ECINEQ
Society for the Study of Economic Inequality

Working Paper Series

Missing top incomes and tax-benefit microsimulation: evidence from correcting household survey data using tax

Marko Ledic

Ivica Rubil

Ivica Urban

ECINEQ 2022 609

ECINEQ 2022 609
2022 March
www.ecineq.org

Missing top incomes and tax-benefit microsimulation: evidence from correcting household survey data using tax records data

Marko Ledic
Faculty of Economics & Business, University of Zagreb
Ivica Rubil
The Institute of Economics, Zagreb
Ivica Urban
Institute of Public Finance, Zagreb

Abstract

Using the microsimulation model EUROMOD for Croatia, we compare the results of simulation based on the original survey data (EU-SILC) with those based on the survey data corrected using tax records data and a recent survey correction method. We show that the correction method, although it debiases inequality estimates, may not be able to correct the in-come structure by source if some income sources are severely under-represented. In Croatia, this is the case for income from capital, property, and contractual work. As a solution, we propose to complement the correction method with an ad hoc pre-correction procedure. The corrections bring the aggregate amount, distribution, and structure of survey income closer to those in the tax data. Consequently, the simulated fiscal instruments become more like those in the tax data. Simulation of a hypothetical tax reform shows the results based on the uncorrected data may be misleading in terms of the estimated budgetary impact and the distributional incidence of the reform.

Keyword: top incomes, survey data, tax records, tax-benefit microsimulation, EUROMOD, EU-SILC

JEL Cassification: D31, H24

Missing top incomes and tax-benefit microsimulation: evidence from correcting household survey data using tax records data

Marko Ledić*, Ivica Rubil**, Ivica Urban***

March 11, 2022

Abstract

Using the microsimulation model EUROMOD for Croatia, we compare the results of simulation based on the original survey data (EU-SILC) with those based on the survey data corrected using tax records data and a recent survey correction method. We show that the correction method, although it debiases inequality estimates, may not be able to correct the income structure by source if some income sources are severely under-represented. In Croatia, this is the case for income from capital, property, and contractual work. As a solution, we propose to complement the correction method with an *ad hoc* pre-correction procedure. The corrections bring the aggregate amount, distribution, and structure of survey income closer to those in the tax data. Consequently, the simulated fiscal instruments become more like those in the tax data. Simulation of a hypothetical tax reform shows the results based on the uncorrected data may be misleading in terms of the estimated budgetary impact and the distributional incidence of the reform.

Keywords: top incomes, survey data, tax records, tax-benefit microsimulation, EUROMOD, EU-SILC

JEL codes: D31, H24

* Faculty of Economics & Business, University of Zagreb; Trg J. Kennedyja 6, 10000 Zagreb; mledic@efzg.hr
** The Institute of Economics, Zagreb; Trg J. Kennedyja 7, 10000 Zagreb; irubil@eizg.hr
*** Institute of Public Finance; Smičiklasova 21, 10000 Zagreb; ivica.urban@ijf.hr

Acknowledgements: This work was fully supported by the Croatian Science Foundation, grant number IP-2019-04-9924. The paper uses EUROMOD version I3.0+. EUROMOD is maintained, developed and managed jointly by the Institute for Social and Economic Research (ISER) at the University of Essex and the Joint Research Centre of the European Commission, collaborating with national teams from the EU member states. The authors are indebted to the many people who have contributed to the development of EUROMOD. The process of extending and updating EUROMOD is financially supported by the European Union Programme for Employment and Social Innovation "Easi" (2014–2020). The results published and the related observations and analysis may not correspond to results or analysis of the data producers.

1. Introduction

Tax-benefit microsimulation is an important tool for answering factual and counterfactual questions on budgetary and distributional impacts of tax-benefit policies, suitable for both *ex ante* and *ex post* policy evaluation (Figari, Paulus, and Sutherland 2015). Many governments use this tool to analytically support policy making, and it is used in academic research as well. Since tax-benefit microsimulation models typically use household-level survey data, data deficiencies can in principle affect simulation results. This paper deals with that issue.

The deficiency of survey data that we focus on concerns the fact that surveys usually fail to properly represent the right tail of the distribution. As convincingly evidenced in the inequality literature, top incomes in surveys tend to be underrepresented relative to incomes lower down the distribution, which biases inequality estimates downward (e.g., Atkinson, Piketty, and Saez 2011; Burkhauser et al. 2012; Jenkins 2017). This problem of missing income at the top has multiple causes: lower willingness of the rich to participate in the survey (higher unit non-response) or to answer (some) income questions (higher item non-response); a stronger tendency to underreport their incomes; top-coding, trimming, or censoring (Hlasny and Verme, 2018b; Lustig 2020).

This deficiency of survey data can bias the results of tax-benefit microsimulation in several ways. The possible biases concern both the budgetary impact and distributional impact of tax-benefit instruments. Regarding the budgetary impact, missing income at the top means lower aggregate income, a part of which is the base for personal income taxes.[1] Consequently, depending on the tax-benefit system, the aggregate simulated income tax may be lower than if income were not missing. Even if no income is missing, the simulated aggregate tax can be biased downward if the structure of income by source is unrepresentative. For example, suppose some part of income from a relatively heavily taxed source is missing, say due to underrepresentation of the rich, which is offset by too much of income from a relatively lightly taxed source, due to overrepresentation of the non-rich. If so, the aggregate amount of total (i.e., from all sources) income will be correctly represented, but the aggregate tax will be biased downward. However, perhaps the more likely situation is that a part of the aggregate income is missing at the top, but in different proportions for different sources: say, the rich participate in the survey less than the rest, and capital income is missing relatively more than employment income, for the former is more concentrated at the top.

[1] For ease of presentation, we speak of personal income taxes only, but the same may apply to social insurance contributions.

When income is missing predominantly at the top, there can as well be consequences for the measurement of the distributional impact of the tax-benefit system. Depending on the tax-benefit design and the shape of the income distribution before taxes and benefits, the missing income issue can make tax-benefit instruments appear more (or less) redistributive or progressive than they really are. This holds for baseline simulations of the actual tax-benefit system, as well as for reform simulations. Thus, to reduce the possible biases in microsimulation results, corrections of survey data are warranted.

There are several approaches to correcting survey income, motivated mostly in the inequality literature aiming to gauge accurate levels and trends of inequality measures. Reviewing them, Lustig (2020) distinguishes between within-survey approaches, which rely on information from the survey,[2] and survey-cum-external data approaches,[3] which use both survey and external information.[4] In the latter, the external information is usually from tax records, which are supposed to represent well, if not perfectly, the income distribution at the top.

In this paper, we use the latter approach. For Croatia in 2017, we leverage data on the population of individual tax returns and the method of Blanchet, Flores, and Morgan (2022), complemented with our own *ad hoc* pre-correction procedure, to make corrections to the income data in the EU-SILC[5] survey, used in the European Union tax-benefit microsimulation model EUROMOD.[6] Central to the paper is a comparison of the EUROMOD results based on the corrected data with those based on the original data to see how the corrections affect the results of microsimulation. We compare the simulated aggregates and distributions of fiscal instruments, as well as their distributive features (progressivity and redistributive effects), in both baseline and reform settings.

In correcting the survey income data for missing top incomes, the Blanchet-Flores-Morgan method (Blanchet, Flores, and Morgan 2022) uses information from the tax data to first adjust the population weights by inflating them in the right tail above an endogenously determined income level and reducing them in the rest of the distribution. These adjustments are meant to reduce the deficiencies arising from unit non-response, item non-response, and underreporting, which cannot be rectified by increasing the sample size. After reweighting, the method

[2] E.g., Atkinson and Micklewright (1983); Korinek, Mistiaen, and Ravallion (2006; 2007); Jenkins (2017); Hlasny and Verme (2018a; 2018b); Hlasny (2021).

[3] E.g., Burkhauser et al. (2018a; 2018b); Medeiros, Castro Galvao, and Azevedo Nazareno (2018); Bartels and Metzing (2019); Chancel and Piketty (2019); Blanchet, Flores, Morgan (2022); Ooms (2021).

[4] Lustig (2020) also considers approaches that amount to dropping survey data completely and using a completely different source instead, such as tax records, national accounts, and other data. But, strictly speaking, such approaches are not correction approaches at all.

[5] European Union Statistics on Income and Living Conditions

[6] See https://euromod-web.jrc.ec.europa.eu/

artificially oversamples at the top and replaces top survey incomes with incomes at the same percentiles in the tax data. This is meant to reduce the deficiencies that can be reduced only by increasing the sample size: first, to reduce the sparsity of the data at the top, as this sparsity reduces the precision of the estimates of statistics based on top observations; and second, to extend the support of the survey income distribution, as the maximum income is smaller in the survey than in the tax data.

For accurate tax-benefit microsimulation, survey income must be correct in terms of both amount and structure by source, since in many countries the tax treatment of income depends on its source (type). Although, in principle, the Blanchet-Morgan-Flores method can correct survey data in both respects, it may fail to do so when some income sources are missing to a large extent. Intuitively, the reason is: when the survey income is unrepresentative in several respects, the method has too many "tasks" to do, and the more tasks there are, the more likely some will be mutually conflicting. The tasks are set by specifying appropriate constraints to reweighting, and some of them may be incompatible and must be dropped. But after dropping them, there is no guarantee that all the necessary corrections will be done. This is what happens in our case with incomes from contractual work, property, and capital. To correct the data, we propose a procedure for *ad hoc* pre-correction of these income sources, after which the Blanchet-Morgan-Flores method performs the correction satisfactorily without specifying too many constraints. In a nutshell, our pre-correction amounts to "sowing" the survey with the largely missing income sources, with the knowledge of where and how much to sow being based on comparing the survey and tax data.

We begin our analysis by showing how the information on the income concept relevant for tax-benefit microsimulation is unrepresentative in the original, uncorrected survey, both in terms of amount and structure by source. Turning to survey correction, we show that in our case, the Blanchet-Flores-Morgan method fails to perform the required corrections. Against that background, we next show that by combining our *ad hoc* pre-correction and the Blanchet-Flores-Morgan method, one can deliver the required corrections. The analysis is based on comparing the uncorrected and the two versions of corrected survey data with the tax data in terms of various statistics such as the aggregate amounts and inequality and concentration measures.

We then proceed to examine how the survey corrections – considering only the combination of the pre-correction and Blanchet-Flores-Morgan method – affect the results of tax-benefit microsimulation with EUROMOD. We first simulate the baseline tax-benefit system in the tax year 2017 using both the uncorrected and corrected data and show that the corrections substantially affect the results. Most notably, the amount of personal income tax and surtax simulated

with the uncorrected data falls substantially short of the amount recorded in the tax data, while the gap closes when the corrected data are used. Moreover, the unrepresentative structure of income by source in the uncorrected data makes the structure of the simulated personal income tax different from the one in the tax data. Again, the discrepancy largely vanishes once we use the corrected data. When using the corrected data, the concentration of the personal income tax with surtax, as well as the social insurance contributions, with respect to the relevant income concept also gets closer to what it is in the tax data. When it comes to the distributive impact of the baseline tax-benefit system, there is a difference between the corrected and uncorrected data in terms of the size and the contributions of different fiscal instruments.

Besides the baseline simulation, we also simulate a reform to the personal income tax. The reform is hypothetical and designed to emphasise the impact of the survey corrections. It consists in: (i) extending the baseline schedule with two income brackets (with the respective tax rates of 24 and 36 percent) by adding a third bracket with a higher rate (48 percent); (ii) doubling the rate (i.e., a flat rate of 12 percent) applied to income from property and capital. We show that the reform simulation with the uncorrected data is misleading in terms of the budgetary and distributional impacts of the reform. With the uncorrected data, the revenue increase is significantly smaller than it should be, and its structure departs from the structure one would reasonably expect, given the reform and the baseline system. As for the distributional impact of the reform, it is stronger with the corrected data, since the reform affects the top of the distribution more than the rest, and the corrected data are more representative at the top.

Our main contribution is to the practice of tax-benefit microsimulation based on survey data. Thus far, the motivation for correcting survey income was mainly to obtain unbiased inequality estimates. We provide evidence that tax-benefit microsimulation practice should consider correcting the underlying data to obtain unbiased results on budgetary and distributional impacts of fiscal instruments. We also show how the recently developed Blanchet-Flores-Morgan correction method and tax records data can be leveraged to perform the needed corrections. Relatedly, we show that the method may fail to deliver the required corrections when too much of some income sources is missing and propose a way to pre-correct the survey so that the Blanchet-Flores-Morgan method can yield the required corrections.

We are not the first to do survey income correction to improve tax-benefit microsimulation, but the literature is rather scarce (Creedy 2004; Creedy and Tuckwell 2004; Myck and Najsztub 2015; Jara and Oliva 2018; Brzezinski, Myck, and Najsztub 2021). Our paper differs in several respects. First, method-wise, while those papers use either reweighting or income replacement, the Blanchet-Flores-Morgan method combines reweighting with oversampling at the top plus

income replacement to tackle non-sampling and sampling related deficiencies, respectively. We are also the only ones to consider how to correct the structure of income by source, which is important for accurate tax-benefit simulation. Second, analysis-wise, we do an extensive, detailed analysis of how the corrections amend the distributive impact of the tax-benefit system, considering not only the system as a whole, but also different types of instruments (taxes, social insurance contributions, social benefits).[7] Third, data-wise, of the papers that use tax records data as external information, we use the most comprehensive data: we have the entire population of income earners, and not only those, say, who paid a positive amount of tax or voluntarily submitted a tax return; all income sources are included, along with the information on their taxable and non-taxable parts. Such data enable us to do the pre-correction procedure that we propose.

The paper is organised in this way. In section 2, we describe the data and the Croatian tax-benefit system and define the relevant income concepts. In section 3, we present the Blanchet-Flores-Morgan correction method, the way we implement it, and the impact of the corrections on income and socio-economic characteristics. In section 4, we analyse the impact on the results of tax-benefit microsimulation. In section 5, we summarise the paper, discuss its implications, and conclude.

2. Data, tax-benefit system, and definitions

In this section, we first describe the survey and tax data that we use. Then we describe in broad lines the Croatian tax-benefit system in 2017. Finally, we define the relevant income concepts with reference to the data sources and the tax-benefit system.

2.1. Data

The survey data come from the European Union Statistics on Income and Living Conditions (EU-SILC) for the year 2018, where income information refers to 2017. EU-SILC is the data source for the official EU statistics on income distribution, inequality, living conditions, and poverty, reported by Eurostat. Most importantly for our purpose, it is the survey data used by EUROMOD, the tax benefit model that we use (section 4.1). The Croatian component of EU-SILC is collected and administered by the Croatian Bureau of Statistics.[8]

[7] Jara and Oliva (2018) perform a distributive analysis, but a very simple one.
[8] The survey's Croatian name is *Anketa o dohotku stanovništva*.

EU-SILC contains information on income and several household and individual characteristics. Income information is detailed, collected through a series of questions on the receipts from various sources such as employment, self-employment, other work arrangements (e.g., contractual work), pensions, assets (physical and financial), as well as social benefits. All information in the Croatian component of EU-SILC is based on the respondents' self-reports, which lowers the quality of the data compared to the countries where the information has increasingly been based on administrative records (e.g., the Scandinavian countries) (see Törmälehto, Jäntti, and Marlier 2017). Regarding the sample size, the 2018 sample we use includes 8,383 households or 21,243 individuals.

The tax data come from administrative records of the Tax Authority unit of the Croatian Ministry of Finance and refer to the tax year 2017. The records' main purpose is the calculation of taxpayers' tax obligations. It is a large database that we have compiled using multiple data files, each containing specific records at the level of individual persons. The separate data files are merged into a single database using anonymised person-specific identification numbers.

The tax records provide information on individual income receipts, both their taxable and non-taxable portions, from multiple sources, including employment, self-employment, contractual work, pension, property, and capital. Unlike in some other countries, where the tax records cover only the population of actual taxpayers, namely those with a positive tax liability, in Croatia a person is covered if (s)he earned any income that is in principle subject to personal income tax, that is, any income from a source considered taxable as per the Act on Personal Income Tax, namely income from employment, self-employment, contractual work, pension, property, and capital. Whether (s)he in fact paid any tax depends on the amount earned and other factors, such as the number of dependent family members and the place of residency.[9] The universe of persons covered by the records numbers about 3.2 million.

Besides the income information, there are also information on the amounts of personal income tax, social insurance contributions, enabling us to benchmark the aggregate amounts based on survey-based simulation in EUROMOD. Some personal characteristics are also provided (e.g., place of residency, sector/industry of activity, number of supported family members), but the available information of this kind is very limited as compared to EU-SILC.

[9] For example, a person earning wages will appear in the records since employment income is a taxable income source, even if his wages are low enough (and/or he has many dependent children) to result in a positive tax liability. On the contrary, a person living on social transfers will not appear in the records, as social transfers are fully non-taxable.

Unfortunately, there is no information required to ascertain which individuals belong to the same family. For this reason, we cannot use the tax data for those parts of the analysis where households are the relevant units. Concretely, we cannot examine the distributive impact of the tax-benefit system, a piece of analysis that we can do using the survey data only. Therefore, these results will be compared only between the original and corrected versions of the survey data (see Sections 4.2 and 4.3).

2.2. Tax-benefit system

We describe only the main instruments of the Croatian tax-benefit system in the year 2017. We focus on the personal income tax with a related surtax and social insurance contributions, both employees' and employers', which are expected to be affected by the missing survey income and the corrections that we do. Concerning social benefits, they are not expected to be affected, and only two main benefits are described. Indirect taxes are not touched upon at all, as the Croatian component of the microsimulation model EUROMOD that we use does not simulate them.[10]

The personal income tax is levied on income from diverse sources: employment, self-employment, contractual work, property and capital income, and pensions.[11] The tax unit is an individual earning income from a taxable income source as per the Act on Personal Income Tax.[12] The tax base is gross income lowered by a basic personal allowance and a supplemental allowance for dependent family members. For wage employment income, the employees' social insurance contributions are a deduction. There are general schedule and source-specific schedules. The general schedule is applied, on a yearly basis, to income from employment, self-employment, contractual work, and pensions. It consists of two brackets, a lower and an upper bracket, with rates of 24 and 36 percent, respectively. Pensioners' tax liability is 50 percent of the liability according to the general schedule. The source-specific schedules are applied to property and capital incomes, with a flat rate of 12 percent. There is a personal allowance (basic allowance for everyone plus additional allowance for dependent family members) for incomes taxed through the general schedule, but not for those taxed through the specific schedules.

The personal income tax is accompanied by a surtax, for which the base equals the personal income tax obligation, and the rate is set by the authorities at the level of towns and

[10] For a comprehensive description of the tax-benefit system, see Urban, Bezeredi, and Pezer (2020) or Nguyen and Rubil (2021). The microsimulation model is introduced in section 3.2.

[11] In general, social benefits are not taxed.

[12] Official Gazette, no. NN 115/16, NN 106/18, NN 121/19, NN 32/20, and NN 138/20.

municipalities. Statutory restrictions on the maximum rate apply: 10 percent for municipalities; 12 (15) percent for towns with a population below (above) 30 thousand.[13] We will use the label "PITS" to refer to the sum of the personal income tax and surtax.

The social insurance contributions, labelled "SIC/SICs" hereafter, include general health, occupational health, employment insurance and pension insurance contributions. SICs are paid on income from wage employment, self-employment, and contractual work. In the case of self-employment income, the SIC base is a lump sum amount, depending on the type of self-employed person. For the other income types, in general, the base is equal to gross income, with a floor for all types of contributions, and a ceiling for pension contributions. In the case of employment income, employees pay pension insurance contributions, while employers pay all other contributions. Income from contractual work is subject only to the general health contributions (paid by the service purchaser) and pension contributions (paid by the contractual worker). The rates vary by type of SIC: 15 percent for the general health contribution; 0.5 percent for the occupational health contribution; 1.7 for the employment contribution; and 20 percent for the pension insurance contributions. On pensions higher (lower) than the average national net monthly wage pay, a special pensioner health contribution is paid, at a rate of 3 (1) percent.[14]

Among several social benefits, the two most important are the child benefit and the guaranteed minimum benefit. The child benefit is an income-tested benefit for households with children aged up to 15 or 19 if in education. There are three levels of the amount per child, depending (inversely) on household income per member. There are also certain pro-natalist supplements for households with three or more children.

The guaranteed minimum benefit is an asset- and income-tested benefit scheme for households with incomes below a "minimum subsistence" amount. The latter amount varies depending on household type; for example, a single or a couple; with or without children; lone- or two-parent family. The benefit amount equals the gap between the minimum subsistence amount and the net household income per member, averaged over the preceding three months. In addition, the beneficiaries are eligible to receive additional compensations for electricity and

[13] There are no restrictions on the minimum rate. About half of all towns and municipalities have a surtax rate of zero. Zagreb, the capital, with its surtax rate of 18 percent, is an exception to the restrictions on the maximum rate.

[14] For the poorer group, the 1-percent contribution is paid from the central government budget. The contributions for certain other social groups, such as the unemployed and persons on maternity/parental leave, are also paid from the central government budget.

housing costs, provided by cities and municipalities, subject to certain a ceiling related to the amount of guaranteed minimum benefit.

2.3. Definitions of income concepts

We use three income concepts, namely *fiscal* income, *gross* income, and *disposable* income. Here we describe them in turn.

What we call *fiscal* income is precisely income as recorded in the tax data: income before PITS and SICs. Perhaps it would be more self-explanatory if we called it gross income – which in fact it is: income gross of PITS and SICs – but we avoided this term on purpose, reserving it for another income concept (to be introduced shortly). We recognise six categories or sources of fiscal income: employment income, self-employment income, pension income, contractual work income, property income, and capital income.[15] The first three sources of fiscal income are self-explanatory, while the last three deserve clarification. Contractual work income is income earned from work that cannot be classified as either employment or self-employment.[16] Property income refers to income derived from physical property (e.g., renting real estate) and property rights. And capital income captures income from financial capital (e.g., dividends or interest). For the sum of all six sources, we use the terms total fiscal income and fiscal income interchangeably.

Fiscal income, both total and disaggregated by sources, is available in the survey data as well.[17] This enables us to compare fiscal income between the two sources, and we do that in terms of aggregate amount, as well as in terms of how total fiscal income and its sources are distributed. As we said in the data section, in the tax data we cannot form households, and thus when making distributional comparisons, any distributional statistics that we use is computed by treating individuals as the units of analysis.

The inability to form households in the tax data is also the reason why the other two income concepts – gross income and disposable income – are available in the survey data only. In our use, both are equivalised household incomes, so that neither of them can be constructed without knowing which individuals belong to different households. In both cases, the household total income is divided by the number of adult equivalents according to the OECD-modified

[15] This categorisation is well in line with the Act on Personal Income Tax; see Official Gazette, no. NN 115/16, NN 106/18, NN 121/19, NN 32/20, and NN 138/20.

[16] Contractual work involves provision of services agreed upon between a contractual worker and a service purchaser. Examples include a musician's concert performance and an engineer's project study.

[17] That is, in both data there is the information required to create the six sources of fiscal income.

equivalence scale. Substantively, gross income is defined as income before taxes (PITS and property taxes) and SICs (excluding employers' SICs).[18] More precisely, it is equal to fiscal income plus private transfers received and market income from sources not considered taxable by the Act on Personal Income Tax. Disposable income is obtained from gross income by subtracting taxes and SICs (excluding employers' SICs) and adding social benefits. Both definitions are standard in the literature, as is the usage of the two income concepts in distributional analyses of the kind we do.

3. Correcting survey data using tax data

In this section, we first describe the method of Blanchet, Flores, and Morgan (2022) (hereafter: the B-F-M method). Then we describe our implementation of the method and the pre-correction that we do to improve its performance.

3.1. B-F-M method

The method can correct survey data for under-representation of top incomes, which, as argued by B-F-M, can arise from two types of survey errors: non-sampling and sampling errors. To explain them, suppose that the sampling design is such that every unit (household or individual) in the population has the same *a priori* probability of being included in the sample; that is, there is no intention that some population groups be under-represented.[19]

Non-sampling errors are those that have nothing to do with the sample size and thus cannot be reduced by increasing it. There are, in principle, three reasons for these errors. First, the phenomenon of lower *ex post* probability of being sampled for units with higher income, which is an example of differential unit non-response. It happens when the richer units are more difficult to reach or when, once reached, they are more likely than the poorer units to refuse to be interviewed. Second, the phenomenon of lower probability for the richer to answer questions on income, which is an example of differential item non-response. And third, the phenomenon of income under-reporting rising with income. As said, increasing the sample size does not help

[18] In section 4.2 we nevertheless consider how the survey corrections affect the aggregate simulated amount of the employers' SICs. These contributions are not included only where we examine how the corrections affect the distributive impact of the tax-benefit system.

[19] Note that this assumption is not inconsistent with the practice of excluding some population groups – usually those in institutions like hospitals, nursing homes, or prisons – from the *population of interest*. Individuals from such groups have a zero a priory probability of being sampled, but in the population of interest everyone has the same positive a priori probability of being sampled.

with these issues, as it cannot make the richer easier to reach, more willing to be interviewed, and, if interviewed, to report (truthfully) their incomes.

On the contrary, sampling errors can be reduced by sampling more units because they originate in insufficient sample size. When the sample is relatively small, some small groups in the population – and the rich are such a group – can hardly be well represented, as relatively few units are sampled. This data sparsity at the top of the income distribution increases the sampling variability (i.e., reduces the precision) of the estimates of statistics depending on the sparse data, but, perhaps more importantly, the estimates may also be biased (e.g., Taleb and Douady 2015). A related issue is that the support of the true income distribution, as represented by the tax distribution, may extend beyond the maximum income in the survey distribution. Increasing the sampling size helps resolve these issues. The distribution of unit and item non-response and the nature of income under-reporting are unaffected by the sample size. Thus, increasing the sample – across the entire distribution or by oversampling at the top – mechanically increases the number of top income observations. This reduces the degree of sparsity and raises the probability of capturing the very top of the true distribution, thereby increasing the overlap in the support at the top.

The B-F-M method is meant to deal with both sampling and non-sampling errors, and it works as follows. Let us, for simplicity, use the term "survey (tax) distribution/density" for "distribution/density of income in the survey (tax) data." For now, to convey the basic intuition, we focus only on the distribution of the income concept to be corrected, disregarding all other survey content, such as individual and household characteristics and subcomponents (sources) of the income concept being corrected.

The method deals with non-sampling errors by way of reweighting (i.e., adjusting population weights) the survey. Consider Figure A with three densities: the survey, tax and reweighted survey density. Using a data-driven procedure, the method first looks for the point (percentile) in the survey distribution above which the latter distribution should be corrected using the tax distribution; that is, above which the two distributions will be merged. This point is called the *merging* point and is indicated by the income level y^{m} in Figure 1.[20] To the right of it, the reweighted survey density resembles the tax density: the original population weights of those to the right of the point y^* are adjusted upwards, and those to the left of it downwards. The weight adjustments add up to zero, which is necessary for the density integral over the support

[20] For details on the determination of the merging point, see B-F-M. Here it suffices to say that the merging point is determined in such a way that the reweighted survey density be continuous.

of survey distribution to remain unchanged (i.e., equal to one). This shows that it makes no sense to state that a survey is unrepresentative at the top only, because if the rich are under-represented, the rest are over-represented by necessity. The result of reweighting is shown in figure A as the transformation of the survey density into the reweighted survey density, whereby the latter becomes perfectly aligned with the tax density for all y to the right of y^{m}.

[Figure 1 near here]

As said, the above intuition considers only the income distribution. But there is other survey content as well, namely a set of demographic and socio-economic characteristics, as well as various income sources making up the income concept subject to correction. Typically, surveys are representative in terms of some but not all of these variables. The reweighting within the B-F-M method is in principle flexible enough to preserve survey representativeness for the variables in terms of which the survey is already representative, as well as to enforce it for the variables in terms of which the survey is not representative. In the former case, information from the very survey is sufficient, whereas in the latter case, trustable external information is required.[21] In both cases, the information is leveraged to form a set of constraints to reweighting. In a nutshell, the reweighting procedure amounts to choosing new weights as a solution to a program minimising the discrepancy between the new and original weights subject to the constraints.[22] Formally, in a survey with sample size n, where the original and new weights are, respectively, d_i and w_i, the latter are found by solving the program

$$\min_{w_1,\ldots,w_n} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \quad \text{subject to} \quad \begin{pmatrix} \text{constraint } 1 \\ \vdots \\ \text{constraint } M \end{pmatrix}. \tag{1}$$

When it comes to the constraints, one always applies: the sum of new weights must equal the sum of original weights, which equals the population size. Other constraints depend on what is to be preserved in the survey or enforced on it. For example, to enforce that the survey total of a variable x, which is considered incorrect, be equal to a true total t obtained from a trustable external source, one would impose the constraint $\sum_{i=1}^{n} w_i x_i = t$. If, on the contrary, one believes the survey total is correct and would like to preserve it after reweighting, one would impose the constraint $\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} d_i x_i$. These are simple ones, but the framework is general enough to allow more complex constraints. For example, suppose one wishes to impose that the share of capital income in total income in each income group $g = 1, 2, \ldots, G$ be equal to the share s^g

[21] For example, if the composition of survey income by source is unrepresentative, the external information on income, say tax records, must contain information on the composition of income by source.

[22] It is grounded in survey calibration theory (e.g., Deville and Särndal 1992).

observed somewhere outside the survey. Denoting capital income by y^{cap}, total income by y^{tot}, and a dummy for belonging to income group g by b_i^g, the relevant constraint is $\sum_{i=1}^{n} w_i b_i^g y_i^{\text{cap}} = \sum_{i=1}^{n} d_i b_i^g s^g y_i^{\text{tot}}$, for every group g.[23]

Reweighting as just described can correct the survey for non-sampling errors. It can do so by virtue of being able to inflate the weights of the under-represented and to deflate the weights of the over-represented appropriately. However, reweighting cannot help with sampling errors.[24] It does not include increasing the sample size, it just reweights the present sample. Observations at the top remain sparse even though their weights are inflated. And if the support of the tax distribution extends beyond that of the survey distribution, the same holds for the reweighted survey distribution: raising the weight of the richest individual in the survey does not increase the income level itself to the maximum observed in the tax data.

B-F-M deal with sampling errors in this way. After reweighting, the sample is increased by artificial over-sampling and income replacement at the top, which can be described as follows. A specified number of observations – a fraction of the existing ones – are added at the top of the reweighted survey by multiplication of the existing observations. Incomes are then assigned to them according to their relative position in the income distribution: the observation whose income corresponds to quantile q of the reweighted survey distribution gets the income of the observation at the same quantile of the tax distribution. The values of other survey variables are assigned to each new observation by "transplantation" from a similar observation, namely a randomly chosen observation out of a specified number of nearest neighbours by income. Finally, the weights are appropriately readjusted to ensure that they add up to the population size.

3.2. Implementation

We use the B-F-M method to correct the Croatian component of EU-SILC survey using tax records data from the Croatian Tax Administration. Here we describe how we implement it.

In principle, when correcting survey data to obtain more accurate results of tax-benefit microsimulation, the choice of income concept to be corrected depends on the tax-benefit system. For example, if the tax treatment of income is the same for all income sources/types, then it suffices to correct total income as the sum of incomes from multiple sources. But, typically, the parameters of personal income tax and social insurance contributions vary with an income source. The differences need not be parametric only; different income sources can be subject to

[23] These examples are from B-F-M; see the paper for other examples and analytical details.
[24] It can do so somewhat, but only asymptotically (Deville and Särndal 1992).

different types of taxes and social insurance contributions. In such cases, for accurate microsimulation, not only total income, but also incomes from different sources, must be accurate. Only in one specific, rather unlikely, case correct total income would automatically mean the multiple sources are correct as well – when the shares of different income sources in total income are correct. This holds when in the survey all income sources are under-captured by the same relative amount. Otherwise, correcting total income does not suffice, and one must do corrections of the multiple sources as well.

In the case of Croatia, both the tax-benefit system rules and the nature of income under-coverage in EU-SILC call for correction of both total income and its different sources. The Croatian tax-benefit system, as we have seen above (see section 2.2), does not treat all income sources equally. When it comes to the personal income tax, while capital and property incomes are taxed at a flat 12-percent rate, income from other sources (employment, self-employment, contractual work, pension) is taxed through a two-rate general schedule, with a 24-percent rate for the lower income bracket and a 36-percent rate for the upper bracket. Importantly, the latter general schedule includes personal allowance.[25] Besides, capital and property incomes are not subject to SICs, and the rates for the other sources vary by source (among other factors).

Concerning income under-coverage, its extent is not the same for all income sources. From the perspective of tax-benefit microsimulation, in Croatia, the relevant income concept is fiscal income (see section 2.3), total as well as disaggregated by source. Comparing the aggregate amounts between the survey and tax data, we see there are substantial differences across the sources of fiscal income in the discrepancy between the aggregates in the two data sets. While the survey employment income is only 7 percent below the figure in the tax data, the gap is much larger for contractual work income (60 percent) and property income (57 percent), and especially for capital income (98 percent).[26] On the contrary, the survey aggregates of self-employment and pension incomes overshoot the tax-data aggregates by 22 and 5 percent, respectively.

Thus, we must correct both total fiscal income and its sources. The B-F-M method, as described in the preceding section, corrects a specified income concept. If one wants to correct its multiple sources too, in principle, one could do so by specifying appropriate constraints. For example, in the preceding section we have shown a constraint that could help enforce a specific share of capital income in the income concept being corrected. Specifying analogous constraints

[25] There is a basic allowance to which everyone is eligible, and additional allowances for dependent family members.

[26] On missing capital income in other countries, see Yonzan et al. (2020) and Ooms (2021).

for other income sources would, in principle, do the job of correcting them all. But the numerical optimisation to solve program (1) for a set of new weights (section 3.1) may fail to converge if there are (too) many constraints. Simply, the more constraints there are, the more likely it is that some are incompatible with one another. Especially because one of the constraints specifies how much an original weight may be maximally deflated or inflated to prevent reweighting to produce too large or negative new weights for some observations.

Our attempt at correcting total fiscal income and its six sources has failed in the sense that the optimisation did not converge.[27] Our initial preferred setup included the following constraints. First, two constraints are meant to preserve the gender and age distributions, which are in line with the census, and a constraint to preserve the number of individuals earning employment income, which is in accordance with that in the tax data. Second, a set of constraints meant to enforce the shares of the six income sources in total fiscal income to be as in the tax data; these were specified for several income brackets along the distribution. The optimisation failed to converge even when we reduced the number of income brackets for the latter set of constraints, thereby reducing the number of constraints. Nor there was a success after allowing larger maximum adjustments of the weights.[28] We achieved convergence only when we kept the first set of constraints and dropped the second. In the resulting dataset, total fiscal income is corrected – the shortfall from the tax data is closed. But its composition by source improved only a little: the aggregate amounts of contractual work, property, and especially capital incomes remained after the correction substantially below the corresponding amounts in the tax data, although there were some improvements.[29] The reason is this: absent the appropriate constraints, which we had to drop to achieve convergence, the reweighting procedure was tuned to correct total fiscal income and not its sources. And the subsequent step, the artificial oversampling and income replacement at the top, was unable by construction to do anything but reduce data sparseness at the top and extend the support of the survey distribution of total fiscal income.

To correct both total fiscal income and its sources, we opted for an additional *ad hoc* correction, a pre-correction applied before running the B-F-M method. The idea is to pre-correct those sources of fiscal income that are very under-represented in the survey data relative to the

[27] We used the Stata module BFMCORR, written by Blanchet, Flores, and Morgan and available at the Boston College Department of Economics Statistical Software Components archive (Blanchet, Flores, and Morgan 2018): https://ideas.repec.org/c/boc/bocode/s458567.html.

[28] By default, BFMCORR prevents that a new weight falls below 1/5 of the original weight or increase above five times the original weight. We set the floor and ceiling to 1/10 and 10, respectively

[29] Concrete and detailed results are presented and discussed in section 3.3.

tax data. In our case, these are contractual work, property, and capital incomes. Comparing the survey and tax data, we observe that the main reason why income from these sources is severely under-represented is that there are too few people in the survey reporting a positive receipt from these sources. Concretely, relative to the tax data (=100), the number of individuals reporting a positive amount of capital, property, and contractual work incomes is 17, 43, and 29 percent, respectively.[30] The pre-correction is not meant to be perfect: it does not alone close the gap in the aggregate amounts of the three income sources between the survey and tax data. One reason is that we do not touch the incomes of those that reported receiving them. Rather, the pre-correction closes the gap in the number of individuals receiving a positive income from each of these sources. The idea is that the pre-correction corrects the survey data to such a degree that the B-F-M method can subsequently complete the job and deliver the desired result.

The pre-correction can be presented as consisting in four steps. Here we refer to capital income, but the same procedure applies to contractual work and property incomes as well. The steps are:

Step 1. We sort all individuals in the survey and tax data in ascending order by "reduced income," which is the sum of employment, self-employment, and pension incomes. In each data, we group individuals into 20 income brackets ($i = 0, 1, \dots, 19$), where one group ($i = 0$) contains only those with zero reduced income.

Step 2. We divide each income bracket (except the one where everyone has zero reduced income) into three subgroups according to the main (i.e., largest) income source (employment, self-employment, or pension income). This gives 19 subgroups of individuals whose main income is employment income ($i = A1, \dots A19$), and similarly for self-employment income ($i = B1, \dots, B19$) and pensions income ($i = C1, \dots, C19$). Together with the zero-reduced-income bracket ($i = 0$), there are 58 subgroups in total.

Step 3. For each subgroup i in the tax data, we calculate the share of individuals with positive capital income, denoted by R_i, respectively. Now, we calculate the *desired* number of survey individuals with a positive capital income, that is, how many survey individuals *should* have a positive capital income: $D_i = R_i N_i$, where N_i is the size of subgroup i in the survey data. Finally, we calculate how many individuals in the survey *should but do not* have a positive capital income: $\Delta_i = D_i - A_i$, where A_i is the *actual* number of survey individuals with a positive capital income.

[30] The figures are substantially higher for employment, self-employment, and pension incomes: 90, 89, and 83 percent, respectively.

Step 4. For each subgroup *i* in the tax data, we take the set of individuals with a positive capital income and top-code the incomes at the 99th percentile of the distribution in this set. Then we randomly draw Δ_i values of capital income from the top-coded distribution and assign each of them to one of Δ_i *adult* individuals with zero capital income randomly drawn from subgroup *i* in the survey.[31]

Loosely speaking, the pre-correction can be described as "sowing" poorly represented income sources – in our case, contractual work, property, and capital incomes – across the cells defined by the level and primary source of what we termed "reduced income."

The pre-corrected data are used as input to the B-F-M method to obtain the final corrected data. The same constraints are specified as in the case without the pre-correction (see above). Other settings of the B-F-M method are also common to both cases: the share of population above the merging point that is added while artificially over-sampling at the top is set to 10 percent; the number of nearest neighbours to randomly choose from while assigning characteristics to the additional observations is set to ten.

Before proceeding, let us set some labels to be used from here on. The original, uncorrected survey data will be referred to as "ORIG" data. ORIG data pre-corrected according to the four-step procedure will be referred to as "pre-corrected ORIG" data. The data corrected using the B-F-M method only, without the pre-correction, will be referred to as "BFM" data. And the data to which we apply both the pre-correction and B-F-M method will be referred to as "PC-BFM" data.

3.3. Results of survey corrections

Appendix table A2 summarises basic information on the B-F-M corrections of ORIG data and the pre-corrected ORIG data, resulting in BFM and PC-BFM data, respectively. In both corrections the pre-correction sample increases, because of the artificial oversampling at the top, to over 28 thousand households (column 4). The merging point is at the 94th percentile of the tax data (column 5), which corresponds to the 95.4th and 94.2nd percentile of ORIG and the pre-corrected ORIG data, respectively (column 6). Equivalently, 4.6 and 5.2 percent of the population in, respectively, ORIG and the pre-corrected ORIG data are above the merging point (column 8). Subtracting each of the latter percentages from 6 percent (column 7), which is the share of population above the merging point in the tax data, we get how much of total population

[31] Appendix Table A1 shows an illustration of the pre-correction procedure for capital income and a subset of subgroups.

is missing above the merging point in the survey: 1.4 percent in ORIG data and 0.2 percent in the pre-corrected ORIG data (column 9). Here we see that the pre-correction reduces the missing population above the merging point. Tiny part of the missing population are individuals with an income outside the support of the fiscal income distribution (columns 10 and 11).

In Appendix Table A3, we look at a few basic individual characteristics in ORIG, BFM, and PC-BFM data. We consider the entire distribution, as well as the parts below and above the merging point. We have seen that the merging points are not the same (see the preceding paragraph), but here we approximate them by the 94th percentile. First, note that at the entire-sample level, there are virtually no differences between the three data sets in the gender, age, and activity compositions. This is because of the constraints to reweighting that we specified (see section 3.2). We observe small differences only for the educational composition. The same holds for the bottom 94 percent because of the size of this group relative to the entire population. For the top 6 percent there are some differences between the data sets. The differences between BFM and ORIG data tend to be somewhat bigger than those between PC-BFM and ORIG data. This is because when the pre-correction is done, less work is left for the B-F-M method to do while reweighting the sample, and consequently, the composition by characteristics is less affected.

We now examine how the corrections affect the amount of fiscal income and its sources. The aggregate amounts are displayed in panel A of Table 1. According to the tax data, the actual aggregate of total fiscal income amounts to HRK 191.5 billion, but only 90 percent of it was captured in the uncorrected survey data (ORIG). Panel A of Figure 2 shows the distribution of the missing fiscal income across selected quantile groups based on the distribution of total fiscal income. The white bars, which refer to ORIG data, show that the missing fiscal income is missing predominantly from the top of the distribution: out of the HRK 19.6 billion missing, as much as 17.3 billion (bar height) or 88.2 percent (underlined percentages above bars) pertains to the top 20 percent. Indeed, almost all of it pertains to the top 10 percent (85.2 percent), most of it to the top 5 percent (75.9 percent), and about half of it to the top 1 percent (52.5 percent).

[Table 1 near here]

[Figure 2 near here]

Considering the sources of fiscal income, we see that the total missing income is due to missing employment, contractual work, property, and capital incomes, while self-employment and pension incomes are somewhat over-captured in ORIG data (see Table 1). In the case of self-employment income, the excess amount might partly be due to the income truthfully reported in the survey but hidden from the tax authority. And in the case of pension income, the

excess might originate from some social benefits for low-income pensioners being reported as pensions.

Although employment income is the largest income source, its contribution to the total missing income is the second largest. The largest contribution pertains to capital income: of the HRK 9.7 billion of it in the tax data, very little is captured in ORIG data, only 2 percent. Incomes from contractual work and property rentals are greatly under-captured in ORIG data, with only 40 and 43 percent of the actual amounts being captured, respectively. Panels B to D of Figure 2 show that the missing employment, contractual work, property, and capital incomes, respectively, pertains mostly to the top of the distribution. Of the whole missing employment income, 94.7 percent is missing in the top 10 percent, and the corresponding figures for contractual work and property incomes are 71.2 and 52.2 percent, respectively. Especially concentrated at the top is the missing capital income: 87.1, 84.3, and 75.8 percent of the missing aggregate comes from, respectively, the top 10, 5, and 1 percent.[32]

Applying the B-F-M method brings the missing fiscal income partly or fully back. Either without the pre-correction described in section 3.2 (resulting in BFM data) or with it (resulting in PC-BFM data), the aggregate fiscal income becomes practically equal to the aggregate reported in the tax data, with very little overshooting (see Table 1). However, while the two versions of the corrected data lead to practically the same improvements in terms of total fiscal income, there are substantial differences between them when the sources of fiscal income are considered. In general, the B-F-M method brings the survey data closer to the tax data more when the pre-correction is done than when it is not done. For example, whereas the under-coverage of employment income in ORIG data (index of 93, relative to 100 in the tax data) turns into an over-coverage of the same size in BFM data (index = 107), in PC-BFM data the over-coverage practically vanishes (index = 101). Moreover, for self-employment income, the over-coverage, in fact, increases considerably with the Blancher-Flores-Morgan correction when the latter is not coupled with the pre-correction, but considerably less so when the pre-correction is done. The advantages of PC-BFM compared to BFM data are also obvious for contractual work and property income. But the difference between BFM and PC-BFM data is especially striking in the case of capital income: BFM data represent a negligible improvement over ORIG data, while PC-BFM data represent a strikingly large improvement, where almost 90 percent of the capital income recorded in the tax data is captured in PC-BFM data.

[32] For evidence on missing capital income at the top in other countries, see Yonzan et al. (2020) and Ooms (2021).

Going back to the distribution of missing fiscal income (total) on panel A of Figure 2, we see that the grey and black bars, representing the missing income (or excess income if negative) in BFM and PC-BFM data, respectively, are very small for top quantile groups, with little difference between the two corrected data. But the differences between them must again be emphasised when it comes to specific income sources. Consider employment income first. For the top 20, 10, 5, and 1 percent, there is too much income in BFM data; indeed, for the top 5 and 1 percent the gap with respect to the tax data is larger than before the corrections (i.e., the excess after the correction is larger than the shortfall before the correction), suggesting that the corrections should not be done. But this is not the case for PC-BFM data, where there is a clear improvement. The advantage of PC-BFM data over BFM data is also evident for the remaining income sources and most of the quantile groups considered, especially in the case of capital income. That is because, when some income sources are missing to a large extent, which is the case with contractual work, property, and capital incomes, the structure of fiscal income by source is severely unrepresentative, and appropriate constraints must be specified for the B-F-M method to properly reweight the survey. The structure by source along the distribution can be enforced only by specifying numerous constraints to the reweighting. Yet the more constraints there are, the more likely it becomes that some of them will be mutually incompatible, resulting in a failure to find the right new weights. That happened in our case, and we had to drop the constraints meant to impose the right structure of fiscal income at different parts of its distribution, which resulted in BFM data. But, as we can see, the structure improved greatly once we made the pre-correction before applying the B-F-M method.

The tax and survey data may differ not only in terms of the coverage of aggregate fiscal income, but also in terms of the number of income recipients. Panel B of Table 1 shows the number of persons receiving a positive amount of fiscal income, total and by sources. In general, the discrepancy relative to the tax data is bigger than in the case of amounts. Of all recipients of fiscal income from any source, only 82 are covered by ORIG data. The discrepancy, both absolute and relative, is largest for the population receiving capital income: while the size of this population is 870 thousand according to the tax data, only 152 thousand or 17 percent are represented in ORIG data. The relative shortfalls from the actual number of recipients are also large for contractual work and property incomes. In our case, the B-F-M correction is of little help in reducing the discrepancies if one does not do the pre-correction as well. In contrast, when preceded by the pre-correction, the Blanchet-Flores-Morgan correction helps in reducing the discrepancies. In PC-BFM data, the shortfall of the population receiving income from any source is reduced to 9 percent, compared to the shortfall of 18 percent in ORIG data. Again, the

improvements are striking for the populations of recipients of contractual work, property, and capital incomes, especially the latter.

In Table 2, we consider the composition of the aggregate fiscal income, that is, the shares of the income sources in the total. The composition of fiscal income by sources in ORIG data differs notably from the true composition in the tax data. This is due to the fact, evident from panel A of Table 1, that the relative difference in the aggregate amount of income between ORIG data and the tax data is not the same across the sources. The shares of employment, self-employment, and pension incomes are larger in ORIG than in the tax data, and the opposite holds for contractual work, property, and capital incomes (panel A of Table 2). The corrections yielding BFM data bring some improvement, though very little, while the corrections resulting in PC-BFM data lead to a considerable improvement. This is confirmed by the values of the Hellinger distance (HD), used here as a measure of (dis)similarity between two compositions (panel B of Table 2). For two compositions, labelled A and B, each with K components, denoted $\{c_1^A, \dots, c_K^A\}$ and $\{c_1^B, \dots, c_K^B\}$, the HD between them is defined as

$$HD(A,B) = \sqrt{\sum_{k=1}^{K} \left(\sqrt{c_k^A} - \sqrt{c_k^B}\right)^2} \Big/ \sqrt{2} \ . \qquad (2)$$

It ranges from 0 to 1, where 0 indicates identical compositions, and 1 indicates completely different compositions.[33] While for BFM data the HD is only marginally smaller than for ORIG (0.144 vs. 0.147), for PC-BFM data the indicator is largely reduced relatively to ORIG, coming very close to zero (0.013).

[Table 2 near here]

Besides the composition by sources at the level of the entire distribution, we also consider the composition for selected quantile groups of the distribution of total fiscal income. In Figure 3, we can see that, in general, the composition in the tax data is most closely resembled by the one in PC-BFM data, and this is confirmed by the HD (reported in parentheses next to bar labels), which on all panels takes the lowest value for PC-BFM data. While PC-BFM data are a clear winner here, there is not much difference between ORIG and BFM data: by the HD, the former wins on three panels and the latter on four panels, but in general, their HDs are quite similar.

[Figure 3 near here]

In Table 3, we look at how the corrections affect the level of inequality in the fiscal income distribution. All indicators in this table are calculated from data where the unit observation is

[33] According to the Hellinger distance, the compositions A and B are completely different, that is, maximally distanced from one another, if all components in A with a positive share have a zero share in B, or *vice versa*.

an individual person, and all individuals are considered, including those with no fiscal income at all, such as children or inactive adults (other than pensioners). Inequality of (total) fiscal income, as measured by the Gini coefficient and the income shares of the top 10, 5, and 1 percent, is substantially biased downward in ORIG data: the Gini coefficient is 0.607 compared to 0.625 in the tax data, and the top income shares are all lower by about 5 percentage points compared to the respective shares in the tax data. In BFM and PC-BFM data, the Gini is substantially higher and closer to that in the tax data. In both cases, it exceeds the Gini in the tax data, but the overshoot is smaller in PC-BFM data. The top income shares after the survey corrections resemble those in the tax data almost perfectly, with only minor differences between the BFM-based and PC-BFM-based shares. Thus, when it comes to inequality of total fiscal income, it does not matter substantially whether the pre-correction is done or not, as in both cases the B-F-M method reduces (in the case of Gini) or virtually removes (in the case of top income shares) the bias that is present in the uncorrected data. This is because the B-F-M method is geared to correct the distribution of total income at the top, and it does that irrespective of the type of pre-correction that we do and that distinguish BFM data from PC-BFM data. The pre-correction is helpful only when one is interested in fixing not only the distribution of total income at the top, but also the distributions of its multiple sources along the whole distribution.

[Table 3 near here]

In sum, the original, uncorrected EUROMOD data suffer from significant under-coverage of fiscal income, mainly due to missing employment and capital income at the top of the fiscal income distribution. Correcting the data by the B-F-M method makes up for the missing fiscal income and reduces the bias in the estimates of inequality measures. But if one aims to fix not only the distribution of total fiscal income at the top, but also the whole distributions of its multiple sources, which is required for the correct structure of fiscal income by source, the B-F-M method might need to be accompanied by pre-corrections of the income sources that are missing from the survey to a large extent.

4. Impact of corrections on tax-benefit microsimulation

In this section, we report on the impact of the survey data corrections on the results of tax-benefit microsimulation. We first describe the microsimulation model EUROMOD (section 4.1). Then we turn to the impact on simulation of taxes and benefits for both the baseline system (section 4.2) and a hypothetical tax reform (section 4.3).

4.1. Tax-benefit microsimulation model

For simulation of taxes, SICs, and social benefits, we use the tax-benefit microsimulation model EUROMOD (Sutherland and Figari 2013).[34] This is a static non-behavioural[35] model that currently simulates the tax-benefit systems of all EU member states. It can simulate the baseline systems for several years, as well as reforms to the tax-benefit instruments modelled, including the introductions of new instruments. The model is survey-based and uses the EU-SILC survey data (see section 2.1), namely the data that we have called ORIG so far.

When it comes to the Croatian tax-benefit system (see section 2.2), the model simulates several instruments, including PITS, SICs, and several social benefits (e.g., child benefit, the guaranteed minimum benefit, unemployment benefit, maternity leave benefit, etc.). For lack of information in EU-SILC, some taxes and benefits are not modelled, but rather taken as reported in the survey or completely left out. An example is personal property taxes that appear in our tables in the results section, which are taken as reported by the survey respondents.[36] Although the model has a module for simulation of indirect taxes, it still does not include Croatia.[37]

ORIG data set used in EUROMOD does not contain information on the taxable and non-taxable receipts of certain sources of fiscal income. The taxable/non-taxable division is relevant in Croatia for employment, pension, and capital incomes. Using ORIG data, EUROMOD assumes that the whole amounts of these income sources are taxable, which leads to excessive simulated amounts of PITS and SICs paid on these income sources relative to the amounts that would be simulated if one knew the size of the taxable receipts. In what follows, we do nothing about it when using ORIG data, since we consider this data set as a benchmark for comparison and thus want it to give results that anyone else using EUROMOD out-of-the-box would obtain. In addition, failing to split incomes into the taxable and non-taxable receipts does not lead to excessive PITS and SICs relative to the amounts recorded in the tax data, since a lot of fiscal income is missing in ORIG. However, when using PC-BFM data, where the missing income is brought back practically completely, assuming the whole income is taxable would lead to excessive amounts of PITS and SICs relative to those recorded in the tax data. For that reason, we

[34] See also the model website: https://euromod-web.jrc.ec.europa.eu. We use the model version I3.0+.

[35] A static model refers to a single period, rather than a multi-period dynamics. A non-behavioural model does not consider behavioural reactions of households/individuals (e.g., changes in labour supply) to changes in tax-benefit instruments.

[36] For a comprehensive description of how the Croatian tax-benefit system is modelled within EUROMOD, see the country report by Urban, Bezeredi, and Pezer (2020).

[37] See De Agostini et al. (2017) and Acoğuz et al. (2020).

leverage the tax data to approximate the taxable/non-taxable split. The procedure, similar in spirit to the pre-correction procedure from section 3.2, is presented in Appendix.

4.2. Impact on baseline simulation of taxes and benefits

Having examined the impact of the survey corrections on fiscal income and its sources, we now turn to the impact on the results of tax-benefit microsimulation.

Table 4 displays the aggregate amounts of taxes, social insurance contributions, and social benefits. For the largest source of fiscal revenue, the social insurance contributions (SICs), the amount simulated in ORIG data is practically in line with the actual amount recorded in the tax data. This is driven by the employers' and employees' SICs, which make up about 95 percent of total SICs, whose amounts are practically fully aligned with those in the tax data. This might be surprising given the finding from the previous section that a non-negligible portion of employment income is missing from ORIG data. The reason behind these seemingly inconsistent findings is that, according to the SIC rules, there is a ceiling to the base of the pension contributions (paid by employees; see section 2.2.), so that the SIC amounts stop rising with income once income reaches the ceiling. Thus, the SICs collected from employment income should not be missing to the same extent as employment income itself. After the survey corrections, in PC-BFM data, there is almost perfect alignment between the employers' and employees' SICs in the tax data. The same holds for total SICs given the relative size of the employers' and employees' SICs.

[Table 4 near here]

Unlike the SICs, the personal income tax and surtax (PITS) is under-simulated in ORIG data relative to the tax data, only 74 percent. Somewhat more than half (54 percent) of the missing PITS pertains to that collected through the general schedule (see section 2.2), which applies to income from employment, self-employment,[38] pension, and contractual work, with employment income as the largest source due to its relative size. Relative to the general-schedule PITS in the tax data, we are able to simulate only 84 percent using ORIG data. A further 38 percent of the missing PITS is due to missing PITS from capital income. Since, as we have seen in the preceding section, capital income is practically completely missing from ORIG data, the simulated PITS from this source is also almost fully missing: the simulated amount is only 2 percent of the amount in the tax data. Somewhat less than 8 percent of the missing PITS pertains

[38] A small part of self-employment income is taxed through the lump-sum schedule.

to the PITS from property income and the rest to the PITS collected through the lump-sum schedule.

The survey corrections increase the simulated amounts of PITS. For total PITS, we see that the whole gap relative to the tax data is closed; there is only a little overshoot of 2 percent. The same holds for the PITS collected through the general schedule, again with a small overshoot. In the case of PITS from property income, there is, however, a large overshoot of 29 percent. Still, the gap created by this overshoot is smaller than the 45-percent shortfall in the case of ORIG data. The improvement achieved by the corrections is strikingly evident in the case of PITS from capital income, where the gap falls from 98 to only 15 percent. The lump-sum-schedule PITS remains virtually unaffected by the corrections.

We next examine how the SICs and PITS are concentrated across the distribution of fiscal income. Table 5 shows the coefficients of concentration of the SICs and PITS, total and disaggregated, with total fiscal income as the ranking variable. Comparing first ORIG data with the tax data, we see that the SICs and PITS simulated in ORIG data tend to be less concentrated than the actual SICs and PITS in the tax data: only for the SICs paid by self-employed persons and pensioners, the concentration index is higher in ORIG data. Comparing ORIG and PC-BFM data, the indices based on the latter are closer to those based on the tax data in eight out of 11 cases, with one draw. That PC-BFM data win here is closely related to the fact that in these data, the composition of fiscal income by source across its distribution is most like the structure observed in the tax data, as we have seen in Figure 2. This is because the distributional composition of fiscal income by source determines the distributional composition of SICs and PITS by type, and the latter in turn impacts the concentration indices for the different types of SICs and PITS, as well as total SICs and PITS.

[Table 5 near here]

The changes in the distribution of fiscal income and in the concentration of fiscal instruments with respect to fiscal income that we have discussed can, in principle, change the indicators of the redistributive impact of the net fiscal system.[39] We complete the analysis by considering them. While thus far the unit of observation was an individual, the results to be presented are based on households as the units of observation, a standard practice in analysing the distributive impact of the fiscal system. In addition, so far, we have used fiscal income (total and disaggregates by sources), whereas now we use two other income concepts: gross income and disposable income (see section 2.3). Since in the tax data there is no information required to

[39] By the net fiscal system, we mean considering all fiscal instruments together, i.e., taxes and social insurance contributions (revenues) and social benefits (expenses).

consider households as the unit of analysis, nor there is information on the amounts of property tax and social benefits, here we can only present the results based on the survey data.

The results are given in Table 6. The Gini coefficients for both gross (G_X) and disposable income (G_Y) increase substantially after the survey corrections (panel A of Table 6). The increase is smaller for gross income than for disposable income (10 vs. 14 percent), leading to a reduction in the redistributive effect of the net fiscal system, both absolute (R) and relative (R/G_X), by 8 and 17 percent, respectively (panel B of Table 6). Thus, the survey corrections affect the size of the redistributive effect substantially.

[Table 6 near here]

To account for the sources of the redistributive effect R, we decompose it into two sub-effects: the vertical effect V and the horizontal effect H (Kakwani 1984; 1986). The former measures what could be dubbed the "potential" size of the redistributive effect, which would arise if the net fiscal system caused no reranking among tax-benefit units (here households).[40] The decomposition, shown in panel B of Table 6, reveals that in both ORIG and PC-BFM data, V is much larger than H, and practically the whole difference in R between the two data sets is accounted for by the difference in V.

V can be further decomposed into the contributions pertaining to each of the fiscal instruments considered (or groups thereof) (Lambert 2001). We decompose it into the contributions of taxes (PITS and personal property taxes), SICs, and social benefits. Applying Lambert's (2001) decomposition, we decompose V as follows:

$$V = v_T + v_S + v_B \,, \tag{3}$$

with

$$v_j = \frac{t_j}{1 - (t_T + t_S - t_B)} K_j \tag{4}$$

for $j = T, S, B$, where t_j and K_j denote, respectively, the average tax rate and the Kakwani progressivity index (Kakwani 1977) of an instrument j, while T, S, and B stand for taxes, SICs, and social benefits, respectively.

The v_j's are shown in panels C, D, and E of Table 6, alongside the corresponding t_j's and K_j's. The last row of each panel shows the relative contribution (expressed in percent) of v_i to V. In ORIG data, the largest contribution pertains to social benefits (42 percent), followed by the contribution of taxes (35 percent). Social benefits contribute most due to their being much

[40] In this case the horizontal effect, measuring the effect of reranking, would be zero.

more progressive (according to the Kakwani index) than taxes and SICs, although their average rate is the smallest. After the survey corrections, the structure of V changes in that the contribution of taxes V_T rises relative to V_S and V_B: the relative contribution of V_T rises to 45 percent. This is driven by an increase in the average rate t_T upon the corrections, and for the following reason: the corrections increase the sum of PITS and property taxes (the denominator of t_T) relatively more than total fiscal income (the denominator of t_T).

Note that the Kakwani progressivity index for taxes (K_T) falls after the corrections. The corrections lead to an increase in the concentration of taxes at higher parts of the gross income distribution, as measured by D_T^X, since most of the missing income is brought back at higher parts of the gross income distribution. But this increase is smaller than the increase in the gross income Gini (G_X), and thus K_T as the difference between the two must fall.

To summarise, the corrections of survey incomes, discussed in the previous section, have a notable impact on the simulated amount of personal income tax. Once the missing income is brought back via corrections, so is the missing amount of personal income tax. Moreover, the corrections affect the concentration of taxes, SICs, and social benefits with respect to the distribution of gross income. Finally, the distributive impact of the net fiscal system is affected, as well as the relative contributions of the fiscal instruments considered.

4.3. Impact on reform simulation

Having examined in the preceding section how the corrections of survey incomes affect the statistics based on baseline tax-benefit microsimulation, we now extend the analysis by considering the simulation of a hypothetical reform of the tax-benefit system.

We consider a reform of the personal income tax. Two changes are introduced relative to the baseline. As described in section 2.2, the baseline general schedule consists of two tax brackets: one with a 24-percent rate that applies to the part of the tax base[41] up to HRK 17,500 per month, the other with a 36-percent rate that applies to the part of the base exceeding HRK 17,500 per month. Our reform adds a third bracket, and in the following way. The first bracket remains as is. In the second bracket, the 36-percent rate now applies to the part of the tax base between HRK 17,500 and 35,000 per month. And the new, third bracket, pertains to the part of the base above HRK 35,000 per month, which is taxed at a rate of 48 percent. Besides the general schedule, the reform introduces a change to the income source-specific tax rates.

[41] Tax base is not the same as gross income. The former is obtained by subtracting from the latter the SICs and the personal allowance (basic and for dependent family members).

Precisely, the baseline rate of 12 percent that applies to property and capital income is doubled to 24 percent. In every other respect, the tax-benefit system remains unchanged.

Before proceeding to simulation and analysis, we stress that we did not design the reform to reflect a proposal discussed in the Croatian public. Instead, it is purely hypothetical, designed with anticipation of certain differences in the simulation results between the original and corrected survey data that would best highlight the consequences for reform simulations of using the alternative survey data. Therefore, the reform includes changes aimed to affect those at the top of the distribution of incomes taxable under the general schedule, as well as those with property and capital incomes. As we have established, most of the missing incomes pertain to such persons.

We examine the budgetary and distributive impacts of the reform, beginning with the former. Table 7 shows the simulated baseline and reform amounts of revenue from PITS, total and by type, based on ORIG and PC-BFM data. For both datasets, the baseline amounts of PITS are reproduced from Table 4. Here we consider only PITS because the amount of SICs does not change by construction, and the change in social benefits is trivially small. Regarding the total PITS (the first row), we observe a striking difference in the absolute (in HRK billion) impact of the reform: while the simulation based on ORIG data suggests that the reform increases the PITS revenue by 0.38 billion only, the simulation based on PC-BFM data indicates the increase is more than five times larger, about 2.18 billion. The respective relative impacts (in percent of the total baseline PITS) also differ substantially (4.3 vs. 17.8 percent). Thus, by simulating the reform with the original, uncorrected data, one would severely underestimate the budgetary impact of the reform.

[Table 7 near here]

What accounts for the difference in the PITS revenue increase upon the reform between ORIG and PC-BFM data? The last two columns of Table 8 show how does the difference of 1.79 billion (= 2.18 billion – 0.38 billion) arise. We see that more than half of it (0.98 billion or 55 percent) is due to a larger increase in the capital-income PITS when the simulation is based on PC-BFM rather than ORIG data. The rest is due to larger increases in the general-schedule and property-income PITS, which account for 24 and 21 percent of the total difference, respectively. These findings are generally in line with what one would expect given our findings from section 4.2 on the structure of the missing income and on where in the distribution income is missing from. Specifically, recall that the single largest part of the missing income is capital income, that the second largest part is employment income (taxed through the general

schedule), and that most of the missing income is missing from the upper parts of the distribution of fiscal income.

Relatedly, if we relied on the simulation with ORIG data, we would be substantially misled regarding the structure of revenue increase upon the reform. To see that, compare columns 3 and 6, which display the structure of the revenue increase for ORIG and PC-BFM data, respectively. For ORIG data, about 3/4 of the increase is accounted for by an increase in the property-income PITS, while the general-schedule and capital-income PITS account for about 20 and 7 percent, respectively. For PC-BFM data, the structure is notably different: the single largest relative contribution is that of capital-income PITS (about 46 percent), followed by the contributions of the property-income PITS (about 30 percent) and general-schedule PITS (about 23 percent). Thus, by using ORIG data, one would markedly overestimate the contribution of property-income PITS and underestimate the contribution of capital-income PITS. This is because, although both incomes sources are concentrated more in the upper parts of the distribution, property income is less concentrated than capital income. (The same holds for the missing property and capital incomes.)

We next examine the distributive impact of the reform. The results are displayed in Table 8. The table is abbreviated in comparison to Table 6 in that the indicators for SICs and social benefits are dropped to focus on taxes as the object of reform. For all indicators in the table, the reform/baseline ratio is farther from one when the simulation is based on PC-BFM data, although in some cases, the difference is small. In other words, the reform has a stronger impact on the indicators considered − both the increases and decreases are somewhat larger in relative terms − when the simulation is based on the PC-BFM data. For example, while the ORIG-based simulation suggests that the Gini coefficient of disposable income (panel A of Table 8) falls only marginally (reform/baseline = 0.997), according to the PC-BFM-based simulation, the fall is not so small (ref./bas. = 0.984). As another example, consider the redistributive effect of the net fiscal system (panel B of Table 8), which increases by only a bit more than one percent (ref./bas. = 1.013) with the ORIG-based simulation, but more than ten percent (ref./bas. = 1.092) with the PC-BFM-based simulation.

[Table 8 near here]

Moreover, with both ORIG and PC-BFM data, the increase in the redistributive effect is predominantly due to an increase in the vertical effect (panel B of Table 8), and this increase is well-matched with the increase in the part pertaining to taxes (panel C of Table 8).[42] Note that

[42] Obtained by applying the Lambert (2001) decomposition; see Section 4.2.

the latter rises due to a rise in the average rate, even though the reform makes taxes somewhat less progressive according to the Kakwani index. That taxes become less progressive is due to the doubling of the PITS rate for property income, which, although concentrated among richer households more than among poorer, is present in non-negligible amounts along the *entire* distribution, not only at the top. In any case, as already said, all impacts of the reform are magnified to some extent when the simulation is done using PC-BFM data rather than ORIG data.

Finally, we complete the analysis by looking at the distributive impact of the reform in a more direct way. More concretely, we examine how the reform changes the disposable income of households at different parts of the gross income distribution. Figure 4 plots the mean equivalised disposable income across selected quantile groups based on the distribution of equivalised gross income. In panel A, the distribution is partitioned into ten decile groups. A first thing to note is that neither group benefits from the reform: the mean disposable income does not increase in any of them.

[Figure 4 near here]

Second, we see that the bottom nine decile groups are practically unaffected in ORIG data and affected very little in PC-BFM data, with losses of about half percent. This small difference between the data sets arises perhaps mainly because of the already noted fact that property income, for which the reform doubles the PITS rate, contributes to household incomes across the whole distribution (although more so towards the top). But only in PC-BFM data is property income present in quantities that are sufficient for the doubling of the PITS rate to cause non-zero losses in disposable income across the bottom 90 percent of the distribution.

Third, in ORIG data, the loss rises to close to 1 percent only in the top decile group, whereas in PC-BFM data it rises to about 3.5 percent. This difference between the top decile group and the rest is expected given the upward income adjustments at the top thanks to the corrections.

To get a finer picture of the reform impact within the top decile group, in panel B we zoom in on this group by partitioning it into top ten percentile groups. In ORIG data, the loss is smaller than 1 percent and rather homogeneous for all but the 100th percentile group, where it reaches 2 percent. In PC-BFM data, is roughly in the range of 1−2 percent over the 91st–97th percentile groups and starts increasing thereafter, to about 3 percent in the next two groups, and to as much as 9 percent at the very top. Thus, the loss is generally larger in PC-BFM data than in ORIG data, and the gap tends to increase along the gross income distribution, becoming large at the very top.

In sum, missing incomes (at the top and elsewhere in the distribution) in survey data may bias the microsimulation output not only for a baseline tax-benefit system, but also for a reform

to it, provided the reform involves changes that would affect the missing incomes if they were not missing. As we have seen, the biases concern both the budgetary and distributive impacts of reforms.

5. Summary, discussion, and conclusion

Various types of adjustments to household survey data, aimed at correcting for the known under-coverage of top incomes in surveys, have been proposed in the inequality literature to reduce the underestimation of survey-based inequality measures. Although top income under-coverage can, in principle, lead to biases in the results of tax-benefit microsimulation based on survey data, there have been few attempts to examine the consequences of missing top incomes for tax-benefit microsimulation.

We have contributed by providing such an analysis in the context of Croatia in 2017, using the Croatian component of the microsimulation model EUROMOD. The model is based on the EU-SILC survey data, which we have corrected using administrative individual-level data on the population of income receivers and applying the recent correction method by Blanchet, Morgan, and Flores (2022). As a methodological contribution, we have shown that to correct survey data for use in tax-benefit microsimulation, out-of-the-box usage of the Blanchet-Flores-Morgan method may not provide the required corrections. If some income sources (in our case, capital and property income) are severely missing from the survey, the method may have to be complemented by a purposefully designed *ad hoc* pre-correction.

We have shown that a significant amount of income is missing in the survey data relative to the tax data (about 10 percent). Most of this income is missing from the top of the distribution and comes from capital, of which only a negligible portion is present in the survey data (about 2 percent). The microsimulation results show that if the missing income is not brought back by the survey corrections, the simulated amount of personal income tax and surtax, especially the portion pertaining to capital income, will fall significantly short of the actual amount from the tax data. On the contrary, after the survey corrections, the simulated amount of personal income tax and surtax, including the part paid on capital income, differs very little from the amount recorded in the tax data. The distributive impact of the net fiscal system also depends on whether the survey data are corrected, with the redistributive effect increasing after the corrections. In addition, after the corrections, the contribution of taxes to the vertical effect increases relative to the contributions of other fiscal instruments (social insurance contributions and social benefits). Finally, we have shown that the survey corrections affect not only the baseline simulation

results, but also the results of reform simulation. Simulating a hypothetical reform to the personal income tax, designed to increase the tax burden of top income earners and receivers of income from capital and property, we have found that the simulation based on the uncorrected data seriously underestimate the budgetary and distributive impact of the reform.

These findings have implications for the practice of tax-benefit microsimulation. As most tax-benefit microsimulation models are based on survey data, and any survey data are likely to suffer to some extent from the issue of top income under-coverage, many existing models are likely to give biased results. Thus, it would be beneficial for both researchers and policy analysts to use these models to explore if and to what extent this issue is present in their models. Insomuch as there are issues of this sort, taking steps towards fixing them would improve the simulation accuracy and, consequently, the reliability and usefulness of the models for making research and policy insights. As our results suggest, this concerns both the budgetary and distributive impacts of fiscal instruments in the baseline as well as in reform settings.

Fixing the issues will depend on (i) the availability of reliable data, external to the survey, on the true income distribution, and (ii) on the extent of missing income, overall and by sources. As shown in recent inequality literature, suitable information may be found in tax records data. If available, these data can be used to correct the survey data following methods such as the one by Blanchet, Flores, and Morgan (2022), as we have done in this paper. An important lesson of our paper is that one may have to do certain income corrections prior to applying the Blanchet-Flores-Morgan method if there are income sources for which too much income is missing from the survey. We emphasise, however, that such corrections are not needed, in principle, if income from different sources is treated alike by the tax system. As we have shown, this is not the case in Croatia, where, for example, capital and property incomes, which are largely missing from the survey, are taxed at a flat rate markedly lower than the rates applicable to other income sources. It should also be said that doing such corrections imposes certain additional data requirements: besides information on total income, the tax data must contain information on income disaggregated by sources. Moreover, tax data often come in the form of tabulations of varying detail, rather than, as in our case, as individual-level micro-data. Although there are ways to synthesise micro-data from tabulated information (Blanchet et al. 2018; Blanchet, Fournier, and Piketty, forthcoming), it remains for further research to explore how precise the pre-correction of the sort we have done can be when based on synthetised, rather than genuine, micro-data.

This paper also has some broader implications concerning the quality of income information in EU-SILC and, consequently, the validity of research findings and official statistical

indicators based on this data. We have shown that the Croatian EU-SILC survey suffers from significant underrepresentation of top income receivers, a finding in accordance with those of Carranza, Morgan, and Nolan (2021), who have shown that this holds for some other countries too. The latter authors have also shown that the extent of underrepresentation, as measured by the increase in the Gini coefficient after survey corrections, tends to be smaller in countries where income information in EU-SILC is based more on administrative registers and less on the survey respondents' own reports.[43] That income information in EU-SILC is of higher quality when based on administrative sources was previously shown in Törmälehto, Jäntti, and Marlier (2017). Thus, our and other evidence suggest that the countries that rely little upon or not at all on administrative sources (including Croatia) should consider doing so for the sake of improvement in the quality of income data in EU-SILC. Alternatively, as suggested by Carranza, Morgan, and Nolan (2021), the national statistical offices, as the institutions collecting and administering EU-SILC, should consider correcting the survey on their own. Seconding this suggestion, we add that it would be useful if, should these corrections be done, EUROSTAT published two sets of estimates of the inequality measures and other statistics that may be sensitive to the data corrections.

Likewise, and directly related to the subject matter of this paper, we envision the practice of tax-benefit microsimulation using EUROMOD, as well as other survey-based models, would be enriched if researchers and policy analysts provided two sets of results: one based on the data as is, the other on corrected data. Exactly how the corrections should be done remains an open question, subject to further research, and we believe this paper is a useful example of how to go about doing so.

[43] See the paper for the classification of the EU countries by this criterion.

References

Acoğuz, E., B. Capéau, A. Decoster, L. De Sadeleer, G. Güner, K. Manios, A. Paulus, and T. Vanheukelom, 2020, "A new indirect tax tool for EUROMOD, final report", Joint Research Centre Project No. JRC/SVQ/2018/B.2/0021/OC

Atkinson, A. B., and J. Micklewright, 1983, "On the reliability of income data in the family expenditure survey 1970–1977", Journal of the Royal Statistical Society Series A, 146, 33–61.

Atkinson, A. B., T. Piketty, and E. Saez, 2011, "Top Incomes in the Long Run of History," Journal of Economic Literature, 49, 3−71.

Bartels, C., and M. Metzing, 2019, "An integrated approach for a top-corrected income distribution", Journal of Economic Inequality, 17, 125–143

Blanchet, T., I. Flores, and M. Morgan, 2018, "BFMCORR: Stata module for correcting surveys using tax data", Statistical Software Components S458567, Boston College Department of Economics, revised 25 Dec 2018.

Blanchet, T., I. Flores, and M. Morgan, 2022, "The weight of the rich: improving surveys using tax data", Journal of Economic Inequality, https://doi.org/10.1007/s10888-021-09509-3.

Blanchet, T., J. Fournier, and T. Piketty, forthcoming, "Generalized Pareto curves: theory and applications", Review of Income and Wealth.

Blanchet, T., B. Garbinti, J. Goupille, and C. Martinez-Toledano, 2018, "Applying generalized Pareto curves to inequality analysis", American Economic Association: Papers & Proceedings, 108, 114−118.

Brzezinski, M., M. Myck, and T. Najsztub, 2021, "Sharing the gains of transition: evaluating changes in income inequality and redistribution in Poland using combined survey and tax return data", European Journal of Political Economy, https://doi.org/10.1016/j.ejpoleco.2021.102121.

Burkhauser, R. V., S. Feng, S. P. Jenkins, and J. Larrimore, J., 2012, "Recent trends in top income shares in the USA: reconciling estimates from March CPS and IRS tax return data", Review of Economics and Statistics, 94, pp. 371–88.

Burkhauser, R. V., N. Herault, S. P. Jenkins, and R. Wilkins, 2018a, "What has been happening to UK income inequality since the mid-1990s? Answers from reconciled and combined household survey and tax return data", Oxford Economic Papers, 70, 301–326.

Burkhauser, R. V., N. Herault, S. P. Jenkins, and R. Wilkins, 2018b, "Survey under-coverage of top incomes and estimation of inequality: what is the role of the UK's SPI adjustment?", Fiscal Studies, 39, 213–240.

Carranza, R., M. Morgan, and B. Nolan, 2021, "Top income adjustments and inequality: an investigation of the EU-SILC", INET Oxford Working Paper No. 2021-16.

Chancel, L., and T. Piketty, 2019, "Indian income inequality, 1922–2015: from British Raj to billionare raj?", Review of Income and Wealth, 65, S33–S62.

Creedy, J., 2004, "Survey reweighting for tax microsimulation modelling", Research on Economic Inequality, 12, 229–249.

Creedy, J., and I. Tuckwell, 2004, "Reweighting the New Zealand Household Economic Survey for tax microsimulation modelling, Australian Journal of Labour Economics, 7, 71–88.

De Agostini, P., B. Capéau, A. Decoster, F. Figari, J. Kneeshaw, C. Leventi, K. Manios, A. Paulus, A. Sutherland, and T. Vanheukelom, 2017, "EUROMOD extension to indirect taxation", EUROMOD Technical Note Series EMTN 3.0.

Deville, J.-C., and C.-E. Särndal, 1992, "Calibration estimators in survey sampling", Journal of the American Statistical Association, 87, 376–382.

Figari, F., A. Paulus, and H. Sutherland, 2015, "Microsimulation and policy analysis", in A. B. Atkinson and F. Bourguignon (eds.), Handbook of Income Distribution, vol. 2, pp. 2141–2221. Amsterdam: Elsevier.

Hlasny, V., and P. Verme, 2018a, "Top incomes and the measurement of inequality in Egypt", World Bank Economic Review, 32, 428–455.

Hlasny, V., and P. Verme, 2018b, "Top incomes and inequality measurement: a comparative analysis of correction methods using the EU SILC data", Econometrics, 6.

Jara, H. X., and N. Oliva, 2018, "Top income adjustments and tax reforms in Ecuador", WIDER Working Paper Series, wp-2018-165, World Institute for Development Economic Research (UNU-WIDER).

Jenkins, S. P., 2017, "Pareto models, top incomes and recent trends in UK income inequality", Economica, 84, 261–289.

Kakwani, N. C., 1977, "Measurement of tax progressivity: an international comparison", Economic Journal, 87, 71–80.

Kakwani, N. C., 1984, "On the measurement of tax progressivity and redistributive effect of taxes with applications to horizontal and vertical equity", Advances in Econometrics, 3, 149–168.

Kakwani, N. C., 1986, Analyzing redistribution policies: a study using Australian data, Cambridge: Cambridge University Press.

Korinek, A., J. A. Mistiaen, and M. Ravallion, 2006, "Survey nonresponse and the distribution of income", Journal of Economic Inequality, 4, 33–55.

Korinek, A., J. A. Mistiaen, and M. Ravallion, 2007, "An econometric method of correcting for unit nonresponse bias in surveys", Journal of Econometrics, 136, 213–35

Lambert, P. J., 2001, The distribution and redistribution of income, 3rd edition, Manchester: Manchester University Press.

Medeiros, M., J. de Castro Galvao, and L. de Azevedo Nazareno, 2018, "Correcting the underestimation of top incomes: combining data from income tax reports and the Brazilian 2010 census", Social Indicators Research, 135, 233– 244.

Myck, M., and T. Najsztub, 2015, "Data and model cross-validation to improve accuracy of microsimulation results: Estimates for the polish household budget survey", International Journal of Microsimulation, 8, 33–66.

Nguyen, N. T. V., and I. Rubil, 2021, "Fiscal policies, inequality, and poverty in Croatia", EIZ Working Papers, EIZ-WP-2104.

Official Gazette (Narodne novine), no. NN 115/16, NN 106/18, NN 121/19, NN 32/20, NN 138/20.

Ooms, T. C., 2021, "Correcting the underestimation of capital incomes in inequality indicators: with an application to the UK, 1997–2016", Social Indicators Research, 157, 929–953.

Sutherland, H., and F. Figari, 2013, "EUROMOD: the European Union tax-benefit microsimulation model", International Journal of Microsimulation, 6, 4–26.

Taleb, N. N., and R. Douady, 2015, “On the super-additivity and estimation biases of quantile contributions”, Physica A: Statistical Mechanics and its Applications, 429, 252–260.

Törmälehto V.-M., M. Jäntti, and E. Marlier (eds.), 2017, The use of registers in the context of EU-SILC: challenges and opportunities, Luxembourg: Eurostat.

Urban, I., S. Bezeredi, and M. Pezer, 2020, “EUROMOD Country Report: Croatia (HR) 2017−2020”, Joint Research Centre, European Commission.

Yonzan, N., B. Milanović, S. Morelli, and J. Gornick, 2020, “Drawing a line: comparing the estimation of top incomes between tax data and household survey data”, Stone Center on Socio-Economic Inequality Working Paper Series, no. 27.

Appendices

Appendix 1. Additional tables

[Table A1 here]

[Table A2 here]

[Table A3 here]

Appendix 2. Procedure to approximate taxable and non-taxable parts of employment, pension, and capital incomes

The procedure for employment income and pension income is the following. The first two steps are common to both income sources. The last three steps are source-specific but analogous, and thus we present them for employment income.

Steps 1 and *2*: Same as in section 3.2.

Step 3: For each subgroup i in the tax data, we calculate the share of individuals with positive non-taxable employment receipts, V_i. Then we calculate the desired number of individuals with positive non-taxable employment receipts in the survey, B_i:

$$B_i = V_i \cdot N_i,$$

where N_i is the size of subgroup i in the survey data.

Step 4: In both the tax and survey data, create new variable, the share of non-taxable employment receipts of person j in his/her employment income. In the survey data, this variable is zero for everyone. Denote this variable by S_j^{tax} and S_j^{sur} in the tax and survey data, respectively.

Step 5: For each subgroup i in the tax data, we take the set of individuals with positive non-taxable employment receipts, and randomly draw B_i values of S_j^{tax}. Then we randomly draw B_i adult individuals from subgroup i in the survey and replace their $S_j^{\text{sur}} = 0$ with one of the positive values drawn from the tax data.

For capital income, the procedure is integral to the pre-correction procedure in section 3.2. Here we just modify slightly the last step.

Steps 1, *2*, and *3*: Same as in section 3.2.

Step 4. Same as in section 3.2, with the following addition: Together with the values of capital income, we also take the shares S_j^{tax} of the same individuals randomly drawn from the tax data and assign them to the same individuals randomly drawn from the survey data (i.e., replace their $S_j^{\text{sur}} = 0$ with $S_j^{\text{tax}} > 0$).

Tables and figures

Table 1. Aggregate amount and number of recipients of total gross fiscal income and its sources

	Tax data	Survey data			Survey data (tax data = 100)		
		ORIG	BFM	PC-BFM	ORIG	BFM	PC-BFM
Panel A. Aggregate amount (HRK billion)							
Fiscal income, total	191.5	171.9	192.8	194.4	90	101	102
Employment income	128.9	120.5	137.8	130.4	93	107	101
Self-employment income	6.7	8.1	10.3	7.0	122	155	105
Pension income	37.7	39.5	39.9	39.9	105	106	106
Contractual work income	3.6	1.4	2.1	3.6	40	59	100
Property income	4.9	2.1	2.4	5.0	43	50	102
Capital income	9.7	0.2	0.3	8.5	2	3	88
Panel B. Number of persons with positive amount (thousands)							
Fiscal income, total	3,236	2,657	2,662	2,946	82	82	91
Employment income	1,607	1,444	1,444	1,444	90	90	90
Self-employment income	119	107	113	107	89	94	89
Pension income	1,303	1,078	1,074	1,076	83	82	83
Contractual work income	257	75	79	250	29	31	97
Property income	312	133	145	298	43	46	95
Capital income	870	152	170	870	17	20	100

Notes. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). In panel B, the number of recipients of fiscal income is not the sum of the number of recipients of its sources, as one person can receive income from more than one source.

Table 2. Composition of gross fiscal income by source

	Tax data	Survey data		
		ORIG	BFM	PC-BFM
Panel A. Composition of gross fiscal income (% of gross fiscal income)				
Employment income	67.3	70.1	71.5	67.1
Self-employment income	3.5	4.7	5.3	3.6
Pension income	19.7	23.0	20.7	20.5
Contractual work income	1.9	0.8	1.1	1.9
Property income	2.5	1.2	1.3	2.6
Capital income	5.1	0.1	0.1	4.4
Panel B. Hellinger distance from composition in tax data				
HD(survey data, tax data)	$\stackrel{\text{def}}{=} 0$	0.147	0.144	0.013

Notes. In panel A, the figures across sources add up to 100% (with rounding errors). ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). HD – Hellinger distance; HD = 0 (HD = 1) indicates complete similarity (dissimilarity) between two compositions (see section 3.3).

Table 3. Inequality of fiscal income and concentration of its sources along its distribution

	Tax data	Survey data		
		ORIG	BFM	PC-BFM
Gini coefficient	0.625	0.607	0.637	0.631
Top 10% share (%)	42.1	37.2	41.9	41.6
Top 5% share (%)	28.6	23.2	28.4	28.2
Top 1% share (%)	12.3	7.7	12.2	12.1

Notes. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). Units of observation are individuals. All individuals are considered, including those with no income, like children.

Table 4. Aggregate amount of taxes, social insurance contributions, and social benefits

	Tax data	Survey data		Survey data (Tax data = 100)	
		ORIG	PC-BFM	ORIG	PC-BFM
Social insurance contributions (SIC), total	46.775	46.333	47.142	99	101
SIC, employers	20.552	19.941	20.548	97	100
SIC, employees	24.085	23.982	23.802	100	99
SIC, self-employed	1.588	1.830	1.863	115	117
SIC, contractual workers	0.474	0.252	0.633	53	133
SIC, pensioners	0.076	0.327	0.296	431	391
Personal income tax and surtax (PITS), total	11.965	8.912	12.202	74	102
PITS, yearly general schedule	10.215	8.566	10.494	84	103
PITS, property income	0.514	0.281	0.663	55	129
PITS, capital income	1.181	0.028	1.009	2	85
PITS, lump-sum schedule	0.054	0.037	0.036	67	66
Personal property taxes	0.228	0.200	0.208	88	91
Social benefits	9.121	5.573	5.244	61	57

Notes. Amounts are in HRK billion. ORIG – uncorrected survey data; PC-BFM – corrected survey data (see section 3.2). In the first column, the figures for personal property taxes and social benefits do not come from the tax data: the former comes from the Ministry of Finance, and the latter from the Croatian Bureau of Statistics.

Table 5. Concentration of fiscal instruments across distribution of fiscal income

	Tax data	Survey data	
		ORIG	PC-BFM
Social insurance contributions (SIC), total	0.708	0.700	0.708
SIC, employer	0.723	0.708	0.725
SIC, employee	0.711	0.699	0.708
SIC, self-employed	0.459	0.614	0.510
SIC, contractual work	0.731	0.655	0.742
SIC, pensioner	0.728	0.762	0.759
Personal income tax and surtax (PITS), total	0.874	0.865	0.881
PITS, yearly general schedule	0.891	0.879	0.903
PITS: property income	0.603	0.584	0.624
PITS: capital income	0.876	0.534	0.850
PITS: lump-sum schedule	0.257	0.047	0.027

Notes. ORIG – uncorrected survey data; PC-BFM – corrected survey data (see section 3.2).

Table 6. Distributive impact of fiscal system

	ORIG	PC-BFM	Ratio
Panel A: Inequality and concentration of gross and disposable incomes			
Gini coefficient of gross income, G_X	0.356	0.393	1.10
Gini coef. of disposable income, G_Y	0.292	0.333	1.14
Concentration coef. of disposable inc., D_Y^X	0.286	0.329	1.15
Panel B: Redistributive, vertical, and horizontal effects of net fiscal system			
Redistributive effect, R	0.065	0.060	0.92
R/G_X	0.182	0.151	0.83
Vertical effect, V	0.071	0.064	0.91
V/G_X	0.198	0.163	0.82
V as share of R (%)	109	107	
Horizontal effect, H	0.006	0.005	0.79
H/G_X	0.016	0.012	0.71
H as share of R (%)	9	7	
Panel C: Taxes			
Concentration coefficient, D_T^X	0.776	0.797	1.03
Kakwani progressivity index, K_T	0.419	0.405	0.97
Average tax rate, t_T	0.050	0.060	1.20
Contribution to vertical effect, v_T	0.025	0.029	1.15
v_T/G_X	0.071	0.074	1.05
v_T as share of V (%)	35	45	
Panel D: Social insurance contributions			
Concentration coefficient, D_S^X	0.443	0.450	1.02
Kakwani progressivity index, K_S	0.086	0.058	0.67
Average tax rate, t_S	0.153	0.135	0.88
Contribution to vertical effect, v_S	0.016	0.009	0.59
v_S/G_X	0.044	0.024	0.53
v_S as share of V (%)	23	14	
Panel E: Social benefits			
Concentration coefficient, D_B^X	–0.366	–0.363	0.99
Kakwani progressivity index, K_B	0.723	0.756	1.05
Average tax rate, t_B	0.034	0.028	0.83
Contribution to vertical effect, v_B	0.030	0.026	0.87
v_B/G_X	0.083	0.065	0.79
v_B as share of V (%)	42	41	

Notes. ORIG – uncorrected survey data; PC-BFM – corrected survey data (see section 3.2). Taxes refer to the personal income tax with surtax and personal property taxes. Social insurance contributions refer to the employee contributions (for pension insurance). The indicators are defined as: $R = G_X - G_Y = V - H$; $V = G_X - D_Y^X$; $V = v_T + v_S + v_B$; $H = G_Y - D_Y^X$; $v_z = (1 - t_T - t_S + t_B)^{-1} t_z K_z$ for $z = T, S, B$; $K_z = D_z^X - G_X$ for $z = T, S$; $K_B = G_X - D_B^X$; $t_z =$ (aggregate z)/(aggregate X) for $z = T, S, B$. For all concentration coefficients, the ranking variable is gross income.

Table 7. Budgetary impact of reform

	Survey data						Difference in reform impact b/w ORIG and PC-BFM data	
	ORIG			PC-BFM				
	Baseline	Reform	Reform impact	Baseline	Reform	Reform impact	Amount	Structure
	[1]	[2]	[3]=[2]−[1]	[4]	[5]	[6]=[5]−[4]	[7]=[6]−[3]	[8]
PITS, total	8.912	9.296	0.384	12.202	14.379	2.177	1.793	100
PITS, general schedule	8.566	8.641	0.075	10.494	10.999	0.505	0.43	24.0
PITS, property income	0.281	0.562	0.281	0.663	1.326	0.663	0.382	21.3
PITS, capital income	0.028	0.056	0.028	1.009	2.018	1.009	0.981	54.7
PITS, lump-sum schedule	0.037	0.037	0	0.036	0.036	0	0	0

Notes. The reform is described in section 4.3. ORIG – uncorrected survey data; PC-BFM – corrected survey data (see section 3.2). All figures are in HRK billion, except those in column 8. The reform does not affect the amounts of social insurance contributions and social benefits.

Table 8. Distributive impact of reform

	ORIG			PC-BFM		
	Baseline	Reform	Ratio	Baseline	Reform	Ratio
Panel A: Inequality and concentration of disposable income						
Gini coefficient, G_Y	0.292	0.291	0.997	0.333	0.328	0.984
Concentration coefficient, C_Y^X	0.286	0.285	0.997	0.329	0.323	0.984
Panel B: Redistributive effect of net fiscal system						
Redistributive effect, R	0.065	0.066	1.013	0.060	0.065	1.092
R/G_X	0.182	0.184	1.013	0.151	0.165	1.092
Vertical effect, V	0.071	0.071	1.011	0.064	0.070	1.084
V/G_X	0.198	0.200	1.011	0.163	0.177	1.084
Horizontal effect, H	0.006	0.006	0.994	0.005	0.005	0.986
H/G_X	0.016	0.016	0.994	0.012	0.012	0.986
Panel C: Vertical effect of taxes						
Concentration coefficient, C_T^X	0.776	0.769	0.991	0.797	0.789	0.990
Kakwani progressivity index, K_T	0.419	0.412	0.984	0.405	0.396	0.980
Average tax rate, t_T	0.050	0.052	1.042	0.060	0.071	1.177
Contribution to vertical effect, v_T	0.025	0.026	1.027	0.029	0.034	1.168
v_T/G_X	0.071	0.073	1.027	0.074	0.087	1.168

Notes. The reform is described in section 4.3. ORIG – uncorrected survey data. PC-BFM – corrected survey data (see section 3.2). For definitions of the indicators, see the notes to Table 6. For all concentration coefficients, the ranking variable is gross income. The baseline figures are reproduced from Table 6. The figures for social insurance contributions and social benefits do not change.

(Appendix) Table A1. Illustration of pre-correction: example with capital income

Subgroup (i)	Income interval (thousand HRK)	Share of persons with positive capital income		Mean capital income		Number of persons in subgroup in ORIG sample (N_i)	Number of individuals in ORIG with positive capital income			Persons with positive capital income in PC-ORIG	
		ORIG	Tax data (R_i)	ORIG	Tax data		Actual (A_i)	Desired (D_i)	Difference (Δ_i)	Share	Mean capital income (HRK)
0	0	0.009	0.196	952	3,171	7,238	77	1,417	1,340	0.184	2,661
A1	0–5	0.009	0.100	284	7,641	40	1	4	3	0.042	641
A2	5–10	0.000	0.092	0	13,026	81	0	7	7	0.091	1,950
A3	10–20	0.017	0.090	384	8,142	246	4	22	18	0.092	1,710
A4	20–30	0.021	0.097	7,131	4,167	285	6	28	22	0.107	6,955
A5	30–40	0.017	0.109	783	9,741	325	6	36	30	0.105	3,492
A6	40–50	0.021	0.121	640	5,249	731	11	89	78	0.126	9,195
A7	50–60	0.016	0.143	778	5,368	938	19	134	115	0.121	1,422
A8	60–70	0.027	0.174	580	17,436	766	23	133	110	0.162	17,216
A9	70–80	0.023	0.197	1,029	10,658	466	13	92	79	0.173	1,658
A10	80–90	0.049	0.224	990	11,467	495	20	111	91	0.210	13,545
A11	90–100	0.060	0.245	1,158	13,782	379	22	93	71	0.247	2,144
A12	100–120	0.055	0.287	1,288	17,537	844	51	242	191	0.277	16,812
A13	120–150	0.084	0.336	1,155	17,377	496	41	166	125	0.308	5,318
A14	150–180	0.150	0.397	958	23,335	218	30	87	57	0.418	10,231
A15	180–210	0.100	0.450	773	38,343	105	11	47	36	0.430	16,781
A16	210–250	0.139	0.489	4,727	38,415	87	12	43	31	0.473	5,400
A17	250–300	0.230	0.536	1,581	46,949	59	14	32	18	0.518	119,502
A18	300–400	0.359	0.553	1,490	95,373	30	8	17	9	0.601	6,865
A19	> 400	0.370	0.612	3,253	145,353	20	8	12	4	0.509	14,212

Notes. Table illustrates the pre-correction procedure presented in section 3.2. Only 20 groups out of 58 are shown. The example refers to capital income. For details, see section 3.2.

(Appendix) Table A2. Summary information on B-F-M correction of survey data

Survey data before applying B-F-M method	Survey data after applying B-F-M method	Sample size (households) before and after applying B-F-M method		Merging point (percentile of fiscal income distribution)		Population above merging point (% of total population)		Size of missing population above merging point (% of total population)		
		Before	After	Tax data	Survey data before applying B-F-M method	Tax data	Survey data before applying B-F-M method	Total	Inside support of fiscal income distribution	Outside support of fiscal income distribution
[1]	[2]	[3]	[4]	[5]	[6]	[7]=100–[5]	[8]=100–[6]	[9]=[7]–[8]	[10]=[9]–[11]	[11]=[9]–[10]
ORIG	BFM	8,383	28,206	94	95.4	6	4.6	1.4	1.36	0.04
Pre-corrected ORIG	PC-BFM	8,383	28,676	94	94.2	6	5.8	0.2	0.19	0.01

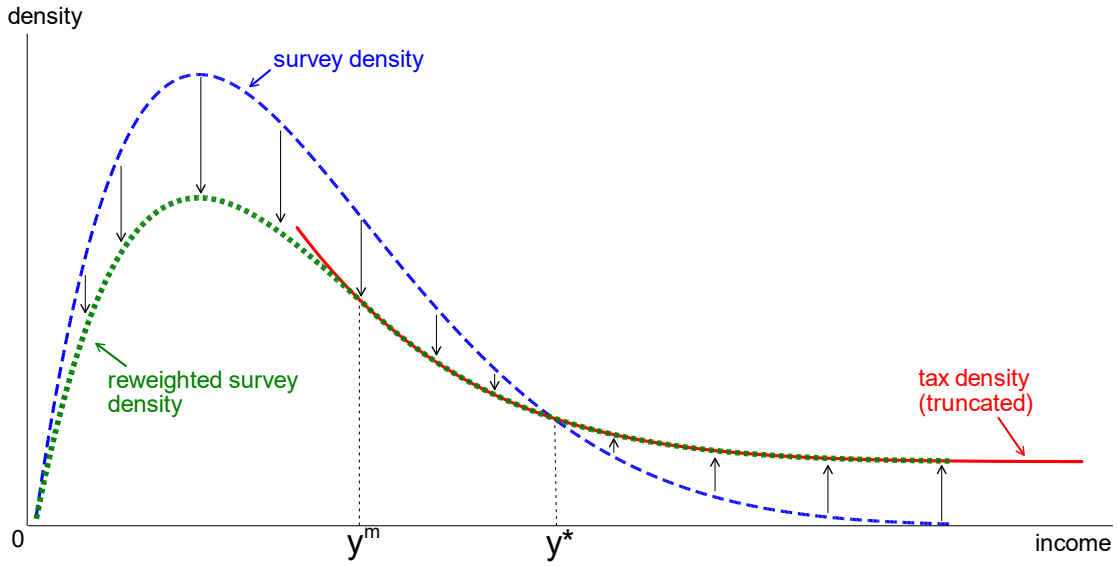
Notes. The table summarises information on two survey corrections using the Blanchet-Flores-Morgan (B-F-M) method, described in Section 3.1. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2).

(Appendix) Table A3. Distribution of socio-economic characteristics

	Entire sample			Bottom 94%			Top 6%		
	ORIG	BFM	PC-BFM	ORIG	BFM	PC-BFM	ORIG	BFM	PC-BFM
Panel A: Gender composition (%)									
Male	48.3	48.32	48.32	47.13	47.07	47.11	67.04	67.89	67.26
Female	51.68	51.68	51.68	52.87	52.93	52.89	32.96	32.11	32.74
Panel B: Activity composition (%)									
Employed	33.68	33.74	33.74	30.62	30.57	30.70	81.71	83.37	81.48
Self-employed	3.48	3.56	3.49	3.12	3.17	3.21	9.15	9.69	7.97
Pensioner	26.57	26.48	26.53	27.76	27.78	27.66	7.91	6.06	8.88
Unemployed	9.19	9.04	9.11	9.70	9.57	9.63	1.24	0.88	0.92
Inactive and other	27.07	27.18	27.12	28.80	28.91	28.80	0.00	0.00	0.76
Panel C: Educational composition (%)									
Low	3.82	3.79	3.81	4.05	4.01	4.01	0.35	0.39	0.73
Middle	59.21	58.10	58.52	60.58	59.65	59.90	37.81	33.89	36.97
High	14.75	15.77	15.41	11.74	12.58	12.42	61.84	65.72	62.19
Panel D: Age composition (%)									
0–19 years	19.64	19.67	19.65	20.89	20.92	20.90	0.00	0.00	0.11
20–29	11.93	11.91	11.91	12.24	12.20	12.17	7.06	7.38	7.71
30–39	13.57	13.55	13.58	12.81	12.81	12.81	25.44	25.13	25.66
40–49	13.34	13.34	13.34	12.33	12.27	12.39	29.11	30.19	28.26
50–59	14.49	14.47	14.47	13.90	13.89	13.97	23.61	23.47	22.21
60–69	13.58	13.61	13.60	13.75	13.77	13.70	10.97	11.15	11.93
70–79	8.58	8.57	8.58	8.95	8.98	8.94	2.81	2.17	2.91
80 or more	4.88	4.88	4.87	5.13	5.16	5.11	1.00	0.50	1.20

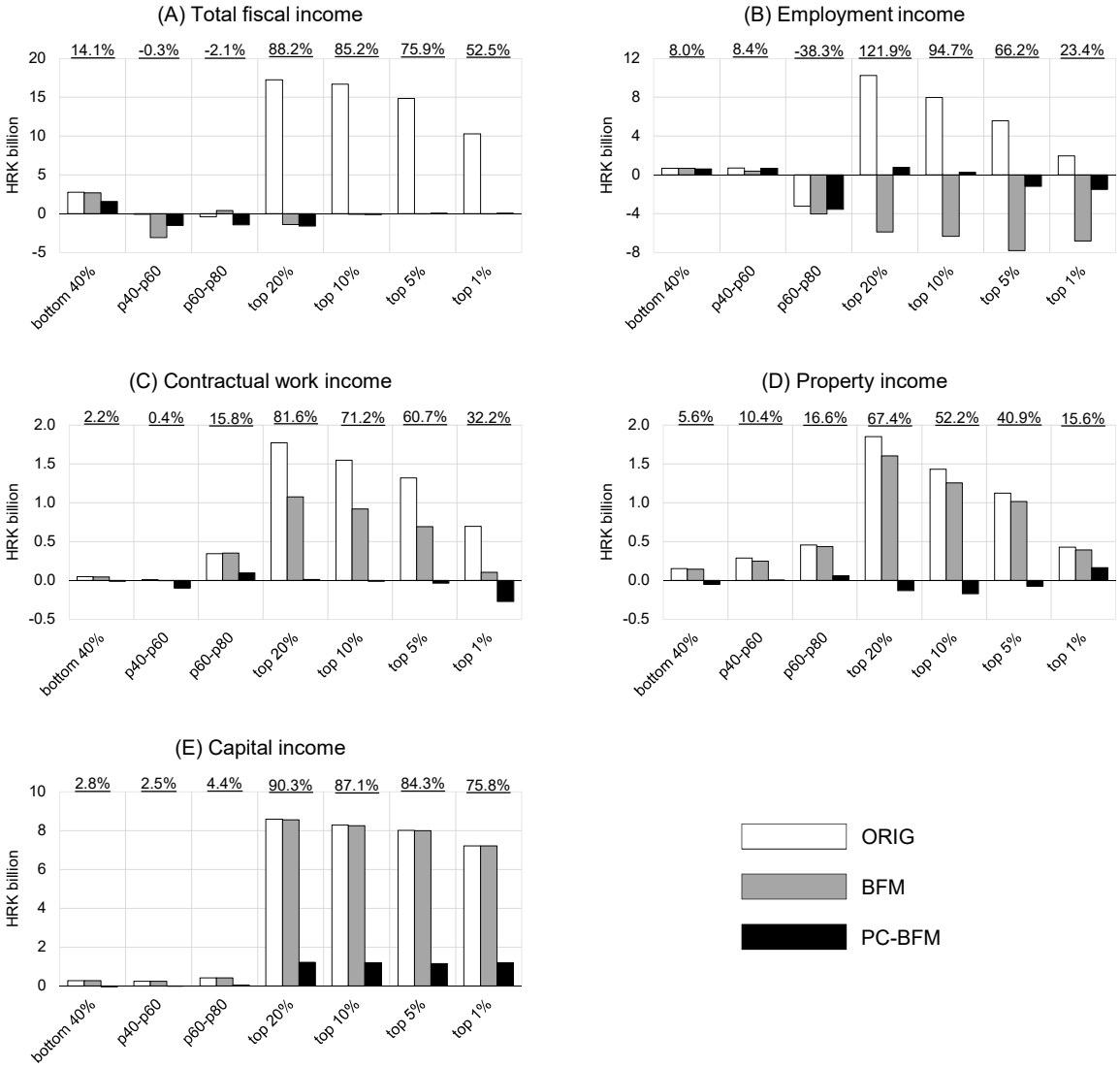
Notes. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). The division into the bottom 94% and the top 6% is based on the distribution of gross fiscal income. This division approximates the division into the parts above and below the merging point, as in the corrections resulting in the BFM and PC-BFM data the merging points are, respectively, the 95.4th percentile and 64.2nd percentiles of the distribution of fiscal income. Low education – primary or no education; middle education – secondary education; high education – tertiary education.

Figure 1. Illustration of survey reweighting as part of Blanchet-Flores-Morgan method



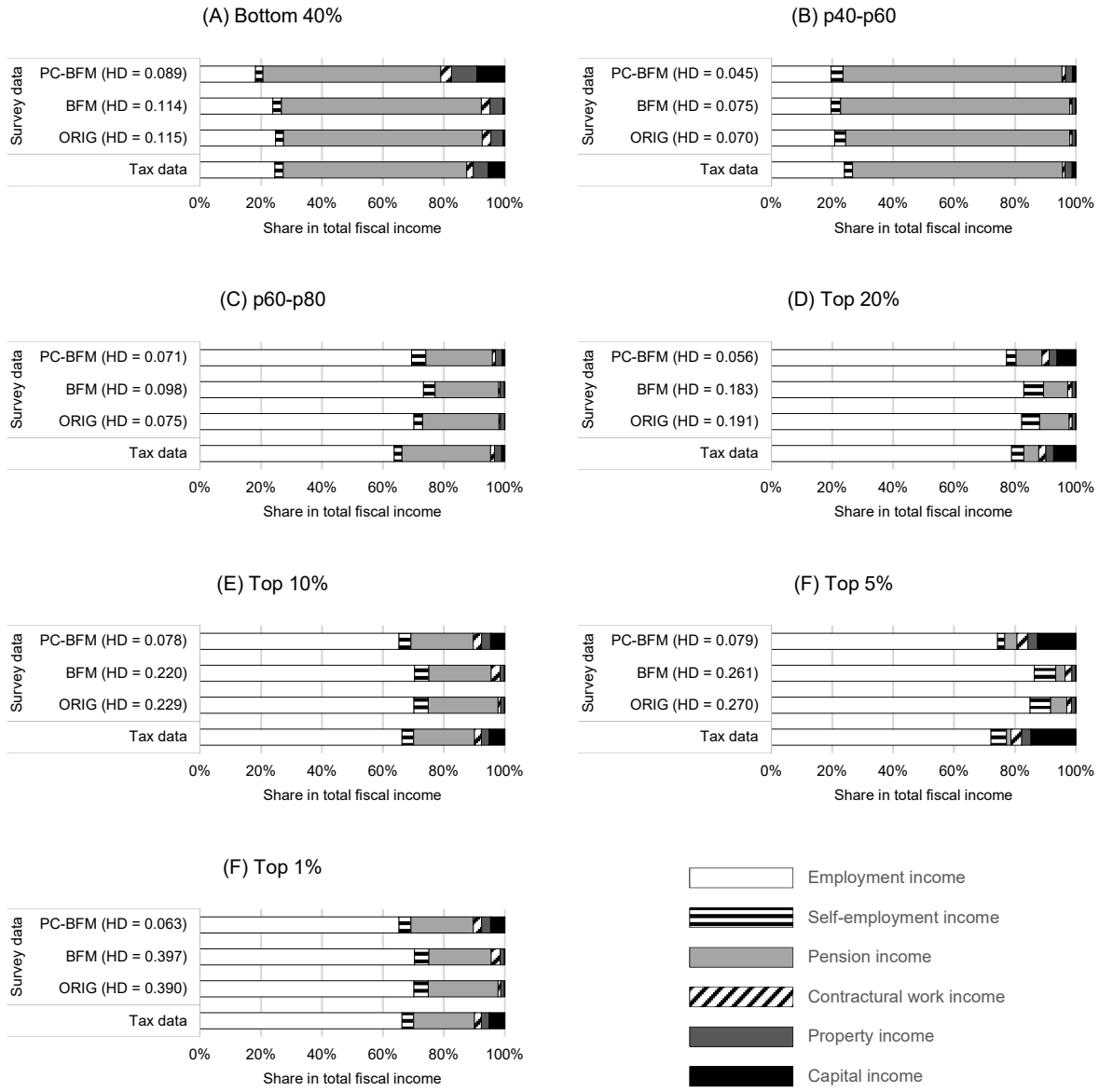
Notes. The figure illustrates how does reweighting work in the survey correction method of Blanchet, Flores, and Morgan (2022). The weights of those to the right (left) of y^* are adjusted upwards (downwards). The result is the transformation of the survey distribution (solid blue) into the reweighted survey distribution (dotted green), where the latter perfectly resembles the tax distribution (dashed red) to the right of the merging point y^m. The figure is an adapted version of Figure 2 in Blanchet, Flores, and Morgan (2022).

Figure 2. Distribution of missing income



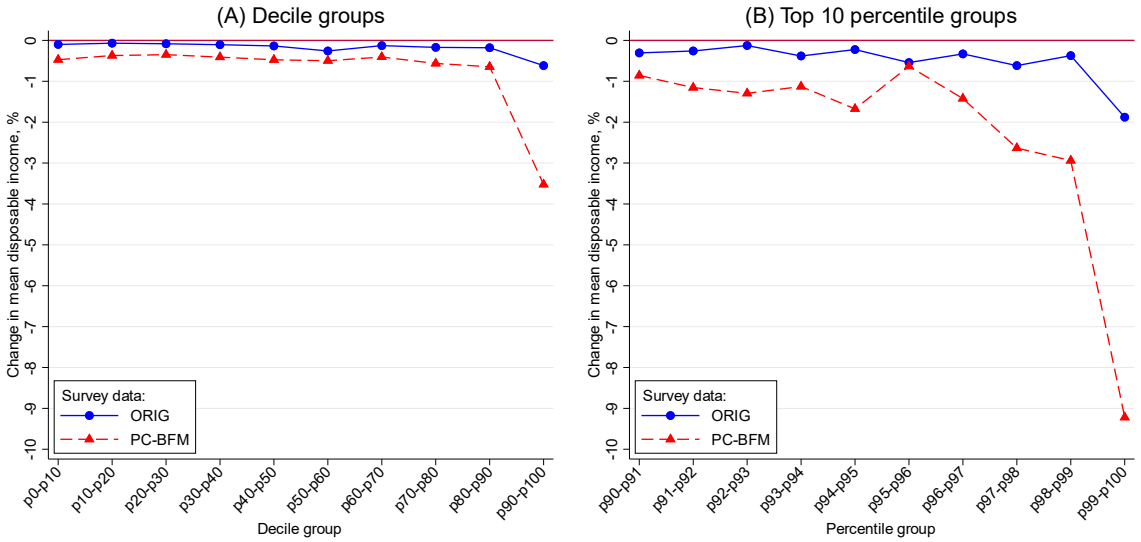
Notes. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). p40, p60, and p80 denote the 40th, 60th, and 80th percentiles of the distribution of total fiscal income. Each underlined percentage indicates what share total missing amount pertains to the corresponding quantile group. A negative percentage indicates that there is too much income in the survey compared to the tax data. In each panel, the percentages for the first four quantile groups from the left (these groups cover the whole distribution) add up to 100%, referring to the aggregate missing amount. For example, in panel E, the distribution of the aggregate missing capital income is this: 2.8%, 2.5%, 4.4%, and 90.3% of the aggregate pertains to the bottom 40%, p40–p60, p60–p80, and top 20%, respectively. The last three percentages are for smaller groups at the top. For example, in panel E, 75.8% of the aggregate missing capital income pertains to the top 1%.

Figure 3. Composition of fiscal income by source along distribution of fiscal income



Notes. ORIG – uncorrected survey data; BFM and PC-BFM – corrected survey data (see section 3.2). HD – Hellinger distance, a measure of similarity between two compositions, with HD = 0 (HD = 1) indicating complete similarity (dissimilarity). Each HD is the distance between the composition in the survey data (ORIG, BFM, or PC-BFM) and the composition in the tax data. p40, p60, and p80 are percentiles of the fiscal income distribution.

Figure 4. Change in disposable income upon reform



Notes. The reform is described in section 4.3. ORIG – uncorrected survey data; PC-BFM – corrected survey data (see section 3.2). p0, p10, …, p99, p100 are percentiles of the distribution of equivalised gross income.