STONE CENTER ON SOCIO-ECONOMIC INEQUALITY
WORKING PAPER SERIES

No. 49

The Inequality (or the Growth) We Measure: Data Gaps and the Distribution of Incomes

Facundo Alvaredo
Mauricio De Rosa
Ignacio Flores
Marc Morgan

March 2022

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

STONE CENTER
ON SOCIO-ECONOMIC INEQUALITY
HOME TO THE US OFFICE OF LIS

The Inequality (or the Growth) we Measure: Data Gaps and the Distribution of Incomes*

Facundo Alvaredo Mauricio De Rosa Ignacio Flores
Marc Morgan

March 21, 2022

Abstract

There is a large gap between income estimates used in inequality studies and macroeconomic statistics. This makes it hard to assess how economic growth is distributed across the population, and to what extent mainstream distributional statistics are an accurate representation of income flows. We take stock of these discrepancies by confronting estimates of the income distribution from surveys, administrative records and aggregates from the system of national accounts, thoroughly documenting them over the past two decades for ten Latin American countries. We find that surveys only account for around half of the macroeconomic income in the region. Measurement gaps account for just over half of the overall gap on average, while the rest is due to conceptual differences across data sets. Measurement gaps have been growing fast for many countries, the bulk being due to non-covered capital income. We also compare the top tails in administrative data and surveys, finding diverging averages –especially for non-wage incomes– and different shapes. We discuss the degree to which inequality levels and trends could be affected.

*We are grateful to the Statistics Division of the United Nation Economic Commission for Latin America and the Caribbean (ECLAC) for sharing their harmonized household surveys dataset. We also thank the statistics division of the Peruvian tax agency (SUNAT) for preparing income tax tabulations; and Catalina Galdamez Vanegas for kindly sharing with us income tax tabulations from El Salvador. We gratefully acknowledge financial support from the European Research Council (ERC Grant 856455), the French Research Agency (EUR Grant ANR-17-EURE-0001), and the Institute for New Economic Thinking. Alvaredo: PSE/EHESS/WIL, INET at Oxford, IIEP-UBA-CONICET (alvaredo@gmail.com). De Rosa: PSE, World Inequality Lab, Instituto de Economía - Universidad de la República (mauricio.derosa@fcea.edu.uy). Flores: City University of New York, PSE-World Inequality Lab (iflores@cg.cuny.edu). Morgan: University of Geneva, PSE-World Inequality Lab (marc.morgan@unige.ch).

1 Introduction

The development of economic statistics is a lengthy historical process that involves the views of the dominant doctrine, the construction of a body of conventions, and the limits of available data. The production of such statistics engage governments, central banks, official statistics offices, research institutions at different stages of the process. The macroeconomic aggregates from the System of National Accounts (SNA), such as Gross Domestic Product (GDP) and National Income (NI), are the most widely used measures of economic activity and are considered as benchmark numbers. In the early years of the SNA, national accountants were also experts in distributional issues, as the inter-linkages between the estimation of national income and its distribution were clearly recognized.[1] Over the last decades, this relationship was broken and the two fields went separate ways. The focus of the SNA has so far been on the aggregates of main institutional sectors in the economy, distinguishing the household sector, the corporate sector, the government sector and the foreign sector. At the same time, the applied analysis of the distribution of income has mostly relied on surveys and administrative data, and this has usually been approached quite independently from the SNA.

One of the observed results of such a disconnect has been the development of a large and sometimes increasing gap between aggregates from inequality studies based on microeconomic data, i.e. surveys and administrative records, and the SNA. The discrepancies can be seen in the levels of income, as well as in their growth rates (see, for example Ravallion (2003); Deaton (2005); Bourguignon (2015); Nolan et al. (2019)), and can attain particularly high levels in developing countries.[2] If it cannot be not surprising that national income is larger than the income concepts traditionally used to study inequality, it has also been growing faster. It has been argued that the discrepancies make it hard to assess how macroeconomic growth is distributed across income groups, and to what extent existing distributional statistics are a proper representation of the income flows in an economy. Recent work has embarked on a process of combining the various available data sources (surveys, administrative records, rich lists) aimed to produce comparable distributional results upscaled to SNA totals. These include, among others, World Inequality Lab (2020), Fixler et al. (2017) and a project coordinated by the OECD (Zwijnenburg, 2019). While the existing gaps have sometimes strengthened the feelings of uncertainty about inequality measurement, these new approaches have taken for granted the numbers provided by the national accounts, a practice that does not always contribute to diminish those feelings.

Yet the discrepancies between different sources of income statistics have long been recognized in different parts of the world. None more so than in Latin America. As an important

[1] See, among others, Kuznets et al. (1941); Kuznets (1953).

[2] Discrepancies can also be observed in wealth and consumption data, but these are beyond the scope of this paper.

precedent, it is worth citing CONADE (1965), which set out to estimate the distribution of income in Argentina in great detail for the years 1953, 1959 and 1961, making use of surveys, population and industrial censuses, income tax registries, and social security records, and attempting a reconciliation with the national accounts. A few decades later, in a seminal study, Altimir (1987) critically analyzed available tax, social security and census data, as well as a variety of household surveys, systematically comparing the latter with the SNA, and concluding that there was a 15-30% gap with aggregate household income, which could be significantly higher for income sources such as property income. These results were explicitly assumed to be an indicator of the underestimation of each type of income in the surveys, and thus Altimir applied adjustments with notable implications for inequality analysis (e.g. an increase of the Gini index of 10-15%). Altimir's approach was adopted by the United Nations-Economic Commission for Latin America and the Caribbean (ECLAC), but the adjustment had many caveats, and was recently discontinued. This experience clearly illustrates the need (as well as the demand) for a reconciliation between micro and macro datasets – or at least the need to fully understand its potential consequences – and of the significant challenges of such an endeavor.

Data availability is arguably one of the main restrictions to properly study the distributional aggregates that feed the research on income distribution. In Latin America, most of this research has used survey data to analyze the evolution of inequality, finding that it has experienced a downturn since the early 2000s, in the context of vigorous economic growth and redistributive public policies (López-Calva and Lustig, 2010; Cornia, 2014; Rodríguez-Castelán et al., 2016; Messina and Silva, 2017; Gasparini et al., 2018). However, question marks over the reliability of household surveys persist, as evidence on top incomes from tax records accumulates (Alvaredo, 2010; Alvaredo and Londoño-Vélez, 2013; Alvaredo et al., 2017; Burdín et al., 2022; Cano, 2015; Rossignolo et al., 2016; Morgan and Souza, 2019; Flores et al., 2020; Zuñiga-Cordero, 2018). More recently, whenever survey aggregates are compared to SNA aggregates, capital incomes appear to be remarkably less covered than labor incomes (Törmälehto, 2011; Bourguignon, 2015; Flores, 2021).

Underpinning the recent projects that seek to put together micro data sets with macro aggregates is the conviction that the statistical combination of data from several sources, based on researchers' own judgement and the resulting imputations, would allow for an acceptable correction and mitigation of the problems. Given the discrepancies at stake in Latin America, this undertaking becomes a sensitive issue, with potentially large revisions to currently accepted inequality trends. To illustrate this, Figure 1 shows recent estimates by De Rosa et al. (2020) of the top 10% income shares and the Gini coefficients of survey income (estimated from household surveys) and national income (estimated from combining surveys and tax data, and imputing the remaining gap in national income) for ten Latin American economies. The conclusion one can draw from this figure is that

micro-macro gaps seem to matter significantly for the inequality we measure in Latin America, notwithstanding the researchers' judgements about how to best impute the gaps to the distribution. The stylized conclusion is that the inequality we measure now seems to be higher and more heterogeneous than we previously thought. The point of this paper is to dig into the aggregate differences between survey incomes and the incomes in the SNA, as well as those from administrative sources.

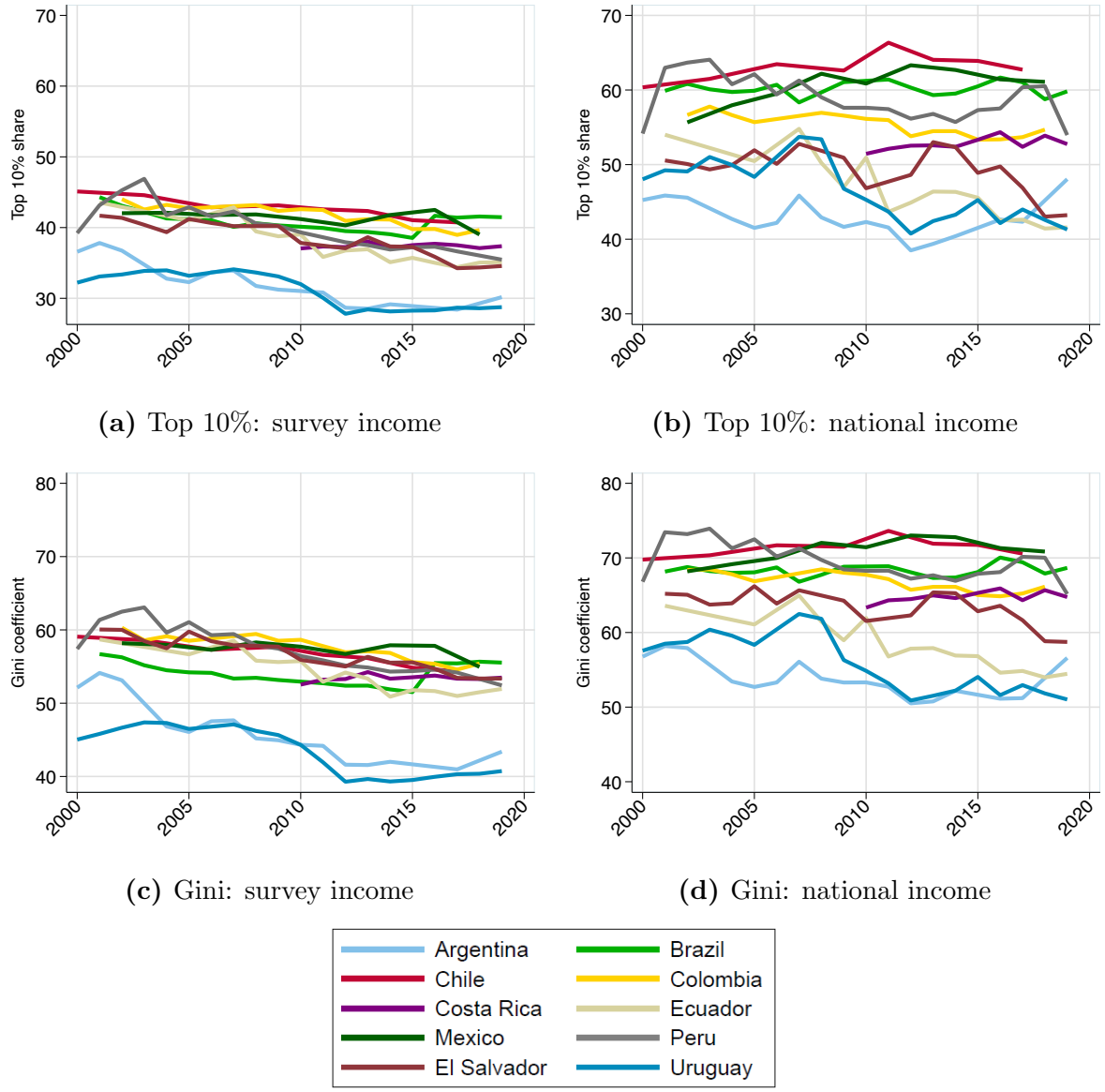
Designing the SNA meant accepting that the standard could not be set at the level of the best: it had to be feasible in less advanced systems, and this may have indeterminate consequences when used for distributional analysis. This is a central concern in this paper. The ultimate aim is to make the reader aware of the magnitude of the challenge, under current circumstances, when a reconciliation between micro-level and macro-level statistics on income is attempted in Latin America, and of the consequences of such an enterprise for inequality statistics. For this we need to take a step back, and provide a renewed view of the scenario before the combination of datasets are put forward.

To achieve this we first map the available data sources on income in the region. These include the SNA, household surveys, income tax data and social security records. We then perform a detailed accounting of the discrepancies between macroeconomic and microeconomic aggregates in terms of income coverage, population coverage and distributional statistics for most countries in the region. Finally, we compare the shape and average incomes of top tails in surveys and tax data. This provides a starting point to establish the suitability of approaches that combine these different data sources in order to re-examine inequality trends in the continent, such as those aforementioned projects.

We document for ten countries – covering 80% of the region's population – that official inequality estimates coming from household surveys only account for around half of national income.[3] Important differences in this ratio exist between some countries, ranging from 50-60% in Brazil, to 25-30% in Mexico. Of particular significance is the fact that in most countries, the total survey income is a declining share of national income over the course of the last two decades. Figure 2 provides a preview of these results. These are further commented in what follows, where we decompose the gap into two quantifiable components, which we call the "measurement gap" – the gap between measurable household income in surveys and equivalent household income in the SNA – and the "conceptual gap" – the portion of national income that is not directly received by households or measurable in survey questionnaires. We find that the measurement gap accounts for roughly 52% of the survey-SNA gap on average, with substantial variation between countries, and a general increasing tendency in recent years. We estimate that overwhelming majority of

[3]These statistics are "official", as opposed to "experiemental", in that they have been routinely published and cited by government departments, national statistics offices and supranational organizations for decades.

Figure 1: Top 10% share and Gini coefficient of survey income and national income



(a) Top 10%: survey income

(b) Top 10%: national income

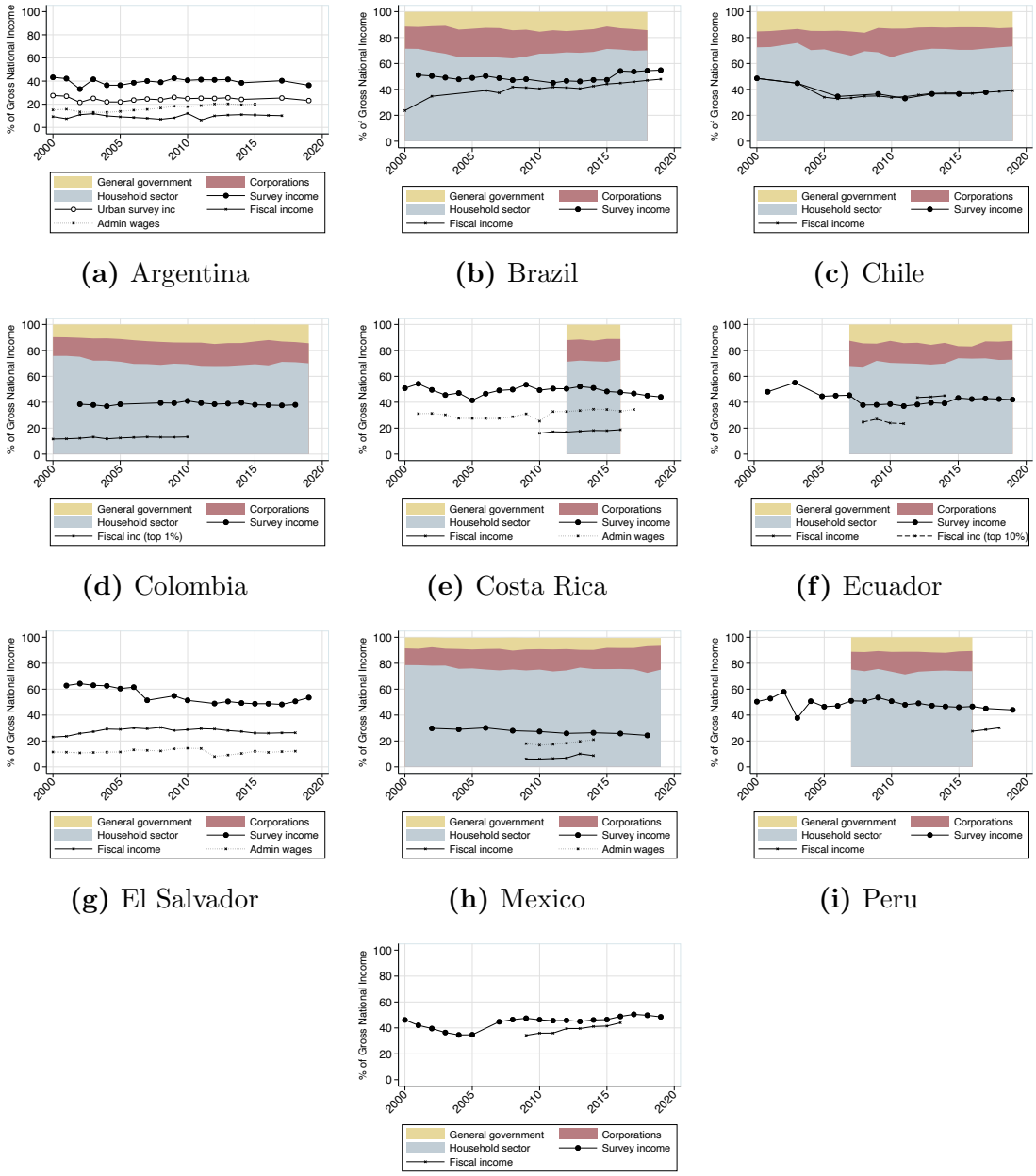
(c) Gini: survey income

(d) Gini: national income

Notes: Income corresponds to pre-tax income of individuals, with the income of couples being split equally. Panels (a) and (c) show the distribution of income as reported in household surveys, while panels (b) and (d) refer to the distribution of national income from the SNA using a combination of surveys, administrative data and aggregate macroeconomic income accounts. Source: De Rosa et al. (2020)

this gap is due to missing capital income received by households, consistent with other recent literature cited above. Moreover, assuming that administrative data better accounts for incomes at the right tail of the distribution, we find an increasing undercoverage of top incomes in surveys, especially when non-wage incomes are considered. Additionally, the top tail of the tax data distribution not only depicts higher concentration levels than surveys, but also a higher degree of top incomes inequality, meaning that the income of individuals at the top in surveys are progressively less covered as one moves up the distribution. These findings have notable implications for analyses of inequality levels and trends within and between countries, which should be given greater attention by the literature.

Figure 2: Comparing total income in national accounts, surveys and administrative data



(a) Argentina **(b)** Brazil **(c)** Chile

(d) Colombia **(e)** Costa Rica **(f)** Ecuador

(g) El Salvador **(h)** Mexico **(i)** Peru

(j) Uruguay

Notes. Own elaboration based on UN national accounts data, ECLAC harmonized surveys, and countries' administrative records; for Colombia, the top 1% is taken from Alvaredo and Londoño-Vélez (2013)), and for Ecuador the top 10% is taken from Cano (2015) and fiscal income comes from Rossignolo et al. (2016). Survey income and fiscal income represent total pretax income in both sources, while admin. wages represents total pretax wage income in administrative wage data. Shaded areas are the balance of primary incomes of the household sector (B.5g, S.14), corporations (B.5g, S.11 + S.12) and the general government (B.5g, S.13).

The remainder of the paper is structured as follows. Section 2 outlines the data we assess and the conceptual framework of each data source, covering the main income variables and the links between each source. We also address the extent to which the SNA is a benchmark for economic indicators. Section 3 presents our findings on aggregate data discrepancies, mapping the evolution of total income across sources, the evolution of gaps by income

component and the possible explanations for the observed gaps. Section 4 documents the top income deficit in surveys and the different shape of the top tails of the survey distribution and the tax data distribution. Section 5 concludes with a discussion of the implications of these discrepancies for inequality analysis, asking whether a reconciliation of these data sets is possible.

2 An inventory of data sets

We rely on four main sources to study aggregate income and its distribution: household surveys, income tax registers, social security records, and the national accounts. Yet there are still other sources that could and should be considered: population and economic censuses, banking information, firm-level data, etc; these are beyond the reach of this paper. Table 1 presents the availability of the microeconomic data sources for the countries in our study. The following subsections elaborate on both the microeconomic and macroeconomic databases we use, highlighting their strengths and weaknesses, as well as assessing their conceptual compatibility.

2.1 Micro-data: segments of a distribution

Microdata refers here to datasets for which information on income is collected at the individual level. Unlike macroeconomic data, which comprises aggregate income by institutional sectors in the economy, this approach allows for direct distributional analysis. Microdata includes both household surveys and administrative records (from income tax declarations and wage data from social security contributions). Historically, surveys have been the most widely used source to study the income distribution and its covariates. They mainly rely on randomized sampling and post-stratification techniques to represent the whole population. It is generally accepted that surveys are a reliable representation of a wide segment of the income distribution, but are a less reliable indicator of the tails of the distribution. On the other hand, administrative records do not generally need to rely on sampling because they cover the universe of tax payers and formal wage earners.[4] However, by definition, administrative records mainly focus on the formal sector and are also subject to misreporting. For these reasons, tax data has been typically used to better study the dynamics of top incomes, often allowing to extend the time coverage of estimates

[4]Some public data sets based on administrative records do not provide information of the universe of reference, but of a representative sample, and, in this sense, they also require a sampling strategy. Examples include the Survey of Personal Incomes Public Use Tape, in the UK, and the Longitudinal Sample of Registered Employment (*Muestra Longitudinal de Empleo Registrado*) in Argentina.

far beyond what surveys enable.[5] In section 4, we confront the distributions described by both administrative sources and survey data where they overlap.

2.1.1 Data from households surveys

We use the survey micro-data harmonized by the Statistics Division of the UN Economic Commission for Latin America and the Caribbean (ECLAC) for ten countries over the years 2000 to 2019. These countries include Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Mexico, Peru, and Uruguay. ECLAC's harmonization process builds on the original surveys produced by the official statistics institutes of the countries listed in Table 1. It seeks to create comparable annual income variables across countries, including the decomposition in terms of labor, capital and mixed incomes, pensions, owner-occupied rental income, transfers and other incomes.[6] In all cases but two, post-tax incomes are recorded on an individual basis, the exceptions being Brazil and Costa Rica, where gross (pre-tax) incomes are recorded.[7] Owner-occupied rental income and some capital incomes are collected at the household level, and distributed among the adults (aged 20 years and over) of the household.

The household surveys provided by ECLAC thus represent one of the key data inputs for this study. More broadly, national surveys are an extremely important reference point in their own right in Latin America, since they are the only source publicly available in almost all the countries. Official statistics on inequality, poverty, unemployment, etc., are drawn from them. The countries that remain excluded from this study are mostly from Central America and the Caribbean. They either do not report distributional data at all (Belize, Cuba, Haiti, Jamaica, Suriname and Trinidad and Tobago), do not run household surveys on a regular basis (Bahamas, Nicaragua, Venezuela), or only run surveys but do not have any kind of publicly accessible administrative data (Bolivia, Dominican Republic, Honduras, Panama and Paraguay).[8]

[5] Tax data estimates can be brought back to the early years of the 20$^{\text{th}}$ century, when comprehensive income tax systems were created, whereas regular household surveys date commonly from the 1970s onwards.

[6] The only exceptions concerning the frequency of the surveys are Chile and Mexico, which collect data every two to three years.

[7] Gross incomes in Brazil and Costa Rica are before personal income tax and employee social contributions.

[8] For more details on these countries see appendix table A.1.

Table 1: Countries and data sets available for comparison

	Survey microdata		Administrative data			
Country	Source	Availability	Source	Availability	Population (% of total)	Definitions
Argentina	Encuesta Permanente de Hogares (EPH) and EPH-Continua from 2003, Insituto Nacional de Estadística y Censos (INDEC)	2000-2014, 2016-2019	Income tax tabulations, Administración Federal de Ingresos Públicos (AFIP), Employee microdata, Ministerio de Trabajo, Empleo y Seguridad Social	2000-2017, 2000-2015	2% 40%	Survey is representative of urban areas (28-31 cities). Income tax data is total pre-tax fiscal income. Employee microdata includes only private sector wages.
Brazil	Pesquisa Nacional por Amostra de Domicílios (PNAD), Instituto Brasileiro de Geografia e Estatística (IBGE)	2001-2009, 2011-2019	Income tax tabulations, Receita Federal (RFB)	2000, 2002, 2006, 2007-2019	14%	Income tax data is total pre-tax fiscal income.
Chile	Encuesta de Caracterización Socioeconómica Nacional (CASEN), Ministerio de Desarrollo Social	2000-2009 (triannual), 2011-2017 (biannual)	Income tax tabulations, Servicio de Impuestos Internos (SII)	2000-2018	70%	Wages reported separately from other fiscal incomes in 2000-2004.
Colombia	Encuesta continua de hogares (Gran Encuesta Integrada de Hogares from 2008), Departamento Administrativo Nacional de Estadística (DANE)	2002-2005, 2008-2018	Alvaredo and Londoño-Vélez (2013)	2000-2010	1%	Income tax data is total pre-tax fiscal income.
Costa Rica	Encuesta Nacional de Hogares, Instituto Nacional de Estadística y Censos (INEC)	2000-2019	Wage income, Non-wage income Zuñiga-Cordero (2018)	2000-2017 2010-2016	28% 5%	Wage earners from social security records, Independent workers from income tax declarations.
Ecuador	Encuesta Periódica de Empleo y Desempleo (EPED) and Encuesta de Empleo, Desempleo y Subempleo (ENEMDU) from 2003, Insituto Nacional de Estadística y Censo (INEC)	2001, 2003 2005-2019	Cano (2015) Rossignolo et al. (2016)	2008-2011 2012-2014	14% 38%	Distributional data on total fiscal incomes is only available from Cano (2015) for the 10%.
El Salvador	Encuesta de Hogares de Propósitos Múltiples, Dirección General de Estadística y Censos (DIGESTYC)	2000-2007, 2009, 2010, 2012-2019	Tax tabulations (wages), Tax tabulations (diverse income) Dirección General de Impuestos Internos (DGII)	2000-2018	4% (wages) 4% (diverse)	Wages of salaried workers are reported separately from income from diverse sources.
Mexico	Encuesta Nacional de Ingresos y Gastos de los Hogares, Instituto Nacional de Estadística, Geografía e Informática (INEGI)	2002-2018 (biannual)	Income tax microdata, Servicio de Administración Tribuataria (SAT)	2009-2014	20% (wages) 2% (diverse)	Wages of salaried workers are reported separately from income from diverse sources.
Peru	Encuesta Nacional de Hogares - Condiciones de Vida y Pobreza, Instituto Nacional de Estadística e Informática (INEI)	2000-2019	Income tax tabulations, Superintendencia Nacional de Aduanas y de Administración Tributaria (SUNAT)	2016-2018	25%	Income tax data excludes entrepreneurial incomes.
Uruguay	Encuesta Continua de hogares (ECH), Instituto Nacional de Estadística (INE)	2000-2005, 2007-2019	Income tax microdata, Dirección General Impositiva	2009-2016	75%	Income tax data is total pre-tax fiscal income.

Note: Own elaboration. Population in administrative data is shown as a share of the total population of the country for the latest years.

2.1.2 Data from administrative records

Available distributional data from administrative sources in Latin America can be classified into four groups:

(i) microdata covering salaried employees with their wages declared at source by their employers, for social security records (e.g. Argentina, Costa Rica and Mexico);

(ii) microdata covering people with non-wage income sources (e.g. Costa Rica and Mexico);

(iii) grouped data (tabulations) based on the universe of tax payers, or those required to declare incomes, arranged by ranges of income (e.g. Argentina, Brazil, Chile, Peru, El Salvador and Uruguay);

(iv) grouped data (tabulations) based on the universe of formal wage earners, arranged by ranges of wages (e.g. Chile and El Salvador);

We exploit two new administrative sources for countries where, to our knowledge, tax data was never available for public-use purposes. One case is Peru, for which the tax authorities kindly prepared tabulated income statistics for this study. The data covers three years (2016-2018). It excludes entrepreneurial incomes, but includes pre-tax wages, dividends, rents, interests and other incomes. The other case is El Salvador, for which we gained access to two types of income tax tabulations, covering 2000-2018. One of the tables includes pre-tax wage income, while the other only includes individuals reporting income from diverse sources.

The rest of the countries in Table 1 can be divided in two groups. On the one side, those regularly publishing and updating their administrative records (Argentina, Brazil, Chile, Mexico and Uruguay). On the other side, those that gave access to microdata to other researchers at some point, but do not produce distributive information from tax registers on a regular basis (Colombia, and Ecuador). For these cases, we use estimates prepared by the authors of previous studies (Alvaredo and Londoño-Vélez, 2013; Cano, 2015; Rossignolo et al., 2016), which are restricted to the top fractiles of the distribution. For Costa Rica we avail of grouped data from Zuñiga-Cordero (2018), given the restricted access to administrative microdata on wage and independent income. Overall, as Table 1 reveals, there is a wide range in the proportion of the population covered in the available tax statistics in each country, with less than 5% in Colombia, El Salvador and Mexico (the latter for diverse income), to over 70% in Chile and Uruguay.

2.2 Macro-data: a reference for aggregates

Macroeconomic data refer here to aggregates that follow the UN System of National Accounts (SNA). These are generally used to monitor domestic and national economic activity and are centered around the concept of Gross Domestic Product – or Value Added – which can be defined in three ways, giving rise to three sets of tables in the SNA: the production approach, the expenditure approach, and the income approach. We focus on the latter, which distinguishes flows between five institutional sectors – the foreign sector, financial corporations, non-financial corporations, the government sector and the household sector. Noteworthy items, for our purposes, are the income of salaried workers (recorded as "compensation of employees") and capital incomes (recorded as "property incomes").

The information from the SNA was obtained by scrapping the UN Statistics Division database (`http://data.un.org`) and the websites of each country's national statistics office. Although the macro aggregates produced by national accountants are often considered among the most reliable and internationally comparable sources, detailed information on the income approach is scarce in the region. Even in countries that produce this kind of data regularly, statistics offices can update their estimates with two to five years of lag. The level of aggregation also varies across countries. For instance, despite the fact that United Nations (2008) recommends distinguishing the Operating Surplus of Households (the income produced by owner-occupied housing and rented dwellings) from Mixed Income (the income of the self-employed), three countries – Chile, Ecuador and Bolivia – report both in the same aggregate. Furthermore, we observe large disparities in the level of detail provided for other relevant variables, such as property income and the consumption of fixed capital (capital depreciation). A lower level of detail in the decomposition of aggregates hinders our capacity to accurately match and compare income concepts across data sets and countries. We are thus forced into a trade-off between the precision of our estimates at the individual country level and their comparability at the regional level.

2.3 Matching micro and macro concepts

There are multiple ways to match incomes across data sets, with options ranging from the most aggregated definition of income to the most dis-aggregated. In this paper, our specific choice depends on a trade-off between the level of detail of the income components and the conceptual consistency of the definitions. Our ability to properly compare incomes depends directly on whether national statistics offices provide sufficient detail in their accounts to disentangle income components.

Table 2 displays the matching we perform for the empirical estimates presented in the

next section. We match five types of income from our harmonized surveys to those in the SNA, in columns 1 and 2, respectively. Since concepts are generally wider in scope in the SNA, column 3 lists the associated income components in the survey, while column 4 lists non-matched or problematic items. In column 4, the items followed by an SNA code (e.g. D61 for social contributions) can be subtracted from the items in column 2 for a better matching (depending on the detail provided by national agencies), while those without an SNA code cannot be separated from the associated aggregates.

Table 2: Mapping households' income-concepts across data sets

Income in Survey [1]	**National Accounts (SNA08)** [2] = [3] + [4]	**Matching definitions in SNA and Survey** [3]	**Non-matching definitions in SNA and Survey** [4]
Salaried work	Compensation of employees (D1)	Wages, salaries (D11)	Social security contributions (D61)
Rental income	Operating surplus (B2)	Rent of owner occupiers	Rental income from dwellings
Investment income	Property income (D4)	Interests received (D41r), Dividends (D42)	Interests paid (D41u) Rent of natural resources (D45) Investment income of insurance policy holders (D441) Investment income of pension funds (D442) Investment income of investment funds (D443)
Non-salaried work	Mixed income (B3)	Self-employed income	Rent of non-dwelling buildings
Benefits	Social transfers (D62)	Pension benefits, Other cash benefits	Sick-leave Unemployment insurance

Note: Based on United Nations (2008) and OECD (2013). All incomes are gross of capital depreciation. SNA item codes are in brackets.

Different types of incomes have different degrees of conceptual overlapping. Labour income from salaried work, for instance, is among the least problematic. In general, one can easily subtract social security contributions from the compensation of employees in the SNA, so that only wages and salaries are compared with surveys reported net-of-contribution wages.[9] Social benefits are relatively straightforward too. Most countries do not distinguish them by type in their national accounts, so we achieve matching consistency by adding all the social transfers together in surveys (pensions and other cash benefits). Often, however, unemployment insurance may not be adequately captured in surveys.[10] Where it is reported, ECLAC's harmonized household surveys confound it with other incomes from employment, such as sick-leave and other wage-related incomes. However, unemployment insurance and sick-leave are included in social transfers in the SNA, and not in wages and salaries. This creates a minor conceptual inconsistency in the matching of aggregates from both sources.

A slightly more complicated case is the income from non-salaried work, which is included

[9]The only exceptions are: Argentina for the whole period, and Costa Rica before 2011. In the former, aggregate social security contributions are never reported, we thus compare survey wages directly to the compensation of employees. In the latter they are only available since 2012, we thus assume a constant ratio between contributions and to compensation of employees before that year.

[10]This is the case notably for Brazil before 2016. Unemployment benefits are thus imputed using information on periods of unemployment reported in the survey and statutory payment levels from the ministry of labour, again following Morgan and Souza (2019).

in the definition of mixed income in the SNA. The measurement of this aggregate is riddled with inconsistencies across countries, as well as being subject to the highest degree of mis-reporting out of all income items (ILO, 2019). A particular issue for us is that the SNA guidelines (United Nations, 2008) also include effective rents from non-dwelling buildings owned by households, as self-employed units, in the mixed income aggregate. The ECLAC's harmonized surveys report all rental income collectively with other capital incomes so the item is indistinguishable by construction. However, this mismatch is likely to be very small in practice, compared to other comparability issues with measuring the income of the self-employed across countries (e.g. inclusion of some of the self-employed in the corporate sector, inclusion of a part of employer income in compensation of employees, under-reporting of income, especially among informal self-employed workers, etc.).[11] A further complication is that three countries in the region (Chile, Ecuador and Bolivia) report the household sector's operating surplus together with the mixed income aggregate, which limits the analysis to a relatively less precise and more aggregate level.

Something similar occurs when comparing imputed rents to owner-occupiers from surveys to the operating surplus of households from the SNA, which also includes households' actual rental income from leased dwellings. In most cases, we are unable to disentangle what rents are imputed or realized in the SNA. However, thanks to more detailed SNA data from the expenditure approach in Brazil, we can estimate that imputed rents account for 93% of the aggregate on average between 2000 and 2015. In some countries, survey questionnaires do not even ask questions on imputed rents, and actual rental income is often reported together with other capital incomes. This creates a mismatch with how the SNA reports this item (in operating surplus rather than property income), yet Brazilian data suggests that the magnitude of the mismatch should be of second order. When rental income from owner-occupiers is not reported in surveys, ECLAC's harmonization process computes its value based on information from similar rented dwellings in the sample for each country. However, the absence of mortgage interest payments is problematic. In the SNA these are included in interests paid on the uses side of the accounts (D41u). A household with a mortgage equal to its imputed rent is not an owner-occupier but an "acquirer", which is functionally equivalent to a renter (Bourguignon, 2015).

The most complex conceptual match is that of investment income from surveys to property incomes from the SNA, which includes many items that are not considered in the survey at all (returns on investment and pension funds, and imputed investment income to insurance policyholders). These items are, in theory, well identified in the SNA, even if they correspond to imputed incomes. However, they usually correspond to a level of dis-aggregation absent in the accounts of most Latin American countries. For those

[11] At this stage we cannot properly verify the extent of all these comparability problems and their variation across countries. We thus leave this avenue open for future investigations.

countries where the detail exists at least for a few years (Brazil, Colombia, Chile, Costa Rica, Ecuador and Mexico) we compare capital incomes in a more conceptually consistent way. In appendix A.2, we show that the non-overlapping concepts are lower than 20% of the aggregate (10% on average), which also suggests a second order issue for those countries where we cannot disentangle these concepts properly. In the case of rent form natural resources, which are usually found among uses in the household sector of national accounts, surveys fail to report them all-together.

Another conceptual difference that affects all factor incomes (labour, capital and mixed incomes) is related to taxes. In the SNA, all of them are recorded as pre-tax, while in the survey the situation is less clear. Most incomes are generally assumed to be declared post-tax (except in Brazil and Costa Rica), especially in the case of incomes that pay the personal income tax at source, like formal wages. In order to solve this issue, in what follows, we use effective income tax rates from administrative data to add income taxes paid across the distribution in the survey (for more details see De Rosa et al. (2020)).

3 Contrasting aggregates

In this section, we quantify the discrepancy between surveys and national accounts. Coverage is highly heterogeneous across income sources. That is, the gap is not the same for all incomes, for some it is almost negligible, while for some others can be substantial. Since different incomes are distributed differently, the impact on measured inequality levels and trends will depend on both the volume of missing incomes and their distribution.

3.1 Micro-data vs Macro-data

From the micro-data perspective, there are a number of reasons why surveys (and administrative data) may underestimate the total income of the household sector. We review the main causes of this phenomenon.

Household survey samples are, in principle, randomly selected from a target population, which is usually meant to be all resident household units. Despite big efforts to enforce randomness, many sources of biases coexist – heterogeneous response rates, non-random misreporting, small samples – and result in both un-representative samples and biased estimates. In order to address these issues, a long tradition of post-sampling adjustments was developed by data producers. The most common techniques use external data, such as population censuses, to re-weight observations in such a way that minimizes the distance between original and adjusted weights, while improving the representativeness of a series

of characteristics. However, these are traditionally socio-demographic, such as age and gender, but not income.[12]

It is only recently – and mostly in developed countries – that survey designs started to address income representativeness to improve the coverage of the top tail. We highlight two main techniques. First, oversampling at the top, which basically consists in increasing the sampling size of the targeted group disproportionately, e.g.in places that are known to be wealthier and expected to have lower response rates. Second, the non-anonymous linkage of surveys and the administrative records of their respondents, especially regarding wages.[13] Unsurprisingly, the second approach has shown to be especially effective at enforcing the consistency of macro and micro estimates of income (Törmälehto, 2011; Flores, 2021). Yet none of the Latin American countries currently employ either of these techniques. We thus ask whether the lack of mechanisms to counter these biases is likely to affect measures of income inequality not only in levels, but also in trends.

Figure 2 provides a visual comparison of aggregates across three sources. It shows the decomposition of gross national income from the SNA into the household sector, the general government and the corporate sector. It also plots the total income reported in household surveys, and the total income reported in administrative data (both total fiscal income and wages in the formal sector when available), as a percentage of gross national income. Three countries, Argentina, Uruguay and El Salvador, do not report aggregates from the income approach in the SNA.[14] One result is clear nonetheless: the gap between raw surveys and national accounts is very large, with total survey income covering usually around 50% of national income. Mexico appears as an extreme case, where the gap reaches close to 80%. Comparing across micro-level sources, we find that administrative data cover similar levels of income (across a smaller population) than surveys (especially in Brazil, Chile, Ecuador and Uruguay). In some countries (such as Argentina, Costa Rica and Mexico), administrative wage data give higher values, as they cover the universe of formal employees, and not just a sub-sample required to file an income tax return. Of particular significance is the fact that in most countries, the total income in surveys is a declining share of income in the national accounts over the course of the last two decades.[15]

The crucial point is to know what part of this difference is relevant in the comparison of sources. We can thus go further and decompose the total survey income—national income

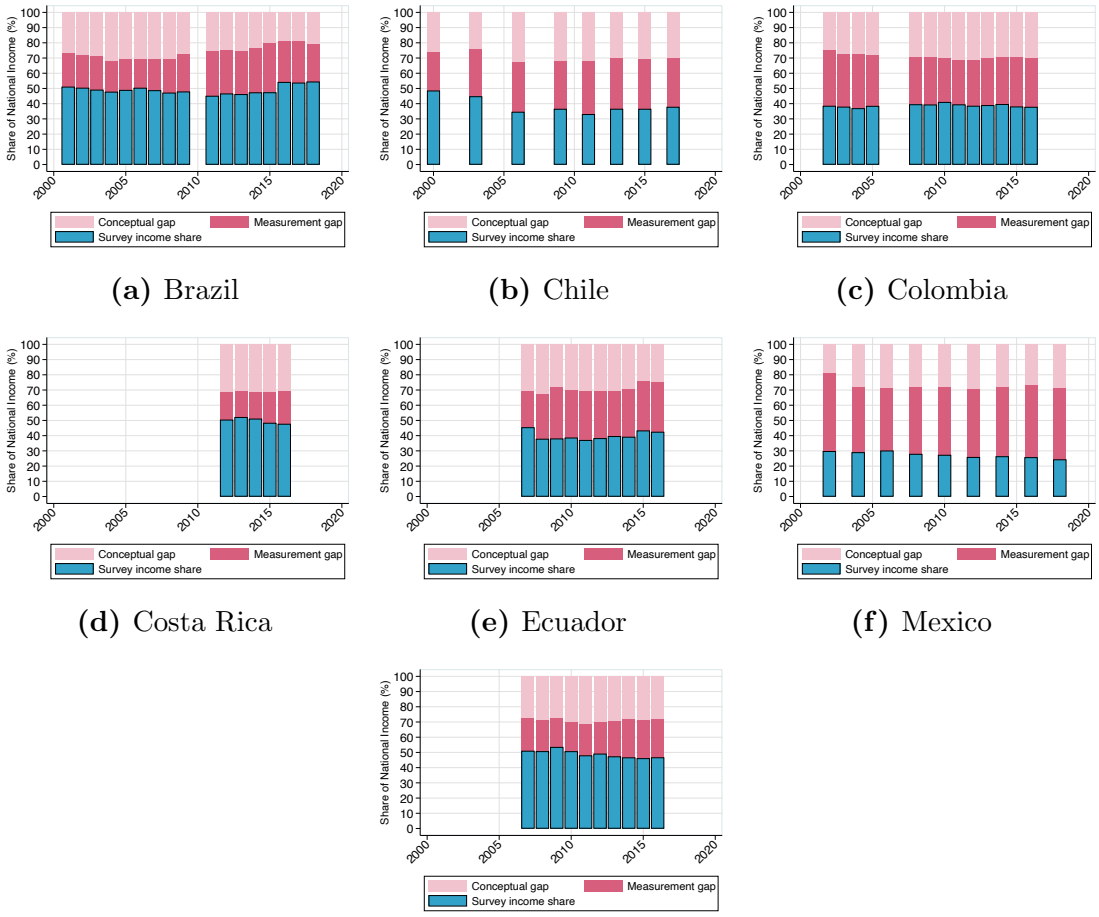
[12]Other relevant issues are top coding and censoring at the top of the distribution. However since the data we use in our empirical assessments are not subject to it, we ignore them in this work.

[13]Several countries participating in the European Union Statistics on Income and Living Conditions project (EU-SILC) have, over the last decade, progressively moved from the standard interview-based collection of information, to a mixed-strategy where incomes are directly obtained from fiscal registers for individuals in the sample. See Atkinson and Marlier (2010); Jantti et al. (2013), as well as the papers presented at the EUROSTAT Workshop on the Use of Registers in the Context of EU-SILC, Vienna, 2012.

[14]Uruguay recently reported aggregates from the income approach, but only for 2012 and 2016.

[15]Brazil since 2015 and Uruguay since 2006 are two exceptions, when both country surveys experienced substantial methodological changes in the sampling strategy, and thus the coverage of incomes.

Figure 3: Decomposing the Survey Income—National Income gap



(a) Brazil **(b)** Chile **(c)** Colombia

(d) Costa Rica **(e)** Ecuador **(f)** Mexico

(g) Peru

Notes. Own elaboration based on ECLAC harmonized surveys and the UN national accounts data for countries with sufficient breakdown in the SNA income approach to perform the calculation. Survey income is total pretax income. Measurement gap refers to the part of the gap explained by the under-coverage of household sector incomes for the matched income concepts. Conceptual gap refers to the part of national income not received/reported by households directly.

gap for each country i in each year t as follows:

$$Total\ gap_{it} = measurement\ gap_{it} + conceptual\ gap_{it}$$

where the measurement gap is the part of the gap associated to matched income items (see table 2); and the conceptual gap is the part related to non-matched household incomes (mostly, but not exclusively, those flowing to other institutional sectors). In practice, the latter is calculated as a residual. Figure 3 presents this decomposition for years where both surveys and national accounts overlap and for the seven countries that have sufficient detail in their SNA. Blue bars correspond to the share of national income that is covered by surveys, which was previously depicted in Figure 2.[16] Rose and pink bars divide the difference into the measurement gap and the conceptual gap, respectively. With Mexico having the largest gap to cover, it is not all surprising that its measurement gap is the

[16]This total is the sum of the incomes in figure A.3.

largest out of all the countries. This gap appears to be smallest in Costa Rica and Peru, where it accounts for less than half of the discrepancy. In the remaining countries it is roughly half of the gap. The measurement gap is thus a significant part of the discrepancy between surveys and national accounts, which raises question marks over the survey's capacity to accurately represent the distribution of income. In order to ascertain this we need to break down the discrepancy by income component and estimate the incidence of each component in the distribution of total income. This is what we turn to in the following sections.

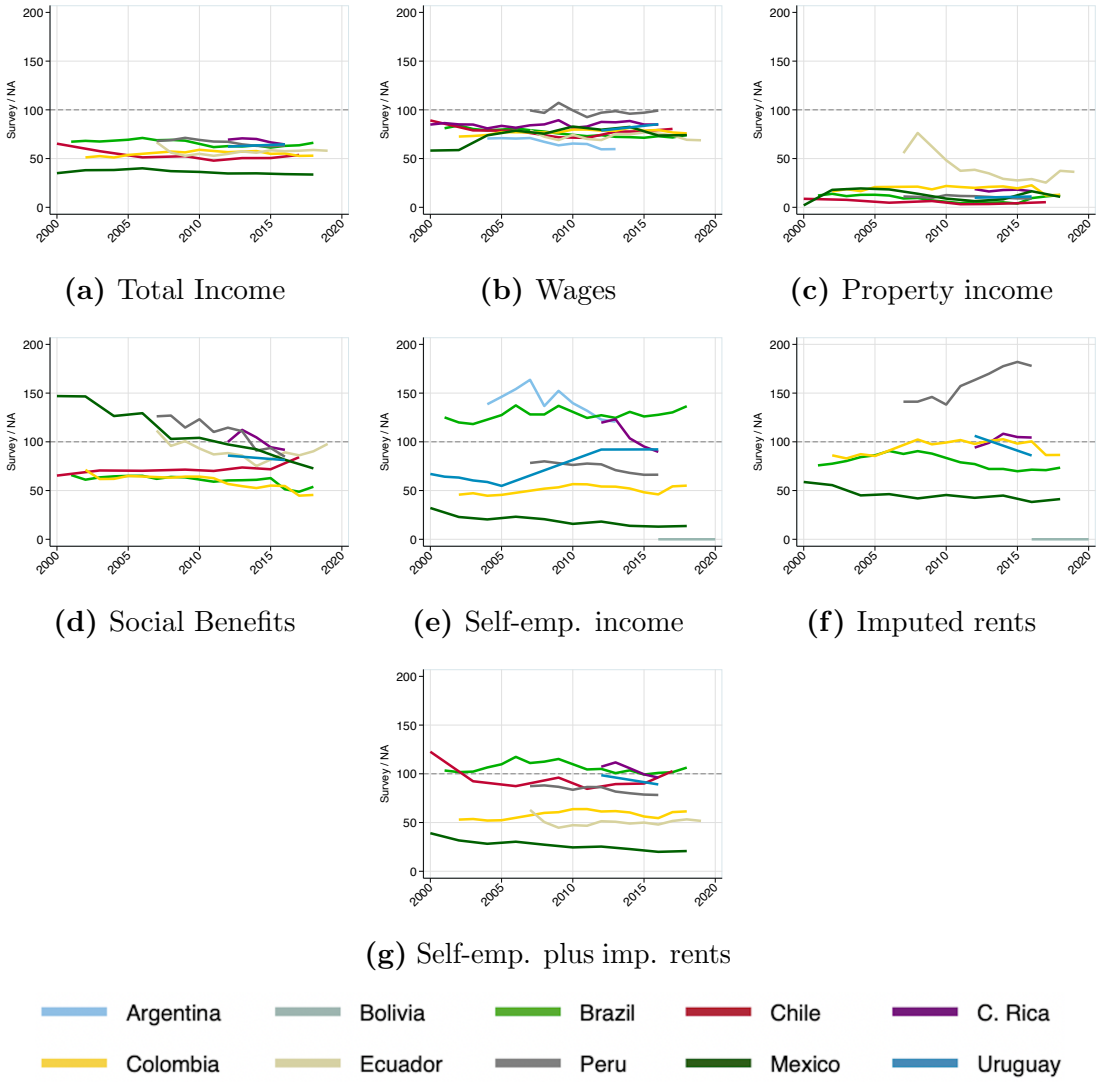
3.2 Heterogeneous coverage of income items

A simple, yet insightful exercise, is to compare the total amount reported, by type of income, in both household surveys and macro data. When definitions are comparable across datasets, the observed measurement gap can be interpreted as an underestimation of income, assuming that national accounts are considered an accurate representation -as discussed, this is a very debatable assumption that we use for the sake of the presentation of numbers.

Figure 4 displays coverage of income components in surveys with respect to corresponding items from the national accounts, based on the matching of concepts presented in Table 2. As previously commented, the sum of all matching incomes is clearly underestimated in all cases, with coverage rates ranging between less than half, for the case of Mexico, to close to 80% in the case of Brazil (subfigure 4a). From the rest of the subfigures, one can see the underestimation is not uniform across income items, and that some of them seem to contribute much more than others to the overall underestimation.

The coverage rates of wages and property incomes are polar opposites (subfigures 4b and 4c). The former are relatively high, with most countries bunching close to complete coverage, whereas property incomes are severely underestimated in all cases, with the exception of Ecuador, for early years. In the majority of cases, coverage is less than 10%. Other incomes, such as benefits, self-employment income and imputed rents – subfigures 4d, 4e and 4f – display relatively more heterogeneous coverage across countries. This includes ratios above one, which suggest that surveys may overestimate certain income components. This, in and of itself, is not wholly unexpected. If certain types of individuals/households with certain types of income are not covered by the survey (due to sparse samples not capturing rare populations and non-response, for example), then the income of certain other individuals/households in the covered sample may well be over-represented. This could affect the populations reporting social benefits or self-employed income. Moreover, the survey reports incomes of a specific reference period, usually a month, or a week that is aggregated to the month of reference, which may not carry over to the entire year.

Figure 4: Discrepancies by income component in surveys with respect to NA



(a) Total Income **(b)** Wages **(c)** Property income

(d) Social Benefits **(e)** Self-emp. income **(f)** Imputed rents

(g) Self-emp. plus imp. rents

Notes. Wage incomes are relatively well represented in surveys, while capital incomes are heavily underestimated. The coverage of other types of income is more heterogeneous, with both under- and over-estimation, depending on the case. Conceptual matching follows the benchmark in table 2. For a further decomposition of capital incomes, see appendix A.2. Chile and Ecuador report the corresponding aggregates of self-employment income and imputed rents together in the same item, they are not included here. Own elaboration based on ECLAC harmonized surveys and UN National Accounts.

Thus, when annualizing incomes – that is, multiplying declared monthly incomes by twelve – we may be attributing too much income to a certain class of activity whose realized annual income is much more volatile than an assumed persistent monthly earning (e.g. self-employed income).[17] If all types of income were to be then adjusted proportionally, the overall impact would depend on their magnitude and distribution. We turn to the former in the following section.

[17]The 'over-estimation' of imputed rents for some countries is more likely to be due to the methods employed by ECLAC (see section 2.3). We recommend that future revisions of this estimate be calibrated to the national accounts estimate of imputed rent where possible.

3.3 The size of 'missing' survey income

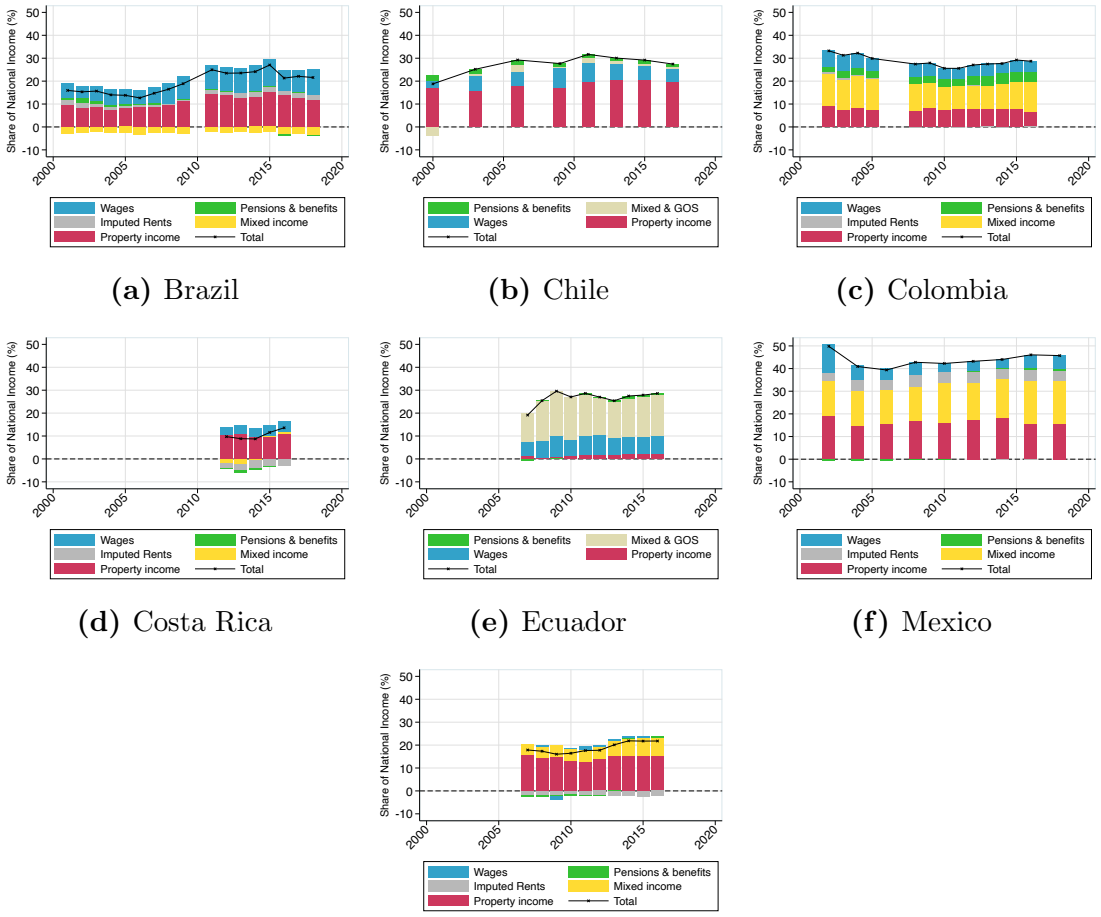
Independently on how to achieve the consistency of micro and macro data, the study of income gaps, and their composition, provides insights into how poor a guide official inequality estimates, and their trends, can potentially be.

Figure 5 summarizes these aggregates for the countries that have sufficiently detailed data. It displays the amount of each income item that survey's fail to capture. As we can see, the overall magnitudes are significant, ranging from 10% to 40% of national income across countries. Not surprisingly, the country with the highest discrepancies is Mexico, for which the magnitude of missing household sector incomes amounts to 40% of national income. Between 10% and 20% are due to the underestimation of property incomes. A similar amount is due to the underestimation of self-employment income, while the remaining magnitudes seem to be less significant. In all countries, although wages tend to be the least underestimated item (see figure 4), they show a relatively stable, if not increasing, tendency over time. In the cases of Brazil, Chile and Peru, the amount of missing survey income increases considerably during the period. In these cases, especially for Brazil and Chile, property incomes seem to play a major role in this evolution. On the contrary, in Colombia and Ecuador, it is the underestimation of mixed incomes and wages that seems to be driving the overall underestimation of incomes.

For almost all countries we also observe a minor portion of incomes that are over-represented in surveys. These are incomes whose share of national income are negative, and thus make the total missing income lower than what it would otherwise be. The overall pattern seems to suggest that, consistent with figure 4, self-employed income and imputed rents are the most susceptible to being over-represented in surveys, followed by pensions (see the previous section). While these household incomes are much less than the incomes under-represented, they should nevertheless be accounted for in any procedure seeking to make surveys macro-consistent.

To summarize, data gaps affect different income-types differently. Moreover, property incomes, which are always more concentrated at the top than other types of income (see appendix A.3), explain a large part of these gaps. We thus ask to what extent are the gaps presented thus far the result of an underrepresentation of the top tail in household surveys? This is what we turn to in the next section.

Figure 5: What's missing (or spare) in surveys



(a) Brazil **(b)** Chile **(c)** Colombia

(d) Costa Rica **(e)** Ecuador **(f)** Mexico

(g) Peru

Note. Own elaboration based on ECLAC harmonized surveys and UN-Data national accounts. GOS stands for gross operating surplus. In the national accounts of Chile and Ecuador mixed income and gross operating surplus are combined in a single aggregate.

4 Distributional implications

In the previous section, the micro-macro gap that emerged from contrasting income aggregates was analyzed both by decomposing it into a measurement gap and a conceptual gap, as well as in terms of income components and their magnitudes. This section studies and compares the distributions described by different micro-data sources where they overlap, which is in the top tail.

4.1 Top income levels in surveys vs tax data

As has been profusely documented in the literature (see Section 1), income differences are particularly important in the top tail of the distribution when survey and administrative data are compared; hence the importance of going beyond the analysis of aggregates.

Indeed, one of the key reasons given in the recent applied literature for combining survey and tax data is to properly account for incomes in the top tail of the distribution, which are assumed to be better captured by the latter. Thus, it is important to compare them in order to assess the likely effect of using administrative tax data to provide more accurate inequality estimates. It should be reminded, though, that the absence of the rich from surveys per se does not necessarily imply that measured inequality levels are biased downwards (see, for instance, Deaton, 2005).

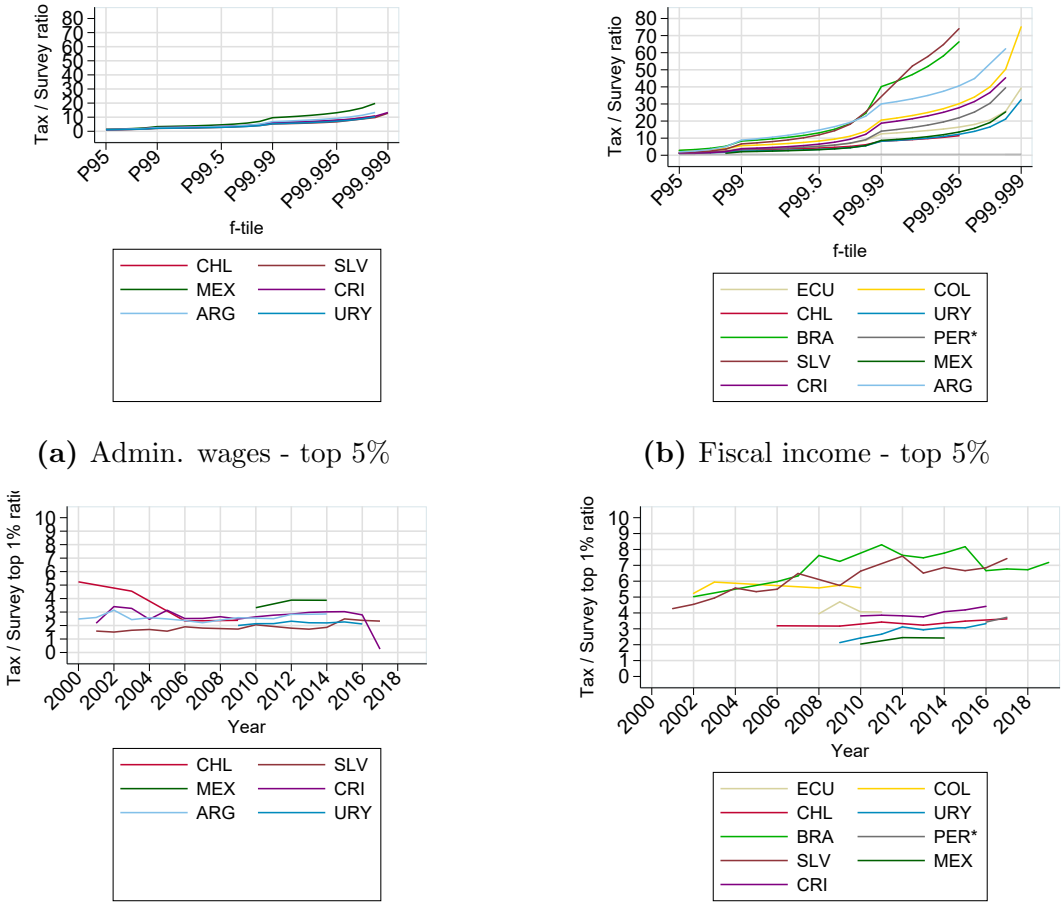
One straightforward way to proceed would be to systematically compare top income shares across data sources, but this requires adjustments to the tax data to account for incomes and population not captured in this sort of administrative data. At least, it would involve considering population and income controls (Atkinson et al., 2011), which are usually taken from censuses and national accounts and/or survey data, respectively.

Considering that the objective of this paper is to compare data sources without combining them in any way, and since estimating top shares involves precisely that kind of procedure, one possibility is simply to compare income levels in tax data and surveys for top income groups. This is depicted in Figure 6, distinguishing between countries for which tax data only accounts for wages (panel a), and the ones for which all incomes are considered (panel b; see Table 1).[18] Overall, incomes in tax records are higher for top fractiles. For all countries, pre-tax wages in the tax records surpass household pre-tax wages at some point within top 5% and are considerably larger within the top 1%. Ratios are substantially higher in panel b, which point to the fact that capital incomes are considerably less covered than wages in surveys, in line with previous research (e.g. Burdín et al. (2022); Morgan and Souza (2019).

Perhaps more importantly, tax and survey pre-tax incomes diverge for top fractiles over time. In panels c and d of Figure 6 the top 1% ratios are depicted, showing an increasing gap between both data sources, which is clearer when total fiscal incomes are considered. This may suggest that non-wage incomes (especially capital incomes) play an important role in this divergence. If in fact administrative data better capture incomes at the top, this pattern indicates that the poor performance of surveys in top fractiles is indeed worsening, which has significant implications for assessing inequality trends.

[18]For this analysis, tax data was interpolated based on a Generalized Pareto distribution to account for all income fractiles in the top 5%. These fractiles are defined relative to an external population control, which is taken from official population projections from country statistics offices.

Figure 6: Tax-survey pre-tax income ratio



(a) Admin. wages - top 5%

(b) Fiscal income - top 5%

(c) Admin. wages - top 1%, 2000-2019

(d) Fiscal income - top 1%, 2000-2019

Note. Own elaboration based on last available tax data point and ECLAC harmonized survey data. In panels a and b, the ratio of average incomes in tax/survey for each fractile is depicted among the top 5%. Ratios for the year 2010, or closest neighboring years are depicted. Panels c and d depict the pre-tax incomes ratio for the top 1%. In panel b, Brazil and El Salvador were caped at P99.995 as they increased exponentially beyond that point. Argentina is excluded from panel (d) due to its erratic pattern in fiscal income. (*) In the case of Peru, tax data on total income excludes entrepreneurial income.

4.2 The shape of the top tail in surveys vs tax data

To supplement the preceding analysis, we proceed by comparing the shape of the top tails directly (Atkinson, 2017; Cowell, 2011). In Figure 7 we plot – both in tax[19] and survey data – a function of income against a transformation of the survival function S, defined as $1 - F$, i.e. the complement of the cumulative distribution. The y axis depicts the log of income as proportion of mean income, while the x axis depicts the log of $1/s$.

By construction, if the fitted functions are linear, the steeper the line is the more concentrated is the income (note that the slope of the curves is equivalent to the inverse of the Pareto coefficient). More importantly, when the data points are in a straight line, the distribution is a Pareto-I, while a concave one is the result of what Atkinson (2017) calls a

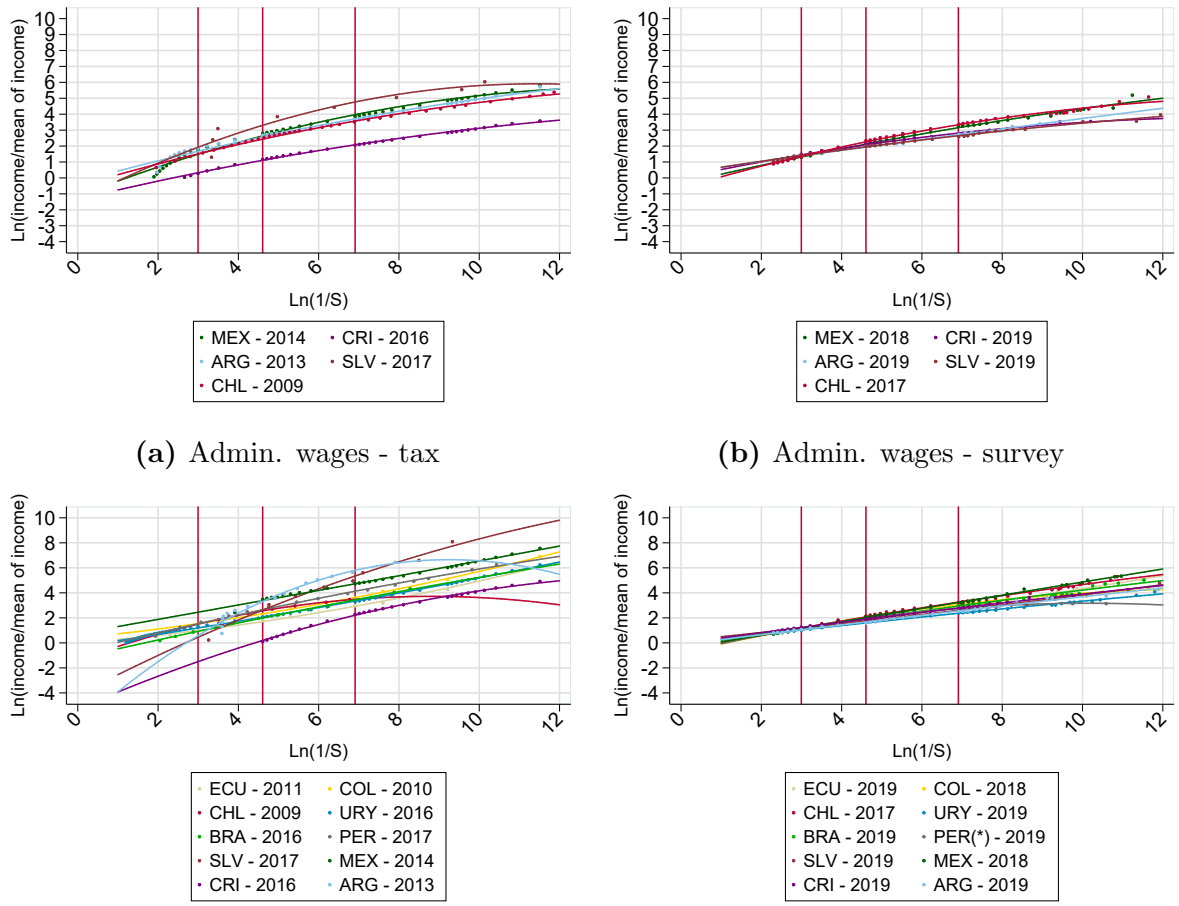
[19] Figure A.4 depicts top tails for first and last available years in the tax data by country.

baronial top tail, and a convex one is *regal*. The latter two shapes are departures from the familiar Pareto distribution, the first representing a distribution in which top positions tend to be more homogeneous, while the opposite holds in the *regal* shape, i.e. individuals in top positions tend to be further apart from each other. The main advantage of this approach is that it provides a way to directly compare top tails without the need to assume any sort of income control. This allows us to visually inspect not only concentration, but also the tail's shape. When only tabulations are available, tax bracket thresholds are plotted (without interpolation), while in the cases with micro data (both for surveys and for a some countries' tax data) selected data points are depicted. In all cases, a second degree polynomial function was adjusted to more clearly visualize the tails.

Three key features stand out. First, as expected, there is more concentration in tax data than in surveys, which are substantially closer to the x axis. Second, there is more heterogeneity in tax data, which may be the result of different shapes of the top tails across countries, or of the different quality and structure of the data. Third, while in survey data the top tails are mostly *baronial*, this is not the case in the tax data, in which several countries present a clear *regal* shape, given by the convexity of the survival functions.

To see it more clearly, the quadratic coefficients are presented in Figure 8. Despite being quite noisy, several facts emerge. The tax data of Ecuador, Uruguay, Mexico and Colombia present *regal* top tails, while the remaining countries have *baronial*-shaped tails. In contrast, barring a few exceptions, survey data presents *baronial* top tails, with incomes more similar to each other at the top. Overall, these sets of results illustrate incomes are substantially higher at the top of the distribution in administrative data than in surveys, driven especially by non-wage income; and that income dispersion within the top tail is generally greater in administrative data.

Figure 7: Pre-tax income in relation to rank



(a) Admin. wages - tax

(b) Admin. wages - survey

(c) Fiscal income - tax

(d) Fiscal income - survey

Note. Own elaboration based on country tax data and ECLAC harmonized survey data. For survey data, incomes are equal-split (broad) pre-tax income. Last available year for each country. The y axis depicts the log of income as proportion of mean income, while the x axis depicts the log of $1/S$, with S being the survival function. Vertical lines represent the thresholds for top 5 %, top 1% and top 0.1% incomes, respectively. (*) In the case of Peru, business incomes were not considered as they are not accounted for in tax data.

Figure 8: Fitted survival function's quadratic coefficients

(a) Admin wages - tax

(b) Admin wages - survey

(c) Fiscal income - tax

(d) Fiscal income - survey

Note. Own elaboration based on country tax and ECLAC's harmonized survey data pre-tax incomes. β coefficient of the regression $\log(income/meanincome) = \beta \log(1/s)^2 + \alpha \log(1/s) + \epsilon$. (*) In the case of Peru, business incomes were not considered as they are not accounted for in tax data.

5 Final remarks: implications for inequality research

What lessons can we draw from the preceding analysis for applied inequality research? We noted a large distance between the aggregates used in inequality studies and those from the SNA in Latin American countries. We also noted a growing undercoverage of household incomes in surveys vis-à-vis the SNA, especially property incomes. We documented that this overall undercoverage of magnitudes is due to a mixture of measurement and conceptual gaps. The former relates to item and unit misreporting, sparseness of the survey sample, especially in capturing 'rare' populations such as the rich, as well as heterogeneous non-response rates of individuals/households. The latter concerns unaligned income definitions between household surveys and the SNA, i.e., incomes in the SNA that by definition are not covered in surveys. For Latin American countries with detailed enough data we showed that at least half the survey–SNA income gap is due to measurement, and a significant portion of this gap is indeed due to missing capital incomes of households, but not only. Underrepresented items also affect wages, self-employed income, imputed rents, and pensions, to differing degrees across countries. The sheer magnitude of these gaps should not leave anyone indifferent.

To what extent then is income inequality being underestimated in household surveys? To approximate an answer to this question we require to know both the volume and the incidence of misreported income items. Although there is more uncertainty in the latter, we know that an important characteristic of survey measurement error is its heterogeneity across the distribution. Discrepancies thus have distributional implications, almost by construction. We showed, for the Latin American case, how these different income items are distributed – as reported in surveys – among the populations of the different countries in comparison to the distribution of total income. The general pattern that emerged is that those income items showing the largest magnitude of discrepancy were those whose distribution were the most unequal, namely capital incomes. This seems to be behind the increasing gap between top incomes in surveys as compared to administrative data that we document. Other income items showed less concentration, but were relatively less underestimated vis-à-vis the SNA. Some items for some countries seem to be overrepresented in surveys relative to the SNA, such as imputed rent, which is generally the least unequally distributed item across our set of countries. Based on these discrepancies, it can thus be anticipated that income inequality is being underestimated in household surveys in Latin America, even before making any survey adjustments. Yet, even if this is the case, the following questions should address to establish whether producing accurate trends is more important than levels.

It must be said that this positioning of surveys in relation to the SNA assumes that the latter are an accurate benchmark for incomes flowing to households. From a conceptual

viewpoint, and on the basis of international standards, the accuracy of the SNA should be relatively high. However, we are aware that in the case of Latin American countries particularly, but not exclusively, the construction of the national accounts must be a cause of concern. Indeed, the SNA remains somewhat of a black box, in the sense that, faced with a lack of information, accountants need to make judgements and assumptions for the calculation of sectoral incomes (i.e. wages, mixed income, interests, dividends, operating surplus...), even where the 'income approach' of the SNA is available, which is not always the case for the countries we study. Some countries have responded to this lack of transparency by publishing detailed methodological documents on the construction of the SNA across sectors, notably in Brazil (IBGE, 2016). Necessarily the NA levels of income, the resulting growth rates, and the capital shares in the developing Latin American countries require a thorough re-examination.

Despite these shortcomings, for roughly thirty years, the UN-ECLAC re-scaled reported incomes in surveys to comparable incomes registered in the SNA, following in the footsteps of Altimir (1975, 1987). This adjustment was done proportionally with the exception of capital incomes, which where imputed only to the top quintile. Yet, it was precisely because of the diverging trends between surveys and the SNA over time and the increasingly costly procedure to reconcile both sources that this adjustment was abandoned by ECLAC. Re-evaluating such an adjustment in light of our findings presents additional challenges. On the one hand, a proportional adjustment is certainly far from perfect, since the true distribution of income components could be very different from the one described by surveys. A better alternative, for example, would require high-quality administrative data, with income decomposed by item, whose availability to researchers is quite rare. From our comparison of surveys and administrative data available in the region, we showed that surveys largely fail to reproduce the shape of the right tail of the distribution in tax data, in some cases more significantly than others, without resorting to incomes in the SNA. The comparison of wages and total incomes in tax and survey data's right tails suggests that part of the decomposition of income in the right tail is also mis-measured in surveys, in line with previous research. To a certain degree, administrative data of the sort available for Latin American countries could be used to implement a finer survey adjustment, especially at the very top of the distribution.

The application of adjustment methods were not the scope of this paper, which concentrated on the necessary step of critically comparing incomes across different data sources to attest whether a reconciliation of such sources is indeed possible. On the basis of our findings we judge that such a reconciliation can be made (this is obvious from the theoretical point of view) but substantial investment is still needed. Indeed a specific contribution of this paper is to also draw attention to the greater needed investment in the level of detail and transparency of the national accounts by national producers.

Attempting to fill the income gap between official household surveys and official macro statistics may seem like too much of a stretch in the current state of affairs. Yet, it should be pointed out that no data source is perfect, and that in any empirical distributional analysis assumptions must be made, whether they are implicit (e.g. taking surveys to represent the national distribution of household income) or explicit (e.g. imputing the micro-macro gaps according certain allocation rules).

In the meantime, an alternative to this view is the vector approach to inequality statistics, whereby a series of internally consistent indicators are compared to one another to determine the plausible direction of inequality over a certain period of time, that is, the study of the evidence provided by different and competing data sources. This approach places less emphasis on inequality levels and more emphasis on trends, which in itself is cost efficient, as less work is needed to reconcile difference data sources. On the other hand, this approach may suffer from contradictory evidence among the difference series, leading to inconclusive evidence on inequality trends. From this perspective, reconciliation of different income sources is merited, especially given that the official macroeconomic accounts of a country are themselves the reconciliation of different sources, among them some of the routinely used microdata sources in inequality studies.

References

Altimir, O. (1975). Estimaciones de la distribución del ingreso en América Latina por medio de encuestas de hogares y censos de población: una evaluación de confiabilidad. CEPAL.

Altimir, O. (1987). Income distribution statistics in Latin America and their reliability. *Review of Income and Wealth 33*(2), 111–155.

Alvaredo, F. (2010). The rich in Argentina over the Twentieth Century, 1932-2004. In A. B. Atkinson and T. Piketty (Eds.), *Top incomes: A global perspective*, pp. 253–298. Oxford University Press.

Alvaredo, F., S. Garriga, and F. Pinto (2017). Household surveys, administrative records, and national accounts in mexico 2009–2014. is a reconciliation possible? Paris School of Economics.

Alvaredo, F. and J. Londoño-Vélez (2013). High Incomes and Personal Taxation in a Developping Economy: Colombia (1993-2010). CEQ Working Paper No. 12.

Atkinson, A. B. (2017). Pareto and the upper tail of the income distribution in the uk: 1799 to the present. *Economica 84*(334), 129–156.

Atkinson, A. B. and E. Marlier (2010). *Income and living conditions in Europe*. Luxembourg: Publications Office of the European Union.

Atkinson, A. B., T. Piketty, and E. Saez (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature 49*(1), 3–71.

Bourguignon, F. (2015, 12). Appraising income inequality databases in Latin America. *The Journal of Economic Inequality 13*(4), 557–578.

Burdín, G., M. De Rosa, A. Vigorito, and J. Vilá (2022). Falling inequality and the growing capital income share: Reconciling divergent trends in survey and tax data. *World Development 152*, 105783.

Cano, L. (2015). Income mobility in Ecuador new evidence from individual income tax returns. UNU-WIDER Working Paper No. 2015/040.

CONADE (1965). *Distribución del Ingreso y Cuentas Nacionales en la Argentina, Vols. I-V*. Buenos Aires: CONADE-CEPAL.

Cornia, G. A. (2014). Inequality trends and their determinants: Latin america over 1990 - 2010. In G. A. Cornia (Ed.), *Falling Inequality in Latin America: Policy Changes and Lessons*. Oxford University Press.

Cowell, F. A. (2011). *Measuring Inequality* (Third ed.). Oxford: Oxford University Press.

De Rosa, M., I. Flores, and M. Morgan (2020, October). Income inequality series for latin america. World Inequality Lab – Technical Note n2020/02.

Deaton, A. (2005, 2). Measuring Poverty in a Growing World (or Measuring Growth in a Poor World). *Review of Economics and Statistics 87*(1), 1–19.

Fixler, D., D. Johnson, A. Craig, and K. Furlong (2017). A consistent data series to evaluate growth and inequality in the national accounts. *Review of Income and Wealth 63*, S437–S459.

Flores, I. (2021). The capital share and income inequality: Increasing gaps between micro and macro-data. *The Journal of Economic Inequality 19*(4), 685–706.

Flores, I., C. Sanhueza, J. Atria, and R. Mayer (2020). Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017. *Review of Income and Wealth 66*(4), 850–874.

Gasparini, L., J. Bracco, L. Galeano, and M. Pistorio (2018). Desigualdad en países en desarrollo: ¿ajustando las expectativas? Documentos de Trabajo del CEDLAS; no. 224.

IBGE (2016). *Sistema de Contas Nacionais Brasil: Ano De Referência 2010*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística.

ILO (2019). The Global Labour Income Share and Distribution. ILO Department of Statistics, Methodological description.

Jantti, M., V.-M. Törmälehto, and E. Marlier (2013). *The use of registers in the context of EU? SILC: challenges and opportunities.* Luxembourg: Publications Office of the European Union.

Kuznets, S. (1953). *Shares of Upper Income Groups in Income and Savings*. NBER.

Kuznets, S., L. Epstein, and E. Jenks (1941). *National Income and Its Composition, 1919–1938, Volume I.* National Bureau of Economic Research.

López-Calva, L. F. and N. Lustig (2010). *Declining inequality in Latin America : a decade of progress?* Washington: Brookings Institution Press.

Messina, J. and J. Silva (2017). *Wage Inequality in Latin America: Understanding the Past to Prepare for the Future*. The World Bank.

Morgan, M. and P. Souza (2019). Distributing growth during late-development: New evidence on long-run inequality in brazil. CLADHE IV, Universidad de Santiago, Chile. July 23-25 2019.

Nolan, B., M. Roser, and S. Thewissen (2019, 9). GDP Per Capita Versus Median

Household Income: What Gives Rise to the Divergence Over Time and how does this Vary Across OECD Countries? *Review of Income and Wealth 65*(3), 465–494.

OECD (2013). *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. OECD Publishing.

Ravallion, M. (2003, 8). Measuring Aggregate Welfare in Developing Countries: How Well Do National Accounts and Surveys Agree? *Review of Economics and Statistics 85*(3), 645–652.

Rodríguez-Castelán, C., L. F. López-Calva, N. Lustig, and D. Valderrama (2016). Understanding the dynamics of labor income inequality in latin america. Technical report, Policy Research Working Paper 7795, The World Bank.

Rossignolo, D., N. Oliva, and N. Villacreses (2016). Cálculo de la concentración de los altos ingresos sobre la base de los datos impositivos: un análisis para el Ecuador. Serie Macroeconomía del Desarrollo No. 184.

Törmälehto, V.-M. (2011). LIS and national accounts comparison. LIS Technical Working Paper Series No. 2.

United Nations (2008). *The System of National Accounts 2008*. European Commission, IMF, OECD, UN, WB.

World Inequality Lab (2020). Distributional National Accounts Guidelines: Methods and Concepts Used in the World Inequality Database. WID.world Working Paper series.

Zuñiga-Cordero, A. (2018). Income inequality in costa rica: An inquiry towards distributional national accounts (dina). Master's dissertation, Paris School of Economics, Paris.

Zwijnenburg, J. (2019, 5). Unequal Distributions: EG DNA versus DINA Approach. *AEA Papers and Proceedings 109*, 296–301.

A Appendix

A.1 Excluded countries

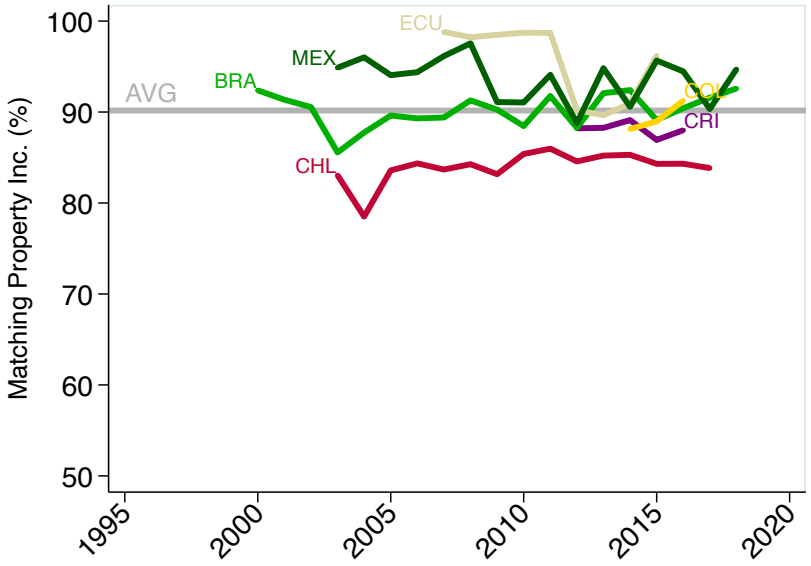
Table A.1: Countries with insufficient data (Excluded)

	Survey microdata		
Country	Source	Sample size, thousands of individuals	Availability
Bahamas	Bahamas Living Conditions Survey	6	2001
Belize	-	-	-
Bolivia	Encuesta de Empleo,Desempleo y Subempleo, Insituto Nacional de Estadística y Censo (INE)	15 – 40	2000-2019
Cuba	-	-	-
Dominican Republic	Encuesta Nacional de Fuerza de Trabajo (ENFT)	15 – 30	2000-2019
Guatemala	Encuesta Nacional de Condiciones de Vida and Encuesta Nacional de Empleo e Ingresos	10 – 70	2000, 2002-2004, 2006, 2011, 2014
Guyana	-	-	-
Haiti	-	-	-
Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples (EPHPM), Institutio Nacional de Estadisticas (INE)	20 – 100	2001-2018
Jamaica	-	-	-
Nicaragua	Encuesta Nacional de Hogares sobre Medición de Nivel de Vida, Instituto Nacional de EStadística y Censos de Nicaragua	20 – 35	2001, 2005, 2009, 2014
Panama	Encuesta de Hogares, Instituto Nacional de Estadística y Censo (INEC)	40 – 55	2000-2019
Paraguay	Encuesta Integrada de Hogares (EIH) and Encuesta Permanente de Hogares (EPH) from 2002, Dirección General de Estadística, Encuestas y Censos (DGEEC)	15 – 40	2001-2019
Suriname	-	-	-
Trinidad and Tobago	-	-	-
Venezuela	Encuesta de Hogares Por Muestreo (EHM), Oficina Central de Estadística e Informática	80 – 240	2000-2006

Note. Own elaboration.

A.2 Consistency of capital incomes from surveys and national accounts

Figure A.1: Share of conceptually consistent property incomes



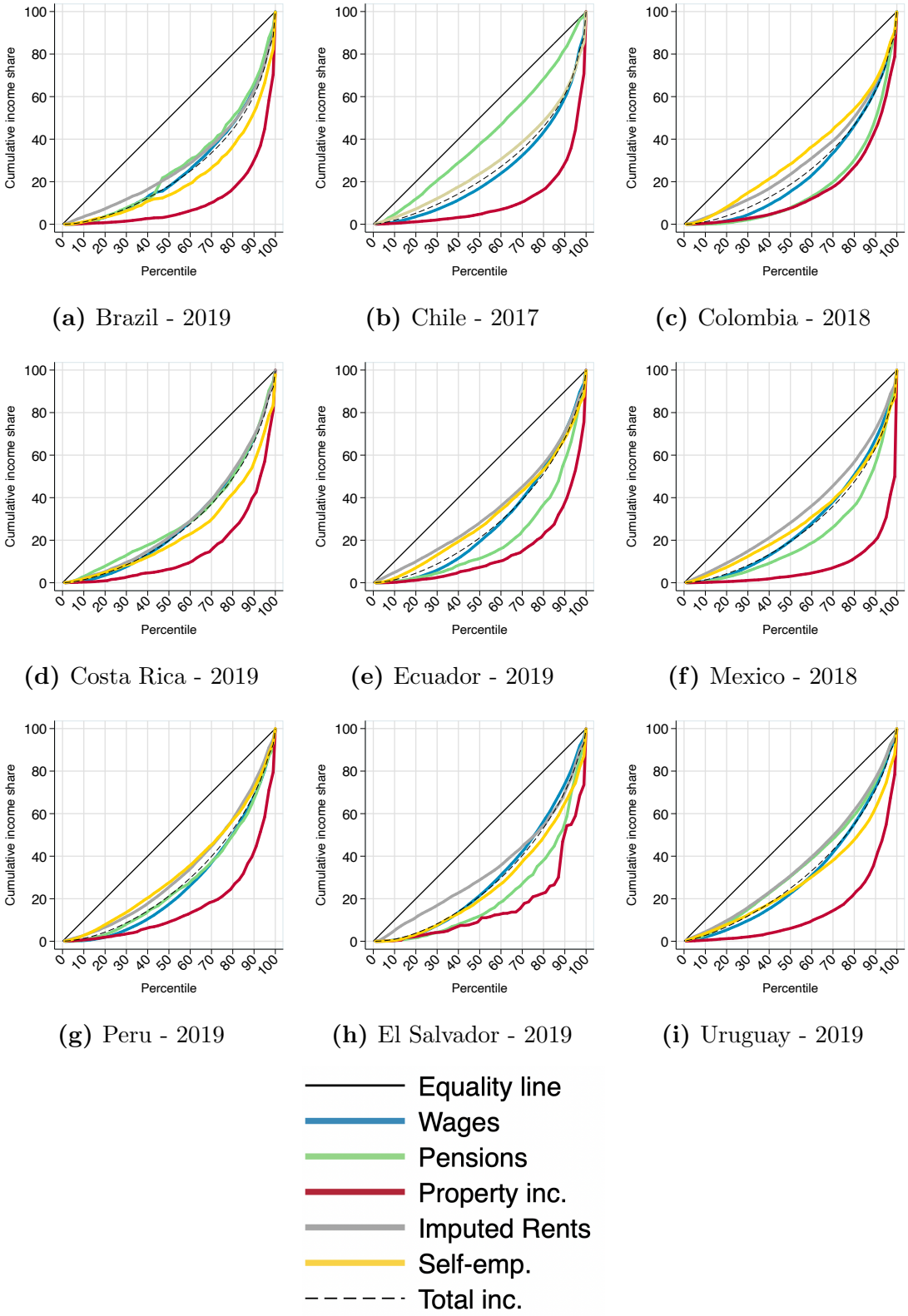
Notes. The share of property incomes from SNA that matches the definition of surveys' capital incomes (dividends and interests) is mostly above 80% of total property income, closer to 90% in most cases. Conceptual differences thus seem to play a minor role in the underestimation of capital incomes displayed in figure 4c. The level of detail that is necessary to observe this is rare in Latin America. Non-matching concepts for the household sector are SNA items D.44, which is composed by D.441, D.442 and D.443 (see table 2). Own elaboration based on the public national accounts reported by each country's relevant institutions.

A.3 The composition and distribution of income in household surveys

To best understand the distributional impact of the under- or over-estimation of each item on the inequality estimates, figure A.2 displays Lorenz curves of total income (dashed line), along with the cumulative distribution of each income item (colored lines). They all rank individuals by increasing total income. In all cases, we confirm that capital income is the most unequally distributed component. This is especially true in the biggest economies in the region: Brazil, Chile and Mexico. In these cases, the top 10% richest households receives between 70% and 80% of the capital income declared in the survey. Other countries such as Colombia, Costa Rica, Peru, and Uruguay are closer to 60%.

Focusing on the bottom left corner of the subpanels of figure A.2, one can distinguish the income components that are more relevant for poorer households, such as self-employment income and imputed rents. For lower deciles, the curve representing both components is higher than the ones representing other incomes in most cases, but particularly in Colombia, Ecuador, Mexico, and Peru. They appear to be less unequally distributed, since they are found in both poorer and richer households. For instance, the income of both shopkeepers, or street vendors, and doctors is usually included as self-employment income. Similarly, the imputed rent to homeowners includes both housing projects acquired through vouchers and luxury dwellings. In the case of Mexico, for instance, where both types of incomes are greatly underestimated (figures 4e and 4f), if we were to adjust the survey –say, by scaling these incomes proportionally– to include the missing part, it would thus have and equalizing impact on overall inequality estimates. Similar conclusions can be reached for Colombia, whereas, in the case of Peru, for which imputed rents are over-estimated, the same adjustment would probably increase inequality. In the cases of Brazil and Costa Rica, self-employment income is overestimated and seems to be more unequally distributed than other incomes (with the exception of capital income). A proportional adjustment of these incomes would thus probably have an equalizing effect. In contrast, adjusting to match higher aggregates of capital income would lead to substantial increases in the level of inequality in all countries.

Figure A.2: Income incidence in latest survey



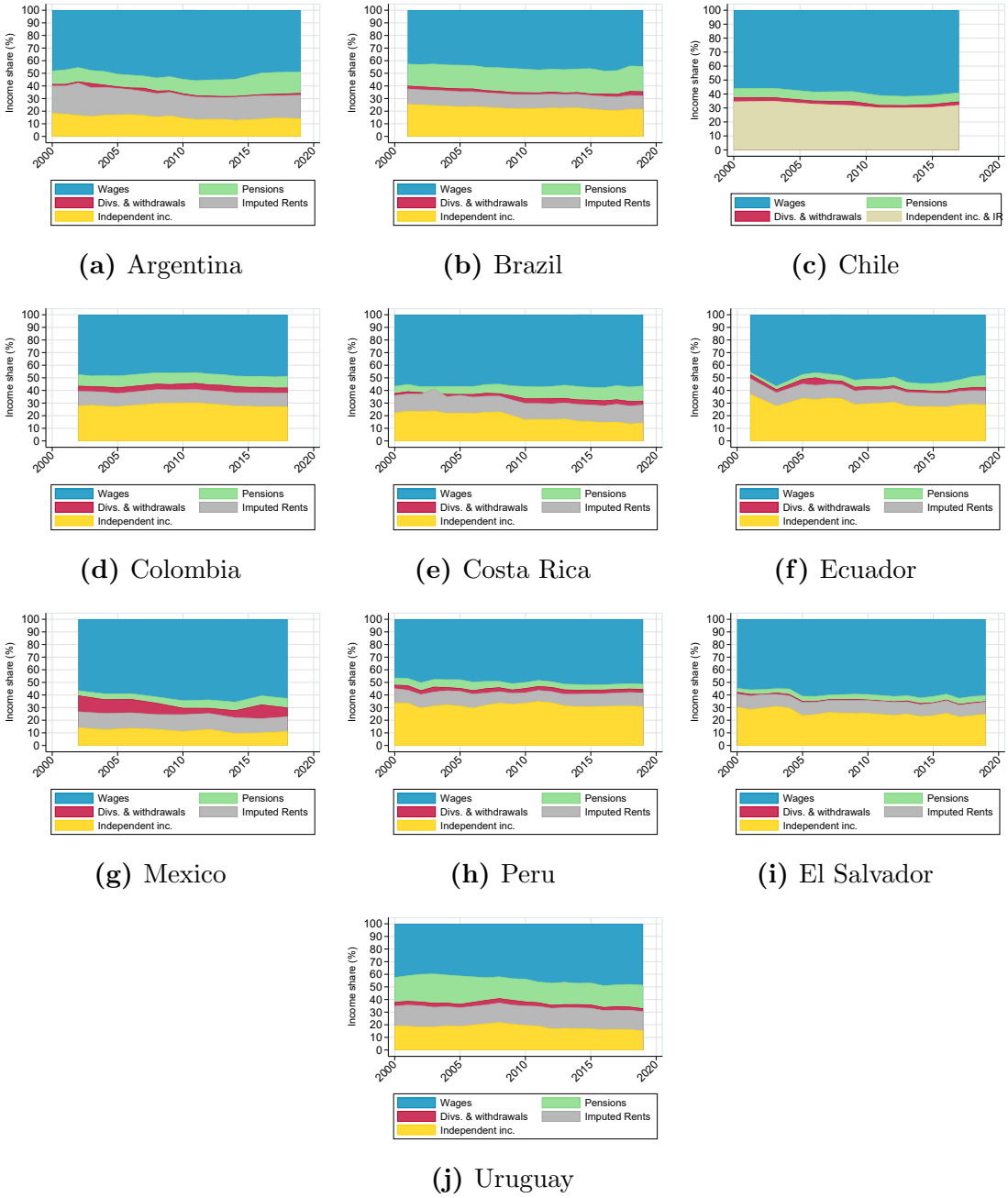
(a) Brazil - 2019 **(b)** Chile - 2017 **(c)** Colombia - 2018

(d) Costa Rica - 2019 **(e)** Ecuador - 2019 **(f)** Mexico - 2018

(g) Peru - 2019 **(h)** El Salvador - 2019 **(i)** Uruguay - 2019

Note. Individuals are ranked by increasing total income. Income is pretax, net of pension contributions. Own elaboration based on ECLAC's harmonized surveys.

Figure A.3: Total income composition



(a) Argentina **(b)** Brazil **(c)** Chile

(d) Colombia **(e)** Costa Rica **(f)** Ecuador

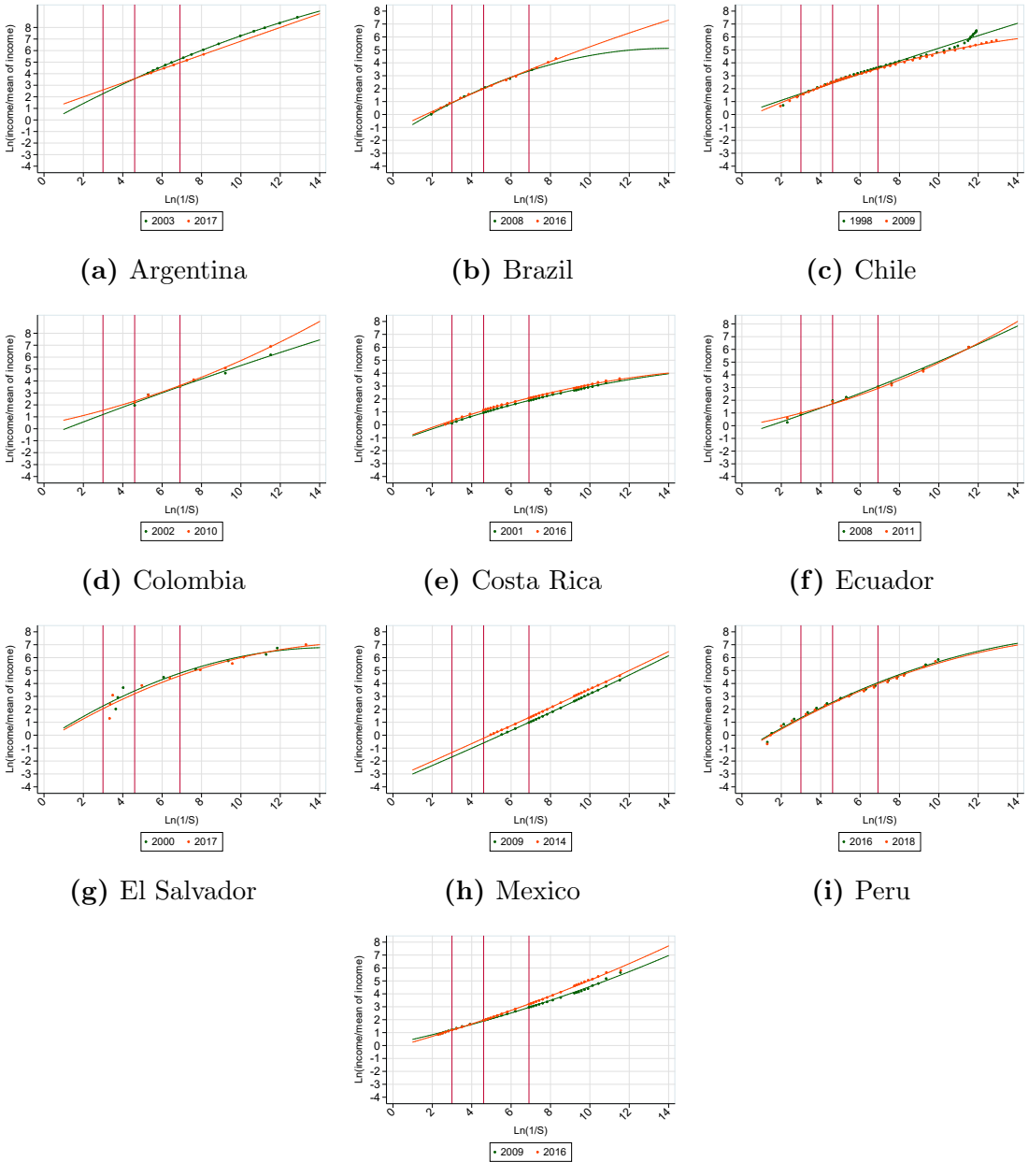
(g) Mexico **(h)** Peru **(i)** El Salvador

(j) Uruguay

Note. Own elaboration based on ECLAC's harmonized surveys. Income is pretax, net of pension contributions.

A.4 Further tables and figures

Figure A.4: Income in relation to rank, by country



(a) Argentina **(b)** Brazil **(c)** Chile

(d) Colombia **(e)** Costa Rica **(f)** Ecuador

(g) El Salvador **(h)** Mexico **(i)** Peru

(j) Uruguay

Note. Own elaboration country's tax records. First and last available years. The y axis depicts the log of income as proportion to mean income, while the x axis depicts the log of $1/S$, with S being the survival function. Vertical lines represent the thresholds for top 5 %, top 1% and top 0.1%, respectively.