilde{x}_1$ and $\tilde{x}_2$ in the projection, which gives the value $\beta_1 t_{x_1} + \beta_2 t_{x_2}$. We add back the unexplained part of $\tilde{y}$ (dashed blue line) to get the calibrated total $\tilde{y}^*$.

Appendix D Further Monte Carlo Simulations

This section presents three supplementary experiments to those presented in section 3.1. Each one of them includes punctual changes in the parameters underlying the benchmark experiment, which is a useful way to isolate possible effects and thus to anticipate the method's performance in different scenarios.

Figure 14 displays results of an experiment which only differs from the benchmark in that the misreporting bias is stronger. That is, the probability of misreporting starts increasing from percentile 85 (P85) instead of 95 (P95). This mechanically affects the accuracy of estimates produced using the raw survey, which is expected given that more people are actually misreporting their income. Indeed, the variance of each of the estimates from the raw sample increases substantially, which is visible by comparing the width of their kernel densities to the corresponding ones in the benchmark setting. Although replaced surveys still appear to get somewhat closer than raw estimates to true values after

Figure 14: Experiment with more Misreporting



(a) Average Income

(b) Top 1% Share

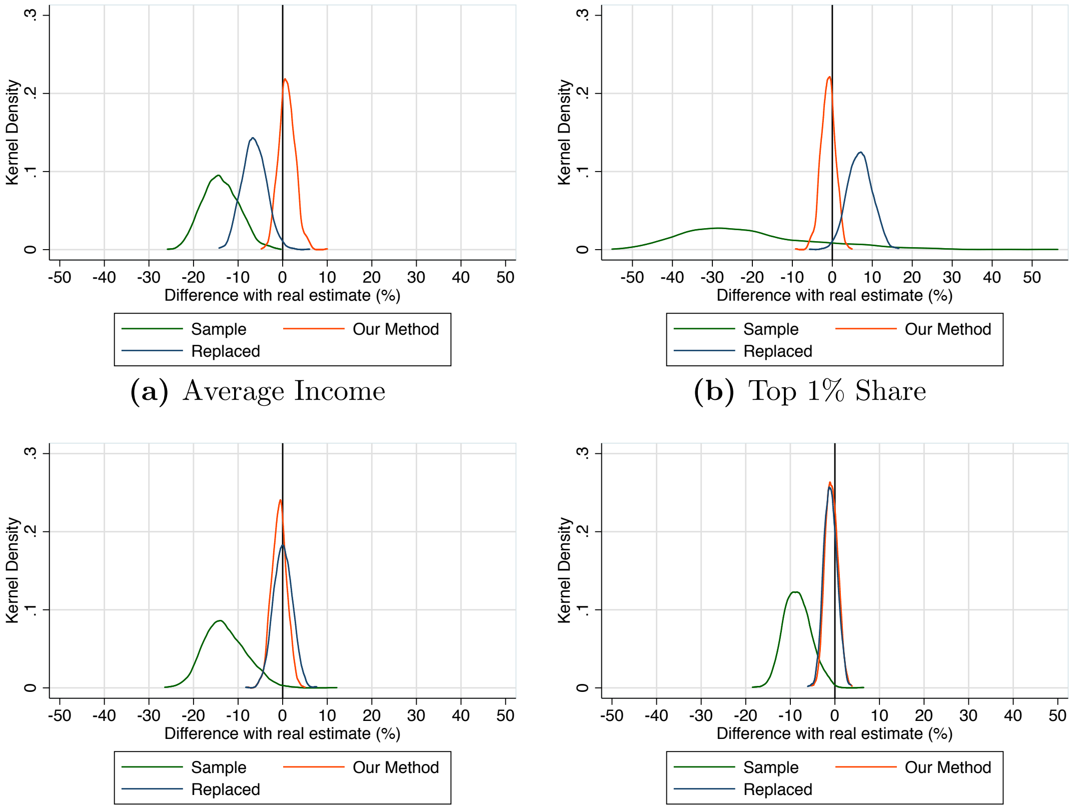
(c) Top 10% Share

(d) Gini Coefficient

correction, they are also substantially affected by the increased variability. However, this setting only affects the performance of our method marginally, as the resulting densities of estimates are almost indistinguishable from those displayed in figure 5, which is a good proof of adaptability. Other experiments were conducted, where we assumed stronger non-response biases. But we do not display results because they are almost identical to those presented in figure 14. This can be explained by the fact that both biases have a similar effect — only in distributive terms, as opposed to individual representativeness — on resulting distributions (see Appendix A.1).

Figure 15 depicts another setting, where the only difference with the benchmark experiment is that the replacing procedure uses the top 5% instead of only the top 1%. The estimates produced by our adjustment method are virtually the same than the benchmark. Since, by definition, biases are active in the top decile, the increase in the replaced population results in estimates that are more accurate, especially in the case of the top 10% share and the Gini coefficient (see figures 15c and 15d), which appear to be substantially closer to our estimates and thus to true values. However, the same is not true for both the estimated average income and the top 1% share, which still tend to be substantially underestimated and overestimated, respectively (figures 15a and 15b).

Figure 15: Experiment with Replacing the Top 5%



(a) Average Income

(b) Top 1% Share

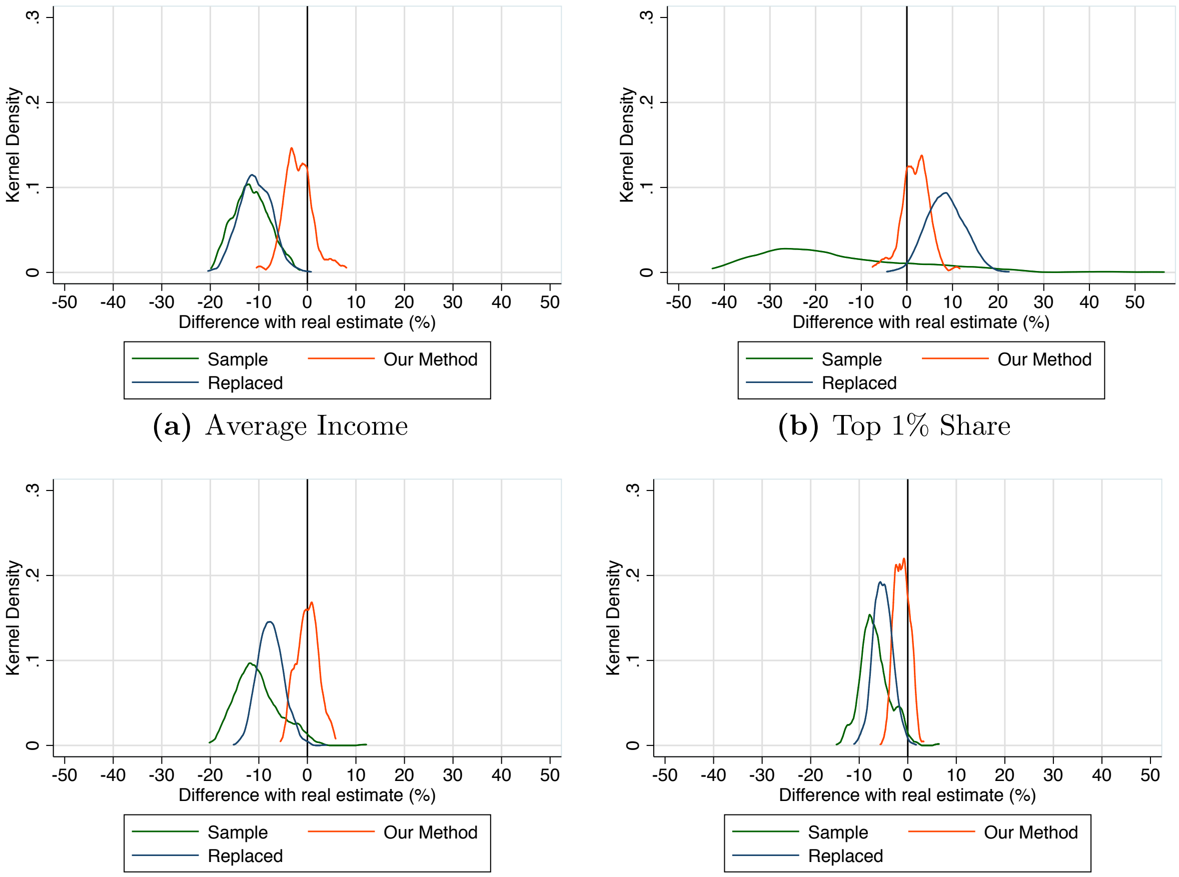
(c) Top 10% Share

(d) Gini Coefficient

Although we could go further and try to find the exact portion of the population that has to be replaced to get a similar result to that obtained with our method, we judge this to be an unnecessary exercise. As we argue in section A.2, the equivalence between our method and replacing can be found in some cases, yet it would only would be valid in a purely distributional perspective because replacing implies extremely unrealistic assumptions at the individual level and, thus, does not preserve the consistency of the resulting observations.

Figure 16 represents a somewhat extreme case where we limit access to tax declarations to only the top 5% of respondents, thus forcing our method to extrapolate adjustment factors in a large majority of cases. Our estimates appear to be less precise than in the benchmark, where a larger part of the information was used, yet they still perform better than both the raw survey and the replacing alternative. The resulting distribution of estimates is to our judgement rather satisfactory, since estimates remain closely centered around true values. This experiment shows that under extreme circumstances, where tax data covers a very small part of the population, estimates resulting from our correction method are still accurate, yet reasonably less precise.

Figure 16: Experiment with Poor Tax Data



(a) Average Income

(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

Appendix E Data Details and Supplementary Results

E.1 Country Specific Income Concepts and Observational Units

E.1.1 Brazil

To reconcile incomes in surveys with those in tax data, we use the latter as the benchmark for the top of the distribution. We thus require that the survey definition of income, from the micro-data, be consistent with the definition of income in the tax tabulations in order for the comparison to make sense. The total income assessed in tax data is pre-tax-and-transfer income, but including pensions and unemployment insurance. It is the sum of three broad fiscal categories: taxable income, exclusively taxed income and tax-exempt income (reported in Table 9 of the tax report *Grandes Números DIRPF*). We describe each of these in turn before describing how we construct the survey definition of income.

Taxable income comprises of wages, salaries, pensions and property rent. These are incomes that are subject to assessment for the personal income tax. Exclusively-taxed income is income that has been already been taxed at source according to a separate tax schedule. It also contains capital income and labour income components. The labour component is the sum of the 13th monthly salary received by the contributor and their dependents, wages received cumulatively by contributors or dependents, and worker participation in company profits. The capital component comprises of the sum of fixed income investment income, interests on own capital ("juros sobre capital próprio"), variable income investment income, capital gains and other capital income. Non-taxable incomes are the last fiscal category, whose decomposition is presented in Table 20 of the tax reports. These are incomes that are declared but which are not subject to any personal taxation when received. Close to one-fifth of these exempt incomes can be classified as labour income. These comprise of compensation for laid-off workers, the exempt portion of pension income for over 65s, withdrawals from employment security fund, scholarships, and other labour incomes. The remaining items can be classified as capital income (distributed company profits, dividends, interests from savings accounts/mortgage notes) or mixed income (the exempt portion of agricultural income).

We construct survey income to be as close to the tax definition as possible. The total income we analyse from the PNAD surveys is the sum of labour income, mixed income and capital income. Labour income is the sum of all reported income from primary, secondary or all other jobs (variables V9532, V9982, V1022) for all employed individuals who do not classify themselves as own-account (self-employed) workers or employers. For employers, we assume that labour income is the portion of their work income that is below the annual exemption limit for the DIRPF, as set by the Receita Federal. Thus, values above the first tax paying threshold are taken to be capital withdrawals. Also in labour income are pensions (V1252, V1255, V1258, V1261), work allowances (V1264), *abono salarial* and unemployment insurance. Of the latter two, the first is imputed as one minimum wage for eligible formal private sector employees, while the second is imputed for respondents who claimed to have received unemployment benefits at some point in the 12 months before the PNAD interview. Benefit levels were imputed as yearly averages of shares of the minimum wage from current legislation. Values of V1273 equal to or below 1 monthly minimum wage are interpreted as social benefits, which are excluded from the analysis.

Mixed income is the reported income of own-account workers. Capital income is estimated as the sum of rent (V1267), financial income, and the capital portion of employer work income (i.e. reported amounts exceeding the annual exemption limit for DIRPF). Financial income (interests and dividends) is taken from other income sources declared (V1273) and estimated as any income from this source that exceeds 1 monthly minimum wage. Finally, we add a 13th monthly salary to the annual calculation of the incomes of formal employees and retirees. In total, the income we calculate from the surveys represents close to 80% of the equivalent (fiscal income) total from the household sector in the national accounts, on average between 2007 and 2015. The total income we use from tax statistics accounts for about 63% of the same fiscal income total from the national accounts over the same period.

Given that the unit of assessment in the tax data can either be the individual or the couple, in cases where the latter opt to declare jointly, we cannot strictly restrict ourselves to the analysis of individual income as it is received by each person. Therefore, we decide follow the tax legislation by identifying the number of married couples appearing jointly on the declaration and splitting their total declared income equally between them when carrying out the generalized Pareto interpolation (Blanchet, Fournier, and Piketty, 2017) from the tabulation. This allows us to bring the analysis to the individual level by assuming that all spouses equally share their income. We use the information available in the tax statistics to estimate the share of joint declarations, which overall represent about 30% of all filed declarations (see (Morgan, 2018)). To be consistent in the comparison, we also use individual income in the surveys, with the income of married couples being split equally between the composite adults. We consider all adults aged 20 or over in our analysis.

E.1.2 Chile

Following the same logic as that applied to the Brazilian case, we construct from the Chilean survey an income definition that is as close as possible to the one used in tax data. The resulting definition is the one we use when merging datasets. However, in Chile, unlike Brazil, the survey reports post-tax incomes. In broad terms, we estimate pre-tax income retrospectively from declared post-tax income. In order to do so, we make a priori assumptions on whether certain types of income pay income taxes or not. Additionally, some self-reported characteristics are used to determine if the income of certain individuals should be treated as taxable or not. For instance, dependent workers that do not have a contract (and will not sign any soon) are considered to be informal, thus they are assumed to not pay the income tax. A similar mechanism is used for independent workers – depending on if they emit invoices (both commercial or for services) we define them as formal or informal. Table E.1 gives a comprehensive view on what types of income are assumed to pay taxes or not. For further comments on the definition of income corresponding to tax data, please refer to Flores et al. (forthcoming).

E.1.3 European Countries

Tax Data For the three European countries we use tabulated tax data from official sources. In the case of Norway and the United Kingdom, the data come directly from institutional sources: "Tax Statistics for Personal Taxpayers" from Statistics Norway (`https://www.ssb.no/en/statbank/list/selvangivelse`) for the former, and the "Survey of Personal Incomes" (SPI) from HM Revenue & Customs (`https://www.gov.uk/government/statistics/income-tax-liabilities-by-income-range`), for the latter.

Table E.1: From Post-Tax to Pre-Tax Income in Chilean Surveys

Type of income	Taxable Income		Tax Exempt Income	
	Variable name	**Code**	**Variable name**	**Code**
Labor Income	Wage (1ry occup.).	y1a	Occasional work.	y16a
	Wage (2ry occup.).	y6, y10	Unemp. insurance.	y14c
	Inc. from previous months (if dependent).	y14b	Tips, travel expenses.	y3c, y3e
	Extra hours, commissions & allowances.	y3a, y3b, y3d y3f	Christmas bonus.	y4a
	Rewards & additional salary.	y4b, y4c, y4d	Inc. of the inactive.	y11a
			Wage of informals.	o17, o14
Pensions	Old age pension.	y27am		
	Disability pension.	y27bm		
	Widow's pension.	y27cm		
	Orphan's pension.	y27dm		
Mixed Income	Inc. of indep. (1ry occup.)	y7a	Inc. of indep. (2ry occup.).	y6,y10
	Inc. from previous months (if indep.).	y14b	Inc. of non-qualified, informal, small minery & craftsmen.	oficio1, oficio4, o14
Capital Income	Rent (agricultural).	y12b	Rent (urban).	y12a
	Interest.	y15a	Rent (seasonal).	y16b
	Dividends.	y15b		
	Withdrawals.	y15c		
	Rent (equipment).	y16a		

Notes: Codes correspond to those of CASEN 2011-2013. Formality is defined as conditional to having a contract and/or emitting ”*boletas de honorarios*" (invoices by independents). Information on formality is only available for primary occupation. Formality is assumed to be the same for 1ry and 2ry occupations. In the survey, income is post-tax. Pre-tax formal income of contract-workers is calculated using tables of IUSC (*Impuesto Único de Segunda Categoría*) retrospectively. Pre-tax income of formals emitting invoices is added of mandatory provisional deductions (e.g. 10%) and standard presumptive expenses (e.g. 30%). Pre-tax capital income is calculated using the IPC (*Impuesto de Primera Categoría*) single tax-rate (e.g. 20%). Rent of urban properties is assumed to be untaxed because of law D.F.L.2 (1959)

The tax unit for both countries is the individual. As explained in Section 3.2.2, we interpolate the tabulations using a generalized Parteo interpolation Blanchet, Fournier, and Piketty (2017). For France, we use detailed tabulations produced by Garbinti, Goupille-Lebret, and Piketty (2016) from the micro-files of French taxpayers. These are available in the Appendix C Tables of their Data (see `http://piketty.pse.ens.fr/en/publications`). We use the individual-level tabulations that present the distribution of gross total fiscal income for 127 percentiles.

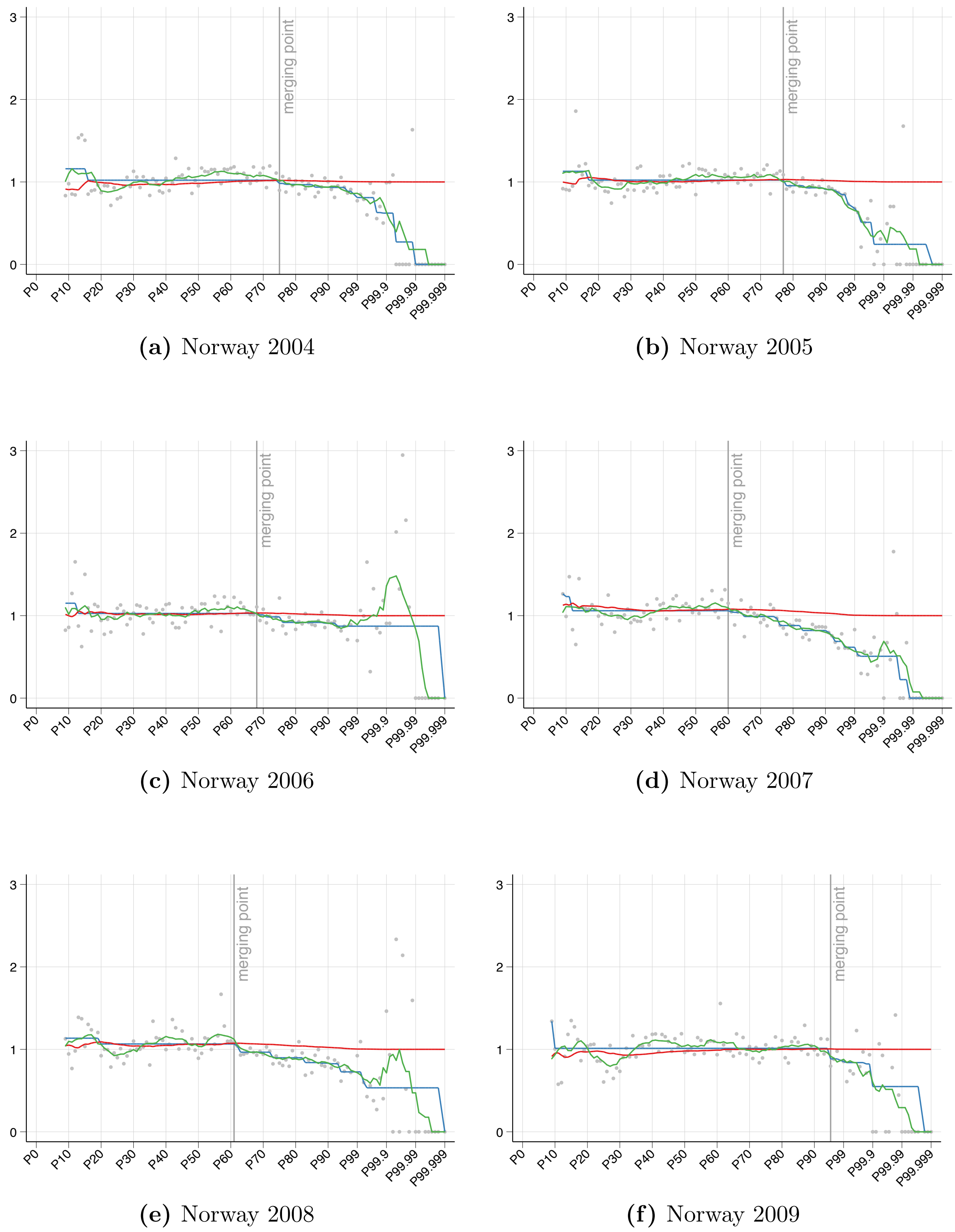
EU-SILC Data The advantage of using EU-SILC data is that it is a harmonized household survey dataset for European countries. However, given that we anchor our estimation method to the tax data, the definition of income used from surveys must match that accounted for in tax statistics. To do so we take the sum for each observation of employee cash or near cash income (variable PY010), self-employment cash income (PY050), Pensions received from private plans (PY080), a host of benefits related to unemployment, old-age, suvivors, sickness and disability (PY090, PY100, PY110, PY120, PY130), and capital income components (rent from property or land (HY040) and interests, dividends, profit from capital investments (HY090)). These capital incomes are reported at the household level. We individualise them by equally splitting the income among spouses and civil partners. For Norway and the UK, consistent with the fiscal income in tax data, we take gross incomes (before income taxes and individual social contributions levied at source). Since fiscal income in the French tax data is before income tax but after social contributions levied at source, we take net income values from the French SILC dataset. Income taxes are not levied at source in France for the period we analyse so the definition of net income in SILC is apt to be used for this case. We also select the reference population to be kept in accordance with the tax statistics. In Norway, the tax tabulations refer to individuals aged 17 and over, so we discard individuals under the age of 17 in the survey. For the UK, the tax data does not provide comparable information, so we follow the practice by Atkinson (2007) in taking a reference population of individuals aged 15 and over. In France, consistent with the use of the population aged 20 and over in Garbinti, Goupille-Lebret, and Piketty (2016), we keep persons aged 20 and over in the survey.

E.2 Further Tables and Figures

E.2.1 Shape of the Bias

Figures E.1-E.5 show the shape of the bias we estimate for the other years among our sampled countries. Each coverage of the data points are determined by the trustable span of the tax data in each country, which is defined as the portion of the population that are subject to positive income tax payments.

Figure E.1: Merging Points in Norway, 2004-2013



(a) Norway 2004

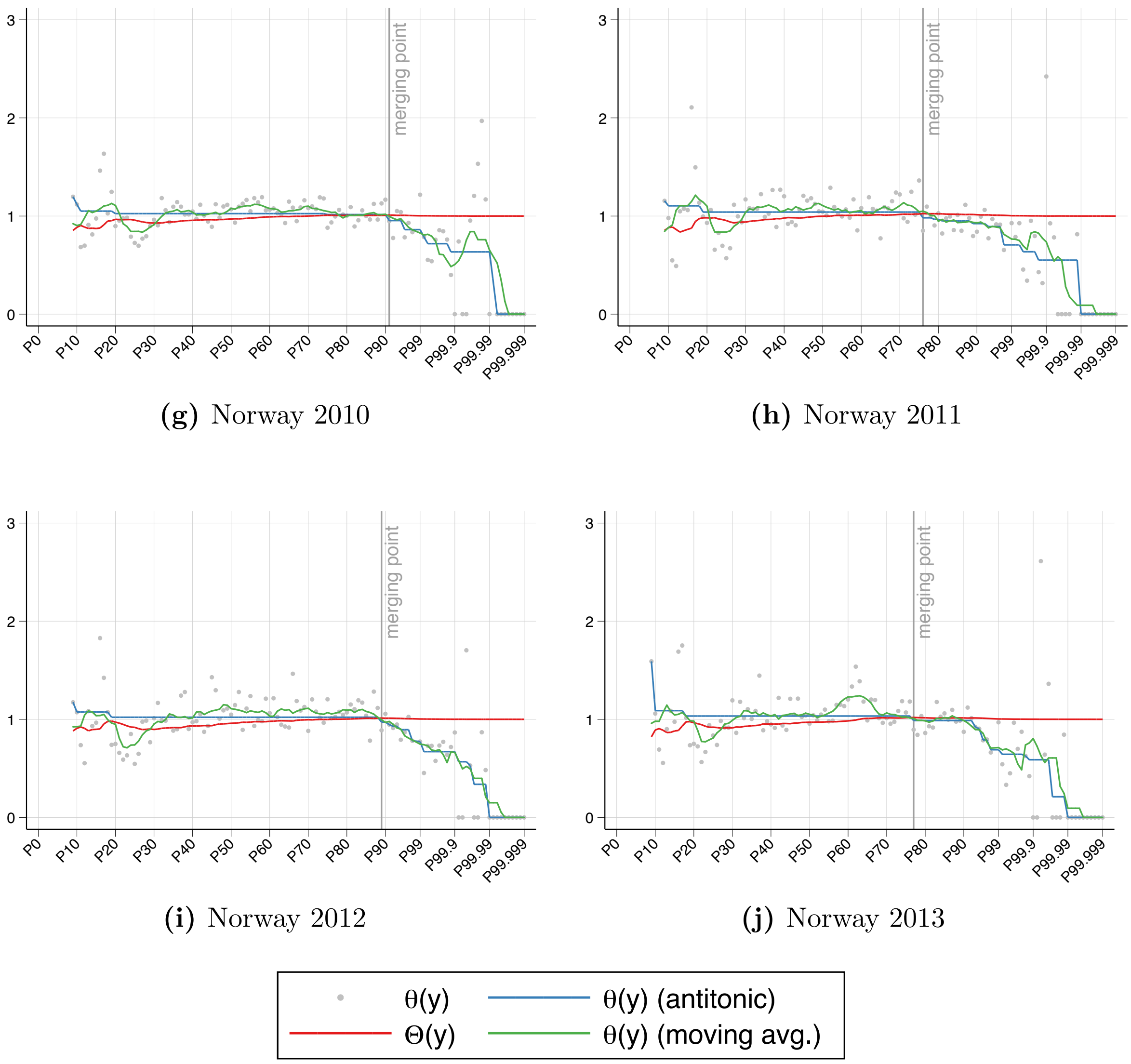
(b) Norway 2005

(c) Norway 2006

(d) Norway 2007

(e) Norway 2008

(f) Norway 2009



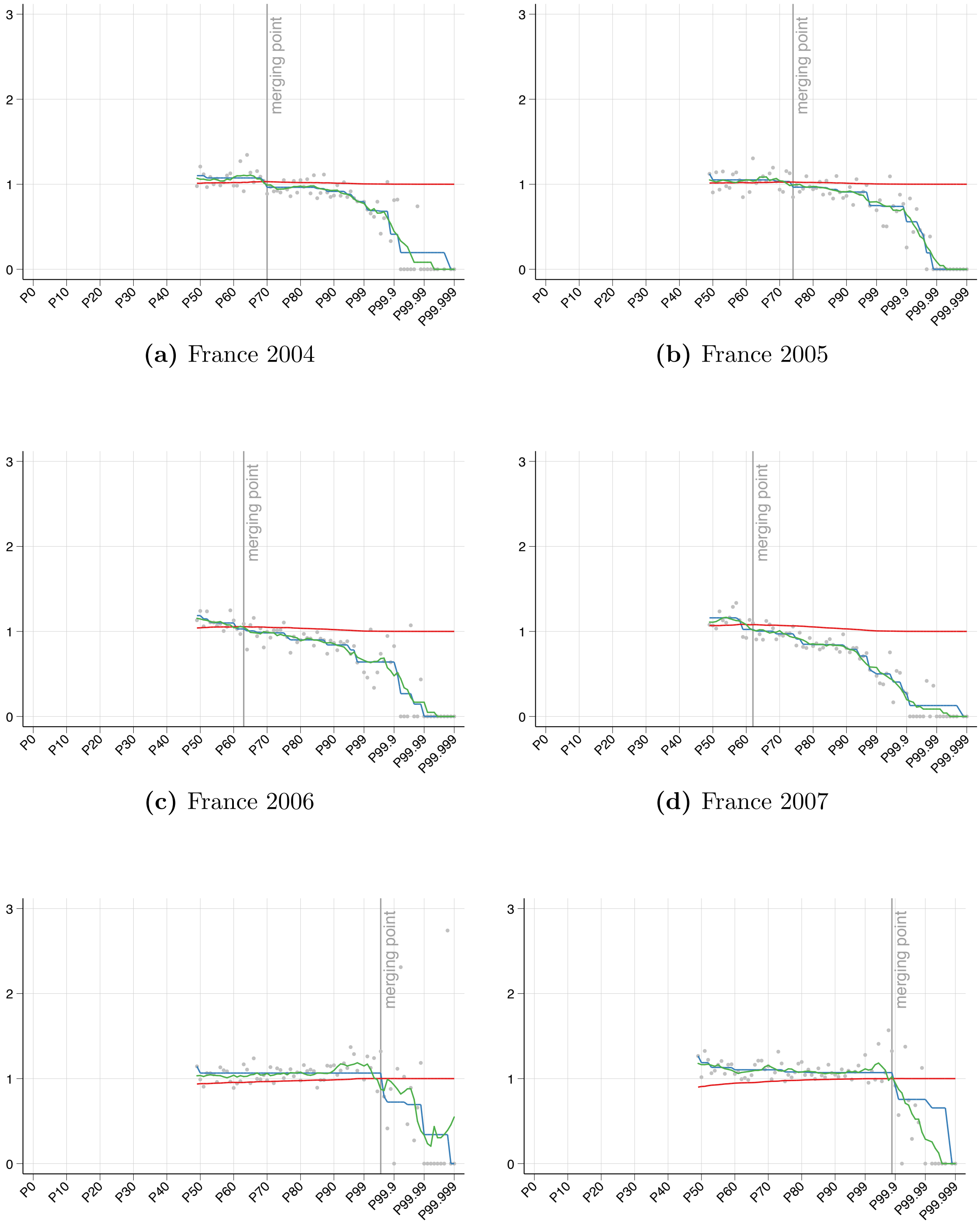
(g) Norway 2010

(h) Norway 2011

(i) Norway 2012

(j) Norway 2013

Figure E.2: Merging Points in France, 2004-2013



(a) France 2004

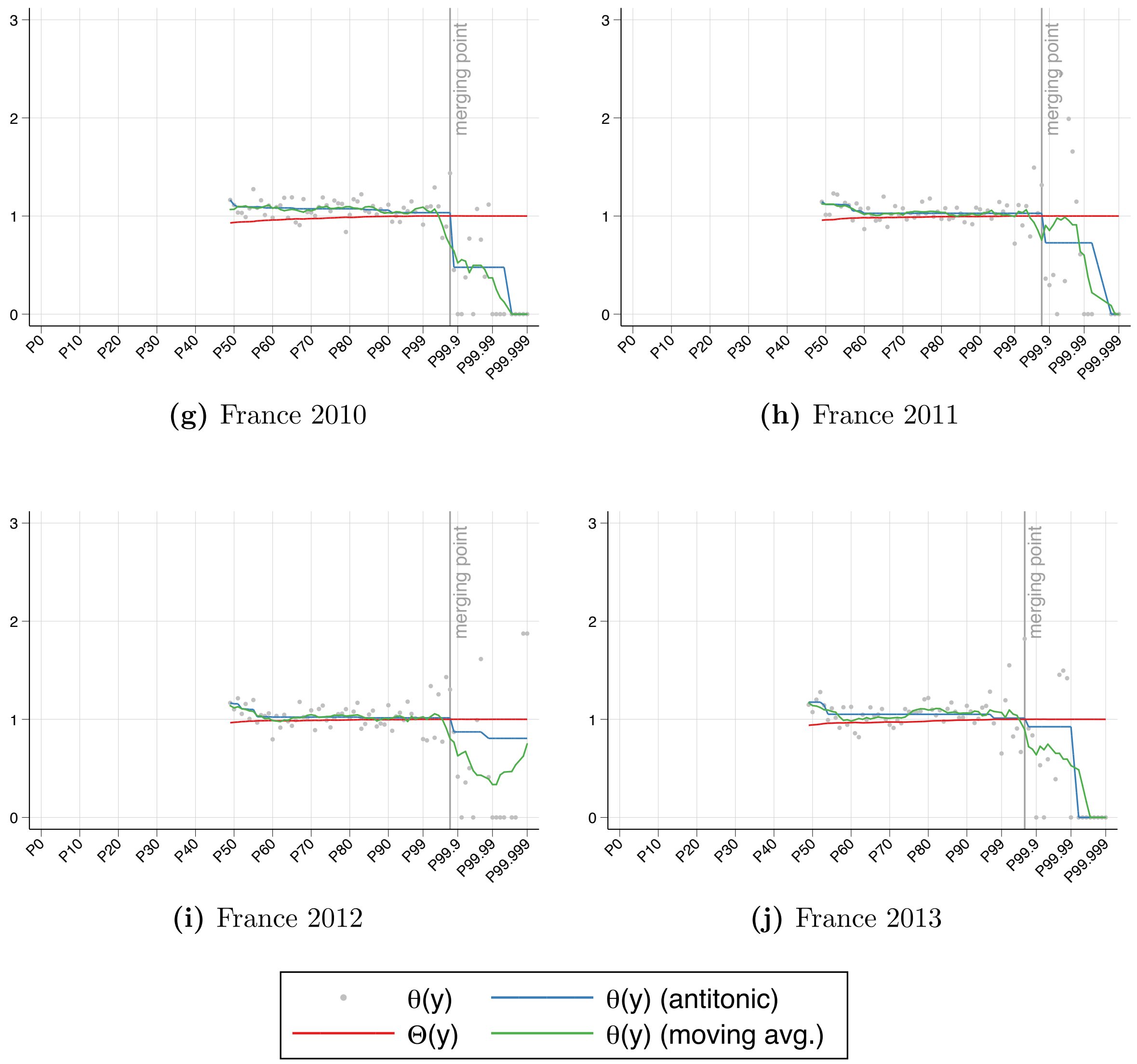
(b) France 2005

(c) France 2006

(d) France 2007

(e) France 2008

(f) France 2009



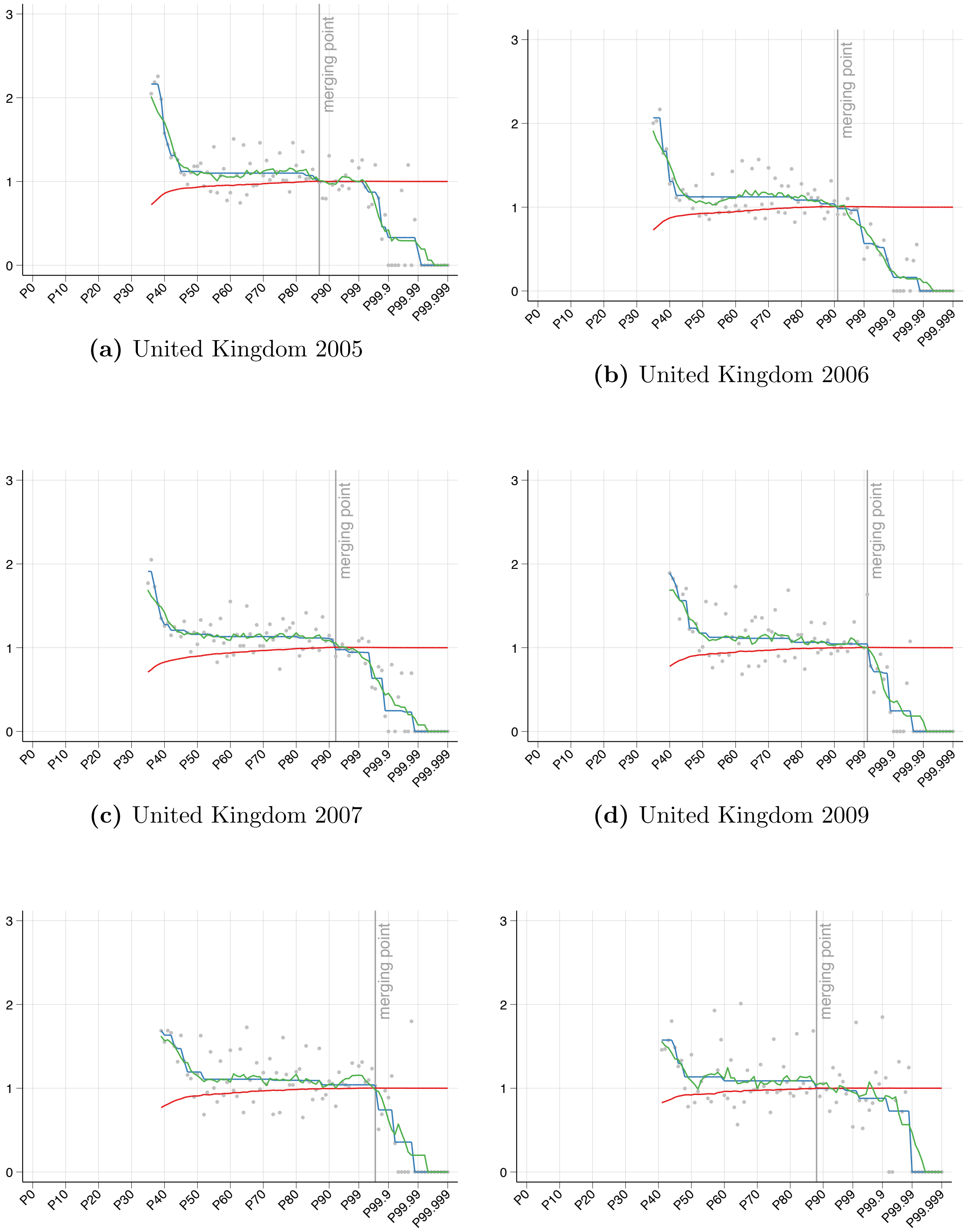
(g) France 2010

(h) France 2011

(i) France 2012

(j) France 2013

Figure E.3: Merging Points in United Kingdom, 2005-2013



(a) United Kingdom 2005

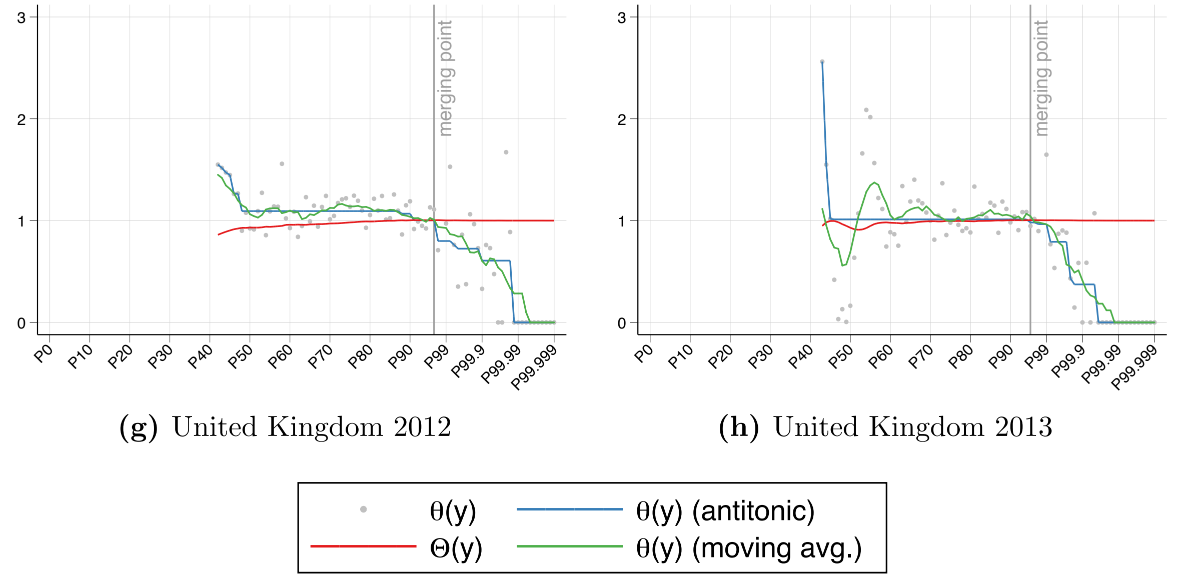
(b) United Kingdom 2006

(c) United Kingdom 2007

(d) United Kingdom 2009

(e) United Kingdom 2010

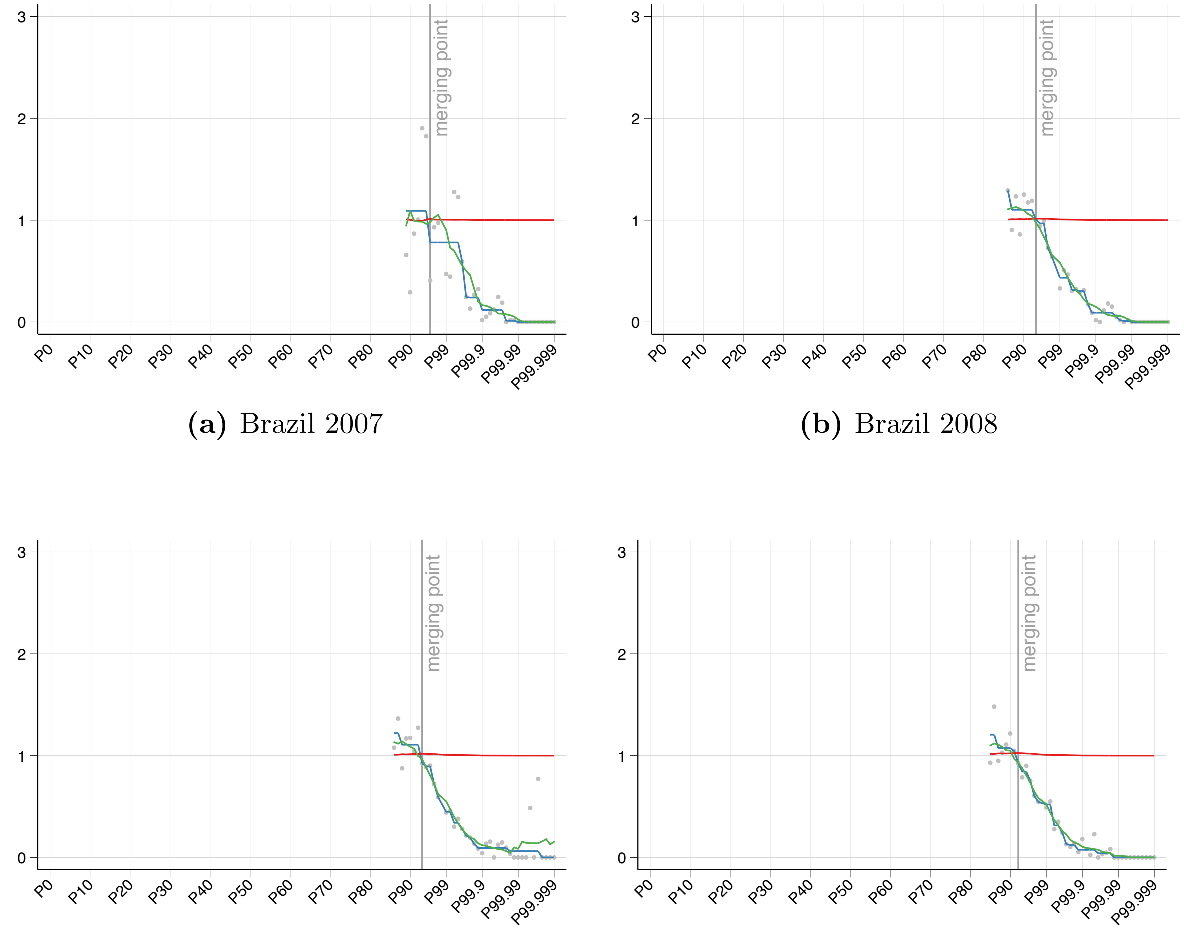
(f) United Kingdom 2011



(g) United Kingdom 2012

(h) United Kingdom 2013

Figure E.4: Merging Points in Brazil, 2007-2014

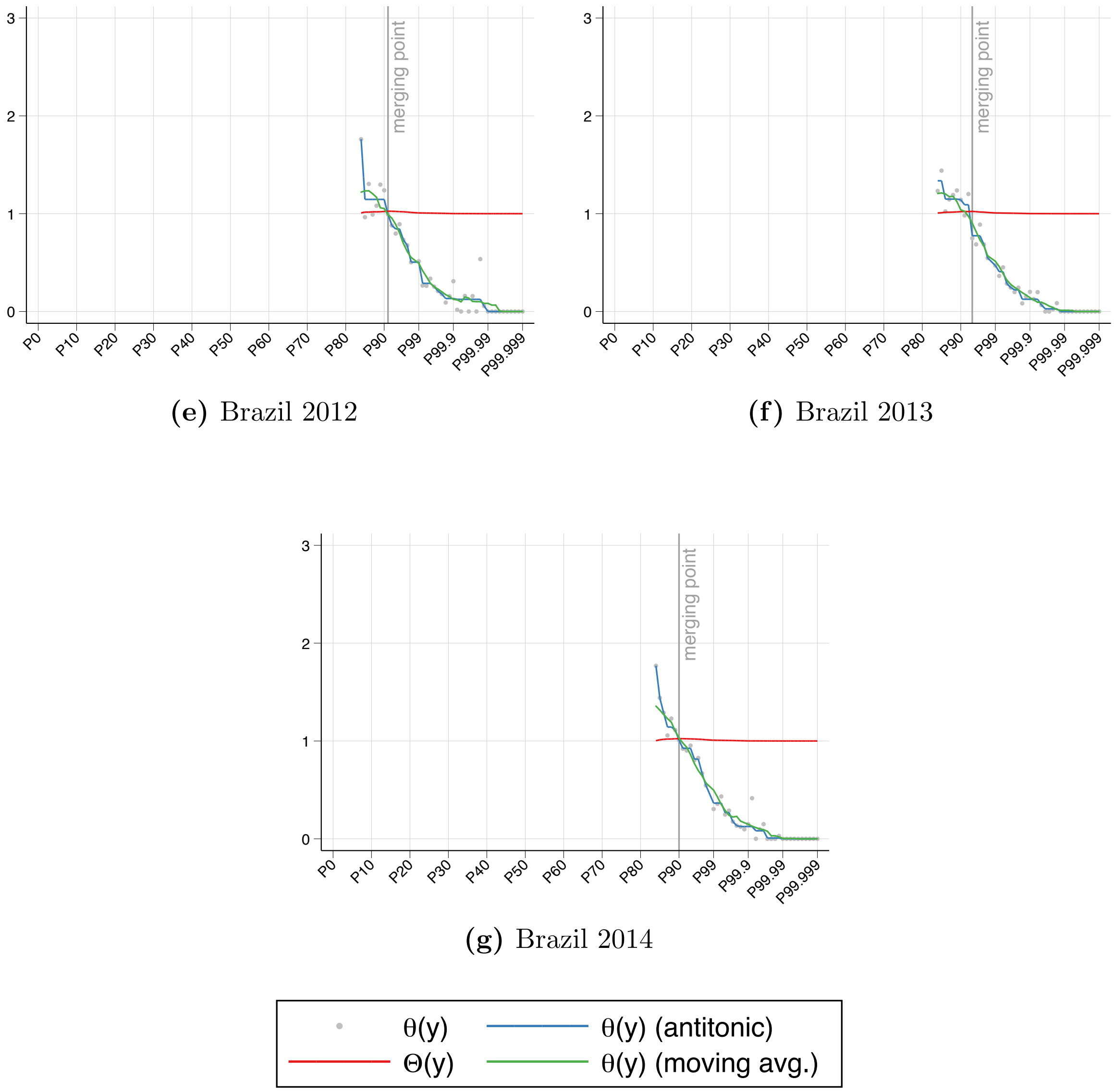


(a) Brazil 2007

(b) Brazil 2008

(c) Brazil 2009

(d) Brazil 2011

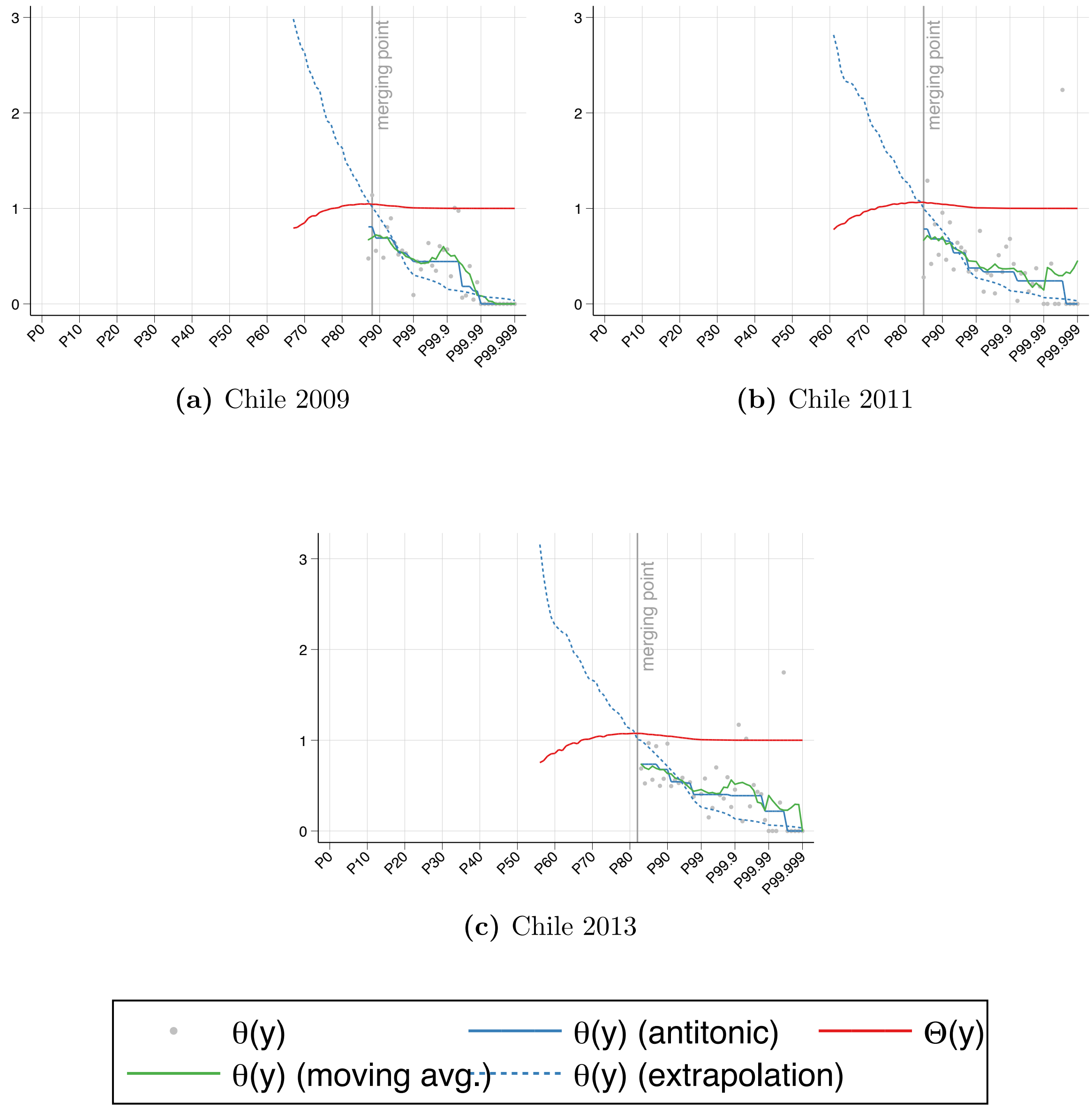


(e) Brazil 2012

(f) Brazil 2013

(g) Brazil 2014

Figure E.5: Merging Points in Chile, 2009-2013



(a) Chile 2009

(b) Chile 2011

(c) Chile 2013

E.2.2 Structure of the Corrected Population

Tables E.2-E.6 show the structure of the corrected population for all years in all sampled countries.

Table E.2: Structure of Corrected Population in Brazil, 2007-2015

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
2007	1.0%	0.7%	0.33%	98.2%	1.8%
2008	1.0%	0.6%	0.44%	97.2%	2.8%
2009	1.0%	0.5%	0.51%	99.3%	0.7%
2011	2.0%	1.4%	0.57%	95.9%	4.1%
2012	3.0%	2.3%	0.70%	98.3%	1.7%
2013	2.0%	1.4%	0.62%	97.1%	2.9%
2014	2.0%	1.2%	0.76%	98.8%	1.2%
2015	2.0%	1.3%	0.70%	97.2%	2.8%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table E.3: Structure of Corrected Population in Chile, 2009-2015

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2009	11.0%	7.2%	3.8%	99.6%	0.4%
2011	14.0%	8.5%	5.5%	99.9%	0.1%
2013	17.0%	10.6%	6.4%	99.9%	0.1%
2015	17.0%	11.1%	5.7%	99.99%	0.01%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table E.4: Structure of Corrected Population in France, 2004-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2004	29.0%	26.8%	2.17%	99.9%	0.1%
2005	25.0%	23.1%	1.95%	98.5%	1.5%
2006	36.0%	32.5%	3.50%	99.5%	0.5%
2007	37.0%	32.0%	4.99%	99.96%	0.04%
2008	0.4%	0.3%	0.11%	97.6%	2.4%
2009	0.1%	0.1%	0.02%	89.8%	10.2%
2010	0.2%	0.1%	0.11%	94.5%	5.5%
2011	0.2%	0.1%	0.06%	94.3%	5.7%
2012	0.2%	0.2%	0.03%	96.5%	3.5%
2013	0.3%	0.3%	0.03%	72.3%	27.7%
2014	0.1%	0.0%	0.05%	99.0%	1.0%

Notes: From 2008, the French survey was supplemented with register data for increased precision in the responses. Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table E.5: Structure of Corrected Population in Norway, 2004-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2004	24.0%	22.5%	1.49%	99.3%	0.7%
2005	22.0%	19.7%	2.27%	99.8%	0.2%
2006	31.0%	28.8%	2.16%	99.9%	0.1%
2007	39.0%	34.2%	4.75%	99.5%	0.5%
2008	38.0%	33.4%	4.59%	99.95%	0.05%
2009	4.0%	3.5%	0.54%	99.4%	0.6%
2010	8.0%	7.1%	0.88%	99.0%	1.0%
2011	23.0%	21.1%	1.93%	99.0%	1.0%
2012	10.0%	8.9%	1.13%	98.6%	1.4%
2013	22.0%	20.5%	1.49%	99.1%	0.9%
2014	5.0%	4.6%	0.39%	96.0%	4.0%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

Table E.6: Structure of Corrected Population in United Kingdom, 2005-2014

Year	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
2005	12.0%	11.7%	0.26%	99.5%	0.5%
2006	8.0%	7.3%	0.72%	96.9%	3.1%
2007	7.0%	6.5%	0.53%	95.5%	4.5%
2009	0.8%	0.5%	0.33%	85.5%	14.5%
2010	0.4%	0.3%	0.14%	84.9%	15.1%
2011	11.0%	10.8%	0.18%	93.0%	7.0%
2012	3.0%	2.6%	0.37%	92.2%	7.8%
2013	4.0%	3.6%	0.45%	86.1%	13.9%
2014	3.0%	2.5%	0.54%	93.6%	6.4%

Notes: Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum).

E.3 Detailed Distribution

Table E.7 depicts a more detailed picture of the impact of our adjustment method on the income distribution of our 5 countries, compared to the raw survey results and those from the replacing alternative. We take the last available year as an illustration. With respect to income shares across the distribution, the main conclusions drawn from the analysis of top shares in Section 3 can be generally extended, more or less, to other top shares, from the top 10% to the top 0.001% shares. As is to be expected, both the middle 40% and Bottom 50% shares are reduced in all countries after our adjustment. This is consistent with the mechanics of our method, where higher aggregate weight for top fractile incomes must be compensated by a lowering of the amount of middle and lower incomes observed in the population. Replacing produces results in the same direction, except that, by not decreasing the weight of lower incomes, it results in higher shares for the Bottom 50% than those from our method in all countries. The same is true for the Middle 40% for Brazil and Chile, but not for the three European countries. Overall, replacing produces inconsistent results across the distribution, which are difficult to explain.

Figure E.6 presents in more detail the impact of our method on total income. For our two country case studies with the largest corrections to total income, we are able to show that the total income in the corrected surveys is closer to the reference total of "fiscal income" from national accounts. For the cases of Chile and Brazil respectively, our correction bridges about 80% and 60% of the gap between survey income and the reference total from national accounts.

Table E.7: Income Shares: Raw Survey and Corrected Survey

Raw Survey

Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	16.9%	8.0%	23.4%	25.2%	14.8%
Middle 40%	45.3%	45.2%	47.0%	48.6%	49.6%
Top 10%	37.7%	46.9%	29.6%	26.2%	35.5%
Incl. Top 1%	*10.2%*	*14.3%*	*7.2%*	*5.8%*	*9.4%*
Incl. Top 0.1%	*2.2%*	*3.4%*	*1.5%*	*1.4%*	*2.5%*
Incl. Top 0.01%	*0.5%*	*0.7%*	*0.4%*	*0.3%*	*0.4%*
Incl. Top 0.001%	*0.09%*	*0.2%*	*0.1%*	*0.03%*	*0.04%*
Average income	€8,691	€8,101	€23,367	€37,431	€22,389
Gini	0.505	0.64	0.40	0.37	0.52

Corrected Survey (Our Method)

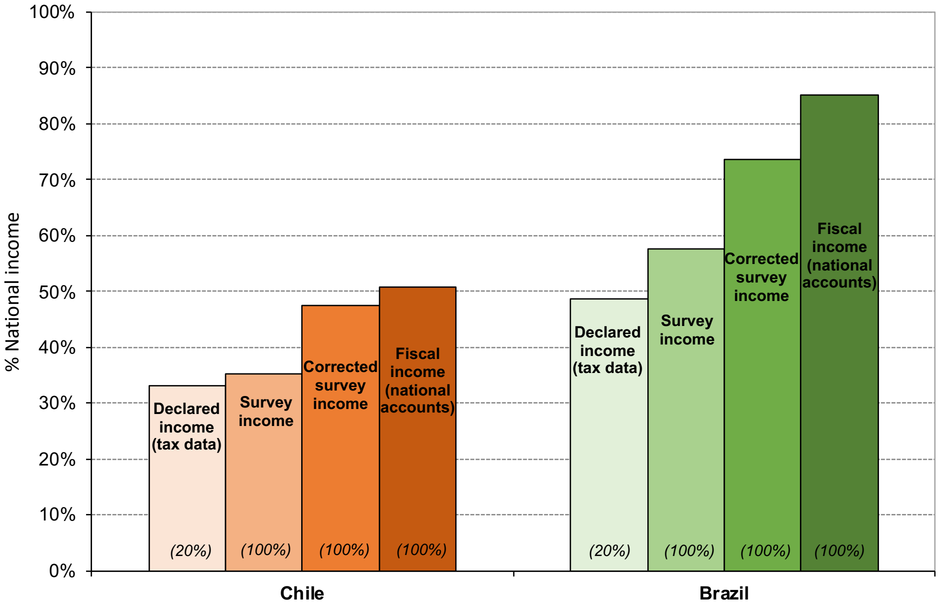
Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	12.7%	6.7%	23.2%	24.6%	13.9%
Middle 40%	35.1%	40.1%	46.5%	47.7%	46.6%
Top 10%	52.3%	53.2%	30.3%	27.6%	39.6%
Incl. Top 1%	*23.7%*	*16.7%*	*8.2%*	*7.1%*	*13.7%*
Incl. Top 0.1%	*11.2%*	*4.5%*	*2.2%*	*2.2%*	*5.4%*
Incl. Top 0.01%	*5.6%*	*1.3%*	*0.6%*	*0.7%*	*2.1%*
Incl. Top 0.001%	*2.8%*	*0.4%*	*0.2%*	*0.26%*	*0.89%*
Average income	€11,935	€11,097	€23,621	€38,320	€24,081
Gini	0.619	0.69	0.41	0.38	0.55

Corrected Survey (Replacing)

Income groups	**Brazil**	**Chile**	**France**	**Norway**	**UK**
Bottom 50%	14.4%	7.9%	24.0%	25.7%	14.8%
Middle 40%	36.4%	41.0%	45.9%	47.1%	46.4%
Top 10%	49.2%	51.2%	30.0%	27.2%	38.8%
Incl. Top 1%	*26.7%*	*21.1%*	*7.9%*	*7.1%*	*14.0%*
Incl. Top 0.1%	*12.6%*	*5.7%*	*2.2%*	*2.2%*	*5.5%*
Incl. Top 0.01%	*6.3%*	*1.6%*	*0.6%*	*0.7%*	*2.1%*
Incl. Top 0.001%	*3.1%*	*0.5%*	*0.2%*	*0.26%*	*0.90%*
Average income	€10,647	€8,792	€23,439	€37,956	€23,578
Gini	0.624	0.70	0.44	0.40	0.57

Notes: The table presents the distribution of pre-tax fiscal income per adult, in the survey before the correction and after the correction using our method and the replacing alternative used in Section 3. Average incomes are expressed in French Euros PPP. Brazil and Chile refer to 2015, while all the European countries refer to 2014.

Figure E.6: Discrepancy of income across datasets in Chile and Brazil: 2015



Reading: in 2015 the total income declared in tax data in Brazil, which covers 20% of the population represents 49% of national income. The total income in the raw survey represents 58% of national income and 74% in the corrected survey, which are both representative of the entire population. The equivalent income calculated from national accounts represents 85% of national income. Authors' calculations using data from surveys, income tax declarations and national accounts.

References

Aaberge, Rolf and A. B. Atkinson (2010). “Top incomes in Norway”. In: *Top incomes: a global perspective* 2.

Abowd, John M. and Martha H. Stinson (2013). “Estimating measurement error in annual job earnings: A comparison of survey and administrative data”. In: *Review of Economics and Statistics* 95.5, pp. 1451–1467.

Alstadsæter, Annette et al. (2016). *Accounting for business income in measuring top income shares: Integrated accrual approach using individual and firm data from Norway.* Tech. rep. National Bureau of Economic Research.

Alvaredo, Facundo (2011). “A note on the relationship between top income shares and the Gini coefficient”. In: *Economics Letters* 110.3, pp. 274–277. DOI: 10.1016/j.econlet.2010.10.008. URL: http://dx.doi.org/10.1016/j.econlet.2010.10.008.

Alvaredo, Facundo et al. (2017). “Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world”.

Angel, Stefan, Richard Heuberger, and Nadja Lamei (2017). “Differences Between Household Income from Surveys and Registers and How These Affect the Poverty Headcount: Evidence from the Austrian SILC”. In: *Social Indicators Research* 138.2, pp. 1–29. ISSN: 15730921. DOI: 10.1007/s11205-017-1672-7.

Atkinson, A. B. (2007). “The distribution of top incomes in the United Kingdom 1908–2000”. In: *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries.* Ed. by A. B. Atkinson and Thomas Piketty. Vol. 1. Oxford University Press, pp. 82–140.

Atkinson, A. B. and Thomas Piketty (2007). *Top incomes over the twentieth century: a contrast between continental European and English-speaking countries.* Oxford University Press, p. 585. ISBN: 9780199286881. URL: https://global.oup.com/academic/product/top-incomes-over-the-twentieth-century-9780199286881?lang=en&cc=fr.

— (2010). *Top incomes: a global perspective.* Oxford University Press, p. 776. ISBN: 9780199286898. URL: https://global.oup.com/academic/product/top-incomes-9780199286898?cc=fr&lang=en&#.

Ayer, Miriam et al. (1955). “An Empirical Distribution Function for Sampling with Incomplete Information”. In: *Ann. Math. Statist.* 26.4, pp. 641–647. DOI: 10.1214/aoms/1177728423. URL: https://doi.org/10.1214/aoms/1177728423.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). “Generalized Pareto Curves: Theory and Applications”.

Bollinger, Christopher R (1998). “Measurement error in the Current Population Survey: a nonparametric look.” In: *Journal of labor economics* 16.3, pp. 576–594. ISSN: 0734-306X. DOI: 10.1086/209899.

Bound, John and Alan B. Krueger (1991). “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?” In: *Journal of Labor Economics* 9.1, p. 1. ISSN: 0734-306X. DOI: 10.1086/298256.

Bourguignon, François (2018). “Simple adjustments of observed distributions for missing income and missing people”. In: *The Journal of Economic Inequality*, pp. 1–18.

Brunk, H D (1955). “Maximum Likelihood Estimates of Monotone Parameters”. In: *Ann. Math. Statist.* 26.4, pp. 607–616. DOI: 10.1214/aoms/1177728420. URL: https://doi.org/10.1214/aoms/1177728420.

Burkhauser, Richard V, Markus H Hahn, and Roger Wilkins (2016). “Top Incomes and Inequality in Australia: Reconciling Recent Estimates from Household Survey and Tax Return Data”.

Burkhauser, Richard V et al. (2016). “What has Been Happening to UK Income Inequality Since the Mid-1990s? Answers from Reconciled and Combined Household Survey and Tax Return Data”. URL: http://www.nber.org/papers/w21991.

— (2018). “Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK’s SPI Adjustment?” In: *Fiscal Studies* 39.2, pp. 213–240. ISSN: 14755890. DOI: 10.1111/1475-5890.12158.

Chancel, Lucas and Thomas Piketty (2017). “Indian income inequality, 1922-2014: From British Raj to Billionaire Raj?” URL: http://wid.world/document/chancelpiketty2017widworld/.

Cristia, Julian and Jonathan A. Schwabish (2009). “Measurement error in the SIPP: Evidence from administrative matched records”. In: *Journal of Economic and Social Measurement* 34.1, pp. 1–17. ISSN: 07479662. DOI: 10.3233/JEM-2009-0311.

Czajka, Léo (2017). “Income Inequality in Côte d’Ivoire: 1985-2014”. In: *WID.world Working Paper* July.

Deming, W. Edwards and Frederick F. Stephan (1940). “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known”. In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444. DOI: 10.1214/aoms/1177731829. URL: https://projecteuclid.org/euclid.aoms/1177731829.

Deville, Jean-Claude (2000). “Generalized calibration and application to weighting for non-response”. In: *COMPSTAT: Proceedings in Computational Statistics 14th Symposium held in Utrecht, The Netherlands, 2000.* Ed. by Jelke G Bethlehem and Peter G M van der Heijden. Heidelberg: Physica-Verlag HD, pp. 65–76. ISBN: 978-3-642-57678-2. DOI: 10.1007/978-3-642-57678-2{_}6. URL: https://doi.org/10.1007/978-3-642-57678-2_6.

Deville, Jean-Claude and Carl-Erik Särndal (1992). “Calibration Estimators in Survey Sampling”. In: *Journal of the American Statistical Association* 87.418, pp. 376–382. DOI: 10.1080/01621459.1992.10475217.

Diaz-Bazan, Tania (2015). “Measuring Inequality from Top to Bottom”. In: *Policy Research Working Paper* 7237.

DWP (2015). “Households Below Average Income: An analysis of the income distribution 1994/95 – 2013/14”. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437246/households-below-average-income-1994-95-to-2013-14.pdf.

Eeden, Constance van (1958). “Testing and Estimating Ordered Parameters of Probability Distributions”. PhD thesis. University of Amsterdam.

Fairfield, Tasha and Michel Jorratt De Luis (2016). “Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile”. In: *Review of Income and Wealth* 62, S120–S144.

Fleming, Kirk G (2007). “We’re Skewed—The Bias in Small Samples from Skewed Distributions”. In: *Casualty Actuarial Society Forum* 2.2, pp. 179–183.

Flores, Ignacio et al. “Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017”. In: *Review of Income and Wealth*. DOI: 10.1111/roiw.12441. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12441. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12441. Forthcoming.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2016). “Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)”. URL: http://piketty.pse.ens.fr/files/GGP2016DINA.pdf.

— (2018). “Income inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA)”. In: *Journal of Public Economics* 162, pp. 63–77.

Hlasny, Vladimir and Paolo Verme (2017). “The impact of top incomes biases on the measurement of inequality in the United States”.

— (2018). “Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data Vladimir”. In: *Econometrics* 6.30, pp. 1–38. DOI: 10.3390/econometrics6020030.

Hoerl, Arthur E and Robert W Kennard (2000). “Ridge Regression: Biased Estimation for Problems Nonorthogonal”. In: *Technometrics* 42.1, pp. 80–86. ISSN: 0040-1706. DOI: 10.1080/00401706.2000.10485983.

Jenkins, Stephen P (2017). “Pareto Models, Top Incomes and Recent Trends in UK Income Inequality”. In: *Economica* 84.334, pp. 261–289. ISSN: 14680335. DOI: 10.1111/ecca.12217.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion (2006). “Survey nonresponse and the distribution of income”. In: *Journal of Economic Inequality* 4.1, pp. 33–55. ISSN: 15691721. DOI: 10.1007/s10888-005-1089-4.

Kuznets, Simon (1953). *Shares of Upper Income Groups in Income and Savings*. NBER. ISBN: 087014054X. DOI: 10.2307/2343040. URL: http://www.jstor.org/stable/10.2307/2343040?origin=crossref.

Lesage, Éric, David Haziza, and Xavier D'Haultfoeuille (2018). "A cautionary tale on instrument vector calibration for the treatment of unit nonresponse in surveys".

Medeiros, Marcelo, Juliana de Castro Galvão, and Luìsa de Azevedo Nazareno (2018). "Correcting the Underestimation of Top Incomes: Combining Data from Income Tax Reports and the Brazilian 2010 Census". In: *Social Indicators Research* 135.1, pp. 233–244.

Morgan, Marc (2018). "Essays on Income Distribution: Methodological, Historical and Institutional Perspectives with Applications to the Case of Brazil (1926–2016)". PhD Dissertation in Economics. Paris: Paris School of Economics & EHESS.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2018). "From Soviets to oligarchs: inequality and property in Russia 1905-2016". In: *The Journal of Economic Inequality* 16.2, pp. 189–223.

Okolewski, Andrzej and Tomasz Rychlik (2001). "Sharp distribution-free bounds on the bias in estimating quantiles via order statistics". In: *Statistics and Probability Letters* 52.2, pp. 207–213. ISSN: 01677152. DOI: 10.1016/S0167-7152(00)00242-X.

Pareto, Vilfredo (1896). *Écrits sur la courbe de la répartition de la richesse.*

Paulus, Alari (2015). "Income underreporting based on income expenditure gaps: Survey vs tax records". URL: http://hdl.handle.net/10419/126467.

Pedace, Roberto and Nancy Bates (2000). "Using administrative records to assess earnings reporting error in the survey of income and program participation". In: *Journal of Economic and Social Measurement* 26, pp. 173–192. ISSN: 07479662.

Piketty, Thomas (2003). "Income Inequality in France, 1901–1998". In: *Journal of Political Economy* 111.5, pp. 1004–1042. DOI: 10.1086/376955. URL: http://www.journals.uchicago.edu/doi/10.1086/376955.

Piketty, Thomas and Emmanuel Saez (2003). "Income Inequality in the United States, 1913–1998". In: *Quarterly Journal of Economics* CXVIII.1.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States". In: *Quarterly Journal of Economics* 133.May, pp. 553–609. DOI: 10.1093/qje/qjx043.Advance.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2017). "Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015". URL: http://www.nber.org/papers/w23368.pdf.

Singh, A C and C A Mohl (1996). "Understanding Calibration Estimators in Survey Sampling". In: *Survey Methodology* 22.2, pp. 107–115.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge (2015). "What Are We Weighting For?" In: *Journal of Human Resources* 50.2, pp. 301–316. ISSN: 0022-166X. DOI: 10.3368/jhr.50.2.301. URL: http://jhr.uwpress.org/lookup/doi/10.3368/jhr.50.2.301.

Taleb, Nassim Nicholas and Raphael Douady (2015). “On the super-additivity and estimation biases of quantile contributions”. In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260. ISSN: 03784371. DOI: `10.1016/j.physa.2015.02.038`. URL: `http://dx.doi.org/10.1016/j.physa.2015.02.038`.