Measuring Inequality from Top to Bottom

Tania Diaz-Bazan

World Bank*

October 6, 2014

Abstract

This paper presents a new methodology to measure inequality that optimally combines household survey information and tax records to construct a complete income distribution. Combining the two data sources is necessary because, on the one hand, household surveys do not accurately represent the wealthiest segment of the population, while tax records do; on the other hand, the opposite is true for the lower end of the income distribution: tax records only include incomes above a certain threshold. The key innovation of the proposed methodology – and the main difference from the existing literature (e.g. Atkinson 2007) – is the choice of an optimal income threshold b. The Gini coefficient for the population is then computed combining the conditional income distributions for incomes below b (using household survey data) and above b (using tax records). Central to this methodology is the fact that b is not chosen arbitrarily: it should be determined in such a way as to minimize reliance on household survey data to compute the top of the income distribution. In practice, the optimal b corresponds to the minimum income level that triggers mandatory tax filing. The proposed methodology is applied to the case of Colombia.

*I am grateful to my thesis advisor Alberto Porto for early discussions on the topic, and for his constant and helpful supervision in developing the idea. I am also grateful to Stefano Giglio for his thoughtful suggestions and to Louise Cord, Lea Gimenez Duarte, and Carlos Rodriguez Castelan for many helpful discussions. Email: tdiazbazan@worldbank.org.

Electronic copy available at: https://ssrn.com/abstract=2506219
Electronic copy available at: http://ssrn.com/abstract=2506219

1 Introduction

Welfare measures such as poverty and inequality are central to the study of economic development. The causes and consequences of inequality have always been a key concern for governments and scholars, but interest in the evolution of inequality has recently spiked. This paper focuses on how to optimally measure inequality.

Household surveys have traditionally been the main data source to estimate welfare measures like poverty and inequality. However, they are subject to measurement error, underreporting, and non-response, among other problems (Székely and Hilgert 1999, Cowell and Victoria-Feser 2003, Atkinson 2007, Burkhauser et al. 2009, Burkhauser et al. 2012). These issues represent a particularly severe problem when measuring inequality: they mostly affect the top of the income distribution, and this can lead to an incorrect estimation of the levels and trends in the Gini coefficient. If non-response rates were constant over time, the measurement of changes in inequality over time would not be greatly affected by these problems (Gasparini et al. 2000). However, it is likely that these rates do fluctuate over time.[1]

The income underreporting level at the top of the distribution is not constant and homogeneous – it can vary by economic strata, by source of income, and over time. Yet there is no mechanism available to detect and model precisely the level of underreporting in household surveys. The most notable pattern is that richer households are more hesitant to disclose their income and assets, which would explain why underreporting tends to affect the top of the income distribution, ultimately resulting in an underestimation of the Gini coefficient (Székely and Hilgert 1999). Time variation in underreporting at the top of the income distribution will also affect the estimated time series of the coefficients (see Burkhauser et al. 2004, 2009, Piketty and Saez 2006b, and Atkinson 2007). Finally, household surveys may fail to represent all types of income sources (for example capital gains), which may result in further bias (Burkhauser et al. 2012).

[1]Statistical institutes have partly succeeded in addressing the non-response issue in household surveys through the use of multivariate regressions with demographic and socioeconomic characteristics.

Electronic copy available at: https://ssrn.com/abstract=2506219
Electronic copy available at: http://ssrn.com/abstract=2506219

An alternative data source used to study top incomes in a population is income tax records, which contain extensive information on firms and individuals that are often not captured by traditional data collection mechanisms such as household surveys. Therefore, tax records can provide valuable input to improve the measurement of income inequality, with the caveat that they also present important limitations. Not only do tax records contain little information about the bottom segments of the income distribution (where individuals are not required to file a tax return), but also filers have the financial incentive to report their income in a way that limits their tax liabilities, therefore reducing the marginal tax rate (Burkhauser et al. 2012), and the estimation could be biased due to tax evasion. That said, examining the top of the income distribution is essential to understanding inequality, and tax records are a fundamental source for this purpose.

As shown by the key contributions of Atkinson (2007) and Alvaredo (2011), the share of income concentrated at the top of the income and wealth distribution can have dramatic effects on inequality. Recent studies have therefore proposed to rely on both household surveys and tax records to improve the estimation of the Gini coefficient. In particular, Atkinson (2007) proposes an approximate methodology to estimate the Gini coefficient. The first stage of his methodology consists in choosing a group of top income earners (say, the top 0.1 percent). Then, if this group of top earners is approximately an infinitesimal fraction of the population, with share S^* of total income, the Gini coefficient can be approximated by the following calculation: $S^* + (1 - S^*)\mathcal{G}$. In this formula, $\mathcal{G}$ is the Gini coefficient of the population that excludes the top earners. More recently, Alvaredo (2011) extends this methodology to the case in which the group of top earners chosen is not infinitesimal relative to the size of the population (for example, the top 1 or 5 percent of the income distribution).

These methodologies have two salient features worth noting. First, the choice of which group of top earners to consider in the formula is arbitrary – the top income group is not optimally chosen based on specific criteria. Second, the Gini coefficient $\mathcal{G}$ for the population that does not include the chosen top earners group is computed using the household surveys. This means that, since we cannot determine precisely

Electronic copy available at: https://ssrn.com/abstract=2506219

the extent of underreporting at the top of the income distribution, it is not clear that the $\mathcal{G}$ estimated from household surveys will correctly capture the Gini coefficient of the population that excludes the selected top earners. For example, suppose that we choose to apply the methodology focusing on the top 0.1 percent of the population, so S^* in the formula above would be the share of income of the top 0.1 percent computed from tax records. Suppose also that the household survey is only representative of the bottom 95 percent of the population, because individuals above the 95th percentile do not correctly report their income. In that case, the $\mathcal{G}$ computed from the household survey would not correspond to the true Gini coefficient of the population outside the selected group of top earners (in this case the bottom 99.9 percent), since the 95-99.9 percentiles are not covered at all in the household survey. The methodology would then suffer from omitting a nontrivial segment of the population from the analysis.

In this paper we build upon the existing methodologies (particularly, Atkinson 2007 and Alvaredo 2011) and propose a new methodology aimed to optimally combine the two data sources, accounting explicitly for the underreporting of the top of the income distribution in the household survey. The methodology is based on finding the optimal income threshold b such that the following statements hold true: (1) the segment of the population with incomes below b is well represented in the household survey data; (2) those with incomes above b are captured in the tax records; and (3) b minimizes the distortions attributed to household survey underreporting at the top of the income distribution. This paper shows that if b can be found (possibly below the income level of the top 0.1 or 1 percent), the population's full unconditional distribution of income can be estimated by combining the Gini coefficients of the population segments with incomes below b (using household survey data) and above b (using tax records). Since the underreporting distortions in the household survey are stronger for higher incomes, the optimal b that satisfies these requirements is the lowest level of income that triggers mandatory tax filing. This is the key element of the proposed methodology. Note that this income threshold can be significantly below the 1 percent traditionally used in the existing literature.

Electronic copy available at: https://ssrn.com/abstract=2506219

The proposed methodology, therefore, solves the issue of arbitrarily choosing the top income group, instead relying on the optimal income threshold b, which gives the maximum weight possible to tax records at the top of the income distribution. In addition, the methodology explicitly takes into account the underreporting of household surveys at the top of the income distribution, and only relies on household survey data to compute the Gini coefficient for the segment of the income distribution that lies below the threshold b.

This paper is organized as follows. After presenting the methodology and its properties in Section 2, the paper demonstrates an application to the case of Colombia (Section 3). Although in recent years an increasing number of governments have made income tax data available to the public, Latin American countries have lagged behind. Colombia is the first country in the region to make personal income tax micro-data available to researchers, which determined the case study selection in this analysis. Section 4 offers conclusions.

2 Methodology

This section describes a simple procedure to estimate the entire income distribution in a population by combining the information in two datasets that are only informative about conditional income distributions. In this case, we focus on the fact that household surveys are representative of the bottom segment of the income distribution but not the top, while tax records capture the top end of the income distribution but not the bottom. The proposed methodology applies a result of Dagum (1997), who showed how to decompose a population's Gini coefficient into a combination of the Gini coefficients of its subpopulations. A related methodology has been applied by Alvaredo (2011) to measure income inequality in the United States and Argentina. The end of this section summarizes the main differences between the proposed methodology and the previous literature.

Electronic copy available at: https://ssrn.com/abstract=2506219

2.1 Setup and Assumptions

Suppose that the cumulative distribution function of income y in a population is $F(y)$. We would like to obtain an estimate of $F(y)$ and some related statistics (like the Gini coefficient, $\mathcal{G}[F]$). However, we can only obtain consistent estimates of some *conditional* distributions of y. In particular, we assume:

Assumption 1. *Define*

$$F_1(y) = Pr\{Y \leq y | Y \leq b\}$$

$$F_2(y) = Pr\{Y \leq y | Y > b\}$$

for some b. We observe $\hat{F}_1(y)$ and $\hat{F}_2(y)$, pointwise consistent estimators of $F_1(y)$ and $F_2(y)$ respectively.

The first assumption simply states that out of the available information (for example, household surveys or tax records) we can obtain consistent estimators of the conditional distribution of income; $\hat{F}_1$ estimates the conditional distribution below an income threshold b, while $\hat{F}_2$ estimates the conditional distribution for incomes above b.

Assumption 2. *$\hat{F}(b)$, a consistent estimator of $F(b)$, is observed.*

The second assumption is necessary to combine the information in $\hat{F}_1$ and $\hat{F}_2$ correctly into an estimator of F. This assumption does not require knowing the full distribution F, only the value at point b.

2.2 Results

We provide two simple results. The first proposition shows that under assumptions 1 and 2, we can obtain a consistent estimate of the distribution F. The second proposition shows that it is possible to express the Gini coefficient of the income distribution F as a combination of Gini coefficients constructed using the conditional distributions F_1 and F_2.

Electronic copy available at: https://ssrn.com/abstract=2506219

Proposition 1. *Under Assumptions 1 and 2, the estimator $\hat{F}(y)$ constructed as:*

$$\hat{F}(y) \equiv \begin{cases} \hat{F}(b)\hat{F}_1(y) & y \leq b \\ \hat{F}(b) + (1 - \hat{F}(b))\hat{F}_2(y) & y > b \end{cases}$$

is a pointwise consistent estimator of $F(y)$.

Proof. In the population, by the definition of conditional distribution we have, for $y \leq b$,

$$F_1(y) \equiv Pr\{Y \leq y | Y \leq b\} = \frac{Pr\{Y < y\}}{Pr\{Y < b\}} = \frac{F(y)}{F(b)}$$

so

$$F(y) = F_1(y)F(b)$$

and for $y > b$

$$F_2(y) \equiv Pr\{Y \leq y | Y > b\} = \frac{Pr\{Y \leq y \text{ and } Y \geq b\}}{1 - Pr\{Y \leq b\}} = \frac{F(y) - F(b)}{1 - F(b)}$$

so

$$F(y) = F(b) + (1 - F(b))F_2(y)$$

Given that under Assumptions 1 and 2 we have consistent estimators for all the right-hand side variables, the consistency of F(y) follows immediately. □

Proposition 1 tells us how we can reconstruct the underlying unconditional income distribution if we know the two conditional distributions above and beyond an income threshold b and the fraction of the population with income lower than b. From the estimated $\hat{F}(y)$, we can then study the properties of the income distribution.

In some cases, it may be convenient to be able to obtain directly the Gini coefficient for the full population, $\mathcal{G}[F]$. The next proposition shows that the Gini coefficient $\mathcal{G}[F]$ can be obtained as a simple linear combination of the Gini coefficients computed on the conditional distributions. We start from a few definitions.

Electronic copy available at: https://ssrn.com/abstract=2506219

Definition. The Gini coefficient $\mathcal{G}[F]$ is defined as:

$$\mathcal{G}[F] = \frac{1}{\mu} \int_0^\infty F(y)(1 - F(y))dy$$

where $\mu \equiv E[y]$ is the unconditional mean of y (i.e., the population mean).

Definition. The Gini coefficient of a conditional distribution $\mathcal{G}[F_j]$, $j \in \{1, 2\}$, is defined as:

$$\mathcal{G}[F_j] = \frac{1}{\mu_j} \int_0^\infty F_j(y)(1 - F_j(y))dy$$

where $\mu_1 = E[y|y \leq b]$ and $\mu_2 = E[y|y > b]$.[2]

These Gini coefficients are simply the standard Gini coefficients computed on each of the two distributions.

Proposition 2. *The full-population (unconditional) Gini coefficient can be written as:*

$$\mathcal{G}[F] = (1 - F(b))F(b)[\frac{\mu_2 - \mu_1}{\mu}] + \left(F(b)^2 \frac{\mu_1}{\mu}\right) \mathcal{G}[F_1] + \left([1 - F(b)]^2 \frac{\mu_2}{\mu}\right) \mathcal{G}[F_2]$$

where

$$\mu = F(b)\mu_1 + (1 - F(b))\mu_2$$

Proof. See Appendix. □

This proposition tells us that we can estimate the overall Gini coefficient as a linear combination of the Gini coefficients of the conditional distributions, plus an adjustment term that takes into account that the two conditional distributions are obtained from different underlying portions of the unconditional distribution.

[2]Note that given the definitions of F_1 and F_2 above, $F_1(y)$ is 1 for $y > b$, so $\int_0^\infty F_1(y)(1 - F_1(y))dy = \int_0^b F_1(y)(1 - F_1(y))dy$. Viceversa, $F_2(y)$ is 0 for $y \leq b$, so $\int_0^\infty F_2(y)(1 - F_2(y))dy = \int_b^\infty F_2(y)(1 - F_2(y))dy$.

Electronic copy available at: https://ssrn.com/abstract=2506219

2.3 Implementation and Relation to the Literature

A crucial prerequisite in the derivation of Propositions 1 and 2 is that we can obtain a consistent estimate of the conditional distribution of income above and below a threshold b. The specific case of combining household surveys with tax records provides a natural example in which to apply this methodology.

Household surveys tend to be poorly representative of the top of the income distribution because high income individuals tend to underreport or simply not report their income. While the survey data will have a bias that primarily affects the top of the income distribution, the bottom segment of the population might be much better represented in the survey. For simplicity purposes, suppose that all incomes below a threshold a are well captured in the household survey. Similarly, for the case of tax records, only individuals with incomes above a threshold c report taxes, so only this high-income population is well represented. In this case, the dataset will correctly represent the conditional distribution of income above that threshold.

Suppose now that $c < a$. In other words, suppose that the level of income that triggers tax reporting is below the level of income above which we fear significant censoring in the household survey. Then, any income level b between c and a ($c \leq b \leq a$) will satisfy the following two properties: (1) all incomes above b are well represented in the tax records data (since it is above the minimum threshold c that triggers tax filing); and (2) all incomes below b are well represented in the household survey, because censoring only occurs above a and $b < a$.

Any such b would be a valid income threshold to apply the methodology above. In practice, it is hard to know where censoring in the household survey starts (a). This suggests that rather than viewing b as a purely free parameter, one should optimally choose b to minimize the potential bias due to household survey underreporting at the top of the income distribution. If tax records are representative of incomes above the minimum threshold to file (c), the natural choice for the optimal b would be precisely c. In other words, since we fear household survey underreporting at the top of the income

Electronic copy available at: https://ssrn.com/abstract=2506219

distribution, we optimally choose the lowest possible b that still ensures that incomes above b are represented in the tax records. The natural choice is the income level that triggers mandatory tax filing.

This intuition about the optimal threshold b and its implications for the estimation of inequality measures are the distinguishing features of the methodology presented in this paper relative to the existing literature. In particular, the methodology of Alvaredo (2011) – applied to the case of Colombia by Alvaredo and Londoño (2013) – combines an estimate of the Gini coefficient for the bottom 99 percent of the income distribution (G^*) with the share of income of the top 1 percent (S), the fraction of the population (P) in that group ($P = 1\%$), and an estimated coefficient of the inverted Pareto distribution (β) that captures the shape of the income distribution function within the 1 percent. The formula used is: $G = (\beta - 1)/(\beta + 1)PS + G^*(1 - P)(1 - S) + S - P$; G^* is estimated using the household survey, while S and β are obtained from tax records.

In this paper, we build upon the contribution of Alvaredo (2011) and propose a methodology that differs from those in several dimensions. First, their formula obtains an estimate of the Gini coefficient of the bottom 99 percent of the population (G^*) from the household survey. However, as previously noted, household surveys are not representative of the full population, which means that G^* does not capture the true bottom 99 percent of the population. Estimating G^* by computing the Gini coefficient of the bottom 99 percent of those households that are present in the household survey will yield a biased result if there is survey underreporting at the top of the income distribution (since it will use the bottom 99 percent of a sample which is truncated at the top, not of the true income distribution). Alternatively, G^* can be estimated as the Gini coefficient of the entire household survey (without excluding the top 1 percent of the respondents), under the assumption that the household survey is effectively representative of the bottom 99 percent of the income distribution due to underreporting at the top. This approach would be correct if we were certain that the household survey misses precisely the top 1 percent. However, if household survey underreporting is significant at a different income level (for example, below the top 1 percent), the estimated

Electronic copy available at: https://ssrn.com/abstract=2506219

G^* would not correspond to the Gini coefficient of the bottom 99 percent of the true income distribution.

A second difference between the two approaches is that, even assuming an accurate estimation of G^*, the formula only corresponds (up to the approximation discussed in Alvaredo 2011) to the one presented in this paper with a choice of b equal to the 99th income percentile, as opposed to the optimal b discussed above. As previously argued, we believe that choosing a lower b (corresponding to the income threshold that triggers mandatory tax filings) results in a better estimate because it reduces the reliance on household survey data for the top of the income distribution.

As shown empirically in Section 3, both features result in interesting differences between the results of Alvaredo and Londoño (2003), based on Alvaredo (2011), and the ones we obtain in this paper for the same country, Colombia. A final note concerns the term $F(b)$ that appears in the equation of Propositions 1 and 2. This term represents the fraction of the population with incomes below b, and needs to be estimated once the preferred threshold b is chosen. We estimate $F(b)$ using the same methodology employed by Atkinson (2007) and Alvaredo (2011) to estimate the share of income that accrues to the top 1 percent of the income distribution. In particular, $b = 1 - \frac{A}{C}$, where A is the total population above the income tax threshold b (computed using the tax records), while C is the control for population (an estimate of the total population of the country).[3]

3 Application: the case of Colombia

Over the past 20 years, the countries of Latin America and the Caribbean have made notable improvements in reducing poverty and income inequality. Income concentration, measured by the Gini coefficient, has been steadily falling in the region, going from 0.58 in 1996 to 0.52 in 2011 (World Bank 2013). However, in the case of Colombia, the Gini co-

[3]The control for population is computed as the total number of individuals over a certain age (here, 20 years) in the country. The threshold depends on the age at which young people enter the labor force. See Atkinson (2007).

Electronic copy available at: https://ssrn.com/abstract=2506219

efficient remained practically stagnant from 2002 to 2012, and declined only marginally during the second part of the decade. Moreover, in 2012 Colombia's Gini coefficient remained higher than the regional average, placing Colombia among the three most unequal countries in LAC and one of the most unequal in the world (World Bank 2013). Furthermore, these numbers do not take into account the potential bias associated with the top of the income distribution, which, if properly accounted for, could result in a higher Gini coefficient.

In recent years, an increasing number of governments have been granting public access to administrative records and other information. The use of administrative records as a statistical tool is a recent trend that increases transparency and makes available to citizens, analysts, and policymakers a greater wealth of information. Tax records contain extensive information on firms and individuals that is often not captured by traditional data collection mechanisms such as household surveys, and, as such, they can provide valuable input to statistical systems. However, many Latin American countries have yet to make public their income tax data. Colombia is the first country to do so, allowing researchers to use the micro-data from personal income tax records. This section applies the methodology described above to compute the Gini coefficient for Colombia in 2010 using data from both household surveys and administrative tax records.

3.1 Data

Around 1.4 percent[4] of Colombia's working population pays income taxes. While 1.1 million individuals (out of a population of 4.5 million) filed an annual tax return in 2010, the rest of the population (3.4 million) paid taxes through withholding. Unlike other countries like the United States, Colombia's income tax system does not allow for joint returns for married couples. This allows us to compare incomes across data sources–household survey and tax records–using the same unit of analysis (individuals).

Colombians file their income tax returns differently depending on whether they are

[4]Dirección de Impuestos y Aduanas Nacionales (DIAN), Exposición de Motivos.

Electronic copy available at: https://ssrn.com/abstract=2506219

small or large taxpayers. This analysis considers both types of taxpayers: those who are not required to keep accounting books (small taxpayers) and those who are required to do so (large taxpayers). The small taxpayer data (Form 210) is composed of a balanced micro-data panel and tabulations made by the tax agency in Colombia (Dirección de Impuestos y Aduanas Nacionales, DIAN). In the case of large taxpayers (Form 110), the dataset is only composed of a balanced micro-data panel.[5] The data provides annual information about labor income, capital gains, other incomes, deductions, exemptions, and taxes paid.

The Great Integrated Household Survey (GEIH) collects information about labor force conditions, socio-demographic characteristics, and different sources of income. The bottom of the income distribution is well represented in the GEIH, which collects data monthly and provides information at the national, urban, and rural levels as well as at the departmental level. In order to compare annual values from the administrative data with those from the household survey, we multiply by 14 the total individual monthly income (12 months plus a bonus equivalent to 2 months' pay).

3.2 Results

In general, all individuals whose incomes exceed a certain threshold appear as a tax unit in the administrative data.[6] Individuals with incomes below this threshold are not captured by the tax records, but they are represented in the household survey. In 2010, this threshold corresponded to 81 million Colombian pesos per year. As explained above, this income level is used to determine the parameter b of the formulas presented in Propositions 1 and 2: the bottom conditional distribution in the propositions will be obtained from the household survey by focusing on all individuals with incomes below b, and the top conditional distribution will use all available tax records (since the threshold b is chosen to be the minimum income level that appears in the tax record).

Figure 1 depicts the distribution of total individual incomes in the household sur-

[5]Since we are only considering a balanced panel in which individuals are observed if they are present

Electronic copy available at: https://ssrn.com/abstract=2506219

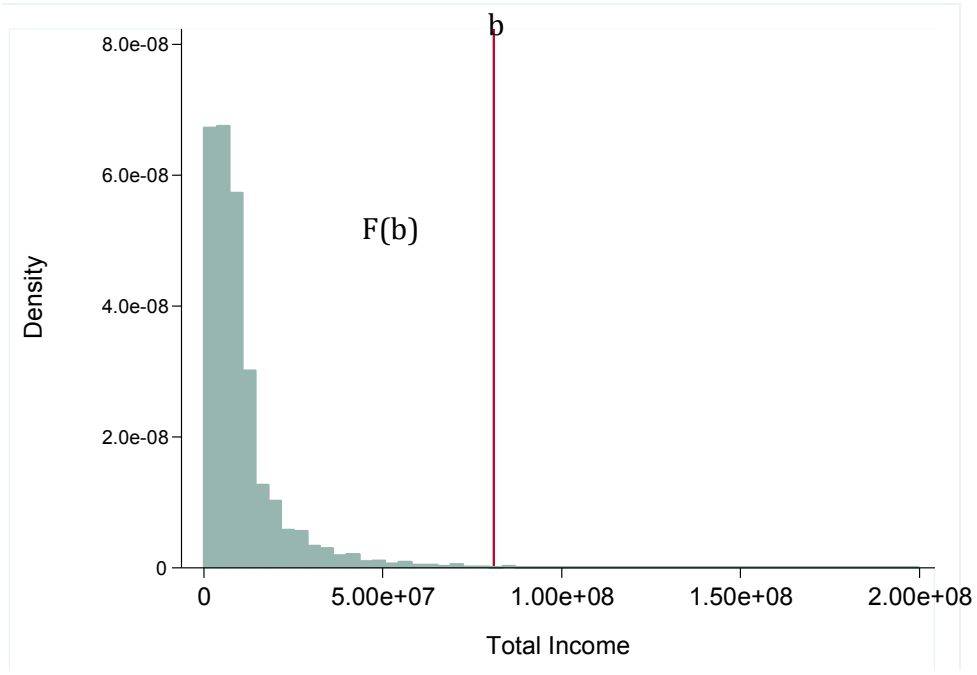


Figure 1: Total Individual Income Distribution for Colombia, Household Survey. Source: author's calculations using the Great Integrated Household Survey (GEIH), 2010. Note: The analysis considers only positive total individual incomes and the population above the age of 20. The red bar indicates the legal threshold for mandatory tax filing, b.

vey.[7] Since tax records are based on individual returns, our estimation only considers the total individual income data collected through household surveys. If more than one person per household files a tax return, they will appear separately, each with their own individual income. We define as a control for population all individuals age 20 and above, since few individuals under 20 years of age[8] contribute income tax revenue; excluding them from the denominator does not significantly affect the results (Atkinson 2007).

The red vertical line in Figure 1 indicates the minimum total individual income

for all the years, the results could be potentially biased.

[6]A caveat is worth mentioning here: the level of tax evasion and its changes over time can affect the results obtained for the Gini coefficient.

[7]We calculate the total income distribution considering all personal incomes: labor income, transfers, remittances, and capital gains, among others. We also include pension claims, since in Colombia pension payments are considered labor income.

[8]The control for population uses population projections data for 2010. The total population is 28,104,576.

Electronic copy available at: https://ssrn.com/abstract=2506219

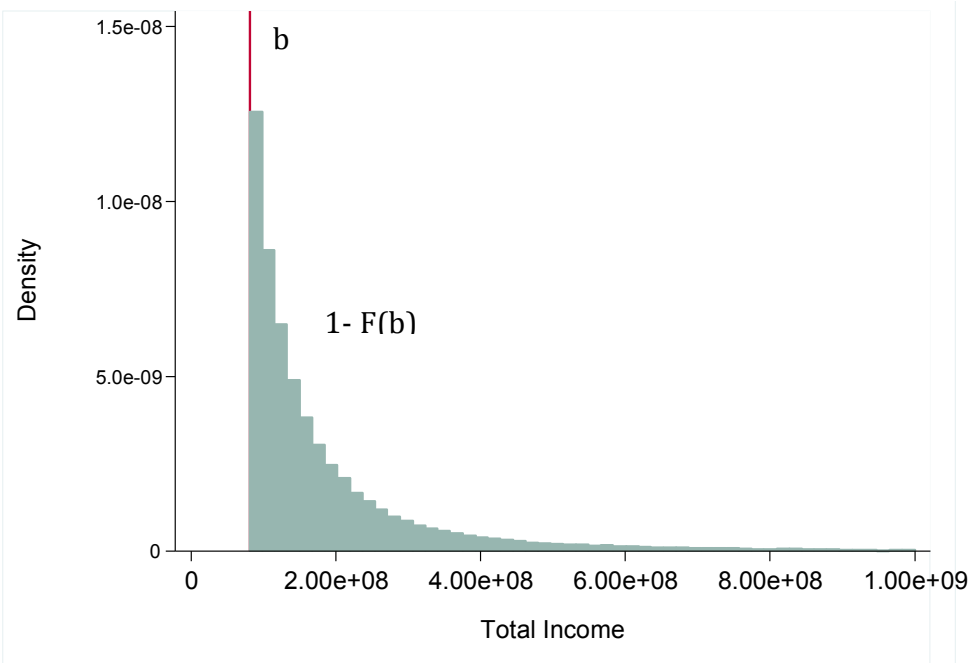


Figure 2: Total Individual Income Distribution for Colombia, Tax Records. Source: author's calculations using income tax data (DIAN), 2010. Note: The analysis considers only positive total individual incomes and the population above the age of 20. The red bar indicates the legal threshold for mandatory tax filing, b.

needed to be part of the tax records data, which, as discussed above, is the threshold b for the methodology presented in this paper. The proportion of individuals who are not required to file a tax return is $F(b)$. The figure illustrates the low representation of the richest individuals in the household survey. Incomes above the threshold b (81 million pesos) are captured by the tax records, as shown in Figure 2.[9] Income is defined to include all sources reported by the individual in the tax return: wages and other labor payments; fees, commissions, and services; interest and financial income; other income (rents) minus 1/6 of the cost of deductions.[10] This definition of income is used in order to make it comparable with the income reported in the household survey, using the same adjustments as Alvaredo and Londoño (2013).

Table 1 shows the estimates of the Gini coefficient using the methodology described

[9]Figure 2 is truncated at the highest income levels for readability.

[10]Using 1/6 of the cost of deductions is ad hoc. The tax file does not provide any information about which deductions are included in this category. For more details, see Alvaredo and Londoño (2013).

Electronic copy available at: https://ssrn.com/abstract=2506219

b	$F(b)$	$\mathcal{G}[F_1]$	$\mathcal{G}[F_2]$	$\mathcal{G}[F]$
79,000,000	0.9854	0.4887	0.4550	0.5979
80,000,000	0.9856	0.4892	0.4544	0.5979
81,031,500	**0.9858**	**0.4897**	**0.4538**	**0.5978**
82,000,000	0.9861	0.4902	0.4532	0.5978
83,000,000	0.9863	0.4905	0.4526	0.5977
106,700,000	0.9900	0.5039	0.4406	0.5960

Table 1: Gini coefficient estimates using different thresholds b. Column (1) reports the chosen coefficient b. Column (2) reports the estimated $F(b)$. Column (3) reports $\mathcal{G}[F_1]$, the Gini coefficient computed using the incomes below b (using the Household Survey), Column (4) reports $\mathcal{G}[F_2]$, the Gini coefficient computed using the incomes above b (using the Tax Records), and Column (5) reports $\mathcal{G}[F]$, the overall estimated Gini coefficient. The bold line corresponds to the optimal choice of b.

in Section 2, setting the threshold b to various levels. The bold row corresponds to the optimal b (the minimum income that mandates filing of the tax record). The first column reports the level of b, the second column the estimated fraction of the population with incomes below b, or $F(b)$. Columns 3 and 4 report the Gini coefficient estimated using the conditional distribution of incomes below b and above b, respectively. Finally, column 5 reports the estimate of the Gini coefficient that combines the two numbers as described in Proposition 2. The resulting preferred estimate of the Gini coefficient is 0.5978. The estimate varies minimally with the choice of b around the optimal threshold of 81 million pesos. However, choosing a higher b–for example focusing on the top 1 percent of the income distribution (b =106 million pesos)–does have a nontrivial impact on the Gini coefficient, by about 0.2 points (see last row of Table 1). Considering that the average annual reduction of the Gini coefficient in Latin American countries over the last 10 years was 0.51 percentage points, the differences in the estimates are not trivial.

An even more striking difference can be seen when comparing these numbers to the ones obtained by Alvaredo and Londoño (2013) for Colombia applying the methodol-

Electronic copy available at: https://ssrn.com/abstract=2506219

ogy of Alvaredo (2011). As mentioned earlier, in this paper we build on his methodology and propose an improvement based on the optimal choice of the income threshold used to combine the household surveys with the tax record. In particular, the main differences between their methodology and the one presented in this paper are: (1) the use of a threshold of 1 percent to combine the two data sets, instead of choosing an optimal threshold b for combining the two; (2) the use of an approximation to combine the Gini coefficient of the bottom of the income distribution obtained from the household survey with the share of income of the top 1 percent of the population; and (3) the calculation of the Gini coefficient for the bottom 99 percent of the income distribution using the bottom 99th percentile of the income distribution obtained from the household survey (ignoring the underrepresentation at the top in the household survey). Applying the Alvaredo and Londoño methodology to the Colombian case, the resulting Gini coefficient would be 0.6076, 0.98 percentage points higher than the estimate obtained using the methodology presented in this paper (the bold row of Table 1). This shows that choosing an optimal threshold to combine the two datasets and explicitly considering the underreporting of incomes in the household survey when computing the Gini coefficients for conditional distributions can have dramatic effects on the measured coefficient.[11]

4 Conclusion

This paper proposes a new methodology to optimally measure inequality by combining household survey and tax record data. The motivation stems from the fact that household surveys poorly represent the top of the income distribution, while tax records cover the top but not the bottom of the distribution. The key innovation of the pro-

[11] When applying the methodology presented in Alvaredo (2011) and Alvaredo and Londoño (2013), instead of estimating the Gini coefficient for the bottom 99 percent of the population using the bottom 99 percent of the household survey's distribution, an alternative could be to simply use the Gini coefficient of the full distribution of the household survey as an estimator of the Gini coefficient for the bottom 99 percent of the true income distribution. This would be correct under the assumption that the household survey underreports incomes exactly above the 99th percentile. The Gini coefficient computed under this assumption (which generally will not be true) would be 0.6468.

Electronic copy available at: https://ssrn.com/abstract=2506219

posed methodology is the choice of an income threshold b used to combine the two data sources; the paper shows that b should not be chosen arbitrarily, but should be chosen optimally to minimize the distortions from household survey underreporting at the top of the income distribution. In particular, we discuss why the optimal income threshold b corresponds to the minimum income level required to file a tax return (i.e., the lowest income captured in the tax records).

After presenting the methodology, we apply it to the case of Colombia. We find that the Gini coefficient for Colombia in 2010 is 0.5978 when computed using an optimal income threshold of 81 million pesos, which is significantly different from the one that would be obtained if using an (arbitrary) threshold corresponding to the top 1 percent of the income distribution (0.5960). The difference of 0.2 points is not trivial, particularly when we consider that over the last 10 years the average annual reduction of the Gini coefficient in Latin America was 0.51 percentage points. The methodology presented in this paper builds on and improves upon the methodologies proposed by Atkinson (2007) and Alvaredo (2011); the difference in the estimated Gini coefficients is as high as 0.98 percentage points. In addition to improving the measurement of inequality, the new proposed methodology is useful for assessing the distributional effects of fiscal policies on the full income distribution, since it also allows to recover the full unconditional distribution of income by combining the conditional distributions observed in the household surveys and in the tax records.

Electronic copy available at: https://ssrn.com/abstract=2506219

References

1. Alvaredo, Facundo. 2011. "A note on the relationship between top income shares and the Gini coefficient." Economics Letters 110 (3): 274–7

2. Alvaredo, Facundo, and Juliana Londoño Velez. 2013. "High Incomes and Personal Taxation in a Developing Economy: Colombia 1993–2010."Commitment to Equity Working Paper No. 12, Tulane University, New Orleans, LA.

3. Atkinson, A.B., 2007. "Measuring Top Incomes: Methodological Issues." In: Atkinson, A., Piketty, T. (Eds.), Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries. Oxford University Press, Oxford.

4. Burkhauser, Richard V., Kenneth A. Couch, Andrew Houtenville, and Ludmila Rovba. 2004. "Income Inequality in the 1990s: Re-Forging a Lost Relationship?," Working Papers 2004–11, University of Connecticut, Department of Economics.

5. Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 2009. "Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data." NBER Working Papers 15320, National Bureau of Economic Research, Inc.

6. ———. 2012. "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data." The Review of Economics and Statistics 94 (2): 371–88.

7. Cowell, Frank A., and Maria-Pia Victoria-Feser. 2000. "Distributional Analysis: A Robust Approach." In Putting Economics to Work, eds. Anthony Atkinson, Howard Glennerster, and Nicholas Stern, pp. 205–262. London: STICERD, London School of Economics.

8. ———. 2003. "Distribution-free inference for welfare indices under complete an incomplete information." Journal of Economic Inequality 1: 191–219.

9. Dagum, Camilo. 1997. "A New Approach to the Decomposition of the Gini Income Inequality Ratio." Empiral Economics 22: 515–31.

Electronic copy available at: https://ssrn.com/abstract=2506219

10. Dirección de Impuestos y Aduanas Nacionales (DIAN). 2012. "Exposición de motivos al proyecto de Ley." Ministerio de Hacienda y Crédito Público, Government of Colombia, Bogotá.

11. Estatuto Tributario Nacional. 2013. Government of Colombia, Bogotá.

12. Gasparini, Leonardo, Mariana Marchionni, and Walter Sosa Escudero. 2000. "La distribución del ingreso en Argentina y en la provincia de Buenos Aires." Ministerio de Economía de la Provincia de Buenos Aires, República Argentina.

13. Székely, Miguel, and Marianne Hilgert. 1999. "What's Behind the Inequality We Measure: An Investigation Using Latin American Data." Working Paper No. 409, Inter-American Development Bank, Washington, DC.

14. World Bank. 2013. "Shifting Gears to Accelerate Shared Prosperity in Latin America and the Caribbean." Document 78507, World Bank, Washington, DC.

Electronic copy available at: https://ssrn.com/abstract=2506219

Appendix

Proof of Proposition 2

Start by writing:

$$\mathcal{G}[F] = \frac{1}{\mu}\int_0^{\infty} F(y)(1-F(y))dy$$

$$= \frac{1}{\mu}\int_0^{b} F(y)(1-F(y))dy + \frac{1}{\mu}\int_b^{\infty} F(y)(1-F(y))dy$$

Substituting the expressions above:

$$\mathcal{G}[F] = \frac{1}{\mu}\int_0^{b} F(b)F_1(y)(1-F(b)F_1(y))dy$$

$$+\frac{1}{\mu}\int_b^{\infty} [F(b)+(1-F(b))F_2(y)](1-F(b)-(1-F(b))F_2(y)])dy$$

Now note that

$$F_1(y)F(b)(1-F_1(y)F(b)) =$$

$$F_1(y)F(b) - F_1(y)^2F(b)^2 =$$

$$= F(b)^2[F_1(y)-F_1(y)^2] + F_1(y)[F(b)-F(b)^2]$$

Similarly, we obtain

$$[F(b)+(1-F(b))F_2(y)][1-F(b)-(1-F(b))F_2(y)]$$

$$= F(b)+(1-F(b))F_2(y)$$

$$-[F(b)+(1-F(b))F_2(y)]^2$$

$$= F(b)+(1-F(b))F_2(y) - [F(b)^2+(1-F(b))^2F_2(y)^2+2F(b)(1-F(b))F_2(y)]$$

$$= F(b)+(1-F(b))F_2(y) - F(b)^2-(1-F(b))^2F_2(y)^2-2F(b)(1-F(b))F_2(y)$$

$$= F(b)+(1-F(b))F_2(y) - F(b)^2+(1-F(b))^2\left[F_2(y)-F_2(y)^2\right]$$

$$-(1-F(b))^2F_2(y)-2F(b)(1-F(b))F_2(y)$$

Electronic copy available at: https://ssrn.com/abstract=2506219

$$= F(b) + \left[(1-F(b)) - (1-F(b))^2 - 2F(b)(1-F(b))\right] F_2(y)$$

$$-F(b)^2 + (1-F(b))^2 \left[F_2(y) - F_2(y)^2\right]$$

$$= F(b) + \left[1 - F(b) - (1 + F(b)^2 - 2F(b)) - 2F(b) + 2F(b)^2\right] F_2(y)$$

$$-F(b)^2 + (1-F(b))^2 \left[F_2(y) - F_2(y)^2\right]$$

$$= F(b) + \left[-F(b) + F(b)^2\right] F_2(y) - F(b)^2 + (1-F(b))^2 \left[F_2(y) - F_2(y)^2\right]$$

$$= [F(b) - F(b)^2] - F(b)(1-F(b))F_2(y) + (1-F(b))^2 \left[F_2(y) - F_2(y)^2\right]$$

Substituting:

$$\mathcal{G}[F] = \frac{1}{\mu} \int_0^b \left\{F(b)^2[F_1(y) - F_1(y)^2] + F_1(y)[F(b) - F(b)^2]\right\} dy$$

$$+\frac{1}{\mu} \int_b^\infty \left\{[F(b) - F(b)^2] - F(b)(1-F(b))F_2(y) + (1-F(b))^2 \left[F_2(y) - F_2(y)^2\right]\right\} dy$$

$$\mathcal{G}[F] = \frac{F(b)}{\mu} \int_0^b \left\{F(b)[F_1(y) - F_1(y)^2] + F_1(y)[1 - F(b)]\right\} dy$$

$$+\frac{1-F(b)}{\mu} \int_b^\infty \left\{F(b) - F(b)F_2(y) + (1-F(b)) \left[F_2(y) - F_2(y)^2\right]\right\} dy$$

Now note that by Fubini's theorem, since income is all nonnegative,

$$\mu = E[Y] = \int_0^\infty (1 - F(y))dy$$

Also

$$\int_0^\infty (1-F(y))dy = \int_0^b (1-F(y))dy + \int_b^\infty (1-F(y))dy$$

$$= \int_0^b 1 - F(b)F_1(y)dy + \int_b^\infty 1 - [F(b) + (1-F(b))F_2(y)]dy$$

$$= \int_0^b 1 - F(b)F_1(y)dy + \int_b^\infty (1-F(b)) - (1-F(b))F_2(y)dy$$

$$= \int_0^b 1dy - F(b) \int_0^b F_1(y)dy + \int_b^\infty (1-F(b))(1-F_2(y))dy$$

Electronic copy available at: https://ssrn.com/abstract=2506219

$$= \int_0^b 1dy - F(b)\int_0^b F_1(y)dy + F(b)\int_0^b 1dy - F(b)\int_0^b 1dy + \int_b^\infty (1-F(b))(1-F_2(y))dy$$

$$= (1-F(b))b + F(b)\int_0^b (1-F_1(y))dy + \int_b^\infty (1-F(b))(1-F_2(y))dy$$

$$= (1-F(b))b + F(b)\int_0^b (1-F_1(y))dy + \int_b^\infty (1-F(b))(1-F_2(y))dy + (1-F(b))b - (1-F(b))b$$

So:

$$\mu = F(b)\mu_1 + (1-F(b))\mu_2$$

Now we are going to focus only on one part of $\mathcal{G}[F]$, that does not involve $F(y)^2$ terms:

$$\mathcal{G}[F] = \frac{F(b)}{\mu}\int_0^b \left\{F(b)[F_1(y) - F_1(y)^2] + F_1(y)[1-F(b)]\right\}dy$$

$$+\frac{1-F(b)}{\mu}\int_b^\infty \left\{F(b) - F(b)F_2(y) + (1-F(b))\left[F_2(y) - F_2(y)^2\right]\right\}dy$$

$$= \frac{F(b)}{\mu}\int_0^b \left\{F_1(y)[1-F(b)]\right\}dy + \frac{1-F(b)}{\mu}\int_b^\infty \left\{F(b) - F(b)F_2(y)\right\}dy$$

$$+\frac{F(b)}{\mu}\int_0^b \left\{F(b)[F_1(y) - F_1(y)^2]\right\}dy + \frac{1-F(b)}{\mu}\int_b^\infty \left\{(1-F(b))\left[F_2(y) - F_2(y)^2\right]\right\}dy$$

The first part (first line) is:

$$\frac{F(b)}{\mu}\int_0^b \left\{F_1(y)[1-F(b)]\right\}dy + \frac{1-F(b)}{\mu}\int_b^\infty \left\{F(b) - F(b)F_2(y)\right\}dy$$

$$= \frac{(1-F(b))F(b)}{\mu}[\int_0^b F_1(y)dy + \int_b^\infty \left\{1-F_2(y)\right\}dy]$$

Now remember that the the second distribution F_2 has support (b,∞), and is 0 below b, so that the mean income of the upper distribution can be computed as (these are conditional means):

$$\mu_2 = b + \int_b^\infty \left\{1-F_2(y)\right\}dy$$

and the average income of the first distribution is:

$$\mu_1 = \int_0^b (1-F_1(y))dy = \int_0^b dy - \int_0^b F_1(y)dy = b - \int_0^b F_1(y)dy$$

Electronic copy available at: https://ssrn.com/abstract=2506219

so

$$\int_0^b F_1(y)dy = b - \mu_1$$

$$\int_b^\infty \{1 - F_2(y)\}\, dy = \mu_2 - b$$

$$\frac{F(b)}{\mu}\int_0^b \{F_1(y)[1 - F(b)]\}\, dy + \frac{1 - F(b)}{\mu}\int_b^\infty \{F(b) - F(b)F_2(y)\}\, dy =$$

$$= (1 - F(b))F(b)[\frac{\mu_2 - \mu_1}{\mu}]$$

So we can write:

$$\mathcal{G}[F] = \frac{F(b)}{\mu}\int_0^b \left\{F(b)[F_1(y) - F_1(y)^2]\right\} dy + \frac{1 - F(b)}{\mu}\int_b^\infty \left\{(1 - F(b))\left[F_2(y) - F_2(y)^2\right]\right\} dy$$

$$+(1 - F(b))F(b)[\frac{\mu_2 - \mu_1}{\mu}]$$

Now we can look at the first part:

$$\frac{F(b)}{\mu}\int_0^b \left\{F(b)[F_1(y) - F_1(y)^2]\right\} dy + \frac{1 - F(b)}{\mu}\int_b^\infty \left\{(1 - F(b))\left[F_2(y) - F_2(y)^2\right]\right\} dy$$

$$= \frac{F(b)^2}{\mu}\frac{\mu_1}{\mu_1}\int_0^b \left\{[F_1(y) - F_1(y)^2]\right\} dy + \frac{[1 - F(b)]^2}{\mu}\frac{\mu_2}{\mu_2}\int_b^\infty \left\{\left[F_2(y) - F_2(y)^2\right]\right\} dy$$

$$= F(b)^2\frac{\mu_1}{\mu}\mathcal{G}[F_1] + [1 - F(b)]^2\frac{\mu_2}{\mu}\mathcal{G}[F_2]$$

So to conclude we have:

$$\mathcal{G}[F] = F(b)^2\frac{\mu_1}{\mu}\mathcal{G}[F_1] + [1 - F(b)]^2\frac{\mu_2}{\mu}\mathcal{G}[F_2] + (1 - F(b))F(b)[\frac{\mu_2 - \mu_1}{\mu}]$$

Electronic copy available at: https://ssrn.com/abstract=2506219