Estimating the Global Income Distribution Exploring a Multiple Imputations approach to correct for missing high incomes and account for statistical uncertainty

Master Dissertation Research Project - AMSE M2/ETE*

Mathias Silva
Supervisor: Michel Lubrano

June 10, 2019

*All data files, R program files, and accompanying data appendix containing all country-level estimates can be publicly accessed on this projects online GitHub repository https://github.com/mathiasgsilva/GID_MI

Contents

1 Introduction

Since the early 1980's the boom of Information and Communication Technologies (ICT), the dissolution of the Soviet Union, the Chinese economic reform, the creation of the World Trade Organization, and the market liberalization policies of the Reagan-Thatcher era all constitute, amongst others, key changes in the global scenario which have potentially restructured many types of economic relationships. In particular, very different trends of economic growth and trade patterns have been experienced by different regions in the world, with exemplary cases such as China which has experienced an important growth spurt with big changes in its insertion in international markets.

One particularly relevant dimension in which this phenomenon of globalization is expected to have caused significant changes is that of the global distribution of incomes and global income inequality, as reviewed in Kanbur (2015). Although there is empirical evidence of decreases in the rate of extreme poverty at global and regional levels in recent decades, the geography and magnitudes of these trends vary a lot across the globe (Chen and Ravallion, 2004, Sala-i-Martin, 2006, Chen and Ravallion, 2010). When trying to empirically study the distribution of incomes at the global level and the distribution of the gains from economic growth in recent decades across the global population, the conclusions are however less clear.

A variety of methodological approaches have been explored in the literature, estimating levels and changes of global income inequality across the period ranging from the late 1980's to the early 2010's which imply very heterogeneous conclusions regarding the dynamics of the global distribution of incomes during these recent decades. One particularly popular of these results is what has been termed the 'elephant' curve of global inequality and growth (Alvaredo et al., 2018, Lakner and Milanovic, 2016), which suggests that over this period the growth rates of the different parts of the global income distribution have been such that the central deciles around the median of the

distribution and the very top fractiles have grown at faster rates than the deciles between the median and these top fractiles. This has drawn particular attention to the role of the highest incomes, and particularly to the global top 1% of incomes, on the global dynamics of inequality as the 'elephant' curve implies that these have been particularly benefited by the economic changes brought about by globalization.

The aim of this dissertation is to explore a combination of statistical and econometric methods which have not been exploited in the previous literature on the Global Income Distribution (GID) with the intention of contributing new tools for exploring the question of how the phenomenon of globalization in the recent decades has affected the interpersonal distribution of incomes across the globe and particularly who have been the most and the least benefited by these changes. Using a sample of countries from the LM-WPID dataset of Lakner and Milanovic (2016) covering all regions of the world and presenting grouped income distribution statistics at the country-year level, I explore a Multiple Imputations framework to deal with two issues that can provide further insight on the robustness of previous results. These are the problems of missing high incomes in survey data and of measuring the statistical uncertainty associated to the estimated global inequality measures. These issues can provide particular insight on the uncertainty concerning the changes in the very top fractiles of the global income distribution along the period.

The remainder of the document is organized as follows. The following section reviews the recent literature on global income inequality and the methodological approaches explored to study the GID. Section 3 introduces the data sources available to study the income distribution at the global level covering a reasonable portion of the global population's incomes and their common features, along with the particular dataset I exploit for my empirical results. Section 4 presents the details of the Multiple Imputations framework I explore to estimate the GID while dealing simultaneously with the issues of missing high incomes and estimating the statistical uncertainty

of the estimated global inequality statistics. Section 5 discusses the main results obtained for the 1993-2008 period, with particular attention to the role of the methods I explore. Finally, section 6 presents the main limitations and conclusions of this dissertation, along with some lines for future research on the topic.

2 Previous studies

Milanovic (2011, chapter 1) and Anand and Segal (2008) propose four notions of what is commonly and vaguely referred to as world income inequality, characterized as follows.

1. *Type zero inequality* refers to inequality across countries in their aggregate incomes,

2. *type one inequality* refers to inequality across countries in their *per capita* incomes, this is a measure of un-weighted international inequality,

3. *type two inequality* refers to inequality across individuals when all individuals in a single country are assigned the per capita income of that country, which constitutes a measure of population-weighted international inequality,

4. and finally *type three inequality*, properly referred to as global interpersonal inequality, refers to inequality across individuals when each individual is assigned his own specific income.

Chronologicaly, the first two concepts of inequality under this taxonomy were the first to be explored by the early economic literature on macroeconomic growth and theories of convergence. The lack of detailed data and a systematic literature on within-country income distributions before the 1970's limited for a certain period the development of approaches to gain insight into other types of world income inequality. Some clear references of the type of methods and inquiries of this first concepts of inequality, as listed in Milanovic (2002), are Theil (1979), Theil and Seale (1994), and Podder

(1994), but also Barro and Sala-i-Martin (1991), Quah (1997), and Jones (1997) amongst others.

Berry et al. (1983), Chotikapanich et al. (1997), Bourguignon and Morrisson (2002), and Milanovic (2002) all serve as a good example of the transition in the literature on world income inequality towards concepts that contemplate within-country distribution in the study of international inequality. This transition was largely led by the availability of income distribution statistics at the country level with both the rise of household income and/or consumption surveys and of recopilatory and homogenizing projects working on the systematization of household survey data at the international level such as the Luxembourg Income Study in 1983, the Deininger and Squire (1996) and subsequent UNU-WIDER's World Income Inequality Database (WIID), and the World Bank's Research Department's *PovCalNet* project, amongst others.

The non-negligible role of within-country inequalities in the dynamics of the GID evidenced by this renewed literature set the focus of the field on the methodological approaches allowing for insight on the global interpersonal distribution of incomes conditional on the still far from ideal data sources available.

Because there's no data on individuals' income being collected at the world level, the fundamental methodological problem is that of using country-level data to estimate the GID. In particular, the main methodological problem concerns how to estimate the within-country distributions from the few summary statistics available in the data and how to aggregate these estimated distributions to obtain a distribution at the global level.

The simplest approach that allows for some inequalities at the country level is the Identical Group Income assumption (i.e., Identical Quantile Income (IQI) assumption). Under this approach a discrete distribution is used to approximate the within-country distribution, largely due to the grouped-data nature of the available datasets. The country's total population is distributed

amongst the groups defined in the data and all individuals within a group are assigned the same income (e.g., the group mean). By construction, this approximation neglects some margins of within-country interpersonal inequality as inequality within income groups is exactly zero, yet it stills captures inequalities between groups, such that this assumption is obviously more sensible when the size of the groups identified by the data are small (e.g., when working with data at the centile income shares level as compared to data at the quintile income shares level). Finally, the GID is simply the population-weighted mixture of all the constructed country-level discrete distributions.

Bourguignon and Morrisson (2002) and Milanovic (2002) are two seminal references in studying the GID exploiting the Identical Group Income assumption. The former exploits historical series on decile income shares and GDP series from National Accounts to construct country-year decile mean incomes series for the 1820-1992 period, while the latter exploits only household survey income or expenditure data to construct decile mean income series for the 1988-1993 period. Despite the heterogeneities in nature and coverage of their data, their estimates of global interpersonal inequality are very similar for the 1992/1993 common point, as seen in table 3.

Lakner and Milanovic (2016) and Anand and Segal (2017) are two more recent extensions of this approach which at the same time explore correction methods for the possible problem of under-reported high incomes in household surveys at the country level. Lakner and Milanovic (2016) explore a correction exploiting the discrepancy between aggregate incomes in household surveys and aggregate household consumption measures from National Accounts, which generally exceed the former. This excess consumption is distributed amongst the entire population either by re-scaling decile mean incomes to match per capita household consumption instead of the survey's mean income or by fitting a one-parameter Pareto distribution within the highest decile and impute the excess consumption only within this decile thus obtaining estimates of the top 5% and top 1% income shares and mean incomes. Anand

and Segal (2017), in exchange use a preceding version of the same dataset Lakner and Milanovic (2016) use but obtain top 1% income share data for a subset of countries from the World Top Incomes Database (WTID)[1]. Although the estimates on global inequality have some slight differences between both studies, particularly in what concerns the income shares of the global richest 10% and 1%, they both present higher levels of global inequality as compared to previous studies using the IQI assumption without correcting for missing high incomes in the data as presented in table 3.

Approaches more sophisticated than the Identical Group Income assumption allow for each individual in the population to be imputed his own specific income, such that inequality between individuals themselves can be estimated and not just inequality between groups of individuals. This implies considering continuous distributions for estimating the country-level income distribution using the type of grouped data provided in any of the large scale datasets used for studying the GID.

One first and rather simplistic approach is to estimate a continuous distribution using non-parametric methods, such that no explicit parametric form is assumed for the distribution of incomes. Sala-i-Martin (2006) and Hong et al. (2019) exploit the IQI assumption using quintile income share and GDP per capita data to fit a Gaussian kernel density function at the country-level assuming a same fixed bandwidth parameter for all country-years. This yields a continuous approximation to each country's income distribution from which 100 percentiles are estimated, the GID is then obtained by fitting a Gaussian kernel density on these population-weighted percentiles of all countries for a given year. In addition, Hong et al. (2019) explore a similar correction for possible missing high incomes as that of Anand and Segal (2017), annexing top 1% and 5% shares data obtained from the WTID to their quintile share data. Table 4 summarizes the results of both papers for select years. As can be seen, the results slightly differ from those obtained

[1]The WTID was the precursor project of what in January 2017 became the much larger World Wealth and Income Database (http://wid.world/).

in those previous studies exploiting the IQI assumption exclusively, yet the uncertainty regarding the overall dynamics of global inequality during the period are still evident.

Because kernel density estimators have poor performance in small samples and there exists no homogenized dataset for studying the GID with more than a few groups or quantiles per country-year, this approach is prone to many limitations when trying to make inference on this distribution. This is more so the case when all country-years are estimated using the same kernel bandwidth, which arbitrarily influences the dispersion of the estimated distribution and therefore its measures of inequality.

Two parametric approaches that present an alternative to the IQI assumption in the context of grouped data have been explored in the literature. On the one side, a parametric functional form satisfying the conditions for a Lorenz curve can be assumed at the country level and fit on quantile income shares data as is usually available in the available large scale datasets for the problem. Castillo et al. (1998) and Sarabia (2008) present the most common functional forms and estimation methods for grouped data used in the literature. Bhalla (2002) explores this approach in the context of the GID, fitting a 3-parameter Lorenz curve on quantile income shares data scaled at GDP per capita as presented in the WIID for each country-year and predicting 100 centiles from the fitted curve to obtain a discrete approximation to the income distribution which minimizes the role of the IQI assumption.

The interest in working with parametric Lorenz curves is that explicit functional assumptions on the specific distribution function of incomes for a country-year may result more restricting than the assumptions on the space of possible Lorenz curves and that grouped income shares data can be directly used in the fitting of Lorenz curves. The drawback of this approach, however, is that the implied parametric distribution function underlying a parametric Lorenz curve may not allow an analytical expression or may be inappropriate representations of the true income distribution generating the data.

The second parametric alternative implies assuming a distribution function for the distribution of incomes at the country level directly. Similar fitting procedures for grouped data as those from the previous parametric alternative can be used in the estimation of the distribution's parameters, and particularly so if the distribution admits a closed form expression for its Lorenz curve. Chotikapanich et al. (1997, 2007, 2012), Pinkovskiy and Sala-i-Martin (2009), Jorda and Niño-Zarazúa (2016), Niño-Zarazúa et al. (2017), and Jorda et al. (2018) explore alternative estimation methods for grouped data and parametric distributions commonly used in the study of income distributions.

The simplest of approaches in this latter direction is to assume a log-normal distribution for all country-year income distributions with country-year specific parameters and exploit moment conditions (e.g., matching the log-normal mean and Gini to the per capita GDP and Gini coefficient reported in the data) to estimate the two parameters of the distribution, as is done in Chotikapanich et al. (1997), Pinkovskiy and Sala-i-Martin (2009), Liberati (2015), and Niño-Zarazúa et al. (2017). More sophisticated distributional assumptions are explored in Chotikapanich et al. (2007, 2012), Jorda and Niño-Zarazúa (2016), and Jorda et al. (2018), all of which assume more general and flexible distributions (of which the log-normal is a particular case) such as the Beta-2 or the Generalized Beta of the second kind (GB2) distribution. These distributions contain at least three parameters and require more sophisticated estimating strategies such as minimizing square of differences between sample quantile means and theoretical ones (which usually implies an additional step of estimating the bounds on the income groups or quantiles in the data) (e.g., Chotikapanich et al. (2012)) or sample quantile income shares and theoretical Lorenz curve coordinates (e.g., Jorda and Niño-Zarazúa (2016), Jorda et al. (2018)).

One particular strength of parametric methods is that under a specified model for an income distribution this can allow for correcting some misreporting issues that household survey data are prone to be affected by, and particularly

the problem of under-reported high incomes. The logic behind the few correction methods explored in this literature begins by assumptions regarding which range of incomes observed in the data are bound to be affected by under-reporting (e.g., the higher quantiles) and deriving an estimation method which either exploits only those observations belonging to the correctly observed range of incomes or which places weights on the observed data depending on the assumed magnitude of under-reporting as is done in Pinkovskiy and Sala-i-Martin (2009) or Jorda and Niño-Zarazúa (2016). Additionaly, parametric methods can serve as a framework from which to derive measures of statistical uncertainty concerning the estimated distribution's parameters or summary statistics as is done in Hajargasht et al. (2012) for the case of no under-reporting. The main drawback of making inference on countries' income distributions using parametric distributions is that while some distribution functions might result an adecquate representation of some country-years true income distribution they may not be so for others and so may fail to capture features of the distribution as, for instance, complex patterns of multimodality. Additionally, in the context of the GID, the small sample size problem inherent to estimating country-year income distributions using grouped data from compilatory datasets restrict the choice of possible distribution functions to those defined over low-dimensional parameter spaces, otherwise estimation would be impossible. This last drawback limits, for instance, the exploration of mixtures of distributions at the country-year level.

Darvas (2016) surveys the literature on the GID with particular interest in comparing the performance of methods from these four methodological approaches (i.e., the IQI assumption, kernel density estimation, parametric Lorenz curves, and parametric distribution functions). Judging by their performance in replicating official national income inequality measures based on quantile income share data at the sub-national level for four different countries (Australia, Canada, Turkey, and the USA), the parametric distribution function approach is the best alternative, even if only relatively restrictive 2-parameter distributions were considered (i.e., the log-normal distribution, the Pareto distribution, and the Weibull distribution). This results provide

support to the argument that, given the current state of data availability for studying the GID, methods based on parametric distribution functions are the best alternative explored as of yet.

As a general result of the previous estimates presented in tables 3, 4, and 5, with important heterogeneities in the data, sample of countries, and methods used, all estimates suggest that global interpersonal income inequality has fallen during the period ranging from the late 1980's to the early 2000's. The magnitude of this fall and of the levels of inequality vary substantially from study to study, however, with the possible drop measured by changes in the global Gini coefficient ranging from .008 points (Anand and Segal, 2017) to .067 points (Lakner and Milanovic, 2016), which imply very different conclusions about the dynamics of global inequality along the recent decades of globalization. Few studies present results on the dynamics of the income shares of the top incomes. However, those that do so consistently evidence a phenomenon of increases in this share for the richest 10% of the global population and even more pronounced for the richest 1%. In terms of regional composition of the GID, a phenomenon consistently evidenced in all previous studies is the important rise in, simultaneously, income levels and income inequality of China, the world's most populated country. Finally, in terms of the contributions of the within- and between-countries inequality components, the common result, with all the same heterogeneities between studies still present, is that within-country inequality has played an increasing role along the period.

Two more recent methods for inference on income distributions using grouped data have yet to be explored in the context of the GID. The first of these is the method of Generalized Pareto Curves proposed by Blanchet et al. (2017, 2018b). This semi-parametric method allows for estimating a continuous income distribution using grouped or tabulated data through a mixture of interpolation and extrapolation methods. Very broadly, the method exploits sample estimates of inverted Pareto coefficients (the coordinates

of the Generalized Pareto Curve (Fournier, 2015)) for the income quantiles[2] reported in the data and applies an interpolation method to obtain these coefficients for all quantiles in the range covered by the data while it applies an extrapolation method based on sampling from a fitted 3-parameter Generalized Pareto distribution for obtaining these coordinates for quantiles above the last one covered in the sample used and, optionally, an analogous extrapolation method to obtain these coordinates for the lower part of the distribution not covered in the data. By smoothness constraints imposed on both of these procedures a smooth quantile or distribution function can be recovered in a distribution-free manner from the estimated Generalized Pareto Curve.

The second of these approaches is that of Approximate Bayesian Computation (ABC)[3] as explored very recently in the context of inference on Lorenz curves using grouped data in Kobayashi and Kakamu (2019). This 'likelihood-free' Bayesian inference method approximates the true posterior distribution of the parameters of an assumed parametric Lorenz curve or distribution function without requiring the likelihood function implied by these to be numerically evaluable, a condition required for the vast majority of conventional Markov Chain Monte Carlo (MCMC) methods used in Bayesian inference. Loosely stated, the approximation is based on comparing simulated income data from the assumed likelihood at each draw of a pre-specified parameter sampling density with the observed income data, under a fixed level of tolerance and a measure of discrepancy if the observed data and the simulated data are close enough then the sampled parameter values at that particular draw are a good approximation of a draw from the true posterior density. Once having approximated a posterior density for the income distribution function's parameters, credibility intervals can be estimated for these and any function of these parameters such as the coordinates of the Lorenz curve or synthetic inequality indices.

[2] In fact, the method can be applied equivalently to quantile, quantile share, or mean quantile income data.

[3] See Sisson et al. (2018) for a first compilation on ABC methods and applications.

Two additional issues have been largely left unattended in the previous literature. On the one side, despite the fact that one key source of data used for the study of the GID is that coming from household surveys, and that these surveys are well known to be prone to problems of under-reporting and under-sampling of incomes at the extremes of the distribution[4], and particularly of the highest incomes, only very few attention has been paid to this incomplete data problem, with some notable exceptions (Lakner and Milanovic (2016), Pinkovskiy and Sala-i-Martin (2009), Anand and Segal (2017), Jorda and Niño-Zarazúa (2016)). The empirical observation concerning the increasing share of global incomes being captured by the richest percentiles of the population along the period calls for further refined insight into the magnitude and composition of this trend, which involves paying particular attention to the problem of missing high incomes in the data.

Despite this issue currently being at the center of the literature on estimating income inequality at the country level using survey data (with Blanchet et al. (2018a), Bourguignon (2018), Higgins et al. (2018), Hlasny and Verme (2018) being some very recent examples of this), the exploration of correction methods for the study of global income inequality is still to be attended to. Anand and Segal (2008) have suggested the use of parametric methods as an alternative worth exploring for correcting the issues of under-reporting and under-sampling of high incomes in the context of the GID, which is what I follow in my method.

The second point left unattended in the literature involves measures of the statistical uncertainty associated to the estimated levels and trends in global inequality as pointed out in Anand and Segal (2008). This issue has received limited attention due to, firstly, the nature of the usual data sources used for the empirical studies of the GID, which is mainly composed of summary statistics at the country-year level without a report of the sample size or estimated standard errors associated to these estimates from individual-level

[4]See Deaton (2005) for a seminal study on this issue and Angel et al. (2018) for a recent innovative empirical assessment of this issue for the Austrian case.

data and, secondly, to the lack of proper statistical methods suitable for this type of problem. One simplistic approach was explored in Milanovic (2002), using a simple jack-knife resampling method on his discrete approximation to the GID to obtain a standard error for his estimates which assumes that the quantile income groups in his data are perfectly estimated. Chotikapanich and Griffiths (2002) and Hajargasht et al. (2012) derive asymptotic parameter covariance matrix estimators under different distributional assumptions in the context of grouped data, however these derivations either require knowledge on the sample sizes used to estimate the observed group-level statistics or have not yet been extended to a framework which simultaneously contemplates the issue of under-reported incomes.

Before developing a method for studying the GID which allows for contemplating this two issues building on the previous literature it is important to understand the nature of the possible data sources available for doing so, as the nature of the data acts as a first limitation in the study of this distribution. A brief overview of the main datasets and compilatory projects aimed at producing data for the study of the GID, and my consequent choice for a specific dataset to use for my empirical results is presented in the following section.

3 Data sources

The usual data sources allowing for intertemporal and global comparisons of income and/or expenditure distributions simultaneously are compiled datasets containing quantile income and/or expenditure shares or means and/or Gini coefficient estimates of the distribution of these, and occasionally the population mean and size, for each country-year. These secondary data sources are usually the product of a vast process of compatibilizations and processing of household surveys, which serve as the main primary source, and are usually available for a period starting in the 1980's, when household survey data started to become more publicly available.

The heterogeneities in population coverage, period, and in the definitions of income and/or expenditures amongst household surveys imply that regardless of the efforts put into producing compatibilized and homogeneized datasets, many limitations still remain. Firstly, because some household surveys contain some measure of household consumption or expenditures only and others contain some measure of household incomes only, these two very different dimensions of welfare must either be treated as equivalent or some artificial conversion method must be used if one seeks to achieve a reasonable level of coverage[5]. Secondly, even if expressed in a common currency the real value of a same amount of income may differ greatly between countries and periods as a result of spatiotemporal price differences and therefore a price index may be employed as a common point for comparisons. The conventional approach in this sense is to analyse values evaluated at a same Purchasing Power Parity (PPP) measure. Finally, because of the generally imperfect coverage of the populations' total incomes in household survey samples a scale correction of country-year measured incomes may be performed by centering their distribution on an external, presumably more reliable, measure of the populations' mean or aggregate income as can be the case of measures derived from National Accounts. As is argued in Anand and Segal (2008), there is no clear consensus on the extent or the most adecquate correction for this issue.

The World Bank's Deininger and Squire (1996) [6] constitutes one of the earliest and largest compilation efforts on income distribution statistics at the country-year level in the literature, covering about 108 countries for the 1947-1995 period. The statistics presented in each country-year observation are the Gini coefficient of incomes and cumulative income quintile shares, with indicators of whether these correspond to the individual or household

[5] As is the general case in the literature, I refer indifferently by incomes to either of these two dimensions of welfare from here on.

[6] http://microdata.worldbank.org/index.php/catalog/1790/study-description

distribution, pre- or net-of-taxes values, and whether the values refers to incomes or expenditures. UNU-WIDER's World Income Inequality Database (WIID) project expands the Deininger & Squire dataset updating and complementing it with data from other compilation projects of lower magnitude such as the Luxembourg Income Study (LIS) which focuses on developped economies. These datasets achieve high levels of coverage at the cost of allowing for definitions of incomes and units of observation that vary a lot between country-years, which introduces a margin for inaccuracies when using these sources for studying the distribution of incomes amongst individuals across time and space.

Milanovic (2002), Milanovic (2012), and Lakner and Milanovic (2016) constitute an alternative line of work in the literature that exploit household survey data sources alone to estimate global measures of inequality using consistent definitions of incomes and individuals as units of observation. The Lakner-Milanovic World Panel Income Distribution (LM-WPID) database is one of the most recent and exhaustive database publicly available from this line of work.

The LM-WPID is mainly the product of the World Bank Research Department's *PovCalNet* household surveys compatibilization and publication project and Milanovic (2002) World Income Distribution (WYD) updated dataset, and covers in total 565 household surveys for the 1988-2008 period in 5-year intervals. Each country-year observation in the data is represented by the average income of ten income decile groups. As explained in the authors' online data repository, the dataset is constructed as follows: average per capita incomes converted to 2005 $PPPs along with decile income shares are obtained from PovCalNet, which are combined to obtain estimated decile mean incomes. Next, these data is merged with WYD data, which has an analogous structure and some differences in coverage. These two sources constitute 98% of the data. The remaining gaps are filled with data from LIS, the British Household Panel Survey, the European Union Survey of Income and Living Conditions and the National Statistical Offices of Finland

and Portugal. Additionaly, in the interest of allowing for two possible scale corrections for the country-year distribution of incomes, two alternative measures from National Accounts, also measured in 2005 $PPPs, are provided. These are GDP per capita and per capita household private consumption.

More recent projects directed towards producing income distribution estimates at the country-year level while considering the possible coverage problems of household surveys, such as the World Inequality Database (WID.world). The main purpose of the WID.world project is to combine survey data, national accounts, and tax data to obtain estimates of income distribution at the country level corrected for high incomes, under the main assumption that tax data offers a better representation of the high incomes than survey data. However, the project has not yet reached a degree of coverage sufficient for a detailed study of the GID, mainly due to the still very limited access to quality household survey data and tax data for the work of such projects.

Due to the coverage of the publicly available LM-WPID dataset in terms of only including information derived from nationally representative household surveys, as well as the relatively extense coverage in time and in number of countries, and the relatively detailed information presented for each country-year income distribution, I use the LM-WPID dataset for the empirical study of the GID[7]. One drawback of this dataset, however, is that it does not report information on whether the unit of observation in the individual-level data are households or individuals, nor does it specify if incomes are measured pre- of net-of-taxes, forcing the analysis to consider this quantities indifferently. This drawback, along with the lack of information on the sample sizes of the data used for computing the group-level estimates reported in the dataset, are shared with the vast majority of these type of datasets. Finally,

[7]Data downloaded on 27/04/2019 from Branko Milanovic's online dataset repository hosted at the City University of New York Stone Center on Socio-Economic Inequality's website (https://www.gc.cuny.edu/Page-Elements/Academics-Research-Centers-Initiatives/Centers-and-Institutes/Stone-Center-on-Socio-Economic-Inequality/Core-Faculty,-Team,-and-Affiliated-LIS-Scholars/Branko-Milanovic/Datasets)

because the data presented is directly derived from household survey data, the reported statistics at the country-year level are prone to be influenced by under-reporting problems which calls for contemplating a correction for this in the estimation strategy.

To introduce some notation and understand the nature of the LM-WPID dataset before presenting the estimation method I propose, let's denote individual/household i's annualized income in country-year j as y_{ij} with true cumulative distribution function over this country-year's population F. The quantile $Q(u)$ denotes the individual income level $y \in Y$ for which a proportion u of country-year's j population has a lower income:

$$Q(u) = F^{-1}(u) = \inf\{y \ : \ F(y) \geq u\} \tag{1}$$

Income deciles are then the resulting vector of evaluating the quantile function Q over the range $d = \{0, .1, .2, ..., .9, 1\}$, where for instance $Q(d_3) = Q(.2)$ denotes the second income decile, $Q(d_1) = Q(0) = 0$ and $Q(d_{11}) = Q(1) = \max(Y)$. Sample estimates of $Q(d_z)$ (denoted $\tilde{Q}(d_z)$) can be obtained as the first observed income level in the ordered sample for which at least a proportion d_z of the observations have a lower income.

Sample estimates of decile non-cumulative population-weighted mean incomes, the variable characterizing the country-year's income distribution in the LM-WPID data are then defined as the weighted sample mean of all observed individual incomes between each interval defined by the income deciles:

$$\tilde{y}_{zj} = \frac{\sum_{i=1}^{n_j} w_{ij} \times y_{ij} \times I(\tilde{Q}(d_{z-1}) \leq y_{ij} \leq \tilde{Q}(d_z))}{\sum_{i=1}^{n_j} w_{ij} \times I(\tilde{Q}(d_{z-1}) \leq y_{ij} \leq \tilde{Q}(d_z))} \ , \ z \in \{2, ..., 11\} \tag{2}$$

where $\tilde{y}_{zj}$ denotes the sample estimate of the d_z-th decile's non-cumulative mean income for a sample of size n_j and sample quantile estimate $\tilde{Q}(d_z)$, I denotes the indicator function, and w_{ij} denotes the weight given to individual i in the sampling scheme such that the estimate is representative of the entire

country-year population[8].

For each country-year, the LM-WPID dataset presents as key variables the estimated vector of ten decile non-cumulative and population-weighted mean incomes, along with the population size of each of these deciles, and the population mean income estimated using individual-level data and population weights only. Table 1 below presents the coverage in terms of number of countries in the LM-WPID for each possible year and year-pairs, clearly evidencing the pattern of increasing household survey data availability in the recent decades.

	1988	1993	1998	2003	2008
1988	72	70	67	66	60
1993	70	112	95	96	87
1998	67	95	118	100	96
2003	66	96	100	130	104
2008	60	87	96	104	118

Table 1: Country coverage by years in LM-WPID

Because of the important difference in coverage of the data between 1988 and 1993, I restrict my analysis to the 1993-2008 period, which constitutes a common sample of 87 countries. Table 2 presents the country and coverage in terms of the entire global population for this common sample. Because of the low coverage in terms of countries of this common sample in some regions, I refrain from analysing the dynamics of the regional composition of the GID and of regional levels of inequality.

Having discussed the nature of the available data for the problem of studying global interpersonal income inequality and the choice for the LM-WPID dataset as the prefered source to be used for my estimations, the following section describes the method of Multiple Imputations used to estimate

[8]In fact, due to lack of documentation on how the estimates are obtained it is unclear how precisely the weighting of individual-level samples is performed for each country-year. However, to the extent of the method to be used for the estimation this can be neglected.

the GID taking into consideration the need for correcting possible under-reported high incomes in the data and for estimating the statistical uncertainty of the final estimates.

4 Proposed method

The small sample sizes of the data presented at the country-year level in the LM-WPID data poses a limitation to the methods that can be used to make inference on the GID from this dataset. Methods requiring considerable sample sizes, such as most non-parametric methods, or individual-level data are therefore not amongst the best choices. It is for this reason, as well as for the recent advances in the literature and the analysis presented in Darvas (2016) supporting the superior performance of parametric distribution methods with respect to other approaches explored in previous studies, that a parametric distribution method should be an adecquate approach for this particular analysis. Additionaly, I approach the estimation of the GID as is the case in all previous studies by first estimating income distributions for each country and then aggregating them in a population-weighted manner to obtain an estimate of the GID.

4.1 The GB2 distribution and its estimation using truncated Lorenz curves

One particularly flexible and general univariate parametric distribution suitable for the problem of modelling income distributions at the country level is the GB2 distribution (McDonald (1984),(Kleiber and Kotz, 2003, chapter 6)). This four-parameter distribution function encompasses as special cases many of the most commonly used parametric income distributions in the literature such as the log-normal, the Singh-Maddala (i.e., Burr XII), and the Weibull distributions, amongst many others, which supports its choice as a sufficiently flexible distribution to model income distributions at the country-year level

without requiring a large number of parameters given the small sample sizes available in the data at this level. The probability density function (pdf) f of the GB2 distribution with shape parameters a, p, and q, and scale parameter b follows:

$$f(y;a,b,p,q) = \frac{ay^{ap-1}}{b^{ap}B(p,q)\left(1+\left(\frac{y}{b}\right)^{a}\right)^{p+q}} \ , \ (y,b,a,p,q) \in R_{+}^{5}$$

where $B(p,q)$ denotes the Beta function, defined as:

$$B(p,q) = \int_{0}^{1} t^{p-1}(1-t)^{q-1}dt$$

Defining

$$u \equiv \frac{\left(\frac{y}{b}\right)^{a}}{1+\left(\frac{y}{b}\right)^{a}} \equiv \frac{y^{a}}{b^{a}+y^{a}}$$

the cumulative distribution function (cdf) F of the GB2 can be expressed as (Kleiber and Kotz (2003)):

$$F(y;a,b,p,q) = \frac{1}{B(p,q)} \int_{0}^{u} t^{p-1}(t-u)^{q-1}dt = IB(u;p,q)$$

which expresses the cdf of the GB2 as the regularized Incomplete Beta Function IB, the cdf of the regularized Beta distribution. Finally, the expectation of this GB2 distribution can be expressed as:

$$E(Y;a,b,p,q) = \frac{bB\left(p+\frac{1}{a},q-\frac{1}{a}\right)}{B(p,q)} \ , \ \ -ap < 1 < aq$$

Before introducing the estimation strategy I use to obtain estimated parameter values of the assumed GB2 distribution for each country-year in the data, it is worth recalling the definition of the Lorenz curve and its potential for studying income distributions and income inequality.

The Lorenz curve of a population's income distribution LC (Lorenz, 1905) can be defined as a function relating income-ordered population cumulative shares $u \in [0,1]$ and the cumulative share of the total population's income

$LC(u) \in [0,1]$ represented by these population shares. The Lorenz curve is a powerful tool for studying the concentration of a population's incomes along different points of the income distribution and has been used as such for countless studies of income distribution and income inequality. Following the presentation in Sarabia (2008), the Lorenz curve for an income variable Y with parametric cdf $F(.;\theta)$ can be expressed using the definition of the first-order moment cdf $F_{(1)}(.;\theta)$ defined as

$$F_{(1)}(y;\theta) = \frac{\int_0^y x dF(x;\theta)}{\int_0^\infty x dF(x;\theta)} = \frac{\int_0^y x dF(x;\theta)}{E(Y;\theta)}$$

where $F_{(1)}(y;\theta)$ can be interpreted as the proportion of total incomes which correspond to the subpopulation of individuals with incomes below y.

For each income-ordered population proportion u, the corresponding income level is $Q(u;\theta) = F^{-1}(u;\theta)$, such that the Lorenz curve corresponding to this parametric distribution can be expressed as

$$\begin{aligned} LC(u;\theta) &= F_{(1)}(F^{-1}(u;\theta);\theta) \\ &= \frac{\int_0^u F^{-1}(x;\theta)dx}{E(Y;\theta)} \ , \ u \in [0,1] \end{aligned}$$

from which properties of continuity, convexity, non-decreasing, and differentiability almost everywhere in $u \in [0,1]$ can be proved, as well as the properties of $L(0;\theta) = 0$ and $L(1;\theta) = 1$.

One important feature of the Lorenz curve is its scale-independence property, as it is by definition a function relating proportions or shares and not income levels themselves. This feature will be central to the estimation strategy I present below.

In a distribution where all individuals in the population receive exactly the same income the Lorenz curve satisfies $LC(u) = u \ , \ u \in [0,1]$, which serves as a perfect equality benchmark from which to measure the degree of

inequality implied by any Lorenz curve.

For the particular case of the GB2 distribution, the first-order moment cdf follows (Kleiber and Kotz, 2003)

$$\begin{aligned} F_{(1)}(y;a,b,p,q) &= F\left(y;a,b,p+\frac{1}{a},q-\frac{1}{a}\right) \\ &= IB\left(u;p+\frac{1}{a},q-\frac{1}{a}\right) \ , \ u \equiv \frac{\left(\frac{y}{b}\right)^a}{1+\left(\frac{y}{b}\right)^a} \equiv \frac{y^a}{b^a+y^a} \end{aligned}$$

such that the Lorenz curve can be expressed as

$$\begin{aligned} LC(u;a,b,p,q) &= F_{(1)}(F^{-1}(u;a,b,p,q);a,b,p,q) \\ &= IB\left(IB^{-1}(u;p,q);p+\frac{1}{a},q-\frac{1}{a}\right) \end{aligned} \tag{3}$$

The expression for the Lorenz curve given in (3) is clear in showing explicitly the scale-invariance of this curve, as the scale parameter b does not enter the function as an argument in the last expression. This property has been exploited in many estimation methods using grouped income data, as it allows estimating the three shape parameters by fitting the GB2 Lorenz curve to the income shares implied in the data, and the scale parameter separately by fitting the expectation of the GB2 to the mean income in the data[9]. The method I use to obtain parameter estimates for the assumed GB2 distribution for each country-year also exploits this feature.

In a setting where the estimated non-cumulative decile mean incomes reported in the LM-WPID data for any country-year $\tilde{y}_k$ are unbiased, then unbiased estimates of cumulative decile income shares can simply be obtained as

$$\tilde{s}_k = \frac{\sum_{i=1}^{k}\tilde{y}_i}{\sum_{i=1}^{10}\tilde{y}_i} \ , \ k \in \{1,...,10\}$$

Since by definition these shares are the sample estimates of the coordinates

[9]See Hajargasht et al. (2016) for a recent development.

of the Lorenz curve for the u_k population proportion of lower incomes, a straightforward fitting strategy to estimate the a, p, and q parameters could define moment conditions matching each of the ten $\tilde{s}_k$ sample estimates to their theoretical counterpart under a GB2 distribution $LC(u_k; a, p, q)$. Equivalently, a least squares estimator can be constructed to obtain parameter estimates which minimize the sum of square distances between these sample quantities and their theoretical counterpart as:

$$(\hat{a}, \hat{p}, \hat{q}) = \min_{a,p,q} \sum_{k=1}^{10} \left(\tilde{s}_k - IB\left(IB^{-1}(u_k; p, q), p + \frac{1}{a}, q - \frac{1}{a} \right) \right)^2 \qquad (4)$$

However, the fitting method must take into account the fact that the observed grouped data at the country-year level is the result of estimating decile mean incomes on household survey data which is prone to a certain degree of under-reporting of high incomes. Assuming that the data are only representative of the income-ordered population proportion t, this means that the sample estimates of cumulative decile income shares $\tilde{s}_k$ are biased estimates of the total population income shares and that a simple fitting method such as that described above will only estimate parameter values for the distribution over this sub-population represented in the data. In this sense, I follow closely the approach proposed in Jorda and Niño-Zarazúa (2016). This method exploits the following logic. If for a given country-year the estimated income share of, for example, the 100 individuals of lowest incomes, representing a share of u_1 individuals in the sample, then these individuals represent the share $u_1 \times t$ of the total population and their incomes represent the share $\tilde{s}_k \times LC(t; a, p, q)$ of the total population's incomes (i.e., their share in the sample weighted by the share of the population's incomes covered in the sample). In terms of the theoretical Lorenz curve this can be expressed as:

$$LC(u_k; t, a, p, q) = \frac{LC(u_k \times t; a, p, q)}{LC(t; a, p, q)}$$

where $LC(u_k; t, a, p, q)$ denotes the theoretical sample income shares of the sample proportion u_k in a sample representative of the proportion t of lowest

incomes in the population. This is, $LC(u_k; t, a, p, q)$ is the theoretical counterpart of the sample estimate $\tilde{s}_k$. The least-squares estimators under this particular assumption on the type of under-reporting becomes then:

$$(\hat{a}, \hat{p}, \hat{q}) = \min_{a,p,q} \sum_{k=1}^{10} \left(\tilde{s}_k - \frac{IB\left(IB^{-1}(u_k \times t; p, q), p + \frac{1}{a}, q - \frac{1}{a}\right)}{IB\left(IB^{-1}(t; p, q), p + \frac{1}{a}, q - \frac{1}{a}\right)} \right)^2$$

This estimator allows for estimating the parameters of the assumed GB2 distribution over the entire country-year population conditional on the assumed known a priori value for the truncation parameter t. It is straightforward that this expression coincides with the expression given in 4 when the data is assumed to be representative of the entire population's incomes (i.e., when $t = 1$) as the Lorenz curve value in the denominator of the theoretical expression will by definition always be 1 in that case.

To obtain an estimate for the scale parameter b, the method used in Jorda and Niño-Zarazúa (2016) cannot be followed, as they assume that the GDP per capita in their data is an unbiased estimate of the population's mean income. In that case, b can be easily recovered by substituting the estimated shape parameters and the GDP per capita value in the expression of the expectation of the GB2 distribution. Given my interest in using only survey data for the estimation, however, I propose an extension of this approach to obtain an estimate of b given the potentially biased mean income estimate in the data for each country-year, denoted $\bar{y}$, and the previously estimated shape parameters conditional on the assumed t. The method goes as follows. Compared to the true population mean income, $\bar{y}$ represents a truncated mean. The theoretical expression of this truncated mean (of which $\bar{y}$ is an unbiased estimator) for an assumed truncation t is:

$$E[Y|Y < F^{-1}(t|a, b, p, q); a, b, p, q] = \int_{-\infty}^{F^{-1}(t|a,b,p,q)} x \frac{f(x|a, b, p, q)}{F(F^{-1}(t|a, b, p, q); a, b, p, q)} dx$$

We have also that:

$$t = F(F^{-1}(t|a,b,p,q);a,b,p,q)$$

Because the theoretical Lorenz curve at t can be defined as:

$$\begin{aligned} L(t|a,p,q) &= F_{(1)}\left(F^{-1}(t|a,b,p,q)|a,b,p+\frac{1}{a},q-\frac{1}{a}\right) \\ &= \frac{\int_{-\infty}^{F^{-1}(t|a,b,p,q)} xf(x|a,b,p,q)dx}{E[Y|a,b,p,q]}, \end{aligned}$$

we can derive an estimate of the population's mean income as:

$$\frac{\bar{y}t}{L(t|\hat{a},\hat{p},\hat{q})} = E[Y|\hat{a},b,\hat{p},\hat{q}],$$

given the previous estimates of the shape parameters of the GB2 and the assumed truncation value t. Recalling the theoretical formula for the expectation of a GB2:

$$E[Y|a,b,p,q] = b\frac{B\left(p+\frac{1}{a},q-\frac{1}{a}\right)}{B(p,q)},$$

we can use this estimate of the population's mean income to derive an estimate of b following:

$$\hat{b} = \frac{\bar{y}t}{L(t|\hat{a},\hat{p},\hat{q})} \times \frac{B(\hat{p},\hat{q})}{B\left(\hat{p}+\frac{1}{\hat{a}},\hat{q}-\frac{1}{\hat{a}}\right)}$$

The estimation strategy for the shape parameters is a non-linear optimization problem, and as such the algorithm used for solving it requires the choice of a vector of starting values (i.e., an initial guess) for the three parameters. In this case, I initialize the algorithm at the vector of initial values $(\hat{a},\hat{p},\hat{q}) = (1.5,1,1.5)$ for all cases. This parameters satisfy the condition for the existence of the mean of the GB2 and are close to many special cases of the GB2 such as the Singh-Maddala, the Fisk, the Dagum, and the Beta-2 distributions. Because the potentially high number of local minima in the objective function of this problem, nothing assures that the obtained parameter estimates are

the global minima even if a set of alternative starting values is explored. This is the case in most estimation strategies for the parameters of the GB2 distribution (e.g., Jorda et al. (2018), Hajargasht et al. (2012)) and there is no clear consensus in the literature on how this initial values should be chosen.

One additional condition imposed on the optimization algorithm involves a restriction on the space of parameter values such that only parameter values for which the expectation of the GB2 distribution and a positive mode exists (i.e., $a \times q > 1$ and $-a \times p < 1$, and $a \geq 1$ respectively).

4.2 A Multiple Imputations approach to estimate the GID

In order to construct an individual-level distribution of incomes for the whole world using country-level grouped data, each individual in the world population needs to be assigned an income level from his country's income distribution. The way in which this income distribution is estimated determines then an income imputation mechanism, from which grouped data is used to assign the unobserved individuals' incomes.

Many of the previous approaches explored in the literature can be defined in terms of their imputation method. For instance, in the more traditional approach in the literature, the IQI assumption, the imputation mechanism involves splitting a country's total population into groups of sizes identified in the data (e.g., income deciles) and each individual is assigned the mean income of his assigned group. The result of this simple imputation is a distribution of individual-level imputed incomes, which can be thought of as an imputed interpersonal distribution of incomes for the country. The GID is then the imputed interpersonal distribution of incomes where the imputation mechanism is such that each individual is imputed an income from his country's distribution under this IQI assumption. In this case, the imputation method yields a unique synthetic sample of individuals' incomes.

Instead of following the more traditional simple imputation approach of assigning one income value for each individual following a specific imputation model, I explore a Multiple Imputations framework. This approach has its first origins in Rubin (1977) and is given extensive technical and applied coverage in Rubin (1987) and Drechsler (2011, chapter 3). The general logic of the Multiple Imputations approach seeks to make valid the use of complete-data inference methods on incomplete data under an imputation method for the missing data. By imputing multiple values from the income imputation model for each individual, many synthetic income samples can be generated and complete-data statistics of interest can be computed over each of these. In particular, as long as the imputation model is stochastic, then each individual will be assigned different incomes across all imputations. This properly translates uncertainty regarding the mechanism determining individuals' incomes to uncertainty on the imputed income samples. The virtue of having as many complete-data estimates as synthetic samples (i.e., as incomes imputed for each individual) is that an averaging of these statistics can be performed to obtain a unique estimated statistic which is influenced by this uncertainty regarding the imputations yet is not entirely determined by any specific synthetic sample.

The multiple imputations approach has been recently explored in the context of top-coded incomes using individual-level survey data in Jenkins et al. (2011). Their imputation method consists in generating many partially-synthetic complete datasets by drawing income values for those individuals for which the top-code income value is observed in the data from a GB2 distribution fitted on the observed data (corrected for the truncation imposed by top-coding).

In terms of estimating the GID using the LM-WPID data, this can be thought of as having missing data issues in two senses. Firstly, because the LM-WPID only presents group-level statistics of each country's income distribution, individuals' incomes themselves are missing. Additionaly, because

household survey data alone is used to estimate this statistics, and this surveys are prone to under-sampling of high incomes, they are provide information only concerning the rest of the population's incomes, implying that they do not treat the problem of this missing incomes. The imputation model used to assign individuals' incomes should contemplate for both of this missing data issues.

I propose a Multiple Imputation strategy adapted to the context of grouped data. The method goes as follows. In the presence of under-reporting or under-sampling of high incomes the decile mean incomes reported in the data at the country-year level are all biased as the survey income deciles themselves are biased estimates of the population income deciles. This implies that the imputations must be done such that individual incomes themselves are imputed for the entire country's population, including those individuals not covered by the survey. To this extent, the estimated GB2 distribution using truncated Lorenz curves described in the previous section can be used as an imputation mechanism. The imputation model from which individual-level synthetic samples are drawn for each country-year is determined then by the assumption regarding the degree of under-reporting for that country-year's household survey data (t) and the corresponding estimated parameters for the assumed GB2 distribution.

Because the solution to the least-squares optimization problem used to estimate the GB2 parameters is entirely determined by the value of t, which shapes the specific optimization problem, it is then possible to define a function $\mathcal{G}_j(t, n, m)$ taking as input a value for t, a population size n, and a number of imputations m, which returns a set of m synthetic datasets, each containing imputed incomes for n individuals. Each of these datasets are an independent random draw of size n from the estimated GB2 distribution for country-year j given the chosen value for t.

Because this framework allows for generating samples for individuals' incomes for all country-years in the LM-WPID data, a synthetic sample

of the GID for a given year can be constructed by randomly taking one synthetic dataset for each country in that year. In order to do so, the samples at the country level must be generated using population sizes n that respect that country's population weight in the total population of all countries in that year in the data. Alternatively, one could scale down the world population such that a sample size n_j is set for China, the world's most populated country, and set the sample size n_l for any other country l such that its size relative to China $\frac{n_l}{n_j}$ matches the relative size of their populations for that year. This is the approach I follow to set the sample sizes, while I determine China's sample size and Luxembourg's sample size (the least populated country in the 1993-2008 common sample) such that the latter has a sample size of 30 observations, which roughly amounts to a sample size of 80000 observations for each sample generated for China.

This method allows then to define a vector function $\mathcal{H}(\mathcal{G}_1(t_1, n_1, m), ..., \mathcal{G}_J(t_J, n_J, m)) \equiv \mathcal{H}(t_1, ..., t_J, n_1, ..., n_J, m)$ which takes as inputs the assumed truncation parameters t_j and population sizes n_j for each of the J countries assumed as representative of the entire global population in a given year, as well as the number of imputations m, and returns m synthetic individual-level samples of the global population's incomes.

Denoting by $H_i \equiv \mathcal{H}(t_1, ..., t_J, n_1, ..., n_J, m)_i$ the i-th of such synthetic samples of the GID, then the Multiple Imputation estimate of any complete-data statistic μ is the simple average of the estimated statistic over each of the m synthetic samples (Rubin, 1987):

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{\mu}(H_i)$$

Additionaly, the estimated variance of this estimate ($\hat{T}_m$) can be obtained as the sum of the average "within-synthetic-sample" variance ($\hat{\bar{v}}_m$), which is the average of the estimated variances of this estimator over each of the H_i synthetic datasets ($\hat{v}(H_i)$, $i = 1, ..., m$), and a weighted average of the

"between-samples" variance $((1 + m^{-1})\hat{b}_m)$, where $\hat{b}_m$ is the variance of the estimated statistic $\hat{\mu}(H_i)$ across all synthetic samples of the GID. This is:

$$\begin{aligned}\hat{T}_m &= \hat{\bar{v}}_m + \hat{b}_m + \frac{\hat{b}_m}{m} \\ &= \sum_{i=1}^{m} \frac{\hat{v}(H_i)}{m} + \sum_{i=1}^{m} \frac{(\hat{\mu}(H_i) - \hat{\mu}_m)^2}{m-1} + \frac{1}{m} \sum_{i=1}^{m} \frac{(\hat{\mu}(H_i) - \hat{\mu}_m)^2}{m-1}\end{aligned}$$

It is important to recall that no explicit distribution function is assumed for individual incomes at the global level in my approach, such that only distribution-free estimators are to be used for any statistic of interest μ. This determines that there may be no sensible analytical formula for estimating the $\hat{v}(H_i)$ component of the total Multiple Imputation variance for some μ. In this cases I apply a simple non-parametric bootstrap method, resampling with replacement many times from H_i and calculating the variance of the estimated statistic across these bootstrap samples.

I consider five common sets of statistics to summarize the levels and changes in the GID and global inequality. Firstly, I consider the Gini coefficient, which can estimated in a distribution-free approach for any synthetic sample H_i as the following coefficient of relative mean differences of individuals' incomes

$$G(H_i) = \frac{\sum_{k=1}^{n} \sum_{j=1}^{n} |y_{k;i} - y_{j;i}|}{2n^2 \bar{y}_i}$$

where $\bar{y}_i$ denotes the mean income in the sample, $y_{k;i}$ denotes the k-th income in H_i, and n denotes the sample size.

Interpreted as a measure of distance between a Lorenz curve and a state of perfect income equality, the Gini coefficient takes values between 0 and 1, with the latter implying maximal inequality. Due to its direct link with the Lorenz curve, the Gini coefficient is a scale-independent measure of inequality, which limits its interpretation as such to concepts of relative inequality. Two important drawbacks of this inequality measure are that,

firstly, this measure is more sensitive to changes in the middle range of the associated distribution than to changes in the extremes of the distribution, and that, secondly, this measure is not perfectly decomposable amongst sub-populations, which in the context of global inequality implies that the Gini coefficient of the global interpersonal distribution of incomes cannot be obtained as some aggregation of Gini coefficients at the country level.

A set of inequality measures which can be more sensitive to changes in other parts of the income distribution is the family of Generalized Entropy (GE) indices, defined in a distribution-free way as

$$GE(H_i;\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{j=1}^{n} \left(\left(\frac{y_{j;i}}{\bar{y}_i} \right)^{\alpha} - 1 \right) \ , \ \alpha \neq 0,1 \\ \frac{1}{n} \sum_{j=1}^{n} \left(\frac{y_{j;i}}{\bar{y}_i} \right) \times \ln \left(\frac{y_{j;i}}{\bar{y}_i} \right) \ , \ \alpha = 1 \\ -\frac{1}{n} \sum_{j=1}^{n} \ln \left(\frac{y_{j;i}}{\bar{y}_i} \right) \ , \ \alpha = 0 \end{cases}$$

This family of indices incorporate a parameter α which can give different weights to changes in different parts of the distribution. In particular, the two special cases where $\alpha = 0$, known as the Mean Log Deviation (MLD), or $\alpha = 1$, known as Theil's T (Theil-T), are of particularly popular use in the literature as they have a clear interpretation in terms of which parts of the distribution are given more weight in the inequality measurement. The MLD is a measure more sensitive to changes in the lower part of the income distribution, while the Theil-T measure is more sensitive to changes in the upper tail.

Thirdly, I explore the global sample mean income, as a summary measure of central position, to analyse changes in the levels of global incomes. The sample mean in this case is simply

$$\bar{H}_i = \bar{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{j;i}$$

A fourth set of measures which can give insight on the shape of an income

distribution and can be given interpretation in terms of income inequality are income group shares. Of particular relevance are the income shares of the highest income groups, like the top 10% and top 1% of individuals in the income-ordered population. One approach to formulate distribution-free estimates of income shares involves the use of empirical distribution function estimates ($\hat{F}_{H_i}$) and distribution-free quantile estimates ($Q(H_i, \tau)$) as follows

$$\begin{aligned} \hat{F}_{H_i}(y_{k;i}) &= \frac{1}{n}\sum_{j=1}^{n} y_{j;i} I(y_{j;i} \leq y_{k;i}) \\ Q(H_i;\tau) &= \inf\{y_i \ : \ \hat{F}_{H_i}(y_i) \geq \tau\} \\ S(H_i;\tau_l,\tau_h) &= \frac{\sum_{j=1}^{n} y_{j;i} I(Q(H_i;\tau_l) \leq y_{j;i} \leq Q(H_i;\tau_h))}{\sum_{j=1}^{n} y_{j;i}} \end{aligned}$$

This is, the income share of the group of individuals with incomes between the τ_l and τ_h percentiles of the sample ($S(H_i;\tau_l,\tau_h)$) is simple the ratio of the aggregate income of this groups with respect to the total income in the sample. More particularly, I estimate income shares of the bottom 50% of the distribution ($(\tau_l,\tau_h) = (0, 0.5)$), the middle 40% ($(\tau_l,\tau_h) = (0.3, 0.7)$), the top 10% ($(\tau_l,\tau_h) = (0.9, 1)$), and the top 1% ($(\tau_l,\tau_h) = (0.99, 1)$).

Finally, a powerful tool to track changes along the income distribution across two periods is the anonymous Growth Incidence Curve (GIC) introduced by Ravallion and Chen (2003). The GIC measures the growth rate of different sectors of an income distribution across two periods and as such can give some insight into how economic growth between these periods is distributed along the distribution. More explicitly, the GIC measures changes in the mean income of a given fractile of the distribution between one point in time and a future one.

A distribution-free estimate of the fractile mean income growth rate ($GIC(H_i, H_j;\tau_l,\tau_h)$) between an initial period $t-1$ and t for the fractile group defined by the (τ_l,τ_h) percentiles over the $H_{i;t} \equiv H_i$ and $H_{j;t-1} \equiv H_j$

synthetic samples, with sizes n_i and n_j respectively, follows

$$\begin{aligned} GIC(H_i, H_j; \tau_l, \tau_h) &= \left(\frac{\sum_{k=1}^{n_j} I(Q(H_j;\tau_l) \leq y_{k;j} \leq Q(H_j;\tau_h))}{\sum_{k=1}^{n_i} I(Q(H_i;\tau_l) \leq y_{k;i} \leq Q(H_i;\tau_h))} \right) \\ &\times \left(\frac{\sum_{k=1}^{n_i} y_{k;i} \times I(Q(H_i;\tau_l) \leq y_{k;i} \leq Q(H_i;\tau_h))}{\sum_{k=1}^{n_j} y_{k;j} \times I(Q(H_j;\tau_l) \leq y_{k;j} \leq Q(H_j;\tau_h))} \right) - 1 \end{aligned}$$

In the context of the GID, only two previous works have explored the GIC (Lakner and Milanovic, 2016, Alvaredo et al., 2018). Despite some notable differences in their estimation methods and the nature of the data they exploit, both studies have put forward the idea of the global GIC following an 'elephant' shape for the period starting in the late 1980's and up to the early 2010's. This elephant shaped curve suggests that growth rates on the lower part of the distribution increase as we move along the distribution, then decrease to their minimum for income groups above the median, who have experienced the least growth in their mean fractile incomes, and then pronouncedly increase for the top income groups. These results suggest that particular detailed attention must be paid to the top income groups, as they evidence that the most heterogenous growth rates are within the top 10%. For these reasons, I define the ordinates of the GIC over the ventiles of the GID up to the 95-th percentile, after which I present estimated fractile mean growth rates at the percentile level.

Two final remarks must be made regarding how the synthetic samples are generated. Firstly, although the GB2 might be a flexible enough distribution to correctly represent many types of incomes distributions it can nonetheless yield unrealistic representations of many others. For example, in economies with significant proportion of the population living with no monetary income (e.g., developping countries with big rural populations), then the mass of zero incomes in the data can force the estimates towards a Pareto type distribution. In this particular case, in fact, the problem is that a single unimodal distribution as the GB2 will hardly represent correctly the duality of having an income distribution for the individuals participating in the monetary economy and a large mass of individuals absent from this with

zero incomes. The magnitude of this problem when estimating the GID can be large, as this is the case of some very large economies as the Chinese one. This problem regarding the assumption of a GB2 model for all country's incomes distributions in my method, in combination to the correction method for missing high incomes, can yield estimated parameters that imply an unrealistically long upper tail on the distribution of incomes. Despite the restriction imposed to guarantee the existence of a mean, the estimated parameters may imply such a long range for the tail of the distribution that numerically infinite values may be generated as an imputed income for some individuals. Because this obviously represents an unrealistic imputation, the solution taken implies simply discarding these numerically infinite draws and re-drawing a valid income imputation from the fitted distribution.

The second remark is closely linked to this drawback. Although numerically infinite values can clearly be identified and substituted in the imputations, this is not the case for some very high but numerically finite imputed incomes. As a consequence of this, some atypical synthetic samples can be generated for the countries where the estimated parameters imply this type of problem. However, the Multiple Imputation framework offers a possible solution to this. Because many synthetic samples are generated for the GID, the distribution across samples of the statistics sensitive to these samples with atypically high incomes will present atypical values with respect to the rest of samples. Because both the Multiple Imputation estimate and its estimated standard error are sensitive to this problem, a simplistic correction to avoid considering samples with unrealistic imputed values involves simply omitting them in the estimation of the Multiple Imputation statistics. Under this simplistic correction to the problem, the Multiple Imputation estimate and its standard error for each statistic are calculated over the less extreme 80% of values of the corresponding statistics across all samples. It is important to notice that this correction yields conservative estimates of both the Multiple Imputation statitistics and their standard errors as it directly supresses both of them.

4.3 Uncertainty regarding the magnitude of under-reporting of high incomes

Having defined an imputation strategy to estimate individual-level income distributions for each country-year's total population and for the GID, it is important to discuss the choice for the truncation parameter t. This parameter is, by assumption, not estimable from the data and must therefore be externally specified. The entire estimation strategy is conditional on this parameter, which expresses the assumed degree of coverage of a country-year's survey data of the corresponding population's incomes. Jorda and Niño-Zarazúa (2016) calibrate this parameter for a small sample of countries so that the estimates perfectly reproduce top income shares presented in the WID.world database, and set this parameter to be the average value of this calibration for all countries not covered in the WID.world ($t = 0.983$).

Although the WID.world data attempt to correct survey data problems of under-reported or under-sampling of high-incomes by exploiting tax data and National Accounts, the coverage of this project is still very limited. Additionaly, neither tax data nor National Accounts are exempt of problems of under-reporting. On the one side, as is pointed out in Anand and Segal (2008), many quantities reported in National Accounts are to some extent verified or corrected using survey data, and it is unclear which items within National Accounts might represent the best unbiased estimate of a population's total household incomes. On the other side, tax evasion works on tax data in an analogous way as under-reporting of high incomes acts on survey data, such that even if some correction may be achieved by complementing survey data with tax data on high incomes, the degree of coverage of the resulting sample is still uncertain.

I explore two alternative scenarios concerning alternative assumptions on the truncation parameter, which I define as follows.

Scenario I: One alternative involves setting a same value for this parameter for all country-years $t_j = t \ \forall j \in \{1, ..., J\}$, for which values on the range

$t \in [0.90, 1]$ are explored. This allows for exploring the sensitivity of the estimated GID to the assumed degree of under-reporting at the country-year level, in a fashion similar to that analysed in Jorda and Niño-Zarazúa (2016). This range of values covers all scenarios between perfect survey coverage ($t = 1$, the standard in the literature) to a scenario where the incomes of the richest 10% of the population are not covered by the survey data ($t = 0.90$). In this approach, as all country-years are assumed to be subject to the same degree of under-reporting, the interpretation of the t parameter at the global level is analogous to that at the country-year level: survey data is assumed to cover the proportion t of incomes of the entire income-ordered global population.

Scenario II: The second scenario introduces a mechanism which explicits the uncertainty about the exact degree of under-reporting at the country-year level through an *a priori* assumed distribution for the t_j parameter. In this scenario, the Multiple Imputation notation introduced in the previous section must be slightly extended to allow for each synthetic dataset produced by $\mathcal{G}_j$ to be associated to a different value of t_j. Denote by $\boldsymbol{t}_j$ an m-dimensional vector of assumed truncation parameter values, then $\mathcal{G}_j(\boldsymbol{t}_j, n, m)$ follows a very similar logic as in the previous notation but now the imputation method follows:

i) $\boldsymbol{t}_j$ is a vector of m independent random draws from a $Beta(\alpha_0, \beta_0) \times (b - a) + a$ distribution, where a and b denote the assumed lower and upper bounds on t_j. This is equivalent to assuming that $\tilde{t}_j \equiv \frac{t_j - a}{(b-a)} \sim Beta(\alpha_0, \beta_0)$. The virtue of this distributional assumption is that a flexible range of distributional shapes can be expressed through different values of the $Beta$ distribution's parameters (α_0, β_0). In particular, I explore three alternative settings for these parameters, considering different assumptions about the expected value for this parameter and the symmetry and skewness of its prior probability distribution. The three respective

Beta densities explored are presented in Figure 7. As is discussed in detail in the following section, one of the drawbacks of the method is its high sensitivity to the values considered for the t parameter, and so I have only considered distributions over the conservative range of $[0.95, 0.999]$.

ii) Each synthetic sample $G_i \equiv \mathcal{G}_j(\boldsymbol{t}_j, n, m)_i$ is then a random draw of the fitted GB2 distribution at the country-year level for the corresponding randomly drawn t_j value.

The interesting feature of this approach is that, because of the direct link between t_j and the consequent parameter estimates for the GB2 distribution for country j, eliciting a prior probability distribution for t_j amounts to eliciting a prior distribution for the set of parameter estimates of the GB2. In this sense, the Multiple Imputation approach simultaneously takes into account the uncertainty regarding the imputation model itself and the uncertainty regarding the imputed values themselves. The previous, more simplistic, scenario assumes that t is known, and therefore that all parameters in the imputation model are known, so it only captures uncertainty regarding the imputed incomes and not on the appropriateness of the imputation mechanism.

5 Results

5.1 Baseline estimates

As a first assessment of the performance of the estimation method at the country level, Figure 1 compares the empirical Lorenz curves of a randomly chosen synthetic sample with the empirical Lorenz curves obtained using the simple IQI assumption for the cases of China, France, USA, Mali, and Morocco in 1993 and in 2008. This Figure clearly shows the underestimation

of inequality obtained when using the IQI as compared to the parametric method I explore, as can be seen by the straight-line behavior of the IQI Lorenz curves at the upper deciles of the distribution. Besides this, both methods yield similar Lorenz curves and dynamics.

As a way of comparing the goodness-of-fit of the method using external data, the scenario where no correction for missing high incomes should give estimated levels of inequality at the country and global levels similar to those found in the previous literature, and particularly to those using similar data sources. To this extent, I compare the country-level estimates with those presented in the World Bank's World Development Indicators, as they use PovCalNet as their data source and this constitutes the larger part of the LM-WPID's data sources. Although the World Bank's estimates cover a subset of the countries in the 1993-2008 LM-WPID common sample I use, Figure 2 evidences the close similarities between estimates in terms of discrepancies between country-level estimates of the Gini coefficient and the top 10% income share. As the Gaussian kernel density estimates presented in the figure show, only a few countries have an estimated Gini or top 10% share which differs with the World Bank's estimates by 5 percentage points or slightly more. These are the cases of Bulgaria, Mauritania, Poland, Kenia, and Ukraine for 1993 and the cases of Bulgaria, Mauritania, Honduras, and Nigeria for 2008.

A final source for comparison with previous studies of the obtained GID estimates under the scenario where no correction for under-sampling is done concerns the shape of the GID and its change over the period. (Lakner and Milanovic, 2016, Figure 2) estimates serve as the most appropriate comparison, as they use the LM-WPID dataset. The Gaussian kernel density estimates over two synthetic samples of the GID for the years 1993 and 2008 presented in Figure 3 closely match their results both in terms of scale and in terms of the shape of the estimated distributions. This baseline estimates show a dynamic for the GID moving from a clear "twin-peaks" shaped distribution for 1993, with an important mass of individuals with

annualized incomes of about $PPP 450, to a less polarized distribution for 2008, with a general shift towards higher levels of incomes for all points in the distribution but with important heterogeneities on the magnitude of this shift.

5.2 Estimates under fixed and common degree of under-reporting

Having assessed the performance of the method in terms of the similarities of the estimates with comparable previous studies at the country and global levels, the results obtained when correcting for under-reported high incomes under the Scenario I specification can be interpreted by comparison to those obtained when no such correction is performed.

Table 6 presents the Multiple Imputation estimates for different assumptions on the t parameter of missing high incomes common to all countries. As a first result, the common assumption of perfect coverage of all incomes ($t = 1$) yields estimates which are compatible with the common results in previous studies presented in tables 3, 4, and 5, and in particular with those of Lakner and Milanovic (2016), Anand and Segal (2017), Niño-Zarazúa et al. (2017) and Jorda and Niño-Zarazúa (2016). These estimates clearly give further insight into the changes observed in figure 3, with an increase in the global annual mean income and a slight decrease in all inequality measures. The changes in income shares are also consistent with these changes, yet provide further insight into the changes at the top of the distribution, where despite the share of the top 10% slightly decreasing the share of the top 1% has increased.

The standard errors of these estimates are all very small, which evidences that very little variation is generated in the estimates within and across synthetic samples by the imputation mechanism in this case. These standard errors, however, reflect clearly how the statistics that vary the most are those

which are highly influenced by the imputed incomes in the upper tail of the distribution such as the Theil-T and the top 10% and top 1% shares. This is to be expected given the generally long tails in the fitted GB2 at the country level, which gives similar probabilities of being sampled in the imputations to high but very different incomes.

Moving on to the estimates obtained when the correction for missing incomes is performed, the first striking observation is the high sensitivity of the estimates to the assumed degree of missing high incomes. Although very little is known about the degree of missing high incomes in household surveys for the vast majority of countries in the sample, and the assumption that all countries share a common and fixed t parameter is certainly a strong one, assuming that any country's household survey's incomes are not representative of the richest 5% of the population ($t = .95$) increases the estimated Gini coefficient by approximately 10 points compared to the scenario where perfect coverage of these surveys is assumed ($t = 1$). This is the case for any of the two years analysed.

One important result concerning the role of the particular imputation mechanism used at the country level is that the statistics that are most sensitive to changes in the assumed value of t are those which are more sensitive to the upper part of the income distribution. To illustrate why this is the case, figure 4 presents empirical Lorenz curves for 10 synthetic samples of the GID under the baseline estimates ($t = 1$) and for 10 synthetic samples of the GID under the estimates corrected for missing high incomes with $t = .95$. A first observation is the existance of some few yet very atypical samples generated by the imputation model with corrections, yielding L-shaped Lorenz curves. This are clearly the result of some unrealisticaly high incomes being imputed for some individuals' at the country level. For the more realistic samples for which Lorenz curves are presented in the figure the share of incomes concentrated by the very top fractiles imply that the imputation model has a very heavy upper tail and as such there is a non-zero probability for individuals to be imputed very (very) high incomes. More

importantly, the largest discrepancies between these curves and the ones from the baseline imputation model occur at the top percentiles of the distribution, which explains why the statistics more sensitive to these differences are the most affected by the corrections for missing high incomes.

Another interesting result concerns the direction of the changes in inequality along the period. Although when a small degree of missing high incomes is assumed the conclusions regarding a drop in global income inequality between 1993 and 2008 are not affected, this is less clear for the higher degrees of correction. This feature of the correction method was already hinted at in Jorda and Niño-Zarazúa (2016), but exploring a wider range of scenarios regarding the assumed value of t and having an estimate of the standard errors of the inequality measures provides a clearer image of its behaviour: when higher degrees of missing high incomes are assumed the change in inequality between 1993 and 2008 is uncertain. This stems mainly from the heavy tail in the imputation model, which generates synthetic samples with very high levels of dispersion within the top percentiles of the distribution. The fact that the Theil-T statistic and the top 10% and top 1% shares are the most affected in relative terms by the changes in t for any of the two years is evidence of this. This increases all inequality measures and their standard errors to the point where, as is the case for $t = .90$, the inequality estimates for 2008 are above those for 1993, yet the fact that in this case the measures of inequality fall within one standard error of each other suggests that the overall change in inequality between both years is uncertain.

The GIC curve presented in figure 5 presents detailed insight into the changes on fractile mean incomes of the GID between 1993 and 2008. Three cases are presented, covering the case where no correction for missing high incomes is done, and two intermediate values for the t parameter. As a first observation, it's important to notice that, despite the heterogeneities on the estimated mean income growth rate for some fractiles of the distribution, the overall 'elephant' shape of the curve holds for all three cases. This is an important result as it evidences that regardless of the assumed degree of

missing high incomes, the overall relative changes along the GID between the two years are largely unchanged and all fractile groups experienced positive growth of their mean incomes along the period. The main differences introduced by the three different imputation models concern the growth rates of the mean incomes of three particular groups. Firstly, the mean income growth rate of the lowest ventile is increasing with the assumed degree of missing high incomes. This is caused by how the shape of the lower tail of the imputation model changes as t decreases. Figure 6 shows Gaussian kernel density estimates of a synthetic sample of the GID for each of the three cases considered in figure 5, plotted over a common range and with the same bandwidth. As can be seen, the difference between the estimated first ventiles for 1993 and 2008 is increased with changes t. The two other main discrepancies between the curves, which concern the top fractiles and those around the 60-th percentile, can be interpreted from this figure in a similar way. It is clear from these figures that the correction for missing high incomes shifts part of the mass of the distribution to the upper tail, extending the range and share of the very top percentiles and as such increases the uncertainty regarding the growth rate in the mean income of the top 1% of the distribution.

This last observation is strictly a virtue of the Multiple Imputations approach. If a simple imputation method had been used to generate the individual-level synthetic samples, the large uncertainty concerning the growth rate of the top 1% under a moderate degree of correction for missing high incomes could not have been evidenced.

5.3 Estimates under country-year specific and uncertain degree of under-reporting

There are many reasons why the assumptions made in the previos section regarding the degree of missing high incomes in the data can be overly restricting of the results. The first involves the unrealistic assumption that all

countries' surveys share an equal degree of under-sampling of high incomes. This is in strong contrast with the recent literature exploiting administrative tax data as a source for corrections at the country-level, which evidences very different degrees of under-sampling and under-reporting of high incomes for the few countries that have been studied (Higgins et al., 2018, Blanchet et al., 2018a, Angel et al., 2018). Secondly, the exact degree of missing high incomes in survey data is unknown for any country. Even if some bounds can be obtained through comparisons with administrative tax data or appropriate items from National Accounts, none of these sources are exempt of measuring errors themselves.

Table 7 presents Multiple Imputation estimates of the GID and global inequality measures under the less strict assumptions defined under scenario II. In this case, the assumption of a common and known t for all countries in the sample is substituted by an assumption of a common probability distribution for the values of t. This assumption reflects uncertainty regarding the specific value of t for any given country, yet it assumes bounds and a probability distribution between them reflecting assumptions about which values are relatively more likely than others. Without any external reference on which to calibrate this distribution for each country, or for the world, three alternative cases of a Beta distribution are explored over the range of values going from a moderate degree of correction ($t = .95$) to a value of $t = .999$ implying a minimal degree of correction for missing high incomes.

These estimates clearly express the differences in the assumed distribution of t, where as more mass of this distribution is shifted towards lower values the discrepancies with the case of no correction increase. This is particularly the case, once again, for the measures more sensitive to changes in the upper tail of the income distribution. Additionaly, the same positive association between the estimated standard errors and the degree of correction for missing high incomes as in the previous scenario is present in these estimates. The case which makes lower values of t in the $[.95, .999]$ range more relatively likely (i.e., $(\alpha_0, \beta_0) = (2, 4)$) is not only the most distant to the baseline

estimates but also the one with the largest estimated standard errors. This reflects the same behaviour concerning the imputation model as was the case in the previous section when considering moderate levels for t.

Figure 8 presents the estimated GIC under the assumptions of this scenario. By comparison with the GIC estimated under the assumption of a known and common degree of missing high incomes in all countries from figure 5, two central observations can be made. Firstly, the uncertainty regarding the degree of missing high incomes introduced in this scenario is translated directly into more variation in the estimated mean fractile income growth rates across synthetic samples. Whereas this variation was very narrow in the previous GIC estimates, it is considerable under this scenario. Such is the case that, except for a few fractile groups, all three cases include the baseline estimates within their range. Secondly, this figure also shares with the GIC presented in the previous section the robustness of the general 'elephant' shape of the GIC to the degree of correction for missing high incomes. In particular, the observation concerning the growing mean income growth rates for the very top percentiles with respect to those immediately before them in the distribution is not put into question by these results.

This last result is clearly determined by how the imputation method was approached. Because the only problem assumed in the data is that of missing high incomes, then a correction method at the country level like the one I explore will always yield higher estimates of the top 1% mean incomes than those without performing any corrections. Because the estimates without corrections for missing high incomes already evidence a change in the GID between 1993 and 2008 which determined a longer upper tail of the distribution, all corrected estimates can only amplify this change.

6 Concluding remarks

In this dissertation I have explored a Multiple Imputations framework for the estimation of the global interpersonal distribution of incomes (GID) and the dynamics of global income inequality between 1993 and 2008. In the context of estimating the GID, the available data sources with a reasonable degree of coverage and comparability present data at a grouped level. I have exploited a particular dataset, the LM-WPID, which compiles summary statistics on the income distribution at the country-year level using household surveys as its unique data source. In addition to only presenting grouped data on mean decile incomes, this dataset is prone to problems of under-reporting or under-sampling of high incomes in the underlying household surveys.

To correct for this issue, I extend on the parametric approach proposed by Jorda and Niño-Zarazúa (2016) which fits a GB2 distribution's Lorenz curve to the decile non-cumulative income shares implied in the data, with a correction re-scaling the ordinates of this Lorenz curve under an assumption on the magnitude of under-sampling of high incomes in the underlying survey data. This fitted GB2 is then used in my approach to generate many synthetic individual-level income samples for each country's population, where each individual is imputed an income drawn from this fitted distribution in each of such samples. This Multiple Imputations approach allows for obtaining many such synthetic income samples for the global population, on which complete-data methods can be used to make inference on the GID. Variations within and across these synthetic samples of the GID can then be used to estimate not only statistics of interest such as the global Gini coefficient or other inequality measures but can also be used to estimate the standard errors of such estimates. In doing so, I attempt a contribution to the literature on the GID by dealing with two issues largely left unattended to by the previous literature: the problem of missing high incomes in survey data and the statistical uncertainty surrounding global measures of inequality.

The two main results from my empirical analysis concern the effects that

even moderate levels of correction for missing high incomes can have on the measurement of global inequality. On the one side, despite exploring many alternative assumptions regarding the degree of missing high incomes in the data, the overall dynamic of the GID between 1993 and 2008 in terms of relative growth along different points of the distribution is not put into question by these. However, the levels and dynamics of global inequality measures are not so clear according to my estimates.

This uncertainty regarding the changes in concrete inequality measures between both years has largely to do with the particular imputation method I exploit. The estimated GB2 distributions at the country level, which defines the imputation model from which individual-level synthetic income samples are generated, can have a heavy upper tail which spans over very high levels of income when moderate degrees of under-sampling of high incomes are assumed. As such, the incomes of the top percentiles in the imputed samples are very likely to be unrealistic representations of the true incomes of individuals.

This problem concerning the imputation model's performance at the tail of the distribution constitutes the main limitation in my approach, and shares many characteristics with current research on the topic using parametric distributions to correct individual-level data for missing high incomes (Bourguignon, 2018, Hlasny and Verme, 2018). This problem has as an additional limitation the possibility to yield very atypical synthetic samples, which I have dealt with in an *ad-hoc* manner by simply omitting them from my estimates. Future research on the topic should explore alternative models for countries' income distributions which better represent their upper tails.

The second important limitation in my approach is that, while I have refered to under-reporting and under-sampling of high incomes as problems of the same nature, this imposes very strong assumptions. Under-reporting strictly refers to the problem of individuals included in the survey's coverage declaring a value for their income different to the true one. This has a

consequence on the distribution of incomes within the data itself. Under-sampling, on the other hand, refers to individuals not being covered by the survey. This has a consequence on the representativeness of the data. In the context of grouped data, the existence of both types of problems is a very intricate issue with no immediate hints on how to deal with.

Finally, a third limitation of my approach concerns the quality of the LM-WPID data. Despite the efforts put into compiling such a dataset from household surveys alone, there are some limitations which cannot be overcome. One particular limitation involves the treatment of income and expenditures indifferently, although the two represent different dimensions of welfare. Another limitation in the dataset which requires further attention is the fact that the data is not treated for zero incomes before the group-level summary statistics are calculated, such that zero incomes are not identifiable in the LM-WPID.

One interesting line for future research involves exploiting external data sources to calibrate the prior distributions for the approach I define under scenario II. This scenario involves a flexible definition of the degree of under-sampled high incomes in a country's data and can be adjusted on the basis of information not necessarily measuring the discrepancy between survey incomes and actual incomes. However, the estimates show that the method is very sensitive to the particular distribution assumed for this parameter. One possible option is to calibrate the 2 parameters (α_0, β_0) for the Beta distribution for a country such that its mode matches the preferable estimate of survey income under-sampling from the comparison with tax data and the comparison with National Accounts. As a second condition so as to allow identification of both parameters one could impose that its expected value matches the average between both such estimates. This might be a feasible approach for some countries, where this data is available or where previous results in the literature can be relied on. For other countries, where this might be unfeasible, regional estimates of this discrepancies could be used.

A second option worth exploring is built around the use of the Human Development Index along with a national constructed index of perceived corruption and social values regarding taxes, financial literacy, and preferences for government intervention in the economy, amongst others. This type of information can be obtained from detailed survey data as the World Values Survey and those published by the Global Barometer Surveys and its partner projects for regions and countries covering the whole world. A plausible assumption is that survey data on incomes from countries with higher levels of human development are less subject to under-reporting of high incomes, and that this is also the case for countries better ranked on an index expressing positive attitudes towards taxation, redistribution policies, and expressing low perceived corruption of the government. By assuming for each country or group of countries a range for the degree of income under-reporting in its household surveys (which could be obtained using as reference the same two data discrepancies as in the previous strategy), a 2-parameter logistic model can be fitted to predict a regional distribution depicting how the degree of under-reporting is distributed over the region's countries given their indices. The 2 parameters of this curve can be estimated using only the country with the best ranking according to both indices and the country with the worst. To do so, the extremes of the assumed range of under-reporting can be assumed to be the values of the logistic curve for this two extreme countries, which yields two conditions from which to estimate the 2 parameters. Given this parameters, all other countries in the region can be predicted a value for the degree of under-reported high incomes in their data given their values in both indices. The Beta prior's parameters can be determined so that this prior resembles the distribution of the predicted values over all countries in the region, and used as the prior in the imputation model for each and all countries' in the region indifferently.

One final line for future research involves studying regional dynamics and regional composition of the GID using this Multiple Imputations framework. This important dimension of the dynamics of global inequality and the GID was not dealt with in this case as it requires working with a sample of the

LM-WPID different than what I have exploited, as the common sample for the 1993-2008 period lacks reasonable coverage in some regions of the globe.

References

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). The elephant curve of global inequality and growth. In *AEA Papers and Proceedings*, volume 108, pages 103–08.

Anand, S. and Segal, P. (2008). What do we know about global income inequality? *Journal of Economic Literature*, 46(1):57–94.

Anand, S. and Segal, P. (2017). Who are the global top 1%? *World Development*, 95:111–126.

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2018). What did you really earn last year?: explaining measurement error in survey income data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Barro, R. J. and Sala-i-Martin, X. (1991). Convergence across states and regions. *Brookings papers on economic activity*, pages 107–182.

Berry, A., Bourguignon, F., and Morrison, C. (1983). Changes in the world distribution of income between 1950 and 1977. *The Economic Journal*, 93(370):331–350.

Bhalla, S. S. (2002). *Imagine there's no country: Poverty, inequality, and growth in the era of globalization*. Peterson Institute.

Blanchet, T., Flores, I., and Morgan, M. (2018a). The weight of the rich: Improving surveys using tax data. *WID. world Working Paper*, 12.

Blanchet, T., Fournier, J., and Piketty, T. (2017). Generalized pareto curves: Theory and applications. *WID.world Working Paper 2017/3*.

Blanchet, T., Garbinti, B., Goupille-Lebret, J., and Martínez-Toledano, C. (2018b). Applying generalized pareto curves to inequality analysis. In *AEA Papers and Proceedings*, volume 108, pages 114–18.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. *The Journal of Economic Inequality*, 16(2):171–188.

Bourguignon, F. and Morrisson, C. (2002). Inequality among world citizens: 1820-1992. *American economic review*, 92(4):727–744.

Castillo, E., Hadi, A. S., and Sarabia, J. M. (1998). A method for estimating lorenz curves. *Communications in statistics-theory and methods*, 27(8):2037–2063.

Chen, S. and Ravallion, M. (2004). How have the world's poorest fared since the early 1980s? *The World Bank Research Observer*, 19(2):141–169.

Chen, S. and Ravallion, M. (2010). The developing world is poorer than we thought, but no less successful in the fight against poverty. *The Quarterly Journal of Economics*, 125(4):1577–1625.

Chotikapanich, D. and Griffiths, W. E. (2002). Estimating lorenz curves using a dirichlet distribution. *Journal of Business & Economic Statistics*, 20(2):290–295.

Chotikapanich, D., Griffiths, W. E., Prasada Rao, D., and Valencia, V. (2012). Global income distributions and inequality, 1993 and 2000: Incorporating country-level inequality modeled with beta distributions. *Review of Economics and Statistics*, 94(1):52–73.

Chotikapanich, D., Griffiths, W. E., and Rao, D. P. (2007). Estimating and combining national income distributions using limited data. *Journal of Business & Economic Statistics*, 25(1):97–109.

Chotikapanich, D., Valenzuela, R., and Rao, D. P. (1997). Global and regional inequality in the distribution of income: estimation with limited and incomplete data. *Empirical Economics*, 22(4):533–546.

Darvas, Z. (2016). *Some are more equal than others: new estimates of global and regional inequality.* Number MT-DP-2016/35. IEHAS Discussion Papers.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and statistics*, 87(1):1–19.

Deininger, K. and Squire, L. (1996). A new data set measuring income inequality. *The World Bank Economic Review*, 10(3):565–591.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media.

Fournier, J. (2015). Generalized pareto curves: Theory and application using income and inheritance tabulations for france 1901-2012. Master's thesis, Paris School of Economics.

Hajargasht, G., Griffiths, W. E., Brice, J., Rao, D. P., and Chotikapanich, D. (2012). Inference for income distributions using grouped data. *Journal of Business & Economic Statistics*, 30(4):563–575.

Hajargasht, G., Griffiths, W. E., et al. (2016). *Inference for Lorenz curves.* University of Melbourne, Department of Economics.

Higgins, S., Lustig, N., Vigorito, A., et al. (2018). The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data. Technical report.

Hlasny, V. and Verme, P. (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the eu silc data. *Econometrics*, 6(2):30.

Hong, S., Han, H., and Kim, C. S. (2019). World distribution of income for 1970–2010: dramatic reduction in world income inequality during the 2000s. *Empirical Economics*, pages 1–34.

Jenkins, S. P., Burkhauser, R. V., Feng, S., and Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1):63–81.

Jones, C. I. (1997). On the evolution of the world income distribution. *journal of Economic Perspectives*, 11(3):19–36.

Jorda, V. and Niño-Zarazúa, M. (2016). Global inequality. how large is the effect of top incomes? *UNU-WIDER Working Paper 2016/94*.

Jorda, V., Sarabia, J. M., and Jäntti, M. (2018). Estimation of income inequality from grouped data. *arXiv preprint arXiv:1808.09831*.

Kanbur, R. (2015). Globalization and inequality. In *Handbook of income distribution*, volume 2, pages 1845–1881. Elsevier.

Kleiber, C. and Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*, volume 470. John Wiley & Sons.

Kobayashi, G. and Kakamu, K. (2019). Approximate bayesian computation for lorenz curves from grouped data. *Computational Statistics*, 34(1):253–279.

Lakner, C. and Milanovic, B. (2016). Global income distribution: From the fall of the berlin wall to the great recession. *The World Bank Economic Review*, 30(2):203–232.

Liberati, P. (2015). The world distribution of income and its inequality, 1970–2009. *Review of Income and Wealth*, 61(2):248–273.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663.

Milanovic, B. (2002). True world income distribution, 1988 and 1993: First calculation based on household surveys alone. *The Economic Journal*, 112(476):51–92.

Milanovic, B. (2011). *Worlds apart: Measuring international and global inequality*. Princeton University Press.

Milanovic, B. (2012). Global inequality recalculated and updated: the effect of new ppp estimates on global inequality and 2005 estimates. *The Journal of Economic Inequality*, 10(1):1–18.

Niño-Zarazúa, M., Roope, L., and Tarp, F. (2017). Global inequality: Relatively lower, absolutely higher. *Review of Income and Wealth*, 63(4):661–684.

Pinkovskiy, M. and Sala-i-Martin, X. (2009). Parametric estimations of the world distribution of income. Technical report, National Bureau of Economic Research.

Podder, N. (1994). A profile of international inequality. *Journal of Income Distribution*, 3(2):5–5.

Quah, D. (1997). Empirics for growth and distribution: polarization, stratification and convergence club. *Journal of Economic Growth*, 2(1):27–59.

Ravallion, M. and Chen, S. (2003). Measuring pro-poor growth. *Economics letters*, 78(1):93–99.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):538–543.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics.

Sala-i-Martin, X. (2006). The world distribution of income: falling poverty and... convergence, period. *The Quarterly Journal of Economics*, 121(2):351–397.

Sarabia, J. M. (2008). Parametric lorenz curves: Models and applications. In *Modeling income distributions and Lorenz curves*, pages 167–190. Springer.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC.

Theil, H. (1979). World income inequality and its components. *Economics Letters*, 2(1):99–102.

Theil, H. and Seale, J. L. (1994). The geographic distribution of world income, 1950–1990. *De Economist*, 142(4):387–419.

Region	Countries	Population 1993 (%)	Population 2008 (%)
Mature economies	Austria, Belgium, Bulgaria, Canada, Switzerland, Czech Republic, Germany, Denmark, Spain, Estonia, Finland, France, United Kingdom, Greece, Hungary, Ireland, Israel, Italy, Japan, Korea Rep., Lithuania, Luxembourg, Latvia, Netherlands, Norway, Poland, Portugal, Romania, Singapore, Slovak Republic, Slovenia, Sweden, Taiwan, United States	0.176	0.157
M. East and N. Africa	Egypt, Jordan, Morocco	0.016	0.017
L. America and Carib.	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Honduras, Mexico, Nicaragua, Panama, Peru, Paraguay, El Salvador, Uruguay, Venezuela	0.077	0.078
Russia, C. Asia, and SE Europe	Azerbaijan, Kyrgyz Republic, Russian Federation, Turkey, Ukraine	0.049	0.04
Sub-Saharan Africa	Burundi, Burkina Faso, Central African Republic, Cote d'Ivoire, Guinea, Kenya, Madagascar, Mali, Mauritania, Niger, Nigeria, Swaziland, Tanzania, Uganda, South Africa, Zambia	0.054	0.064
China		0.213	0.196
India		0.167	0.176
Other Asia	Bangladesh, Indonesia, Cambodia, Laos, Sri Lanka, Malaysia, Pakistan, Philippines, Thailand, Vietnam	0.122	0.127
Total		0.873	0.855

Table 2: Composition and World population coverage of the 1993-2008 LM-WPID common sample.

Paper	Data	Method	Year	Countries	Gini	MLD	Theil-T	Top 10%	Top 1%
Bourguignon and Morrisson (2002)	Historical GDP and decile shares series. 1990 PPP dollars.	IQI	1992	33	0.657	0.827	0.855	53.4	-
Milanovic (2002)	Household Surveys. Income and expenditures alike. 1993 PPP dollars	IQI	1988	91	0.628	-	0.765	46.9	9.3
			1993	91	0.66	-	0.873	50.8	9.5
Lakner and Milanovic (2016)	Household Surveys. Income and expenditures alike. 2005 PPP dollars	IQI. Correction imputing discrepancy with NA household consumption per capita.	1988	75	(0.722,0.763)	1.142	1.022	57.2	11.8
			1993	121	(0.719,0.761)	1.107	1.024	58.1	12
			1998	121	(0.715,0.772)	1.071	1.028	64.12	18.6
			2003	121	(0.719,0.781)	1.076	1.049	60.76	14.2
			2008	121	(0.696,0.759)	1.027	1.003	60.3	15.6
Anand and Segal (2017)	Milanovic (2012) and top 1% shares from World Top Incomes Database (WID.world)	IQI. Correction using Pareto imputation for top 0.1%.	1988	92	0.71	1.014	1.061	56.1	17
			1993	104	0.702	1.013	1.062	56.8	17.2
			1998	109	0.696	0.971	1.100	57.7	19
			2005	119	0.702	1.023	1.150	57.8	20.2

Table 3: Previous estimates of global interpersonal inequality.

Paper	Data	Method	Year	Countries	Gini	MLD	Theil-T	Top 10%	Top 1%
Sala-i-Martin (2006)	GDP and quintile shares. PPP dollars.	Gaussian kernel density estimation	1988	138	0.649	0.842	0.808	-	-
			1993	138	0.64	0.819	0.787	-	-
			1998	138	0.638	0.816	0.785	-	-
Hong et al. (2019)	GDP and quintile shares. PPP dollars. Top 1% and 5% shares from WTID.	Gaussian kernel density estimation.	1988	188	0.687	0.975	0.976	-	-
			1993	188	0.694	0.992	1.016	-	-
			1998	188	0.682	0.965	0.906	-	-
			2003	188	0.671	0.926	0.869	-	-
			2008	188	0.658	0.879	0.942	-	-
Bhalla (2002)	GDP and quantile shares. 1993 PPP dollars	Parametric Lorenz curve.	1988	130	≈ 0.672	-	-	-	-
			1993	130	≈ 0.67	-	-	-	-
			2000	130	≈ 0.652	-	-	-	-
Chotikapanich et al. (1997)	GDP and Gini coefficients. PPP dollars.	Log-normal distribution.	1985	36	0.646	0.802	-	-	-
			1990	36	0.648	0.805	-	-	-
Pinkovskiy and Sala-i-Martin (2009)	GDP and quintile shares. PPP dollars.	Log-normal distribution.	1988	≈ 187	0.648	0.849	0.797	-	-
			1993	≈ 187	0.648	0.841	0.802	-	-
			1998	≈ 187	0.637	0.806	0.785	-	-
			2003	187	0.623	0.775	0.740	-	-

Table 4: Previous estimates of global interpersonal inequality.

Paper	Data	Method	Year	Countries	Gini	MLD	Theil-T	Top 10%	Top 1%
Liberati (2015)	GDP per capita and Gini coefficients. PPP dollars.	Log-normal distribution	1988	164	0.695	-	-	-	-
			1993	186	0.684	-	-	-	-
			1998	187	0.686	-	-	-	-
			2003	188	0.675	-	-	-	-
			2008	189	0.659	-	-	-	-
Niño-Zarazúa et al. (2017)	GDP per capita and quantile shares. PPP dollars.	Log-normal distribution	1985	86	0.708	-	-	-	-
			1995	122	0.705	-	-	-	-
			2005	135	0.68	-	-	-	-
Chotikapanich et al. (2012)	GDP per capita and quantile shares. PPP dollars.	Beta 2 distribution.	1993	91	0.648	0.813	-	50	11.4
			2000	91	0.64	0.795	-	50.9	12.5
Jorda and Niño-Zarazúa (2016)	GDP per capita and quantile shares. PPP dollars.	GB2 distribution. Truncation correction for missing top incomes.	1990	117	-	(1.087,1.507)	(0.934,1.083)	-	-
			1995	129	-	(1.045,1.472)	(0.934,1.107)	-	-
			2005	146	-	(0.927,1.319)	(0.813,0.970)	-	-

Table 5: Previous estimates of global interpersonal inequality.

Figure 1: Baseline estimates Lorenz curves for 5 selected countries

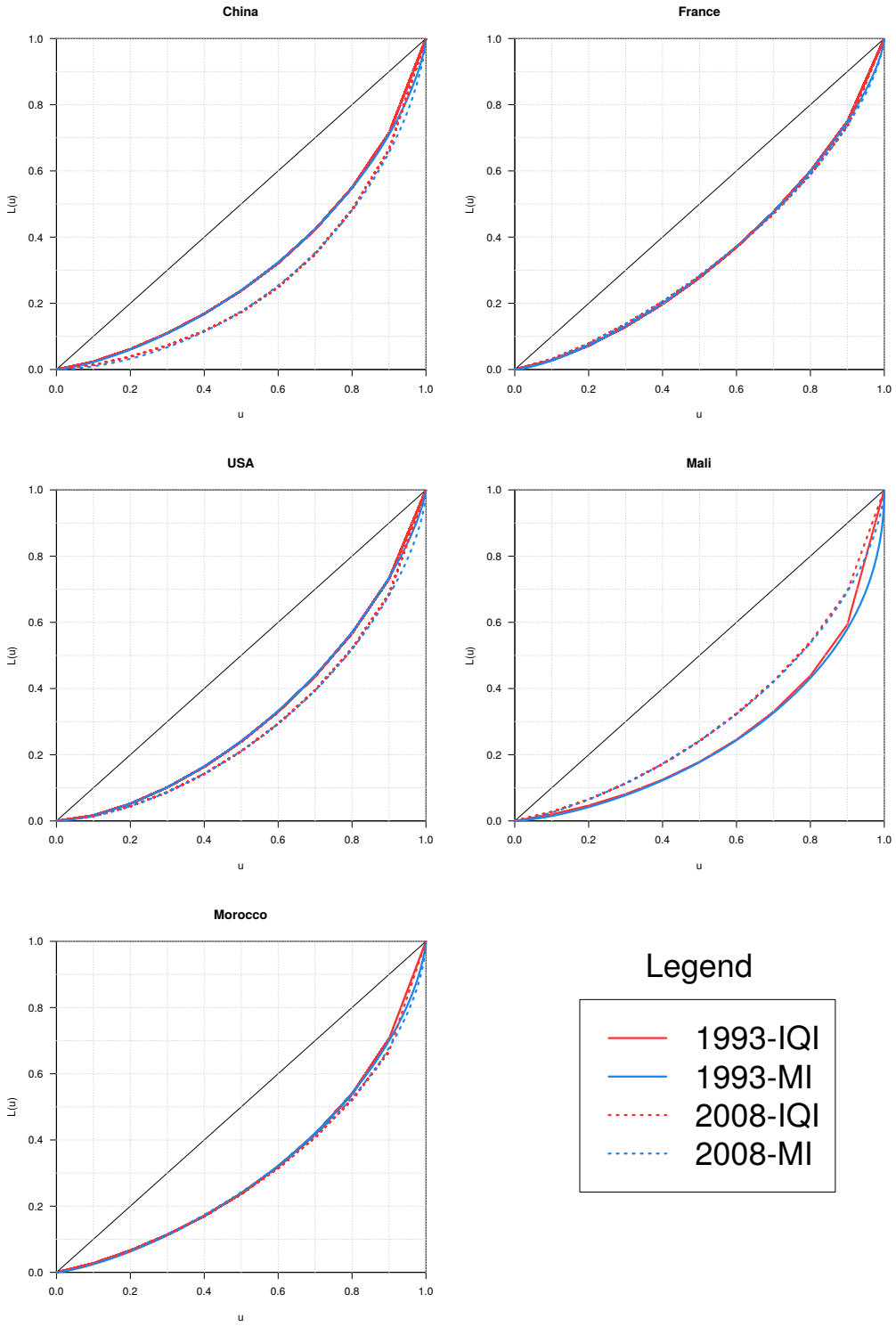


Figure 2: Baseline estimate comparisons with World Bank estimates.

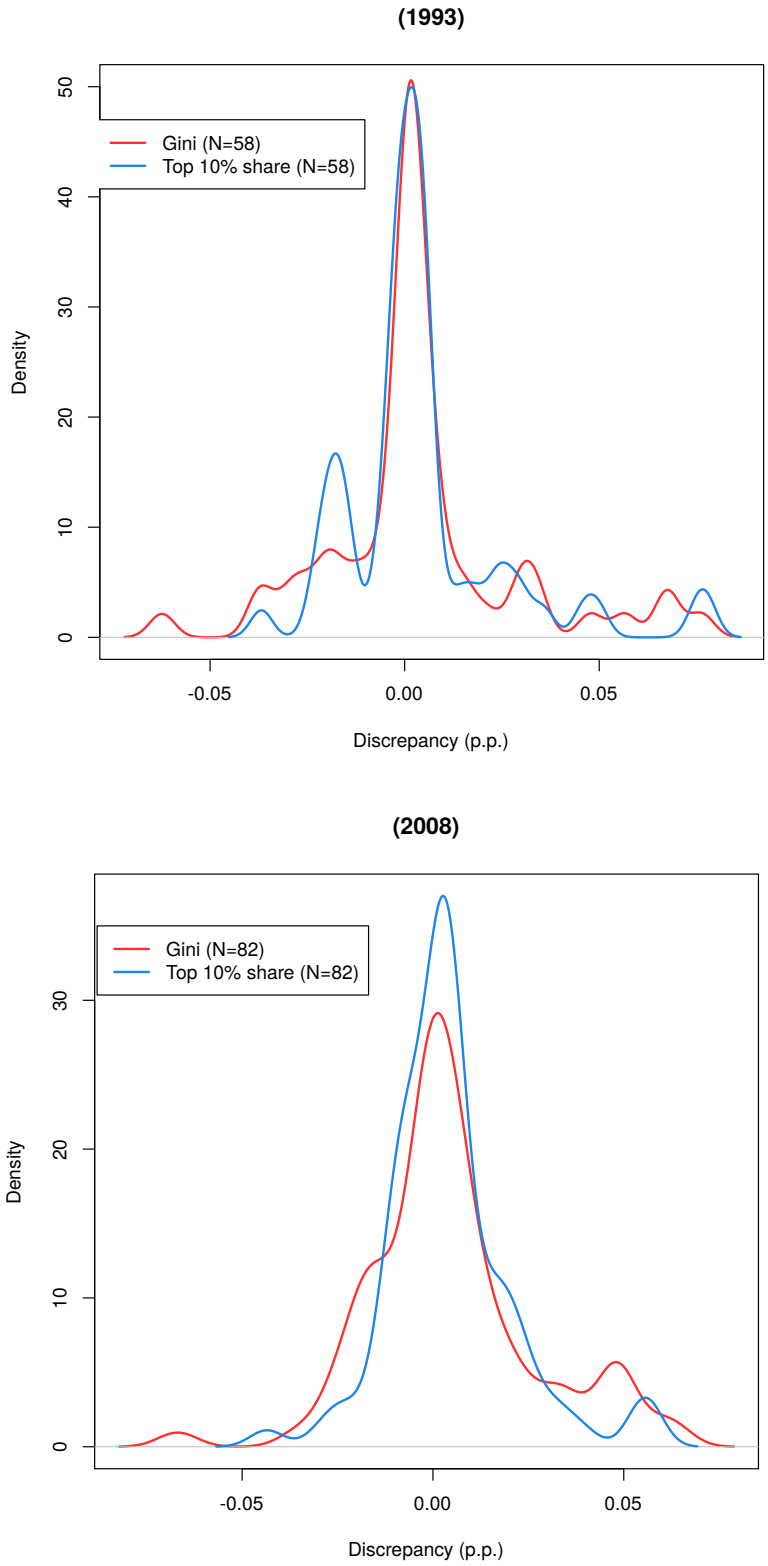


Figure 3: Gaussian Kernel Density estimates over imputed samples of the GID (1993-2008). Baseline imputation method

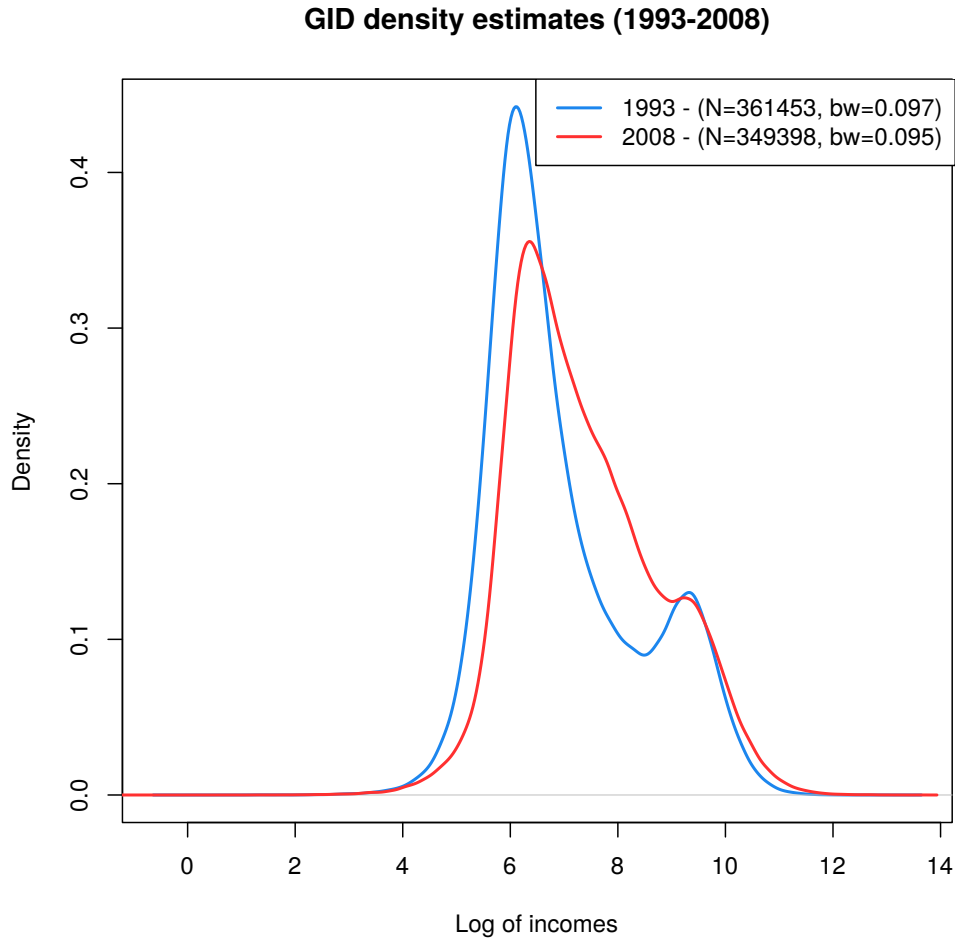


Table 6: Multiple Imputation estimates under Scenario I

	1993						
	t=1	t=.995	t=.99	t=.985	t=.98	t=.95	t=.90
Gini	0.7252	0.7357	0.7441	0.7545	0.7648	0.8292	0.8654
	(0.0006)	(0.0017)	(0.0028)	(0.0047)	(0.0054)	(0.0226)	(0.0183)
MLD	1.1396	1.1824	1.2157	1.26	1.3071	1.6493	1.9334
	(0.0026)	(0.0071)	(0.011)	(0.0202)	(0.0228)	(0.141)	(0.1443)
Theil-T	1.0605	1.1522	1.2437	1.3864	1.5022	2.9134	3.3076
	(0.0042)	(0.0281)	(0.0601)	(0.1229)	(0.1278)	(0.921)	(0.8792)
Mean	3345.8204	3551.8608	3726.933	3944.813	4177.768	6328.238	9229.5714
	(12.7445)	(26.4735)	(43.0928)	(78.8648)	(100.3637)	(960.5122)	(1479.8618)
Income Shares							
Top 10%	0.5833	0.5978	0.6097	0.6248	0.64	0.737	0.7935
	(0.0011)	(0.0027)	(0.0041)	(0.0072)	(0.0081)	(0.0357)	(0.0279)
Top 1%	0.1323	0.1594	0.1818	0.2105	0.2386	0.4253	0.5109
	(0.0013)	(0.005)	(0.0089)	(0.0151)	(0.0171)	(0.0779)	(0.0675)
Middle 40%	0.1617	0.1564	0.1522	0.1466	0.1412	0.1057	0.0865
	(0.0006)	(0.0011)	(0.0016)	(0.0029)	(0.0033)	(0.0143)	(0.0118)
Bottom 50%	0.0592	0.056	0.0536	0.0509	0.0481	0.0328	0.0231
	(0.0002)	(0.0004)	(0.0006)	(0.001)	(0.0011)	(0.0044)	(0.0031)
	2008						
	t=1	t=.995	t=.99	t=.985	t=.98	t=.95	t=.90
Gini	0.6973	0.711	0.7226	0.734	0.7463	0.8144	0.903
	(0.0009)	(0.002)	(0.003)	(0.0041)	(0.0074)	(0.0187)	(0.0259)
MLD	1.0133	1.0626	1.1065	1.1496	1.1989	1.526	2.2505
	(0.0031)	(0.0069)	(0.0111)	(0.016)	(0.0294)	(0.1059)	(0.3569)
Theil-T	0.988	1.0942	1.2095	1.328	1.4985	2.6153	4.7106
	(0.0061)	(0.0282)	(0.0531)	(0.0833)	(0.1692)	(0.6472)	(1.6797)
Mean	4394.4293	4705.5142	4987.0234	5286.559	5623.7231	8362.029	19997.2543
	(17.4262)	(34.018)	(58.8081)	(87.6112)	(167.9759)	(953.6878)	(12441.8841)
Income Shares							
Top 10%	0.5601	0.5783	0.5941	0.61	0.6274	0.7269	0.8578
	(0.0014)	(0.0028)	(0.0045)	(0.0058)	(0.0106)	(0.0271)	(0.0385)
Top 1%	0.1492	0.1782	0.2045	0.2306	0.2605	0.4251	0.661
	(0.0018)	(0.0051)	(0.0082)	(0.0112)	(0.0209)	(0.058)	(0.0907)
Middle 40%	0.2052	0.1962	0.1883	0.1807	0.1726	0.126	0.0656
	(0.0007)	(0.0014)	(0.0022)	(0.0028)	(0.005)	(0.0125)	(0.0176)
Bottom 50%	0.0698	0.0656	0.0623	0.0592	0.056	0.0389	0.0181
	(0.0003)	(0.0005)	(0.0007)	(0.0009)	(0.0016)	(0.0039)	(0.0048)

Average estimated values over the 80% less extreme values for each statistic's distribution across $m = 50$ synthetic samples. Multiple Imputation standard errors over these same samples in parenthesis.

Figure 4: Empirical Lorenz curves over corrected and baseline GID samples.

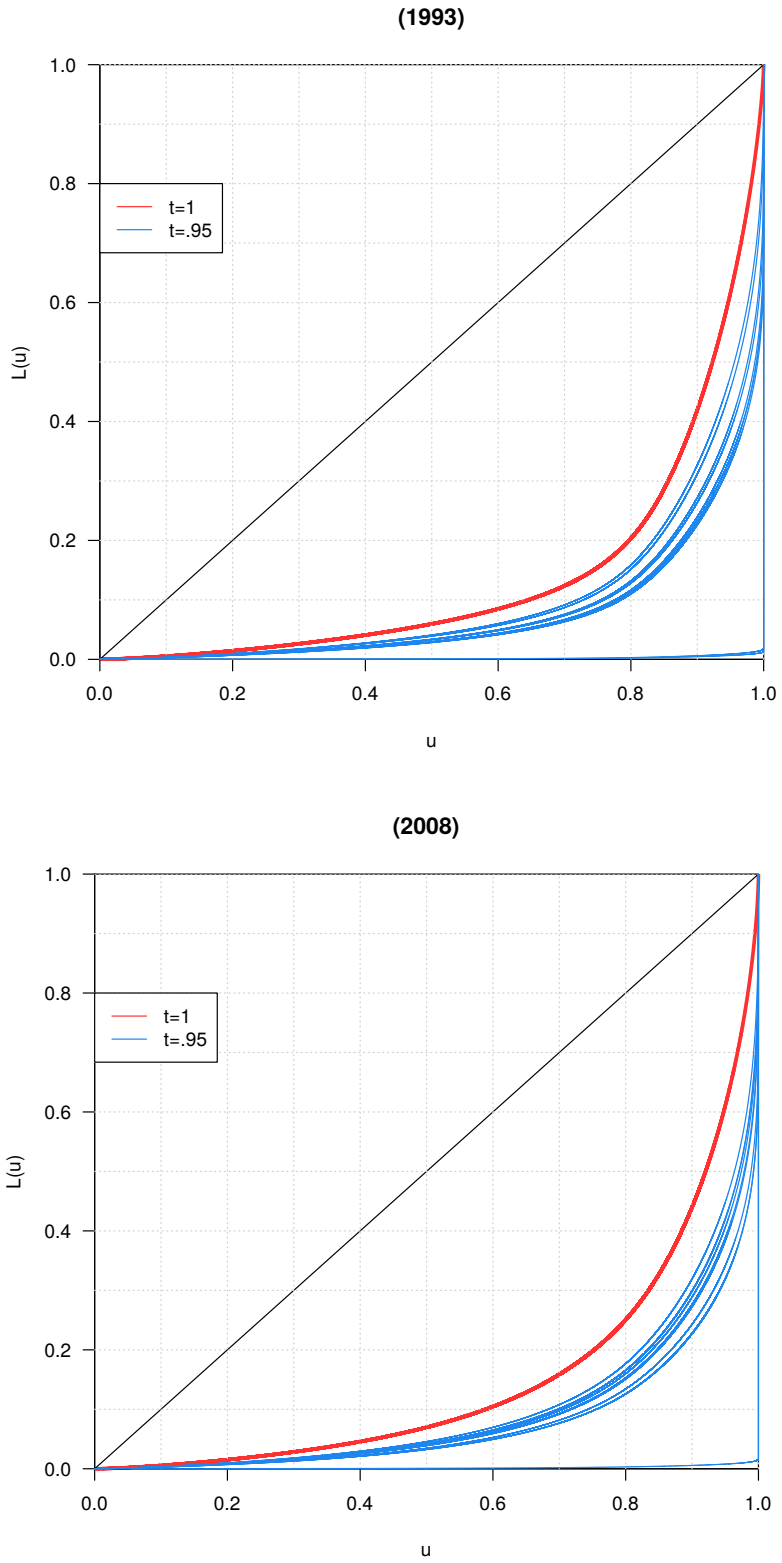
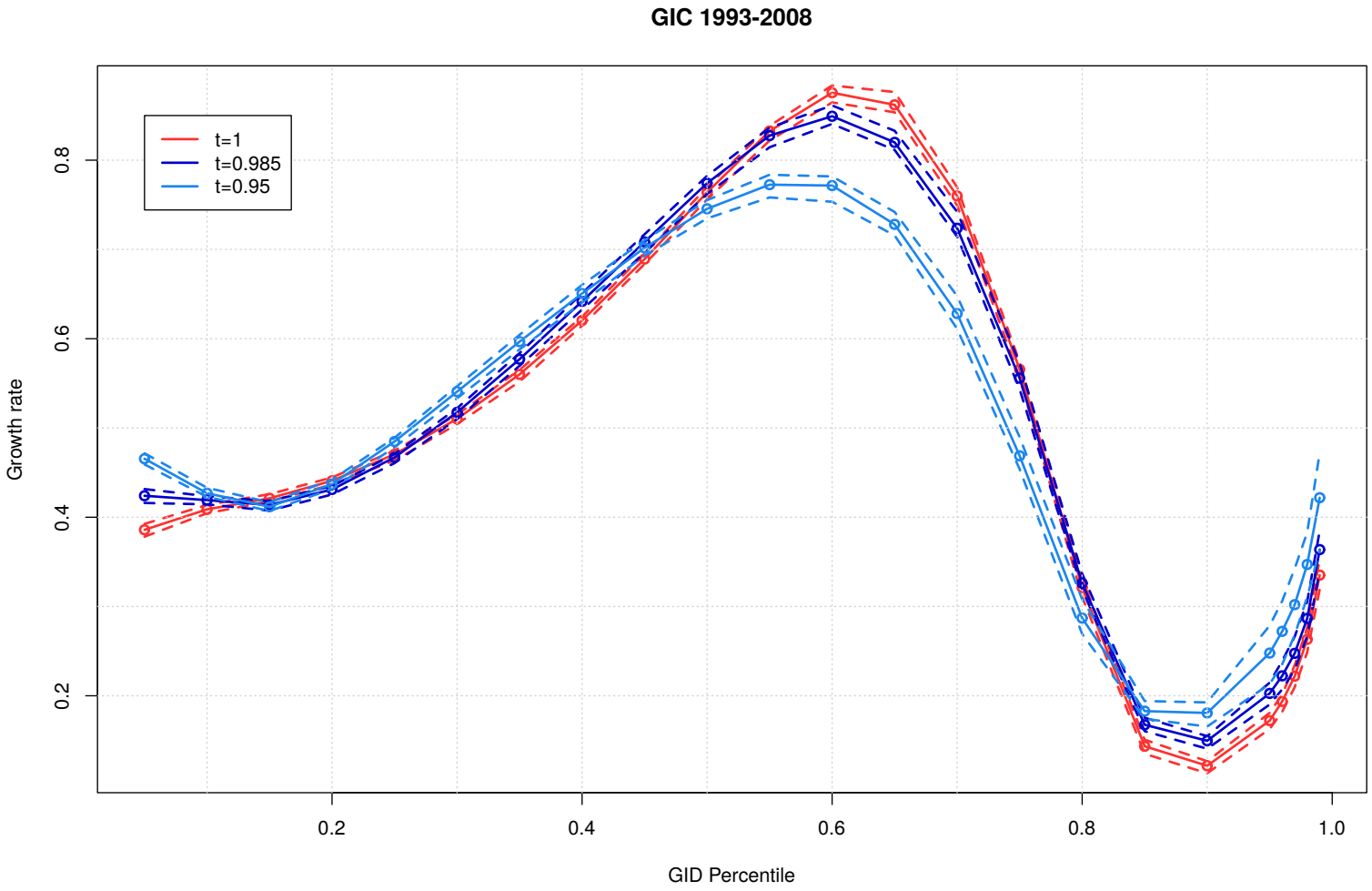
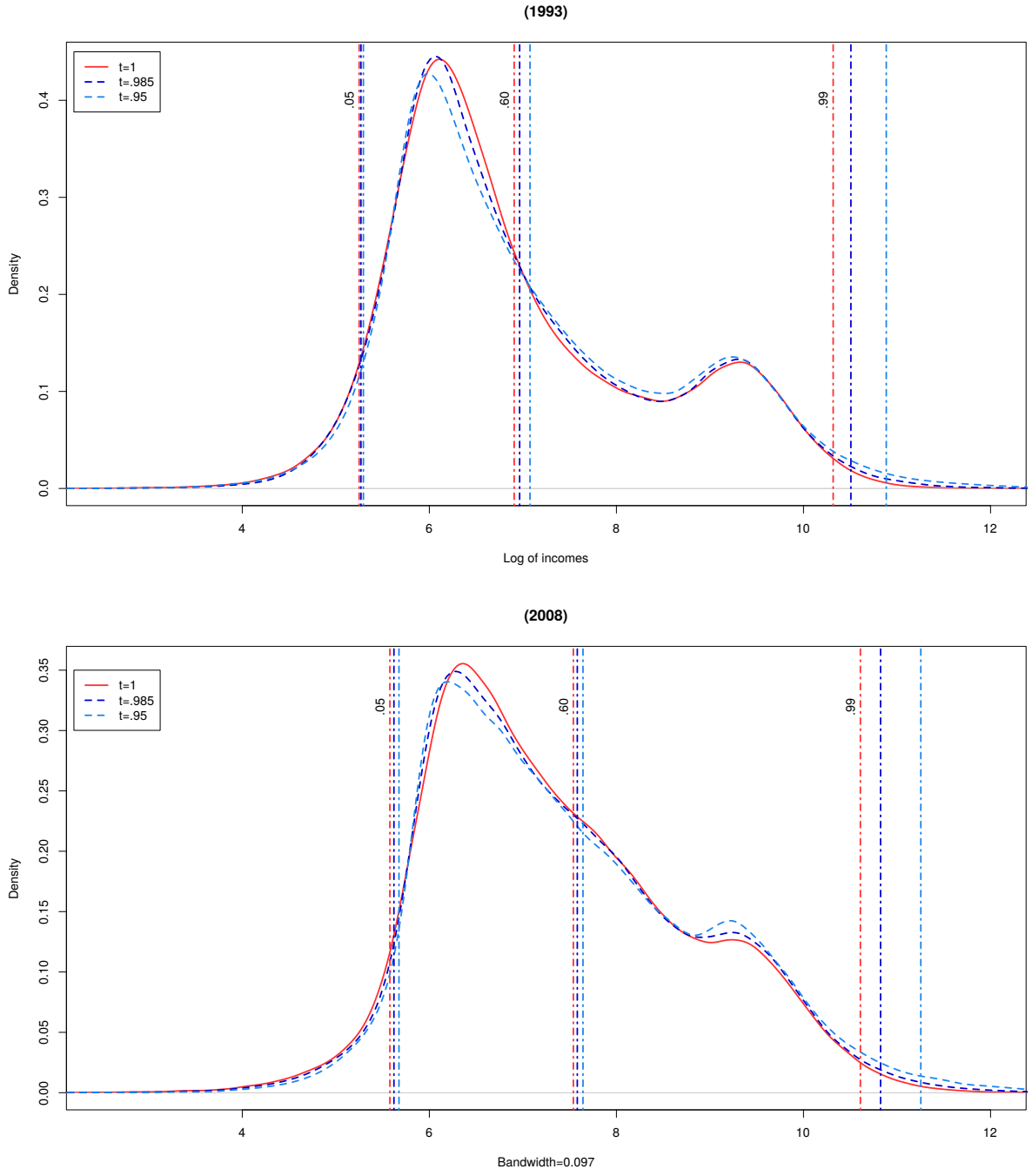


Figure 5: Growth Incidence Curve estimate under Scenario I



Note: Dashed lines show maximum and minimum estimated GIC coordinates for each fractile group over all synthetic samples. $m = 50$

Figure 6: Gaussian kernel density estimates over samples of the GID under scenario I.



Note: Dashed lines show sample estimates of the .05, .60, and .99 quantiles of each distribution.

Figure 7: Beta distributions assumed for t under Scenario II

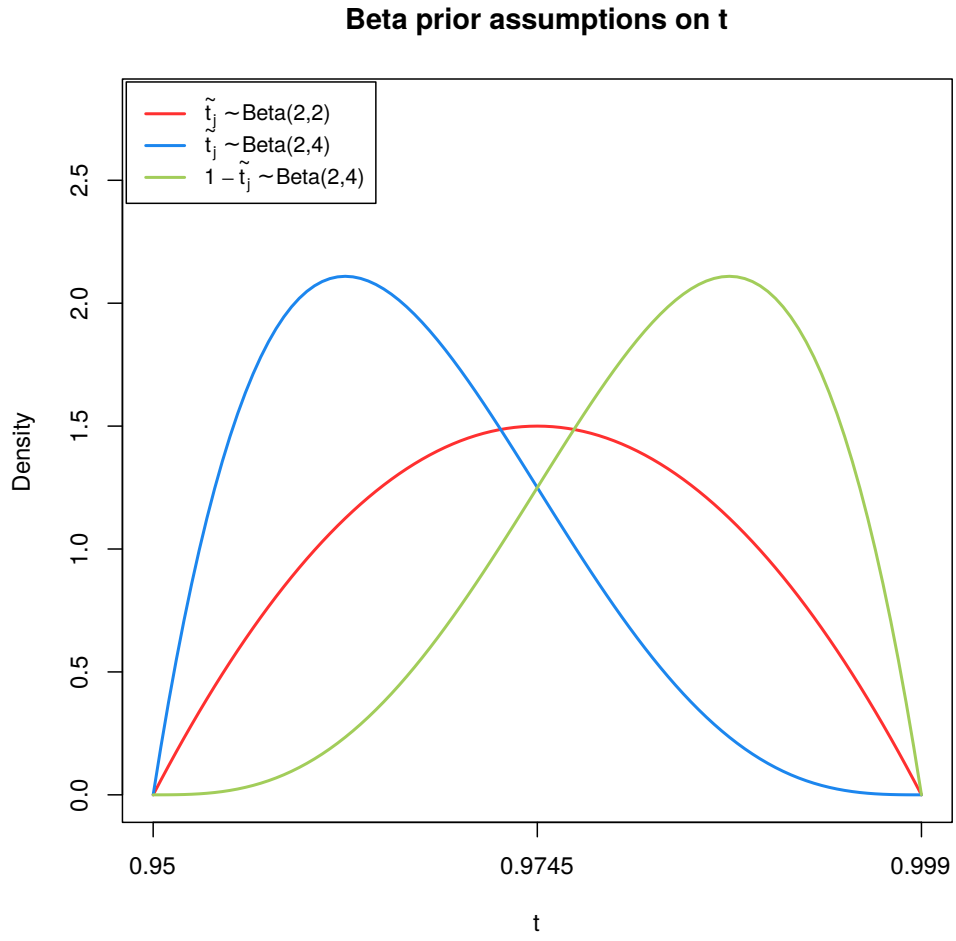
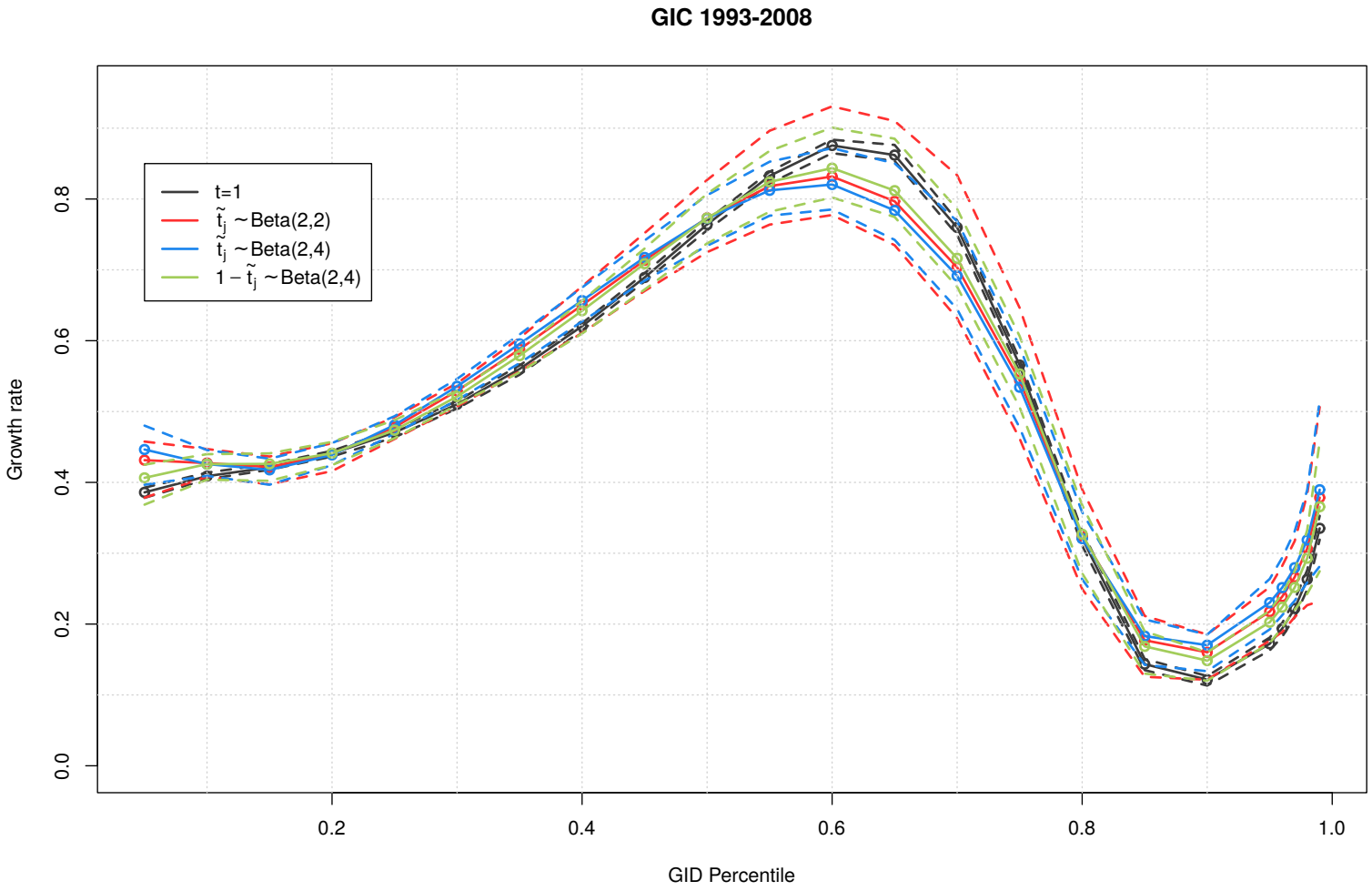


Table 7: Multiple Imputation estimates under Scenario II

	1993			
	t_j=1	$\tilde{t}_j \sim Beta(2,2)$	$\tilde{t}_j \sim Beta(2,4)$	$1-\tilde{t}_j \sim Beta(2,4)$
Gini	0.7252	0.7777	0.7932	0.7601
	(0.0006)	(0.0092)	(0.014)	(0.007)
MLD	1.1396	1.3669	1.4447	1.2856
	(0.0026)	(0.0426)	(0.0704)	(0.0294)
Theil-T	1.0605	1.7237	2.0243	1.4589
	(0.0042)	(0.2722)	(0.4644)	(0.1713)
Mean	3345.8204	4502.4598	4968.7429	4067.7417
	(12.7445)	(196.8613)	(364.7849)	(134.1308)
Income Shares				
Top 10%	0.5833	0.6593	0.6824	0.6331
	(0.0011)	(0.014)	(0.0217)	(0.0106)
Top 1%	0.1323	0.2745	0.3188	0.2264
	(0.0013)	(0.0302)	(0.0464)	(0.0219)
Middle 40%	0.1617	0.1338	0.1257	0.1435
	(0.0006)	(0.0057)	(0.0087)	(0.0043)
Bottom 50%	0.0592	0.045	0.0411	0.0494
	(0.0002)	(0.0019)	(0.0028)	(0.0015)
	2008			
	t_j=1	$\tilde{t}_j \sim Beta(2,2)$	$\tilde{t}_j \sim Beta(2,4)$	$1-\tilde{t}_j \sim Beta(2,4)$
Gini	0.6973	0.7593	0.7746	0.74
	(0.0009)	(0.0096)	(0.0095)	(0.0068)
MLD	1.0133	1.2547	1.3239	1.1738
	(0.0031)	(0.0405)	(0.0426)	(0.0269)
Theil-T	0.988	1.6502	1.8317	1.4028
	(0.0061)	(0.2093)	(0.2388)	(0.1329)
Mean	4394.4293	6053.235	6623.5302	5445.9185
	(17.4262)	(265.6551)	(293.7935)	(169.7988)
Income Shares				
Top 10%	0.5601	0.6459	0.6682	0.6187
	(0.0014)	(0.0145)	(0.0139)	(0.0099)
Top 1%	0.1492	0.2899	0.3251	0.2452
	(0.0018)	(0.0269)	(0.0272)	(0.0198)
Middle 40%	0.2052	0.1639	0.1536	0.1767
	(0.0007)	(0.0067)	(0.0065)	(0.0047)
Bottom 50%	0.0698	0.0524	0.0484	0.0576
	(0.0003)	(0.0022)	(0.0021)	(0.0016)

Average estimated values over the 80% less extreme values for each statistic's distribution across $m = 100$ synthetic samples. Multiple Imputation standard errors over these same samples in parenthesis. $\tilde{t}_j \equiv \frac{t_j - 0.95}{(0.999-0.95)}$

Figure 8: Growth Incidence Curve estimate under Scenario II



Note: Dashed lines show maximum and minimum estimated GIC coordinates for each fractile group over all synthetic samples. $m = 100$