# ANNUALIZING LABOR MARKET, INEQUALITY, AND POVERTY INDICATORS

*Eduardo Lora, Miguel Benítez and Diego Gutiérrez*

CEQ INSTITUTE
COMMITMENT TO EQUITY
Tulane University

Working Paper 113
September 2021

# The CEQ Working Paper Series

The CEQ Institute at Tulane University works to reduce inequality and poverty through rigorous tax and benefit incidence analysis and active engagement with the policy community. The studies published in the CEQ Working Paper series are pre-publication versions of peer-reviewed or scholarly articles, book chapters, and reports produced by the Institute. The papers mainly include empirical studies based on the CEQ methodology and theoretical analysis of the impact of fiscal policy on poverty and inequality. The content of the papers published in this series is entirely the responsibility of the author or authors. Although all the results of empirical studies are reviewed according to the protocol of quality control established by the CEQ Institute, the papers are not subject to a formal arbitration process. Moreover, national and international agencies often update their data series, the information included here may be subject to change. For updates, the reader is referred to the CEQ Standard Indicators available online in the CEQ Institute's website www.commitmentoequity.org/datacenter. The CEQ Working Paper series is possible thanks to the generous support of the Bill & Melinda Gates Foundation. For more information, visit www.commitmentoequity.org.

The CEQ logo is a stylized graphical representation of a Lorenz curve for a fairly unequal distribution of income (the bottom part of the C, below the diagonal) and a concentration curve for a very progressive transfer (the top part of the C).

COMMITMENT
TO EQUITY

CEQ INSTITUTE
COMMITMENT TO EQUITY
Tulane University

# ANNUALIZING LABOR MARKET, INEQUALITY, AND POVERTY INDICATORS

*Eduardo Lora*[*], *Miguel Benítez*[†] *and Diego Gutiérrez*[‡]

CEQ Working Paper 113

SEPTEMBER 2021

## ABSTRACT

Widely, 12-month or 4-quarter average labor market, inequality and poverty indicators computed from repeated cross sections of household surveys are interpreted as annual. This is a valid interpretation only when several very specific criteria are met. Annual measures of indicators such as labor participation rates differ from their 12-month- or quarterly averages except when those who participate in a month or quarter also participate the other 11 months or three quarters. The same apply to unemployment rates and poverty rates. We propose several methods to accurately annualize sub-annual data. Some rely on ancillary questions often included in household surveys, others require econometric techniques such as predictive mean matching. Using data for Colombia we present annual measures of labor participation, occupation, unemployment, per capita labor income, average per capita household income, the Gini coefficients of labor income and per-capita household income, and moderate and extreme poverty rates.

**JEL Codes:** D31, J21 and J64

**Keywords**: Annualization, employment, income distribution, income poverty, Gini coefficient, labor income, labor participation, poverty, unemployment

[*] Universidad Eafit. 5203 Westbard Avenue, Bethesda MD, 20816 (for correspondence and/or off prints), eduardo.a.lora@gmail.com

[†] Inter-American Development Bank. 70 I St SE. Apt 1121, Washington DC, 20003, mbenitezr07@gmail.com

[‡] Inter-American Development Bank. 70 I St SE. Apt 1121, Washington DC, 20003, da.gutierrez95@gmail.com

This paper was prepared as part of the Commitment to Equity Institute's country-cases research program and benefitted from the generous support of the Bill & Melinda Gates Foundation. For more details, click here www.ceqinstitute.org.

# Annualizing labor market, inequality, and poverty indicators

**Eduardo Lora**[1] **Miguel Benítez**[2] **Diego Gutiérrez**[3]

[1] Universidad Eafit. 5203 Westbard Avenue, Bethesda MD, 20816 (for correspondence and/or off prints), eduardo.a.lora@gmail.com

[2] Inter-American Development Bank. 70 I St SE. Apt 1121, Washington DC, 20003, mbenitezr07@gmail.com

[3]Inter-American Development Bank. 70 I St SE. Apt 1121, Washington DC, 20003, da.gutierrez95@gmail.com

**Key Words**: annualization, employment, income distribution, income poverty, Gini coefficient, labor income, labor participation, poverty, unemployment

## Abstract

Widely, 12-month or 4-quarter average labor market, inequality and poverty indicators computed from repeated cross sections of household surveys are interpreted as annual. This is a valid interpretation only when several very specific criteria are met. Annual measures of indicators such as labor participation rates differ from their 12-month- or quarterly averages except when those who participate in a month or quarter also participate the other 11 months or three quarters. The same apply to unemployment rates and poverty rates. We propose several methods to accurately annualize sub-annual data. Some rely on ancillary questions often included in household surveys, others require econometric techniques such as predictive mean matching. Using data for Colombia we present annual measures of labor participation, occupation, unemployment, per capita labor income, average per capita household income, the Gini coefficients of labor income and per-capita household income, and moderate and extreme poverty rates.

## Abbreviations and acronyms

| | |
|---|---|
| DANE | Departamento Administrativo Nacional de Estadística (Colombia's official statistical office) |
| Fedesarrollo | Fundación para la Educación Superior y el Desarrollo |
| GEIH | Gran Encuesta Integrada de Hogares (Colombia's integrated household survey system) |
| i.i.d. | Independent and identically distributed random variable |
| ILO | International Labor Organization |
| PMM | Predictive mean matching |

## 1. Introduction

Monthly or quarterly household surveys are an important source of information to compute many yearly aggregates and indicators. For instance, monthly incomes reported by individuals in household surveys can be used to calculate total annual labor income in an economy. The calculation is trivial: monthly incomes reported by the individuals surveyed each month are expanded (using the sampling probabilities) and then all the (expanded) incomes are aggregated over the 12 months of the year (but it should be kept in mind that design features of survey data, such as sampling construction and weighting may affect the analysis of the data; Scheaffer et al, 2012). Twelve-month or quarterly averages of many indicators, such as labor participation, unemployment and poverty rates are computed regularly by statistical offices based on household surveys (ILO, 2019).

The incidence of fiscal policy variables, such as personal income taxes and transfers, on disposable incomes is also calculated from sub-annual (be it monthly, quarterly or any other frequency) data "annualized" by multiplying the sub-annual values by their sub-annual frequency (Lustig, 2018). If such calculations are interpreted as rolling averages of sub-annual data, there is nothing necessarily wrong with them. However, in many instances such interpretation is not warranted. For instance, if a measure of poverty severity is intended to shed light on the resources needed to bring the poor above the poverty line *in a year*, it is first necessary to adopt some criterion to define *yearly* poverty: is someone considered annually poor if he/she falls below the poverty line *on average* during the whole year, or *at least a month* in the year? The yearly resources needed to eradicate poverty with the first criterion will be fewer than those needed with the second criterion, where there will be more individuals classified as poor.

Furthermore, no such choice of criterion makes sense in some respects. If taxes obligations and returns are yearly based, as is the case in all countries, what sense can it make to compute the effect of taxes on disposable income as an average of monthly observations? Unless someone's income is the same every month of the year, it will be incorrect to calculate the tax owed by dividing by 12 the yearly minimum threshold (or each of the thresholds by tariff rate) that applies to a tax, compute the tax owed each month, and add up the monthly calculations.

The purpose of this paper is to propose simple methods to compute annual measures of labor participation, occupation, unemployment, labor income and its distribution, per capita household income and its distribution, and moderate- and extreme-income poverty from monthly (or, in general, sub-annual) data collected by household surveys that are structured as repeated cross-sections (not as panels). We do not deal with measures of consumption or multi-dimensional poverty because they are usually computed from panel or cross-sectional data that either take into consideration purchase frequency differences across items (foodstuffs vis-à-vis durables, for instance) or change little within a year (housing conditions, years of education, etc.).[4]

The annualization issues discussed in this paper have been largely overlooked in the labor and inequality literatures. However, in the poverty literature, several authors have pointed out that cross-sectional data fail to account for the temporal dimension of poverty when poverty is not a permanent phenomenon (Foster, 2007; Dang and Lanjouw, 2013;

[4] Household's consumption or expenditure measures are usually computed from information collected in surveys that enquire about the amount of consumption or expenditure of each consumption item over a "recall period", which varies across items (for instance, it may be a week for foodstuffs but a year for durables). Although the data collected do not come from a panel, the surveys are designed with the ultimate objective of obtaining an estimate of the consumption expenditure of each household *over the previous year* (Deaton and Grosh, 1998; Deaton, 2003).

Kafle et al, 2017; Jolliffe and Serajuddin, 2015; Bierbaum and Gassmann, 2012; Dang and Dabalen, 2019; Vakis, et al, 2016).

The duration of poverty cannot be inferred from its incidence –that is, the fraction of the population below the poverty line at the time of conducting a survey— a deficiency that is not solved by averaging repeated cross-sectional poverty rates over several periods, as is often done (Dang and Lanjouw, 2013). To address the issue, Jolliffe and Serajuddin (2015) re-estimated poverty rates in Jordan using the responses obtained through repeated visits to surveyed households (panel data). They found that the yearly-average estimation of poverty (called status quo approach) was 3.8 percentage points below the official estimate obtained with the cross-sectional approach. Similar results were found by Kafle et al (2017) for Ethiopia, where panel data from the first two waves of the Ethiopia Socioeconomic Survey were used to assess changes in poverty status based on consumption and asset ownership. As noted by the authors, longitudinal data at the household level provide additional information on the dynamics of wellbeing that could not be captured with cross-sectional data (Kafle et al, 2017).

To overcome the limitations of one-visit surveys, Dang and Lanjouw (2013) constructed synthetic panels with cross-sectional data and used them to estimate the household-level probabilities of falling below the poverty line in the unobserved periods. In their study, Dang and Lanjouw (2013) estimated chronic and transitory poverty rates in Bosnia-Herzegovina, Lao People's Democratic Republic, Peru, United Sates and Vietnam. Other studies have covered Kyrgyz Republic (Bierbaum and Gassmann, 2012), and several African (Dang and Dabalen, 2019) and Latin American countries (Vakis, et al, 2016). Some of these studies have compared the synthetic panel estimates with actual panel data and have found that the method provides accurate estimates of the actual data.

Although these studies deal with matters closely related to the annualization issue addressed by us, none of them, nor any other study to our knowledge, have challenged the standard interpretation of averages of monthly or quarterly (income) poverty rates as "annual". In this line, this paper contributes to the existing literature in at least two ways. First, it critically reviews the usual interpretation of key labor market, income, inequality, and poverty indicators that are computed from repeated cross-section data. Second, it proposes methods to adequately deal with the annualization issue.

The rest of this paper is organized as follows: section 2 discusses the relations between monthly and annual labor indicators and illustrates them with a Monte Carlo experiment and actual data for Colombia; section 3 makes use of the concepts and methods of the previous section and introduces additional methods for the calculation of annual income and annual income inequality at the individual level; section 4 introduces additional methods to extend the calculation of annual income and income inequality to the household level; section 5 discusses the application of the methods to the computation of annual poverty measures; section 6 summarizes the methods proposed and suggest additional avenues to facilitate the calculation of annual indicators and make use of the results.

## 2. Measuring annual labor indicators from monthly data

*Monthly and annual incidence rates*

The *incidence* rate of any phenomenon ($r$) in a period *t* is the ratio between the number of people who experience the phenomenon during that period ($R$) and the size of the group that can experience the phenomenon (N):

$$r_t = \frac{R_t}{N_t} \qquad (1)$$

Unemployment rates and labor participation rates, which are usually measured via monthly household surveys, are examples of *monthly* incidence rates. Subsequently, we will refer to any monthly incidence rate just as 'monthly rate' and we will keep the subindex $t$ only when referring to months (not to years). For unemployment rates, the relevant group is the labor force, and for participation rates the relevant group is usually the working age population.

If $N$ remains constant throughout a year, the 12-month average of a monthly rate is,

$$\bar{r} = \frac{\sum_{t=1}^{t=12} R_t}{12N} \qquad (2)$$

A 12-month average of a monthly rate should not be confused with the *annual incidence* rate ($i$) of the phenomenon. The latter is the ratio between the number of people who experience the phenomenon *at least* a month (A) and the size of the relevant group:

$$i = \frac{A}{N} \qquad (3)$$

Under what conditions is the 12-month average of a monthly rate equal to the annual incidence rate, $\bar{r} = i$? Only when $\sum_{t=1}^{t=12} R_t = 12A$, that is, when the total number of monthly cases of the phenomenon in a 12-month period is 12 times the number of people who experience the phenomenon. Obviously, this requires that anyone who experiences the phenomenon any given month experiences it also the other 11 months. In other words, the 12-month average of the monthly rate and the annual incidence rate (in the same 12 months) are equal only if the duration of the phenomenon is 100% of the 12 months. In general, a monthly rate is the product of the annual incidence rate and the proportion of the year *that those affected* experience the phenomenon –that is, the *duration rate d*,

$$r = id \qquad (4)$$

Duration $d$ is, therefore, the average of the duration $(d_j)$ for every individual $j$ within the population *that experiences the phenomenon* in the year $(d = \frac{1}{A}\sum_{j=1}^{A} d_j)$. The usual cross-section household surveys do not provide sufficient information to directly compute $d_j$ for every individual who belongs to $A$, but only for those who experience the phenomenon the same month of the survey, $\ddot{A}$. For instance, the survey may ask those who are currently unemployed how many months they have been unemployed. Although $A$ is not observed, it can be approximated by the following expression (see below why this is an approximation):

$$A \cong \sum_{j\in\ddot{A}} \frac{1}{d_j} \qquad (5)$$

In other words, every individual in $\ddot{A}$ represents $\frac{1}{d_j}$ individuals because $d_j$ is the probability that, being part of $A$, they are observed the month of the survey. Only in the extreme case where every $d_j$ is 1, is $A = \ddot{A}$. In general, average duration is:

$$d = \frac{1}{A}\sum_{j=1}^{A} d_j = \frac{\ddot{A}}{A} \cong \frac{\ddot{A}}{\sum_{j\in\ddot{A}}\frac{1}{d_j}} \qquad (6)$$

The relation between monthly rates, annual incidence rates and duration applies to any labor phenomenon, be it participating in the labor force, being employed or unemployed. When incidence and duration rates refer to *complete* episodes (not to 12 consecutive months, as we have assumed so far), it is also true in *steady state* that the monthly rate is the product of the incidence rate and the duration rate, as originally demonstrated by Kaitz (1970) for the case of unemployment (see also Sider, 1985).

If the phenomenon occurs randomly to anyone any given month, with probability *p*, regardless of their previous occurrence to them or to anyone else, the monthly rate is *r=p,* and the annual incidence rate is $i = 1 - (1-p)^{12}$. It follows from equation (4) that, under the assumption of complete randomness:

$$d = \frac{r}{i} = \frac{p}{1-(1-p)^{12}} \qquad (7)$$

The duration rate, under the assumption, is very close to the monthly rate, except for small values of *p.* For instance, when *p* is 0.1, *d* is 0.13936, but when *p* is 0.3, *d* is almost the same (0.3042). This implies that the annual incidence rate approaches 100% very fast: while it is 71.76% when *p* is 0.1, it reaches 98.61% when *p* is 0.3. This explains why equations (5) and (6) are approximations.

*Monthly and annual unemployment indicators in a population and a sample: A Monte Carlo example*

In a monthly household survey, under the assumption of complete randomness of a phenomenon, the probability that, in any given month, an individual is observed experiencing the phenomenon (for instance, being unemployed) is the same rate of the phenomenon. The assumption means that the probability that the individual appears in the sample is equal and independent of that of any other individuals (i.i.d.), and it is also independent of whether she is experiencing the phenomenon. It follows from this that, if a household survey provides information on the monthly rate (for instance, the unemployment rate), we can obtain the incidence and the duration rates.

A Monte Carlo experiment can demonstrate that this is the case and that, as stated above, the monthly and the annual incidence rates are related through the duration of the phenomenon. We undertake several Monte Carlo simulations (with 100 repetitions) on unemployment rates. We define a population of 100,000 individuals who participate in the labor market, and who can be either employed or unemployed. Each month, individuals are unemployed with an exogenous probability *p*. Then, we draw monthly random samples equivalent to 1% of the labor force and proceed to calculate both the 12-month average of the monthly unemployment rate and the annual incidence rate, based on the probability of unemployment for each individual. Finally, we compute the duration rate in two ways, as given by equations (6) and (7). For the former, we use the sample information of the individuals unemployed in month 12th, along with their corresponding numbers of months of unemployment in the whole 12 months of the simulation (as if the unemployed had responded to the question "how many months in the last 12 months have you been unemployed?").

The results –which are presented in Table 1— show that the estimation via sampling replicates the parameters of the whole population under different values of the probability of unemployment, *p* (which, by definition, is the same unemployment rate of the whole population). For example, if we assume *p=0.10*, the method estimates an incidence rate of 71,4%, which is close to the population value (71,6%), meaning that, for this hypothetical population, about 71% of the population would be classified as 'unemployed' at least one month of the year. This confirms that the annual incidence rate is not the same as the 12-month average of the monthly rate and, by extension, that the number of people who experience unemployment in a year is not the same as the 12-month average of those who experience unemployment in a month. The simulation confirms that duration *d* is correctly estimated via either equation (6) or equation (7), the latter being more precise for low values of *p*, as it uses the samples of the 12 months, instead of an approximation.

**<Table 1 about here>**

Therefore, we verify that, from the monthly rates of a household survey, the annual incidence rates and the duration of the phenomenon can be obtained accurately under the assumption of complete randomness of the phenomenon. We also verify that the 12-month average of the rates does not correspond to the annual incidence rate.

*Calculating annual indicators in practice: (1) employment*

However, a labor phenomenon, such as employment, is not completely random, not only because it is experienced with higher frequency by some groups (men, educated, prime age individuals, etc.) but because it is more likely to be experienced by those who are already employed or surrounded by other employed people. Therefore, (annual) the incidence and the duration of the phenomenon cannot be inferred from the monthly rates. But, as suggested above, incidence and duration can be estimated based on ancillary questions often included in household surveys which provide direct information on duration or incidence for at least some groups.

In this sub-section we use such direct monthly survey information to calculate the absolute number and the annual incidence and duration of employment and labor participation. We use data from the main Household Survey in Colombia (*Gran Encuesta Integrada de Hogares*, GEIH), an official monthly survey that has been collected by the *Departamento Administrativo Nacional de Estadística* (DANE) since 2006. It provides labor and sociodemographic information from households in 443 Colombian municipalities (40% of total), and it is representative of the universe of households in the whole country, in each of the 23 cities that are capitals of the main departments, and in each of the 31 departments. The GEIH is stratified by geographical area, population size, level of urbanization and level of unsatisfied basic needs by municipality. Since sampling is done by households, "expansion factors", which are the same for all the individuals that belong to the same household. All our calculations below are done with the "expanded" values, using such expansion factors.

Four questions in the survey can be used to identify the number of individuals that have been occupied at least an hour in the last 12 months (see Table 2). In December 2019, approximately 18.9 million individuals declared to have worked all past 12 months, while 3.9 million declared to have worked in that month, but not all the previous 11 months (which means that a total of 22.8 million were occupied in December 2019). An additional 1.5 million, who were inactive in the month, declared to have worked for the last time less than a year ago, and 1.7 million, currently unemployed, declared to have worked for the last time less than 53 weeks ago. Therefore, a total of 25.9 million individuals worked in 2019, according with the information provided by the December 2019 survey. This implies that the average duration of occupation in 2019 was 87.7%, or 10.5 months (ignoring seasonality factors). Replicating this calculation with the whole monthly series of surveys produces the results presented in Figure 1. Each of the series is presented in two ways: monthly and as a 12-month rolling average; the latter roughly isolates seasonality factors. It must be stressed that the 12-month averages of the *monthly* occupied do not correspond to the *annually* occupied (that is, those who worked *some* of the last 12 months), as is often assumed. The ratio between the whole series of the monthly and the annually occupied is 86.4%, which is the average duration of occupation

(10.3 months). This calculation is consistent with equation (4) above (after multiplying both sides of equation (4) by $N$, it becomes $R = Ad$, and therefore $d = R/A$).

**<Figure 2 about here>**

**<Table 2 about here>**

The number of workers who declare to have worked in the last month but not all the previous 11 months can be used to calculate the occupation duration of that group. With the data for December 2019 in Table 2, a duration of 54.9%, or 6.6 months is obtained (3.9/(3.9+1.5+1.7)), which ignores seasonality factors. With the whole series in Figure 1, the duration of those not fully occupied is 55.8%, or 6.7 months. The duration of those not fully occupied could be also approximated using the information provided by those employed in the current month on the numbers of months that they worked in the past 12 months (equation 6). However, this method produces a very different value of the duration of occupation: 8.3 months in the December 2019 survey or 8.4 in the whole series. We hypothesize that this discrepancy results from the fact that the interviewees associate the expression 'number of months' with the number of full-time work months, and not with the number of *different* months when they worked at least one hour. On this assumption, we correct the number of months reported by the average hours worked per week by each person. Namely, we divide the number of months reported in the survey ($\ddot{m}_j$) by $(h_j/\bar{h})$, where $h_j$ is the number of hours worked by individual $j$ in the last week, and $\bar{h}$ is the average hours worked by all employees in the last week, including extra-hours and second jobs. The intuition behind this adjustment is that workers who report having worked a certain number of full months but usually work fewer hours per week than the average (including overtime and hours in second jobs), must have worked more non-full months. The results, which are presented in Figure 2 for the smoothed series, show that this method closely replicates the number of annual employed workers calculated above.

**<Figure 2 about here>**

### *Calculating annual indicators in practice: (2) labor participation*

Labor participants in a month are those who, in the reference period of the survey, *either* worked for at least an hour ("employed") *or* had not worked but were able to work and actively searched for a job ("unemployed"). As we have just seen, two alternative methods based on monthly data can be used to calculate annual occupation. A similar method could in principle be applied to estimate the annual unemployed, but the GEIH does not collect information on the number of months that the unemployed looked for a job in the previous 12 months, or about the number of months that labor participants were inactive. Although GEIH information does not allow to estimate the number of annually unemployed (and, therefore, the annual incidence of unemployment), it does allow to calculate the amount of annual labor participants. These can be identified as those who (i) worked or actively search for jobs in the current month *and/or* (ii) declare to have worked or actively searched for jobs in the last 12 months. Naturally, annual participants will be more than the 12-month average of monthly participants, unless every participant participated the 12 months (recall equation 4).

**<Table 3 about here>**

Table 3 presents the numbers to make the calculation with one monthly survey (December 2019). Participants include those who worked that month (22.8 million), or were unemployed (2.4 million), or were inactive that month and have either worked or searched for a job in the past 12 months (1.8 million). Therefore, based on the December 2019 survey, annual labor participants in 2019 were 26.9 million, while monthly labor participants were 25.2 million (again, ignoring seasonality factors). Since working age population was 39.6 million, these numbers imply that the (monthly) participation rate was 63.6% and the incidence of participation was 67.9%. From equation (4), it follows that the duration of participation was 93.5% or 11.2 months. In the left panel of Figure 3, we present the numbers of monthly and annual participants for the whole series since December 2007, and in the right panel we show the corresponding rates of (monthly) participation and participation incidence. We cannot split the universe of participants between all-time participants and occasional participants to calculate the duration of participation of the latter (as we did above for those occupied some of, or all, the previous 12 months). As in Figure 1, each of the series in Figure 3 is presented in two ways: monthly and as a 12-month rolling average, which roughly isolates seasonality factors. Again, the figure makes clear that the 12-month averages of the *monthly* data do not correspond to the *annual* series, as is often assumed.

**<Figure 3 about here>**

## 3. Measuring annual income and annual income inequality from monthly data

*Annual incomes by individual*

*Total* annual labor income of *all* the occupied at least some month(s) of a year can correctly be estimated as the sum of the monthly labor incomes observed in the 12 monthly surveys of the year (as is usually done by official statistical offices, such as DANE in Colombia, after "expanding" each observation in the sample by its corresponding statistical weight, or "expansion factor"). Annual labor income per worker $\overline{Y^L}$ is, of course:

$$\overline{Y^L} = \frac{1}{A}\left(\sum_{j,t=1}^{t=12} y_{j,t}^L\right) \qquad (8)$$

Where A is the number of *annually occupied* and $y_{j,t}^{L}$ is labor income of individual $j$ in month $t$. As discussed in the previous section, the number of annually occupied (A), which is not observed, can be estimated as $R/d$, where $R$ is the number of monthly occupied, which is observed, and $d$ is the average duration of occupation of the annually occupied, which can be estimated via several methods. Therefore:

$$\overline{Y^L} = \frac{d}{R}\left(\sum_{j,t=1}^{t=12} y_{j,t}^{L}\right) \qquad (9)$$

What does this imply for the computation of individual income distribution measures, such as the Gini coefficient? If some workers are occupied less than the 12 months of the year, the Gini of annual incomes is not the same as the average of the Gini coefficients of the monthly incomes. To compute the former, we need to estimate the annual income of every annually occupied worker, not only of those workers that are observed in the month(s) of the survey. Let's assume that *monthly* labor income by individual does not change from month to month,[5] but the individual may work some months and not others. Therefore, the individual's annual income $Y_j$ is:

$$Y_j = \ddot{y}_j^L m_j \qquad (10)$$

Where $\ddot{y}_j^L$ is monthly income when she works and $m_j$ is the number of *different* months (not necessarily full-time months) she works. As we have seen, the monthly household surveys by DANE include a question on the number of months that must be adjusted to reflect the actual number of *different* months the individual worked in the last 12 months:

[5] We are aware that this assumption does not reflect the reality of non-salaried workers, but we have no reason to assume that income received by an individual in a given month differs systematically from its expected value over the year, apart from seasonality factors, which can be ignored in 12-month rolling averages.

$$Y_j = \ddot{y}_j^L \ddot{m}_j \bar{h}/h_j \qquad (11)$$

With this method $Y_j$ can be estimated for all the individuals who declare to have worked the month of the survey. Those who did not work the month of the survey but did work at least one month of the previous 11 months are not observed but must be included to have the full distribution of individual incomes. This can be done by expanding the weight of each of the individuals who did work the month of the survey by the inverse of their occupation duration, which corresponds approximately, as we saw in the previous section, to $m_j/12$, where $m_j$ must be approximated by $\ddot{m}_j \bar{h}/h_j$ for those who declare to have worked less than 12 months in the last year.

The black lines of Figure 4 show average annual income per worker ($\bar{Y}^L$) calculated from the estimates of $Y_j$, weighting each observation as just explained. The light grey lines of the figure show the same variable but only for those who worked every month of the last year ("fully occupied"). Naturally, those who work 12 months per year make higher annual incomes than the average of all workers, because the latter includes those who work fewer months. Similarly, average annual income per worker is lower than 12 times the average monthly income reported by those who worked the month of the survey – the darker grey line, "monthly basis"—because the former includes those who worked some other months, but not the month of the survey.

**<Figure 4 about here>**

**<Figure 5 about here>**

Using the same black and grey palette, Figure 5 shows the Gini coefficients calculated from the weighted estimates of $\boldsymbol{Y_j}$ and from the monthly data that comes directly from the surveys. The former ones are higher because they include those workers who were not observed the month of the survey, who make lower annual salaries than those who work every month of the year. A Gini coefficient computed directly from a monthly survey refers, by definition, to the distribution of monthly labor incomes among those who worked that month. A 12-month rolling average of those coefficients is still a measure of the distribution of monthly, not annual, incomes. Although both are conceptually correct, for public policy or welfare analysis purposes, the annual-income based Gini is preferable to the monthly-income based one (under the assumption that monthly incomes are fungible within a year).

## 4. Measuring annual income and annual income inequality of *households* from monthly data

### *Annual incomes by family: the method*

Estimating annual indicators at the household level poses additional challenges because, although any household may have a similar probability of being observed in a monthly survey, within each household any individual who does not stay in the same labor status the 12 months of the year will be observed in only one of the possible statuses she may have. This sub-section proposes a method to annualize indicators at the household level, using a machine-learning algorithm. Hereunder we present the problem and the proposed method for the specific cases of household income and annual poverty.

The most common method for computing household income –which is a necessary step to calculate income poverty rates—assumes that 12 times the sum of the incomes earned by all the members *i* of household *j* ($\sum_{i \in j} y_{ijt}$) is the household's annual income. However, this is not necessarily the case if the household´s monthly income is not the same every month. Therefore, annual income ($Y_j$) of household *j* is defined as:

$$Y_j = \sum_{t=1}^{12} \sum_{i \in j} y_{ijt} \qquad (12)$$

If each household member earns the same amount every month, she has some income (and zero otherwise):

$$Y_j = \sum_{i \in j} y_{ijt}\, m_{ij} \qquad (13)$$

Where $m_{ij}$ represents the number of months in which individual *i* earned income. As in the individual-level case above, this calculation must include income earned by individuals that did not receive income the month of the survey but did receive income some other month(s) of the year. In monthly cross-section surveys, such as DANE's, information on $y_{ijt}$ and $m_{ij}$ is not collected for every individual within the household. For example, in a given month, certain household members may have received no income due to temporary episodes of inactivity or unemployment although they may have had income other months of the year. This has profound implications on measuring annual income poverty, since a monthly-poor household is not necessarily poor in the entire year, as discussed in the following section. Hence, calculating annual income indicators at the household level requires the estimation of $y_{ijt}\, m_{ij}$ for all household members.

To do this, we use the *predictive mean matching* method, which is a standard multiple imputation methodology that allows us to impute the possible annual earnings (that is,

$y_{ijt}$ $m_{ij}$) that are unknown in a monthly survey (that samples households correctly). This is a partially parametric method that imputes the missing values of the variable of interest using the observed values of a set of candidates or "possible donors". This process relies on the closest predictive mean by combining the standard linear regression and the nearest-neighbor imputation approaches (Rubin, 1986; Little, 1988).

Without loss of generality, the procedure can be resumed in three steps. First, the method uses a normal linear regression (where each observation is an individual) to obtain linear predictions of the variable of interest from a set of regressors, which in our case will be variables at the individual level (such as age, sex, and education) and at the household level (such as family composition, assets, location, etc.; see the complete list of variables in the Appendix). Second, the predictions obtained in the previous step are used to construct a metric distance to form the set of possible donors by using the nearest neighbor approach. It is important to clarify that the linear regressions are not used to impute the missing values; their only purpose is to construct a metric to identify the most similar observations. Finally, the algorithm randomly chooses one (or more) of the nearest observations to impute values to the target variable of the individual. It is important to note that observed values of the target variable are not altered by the procedure.

One important attribute of this methodology is that it preserves the distribution of the observed values in the missing part of the data, which makes it more robust than the fully parametric linear regression approach. However, it is important to decide how many nearest neighbors may be included with the purpose of creating the set of possible donors. We ran the algorithm using between one and five nearest neighbors and based on the correlations with the original target variable we chose the simple average of the first five k-nearest neighbors.

## 5. Estimating annual income poverty

The decomposition of the (monthly) rate of a phenomenon between its (annual) incidence and duration applies to any labor phenomenon, be it participating in the labor force, being employed or unemployed. It also applies to poverty. Let's define as *monthly income poor* someone whose income in the month is below the monthly poverty line. Then, the monthly poverty rate is the share of the population who experience poverty; the (annual) incidence of poverty is the share of the population that experiences (monthly) poverty at least a month of the year and the (annual) duration of poverty is the fraction of the year that those that experience (monthly) poverty are poor. Therefore, if we stick to the *monthly* definition of income poverty, equation (4) remains valid.

Let's now define as *annually poor* someone whose income over the year is below 12 times the monthly poverty line. Under what conditions would the annual poverty rate be the same as the 12-month average of the monthly poverty rates? In other words, what assumptions must hold to interpret the annual average of the monthly poverty rates as equivalent to the annual poverty rate, as is often implied?

Let $q_{jt}$ be a dichotomous variable that classifies a household $j$ as poor if its per capita income ($y_{jt}$) in month $t$ is below the monthly poverty line ($L$):

$$q_{jt}\begin{cases} 1 & if\ \ y_{jt} < L \\ 0 & otherwise \end{cases} \qquad (14)$$

From the above, we can define the monthly poverty rate ($\varphi_t$) as:

$$\varphi_t = \frac{1}{N}\sum_j q_{jt}k_j \qquad (15)$$

Where $N$ is total population and $k_j$ is the number of persons in household *j.* Consequently, the *average monthly* poverty rate in the 12 months of the year ($\bar{\varphi}$) can be expressed as follows:

$$\bar{\varphi} = \frac{1}{12N}\sum_i \sum_{t=1}^{12} q_{it}k_j \qquad (16)$$

On the other hand, let's *annual* poverty $Q_j$ be a dichotomous variable that takes value 1 when the *annual* per capita income of household $j$ is less than 12 times the monthly poverty line:

$$Q_j \begin{cases} 1 \quad if \sum_{t=1}^{12} y_{jt} < 12L \\ 0 \quad otherwise \end{cases} \qquad (17)$$

We can rewrite $Q_j$ in terms of $q_{jt}$. $Q_j$ equals 1 when the income obtained in the "good" months (when $q_{jt}$ is 0) is not enough to compensate the lack of income during the "bad" months (when $q_{jt}$ is 1). Let's further define monthly poverty *severity* ($s_j$) as the difference between the monthly income per capita of the household when it is poor and the poverty line; and the monthly *income buoyance* ($b_j$) as the difference between the monthly income per capita of the household when it is not poor and the monthly poverty line (note that both $s_j$ and $b_j$ can only take positive values). The household is annually poor when condition (9) below is met, that is, when the product between the average severity of monthly poverty and the number of months that the household is monthly poor is higher than the product of the income buoyancy and the number of months that the household is not poor:

$$s_j \sum_{t=1}^{12} q_{jt} > b_j \left( 12 - \sum_{t=1}^{12} q_{jt} \right) \qquad (18)$$

Reordering, we obtain:

$$\frac{s_j + b_j}{12 b_j} \sum_{t=1}^{12} q_{jt} > 1 \qquad (19)$$

Therefore, equation (8) can be rewritten as:

$$Q_j \begin{cases} 1 \; if \;\; \frac{s_j + b_j}{12 b_j} \sum_{t=1}^{12} q_{jt} > 1 \\ 0 \;\; otherwise \end{cases} \qquad (20)$$

$$\phi = \frac{1}{N} \sum_j Q_j k_j \qquad (21)$$

Given this, the annual poverty rate ($\phi$) and the 12-month average poverty rate ($\bar{\varphi}$) are two different expressions, which only coincide in specific combinations of the parameters of equation (18). If households that are poor any given month are also poor the other 11 months of the year, the two measures are of course the same. Otherwise, annual poverty will depend, not just on the monthly poverty rates, but on the poverty severity and the income buoyancy of those households that are poor at least one month of the year.

Using the *predictive mean matching* results by individual summarized in a previous section, we can compute the annual poverty rate, as well as a set of other poverty indicators, which are presented in Table 4 for Colombia in 2019. While poverty estimated with the 12-month average supposedly affects 35.6% of the population in 2019, the annual estimation using the PMM method indicates that it affects 32.5% of the population. The difference is about 1,5 million individuals. Similarly, while the 12-month average indicates that the incidence of extreme poverty was 10.2% in 2019, the adjusted method indicates that it was 8,1% (1 million fewer individuals).

**<Table 4 about here>**

Also, poverty severity is lower (38% of poverty line) than the 12-month average (40.5%). In addition, the Gini index of household´s income per capita is 53.8% with the annual estimation, which is lower than the 12-month average (54. 7%).

## 6. Concluding remarks

The temporal dimension matters to measure the incidence rate of many phenomena. The snapshot of unemployment or income poverty in one month is different from the incidence of these phenomena in the whole year. Averaging the monthly cross sections of a phenomenon over a year does not provide a correct measure of its annual incidence, unless the people who experience it in a month experience it every month of the year. This study has proposed different ways of estimating the annual version of labor market, income, inequality, and poverty indicators based on monthly (or, in general, sub-annual) repeated cross-sections. The methods proposed were applied to household survey data collected by DANE, the Colombian statistical office. Some of the methods are adequate to estimate individual level indicators, while others are suitable to household level indicators.

We have proposed two methods to estimate annual indicators at the *individual* level. In the first method we use questions from different household survey modules to detect individuals who report having experienced a certain phenomenon in some month of the year. This is the case, for example, of the participation rate, which goes up (in 2019) from 63% when calculated as a monthly average to 68% when estimated based on the additional information interviewees provide on their labor activities during the last 12 months. In the second method, which is useful when the surveys do not contain questions that allow detecting the annual incidence of the phenomenon, we estimate the indicators based on the probability of observing individuals in a certain state of the phenomenon. This is the case, for example, of labor incomes per person, which, for all those who work less than 12 months, are lower when estimated than when calculated as 12 times their monthly values. Thus, per-capita labor income is lower, and the Gini coefficient of individual labor incomes is higher, when estimated as explained than when directly calculated from the monthly data as is usually done.

To estimate annual *household* level indicators –such as income poverty—, we used the predictive mean matching method to impute the possible annual earnings of individuals who are not observed working the month of the interview, but who report having worked previous months of the year. If the annual poverty line is defined as 12 times the monthly poverty line, the method implemented shows that annual poverty in Colombia in 2019 was 32.5%, instead of 35.6% as officially reported based on the monthly poverty line and the monthly household incomes.

Our analyses suggest that monthly household surveys could facilitate the computation of annual indicators if some additional questions were included. For instance, Colombian household surveys do not include adequate questions to calculate the annual incidence of unemployment, a deficiency that could be easily solved asking all working age individuals (not just those currently unemployed) about job searches in the past twelve months. Similarly, some annual aggregates at the family level would benefit from adding questions to all family members in working age about their labor income in the last 12 months.

Adequate measurement of annual income variables at the individual and family level is necessary not only to produce annual income distribution and income poverty measures, but also to improve fiscal incidence analyses. For instance, the incidence of value-added taxes across income groups is distorted by the fact that some families do not report income in the month of the survey –which puts them in the lowest income quantile— while reporting consumption levels that are typical of higher income quantiles. One of the co-authors of this paper has applied the methods of annual income estimation to compute the incidence of VAT in Colombia, finding a substantial difference as a result (Gutiérrez and Mejía, 2021). Similarly, the incidence of personal income taxes is distorted when computed from observed monthly incomes rather than from estimated annual incomes (Lora, 2021). Finally, the incidence of subsidies is not the same when computed on monthly, rather than annual family incomes, as shown in an application to Colombia (Benítez and Mejía, 2021).

## 7. Acknowledgments

This article was prepared for the research project “Reforms for a Post-Covid Colombia: Towards a New Social Contract”, led by Eduardo Lora and Luis Fernando Mejía at Fedesarrollo and funded by an anonymous donor. The authors acknowledge useful comments from Roberto Angulo, Juan Camilo Chaparro, Jairo Núñez, Luis Fernando Mejía, participants in the seminars convened by Fedesarrollo and Banco de la República, and an anonymous reviewer kindly appointed by the Commitment to Equity Institute.

## 8. References

Benítez, M. and Mejía, L.F.: Reforma a la política social. In Lora, E. and Mejía, L.F. (eds.). *Reformas para una Colombia post-covid-19: hacia un nuevo contrato social*. Fedesarrollo, Bogotá (2021).

Bierbaum, M. and Gassmann, F.: Chronic and transitory poverty in the Kyrgyz Republic: What can synthetic panels tell us? *UNU-MERIT Working Papers* (2012).

Dang, H. A. and Lanjouw, P.: Measuring poverty dynamics with synthetic panels based on cross-sections. *World Bank Policy Research Working Paper* 6504 (2013).

Dang, H. A. H. and Dabalen, A. L.: Is poverty in Africa mostly chronic or transient? Evidence from synthetic panel data. *The Journal of Development Studies,* 55(7), 1527-1547 (2019).

Deaton, A.: "Household surveys, consumption, and the measurement of poverty." *Economic Systems Research* 15, no. 2: 135-159 (2003)

Deaton, A. and M. Grosh.: *Designing household survey questionnaires for developing countries lessons from ten years of LSMS experience, chapter 17: Consumption*. No. 218 (1998).

Foster, J. E.: A class of chronic poverty measures. *Poverty dynamics: interdisciplinary perspectives*, 59-76 (2009).

Gutiérrez, D. and Mejía, L.F:. Hacia un IVA más eficiente y equitativo. In: Lora, E. and Mejía, L.F. (eds.) *Reformas para una Colombia post-covid-19: hacia un nuevo contrato social.* Fedesarrollo, Bogotá (2021).

ILO (International Labour Organization): Quick Guide on Interpreting the Unemployment Rate. Geneva (2019).

Jolliffe, D. and Serajuddin, U.: Estimating poverty with panel data, comparably: an example from Jordan. *World Bank Policy Research Working Paper* 7373 (2015).

Kafle, K., McGee, K., Ambel, A. and Seff, I.: Once poor always poor? Exploring consumption- and asset-based poverty dynamics in Ethiopia. Ethiopian *Journal of Economics*, 25(2), 37-76 (2016).

Little, R. J. A.: Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 6: 287–296 (1988).

Lora, E.: Reformas al Impuesto de Renta de las Personas. In: Lora, E. and Mejía, L.F. (eds.) Reformas para una Colombia post-covid-19: hacia un nuevo contrato social. Fedesarrollo, Bogotá (2021).

Lustig, N.: Measuring the distribution of household income, consumption and wealth. *For Good Measure: Advancing Research on Well-being Metrics Beyond GDP* (2018).

Rubin, D. B.: Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business and Economic Statistics 4: 87–94 (1986).

Scheaffer, R. L., W. Mendenhall, III, R. L. Ott, and K. G. Gerow: Elementary Survey Sampling. 7th ed. Boston: Brooks/Cole (2012).

Vakis, R., Rigolini, J. and Lucchetti, L.: Left behind: chronic poverty in Latin America and the Caribbean. *World Bank Publications* (2016).

Table 1. Monte Carlo simulations for annual unemployment rate

| Parameters observed in the universe of 100.000 observations | | | Parameters estimated via monthly sampling (12-month averages, except duration from equation (6), which comes from month 12th) | | | |
|---|---|---|---|---|---|---|
| Unemployment rate (r=p) | Incidence rate (i) | Duration (d) | Unemployment rate (r) | Incidence rate (i) | Duration (d) | |
| | | | | | Equation (6) | Equation (7) |
| 5,0% | 45,8% | 10,9% | 4,8% | 42,6% | 10,7% | 10,9% |
| 10,0% | 71,6% | 13,9% | 10,3% | 71,4% | 13,8% | 14,0% |
| 20,0% | 93,1% | 21,5% | 20,3% | 93,6% | 21,4% | 21,5% |
| \|30,0% | 98,5% | 30,5% | 30,2% | 97,4% | 30,5% | 30,5% |
| 40,0% | 99,8% | 40,1% | 39,8% | 99,2% | 40,0% | 40,1% |
| 50,0% | 100,0% | 50,0% | 49,8% | 100,0% | 50,0% | 50,0% |

Source: Authors´ calculations.

Table 2. Identifying the annual occupied with monthly household survey data

(Colombia, December 2019)

| Classification | Question | Criteria | Number of individuals |
|---|---|---|---|
| Fully occupied | | Worked all past 12 months | 18.863.272 |
| Non-fully occupied | | Did not work all past 12 months, but worked last month (either "worked at least an hour in the last week", or "report having a work" in the past month) | 3.897.458 |
| Inactive but previously occupied | How long did you work for the last time? (P7440) | Less than a year ago | 1.504.447 |
| Unemployed but previously occupied | How many weeks have passed since you worked for the last time? (P7320) | Less than 53 weeks | 1.682.915 |
| Total occupied in the year (annual occupied) | | | 25.948.092 |

Source: Authors´ calculations based on GEIH-DANE

Table 3. Identifying annual labor participants (Colombia, December 2019)

| Classification | Question | Criteria | Number of individuals |
|---|---|---|---|
| Employed | | Either “worked at least an hour in the last week”, or “report having a work” | 22.760.730 |
| Unemployed | | Actively searched for job | 2.398.074 |
| Inactive | "How many weeks have passed since you worked for the last time?" (P7320) and "How long did you searched for the last time?" (P7456) | Worked or searched for job in the last 12 months | 1.750.495 |
| Total | | | 26.909.299 |

Source: Authors´ calculations based on GEIH-DANE

Table 4. Poverty and inequality measures (Colombia, 2019)

| Indicator | Measurement | 12-months average | Annual estimation |
|---|---|---|---|
| Poverty | Incidence (%) | 35,62 | 32,54 |
| | Severity (% of poverty line) | 40,5 | 38,0 |
| | Cost of removing (% GDP) | 2,63 | 2,26 |
| Extreme poverty | Incidence (%) | 10,21 | 8,10 |
| | Severity (% of extreme poverty line) | 41,2 | 39,4 |
| | Cost of removing (% GDP) | 0,32 | 0,24 |
| Inequality | Gini index | 54,66 | 53,81 |

Source: Author´s calculations.

Appendix 1: Variables used in the *predictive mean matching* method

| Predictor variables for income and months worked |
|---|
| Sex |
| Age |
| Years of schooling |
| Household size |
| Zone (urban, rural) |
| Overcrowding (person per room) |
| Access to home services (sewerage, garbage collection, aqueduct, energy, internet) |
| Housing stratification |
| Housing material |
| Possession of domestic appliances (blender, fridge, stove, microwave, heater, TV, DVD, sound equipment, computer, vacuum cleaner, air conditioning, cooling fan) |
| Vehicles (bike, motorcycle, car) |
| Land ownership |
| Presence of children under 12 years |
| City (23 main cities) |

**Figure captions (to be placed below corresponding figure)**

Figure 1. Employment (Colombia, 2008-2019)

The “monthly occupied” series (and its 12-month rolling average) come directly from the surveys (GEIH). Only some of the monthly occupied are “occupied all the last 12 months”. For this reason, the “annually occupied” some of the last 12 months are more than the “monthly occupied”. Here the number of “annually occupied” is calculated based on several ancillary questions in the surveys (first method).
Source: Author´s calculations based on GEIH, DANE.

Figure 2. Monthly and annually occupied (12-month rolling averages, Colombia, 2008-2019)

Here, in addition to the 12-month rolling averages of the three series in Figure 1, we present an alternative computation of the “annually occupied (some of the last 12 months, second method).” The method uses the (adjusted) number on months worked in the last 12 months declared by the interviewees. The first and the second method produce very similar estimates of the number of annually occupied.
Source: Author´s calculations based on GEIH, DANE.

Figure 3. Labor participation (Colombia, 2008-2019)

a. Millions of individuals

b. Rates (of working age population)

Panel a. shows the number of “monthly participants” (and its 12-month rolling average), which comes directly from the surveys, and the number of “annual participants”, which is estimated based on ancillary questions in the survey. Panel b. presents the participation rates corresponding to the two measures.
Source: Author´s calculations based on GEIH, DANE.

Figure 4. Estimation of nominal annual labor income (pesos, Colombia, 2009-2019)

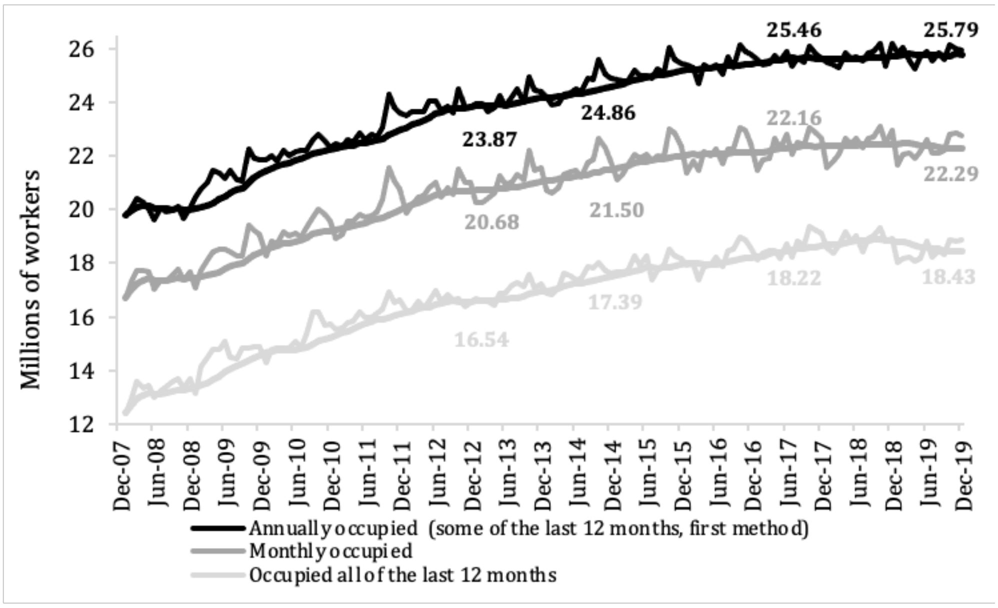
The annual labor income of those “fully occupied” is calculated as 12 times the monthly income declared by the interviewees who said they worked the last 12 months. The “monthly basis” calculation is done the same way, also including those who said they worked less than 12 months. To correctly calculate the actual “annual” labor income of everybody, the latter is adjusted by the actual number of months worked by each worker.
Source: Author´s calculations based on GEIH, DANE.

Figure 5. Gini index of labor income (Colombia, 2009-2019)

The “monthly” Gini is computed directly from the monthly incomes reported by those who worked the month of the survey. The “annual” Gini is computed from the annual incomes estimated for every person who worked at least a month in the last 12 months. Labor income concentration is higher in the latter case because those who did not work all the last 12 months have lower *annual* incomes than those who worked every month. Source: Author´s calculations based on GEIH, DANE.

## Figures

Figure 1. Employment (Colombia, 2008-2019)



.

Figure 2. Monthly and annually occupied (Colombia, 12-month rolling averages, 2008-2019)

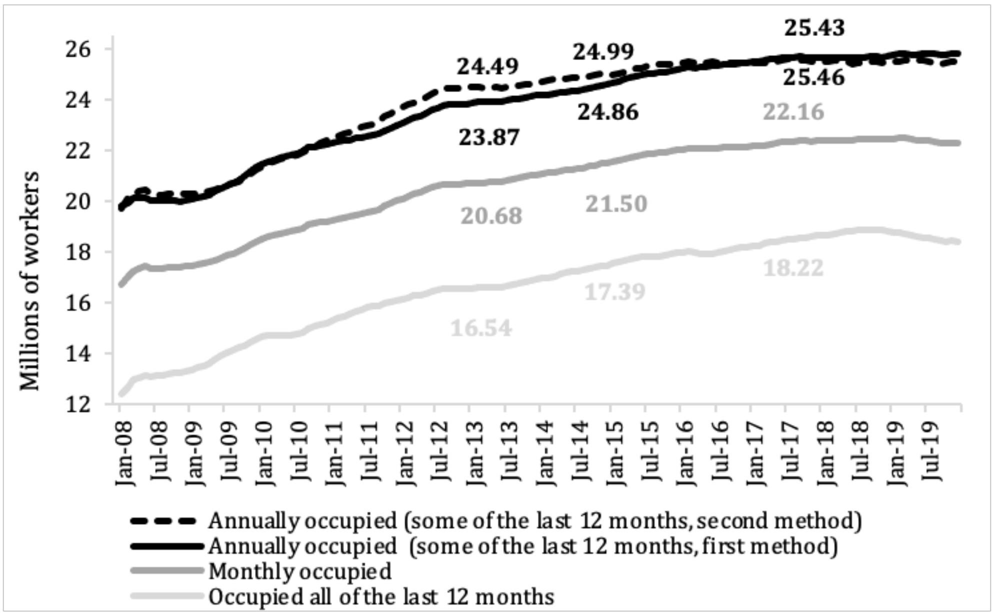
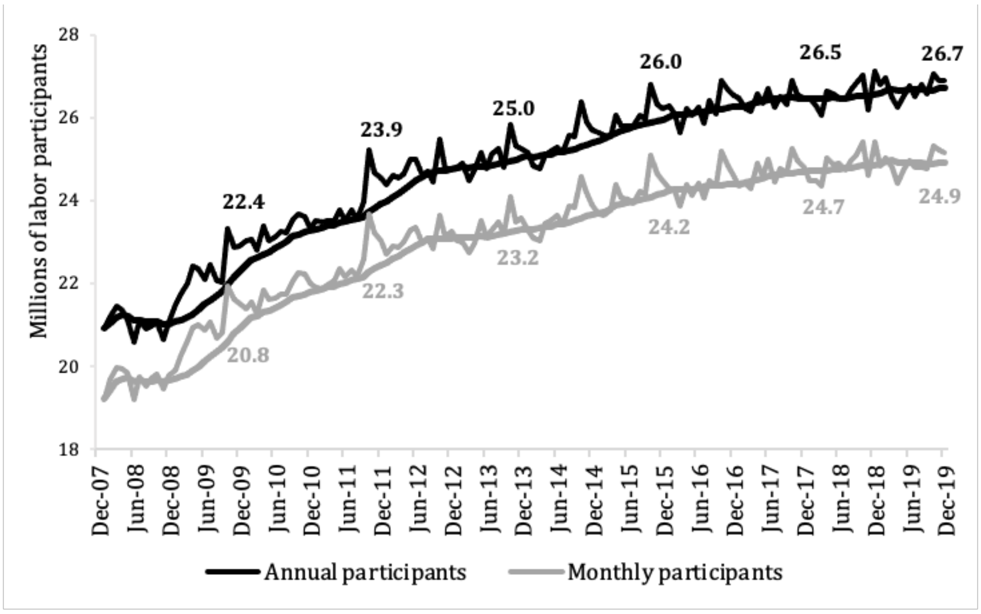


Figure 3. Labor participation (Colombia, 2008-2019)

a. Millions of individuals



b. Rates (of working age population)

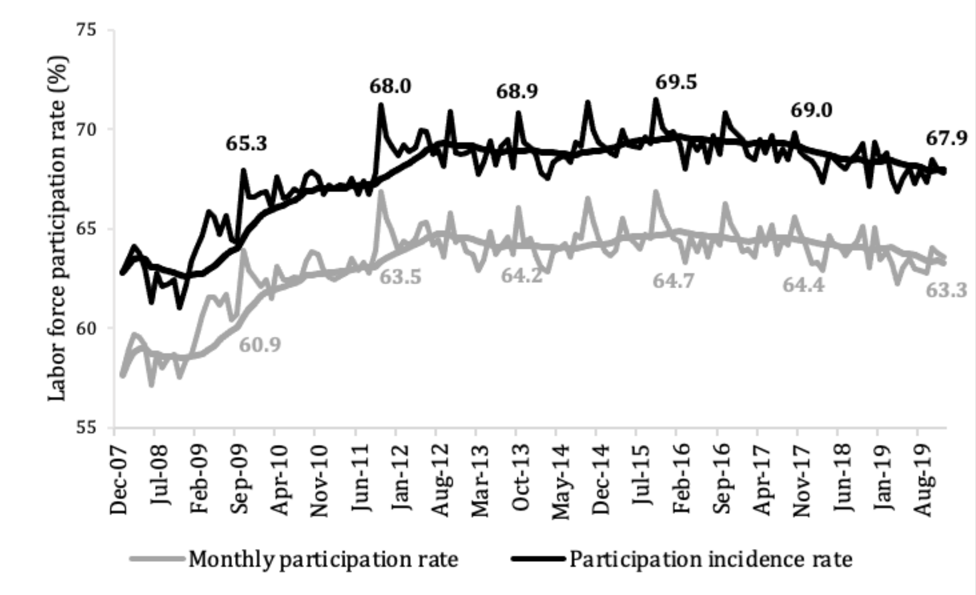


Figure 4. Estimation of nominal annual labor income (pesos, Colombia, 2009-2019)

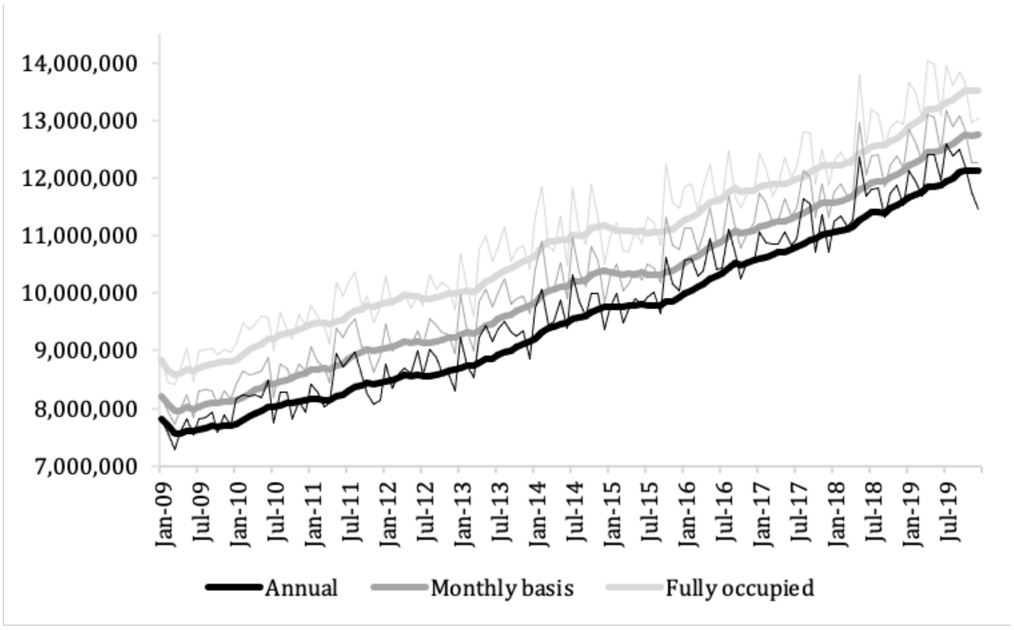
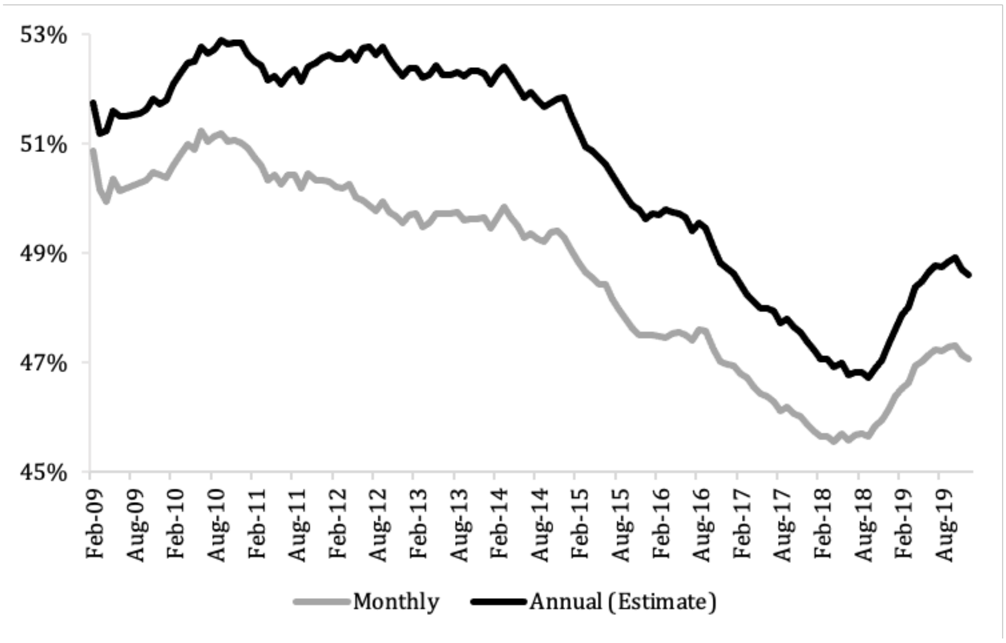


Figure 5. Gini index of labor income (Colombia, 2009-2019)


Feb-09
Aug-09
Feb-10
Aug-10
Feb-11
Aug-11
Feb-12
Aug-12
Feb-13
Aug-13
Feb-14
Aug-14
Feb-15
Aug-15
Feb-16
Aug-16
Feb-17
Aug-17
Feb-18
Aug-18
Feb-19
Aug-19
Monthly
Annual (Estimate)