Maximum Constrained Pseudo-Likelihood Estimation of Income Distributions, Combining Sources.

Alfredo Bustos[1]
Miriam Romo[2]
[1] *Instituto Nacional de Estadística y Geografía (INEGI), H. de Nacozari 2301, Aguascalientes 20276, México, Tel.: +52 449 910 5431; E-mail: alfredo.bustos@inegi.org.mx*
[2] *Instituto Nacional de Estadística y Geografía (INEGI), H. de Nacozari 2301, Aguascalientes 20276, México Tel.: +52 449 910 1151; E-mail: miriam.romo@inegi.org.mx.*

Abstract
Accurate determination of the distribution of income from data collected in households through sample surveys is difficult due to limitations presented by this data source. Among them are the under-declaration of income throughout the entire sample; as well as the lack of representatives, in it, of the small very high-income household population group. The simultaneous use of this source with others, such as fiscal records or household income national accounts aggregates, contributes to reducing, albeit partially, its limitations. Similarly, limitations of supplemental sources will also see their importance reduced. In this note, we seek to extend results, which simultaneously use several data sources to improve the estimation of income distributions, to include the years 2010, 2012, 2014 and 2016, both for Mexico and for the Mexican states. The above will allow us to establish new comparisons with sample results to confirm or rectify the behaviour of various trends during the period. In the same way, we will carry out comparisons for the states over time, and between them. We will show that the adjustments made give rise to an increasing trend in the Gini at the national level, which contrasts with the decrease in the value of the same indicator for Mexico's income and expenditure survey (ENIGH). In other words, our results indicate that inequality increased during the period. Despite this, inequality within states does not seem to have undergone a significant evolution. This seems to suggest that what has driven the growth of inequality in the country is an increase in inequality between states.

Key Words: Maximum Constrained Pseudo-Likelihood (MCPL), Generalized Gamma, Generalized Beta, Gini Coefficient, Household Income Survey.

1. Introduction

The Mexican sources of statistical information on household income are diverse. All of them show advantages and disadvantages in terms of estimating income distributions[1]. Undoubtedly, the best known are population and housing censuses, which are carried out with different frequencies, using different methodologies, by almost every country on the

[1] The interested reader may extend this information in UNECE (2011).

planet. Its universal coverage, when it is the case, results in a high cost of collection and, consequently, an interval of several years between one survey and the next. The universal coverage of surveys through population and housing censuses often leads to the belief that the information collected during these exercises is more accurate. However, this may or not be true depending on the topic or the variable in question. In the case of the income variable, there are limitations that have led the UN Statistics Division to exclude it from the group of important census variables to be collected (see Table 3, DESA (2017)). First, it is not entirely clear that the person providing the information precisely knows the income of each of the residents in the home that they receive, or their sources. It is also not clear if the declaration refers to gross or net income, since it is not easy for the respondent to know deductions experienced for income from work, whether they are tax, social security or of some other nature. The foregoing is aggravated when, under circumstances of high insecurity, the declaration of income is inaccurate or outright omitted.

The collection of information that is carried out through non-specialized surveys in the measurement of household income faces limitations like those of the census. An exception to the above is represented by the case of those for which one of the main topics is the study of household income. They generally seek that each of the recipients provide information on their own income. Additionally, each of them provides information on income received from different sources; among which are those from a job, whether dependent on third parties or not, government or relatives transfers, as well as support and loans from family members or friends, to mention some of them.

In other words, a greater thematic breakdown is contemplated at the expense of a smaller geographic one. In addition to this drawback, other limitations of this approach refer to the representativeness of the sample, an important question in the production of all official statistics. In general, it is difficult for the tiny fraction of the population whose income is extraordinarily high to be represented in a random sample. Given its small size, the error made when considering that the sample does represent the population may be considered negligible, but when the extraordinary size of their incomes is considered large biases become likely. Likewise, as in the census case, it is not clear whether the declaration refers to gross or net income. The foregoing is aggravated when the amount of income is undeclared or omitted.

A source that could represent universal coverage in relation to income is constituted by tax records. In theory, each of the recipients would periodically report, at least once a year, to the tax authority both the amount of their income and the tax contributions made in any given period. In addition to the declaration of the total income received, tax returns generally allow reference to each of the sources that make up said income. However, administrative records are often created with the intention of meeting specific needs, without considering alternative statistical uses that could be given to the information contained in them; tax records are no exception. In general, they do not comply with universal coverage for various reasons, among which are the informal employment situation in which a more or less important fraction of the economically active population works; also, not all formal workers are required to file an annual income statement. Therefore, the registry that collects the annual tax returns of natural persons only refers to that part of the population that is either obliged to present it for satisfying one or more regulatory conditions, or that does so voluntarily from the perspective of a partial tax refund. A complementary tax registry is one that concentrates the employer reports on the payments of wages and salaries, as well as on tax or other withholdings, to their workers. This information is updated monthly, but it should be noted that, as already noted, it covers only formal workers.

In turn, the System of National Accounts (SNA) represents an international effort to make countries' gross economic summaries available. Among its greatest virtues is that of being based on a coherent methodology that gives them great consistency, and which is constantly being updated to adapt to new circumstances (INEGI, 2017). The National Accounts provide a comprehensive description of all economic activity in the economic territory of the country, including activities that involve domestic units (that is, individuals and businesses resident in the country) and external units (residents in other countries).

There are two main types of institutional sectors. On one side are the legal or social states recognized by law that carry out activities and operations in their own name, such as companies, government units and non-profit institutions that serve households, for which there is a set complete accounts and balances of assets and liabilities. The other brings together individuals and households or groups of individuals, as well as individual companies without legal personality, which can be owners of assets and incur liabilities, but have no legal obligation to account for their activities. For purposes of determining the distribution of income, the SNA only provides aggregates at the national level and, in some cases, at the state level. We know efforts are being made to achieve what is called Distributional National Accounts (DINA; see Piketty et al. (2018)). These efforts notably make use of harmonized national accounts and fiscal data series, supplemented by income and wealth surveys, as well as inheritance, property, and wealth tax data.

1.1 Inconsistency between sample and national accounts aggregates.

Income distribution is one of the most frequently referred statistics both in specialized magazines and in newspapers and newscasts. Its applications are multiple in studies of poverty, inequality, and even fiscal matters.

The estimate of total household income provided by surveys in various countries shows a widely varying discrepancy with the total determined by the Systems of National Accounts for the institutional sector of households. In the Mexican case, the sample information estimates a total for current household income that consistently reaches a value less than half that recorded by the Mexican System of National Accounts (SMCN). It has been usual, in this as in many other countries, to assign greater credibility to this last figure. Consequently, we have sought to determine the causes that give rise to the apparent underestimation. Today, two main causes are identified, already mentioned, and identified as underreporting and truncation.

Given the difficulty of assigning the proportion of the deficit represented by each of the previous causes, it has been usual to act as if none or only one of the two were present. Of course, this can lead us to make decisions against which biases must be prevented. For exemplification purposes, consider using the income values as reported in the survey; that is, without any modification. Under these conditions, by not considering underreporting, the proportion of households that will exhibit an income less than a given number, such as the minimum income sufficient to satisfy some household needs, will be exaggerated. In other words, this proportion will turn out to be higher than the one that would be achieved had income been adjusted for underreporting. It is not difficult to think that, under these conditions, measures such as the prevalence of income poverty will be overestimated. Similarly, the absence of extraordinarily large incomes in the sample data could also lead to an underestimation of income inequality.

On the other hand, if it is assumed that the discrepancy is due exclusively to the underreporting of income in the survey, it can be suggested that, according to some criteria, the declared values are increased in such a way as to lead to the expanded sum of the new values to coincide with the total from national accounts. Although the foregoing

may solve the problem of discrepancy, it adds uncertainty for public policy purposes since the proportion of income that would be explained by truncation will be distributed among the households in the sample, which may lead to an excessive increase in their income. respective income. Consider then the case of those households whose declared income is close to, but below an income poverty line. It is possible to think that in some of these cases the new income will be higher than the mentioned poverty line. Households in these conditions will not be included in the estimate of the percentage of those living in poverty. Proceeding in this way can lead to an underestimation of poverty figures, the magnitude of which will not be known. What happens to the measurement of income inequality will depend on the form or method used to adjust but will probably decrease.

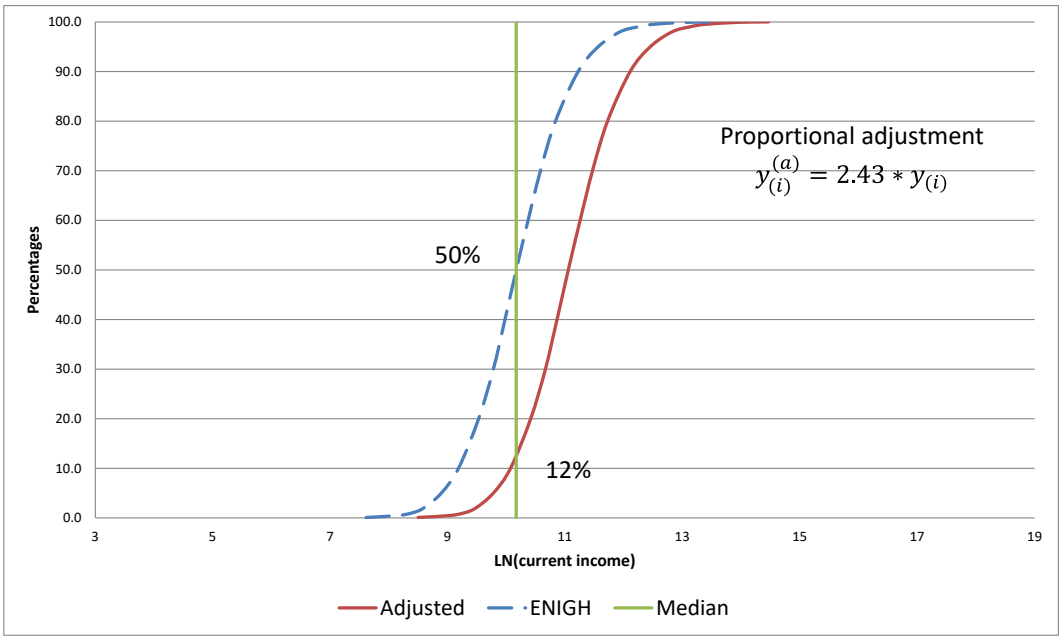


Figure 1: Proportional adjustment of observed income.

Fig. 1 illustrates the effect of poorly fitting income data from household surveys. Suppose for a moment that the median of the income distribution obtained from the sample data provides a reasonable estimate of what we would consider an income poverty line. Roughly, this would imply that half of the country's households, whose incomes are below said line, are considered in poverty. Furthermore, suppose that, since for our purposes it is sufficient for the expanded sum of the adjusted data to coincide with the total estimated by the system of national accounts, we consider adequate the adjustment that multiplies the observed figures by 2.43, as would be the case of Mexico for 2012. Note that, if the poverty line did not show any displacement after carrying out the adjustment, the proportion of households considered to be in poverty would drop to around 12%. It must be clear that such a "success" of social policy is nothing more than the result of an arithmetic trick, satisfying the only condition imposed, and clearly insufficient, on the expanded sum of adjusted values, and nothing else.

Finally, risks are also run when it is assumed that truncation fully explains the discrepancy between the two estimates. In this case, authors such as Campos et al. (2014), following Lakner and Milanovich (2013)[2], have followed the course of not

[2] Unlike Campos and co-authors, they adjust consumption rather than income.

applying any modification to the income of households that are in the first eight deciles of the income survey, so that the discrepancy between the sources is distributed in different proportions among those who occupy the two largest deciles. Regarding lower incomes, those associated with conditions of poverty, we return to the situation discussed in the paragraphs above in which income is not subject to any adjustment. In other words, it is probable that the magnitude of poverty is overestimated without being able to determine the magnitude of said overestimation caused by ignoring the presence of underreporting for those income levels. Of course, measures of income inequality will grow beyond what would seem a reasonable limit.

In the work by Campos et al., the distribution of income obtained following the previous procedure is linked to optimal taxation and possible collection. The possibly distorted image of said distribution can lead to the conclusion that the collection of tax revenues obtained today is much lower than that which would occur if its assumptions, debatable, were fulfilled. If the person in charge of preparing the Federal Government's revenue policy took these results uncritically, it is not difficult to think that he would conclude that the magnitude of tax evasion reaches significant levels. The same decisions that it would make based on such uncertain results could lead, for example, to implement fiscal adjustments that could lead to higher government deficits since the income actually taxed will not reach what was expected, according to these estimates.

In the next section, a brief account of the methodology developed to obtain the results presented later will be made. Throughout this investigation, our original proposal has been used, which has been called the “Maximum Pseudo-Restricted Likelihood” (MCPL) criterion. Some previously published results are also referenced to facilitate understanding of the material. The interested reader may turn to the works Bustos (2015 a, b), Bustos et al. (2017). The second section shows the behaviour of the results of applying this methodology to Mexican data between 2010 and 2016, disaggregating them by state. Finally, the last section of the work includes some considerations on the methodology, on its extensions, as well as some suggestions for future research work.

1.2 Methodology

First, it is necessary to make it clear that our research aims to find an approximation to the distribution of household income in Mexico, taking advantage of the available information in the best possible way. It is not intended, as has been the case with other related works, to produce a database with adjusted income values. Furthermore, unlike other studies that have sought to carry out some type of adjustment for the declaration of income in household surveys, we have decided to start by clearly establishing the optimality criterion that should guide our search for solutions in the estimate of income distribution. In this sense, the contrast is clear, multiple proposals have resulted in adjusted data that satisfy that their expanded sum coincides with the value provided by MSNA. However, it is not possible to make comparisons between them since they have been obtained with ad-hoc arguments, and without making explicit the efficiency measure used. It can be assumed that this is another reason why the measurement of poverty in Mexico at the beginning of the 21st century does not make use of any such adjustment proposal.

In summary, for our purposes the following objective is established: Make optimal use of the information available to estimate a distribution of household income in Mexico that is closer to reality and that is less arbitrary, without making imputations or corrections to the data of the ENIGH.

The proposed criterion seeks to maximize the pseudo-likelihood for the sample, considering the information from other sources, incorporated as constraints that must be

satisfied by the values of the parameter vector. This approach appeared for the first time in Bustos (2015), expressed as follows:

$$\underset{\underline{\theta},\underline{\lambda}}{Max}\left\{\sum_{i=1}^{n}\frac{1}{\pi_{(i)}}\ell\left(\underline{\theta};\underline{Y}_{(i)}\right)-\underline{\lambda}'\left(\underline{h}\left(\underline{\theta}\right)-\underline{c}\right)\right\}, \quad (1)$$

where:

$\ell\left(\underline{\theta};\underline{Y}_{(i)}\right)$ represents the natural logarithm of the density function, evaluated at the i-th sample value $\underline{Y}_{(i)}$, as a function of the vector of parameters $\underline{\theta}$ whose optimal value is to be determined;

$\pi_{(i)}$, to the inclusion probability corresponding to said sampling unit;

$\underline{h}\left(\underline{\theta}\right)$, function(s) of the parameters whose values are restricted to taking specific values given in the vector $\underline{c}$ (eg, SNA) and whose shape is linked to the particular distribution of try.

$\underline{\lambda}$, vector of Lagrange multipliers of appropriate dimensions.

Based on the quarterly household income of the ENIGH-2012, 2 probability distributions were adjusted:

- Generalized Gamma (GG),
- Generalized beta type II (GB2),

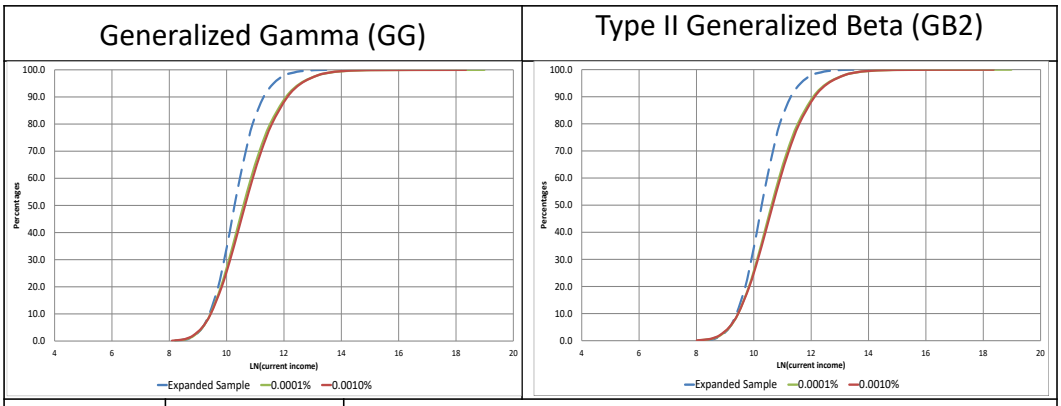
Table 1. Constraints used in numerical examples

Concept	Constraint:	Interpretation
Average income (MSNA)	$h_1(\underline{\theta}) = E\left[Y\middle\vert\underline{\theta}\right] = c_1$	Household income average from the model is equal to the one reported by MSNA.
Income Integral (SAT)	$h_3\left(\underline{\theta}\right) = E(X\vert X > \varphi_\alpha, \theta) = = \frac{1}{\alpha}\int_{\varphi_\alpha}^{\infty} y f_Y\left(y\middle\vert\underline{\theta}\right)dy = c_3$	Average income of households whose income is above threshold φ_α, according to SAT, is equal to average income in $100\alpha\%$ upper tail of fitted model.

The functional forms of $\boldsymbol{\ell}$ and $\boldsymbol{h}$ vary with each distribution selection. In all cases, a constraint was imposed: $\boldsymbol{h}(\underline{\boldsymbol{\theta}})$ = E[Y|$\underline{\boldsymbol{\theta}}$] = Average quarterly income per household, according to SNCM. To consider the information provided by anonymized tax records, a 2nd constraint was included. The following table summarizes the form that the constraints used throughout the development of the project have taken. The first one forces the parameters of the model, including those of the optimal fit, to take values such that the average value of the adjusted distribution coincides with the one that would be determined from the information from the Mexican System of National Accounts of (MSNA)[3]. The second leads to parameter values such that the proportion of households with incomes above a certain threshold, according to the adjusted model, coincides with

[3] For instance, in 2012 Mexico such household average equalled $86,410.57 MXN; this is the value used in constraint 1. Instead, ENIGH reported an average of $38,000 for the same year.

that obtained from the tax records of the Tax Administration System (SAT)[4]. The third also refers to the SAT records, but this time referring to the average income obtained by the group of households that earn more than the mentioned threshold. Said average must coincide both for the adjusted model and for the SAT records. In recent applications of the procedure, the first and the third have been used because the latter is considered more informative than the second[5].



	ENIGH	MCPL			
		GG		GB2	
		0.0010%	0.0001%	0.0010%	0.0001%
OPTIMAL VALUE CMPL	NA	-365,349,780	-364,628,437	-365,541,743	-364,913,147
CONSTRAINTS					
$E(X\|\theta)$	NA	✓	✓	✓	✓
$E(X\|X > \varphi_\alpha, \theta)$	NA	✓	✓	✓	✓
GINI COEFFICIENT	0.45	0.62	0.64	0.61	0.63
RATIO X/I	21.62	52.09	54.16	51.66	53.20

Figure 2: Comparison ENIGH vs. Four MCPL Fits (2 distributions, 2 thresholds each)

As an example of the advantages of incorporating additional information, we refer the reader to the Gini coefficients obtained in Bustos (2015) and in Bustos et al. (2017). In the first case, as there was no access to fiscal records, only the first constraint was used, resulting in values for the measurement of inequality close to 0.8, for the optimal model. For its part, once the SAT information was incorporated, the values reported for the Gini coefficient were between 0.61 and 0.64, for the four reported adjustments. In both cases, the adjustments were made using information from the ENIGH 2012, which reported a value of the Gini coefficient equal to 0.44, underestimating the inequality in accordance with the previous results. Subsequent analysis made it possible to establish that, for example, the threshold corresponding to one millionth of households for the adjustment with a single constraint turned out to have a value 10 times greater than that reported by the tax returns database for that year. In other words, said first adjustment took income corresponding to the lower deciles to transfer them to the upper ones, resulting in an apparent overestimation of the income received by households in the highest income deciles; consequently, according to these results, income inequality would also be

[4] The anonymized tax filings data base SAT made available does not allow identification of households. This made it necessary for us to assume individuals with incomes above a large threshold as households.

[5] It should be clear that income thresholds as well as the number of people with incomes above them is all that is required from the tax authority.

overestimated. It is clear, on the other hand, that the fiscal policies that can be implemented in one case or another would also show significant differences and amounts. From the second set of results it seems clear that the constraints that have been imposed have an important weight in their determination. Indeed, for this case, four adjustments were made from the consideration of two different thresholds for each of the distributions. The table that accompanies the previous fig. 2 shows a summary of relevant results. It will be observed that the optimal values of the criterion used exhibit very little variability. It will also be observed that something similar occurs both for the Gini coefficients and for the relationships between the total income of the first and the 10th deciles, denoted by Ratio X / I in the table. In our opinion, the thresholds used make little difference; the models used make even less difference. This statement is important since the apparent arbitrariness in the selection of models becomes irrelevant, for all practical purposes.

As can be seen in fig. 3, the consideration of gross aggregates such as, for example, the income shares of each of the deciles of the household population seem to reflect an important similarity between the alternative models. The accumulated shares in the income of the first five deciles show a difference of only 0.3 percent. For its part, the participation of the last decile is between 52.8 and 53.8 percent of the total. Those at the top 1% differ by less than 1%.

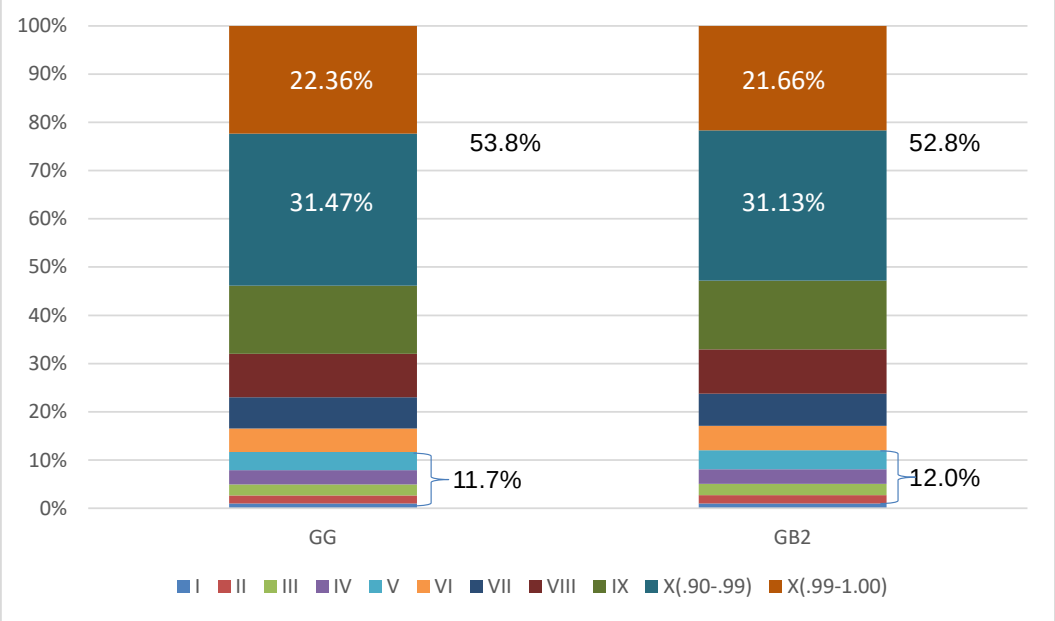


Figure 3: Income distribution by decile according to two models, Mexico, 2012

Complementarily, the large weight of the constraints can also be observed through another example, which arises from a proposal received during the discussion of our results with public officials. The proposal echoed one made by Martínez (1973) who suggested that, in order to correct the under-declaration of low income in household income and expenditure surveys, declared income values should be replaced by household expenditure values, whenever the former are lower than the latter. When the values of other sources of information on income are ignored, the sample results will show an increase in the lower tail, as shown in the fig. 3. On the other hand, when imposing constraints, as we have described, the adjustments obtained using only the

declared income and those that also use the expenditure information do not show any significant difference.

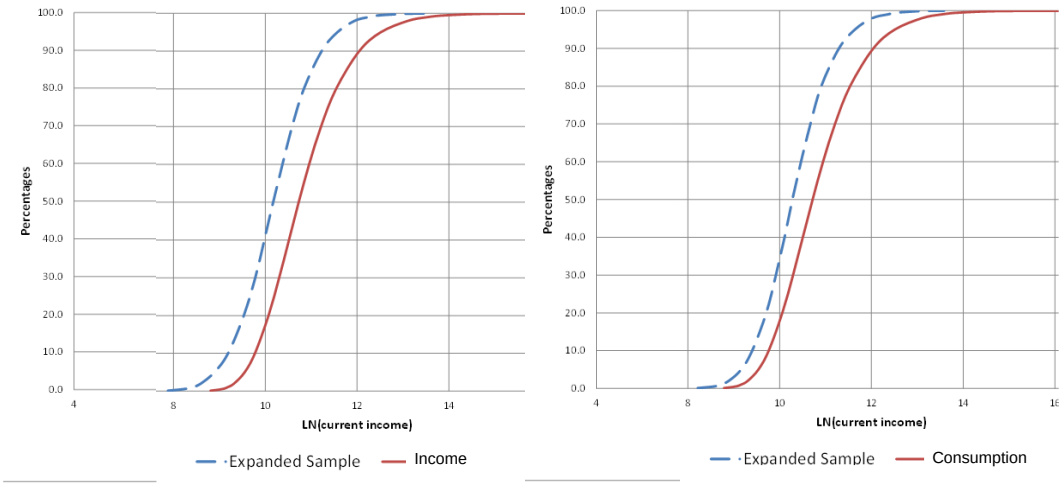


Figure 4: Effect of replacing household income values smaller than expenditure, 2012.

From the foregoing, it follows that the constraints weigh on the results as much or more than the sample information that serves as their input. This seems to have led to another proposal, this one from colleagues from Eurostat, aimed at relaxing the importance of constraints the same way as this is done in Hodrick-Prescott filter. That is, fixing the values of the Lagrange multipliers "by eye", even when this leads to a suboptimal fit. Given that the main motivation of this work is to provide an objective criterion to make the adjustment, this proposal has not been explored in view of the subjectivity it introduces.

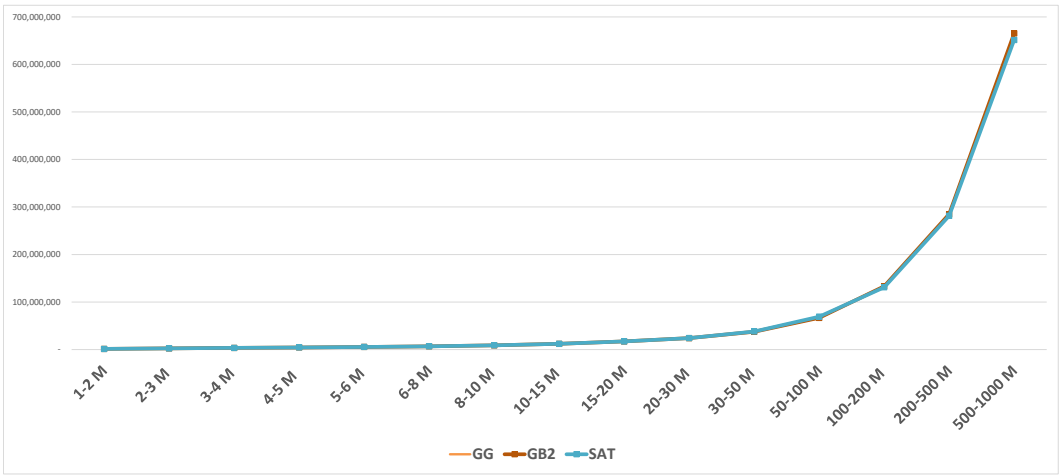


Figure 5: Yearly average according to income range, by source, Mexico, 2014.

Fig. 5 shows an important coincidence between the income averages according to the SAT database and those calculated from the adjusted models, when households are classified in ranges according to their income. Fig. 6, on the other hand, shows the number of households in each income range, and exhibits important differences between both sets of calculations, particularly in the low-income ranges. At least in part, the aforementioned difference seems to be a consequence of the assumption introduced when

making use of the tax information; that is, to make use of a database whose observation unit is the individual and not the household. While this assumption does not seem to have great consequences for the upper tail, the number of individuals whose income is in one of the lower ranges will be less than the number of households whose income falls in the same range. These households cannot be made perceptible from the database of tax filings by individual taxpayers. Indeed, as we move away from high incomes, the effect of different units of observation will become more important. Of course, it is not ruled out that other causes, such as our other assumption about the reliability of the data from the MSNA, or that tax under-declaration, are present in the differences shown. Unfortunately, we lack information that allows us to determine the relative importance of each of them.

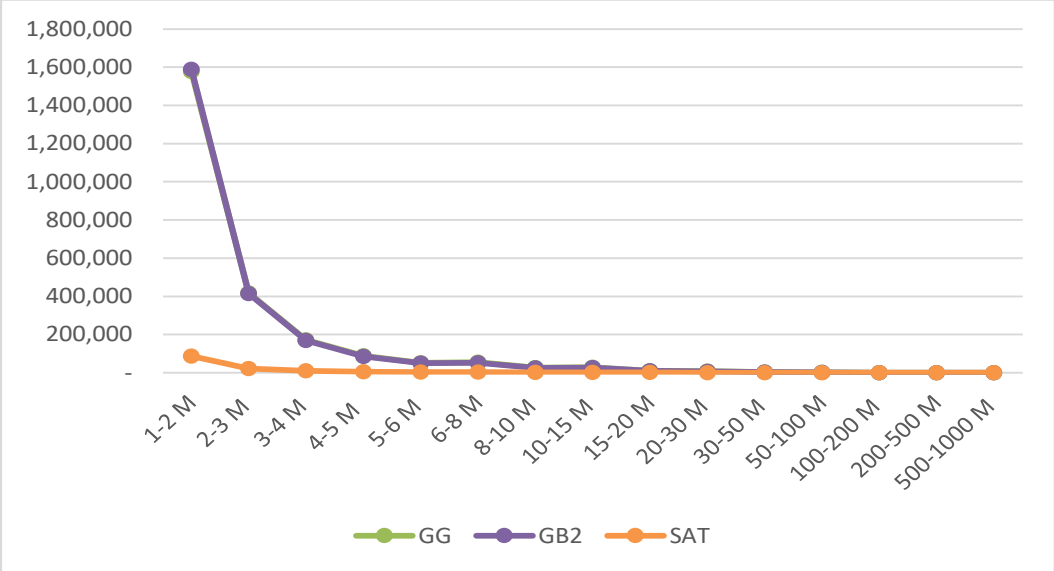


Figure 6: Households by income range, according to source, Mexico, 2014

2. Results

In this section, through a set of statistical summaries, we will illustrate the way in which our results show the evolution of income distribution over both time and space. For these purposes, use will be made of INEGI's databases for ENIGH, or its Socioeconomic Conditions Module (MCS); of SAT's anonymized tax returns; as well as the MSNA's results published for the institutional sector of households. All of them for the years 2010, 2012, 2014 and 2016.

2.1 National comparisons over time

First, we will focus our attention on the study of inequality along various dimensions. Fig. 7 shows the evolution of the Gini coefficient computed from both ENIGH and from the adjustment of a generalized gamma distribution, imposing the constraint associated with the one millionth threshold, throughout the study period. As can be seen, the gap between the two lines has an almost constant magnitude except, perhaps, for the last year. Indeed, while for the survey there is no significant change in the value of the Gini

coefficient, for the adjusted model inequality shows an increase that takes it above 0.65 in 2016 for the first time during the period.

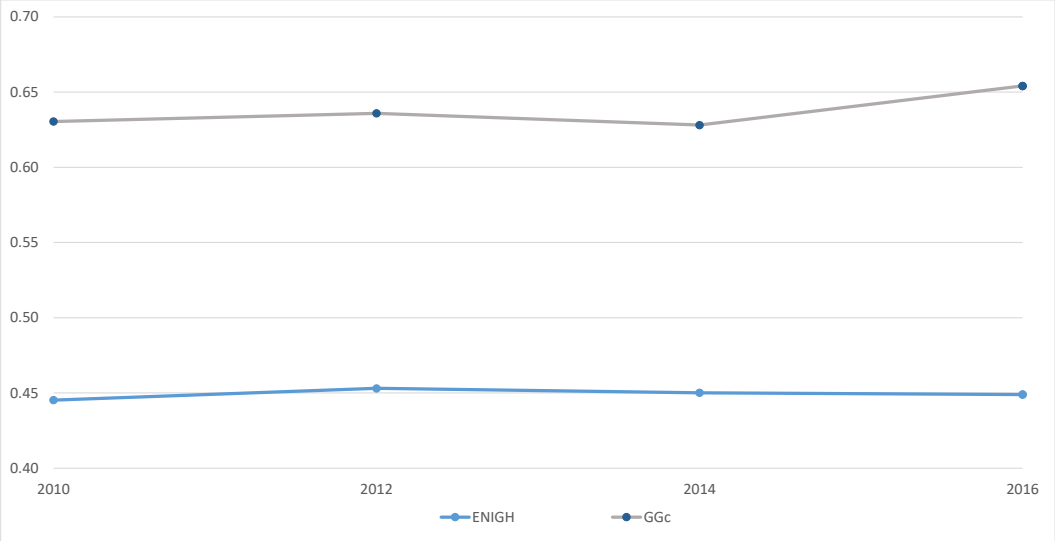


Figure 7: Gini Coefficients, Mexico, 2010-2016

The X-th to I-st decile total income ratio contrasts the incomes of the two sets of households, each one made up of 10% of the total, at both ends of the distribution. In other words, unlike the Gini coefficient, it ignores the results for the intermediate deciles. Fig. 8 seems to indicate that for the survey the gaps between the extremes have been narrowing as this ratio remains below 25. On the other hand, the MCPL adjustments, which begin with a decrease from the first to the second year of those considered, show that this discrepancy has been growing. Income at the top reaching a value 75 times the one at the bottom, for 2016.

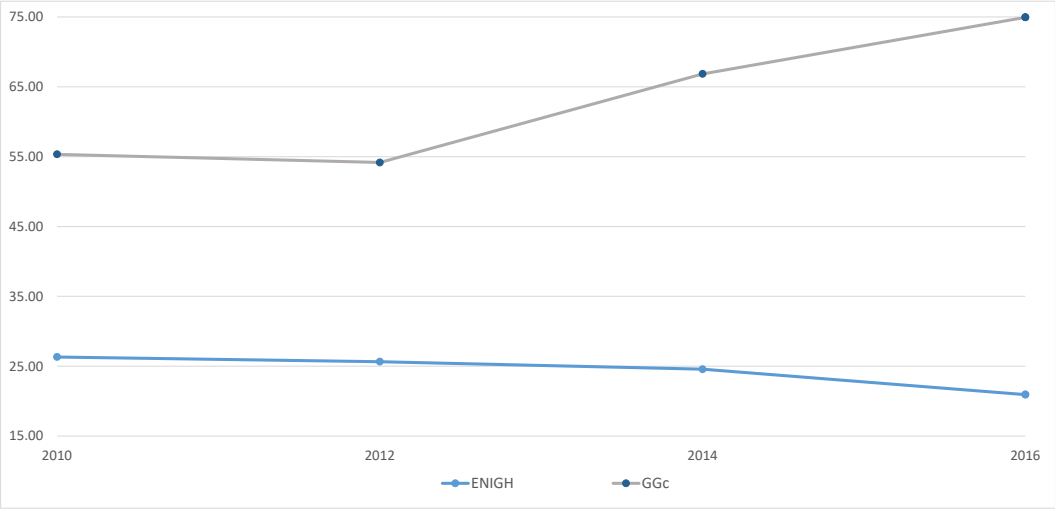


Figure 8: X-th to I-st decile total income ratio, 2010-2016

In other words, the values of the two statistics considered, which are not affected by inflation, exhibit the apparent advantages of making more efficient use of the available information. Indeed, the images that both sets of results paint in our minds, as well as the

conclusions that can be drawn from them for public policy purposes, are contrasting. In the case of the Gini coefficients, it could be thought that changes introduced in the survey methodology from 2015 could have resulted in an anomalous behaviour for 2016. However, when we consider the second set of results, it can be conclude that the gap between both years had already started to increase since 2014; that is, at least one year before the introduction of modifications to the survey. Based on the above, it is difficult to ignore the inequalities observed in both cases for 2016.

2.1 State-level results

It is also important to disaggregate the previous results at least down to the state level. For this purpose, ENIGH's MCS were used for the same years, considering its larger state sample. Based on this information, a model was adjusted for each of the states following the procedure described above. In view of the lack of availability of information for the institutional sector of households at the level of the federal state, it was resorted to proportionally distributing its total national income for households, according to the relative participation provided by the ENIGH itself[6].

In this way, the state version of one of the two constraints that we have been using was obtained. In contrast, SAT databases provides the state of residence of the taxpayer. Based on the above considerations, it is possible to obtain a set of results at the state level. In view of the smaller population sizes involved, the results can be strongly influenced by the fortuitous occurrence of extraordinary events such as, for example, winning a prize in a major draw.

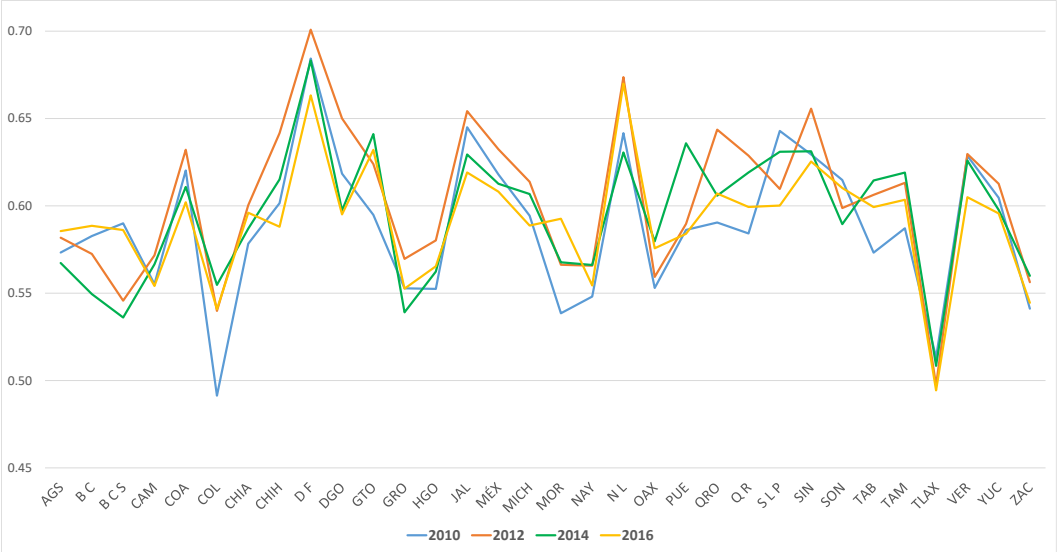


Figure 9: Gini coefficients by state, 2010-2016.

From fig. 9, according to the adjusted models, inequality is not uniformly distributed over the Mexican states. Despite the variability exhibited by the values presented there, the presence of three sets of states can be considered. Those states whose Ginis are above 0.6 for all the years considered belong to the first of them. Next, we would have the set of states for which one or more of the values is below 0.6, while one or more of them is

[6] State GDP could not be used as a benchmark for the states' share of national household income since, for instance, oil extraction value in some states by the national oil company introduced distortions.

above. Finally, there are those states whose Gini coefficients remain below the mentioned value.
On the other hand, while some of the states show very similar behaviours over time, others show a changing circumstance. It is worth noting that, in none of the cases, does 2016 show the most extreme conditions of inequality, contrary to what happens at the national level. This leads us to ask if it is possible that within the states a reduction in inequality may take place when, among them, an increase occurs.

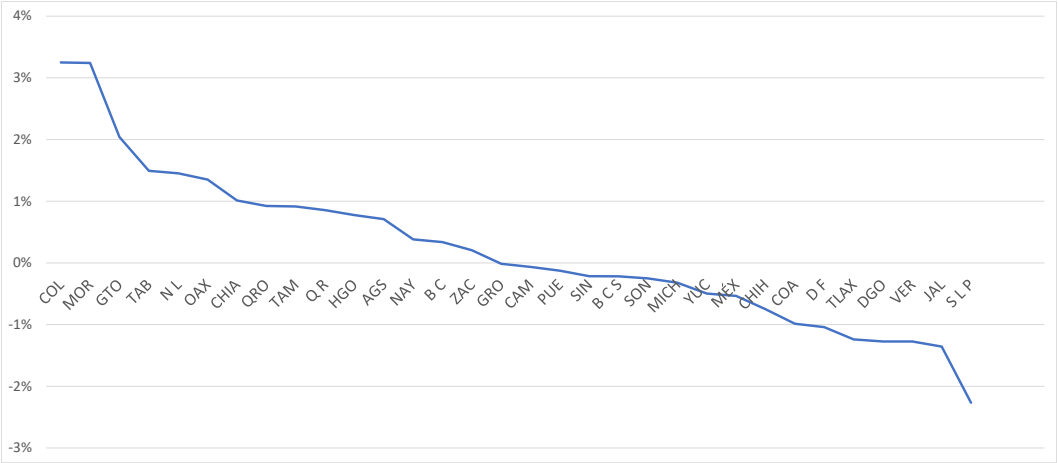


Figure 10: Geometric mean of yearly growth of Gini coefficients, by state, 2010-2016.

Fig. 10 shows the geometric average for the growth rate of the Gini coefficient, by state, for the period 2010-2016. Almost half the states show an increase in the values of this coefficient, so inequality seems to have grown in them. It is necessary to specify that a high rate of growth in inequality may be due to a small value at the beginning of the period, as may be the case for Colima and Morelos. On the other hand, it is noteworthy that two of the states with the highest inequality in 2010, Nuevo León and the Federal District (now Mexico City), show divergent trajectories; the value of the coefficient increased for the first, while for the second the opposite occurred. It is these movements in opposite directions for the states that lead us to think that the growing national inequality can be explained more in terms of the behaviour of inequality between them, than that which occurs within them.
The previous question is also applicable when considering comparisons between income totals for the two extreme deciles. Again, we observe that neither for 2014 nor for 2016, such ratios are uniformly higher for all states. In fact, it is possible to observe that, for this last year and for a large proportion of the states, the values of the ratios turn out to be minimal when compared with those of the whole period.
This is consistent with national and international evidence. For instance, Lakner (2013) points out that "most of global inequality is accounted for by differences between countries, although this contribution has declined over time, suggesting that countries have become more similar. The within-country component of global inequality, however, has increased continuously over this twenty-year period."
Evidence of divergent trajectories of municipal marginalization in Mexico, between 1990 and 2010, has been documented in Vargas and Cortés (2016). They state that their results "… provide clear evidence that marginalization in Mexico has been decreasing, but with an increase in heterogeneity among municipalities in the last 20 years. In other words, although marginalization has decreased in the last two decades, the country has experienced, at the municipal level, a process of territorial divergence, rather than

convergence". Applied to the case of income inequality, this would allow us to give a positive answer to the questions posed above, stating that it is possible that inequality decreases within the states, while it grows between them.

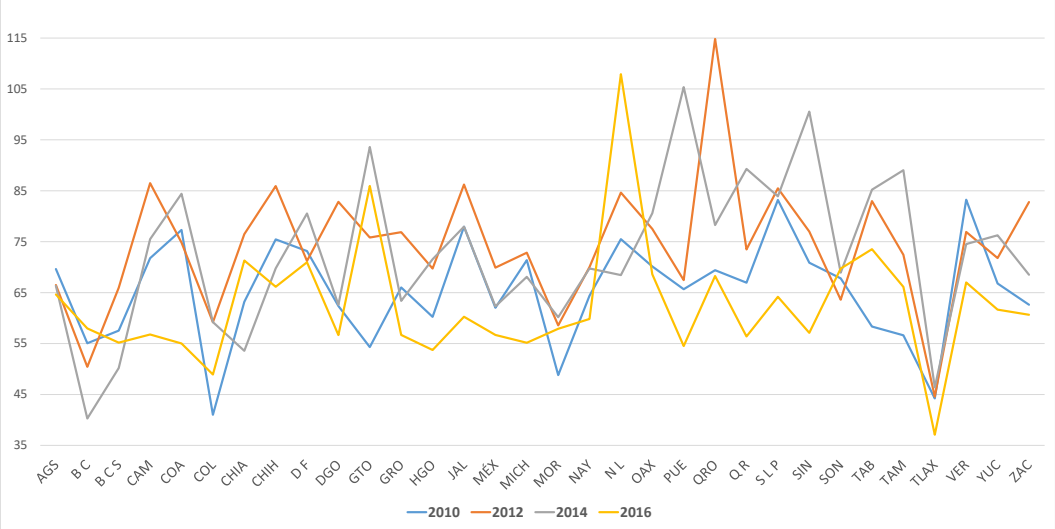


Figure 11: X-th to I-st decile total income ratio, by state, 2010-2016.

3. Concluding remarks

Contrary to what happens at the national level, Mexico's national accounts do not produce information at the state level for the institutional sector of households. For this reason, it is necessary to resort to the assumption that the state participation in household income coincides with that of the sample information. The discussion of our results may be affected by this assumption, but to what extent we do not yet know.

But for some exceptions, although at the national level inequality worsened between 2010 and 2016, within the states there is no significant deterioration. This has led us to postulate that inequality between states is an important factor that must be considered when trying to explain their behaviour at the national level, over time. Currently, however, we do not have sufficient evidence to support or reject this hypothesis.

For the previous results, there are many common elements that allow comparisons: the sample design, treatment for missing income values, similar income percentiles according to the SAT, allocation of the state average of income according to MSNA and ENIGH, to name a few. International organizations have a more complicated task when trying to carry out comparisons between countries based on sample results. Procedures such as oversampling of high-income households or calibration of results with exogenous totals have been carried out by some countries, but not all. This has led to the introduction of standard procedures. For example, to calculate the incidence of poverty in Latin American and Caribbeancountries, ECLAC in the past applied a procedure for correcting the income declared in the survey known as "income adjustment", which sought to minimize the bias by underreporting of income in the surveys taking as a reference the household income and expenditure accounts of the System of National Accounts.

Perhaps its greatest deficiency was found in the assumption that the discrepancy between totals, by source, was distributed among households in a proportional way. They have recently eliminated this procedure and work with survey data without adjusting them (ECLAC (2019)). By resorting to the identification of a

reference population through a method called “moving quintiles” to set poverty lines, they argue that “income adjustment does not improve the plausibility, comparability or reliability of poverty measurements”. Of course, this occurs since the adjustment method does not seem to significantly modify the relative positions of households. However, they also admit that, for the case of measuring inequality, adjustments like the one discussed in this paper can be useful.

Acknowledgements

We are grateful to Gerardo Leyva, Emanuele Baldacci and Fabrice Gras provided useful comments. Remaining errors are totally due to the authors.

References

[1] Boghosian, B. M., Is inequality Inevitable? originally published with the title "The Inescapable Casino", Scientific American 321, 5, 70-77 (November 2019). https://www.scientificamerican.com/article/is-inequality-inevitable/

[2] Bustos, A., *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources*, Workshop on Measuring Inequalities of Income and Wealth, High-Level Expert Group on the Measurement of Economic Performance and Social Progress, Berlin, Sep. 15-16, 2015 https://fr.slideshare.net/StatsCommunications/hleg-thematic-workshop-on-measuring-inequalities-of-income-and-wealth-alfredo-bustos.

[3] Bustos, A., *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources*. *Statistical Journal of the IAOS*, Statistical Journal of the IAOS 31 (2015) 565–577 https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji923.

[4] Bustos, A., Leyva, G., *Towards a More Realistic Estimate of the Income Distribution in Mexico*. Latin American Policy, Volume 8, Issue 1, June 2017, Pages 114-126. https://onlinelibrary.wiley.com/doi/abs/10.1111/lamp.12114

[5] Campos Vázquez, R., Chávez Jiménez, E., Esquivel Hernández, G., “*Los Ingresos Altos, la Tributación Óptima y la Recaudación Posible*”, primer lugar, Premio Nacional de Finanzas Públicas, 2014.

[6] Cortés-Cáceres, F., “*El Cálculo de la Pobreza y la Desigualdad a Partir de la Encuesta Nacional de Ingresos y Gastos de los Hogares”*, Comercio Exterior, Vol. 51, No.10, México, Octubre 2001.

[7] DESA, Principles and Recommendations for Population and Housing Censuses, Revision 3, Statistics Division, Department of Economic and Social Affairs, United Nations, New York, 2017, consultado el 18 de junio de 2020 en https://ec.europa.eu/eurostat/ramon/statmanuals/files/PR_on_pop_and_hous_cens_Rev3.pdf

[8] Economic Commission for Latin America and the Caribbean (ECLAC), “Income poverty measurement: updated methodology and results”, ECLAC Methodologies, No. 2 (LC/PUB.2018/22-P), Santiago, 2019. https://repositorio.cepal.org/bitstream/handle/11362/44920/1/S1900554_en.pdf

[9] Instituto Nacional de Estadística y Geografía (INEGI), Encuesta Nacional de Ingresos y Gastos de los Hogares, ENIGH 2014, Nueva construcción de ingresos y gastos Diseño conceptual y definición de categorías y variables,

2015, https://www.inegi.org.mx/contenido/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825072728.pdf

[10] Instituto Nacional de Estadística y Geografía (INEGI), Sistema de Cuentas Nacionales de México Fuentes y metodologías Año base 2013 Cuentas por Sectores Institucionales, 2017. https://www.inegi.org.mx/contenidos/programas/si/2013/metodologias/SCNM_Metodo_CSI_B2013.pdf

[11] Lakner, Ch., Milanovic, B**.,** "Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession". World Bank policy research paper 6719, 2013.

[12] Martinez, I., "La Distribución del Ingreso en México: Tendencias y Proyecciones a 1980", en Solis L., La Economía Mexicana, Vol. 1, México D.F., Siglo Veintiuno Editores, 1973.

[13] Thomas Piketty, Emmanuel Saez, Gabriel Zucman, Distributional National Accounts: Methods and Estimates for the United States, *The Quarterly Journal of Economics*, Volume 133, Issue 2, May 2018, Pages 553-609, https://doi.org/10.1093/qje/qjx043

[14] UNECE, Canberra Group, Handbook on Household Income Statistics, Second Edition, United Nations, United Nations Economic Commission for Europe, Geneva, 2011.

[15] Vargas, D. y Cortés, F., "Análisis de las trayectorias de la marginación municipal en México de 1990 a 2010", en POR UN MÉXICO SOCIAL: CONTRA LA DESIGUALDAD, Cordera, et al. (Coords.), Universidad Nacional Autónoma de México, 2016.