Hecksher, Marcos; Neri, Marcelo Cortes; Silva, Pedro Luis do Nascimento

Working Paper

New imputation procedures in the measurement of inequality, growth, and poverty in Brazil

WIDER Working Paper, No. 2018/128

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Hecksher, Marcos; Neri, Marcelo Cortes; Silva, Pedro Luis do Nascimento (2018) : New imputation procedures in the measurement of inequality, growth, and poverty in Brazil, WIDER Working Paper, No. 2018/128, ISBN 978-92-9256-570-1, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, https://doi.org/10.35188/UNU-WIDER/2018/570-1

This Version is available at:
https://hdl.handle.net/10419/190175

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2018/128

New imputation procedures in the measurement of inequality, growth, and poverty in Brazil

Marcos Hecksher,[1] Marcelo Neri,[2] and Pedro Luis do Nascimento Silva[3]

October 2018

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: This paper develops a new imputation methodology applied to missing incomes values in PNAD. PNAD is the main Brazilian household survey, but it has no imputation. The imputation process starts by fitting regression models applied to different income sources considering the complex sampling design of the survey. Later this procedure is combined with stochastic methods. In 2015, 2.5 per cent of the sample had per capita incomes imputed, resulting in slightly higher levels of inequality. Inequality and poverty changes were not affected by imputation. We took advantage of the methodology proposed to input rents and to preserve pressure points in income distributions associated with Brazilian institutional features such as minimum wages.

Keywords: Imputation methods, stochastic imputation, inequality, poverty, missing incomes
JEL classification: C81, D31, D63, I3

[1] Instituto de Pesquisa Econômica Aplicada (IPEA) and Instituto Brasileiro de Geografia e Estatística (ENCE/IBGE), Rio de Janeiro, Brazil, corresponding author: mdhecksher@gmail.com; [2] Getulio Vargas Foundation (FGV) Social and FGV EPGE (Escola de Pós-Graduação em Economia), Rio de Janeiro, Brazil; [3] ENCE/IBGE, Rio de Janeiro, Brazil.

This study has been prepared within the UNU-WIDER project on 'Inequality in the Giants'.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-570-1

Typescript prepared by Joseph Laredo.

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

This paper focuses on the measurement of inequality and poverty by addressing some difficulties that arise from nonresponse and from the methods used to capture income in the main source of monitoring data on socioeconomic indicators in Brazil, the Pesquisa Nacional por Amostra de Domicílios (PNAD). The PNAD survey questionnaire asks about individual labour earnings in cash and in kind (products and goods) from the main job, from a secondary job, and, if applicable, from all other jobs. For all these questions, the reference period is a month. The same short reference period is used for eight other income questions that capture income from retirement (2 questions), pensions (2 questions), bonus to remain active when able to retire, rent, donations received, and other sources.

The combination of a short reference period and the fact that the questionnaire allows a proxy response means that many individuals—even those who have a job—may report having no (zero) income during the reference period. It is assumed that the same factors, combined with refusal or inability to provide answers, lead to some nonresponse to the income questions. The PNAD survey microdata therefore contain both zero and missing value codes for the income questions.

If one is analysing data from a single year in the long PNAD series, there is little to lose from discarding households or individuals with missing or zero income from the analysis. However, if one aims to compare income-based summaries over time, the fact that the proportion of records with zero or missing value codes varies over time may affect the analysis. Hence the aim of this paper is to address the problems of missing and zero incomes in the PNAD by imputing to remove such values/codes from the microdata, and subsequently to use the 'complete' data to produce revised estimates of income inequality that are unaffected by the varying proportions of missing and zero incomes.

The methodology implemented preserves the measurement of the second moment of income distribution. We compare mean and inequality measures with and without imputation to assess how social welfare's levels and changes are affected by the imputation procedures proposed. Our analysis of inequality uses concentration curves and also emphasizes its impact on poverty indicators. The idea is to increase the weights given to the bottom part of per capita income distribution, since traditional measures such as the Gini and Theil indexes give more weight to the opposite extreme of the income spectrum.

We take advantage of the methodology developed to add a separate analysis of imputed rent into income-based social measures. We also explore links between the imputation exercises performed and a few policy-related issues. One is measuring the impact of the minimum wage on different labour market segments, on social security, and on social policy as a numeraire that both determines eligibility criteria and benefits size. The imputation procedure preserves pressure points associated with the minimum wage and develops a separate model for each type of income source and working class in the case of labour income.

The rest of this paper is organized as follows. Section 2 provides a broad description of the PNAD survey. Section 3 presents the set of methods used to impute the income and other variables. Section 4 comments on the results of applying such methods for social measures based on individual incomes. It assesses the impacts of the procedures adopted on mean, inequality, and pressure points in distributions of the various income sources. Section 5 evaluates the impact of the imputation procedures proposed in terms of per capita household income-based measures between the years 2001 and 2015. It highlights the main changes to the results of selected measures

of mean income, inequality, and poverty due to imputation. It also takes advantage of the procedure to assess the impact of imputed rents on these measures. Section 6 concludes the paper with some remarks and recommendations for those who aim to analyse the imputed microdata made available by this work.

2 The PNAD survey

The PNAD survey was carried out by the Instituto Brasileiro de Geographia e Estastistica (IBGE; www.ibge.gov.br) from 1967 till 2015, when the survey series was ended and replaced by the PNAD Contínua (Pesquisa Nacional por Amostra de Domicílios Contínua). For almost 50 years the PNAD was the main source of socioeconomic data and indicators for Brazil. The other main sources for such data were the samples collected as part of the decennial population censuses, which enabled a more detailed geographic breakdown, and the Pesquisa de Orçamentos Familiares (POF), which provided income and consumption, but both were too infrequent for many uses.

The PNAD started as a quarterly survey in 1967, but it stopped in 1970 due to the 1970 decennial census. It then restarted in 1971 as an annual survey, carried out during a single quarter (typically from September to November) until the end of the series. It did not take place in 1970, 1974, 1975, 1980, 1991, 1994, 2000, or 2010. In 1974–75 it was replaced by the ENDEF (Estudo Nacional da Despesa Familiar). In 1994 it did not take place for administrative reasons. For all the other years it did not take place because of the decennial census, which included a 'long-form' sample questionnaire.

The PNAD coverage evolved over time, and from 2004 onwards it finally reached the whole national territory. Previously, for the longest part of the series, the survey did not cover the rural areas in the North region of Brazil, mostly for cost and operational reasons.

The survey adopted the same stratified, multistage sampling design for the entire period of its series. Stratification was mainly geographical, to enable the production of results for the main subdivisions of the country: 26 states plus the Federal District of Brasília, plus 9 metropolitan areas centred around the capitals of the states of Pará, Ceará, Pernambuco, Bahia, Minas Gerais, Rio de Janeiro, São Paulo, Paraná, and Rio Grande do Sul. Hence, municipalities were first stratified into 36 geographical strata (9 metropolitan areas, 9 complements of states that have metropolitan areas, and 18 states with no metropolitan area). In each of the 9 metropolitan areas, municipalities played the role of pseudo-strata, since the primary sampling units (PSUs) were the census enumeration areas, which were sampled with probability proportional to size (number of private households) after sorting by municipality, urban x rural classification, and enumeration area code. The same approach was adopted for other municipalities classified as large and included in the sample with certainty. In these strata, the PNAD therefore used a stratified two-stage cluster sample of households, with enumeration areas as PSUs and households as secondary sampling units (SSUs). In all other areas the PNAD used a stratified three-stage cluster sample of households, where municipalities were the PSUs, the census enumeration areas were the secondary sampling units (SSUs) and households were sampled in the last stage.

If a household was sampled, information was obtained for all its members. The survey, however, accepted proxy responding on behalf of children, absent household members, or other members who were unable to respond themselves.

Sampling of both municipalities and enumeration areas was always carried out by systematic sampling with probability proportional to size. For municipalities, the size measure was the

population as estimated from the latest decennial census available. For enumeration areas, size was the number of private households observed in the latest decennial census available. Sampling of households within enumeration areas was always carried out by equal probability systematic sampling after annual updates of the list of households in each sampled enumeration area.

Once the sample of municipalities and enumeration areas had been selected following each decennial census, this sample remained fixed until the next decennial census update. But within each enumeration area the samples of households were refreshed every year, using a procedure designed to yield non-overlapping samples with previous years in the same decade.

Overall, the survey for 2015 included a sample of 9,166 enumeration areas, 151,189 housing units, and 356,904 persons, spread over 1,100 municipalities. Sampling fractions varied from 1/950 to 1/150. More details are available from the methodological documents, especially IBGE (1981) and IBGE (2016b).

3 Imputation methodology

3.1 Main ideas

The imputation methodology adopted to complete the PNAD microdata combines methods already used by the IBGE to impute income variables in some of its other household surveys, with other methods based on econometric analysis.

For the period following the 2000 population census, the IBGE used regression trees (Breiman et al. 1993) to define imputation classes for the hot-deck imputation of missing incomes—this approach was used in both the 2000 and 2010 population censuses, as well as in the Pesquisa Mensal de Emprego (PME)—see, for example, IBGE (2003, 2007, 2016a).

Imputation classes are formed by considering a set of characteristics of the individuals and their households. A routine programmed in R (R Core Team 2017) was used to split the sample into successive binary partitions, using the characteristics considered. Once the full set of potential predictors for income was defined, the routine found, at each stage of the process, the predictor variable and its respective cut-off point that would divide the (sub)sample into two classes to maximize the income homogeneity within each class and the income heterogeneity between the two classes. Having as default parameters an expected number of classes (or terminal nodes) and a minimum number of observations in each of the classes to be formed, the method generated a set of relatively homogeneous classes with respect to the target income variable. These classes were subsequently used for the random selection of donors, whose information would be imputed to the records with missing incomes allocated in the same class. Note that there could be no missing values for any of the predictors, such that all records could be allocated in the classes formed by the routine. This approach is a case of 'random (hot-deck) imputation within classes'—see De Waal et al. (2011) with classes formed by regression trees.

In the 2010 population census and in the PNAD Contínua, the IBGE adopted the method of the nearest neighbour donor imputation (De Waal et al. 2011). The Canceis package, developed by Statistics Canada (2007), is used to implement the approach. A set of predictors for income is defined and a fixed importance weight is assigned to each of them. These weights are then used to calculate the distance between each observation with missing income (to be imputed) and complete observations having valid (unimputed) values of income using the set of predictor variables to measure dissimilarity. A predefined number of records with the smallest distances

from the receiver observation are considered as potential donors for each receiver and, among them, one is randomly selected to serve as donor.

A third method, more directly associated with econometrics, is the deterministic imputation of the predicted value of the income in a Mincerian earnings regression—an example of model-based regression imputation (De Waal et al. 2011). However, in this case, the variance of the imputed values will be smaller than the variance of observed incomes. In addition, the imputed values tend to lose relevant discontinuities observed in the original distribution, such as a high concentration of values on the exact level of the minimum wage and other inflexions in the curves that associate income with their predictors at different levels. Hence this method was modified for use in the present application, to counter these adverse side effects.

3.2 Approach used

What we did for the PNAD series aimed to preserve the econometric intuition of the model-based regression method, as well as to take advantage of the principles behind the two stochastic methods used by the IBGE.

We therefore used a modification of imputation by 'predictive mean matching', where a regression model is fitted for the target income Y, and then for each receiver (a record with missing income) a donor is randomly selected from a set of K nearest neighbours, where these neighbours are located using only the predicted values $\hat{Y}$ obtained from the fitted regression model. Observations with an observed income are sorted by $\hat{Y}$ and are used to find the set of K (=41, say) adjacent observations whose median predicted value $\hat{Y}$ is the nearest to each receiver's predicted value $\hat{Y}$.

As in the nearest neighbour imputation method mentioned above, the predicted value $\hat{Y}$ can be considered as a summary value for the predictor variables, but in which the weights become flexible and correspond to the coefficients of the fitted regression model, making better use of the information available in the data for each edition of the survey. As in the regression tree method, it is possible to fit different regressions not only for each edition of the survey, but also for different relatively homogeneous strata, with this homogeneity defined both in terms of the distribution of income values and in terms of the frequencies of nonresponse and invalid values. This method does not require specific packages for editing and imputation, and can be programmed in any software. We used Stata, to take advantage of the standardized variables available for all the editions of PNAD from the period 1981–2015, obtained from the Datazoom website (www.econ.puc-rio.br/datazoom/index.html).

3.3 Missing and zero incomes

The cases to be treated fall into two types: missing incomes and zero incomes. The first case refers to people reported as earning an income from a certain source, for which the value was not reported, where a missing value code was recorded in the database. The second case refers to households in which no income was reported from any resident, where a zero value was recorded in the derived variables for total and per capita household income.

As the two situations are quite distinct, the corresponding imputation treatments should also be. It was decided to first impute the missing incomes and then to treat the remaining zero household incomes by imputing values for all members of these households.

The imputation marks make clear what changes are needed to correct both cases and what are needed to eliminate only missing incomes. With this, users of the imputed data sets can easily

decide whether to consider or ignore each of the imputed cases according to the analysis to be made.

3.4 Variables imputed

The individual variables imputed are listed in Box 1, sorted according to the successive stages of the imputation process. All labour incomes were imputed together from the same donor in a first block. If a person has a total income from all jobs with a missing value, then the values for each labour income variable come from the same donor, selected among the nearest neighbours in terms of the predicted value in a single regression for the total income from all jobs.

Box 1

Stages of imputation of individual missing values for income variables

Imputation stage	Variables to be imputed
Labour incomes	Monthly income in cash—main job
	Monthly income in goods—main job
	Monthly income in cash—secondary job
	Monthly income in goods—secondary job
	Monthly income in cash—all other jobs
	Monthly income in goods—all other jobs
Retirement incomes	Retirement income from social security institute or from the federal government
	Other types of retirement income
Pensions	Pensions from social security institute or from the federal government
	Other types of pension
Other incomes	Financial investments, social programmes, dividends, etc.
Rents received	Rents received
Donations received	Donations received from non-residents
Permanence bonus	Bonus to remain active when able to retire

Source: Authors' construction.

Then there are a second block for retirement incomes and a third block for pension incomes. Other sources are imputed separately, following the order presented in the table: other incomes (financial, social programmes, etc.), rents received, donations received, permanence bonus.

At each stage of imputation, the values already imputed in previous steps are treated as if they were known (or collected), and can be used to predict the incomes to be imputed in the later stages. The blocks were ordered by decreasing participation in the total income computed by the PNAD. This choice has advantages and disadvantages. The use of the same donor for the different variables of each block helps to preserve the structure of covariance among the incomes within the block. But the fact that different donors can be used at different stages attenuates the covariances of incomes in different blocks.

The definition of the blocks followed from two assumptions. First, that the imputed database is going to be used not only for analysis of total incomes, but also for analysis of the incomes of each block and, to a lesser extent, in the specific components of each block. Second, that there are key differences in the prediction models to be used for each block.

The table is based on the individual income variables investigated in the latest editions of the PNAD, but the blocks can be aggregated for earlier editions with a simpler format.
In addition to the variables collected directly in the interviews, the following derived variables were imputed:

a) individual variables—monthly income in the main occupation, monthly income in all jobs, monthly income from all sources;
b) family variables—monthly family income, monthly per capita family income;
c) household variables—monthly household income, monthly per capita household income, monthly household income with unpaid rent.

This last variable is not among the incomes investigated by the PNAD and is derived by adding an implicit rent associated with the value of housing services to each household's total income. It was estimated for households that do not pay rent on the basis of the values observed for those that do pay.

3.5 Strata and regression model for labour incomes

All imputations were performed after separating the data into previously defined 'imputation strata'. The idea behind this approach is to maximize homogeneity between receivers and potential donors, prior to fitting models and imputing. Imputation strata were thus formed by cross-classifying the variables mentioned in Box 2 in the column 'imputation stratification variables'.

In the sequence, within each imputation stratum, linear regression models were fitted using log(income) as the response, and the predictors listed under the column labelled 'predictor variables' in Box 2.

Box 2

Imputation stratification variables and predictor variables for each stage

Imputation stage	Stratification variables	Predictor variables
Labour incomes	- Year - Respondent (the worker or another resident of the household) - Four groups by position in the main job (unregistered employee; registered employee or public/ military servant; self-employed; and employer)	- Years of schooling - Sex - Age - Status in the household - State - Urban - Metropolitan - Colour/race - Hours normally worked - Occupational group
Retirement incomes	- Year - Respondent - Urban	- Labour income post-imputation - Years of schooling - Sex - Age - Status in the household - State - Colour/race
Pensions	- Year - Respondent	- Previous stages total income - Years of schooling - Sex - Age - Status in the household - State - Colour/race
Other incomes	- Year - 65 years or older	- Previous stages total income - Years of schooling - Sex - Age - Colour/race
Rents received	- Year	- Previous stages total income - Years of schooling - Sex - Age - Colour/race - Urban - Metropolitan
Donations received	- Year	- Previous stages total income - Years of schooling - Age - Colour/race - Urban - Metropolitan - Status in the household - Region
Permanence bonus	- Year	- Labour income post-imputation

Source: Authors' construction.

4 Results of individual incomes imputation process

4.1 Models fitted to potential donors

The approach described in Section 3 was applied to the PNAD data for 2001 and 2015. All individual incomes and the paid rent were imputed. Once individual incomes were imputed, all relevant derived variables such as per capita household income were calculated and added to the data sets.

Tables 1 and 2 show, for each block of individual incomes, the number and the percentage of observations classified as potential donors, receivers, and those who did not participate in the respective imputation stage in 2001 and 2015. Labour incomes are those that have more receivers and donors in both years. Although the increase in the percentage of employed persons raised the percentage of donors in this block from 38.7 per cent in 2001 to 42.8 per cent in 2015, the percentage of receivers remained at 0.6 per cent of the total sample in both years.

Table 1: Number and percentage of observations by imputation stage (2001)

Group	Labour	Retirement	Pension	Other	Received rent	Donation	Permanence bonus
Donors	146,681	30,198	15,119	9,026	4,324	3,936	8
Receivers	2,393	222	87	145	16	15	0
Others	229,763	348,417	363,631	369,666	374,497	374,886	378,829
Donors	38.7%	8.0%	4.0%	2.4%	1.1%	1.0%	0.0%
Receivers	0.6%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Others	60.6%	92.0%	96.0%	97.6%	98.9%	99.0%	100.0%

Source: Authors' computations from PNAD data.

Table 2: Number and percentage of observations by imputation stage (2015)

Group	Labour	Retirement	Pension	Other	Received rent	Donation	Permanence bonus
Donors	152,687	39,539	17,090	26,349	2,995	2,716	51
Receivers	2,276	358	121	181	37	63	2
Others	201,941	317,007	339,693	330,374	353,872	354,125	356,851
Donors	42.8%	11.1%	4.8%	7.4%	0.8%	0.8%	0.0%
Receivers	0.6%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%
Others	56.6%	88.8%	95.2%	92.6%	99.2%	99.2%	100.0%

Source: Authors' computations from PNAD data.

The percentage of donors of other incomes, including social benefits and income from financial investments, grew by 5 percentage points and exceeded that of pensions between 2001 and 2015. However, since the average value of other incomes does not reach a third of the average pension, total pension income was still 94 per cent higher than other incomes in 2015, which justifies its precedence in the imputation process. The last source of income imputed was the rarest, the permanence bonus, which had no imputation in 2001, and in 2015 had only two observations imputed, based on 51 potential donors.

The original distribution of the derived variables in 2001 and 2015 is represented in Tables 3 and 4. Four groups are presented according to the values of these variables in the original database: zero, positive (except code for missing), NA or not available (usually because 'not applicable'), or code for missing (999,999,999,999). The sum of the absolute values in each column results in the total of individual observations of the sample: 378,837 in 2001 and 356,904 in 2015. The percentages in each column add up to 100 per cent.

Table 3: Number and percentage of observations by value of derived income variables (2001)

Value of income	All jobs	All sources	Household	Family
Zero	194,459	154,506	5,902	12,341
Positive	146,681	186,249	363,367	357,739
NA	35,557	35,557	1,696	1,476
Code for missing	2,140	2,525	7,872	7,281
Zero	51.3%	40.8%	1.6%	3.3%
Positive	38.7%	49.2%	95.9%	94.4%
NA	9.4%	9.4%	0.4%	0.4%
Code for missing	0.6%	0.7%	2.1%	1.9%

Source: Authors' computations from PNAD data.

Table 4: Number and percentage of observations by value of derived income variables (2015)

Value of income	All jobs	All sources	Household	Family
Zero	10,312	91,687	1,404	3,911
Positive	152,687	214,049	347,480	345,365
NA	191,711	48,305	592	537
Code for missing	2,194	2,863	7,428	7,091
Zero	2.9%	25.7%	0.4%	1.1%
Positive	42.8%	60.0%	97.4%	96.8%
NA	53.7%	13.5%	0.2%	0.2%
Code for missing	0.6%	0.8%	2.1%	2.0%

Source: Authors' computations from PNAD data.

The individual income variables (of all jobs and all sources) do not apply to people under 5 years of age in 2001 and under 10 years of age in 2015, which explains the high frequency of NA in these two columns and their growth between the two analysed years. In the case of income from all jobs, people of valid age who did not work received a zero value in 2001 and NA value in 2015, which helps to explain the reduction in the frequency of zero values and the increase in the frequency of NA values.

Total and per capita household income do not include income from residents whose status in the household is 'paying guest', 'domestic employee', or 'relative of a domestic employee'. Similarly, family income does not include income from residents whose status in the family is 'paying guest', 'domestic employee', or 'relative of a domestic employee'. Zero values appear when the person, household, or family declares none of the considered incomes. Values encoded as missing in these derived variables appear when there is some missing code among the considered individual incomes; that is, there is an unreported value of some existing but undeclared income.

The imputation process began by fitting the regression models considering only the observations classified as potential donors. The expected theoretical relations between the variables to be imputed and all the others available in the PNAD guided the initial choice of the potential predictor variables to be considered in each model. Then, model selection was performed considering the complex sampling design of the PNAD when testing the statistical significance of the predictor variables in 2015.

The first results indicated the convenience of excluding regressors that did not provide predictive power for the income or, in the case of factor variables, to aggregate categories with similar effects. In all models where the colour/race variable was maintained, for example, the available categories were aggregated, composing a binary variable with a value of 1 for people identified as White or 'yellow' (*amarelo*) and 0 for the other categories.

The criterion of theoretical pertinence, however, led to the option to keep as regressor of the most frequent incomes (labour, retirement, and pension) and paid rent the factor variable that identifies the 27 Brazilian states, without aggregations, although some states did not always present statistical significance. For donation income, the states were aggregated according to the five Brazilian macro-regions. For the other variables to be imputed, this variable was not used.

Tables 5 to 12 present the final models fitted for 2015 as used in the imputation process for that year, to be simplified in years when not all regressors are available. The coefficients presented in the tables are not valid for the strata where the imputations were actually made in the case of the most frequent incomes (labour, retirement, pension, and other). The large set of regression tables with the coefficients used in all strata and years is available for consultation, but here we present only the eight models that guided the selection of regressors.

The base levels defined for the factor variables of the different models were: relation to head of the household—son/daughter; state—São Paulo; occupational group—aggregate of others; region—Southeast; type of residence—apartment; wall material—brick; sewage—general network.

Linear regressions were fitted, almost always having the natural logarithm of the income to be imputed as the dependent variable. The imputed incomes in precedent blocks were accumulated and the natural logarithm of this accumulated income was used as a regressor for the next block, always in an interaction term with a binary variable that indicates the existence of any of the incomes already imputed. In the specific case of permanence bonus, a linear regression of the value of this income (without logarithm) against the income value of all jobs (without logarithm), and a constant were used. In the regression of the natural logarithm of paid rent, which is a household variable, no income was used as an explanatory variable.

In general, the results presented in Tables 5–12 confirm several of the expected relations: income in Brazil is usually higher among the more educated, men, older people, those identified as White or yellow, those responsible for their households, those in urban and metropolitan areas, those living in the richest states, and those concentrated in the Midwest, South, and Southeast regions of the country. Labour income tends to be higher for managers, science and arts professionals, and mid-level technicians, while it is lower for trade, service, and agricultural workers.

It is worth noting that the existence of some individual income among those that have already passed through imputation in previous stages presents a negative coefficient in all the regressions in which it was included as a predictor variable. The interaction of this indicator variable with the total value of incomes that have already been imputed has a positive coefficient in almost all regressions, except for the one whose dependent variable is 'other incomes'. These other incomes include both financial income and social benefits for low-income people.

Table 5: Regression of log labour income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	4.3330	0.0272	159.4	***
Years_of_schooling	0.0692	0.0008	85.6	***
Male	0.3340	0.0043	77.8	***
Age	0.0469	0.0010	45.8	***
Age^2	-0.0004	0.0000	-35.5	***
Head	0.1784	0.0055	32.3	***
Spouse	0.1250	0.0061	20.4	***
Other_status	0.0508	0.0074	6.8	***
Rondônia	0.0082	0.0188	0.4	
Acre	-0.1499	0.0289	-5.2	***
Amazonas	-0.2307	0.0252	-9.2	***
Roraima	-0.0608	0.0359	-1.7	.
Pará	-0.2372	0.0168	-14.1	***
Amapá	-0.0707	0.0297	-2.4	*
Tocantins	-0.0587	0.0218	-2.7	**
Maranhão	-0.4604	0.0350	-13.2	***
Piauí	-0.4865	0.0445	-10.9	***
Ceará	-0.4691	0.0203	-23.1	***
Rio Grande do Norte	-0.2885	0.0217	-13.3	***
Paraíba	-0.3878	0.0297	-13.1	***
Pernambuco	-0.3852	0.0178	-21.7	***
Alagoas	-0.3184	0.0204	-15.6	***
Sergipe	-0.3370	0.0324	-10.4	***
Bahia	-0.3755	0.0147	-25.6	***
Minas Gerais	-0.1051	0.0108	-9.8	***
Espírito Santo	-0.0372	0.0198	-1.9	.
Rio de Janeiro	-0.0716	0.0121	-5.9	***
Paraná	0.0026	0.0111	0.2	
Santa Catarina	0.1066	0.0166	6.4	***
Rio Grande do Sul	-0.0621	0.0118	-5.3	***
Mato Grosso do Sul	0.0570	0.0241	2.4	*
Mato Grosso	0.0978	0.0168	5.8	***
Goiás	0.0044	0.0124	0.4	
Distrito Federal	0.1828	0.0267	6.9	***
Urban	0.1477	0.0107	13.8	***
Metropolitan	0.1171	0.0072	16.3	***
White/yellow	0.1066	0.0043	24.9	***
Working_hours	0.0158	0.0003	59.7	***
Managers	0.5294	0.0115	45.9	***
Science/arts	0.4719	0.0101	46.8	***
Technicians_med.level	0.1980	0.0072	27.3	***
Service_workers	-0.1593	0.0050	-31.7	***
Merchants	-0.1403	0.0063	-22.3	***
Agricultural_workers	-0.2691	0.0141	-19.1	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.3801181); number of Fisher Scoring iterations: 2

Source: Authors' computations from PNAD data.

Table 6: Regression of log retirement income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	6.2947	0.0248	253.8	***
Has_previous_steps_income	-0.4389	0.0395	-11.1	***
ln_previous_steps_income	0.0454	0.0059	7.7	***
Years_of_schooling	0.0743	0.0009	82.0	***
Male	0.1782	0.0056	31.6	***
Age	0.0035	0.0003	11.4	***
Head	0.0633	0.0052	12.2	***
Rondônia	-0.0660	0.0195	-3.4	***
Acre	0.1002	0.0286	3.5	***
Amazonas	-0.0653	0.0216	-3.0	**
Roraima	-0.0269	0.0626	-0.4	
Pará	-0.0946	0.0124	-7.7	***
Amapá	-0.0222	0.0604	-0.4	
Tocantins	-0.1018	0.0169	-6.0	***
Maranhão	-0.0326	0.0177	-1.8	.
Piauí	-0.0269	0.0192	-1.4	
Ceará	-0.0992	0.0119	-8.3	***
Rio Grande do Norte	-0.0414	0.0196	-2.1	*
Paraíba	-0.0536	0.0257	-2.1	*
Pernambuco	-0.0843	0.0151	-5.6	***
Alagoas	-0.0847	0.0160	-5.3	***
Sergipe	-0.0211	0.0190	-1.1	
Bahia	-0.0544	0.0122	-4.5	***
Minas Gerais	-0.0657	0.0121	-5.4	***
Espírito Santo	-0.0467	0.0196	-2.4	*
Rio de Janeiro	-0.0415	0.0158	-2.6	**
Paraná	-0.0726	0.0141	-5.2	***
Santa Catarina	-0.0409	0.0182	-2.2	*
Rio Grande do Sul	-0.0705	0.0126	-5.6	***
Mato Grosso do Sul	-0.0468	0.0265	-1.8	.
Mato Grosso	-0.0691	0.0157	-4.4	***
Goiás	-0.0913	0.0176	-5.2	***
Distrito Federal	0.3904	0.0438	8.9	***
White/yellow	0.0332	0.0055	6.1	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.1990723); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 7: Regression of log pension income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	5.1883	0.0306	169.4	***
Has_previous_steps_income	-1.1590	0.0964	-12.0	***
ln_previous_steps_income	0.1405	0.0142	9.9	***
Years_of_schooling	0.0429	0.0016	27.2	***
Male	0.1059	0.0149	7.1	***
Age	0.0231	0.0004	62.6	***
Rondônia	-0.0563	0.0462	-1.2	
Acre	-0.0621	0.0561	-1.1	
Amazonas	-0.1071	0.0448	-2.4	*
Roraima	-0.1093	0.0778	-1.4	
Pará	-0.1046	0.0327	-3.2	**
Amapá	-0.2324	0.1077	-2.2	*
Tocantins	-0.1672	0.0364	-4.6	***
Maranhão	-0.0310	0.0411	-0.8	
Piauí	-0.1642	0.0556	-3.0	**
Ceará	-0.1982	0.0354	-5.6	***
Rio Grande do Norte	-0.2099	0.0617	-3.4	***
Paraíba	-0.1748	0.0485	-3.6	***
Pernambuco	-0.1761	0.0287	-6.1	***
Alagoas	-0.0952	0.0442	-2.2	*
Sergipe	-0.2714	0.0499	-5.4	***
Bahia	-0.2158	0.0264	-8.2	***
Minas Gerais	-0.1041	0.0212	-4.9	***
Espírito Santo	-0.0662	0.0561	-1.2	
Rio de Janeiro	0.0467	0.0245	1.9	.
Paraná	-0.0499	0.0223	-2.2	*
Santa Catarina	-0.0012	0.0343	-0.0	
Rio Grande do Sul	-0.0840	0.0218	-3.8	***
Mato Grosso do Sul	-0.0580	0.0553	-1.0	
Mato Grosso	-0.0413	0.0416	-1.0	
Goiás	-0.0185	0.0302	-0.6	
Distrito Federal	0.1979	0.0516	3.8	***
White/yellow	0.0804	0.0114	7.0	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.3883273); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 8: Regression of log other income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	4.8429	0.0273	177.6	***
Has_previous_steps_income	-0.1747	0.0894	-2.0	.
ln_previous_steps_income	-0.0377	0.0149	-2.5	*
Years_of_schooling	-0.0098	0.0018	-5.4	***
Male	0.6128	0.0198	31.0	***
Age	0.0137	0.0005	25.9	***
White/yellow	0.0534	0.0145	3.7	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.6824227); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 9: Regression of log rent income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	4.8997	0.1036	47.3	***
Has_previous_steps_income	-2.0715	0.1666	-12.4	***
ln_previous_steps_income	0.2447	0.0222	11.0	***
Years_of_schooling	0.0599	0.0043	13.8	***
Male	0.1262	0.0338	3.7	***
Age	0.0195	0.0013	15.0	***
Urban	0.1831	0.0605	3.0	**
Metropolitan	0.1414	0.0364	3.9	***
White/yellow	0.2040	0.0332	6.1	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.6021916); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 10: Regression of log donation income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	4.4525	0.0932	47.8	***
Has_previous_steps_income	-1.1201	0.1616	-6.9	***
ln_previous_steps_income	0.1476	0.0271	5.4	***
Years_of_schooling	0.0837	0.0045	18.5	***
Age	0.0100	0.0013	7.9	***
White/yellow	0.1325	0.0401	3.3	***
Urban	0.2221	0.0522	4.3	***
Metropolitan	0.1760	0.0406	4.3	***
Head	0.3814	0.0396	9.6	***
North	-0.1789	0.0600	-3.0	**
Northeast	-0.3441	0.0483	-7.1	***
South	0.2418	0.0647	3.7	***
Central-West	0.2554	0.0875	2.9	**

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.613422); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 11: Regression of permanence bonus income for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	493.8128	167.9474	2.9	*
Labour_income	0.0715	0.0119	6.0	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 567785.4); number of Fisher Scoring iterations: 2.

Source: Authors' computations from PNAD data.

Table 12: Regression of log paid rent for potential donors (2015)

	Estimate	Std.Error	t	
(Intercept)	5.5432	0.0635	87.3	***
Bathrooms	0.2631	0.0112	23.4	***
Rooms	0.0945	0.0041	23.0	***
Bedrooms	0.0560	0.0060	9.3	***
Residence_house	-0.3867	0.0140	-27.6	***
Residence_room	-0.3786	0.0455	-8.3	***
Wall_wood	-0.2604	0.0197	-13.2	***
Wall_other	-0.4289	0.0701	-6.1	***
No_concrete_slab_roof	-0.0604	0.0106	-5.7	***
No_piped_water	-0.2534	0.0450	-5.6	***
Sewage_septic_tank_drain	-0.0440	0.0189	-2.3	*
Sewage_septic_tank_no_drain	-0.1655	0.0171	-9.7	***
Sewage_other	-0.1387	0.0280	-5.0	***
Urban	0.2913	0.0608	4.8	***
Metropolitan	0.2763	0.0118	23.5	***
Rondônia	-0.1721	0.0321	-5.4	***
Acre	-0.1368	0.0312	-4.4	***
Amazonas	-0.1319	0.0317	-4.2	***
Roraima	-0.0973	0.0408	-2.4	*
Pará	-0.3369	0.0585	-5.8	***
Amapá	-0.1161	0.0739	-1.6	
Tocantins	-0.2930	0.0387	-7.6	***
Maranhão	-0.4418	0.0523	-8.5	***
Piauí	-0.6452	0.0521	-12.4	***
Ceará	-0.6962	0.0243	-28.7	***
Rio Grande do Norte	-0.5498	0.0373	-14.7	***
Paraíba	-0.6828	0.0614	-11.1	***
Pernambuco	-0.5754	0.0225	-25.6	***
Alagoas	-0.5273	0.0351	-15.0	***
Sergipe	-0.4694	0.0386	-12.2	***
Bahia	-0.6376	0.0235	-27.1	***
Minas Gerais	-0.3953	0.0195	-20.3	***
Espírito Santo	-0.2853	0.0279	-10.2	***
Rio de Janeiro	-0.1815	0.0213	-8.5	***
Paraná	-0.1653	0.0214	-7.7	***
Santa Catarina	0.0486	0.0258	1.9	.
Rio Grande do Sul	-0.1937	0.0216	-9.0	***
Mato Grosso do Sul	-0.0827	0.0246	-3.4	***
Mato Grosso	-0.0018	0.0401	0.0	
Goiás	-0.1444	0.0216	-6.7	***
Distrito Federal	-0.1321	0.0248	-5.3	***

Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for gaussian family taken to be 0.1816835); number of Fisher Scoring iterations: 2.

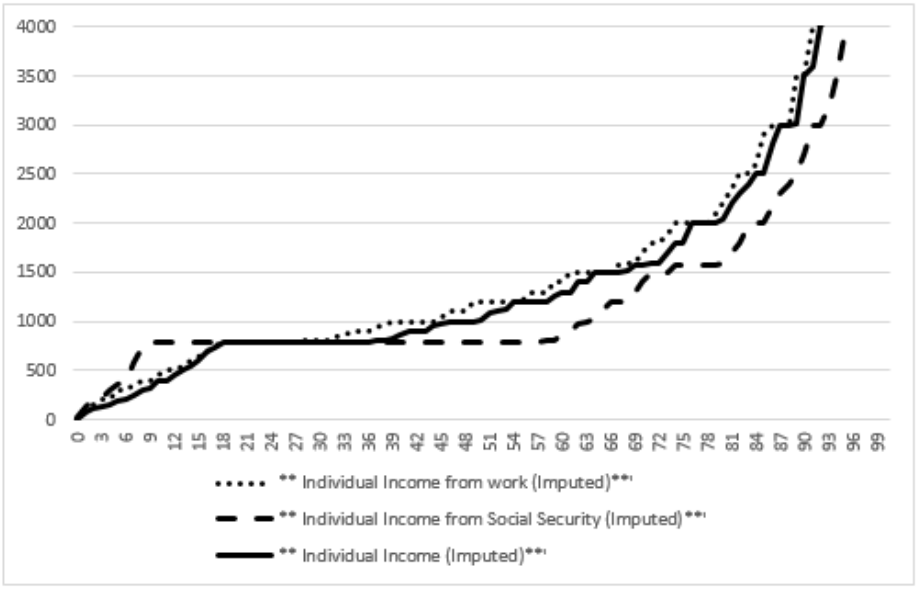
Source: Authors' computations from PNAD data.

4.2 Pressure points in individual income distributions

The imputation methodology developed and used in this paper preserves the accumulation of the distribution mass in certain pressure points—in particular, those associated with institutional features such as the federal minimum wage, which directly affects the two most important income sources in Brazil: labour earnings and social security benefits. Figure 1 illustrates both cumulative distribution functions, as well as showing a curve for the individual income from all sources. The pressure point at exactly the minimum wage of R$788 is bigger for social security (46 per cent) and even income from all sources (17 per cent) than for labour income (10 per cent). The Brazilian 1988 Constitution sets the minimum wage as the floor to social security benefits. In search of other institutional features, we replicate this in Figure 2 with labour income disaggregated by occupational status, taking advantage of having a separate imputation procedure for each one of them. We note that informal employees are also affected by the minimum wage, since the Brazilian judicial system allows them to take their employees to Labour Courts in Brazil (Neri et al. 2001).

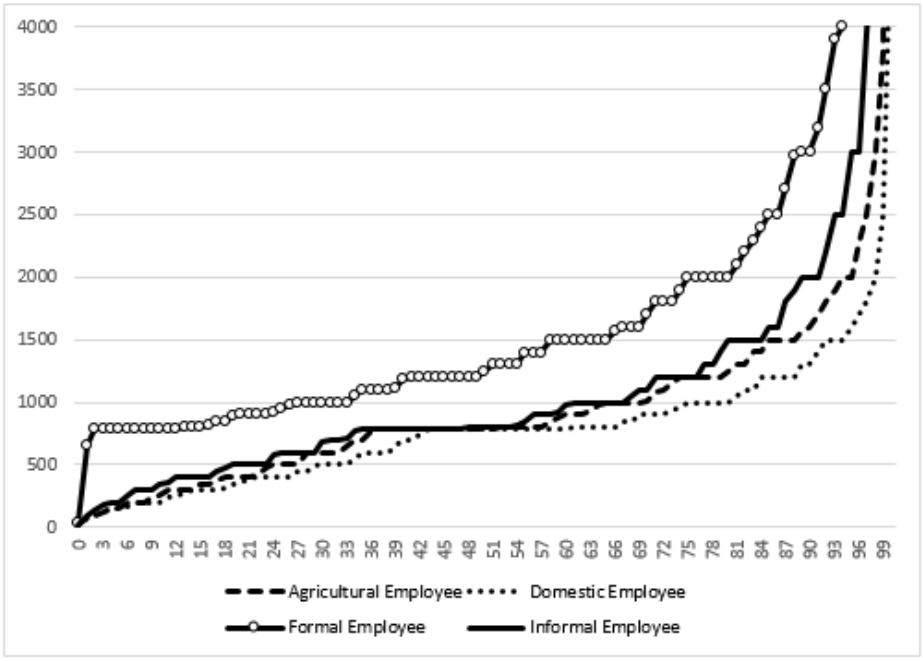
Each point in the domain of a continuous income distribution has a null mass. So the simple fact that these institutional points are gathering individuals suggests that the imputation procedure is keeping these institutional characteristics. Table 13 synthesizes this graphical information, presenting the share of each individual source below, equal to, and above the minimum wage for the overall distribution and the one only with imputed values. We focus on the preservation of the discontinuity at the minimum wage that is a feature of the imputation methodology proposed in this paper. Then, Table 14 presents a picture of 2015 pressure points for the income received from main job disaggregated by different working classes.

Figure 1: Individual income cumulative distributions (2015)



Source: Authors' computations from PNAD data.

Figure 2: Labour income cumulative distribution by occupational status (2015)



Source: Authors' computations from PNAD data.

Table 13: Positive income from work and minimum wage (mw)—income with imputed income and only imputed income (2001 and 2015)

Income concept		With imputation		Only imputation
		2001	2015	2015
Income from all jobs	% Positive	39.27	43.74	-
	< MW	18.92	17.27	16.84
	= MW	8.62	10.14	6.23
	> MW	72.46	72.59	76.93
Income from main job	% Positive	39.21	43.72	-
	< MW	19.3	17.46	17.31
	=MW	8.81	10.31	6.57
	> MW	71.88	72.22	76.12
Income from social security	% Positive	11.87	15.73	-
	< MW	5.56	9.41	6.14
	= MW	47	47.21	37.59
	> MW	47.43	43.38	56.27
Income from all sources	% Positive	49.85	61.43	-
	< MW	16.16	17.82	11.66
	= MW	14.71	17.75	9.24
	> MW	69.13	64.43	79.1

Source: Authors' computations from PNAD data.

Table 14: Positive income from work and minimum wage—income with imputation and only imputed income (2015); statistics by occupational status

Income from main job		With imputation			Only imputed
2015	% Positive	< MW	= MW	> MW	= MW
Total	43.72	17.46	10.31	72.22	6.57
Occupational status					
Agricultural employee	99.57	36.31	15.88	47.81	14.69
Domestic worker	99.79	43.14	17.41	39.45	23.47
Formal worker	100	1.53	11.24	87.22	7.81
Informal worker	100	34.59	12.70	52.71	12.07
Self-employed	99.98	34.37	5.33	60.29	4.73
Employer	99.97	2.89	2.23	94.88	2.02
Public servant	99.99	4.13	11.87	84	3.31

Source: Authors' computations from PNAD data.

4.3 Income distribution in 2015

The process of imputing individual incomes generally resulted in higher average incomes and slightly higher levels of inequality than those previously estimated for 2001 and 2015 without imputation. In the case of the monthly income of all jobs in 2015, detailed here as an example, Table 15 shows that the estimated mean of the imputed values was 36.8 per cent higher than the mean of unimputed values, resulting in a post-imputation mean 0.6 per cent higher than the pre-imputation mean. The standard deviation of the imputed values was 61.7 per cent higher than for the unimputed values, and the standard deviation of all values after imputation was 1.5 per cent higher than for the pre-imputation values.

Table 15: Summary statistics of labour income before and after imputation (2015)

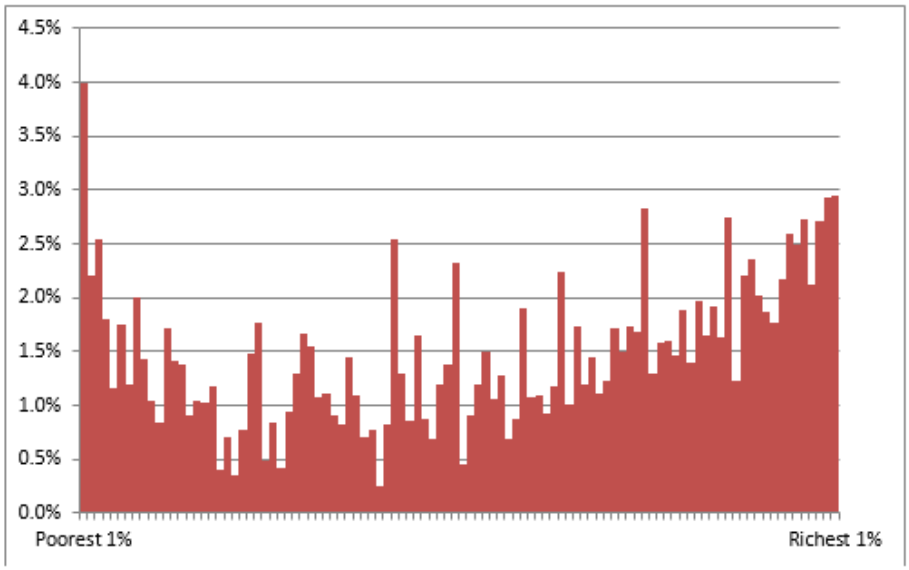
	Obs	Pop	Mean	Std. Dev.	Min	Max
All pre-imp	152,687	87,672,906	1,849	2,887	5	200,000
Imputed	2,276	1,565,545	2,531	4,669	15	120,000
All post-imp	154,963	89,238,451	1,861	2,929	5	200,000

Source: Authors' computations from PNAD data.

The increase in dispersion occurred even though the method, by construction, prevented the imputed values from exceeding the original minima and maxima. Figure 3 shows that the imputed

values were more frequent in the hundredth extremes of the final distribution, and rarer in the middle of the distribution, which is due to the profile of the receivers' predicted values.

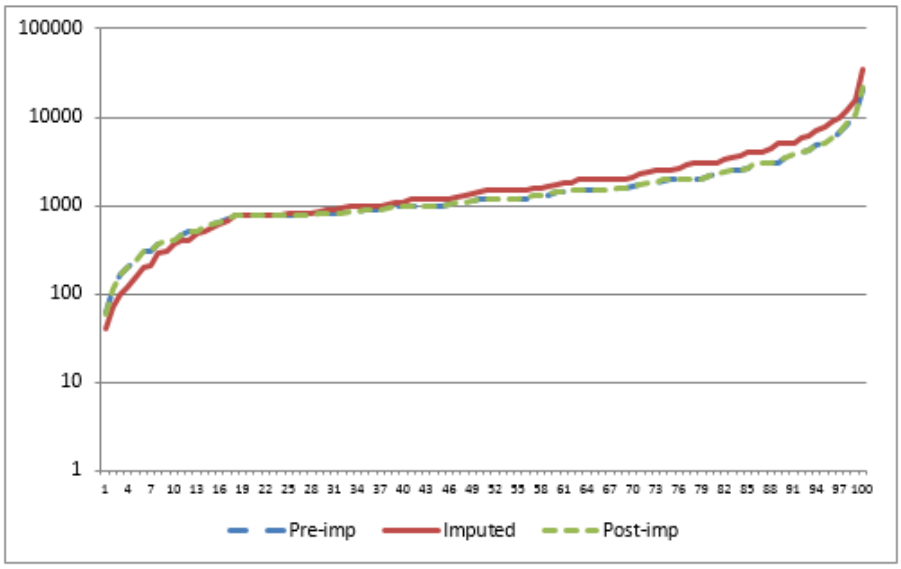
Figure 3: Percentage of imputed labour incomes by hundredth (2015)



Source: Authors' computations from PNAD data.

Figure 4 compares, on a logarithmic scale, the average values of the hundredths (or centile groups) of the labour income before and after imputation and for the receivers of imputed values. It is not possible to distinguish the initial and final distributions in this graph, but the distribution of the imputed values stands out from the others, which highlights the changes caused by the imputation. The three lines overlap from the 19th to the 23rd hundredth in R$788, the exact value of the national minimum wage in 2015. However, the imputed incomes are lower until the 18th and higher from the 24th hundredth. Although the minimum and maximum values of the imputed distribution are not as extreme as the originals (Table 15), the 1st and 100th hundredth means are more extreme in the imputed distribution than in the unimputed and final distributions (Figure 4).

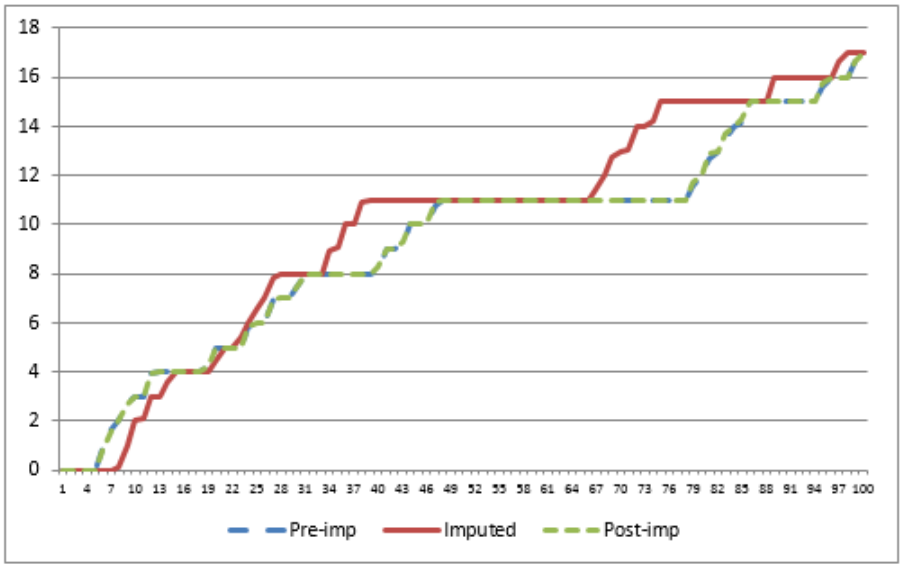
Figure 4: Distributions of labour income before and after imputation (2015)



Source: Authors' computations from PNAD data.

To indicate how the imputation receivers' profile determines this pattern, Figure 5 shows the hundredths of the study years distribution in the same three groups shown in the previous chart. Education is just one of the dimensions used to predict earned income, but the graph helps us to see why the imputed incomes focused more on extreme values than the original ones. Apparently, the propensity that a labour income exists without informing its value is higher among workers with extreme levels of schooling and, therefore, with extreme values of expected income. In addition, mean schooling of receivers (10 years) was higher than that of labour income donors (9.2 years).

Figure 5: Distributions of years of schooling before and after labour income imputation (2015)



Source: Authors' computations from PNAD data.

As a result, labour income inequality was slightly increased by imputation in the 2015 data. Table 16 shows that there was some increase in the point estimates of generalized entropy and Gini and Atkinson indices.

Table 16: Inequality indices of labour income before and after imputation (2015)

	Pre-imp.	Post-imp.
Percentile ratios		
p90/p10	8.621	8.750
p90/p50	2.917	2.917
p10/p50	0.338	0.333
p75/p25	2.538	2.538
GE and Gini indices		
GE(-1)	0.752	0.763
GE(0)	0.430	0.434
GE(1)	0.500	0.505
GE(2)	1.218	1.238
Gini	0.486	0.488
Atkinson indices		
A(0.5)	0.204	0.205
A(1)	0.350	0.352
A(2)	0.601	0.604

Source: Authors' computations from PNAD data.

Table 17 summarizes the results for the mean values and the Gini coefficients of several variables before and after the imputation for 2015. The two least frequent individual incomes (donations and permanence bonus) were the only ones whose Gini coefficients were reduced by the

imputation. The permanence bonus was also the only individual income whose average was reduced by the imputation. The imputation of individual incomes increased per capita household income by 1.5 per cent and its Gini coefficient by 0.002.

Table 17: Mean and Gini of imputed variables before and after imputation (2015)

	Pre-imp.	Post-imp.	Mean Var. %	Pre-imp.	Post-imp.	Gini Diff.
Labour	1,849	1,861	0.6%	0.486	0.488	0.002
Retirement	1,450	1,462	0.9%	0.386	0.389	0.003
Pension	1,036	1,038	0.2%	0.402	0.402	0.000
Other	312	313	0.2%	0.492	0.493	0.001
Received rent	1,341	1,345	0.3%	0.529	0.529	0.000
Donation	574	579	0.9%	0.558	0.557	-0.001
Permanence bonus	983	952	-3.2%	0.457	0.452	-0.005
Per capita household	1,057	1,073	1.5%	0.514	0.517	0.002
Paid rent	602	629	4.5%	0.364	0.417	0.053

Source: Authors' computations from PNAD data.

The results of the paid rent imputation have different patterns from the others and should be interpreted with caution. Table 18 shows that, for this variable, the potential donors are a minority (17.7 per cent) and the receivers of imputed values are the majority (82.3 per cent), which amplifies the impact of the imputation. According to the previous table, the imputation increases the average value by 4.5 per cent and the Gini coefficient by 0.053.

Table 18: Mean and Gini of paid rent post-imputation by household ownership status (2015)

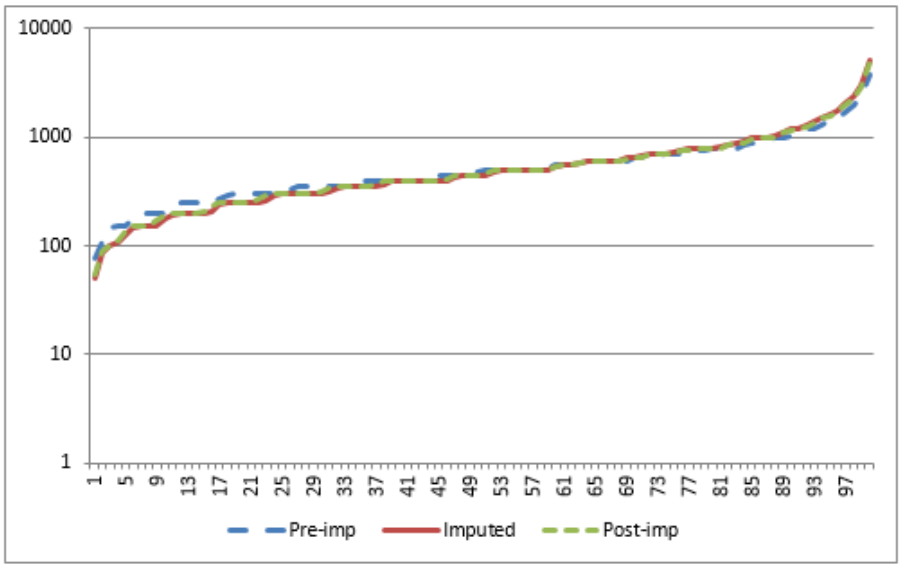
	% Pop.	Mean	Gini
Rented (potential donors)	17.7%	602	0.364
Rented (receivers)	0.1%	820	0.415
Owned—paid for	69.9%	640	0.429
Owned—still paying	4.9%	823	0.380
Ceded—by employer	1.9%	409	0.407
Ceded—not by employer	5.1%	463	0.378
Other	0.3%	536	0.418
All post-imp	100.0%	629	0.417

Source: Authors' computations from PNAD data.

The imputed values of this variable can be interpreted as the expected market prices for the monthly rent of each property, if they were rented. The imputation receiving group is heterogeneous and includes households that reported rent but did not report its value (0.1 per cent of the total after imputation). The characteristics of these households that did not respond to the rental value indicate an average market value 36.3 per cent higher than the average informed values. The properties with the lowest average market value imputed are those provided ("ceded") by the employers of the residents. Those with the highest average value are the owned ones that are still being paid for.

Figure 6 shows how the imputation of paid rent generated a more unequal distribution than the original one. Since the imputation receivers are the majority in the cases of this variable, the curve referring to all post-imputation values is close to the imputed values curve and is clearly distinguishable from the original curve.

Figure 6: Distributions of paid rent before and after imputation (2015)



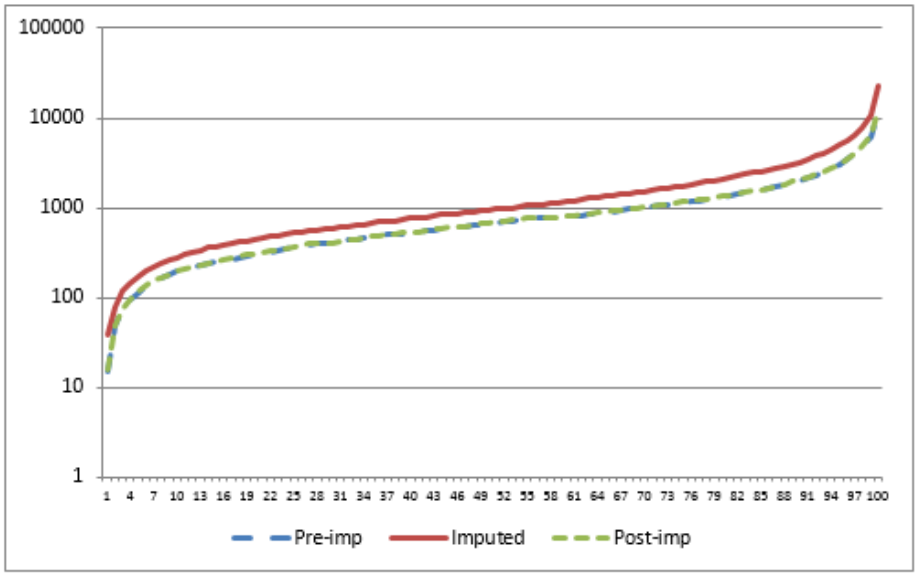
Source: Authors' computations from PNAD data.

5 Results of per capita income imputation procedures: levels and changes

5.1 2015 levels and 2001–15 changes

Household per capita income-based welfare indicators drive most of the social policy debates—in particular those related to poverty and inclusive growth. The change in the distribution of per capita household income by the process of imputation of individual incomes can be visualized in Figure 7. All hundredths had their averages increased, resulting in an overall mean 1.5 per cent higher. The point estimate of the Gini coefficient increased by only 0.00249, from 0.51438 to 0.51687.

Figure 7: Distributions of per capita household income before and after imputation (2015)



Source: Authors' computations from PNAD data.

Table 19 shows that 2.5 per cent of the (weighted) sample had the per capita household income altered by the imputation of individual incomes. Most of the imputations in this derived variable occurred on original missing value codes (2.4 per cent of the population). The original positive values that were increased by the imputation of some individual income accounted for only 0.1 per cent of the population. Only six observations with original per capita household income equal to zero were replaced by imputed values (0.003 per cent of the population). Per capita household income remained zero for 0.4 per cent of the population and not applicable for 0.2 per cent.

Table 19: Transitions of per capita household income with imputation (2015)

Transition	Obs.	Pop.	%Pop.
Unchanged zero	1,398	883,023	0.4%
Unchanged positive	347,227	198,554,677	96.9%
Unchanged NA	592	317,051	0.2%
Imputed on zero	6	5,246	0.0%
Increased positive value	253	155,357	0.1%
Imputed on code for missing	7,428	4,944,051	2.4%
Total	356,904	204,859,405	100.0%

Source: Authors' computations from PNAD data.

The year 2001 marks the beginning of a period of successive reductions in the inequality of per capita household income observed in the PNAD. Therefore, it is interesting to highlight the effects of the imputation process on the variation of the estimated inequality between 2001 and 2015. To compare the data for 2001 and 2015, however, it is necessary to exclude from 2015 the information on the rural areas of the North region of Brazil, which only began to be covered by the PNAD in 2004. In 2015, 3.4 per cent of the observations and 2.1 per cent of the weighted sample lived in the rural areas of the North region.

Table 20 compares the effects of imputation on labour income and per capita household income in 2001 and 2015 without rural North. Imputation mitigated income growth between 2001 and 2015 and practically did not change the reduction in inequality observed in the PNAD.

Table 20: Transitions of per capita household income with imputation (2015)

Year	Income	Mean			Gini		
		Pre-imp.	Post-imp.	Var. %	Pre-imp.	Post-imp.	Diff.
2001	Labour	595	602	1.1%	0.566	0.568	0.003
	Per capita household	297	305	2.5%	0.594	0.597	0.003
2015	Labour	1,863	1,875	0.7%	0.485	0.487	0.002
	Per capita household	1,070	1,086	1.5%	0.513	0.515	0.002

Source: Authors' computations from PNAD data.

The increase in average incomes caused by imputation is higher in 2001 than in 2015. Therefore, after imputation in these two years, real growth in labour income decreases from an annual average of 1.52 per cent to 1.48 per cent and the per capita household income real growth decreases from 2.53 per cent to 2.46 per cent. The Gini index point estimates of labour income and per capita household income increase by 0.003 in 2001 and 0.002 in 2015. Thus, the Gini index fall of both indicators between 2001 and 2015 becomes only 0.001 more intense and can be considered unchanged.

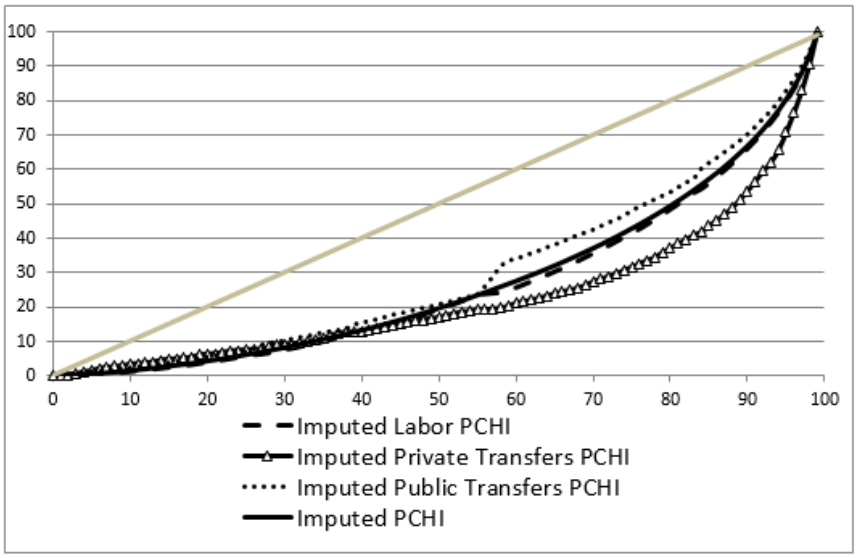
5.2 Concentration curves

Concentration curves are a representation that bears similarities to the Lorenz curve. While the latter refers to the distribution of a single variable throughout the population, the former is constructed from the distribution of two variables in the population. In fact, the Lorenz curve can be understood as a particular case of the concentration curve where the variable used in the

ordering of the population and the output variable coincide. Similarly, the correspondence between the Gini index and the Lorenz curve also appears in the relationship between the concentration curve and the concentration index. One key difference is that the Gini varies between 0 and 1 while the concentration index varies between -1 and 1. If a certain attribute is more directed to the poor—for example, conditional cash transfers—then the indicator is negative.

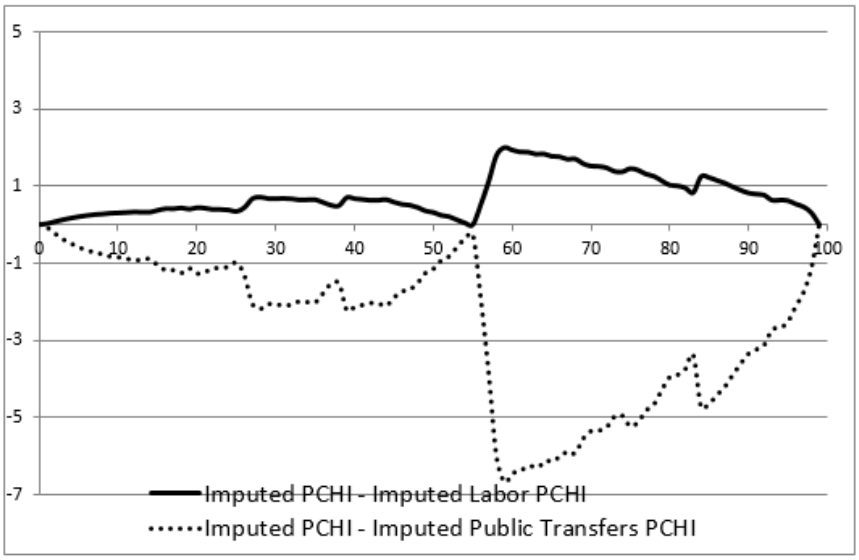
Figure 8 presents the concentration curves of the whole 2015 sample after imputation ordered by total per capita income. It also separates labour income from other private transfers and the total public transfers. We note in public transfers some mass accumulation just before the 60th percentile, which is somewhat surprising because we are working with a more aggregate concept that tends to smooth out these kinks. This impression is confirmed by Figure 9, which presents the differences between concentration curves with respect to total income. The discontinuity close to the 60th percentile is observed not only for public transfers but also for labour income.

Figure 8: Concentration curves ordered by imputed per capita household income (2015)



Source: Authors' computations from PNAD data.

Figure 9: Concentration curves differences in relation to imputed per capita household income (2015)



Source: Authors' computations from PNAD data.

5.3 Poverty

Usual inequality indexes do not give substantial weight to the lower tail of the income distribution. For example, any marginal increase below the 75th percentile in Brazil, according to PNAD 2014, would lower the Gini index. The same statistics for the Theil-L and Theil-T are the 74th and 87th percentiles, respectively (Hecksher et al. 2017). Implicitly, we care relatively little about the poorest segments of our society. New metrics such as the shared prosperity indicator, incorporated in the recent Sustainable Development Goals (SDGs), target the bottom 40 per cent of the distribution. We go one step further and study empirically here the behaviour of inequality in terms of poverty alleviation objectives. In this section, we use standard poverty measures to address the anti-inequality impacts involved in different measures used.

$$P^{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z - Y_i}{Z}\right)^{\alpha} I(Y_i \leq Z) \qquad (1)$$

where
N = population size
Z = poverty line
Y_i = income level for the individual i
a = poverty aversion degree

Brazil adopted an official extreme poverty line of around US$1.25 per day using older purchasing power parity (PPP), which is perhaps too low for the country's level of income. We perform different exercises using different international poverty lines, recently raised by the use of the new PPP estimates, applied to different measures (P^0, P^1), with and without imputation for missing values, as shown in Tables 21 and 22. We focus here on the P^1 measure using the US$3.2 a day PPP poverty line, under which poverty with imputation is 16.8 per cent lower, or 0.9 percentage points lower, in 2015. Poverty differences across time are much smaller, not exceeding 0.4 percentage points. In our benchmark scenario, the difference amounts to 0.1 percentage point.

Table 21: Proportion of poor (P0) with and without imputation (1995, 2001, 2003, 2009, 2014 and 2015) (%)

PPP a day Poverty line	P0 (without imputation)						P0 (with imputation)					
	1995	2001	2003	2009	2014	2015	1995	2001	2003	2009	2014	2015
US$1.25	8.5	8.0	7.7	3.9	2.3	2.8	9.1	7.8	7.6	3.8	2.2	2.8
US$1.9	17.6	16.8	16.5	9.6	7.3	6.8	16.3	14.7	14.6	6.6	3.7	4.7
US$3.2	32.0	29.9	30.5	18.0	12.2	12.6	30.8	28.0	28.7	15.2	8.7	10.5
US$4	39.1	37.8	38.3	24.0	16.4	17.1	37.9	36.0	36.7	21.3	13.0	15.0
US$5.5	50.8	49.1	50.1	34.7	24.6	25.6	49.7	47.6	48.7	32.1	21.3	23.7

Source: Authors' computations from PNAD data.

Table 22: Poverty gap (P1) with and without imputation (1995, 2001, 2003, 2009, 2014 and 2015) (%)

PPP a day Poverty line	P1 (without imputation)						P1 (with imputation)					
	1995	2001	2003	2009	2014	2015	1995	2001	2003	2009	2014	2015
US$1.25	3.9	3.9	3.6	2.1	1.1	1.3	4.5	3.8	3.5	2.0	1.1	1.3
US$1.9	7.7	7.4	7.0	4.5	3.4	3.2	7.5	6.4	6.1	3.1	1.7	2.2
US$3.2	14.4	13.7	13.5	7.5	4.9	5.2	14.3	12.7	12.7	6.3	3.5	4.3
US$4	18.6	17.6	17.5	9.9	6.3	6.8	18.4	16.7	16.7	8.8	5.0	6.0
US$5.5	25.6	24.5	24.7	14.8	9.7	10.6	25.5	23.7	24.0	13.7	8.4	9.8

Source: Authors' computations from PNAD data.

5.4 Growth–inequality decomposition

We analyse the impacts of inequality on poverty changes using a standard Datt-Ravallion decomposition into growth (G()) and redistributive (I()) components to assess their relative roles, ignoring the residual component (R()), which tends to be small.

$$P_{t+n} - P_t = G(t,t+n;r) + I\ (t,t+n;r) + R(t,t+n;r)$$

$$G(t,t+n;r) \equiv P(Z/\mu_{t+n}, L_r) - P(Z/\mu_t, L_r);$$

$$I(t,t+n;r) \equiv P(Z/\mu_r, L_{t+n}) - P(Z/\mu_r, L_t)$$

The fall of poverty in percentage points increases with the level of the poverty line used, which is not a surprising fact because poverty levels are also necessarily higher. But the proportional absolute variation is smaller (not monotonically) for higher poverty lines ranging from a fall of -65.3 per cent to -58.7 per cent in the case of the poverty gap (P^1). Using as a benchmark the intermediary US$3.2 a day line, the fall in the 2001–15 period amounted to -8.4 percentage points or -66.3 per cent. This means that poverty fell more than the expected 50 per cent in the United Nations' first Millennium Development Goal in much less than the 25-year period.

The most important dimension analysed here is the relative share of poverty fall explained by inequality, ranging from 45.9 per cent to 56.3 per cent in the case of P^1 with imputation (Table 24). In the case of our intermediary US$3.2 a day line, 45.9 per cent of total P^1 fall was explained by the inequality component.

The imputation process slightly reinforced the role of inequality for the P^1 fall as defined by the three highest poverty lines, and did the opposite when we considered the two lowest lines. In the case of P^0, the role of inequality was softened by the imputation only for the lowest poverty line, while it was increased for the other four lines (Table 23). These effects were small and did not change the predominant driver (inequality or growth) in any of the analysed cases.

Table 23: Poverty variation between 2001 and 2015 with and without imputation

P0—Without imputation	Inequality—Effect	Growth—Effect	Total
Poverty—US$1.25 new PPP line	40.41%	59.59%	-5.2
Poverty—US$1.9 new PPP line	39.88%	60.12%	-9.9
Poverty—US$3.2 new PPP line	45.38%	54.62%	-17.3
Poverty—US$4 new PPP line	45.34%	54.66%	-20.7
Poverty—US$5.5 new PPP line	47.72%	52.28%	-23.5
P0—With imputation	Inequality—Effect	Growth—Effect	Total
Poverty—US$1.25 new PPP line	39.56%	60.44%	-5.0
Poverty—US$1.9 new PPP line	40.62%	59.38%	-10.0
Poverty—US$3.2 new PPP line	46.35%	53.65%	-17.5
Poverty—US$4 new PPP line	46.48%	53.52%	-21.0
Poverty—US$5.5 new PPP line	49.10%	50.90%	-23.9

Source: Authors' computations from PNAD data.

Table 24: Poverty variation between 2001 and 2015 with and without imputation

P1—Without imputation	Inequality—Effect	Growth—Effect	Total p.p
Poverty—US$1.9 new PPP line	47.89%	52.11%	-4.2
Poverty—US$3.2 new PPP line	45.49%	54.51%	-8.5
Poverty—US$4 new PPP line	46.29%	53.71%	-10.8
Poverty—US$5.5 new PPP line	46.01%	53.99%	-13.9
P1—With imputation	Inequality—Effect	Growth—Effect	Total
Poverty—US$1.25 new PPP line	56.34%	43.66%	-2.5
Poverty—US$1.9 new PPP line	46.74%	53.26%	-4.2
Poverty—US$3.2 new PPP line	45.87%	54.13%	-8.4
Poverty—US$4 new PPP line	46.72%	53.28%	-10.7
Poverty—US$5.5 new PPP line	46.90%	53.10%	-13.9

Source: Authors' computations from PNAD data.

5.5 Imputed rent

Up to this point we have discussed the differences between imputed incomes and original values in the database. We now add imputed rent estimates, which is expected to affect the substantive results found. Poverty with imputed rent estimates is lower (Table 25). For example, in 2015 using the U$3.2 a day PPP line the proportion of poor (P^0) is 40 per cent lower, while the poverty gap (P^1) is 48.9 per cent lower. The amounts of P^0 and P^1 falls between 2001 and 2015 fall, respectively, from 17.3 to 14.0 (Table 26) and from 8.4 to 5.8 percentage points (Table 27) using imputed rents. Using Datt-Ravallion-type decomposition, the share of poverty fall explained by inequality reduces from 45.87 per cent to 30.38 per cent.

Table 25: Proportion of poor (P^0) and poverty gap (P1) (2001, 2003, 2009, 2014 and 2015): imputed per capita household income + imputed rent (%)

PPP a day Poverty line	P^0					P^1				
	2001	2003	2009	2014	2015	2001	2003	2009	2014	2015
US$1.25	4.2	4.2	1.9	0.7	0.9	1.6	1.6	0.7	0.3	0.4
US$1.9	9.1	9.3	4.0	1.8	2.3	3.4	3.4	1.5	0.6	0.8
US$3.2	20.6	21.6	10.5	5.3	6.3	8.2	8.4	3.8	1.8	2.2
US$4	27.3	28.6	15.5	8.3	9.7	11.4	11.8	5.7	2.8	3.4
US$5.5	39.0	40.7	25.1	15.3	17.2	17.5	18.2	9.8	5.4	6.2

Source: Authors' computations from PNAD data.

Table 26: Poverty variation between 2001 and 2015 for imputed per capita household income + imputed rent (%)

P^0	Inequality—Effect	Growth—Effect	Total p.p.
Poverty—US$1.25 new PPP line	20.52%	79.48%	-3.1
Poverty—US$1.9 new PPP line	26.65%	73.35%	-6.5
Poverty—US$3.2 new PPP line	35.81%	64.19%	-14.0
Poverty—US$4 new PPP line	39.10%	60.90%	-17.1
Poverty—US$5.5 new PPP line	38.69%	61.31%	-21.3

Source: Authors' computations from PNAD data.

Table 27: Poverty variation between 2001 and 2015 for imputed per capita household income + imputed rent (%)

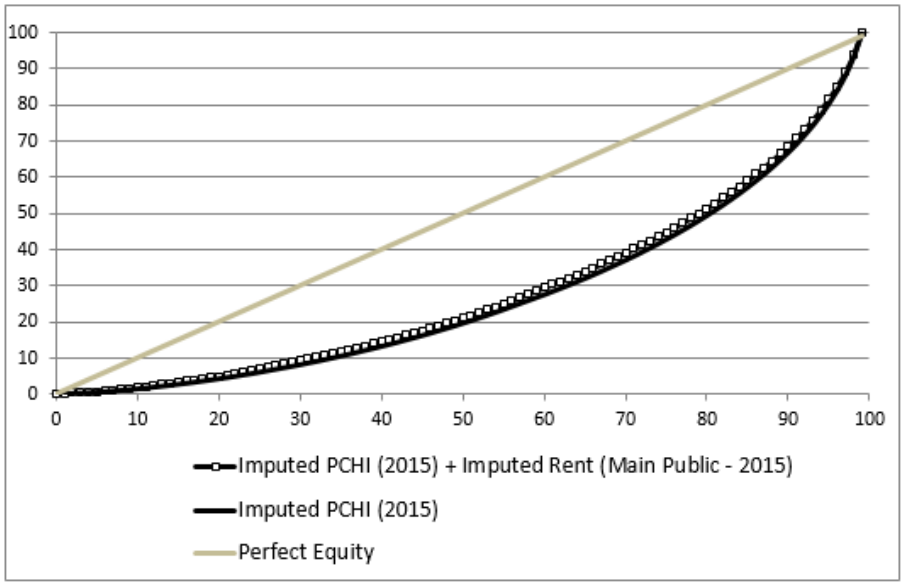
P[1]	Inequality—Effect	Growth—Effect	Total p.p.
Poverty—US$1.25 new PPP line	19.68%	80.32%	-1.2
Poverty—US$1.9 new PPP line	22.45%	77.55%	-2.5
Poverty—US$3.2 new PPP line	30.38%	69.62%	-5.8
Poverty—US$4 new PPP line	33.91%	66.09%	-7.8
Poverty—US$5.5 new PPP line	36.64%	63.36%	-11.0

Source: Authors' computations from PNAD data.

The concentration curves for these two concepts (with and without imputed rent) are shown in Figure 10, in which imputed rent adds to equality. On the other hand, Figure 11 shows that imputed rent reduces the fall of inequality observed between 2001 and 2015, since the curves are closer.

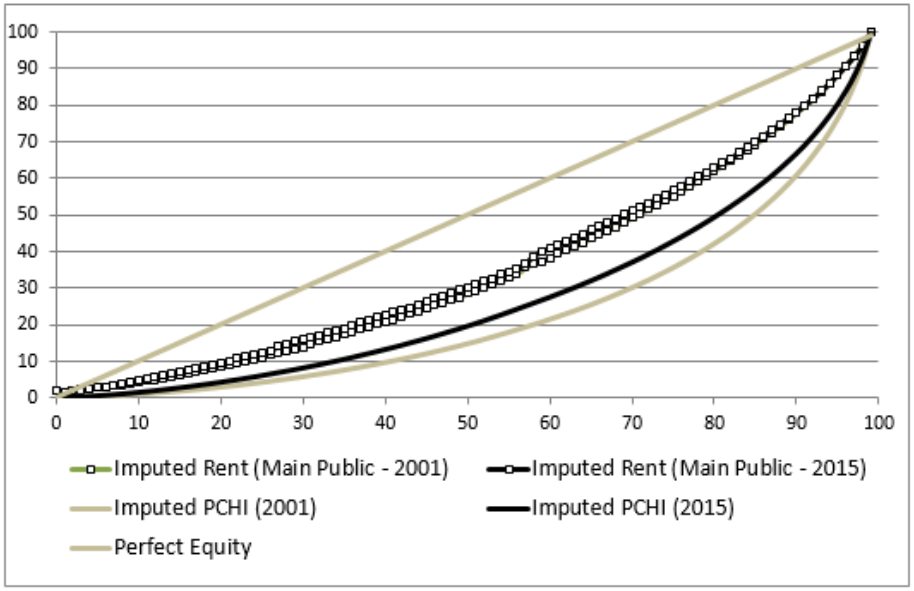
Table 28 presents the evolution of the Gini index for three different concepts of mean per capita household income: without income imputation, with income imputation, and with income imputation and an imputed rent value. The table reveals that both imputation processes reduce Gini's level for all five years in the sample, without affecting substantially the percentage variation between 2001 and 2015. In short, imputing rents does not affect the inequality trends observed from PNAD household surveys.

Figure 10: Concentration curves ordered by imputed per capita household income (2015)



Source: Authors' computations from PNAD data.

Figure 11: Concentration curves ordered by imputed per capita household income (2001 and 2015)



Source: Authors' computations from PNAD data.

Table 28: Gini of different per capita household income concepts (2001, 2003, 2009, 2014 and 2015)

Mean per capita household income concepts—Gini	2001	2003	2009	2014	2015	% Change (2001–15)
Without imputation	0.60266	0.58947	0.55271	0.53109	0.5236	-13.11%
With imputation	0.59697	0.58333	0.54110	0.51569	0.5150	-13.72%
With imputation + imputed rent	0.56800	0.55842	0.52043	0.49507	0.4937	-13.08%

Source: Authors' computations from PNAD data.

6 Final remarks

The primary objective of this paper was to develop an imputation procedure for each source of incomes to people and households with missing or zero values on the microdata databases of the national household survey (PNAD), collected by the IBGE. We also analysed the impact of this income imputation on income distribution estimates in Brazil and its variations over time.

The public microdata from other IBGE household surveys—Censo Demográfico, Pesquisa Mensal de Emprego (PME), Pesquisa de Orçamentos Familiares (POF), and PNAD Contínua trimestral—provided income variables with imputation. It was intended, on this project, to combine some strengths of the techniques already implemented on these other IBGE databases and, with a new method, apply imputations to various editions of the annual PNAD.

Broadly speaking, we developed an approach that considers the changing predictive power of the available variables through time, between groups (e.g. by occupational status), and over different income sources, preserving discontinuities and high-frequency values as the exact minimum wage. The method has the econometric intuition of the deterministic model-based regression imputation but avoids its tendency to reduce variance. For that, it applies a stochastic selection of one donor from a set of neighbours whose predicted incomes are nearest to the predicted income of the receiver observation.

After the imputing procedure, we analysed its impacts on levels and variations of average income and income inequality indexes. As a result of the project, we also produced, from the PNAD's original public microdata, a database with imputed incomes and imputation marks. The approach described was applied to the PNAD data for 2001 and 2015. All individual incomes and paid rent were imputed. Once individual incomes were imputed, all relevant derived variables such as per capita household income were calculated and added to the data sets.

The process of imputing individual incomes generally resulted in higher average incomes and slightly higher levels of inequality than the ones previously estimated for 2001 and 2015 without imputation. In the case of the monthly income of all jobs in 2015, detailed here as an example, the estimated mean of the imputed values was 36.8 per cent higher than the mean of unimputed values, resulting in a post-imputation mean 0.6 per cent higher than the pre-imputation mean. The standard deviation of all values after imputation was 1.5 per cent higher than the pre-imputation values. There was some increase in the point estimates of generalized entropy and Gini and Atkinson indices.

We establish connections between our imputation procedures and Brazilian income policies. In particular, minimum wages as an institutional floor to social security payments or their widespread use among informal employees are preserved in the methodology developed in this paper, which takes advantage of the separate imputation methodology applied to different income sources and different working classes in the case of labour income components.

The increase in mean incomes caused by imputation is higher in 2001 than in 2015. Therefore, after imputation in these two years, real growth in labour income decreases from an annual average of 1.52 per cent to 1.48 per cent and the growth of per capita household income decreases from 2.53 per cent to 2.46 per cent. The Gini index point estimates of labour income and per capita household income increase by 0.003 in 2001 and 0.002 in 2015. Thus, the Gini index fall of both indicators between 2001 and 2015 becomes only 0.001 more intense and can be considered unchanged.

In 2015, 2.9 per cent of the (weighted) sample had the per capita household income altered by the imputation of individual incomes. Most of the imputations in this derived variable occurred on original missing value codes for individual incomes (2.4 per cent of the population). The original positive values that were increased by the imputation of some individual income accounted for only 0.1 per cent of the population and the original zero household income values that were directly imputed were 0.4 per cent.

Poverty levels were reduced by the imputation procedure in more than 90 per cent of the combinations between poverty measures, poverty lines, and years. However, poverty changes—at least in the 2001–15 period—were much less affected. Usual inequality indexes do not give substantial weight to the lower tail of the income distribution. We addressed this issue by analysing the relative impact of inequality on poverty changes. The share of poverty fall explained by the distributive component was not affected by the imputation procedures.

We took the methodology one step further and applied it to separate rent imputation procedures. The procedures developed here showed some impact on the level of poverty and inequality. Although imputed rent did reduce the relative importance of the inequality component of poverty reduction, it did not affect inequality trends as measured by the Gini coefficient.

References

Breiman, L., et al. (1993). *Classification and Regression Trees*. New York: Chapman & Hall.

De Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

Hecksher, M., P.L.N. Silva, and C. Courseil (2017). 'Preponderância Dos Ricos Na Desigualdade De Renda No Brasil (1981-2016): Aplicação Da J-Divergência A Dados Domiciliares E Tributários'. Dissertação De Mestrado ENCE/IBGE.

IBGE (1981). 'Metodologia Da Pesquisa Nacional Por Amostra De Domicílios Na Década De 70'. Rio de Janeiro: IBGE.

IBGE (2003). 'Metodologia Do Censo Demográfico 2000'. Rio de Janeiro: IBGE. Available at: https://servicodados.ibge.gov.br/Download/Download.ashx?http=1&u=biblioteca.ibge.gov.br/visualizacao/livros/liv5295.pdf (accessed 4 October 2018).

IBGE (2007). 'Pesquisa Mensal De Emprego. 2a. Edição'. Rio de Janeiro: IBGE.

IBGE (2016a). 'Metodologia Do Censo Demográfico 2010, 2a Edição'. Rio de Janeiro: IBGE. Available at: https://servicodados.ibge.gov.br/Download/Download.ashx?http=1&u=biblioteca.ibge.gov.br/visualizacao/livros/liv81634.pdf (accessed 4 October 2018).

IBGE (2016b). 'Pesquisa Nacional Por Amostra De Domicílios 2007/2015'. Rio de Janeiro: [S.N.]. Available at: https://servicodados.ibge.gov.br/Download/Download.ashx?http=1&u=biblioteca.ibge.gov.br/visualizacao/livros/liv98887.pdf (accessed 4 October 2018).

Neri, M.C., G. Gonzaga, and J.M. Camargo (2001). 'Salário Mínimo, Efeito Farol E Pobreza'. *Revista de Economia Política*, 21(2): 78–90.

R Core Team, The (n.d.). *R: A Language and Environment for Statistical Computing*. Available at: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (accessed 25 September 2018).

Statistics Canada (2007). *Social Survey Methods. Canceis Version 4.5 Workbook*. Ottawa.