Shorrocks, Anthony; Wan, Guanghua

Working Paper

Ungrouping income distributions: Synthesising samples for inequality and poverty analysis

WIDER Research Paper, No. 2008/16

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

Suggested Citation: Shorrocks, Anthony; Wan, Guanghua (2008) : Ungrouping income distributions: Synthesising samples for inequality and poverty analysis, WIDER Research Paper, No. 2008/16, ISBN 978-92-9230-058-6, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki

This Version is available at:
https://hdl.handle.net/10419/63307

UNITED NATIONS UNIVERSITY
UNU-WIDER
World Institute for Development Economics Research

Research Paper No. 2008/16

Ungrouping Income Distributions

Synthesising Samples for Inequality and Poverty Analysis

Anthony Shorrocks and
Guanghua Wan*

February 2008

Abstract

We describe a new method of facilitating inequality and poverty analysis of grouped distributional data by allowing individual income observations to be reconstructed from any feasible grouping pattern. In contrast to earlier methods, our procedure ensures that the characteristics of the synthetic sample exactly match the reported values. The performance of the algorithm is evaluated first by using household survey records to compare true income observations with their synthetic counterparts, then by comparing the true and generated values of the Gini coefficient and other inequality indices. The results indicate that the new technique is capable of reproducing individual data from grouped statistics with a high degree of accuracy.

Keywords: grouped data, income distribution, inequality, poverty

JEL classification: C81, D31

*both UNU-WIDER, Helsinki

This study has been prepared within the UNU-WIDER project on New Directions.

UNU-WIDER acknowledges the financial contributions to the research programme by the governments of Denmark (Royal Ministry of Foreign Affairs), Finland (Ministry for Foreign Affairs), Norway (Royal Ministry of Foreign Affairs), Sweden (Swedish International Development Cooperation Agency—Sida) and the United Kingdom (Department for International Development).

ISSN 1810-2611 ISBN 978-92-9230-058-6

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

www.wider.unu.edu *publications@wider.unu.edu*

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Lorraine Telfer-Taivainen at UNU-WIDER

1 Introduction

Research on inequality and poverty during the past half century has been greatly influenced by the writings of Amartya Sen. Amongst his many and varied contributions, the Radcliffe Lectures on inequality (Sen 1973) and the *Econometrica* article on poverty measurement (Sen 1976) spurred countless numbers of readers to attempt to confront distributional questions with empirical evidence. Yet when these publications were written a generation ago, anyone wishing to undertake empirical work faced serious handicaps. Computing hardware and software were rudimentary by current standards. Household microdata were in short supply and rarely accessible to independent researchers. Those interested in inequality, poverty, and other distribution-related issues usually had to be content with simple computational procedures applied to summary statistics, grouped frequency tables, and other types of published secondary material.

To compensate for the shortcomings, a variety of ingenious procedures and tools were developed. Many different functional forms were offered as approximations to the empirical income distributions and compared to the alternatives. Examples of this genre are provided by Gastwirth (1972), Salem and Mount (1974), Kakwani (1976), Kakwani and Podder (1973, 1976), Singh and Maddala (1976), Kloek and Van Dijk (1978), McDonald and Ransom (1979, 1981), Harrison (1981) and McDonald (1984).[1] Other authors, including Cowell and Mehta (1982) proposed ways of estimating inequality indices from grouped data. Attention was also given to the optimal way of summarizing data in order to preserve distributional information (Davies and Shorrocks 1989).

Nowadays, the quantity, quality and availability of household datasets have rendered this earlier literature largely redundant. Related issues—such as the variability inevitably associated with finite samples—are more likely to be addressed using semi-parametric or non-parametric techniques like kernel density estimation (see DiNardo *et al.* 1996; Deaton 1997; D'Ambrosio 1999). However, the appetite for extracting distributional information from grouped data has not completely vanished. Independent researchers seldom have the capacity to work simultaneously with many micro datasets, and may be obliged to make use of summary information. Micro information from surveys in the distant past may not have survived, forcing those interested in long-term distributional trends to grapple with published grouped data. In other cases, access to microdata is restricted by concerns about confidentiality or political sensitivity, or because users are charged a high fee.

The continuing need for methods of extracting information from grouped data is well illustrated by the recent flurry of interest in the world distribution of income and its trend over time, as studied by Schultz (1998), Milanovic (2002, 2005), Bourguignon and Morrisson (2002), Sala-i-Martin (2002), Capéau and Decoster (2004), and Dowrick and Akmal (2005), amongst others. While a number of factors contribute to the controversies in this literature—including the coverage of countries, the concept used for average income (or expenditure), and the adjustment made to official exchange rates to compensate for purchasing power variations—limitation on access to microdata is perhaps the greatest single source of the conflicting results. Milanovic (2002, 2005)

[1] See Bandourian *et al.* (2002) and the references cited therein for more recent contributions to this topic.

utilizes a large number of household surveys, but others have to resort to grouped income distribution data, and to adopt simplifying assumptions, conjecturing, for example, that individual countries can be adequately represented by 5-person or 10-person distributions whose incomes correspond to quintile or decile shares. Similar concerns apply with even greater force to studies of long-term poverty trends, since the approximation to a 5- or 10-person distribution clearly limits the nuance of the results.

Our own interest in the question of extracting information from grouped data has been prompted by a desire to analyse alternative poverty trend scenarios for Russia, where income distribution series are only available in grouped form (Shorrocks and Kolenikov 2001).[2] It has been reinforced by exposure to the problems posed by the World Income Inequality Database (WIID) which contains summary observations on 156 countries, most relating to the period 1960-2005.[3] Of the 4,981 observations, 2,945 include information on quintile or decile shares and Gini coefficients. Around 35 per cent of the observations have more details. Comparisons by researchers would be facilitated if all observations had figures for decile shares, and perhaps also for top percentile shares, alternative inequality measures, and poverty rates corresponding to a variety of poverty lines.[4]

Estimates of poverty measures and the Gini inequality index can be obtained from grouped data using the POVCAL software on the World Bank website. In fact, POVCAL was used to generate Gini values for many of the observations inherited by the WIID database from the World Bank. However, POVCAL operates by fitting the general Quadratic and Beta Lorenz functions to grouped data and then applying the formulae reported in Datt (1998). Unfortunately, as shown later, the General Quadratic and Beta forms often generate Lorenz curves that dip below the horizontal axis—in other words, the software can generate negative values even when the data refer to consumption rather than income. Furthermore, the quantile shares associated with the fitted functions can differ significantly from the reported values with which the procedure begins.[5]

This paper describes an improved method for calculating distributional indicators such as inequality values and poverty rates from grouped distribution data. An algorithm allows a sample of 'income' observations to be reconstructed from any valid set of Lorenz co-ordinates.[6] This sample may then be used to compute inequality and poverty statistics, by treating the sample observations as if they had been drawn from a household survey with a homogeneous population of equally weighted households.

---

2 Similar hurdles are faced by those interested in inequality trends in China; see, for example, Chotikapanich, Rao and Tang (2007) and Chotikapanich, Griffiths and Rao (2007).

3 The WIID database may be accessed at www.wider.unu.edu/research.

4 The algorithm described in this paper has been used recently to generate country wealth samples that allow the global distribution of personal wealth to be estimated: see Davies *et al.* (2008, 2007).

5 The POVCAL software can be accessed from www.worldbank.org/LSMS/tools/povcal/. See Minoiu and Reddy (2006) for a detailed critique.

6 Here and elsewhere in the paper, the term 'income' is used generically. Lorenz curves plot the cumulative income shares against the cumulative population shares when observations are ordered in terms of increasing incomes. Only relative incomes matter for Lorenz curves, so the synthetic sample values can be arbitrarily normalized, for example to ensure that the mean is unity.

Two initial restrictions are placed on the synthetic sample, although neither is strictly necessary. First, the observations are constrained to take positive values. This was done to avoid instances in which, say, negative observations are produced for consumption, and also to ensure that values can be computed for all inequality indices in common use. Second, a sample size of 1,000 was chosen for the synthetic distribution. This number was selected primarily in order to produce poverty rates accurate to one decimal point for any given poverty line. A smaller sample size, though feasible, is probably unnecessary given the computing power available nowadays. A larger sample would reduce the downward bias in the inequality value due to averaging incomes within each tenth of a percentile; but the scope for improvement in accuracy in this and other respects is likely to be modest, as confirmed later in this paper by experiments with samples of 2,000 observations.

In principle, many different methods can be used to construct the samples, including parametric and non-parametric techniques employed in the past to estimate distributions from grouped data. Three main criteria were used to discriminate between the alternative procedures: the algorithm should be universally applicable, in the sense that it can accommodate any feasible pattern of grouped data; the characteristics of the generated sample should *exactly* match the reported grouped values; and the procedure should perform well in tests that start with an income sample, compute grouped values, and then use the algorithm to reconstruct the ungrouped data. It is also an advantage to have an algorithm that is both speedy and easy to understand.[7]

The criterion of universal applicability appears anodyne, but turns out in practice to be quite stringent when combined with the requirement that the sample values exactly match the reported data. In particular, it is possible to encounter grouped data for which the mean incomes of adjacent groups are identical. Usually this happens because the published data on percentage shares have been rounded to three, or even two, significant figures. If mean incomes are similar for adjacent groups, then the income values in the relevant ranges must be very compressed, perhaps even identical. While this is unlikely to be true in practice, we took the view that a feasible pattern of grouped data, however implausible, should be respected, and that the chosen algorithm should be capable of handling problematic situations as well as more common arrangements.

The procedure proposed in this paper involves two main stages. Stage I fits a parametric distribution to the grouped observations and then generates a sample from the fitted distribution as an initial approximation to the synthetic observations. Stage II of the algorithm takes the raw sample and adjusts the values until the sample statistics exactly match the 'true' figures. We experimented with several alternative procedures for Stage II, eventually settling on the 'stretching' routine described in Section 2 below. For Stage I, any of the standard distributional or functional forms is a potential candidate: the lognormal (LN), General Quadratic (GQ), Beta, Generalized Beta (GB) and Singh-Maddala (SM) forms were the candidates chosen for this paper.

---

[7] Other conditions might be added to this list. For example, it is probably a good idea to have relatively 'smooth' data that avoids bunching of values or gaps in the income space.

2 The algorithm

Consider a real interval partitioned into m disjoint income classes which are labelled in increasing order. The grouped distribution data for a population of individuals is captured by m+1 Lorenz co-ordinates (p_k^*, L_k^*), where p_k^* (k = 1, … , m) denotes the aggregate proportion of the population in income classes 1 to k; L_k^* (k = 1, …, m) is the corresponding (cumulative) income share; and $(p_0^*, L_0^*) = (0, 0)$.[8] In practice, these Lorenz co-ordinates will typically derive from data reported in the form of the quantile shares, for example decile or quintile shares; but they can also originate from frequency distributions which may record additional details such as the bounds of the income classes. Details of the absolute levels of income are lost in the construction of Lorenz curves, so the overall mean value may be taken to be unity, in which case the mean income of class k is given by

$$(1) \qquad \mu_k^* = \frac{L_k^* - L_{k-1}^*}{p_k^* - p_{k-1}^*}, \quad k = 1, \dots, m.$$

Our aim is to construct a synthetic (and ordered) sample of n equally weighted observations which has a mean value of unity and properties that conform to those of the grouped data. To achieve the required match with the grouped data, the n observations are partitioned into m non-overlapping (and ordered) groups, with group k containing $m_k = n(p_k^* - p_{k-1}^*)$ observations. The value of the ith observation in class k is denoted by $x_{k\,i}$ (k = 1, … , m; i = 1, …, m_k), and the sample mean of class k is signified by μ_k.

The proposed 'ungrouping' algorithm involves two stages. Stage I constructs a rough initial sample with unit mean by generating a set of synthetic values from a parametric form fitted to the grouped data. Suppose, for example, that the underlying distribution is taken to be lognormal and that the sample size is chosen to be 1,000. A value for the standard deviation of log incomes, σ, is obtained by averaging the m-1 estimates:

$$(2) \qquad \sigma_k = \Phi^{-1}(p_k^*) - \Phi^{-1}(L_k^*), \; k = 1, \dots, m\text{-}1,$$

where Φ is the standard normal distribution function (Aitchison and Brown 1957; Kolenikov and Shorrocks 2005: Appendix). The raw sample may then be generated by the percentile points 0.05, 0.15, …, 99.85, 99.95 corresponding to the fitted lognormal. In addition to the lognormal, a number of other parametric forms were considered as candidates for the initial sample. Results obtained using these alternative specifications are discussed in Section 3 below.

Stage II of the algorithm begins with the initial sample and then adjusts the observations until the sample statistics match the 'true' values.[9] Several alternative procedures were

---

[8] Asterisks are used to distinguish the target (true) values from the (non-asterisked) synthetic sample values, which may not match the target figures.

[9] Our procedure makes no use of information on the maximum and minimum values within groups, although frequency tables for income distributions often report the interval endpoints. It might be possible

considered for Stage II, but many of them failed to converge in a reasonable time period, especially when confronted with unusual data properties, such as adjacent income ranges with similar means. The two-step process eventually chosen is both universally applicable and speedy.

The first step adjusts the sample observations in such a way that each of the class k mean incomes, μ_k, is transformed into the corresponding 'true' values, μ_k^*, and appropriate changes made to the intermediate values. To be precise, consider any interval $[\mu_k, \mu_{k+1})$, ($k = 1, \ldots, m\text{-}1$), and convert the initial sample value $x_j \in [\mu_k, \mu_{k+1})$ into the intermediate value $\hat{x}_j$ according to the rule

$$\text{(3a)} \qquad \frac{\hat{x}_j - \mu_k^*}{\mu_{k+1}^* - \mu_k^*} = \frac{x_j - \mu_k}{\mu_{k+1} - \mu_k}, \quad \text{for } k = 1, \ldots, m\text{-}1 \text{ and } x_j \in [\mu_k, \mu_{k+1}),$$

or equivalently

$$\text{(3b)} \qquad \hat{x}_j = \mu_k^* + \frac{\mu_{k+1}^* - \mu_k^*}{\mu_{k+1} - \mu_k}(x_j - \mu_k), \quad \text{for } k = 1, \ldots, m\text{-}1 \text{ and } x_j \in [\mu_k, \mu_{k+1}).$$

Similar adjustments are made at the bottom and top of the distribution using the rule:

$$\text{(4)} \qquad \hat{x}_j = \frac{\mu_1^*}{\mu_1} x_j \quad \text{for } x_j < \mu_1; \quad \hat{x}_j = \frac{\mu_m^*}{\mu_m} x_j, \quad \text{for } x_j \geq \mu_k$$

Note that the transformation given by (3) is well defined because the raw sample from Stage I is both distinct and ordered, hence $\mu_{k+1} > \mu_k$. Note also that the transformation defined by (3) and (4) is (weakly) monotonic, so the sample retains its non-decreasing order.

The above construction ensures that, within each income group, the true mean lies within the range of sample values; in other words:

$$\text{(5)} \qquad \min_i \hat{x}_{ki} \leq \mu_k^* \leq \max_i \hat{x}_{ki}, \quad \text{for } k = 1, \ldots, m.$$

The second step keeps the group bounds fixed and compresses the gaps between the sample values and the upper (resp. lower) bound of the group if the sample mean is below (resp. above) the true value. Specifically, define the lower bound of each group by

$$\text{(6)} \qquad c_1 = 0; \quad c_k = \tfrac{1}{2}\left(\max_i \hat{x}_{k-1,\,i} + \min_i \hat{x}_{ki}\right), \quad k > 1,$$

and convert the intermediate value $\hat{x}_{ki}$ into the final value $\hat{x}_{ki}^*$ according to the rule

---

to refine our algorithm to exploit this additional information, for example by adjusting the data at the start of Stage II to match the true group endpoints.

(7a) $$x^*_{ki} = c_{k+1} - \frac{c_{k+1} - \mu^*_k}{c_{k+1} - \hat{\mu}_k}(c_{k+1} - \hat{x}_{ki}), \quad \text{if } \mu^*_k > \hat{\mu}_k \text{ and } k < m;$$

(7b) $$x^*_{ki} = c_k + \frac{\mu^*_k - c_k}{\hat{\mu}_k - c_k}(\hat{x}_{ki} - c_k), \quad \text{if } \mu^*_k < \hat{\mu}_k \text{ or } k = m.$$

It may be confirmed that this transformation retains the sample ordering both within and across groups, and that the group means compiled for the final sample values, x^*_{ki}, match the true values, μ^*. Within two rounds, therefore, the algorithm produces an ordered sample that exactly replicates the properties of the reported grouped data.

3 Evaluation

The performance of the 'ungrouping' algorithm may be assessed in a variety of ways. One method exploits the additional statistics often attached to grouped data. For example, the values of Gini coefficients (presumably calculated from the original microdata) are sometimes reported alongside published frequency tables. Generating a synthetic income sample from the grouped data enables the Gini index to be estimated and compared to the reported Gini value.

This option was explored in the context of the WIID database using the lognormal as a first approximation. On the whole the results are encouraging, especially when applied to the WIID observations known to be more reliable. In the vast majority of cases, the difference between the 'true' Gini value and the 'synthetic' estimate was less than 0.003 (approximately 1 per cent). As expected, this exercise also suggests that the errors associated with our algorithm shrink as the grouping becomes less coarse (and the number of Lorenz co-ordinates increases).

While this method of assessment has its attractions, a number of problems arise, particularly with regard to reconciling the occasional large discrepancies between the reported Gini figure and the synthetic estimate. It is possible that the published frequency table and Gini value refer to different sets of data for the same country and point of time, or that some of the numbers have been reported incorrectly. Other errors could have been introduced by estimating the Gini values from the grouped data, rather than the original micro-sample. It therefore becomes difficult to evaluate performance without relying heavily on personal judgements concerning the reliability of individual observations.

Another, more stringent, test starts with a suitable micro sample, constructs a set of grouped data, and then examines the degree to which the 'ungrouping' algorithm successfully reconstructs the original data. Information contained in the US Current Population Survey (CPS) for 2000 was used for this purpose. A random sample of 1,000 (positive) income observations was drawn from the CPS microdata and various quantile shares computed from the sample. The ungrouping utility was then applied to the grouped data to generate a synthetic sample of 1,000 which could be compared with the original CPS sample. Three patterns of grouped data were considered, representing the

most common arrangements found in practice: quintile shares; decile shares; and the intermediate case of quintile shares plus the top and bottom decile shares (indicated by the label 'quintile-TB'). To allow for sampling variations, the exercise was repeated first 100 times, and then 200 times. To study the influence of sample size, the experiment was later repeated with a sample consisting of 2,000 observations.

Two methods were used to assess the reliability of the synthetic sample. First, the value of each observation in the (ordered) synthetic sample was compared to its counterpart in the true distribution. Second, inequality values calculated from the reconstructed data were compared with their 'true' values. On the whole, the first exercise is more comprehensive and insightful, because the synthetic data may contain systematic biases. Our limited experience suggests that the generated sample may underestimate incomes on some segments of the Lorenz curve and exaggerate incomes on other segments. However, estimates of the Gini value and other inequality indices may nevertheless closely approximate their true values, giving a spurious impression of accuracy and reliability.

Five alternative specifications were considered as candidates for the distributional forms used in Stage I: the lognormal (LN), General Quadratic (GQ), Beta, Generalized Beta (GB), and Singh-Maddala (SM) functions.[10] The Beta distribution and the General Quadratic Lorenz function both proved to have a major flaw which ruled them out of further consideration: most of the synthetic samples generated during Stage I contained one or more negative observations, despite the fact that all income values are positive in the CPS data. With quintile information, both functional forms fail to ensure non-negative values over 60 per cent of the time. The failure rate rises above 90 per cent with the quintile-TB data, and approaches 100 per cent with a sample size of 2000 (see Table 1). This deficiency eliminated the two functions from further consideration.

For each of the three remaining functional forms the synthetic sets of sample observations were compared to their true counterparts. The results recorded in Table 2 are obtained by expressing both sets of observations in terms of percentage income shares and then computing the absolute deviations. Thus, for example, if the income share of the poorest (richest) person is 0.01 (4.8) per cent and the corresponding synthetic value is 0.015 (3.4) per cent, then the absolute deviation is 0.005 (1.4) per cent. In order to identify any distributional pattern of errors, the absolute deviations are summed within each decile.[11]

---

[10] Details of the lognormal form are given by Aitchison and Brown (1957); the general quadratic Lorenz curve by Villasenor and Arnold (1989); the Beta Lorenz curve by Kakwani (1980); the Generalized Beta by McDonald (1984); and the Singh-Maddala distribution by Singh and Maddala (1976).

[11] Because the observations are expressed as income shares, the sum of absolute deviations is preferred to the mean absolute deviation. The value of the latter is reciprocally related to the sample size; in other words, for a fixed gap between the true and synthetic Lorenz curves, the mean absolute deviation will halve when the number of sampling points doubles.

Table 1: Percentage of times that negative incomes are generated

Grouping pattern	1000 observations		2000 observations	
	Beta	GQ	Beta	GQ
		100 replications		
Quintile	66	65	79	76
Quintile-TB	92	90	97	96
Decile	88	87	94	93
		200 replications		
Quintile	61	65	78	76
Quintile-TB	93	92	98	97
Decile	90	89	96	94

Note: Beta = Beta functional form for Lorenz curve. GQ = general quadratic form for Lorenz curve.
Source: Authors' calculations.

Table 2: Sum of absolute deviations of individual income shares, by decile intervals

	decile	1000 observations						2000 observations					
		LN	SM	GB	ALN	ASM	AGB	LN	SM	GB	ALN	ASM	AGB
	1	0.63	0.17	0.16	0.28	0.15	0.15	0.64	0.17	0.16	0.28	0.14	0.15
	2	0.21	0.09	0.11	0.24	0.13	0.13	0.22	0.08	0.11	0.25	0.13	0.13
	3	0.12	0.11	0.13	0.07	0.05	0.05	0.11	0.10	0.12	0.06	0.04	0.04
	4	0.41	0.17	0.17	0.08	0.06	0.06	0.40	0.16	0.16	0.07	0.05	0.05
Quintile pattern	5	0.74	0.22	0.19	0.14	0.09	0.09	0.73	0.20	0.17	0.12	0.07	0.07
100 replications	6	1.07	0.29	0.21	0.17	0.11	0.11	1.03	0.24	0.17	0.15	0.09	0.09
	7	1.17	0.23	0.17	0.15	0.12	0.13	1.13	0.17	0.13	0.13	0.11	0.12
	8	0.95	0.27	0.40	0.20	0.15	0.17	0.92	0.23	0.41	0.18	0.13	0.15
	9	0.49	0.70	0.98	0.49	0.55	0.84	0.43	0.73	1.02	0.39	0.58	0.89
	10	4.14	2.61	3.03	2.18	2.38	2.96	3.88	2.50	3.05	1.81	2.31	3.02
Total		**9.92**	**4.87**	**5.56**	**3.99**	**3.78**	**4.68**	**9.47**	**4.59**	**5.50**	**3.44**	**3.66**	**4.70**
	1	0.54	0.12	0.12	0.16	0.11	0.11	0.56	0.12	0.11	0.15	0.10	0.10
	2	0.12	0.27	0.30	0.05	0.06	0.06	0.12	0.27	0.30	0.05	0.06	0.06
	3	0.22	0.30	0.32	0.05	0.06	0.06	0.21	0.30	0.32	0.04	0.05	0.05
Quintile-TB	4	0.57	0.36	0.34	0.07	0.06	0.06	0.56	0.35	0.33	0.06	0.05	0.05
pattern	5	0.90	0.37	0.31	0.09	0.07	0.07	0.90	0.36	0.30	0.08	0.06	0.06
100 replications	6	1.23	0.39	0.27	0.11	0.09	0.09	1.20	0.34	0.22	0.10	0.08	0.07
	7	1.32	0.24	0.19	0.14	0.10	0.10	1.28	0.18	0.16	0.13	0.08	0.08
	8	1.05	0.34	0.54	0.16	0.17	0.18	1.01	0.33	0.57	0.13	0.15	0.16
	9	0.52	1.02	1.31	0.25	0.26	0.29	0.47	1.08	1.35	0.18	0.22	0.26
	10	5.00	3.09	3.43	1.95	2.66	3.32	4.74	3.04	3.47	1.54	2.71	3.44
Total		**11.48**	**6.50**	**7.10**	**3.04**	**3.64**	**4.33**	**11.04**	**6.38**	**7.12**	**2.47**	**3.56**	**4.33**

table continues…

	decile	1000 observations						2000 observations					
		LN	SM	GB	ALN	ASM	AGB	LN	SM	GB	ALN	ASM	AGB
	1	0.58	0.12	0.12	0.16	0.10	0.10	0.59	0.12	0.11	0.15	0.10	0.09
	2	0.15	0.22	0.24	0.04	0.04	0.04	0.16	0.22	0.24	0.03	0.03	0.03
	3	0.17	0.25	0.24	0.04	0.04	0.04	0.16	0.25	0.24	0.03	0.03	0.03
	4	0.50	0.30	0.25	0.05	0.05	0.05	0.49	0.29	0.24	0.05	0.04	0.04
Decile pattern	5	0.83	0.32	0.22	0.06	0.06	0.06	0.82	0.31	0.21	0.06	0.05	0.05
100 replications	6	1.16	0.35	0.21	0.09	0.07	0.07	1.12	0.31	0.15	0.09	0.07	0.06
	7	1.25	0.23	0.25	0.10	0.08	0.08	1.21	0.18	0.21	0.08	0.07	0.07
	8	1.01	0.33	0.64	0.12	0.11	0.11	0.97	0.32	0.65	0.11	0.08	0.08
	9	0.51	0.95	1.31	0.21	0.20	0.22	0.45	0.99	1.34	0.16	0.16	0.18
	10	4.61	2.97	3.56	1.93	2.57	3.52	4.35	2.90	3.55	1.53	2.59	3.59
Total		**10.75**	**6.04**	**7.03**	**2.81**	**3.31**	**4.27**	**10.31**	**5.88**	**6.95**	**2.28**	**3.21**	**4.23**
	1	0.63	0.17	0.16	0.28	0.14	0.15	0.64	0.17	0.16	0.28	0.14	0.14
	2	0.22	0.09	0.11	0.24	0.13	0.13	0.23	0.08	0.10	0.24	0.12	0.13
	3	0.12	0.11	0.13	0.07	0.06	0.06	0.11	0.10	0.12	0.06	0.04	0.04
	4	0.40	0.16	0.16	0.07	0.06	0.06	0.40	0.15	0.16	0.07	0.05	0.05
Quintile pattern	5	0.74	0.23	0.20	0.13	0.09	0.09	0.73	0.20	0.18	0.12	0.07	0.07
200 replications	6	1.07	0.29	0.21	0.16	0.11	0.10	1.03	0.25	0.17	0.15	0.09	0.09
	7	1.19	0.23	0.17	0.17	0.12	0.13	1.13	0.17	0.13	0.14	0.11	0.11
	8	1.00	0.26	0.37	0.22	0.15	0.17	0.93	0.22	0.40	0.18	0.13	0.15
	9	0.52	0.69	0.97	0.52	0.54	0.83	0.43	0.73	1.01	0.40	0.58	0.89
	10	4.23	2.58	2.99	2.19	2.33	2.90	3.89	2.52	3.06	1.81	2.33	3.02
Total		**10.10**	**4.81**	**5.46**	**4.03**	**3.72**	**4.60**	**9.51**	**4.59**	**5.48**	**3.45**	**3.67**	**4.69**
	1	0.55	0.11	0.11	0.15	0.10	0.10	0.56	0.11	0.11	0.15	0.10	0.10
	2	0.12	0.27	0.29	0.06	0.06	0.06	0.12	0.26	0.29	0.05	0.06	0.06
	3	0.22	0.30	0.31	0.05	0.06	0.06	0.21	0.29	0.31	0.04	0.05	0.05
	4	0.56	0.35	0.33	0.08	0.07	0.07	0.56	0.34	0.32	0.06	0.05	0.05
Quintile-TB	5	0.91	0.38	0.32	0.09	0.08	0.08	0.89	0.36	0.30	0.08	0.06	0.07
pattern	6	1.23	0.39	0.27	0.12	0.09	0.09	1.20	0.34	0.23	0.10	0.08	0.08
200 replications	7	1.33	0.26	0.19	0.15	0.11	0.11	1.28	0.19	0.16	0.13	0.09	0.09
	8	1.09	0.32	0.50	0.17	0.18	0.19	1.02	0.32	0.55	0.13	0.16	0.17
	9	0.55	1.00	1.28	0.25	0.27	0.30	0.48	1.07	1.34	0.19	0.23	0.27
	10	5.09	3.07	3.38	1.94	2.60	3.24	4.75	3.05	3.47	1.54	2.71	3.43
Total		**11.65**	**6.43**	**6.97**	**3.06**	**3.61**	**4.29**	**11.05**	**6.34**	**7.07**	**2.49**	**3.58**	**4.34**
	1	0.58	0.12	0.11	0.15	0.10	0.09	0.59	0.12	0.11	0.15	0.09	0.09
	2	0.15	0.22	0.24	0.04	0.04	0.04	0.16	0.21	0.24	0.03	0.03	0.03
	3	0.17	0.25	0.24	0.04	0.04	0.04	0.15	0.24	0.23	0.04	0.03	0.03
	4	0.49	0.29	0.24	0.05	0.05	0.05	0.49	0.29	0.23	0.05	0.04	0.04
Decile pattern	5	0.84	0.33	0.23	0.07	0.06	0.06	0.82	0.31	0.21	0.06	0.05	0.05
200 replications	6	1.16	0.35	0.21	0.09	0.07	0.07	1.12	0.31	0.16	0.09	0.06	0.06
	7	1.27	0.25	0.24	0.10	0.08	0.08	1.21	0.18	0.21	0.08	0.07	0.07
	8	1.05	0.31	0.60	0.13	0.11	0.11	0.98	0.30	0.63	0.11	0.09	0.09
	9	0.54	0.93	1.28	0.21	0.20	0.21	0.46	0.98	1.32	0.16	0.16	0.18
	10	4.70	2.95	3.51	1.91	2.51	3.43	4.36	2.91	3.54	1.54	2.59	3.57
Total		**10.93**	**5.98**	**6.90**	**2.79**	**3.24**	**4.19**	**10.34**	**5.84**	**6.88**	**2.29**	**3.22**	**4.22**

Note: LN = lognormal, SM = Singh-Maddala, GB = Generalized Beta; Prefix 'A' = adjusted data.

Source: Authors' calculations.

The six column headings in Table 2 refer to the results for the crude (Stage I) samples obtained from the lognormal (LN), Singh-Madalla (SM) and Generalized Beta (GB) distributions, plus the results (ALN, ASM and AGB, respectively) obtained after the samples are adjusted to match the group details in Stage II of the ungrouping algorithm. The first point to note is that the Stage II adjustment procedure proposed in this paper usually leads to a significant reduction in the errors. This is seen in Table 2 by comparing the deviations for the raw and adjusted samples, holding constant the sample size and grouping assumption. With very few exceptions the adjustment leads to a better match with the true income values, often reducing the average deviation by a factor of two or more. Table 2 also hints at an improvement in the match as the sample size increases from 1,000 to 2,000, although the improvement is not uniform, nor much in evidence when the Generalized Beta distribution is used. Raising the number of sample replications from 100 to 200 appears to have little effect, suggesting that the reported figures are close to their asymptotic values.

Turning to the pattern across deciles, the results—particularly those for the adjusted sample values—show that the errors are heavily concentrated in the top decile and (to a lesser degree) in deciles 1-2 and deciles 8-9.[12] In deciles 3-7, the synthetic income values closely match the true values. The impact of the grouping criterion seems less clear at first. Focusing on the columns corresponding to the unadjusted data, there is little evidence that errors diminish as the grouping pattern becomes less coarse. However, after the sample is adjusted, the absolute deviations decline most of the time as the grouping arrangement changes from quintiles to quintile-TB to deciles.

Table 3: Total absolute deviations of income shares

	1000 observations						2000 observations					
Number of samples	LN	SM	GB	ALN	ASM	AGB	LN	SM	GB	ALN	ASM	AGB
						Quintile						
100	9.92	4.87	5.56	3.98	3.78	4.68	9.47	4.59	5.50	3.44	3.66	4.70
200	10.10	4.81	5.46	4.03	3.72	4.60	9.51	4.59	5.48	3.45	3.67	4.69
						Quintile-TB						
100	11.47	6.50	7.10	3.04	3.64	4.33	11.04	6.37	7.12	2.46	3.56	4.33
200	11.65	6.43	6.97	3.06	3.61	4.29	11.05	6.34	7.07	2.48	3.58	4.34
						Decile						
100	10.75	6.04	7.03	2.81	3.31	4.27	10.31	5.88	6.95	2.28	3.21	4.23
200	10.93	5.98	6.90	2.79	3.24	4.19	10.34	5.84	6.88	2.29	3.22	4.22

Note: LN = lognormal, SM = Singh-Maddala, GB = Generalized Beta; Prefix 'A' = adjusted data.
Source: Authors' calculations.

[12] The *proportional* deviations may not be greatest in the top decile because the base values are larger.

The final issue concerns the choice of distribution function for the raw data sample. Here the result is a little surprising. The summary of Table 2 results reproduced in Table 3 demonstrates that before the Stage II adjustment, the lognormal is unambiguously the worst performer and Singh-Maddala performs the best. After the sample is adjusted, the Singh-Maddala form continues to dominate the Generalized Beta distribution. But the lognormal-based estimates improve so much that they overtake the Generalized Beta results in each of the twelve scenarios identified in Table 3. They also dominate Singh-Maddala in all situations except those corresponding to the coarsest grouping criterion (quintiles alone) and smallest sample size (1,000 observations). More surprisingly, perhaps, the disaggregated results in Table 2 show that the adjusted lognormal values provide particularly accurate values in the top decile. This is precisely the region where the lognormal is not expected to perform well; yet the lognormal dominates both of the other candidate distributions in every instance.

It is not clear why the Stage II adjustment leads to such an exceptional improvement in the lognormal originated estimates. However our findings support the view that the ungrouping algorithm described above, coupled with an initial lognormal fit, is capable of reproducing sample data from grouped statistics with a high degree of accuracy. Our results also lead to the recommendation that the size of the synthetic sample should be chosen as large as possible, since the lognormal is unambiguously best in the runs with 2,000 observations, and since increasing the size of the sample improves the data match in all circumstances.

The second method of assessing the performance of the ungrouping algorithm using the CPS involves a comparison between the true inequality values and the estimates generated via the synthetic sample. Four inequality measures were used for this purpose, the Gini coefficient and three members of the entropy family: the mean logarithmic deviation (*MLD*) the Theil coefficient (*T*), and the squared coefficient of variation (CV^2). For a sample of n observations x_i ($i = 1, \ldots, n$) with mean μ, these indices may be written, respectively, as

$$MLD = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{\mu}{x_i}$$

$$T = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{\mu} \ln \frac{x_i}{\mu} \qquad (8)$$

$$CV^2 = \frac{1}{n}\sum_{i=1}^{n} \left\{ \left(\frac{x_i}{\mu} \right)^2 - 1 \right\}.$$

Tables 4-7 report the mean absolute percentage error for each of these indices, using both the raw synthetic sample and the adjusted sample. The results for the Gini coefficient are very encouraging. Table 4 shows that the errors for the raw synthetic sample are usually 2-3 per cent (and higher still for the lognormal fit). However, the expected error for the adjusted samples never exceeds one per cent and falls below 0.1 per cent for lognormal generated observations constructed from decile share

information. This translates into a confidence interval of around ±0.001 for the ungrouping estimates of a typical income Gini value (say, 0.4).

Results for the three entropy indices are less satisfactory. The best that can be achieved with the Theil coefficient is about a 1 per cent error, which is reasonably acceptable (see Table 6). But the minimum expected error is 2.5 per cent for the squared CV (Table 7), and around 4 per cent for the MLD (Table 5).

Table 4: Mean absolute percentage error: Gini coefficient

Grouping pattern	Unadjusted data			Adjusted data		
	LN	SM	GB	ALN	ASM	AGB
			1000 sample observations, 100 replications			
Quintile	4.01	0.91	0.46	0.18	0.48	0.73
Quintile-TB	6.53	3.24	2.31	0.10	0.19	0.28
Decile	5.41	2.48	1.44	0.08	0.22	0.35
			1000 sample observations, 200 replications			
Quintile	4.07	0.97	0.52	0.20	0.47	0.71
Quintile-TB	6.57	3.29	2.37	0.10	0.19	0.27
Decile	5.47	2.56	1.54	0.08	0.22	0.35
			2000 sample observations, 100 replications			
Quintile	3.65	0.61	0.33	0.15	0.54	0.78
Quintile-TB	6.17	2.95	2.07	0.11	0.22	0.30
Decile	5.05	2.22	1.01	0.07	0.25	0.37
			2000 sample observations, 200 replications			
Quintile	3.65	0.62	0.36	0.15	0.55	0.78
Quintile-TB	6.15	2.91	2.06	0.11	0.22	0.30
Decile	5.04	2.20	1.00	0.08	0.25	0.37

Note: LN = lognormal; SM = Singh-Maddala; GB = Generalized Beta.
Source: Authors' calculations.

In some respects, the pattern of results in Tables 4-7 corroborates the conclusions drawn earlier from Tables 2 and 3. The Singh-Maddala derived data are better on every count than those obtained using the Generalized Beta distribution. The lognormal performs

poorly before the synthetic sample is adjusted, but improves greatly during Stage II of the algorithm, so much so that it leapfrogs above both the Singh-Maddala and Generalized Beta estimates, unless the mean logarithmic deviation is chosen as the inequality index. As regards the grouping arrangement, the estimates again tend to improve (for the adjusted data at least) as one moves from quintiles to quintile-TB to deciles, echoing the slightly ambiguous results obtained earlier.

Table 5: Mean absolute percentage error: mean logarithmic deviation

Grouping Pattern	Unadjusted data			Adjusted data		
	LN	SM	GB	ALN	ASM	AGB
			1000 sample observations, 100 replications			
Quintile	17.66	10.82	11.24	15.32	11.57	12.01
Quintile-TB	12.90	3.94	4.12	10.36	7.98	8.08
Decile	15.03	5.93	6.91	10.30	8.02	8.10
			1000 sample observations, 200 replications			
Quintile	17.47	10.58	10.99	15.12	11.35	11.78
Quintile-TB	12.73	3.73	3.92	10.15	7.76	7.86
Decile	14.85	5.64	6.59	10.09	7.79	7.88
			2000 sample observations, 100 replications			
Quintile	18.23	11.30	11.63	15.35	11.65	12.05
Quintile-TB	13.45	4.20	4.31	10.28	7.96	8.04
Decile	15.59	6.27	7.01	10.22	8.00	8.06
			2000 sample observations, 200 replications			
Quintile	18.07	11.14	11.46	15.20	11.49	11.90
Quintile-TB	13.34	4.18	4.27	10.17	7.86	7.95
Decile	15.46	6.21	6.92	10.13	7.90	7.97

Note: LN = lognormal; SM = Singh-Maddala; GB = Generalized Beta.
Source: Authors' calculations.

Table 6: Mean absolute percentage error: Theil coefficient

Grouping Pattern	Unadjusted data			Adjusted data		
	LN	SM	GB	ALN	ASM	AGB
			1000 sample observations, 100 replications			
Quintile	10.19	1.95	3.85	1.42	3.70	5.32
Quintile-TB	16.52	3.53	2.26	0.95	2.66	3.57
Decile	13.69	2.91	4.05	0.95	2.59	3.86
			1000 sample observations, 200 replications			
Quintile	10.48	2.00	3.71	1.52	3.60	5.22
Quintile-TB	16.78	3.83	2.46	0.96	2.60	3.49
Decile	13.96	3.18	4.05	0.96	2.53	3.78
			2000 sample observations, 100 replications			
Quintile	8.78	2.41	4.65	1.15	4.19	5.67
Quintile-TB	15.05	2.55	1.68	0.84	2.99	3.79
Decile	12.24	1.94	3.98	0.84	2.90	4.01
			2000 sample observations, 200 replications			
Quintile	8.78	2.46	4.62	1.15	4.21	5.68
Quintile-TB	14.97	2.59	1.82	0.88	3.00	3.80
Decile	12.20	2.03	4.03	0.88	2.92	4.02

Note: LN = lognormal; SM = Singh-Maddala; GB = Generalized Beta.
Source: Authors' calculations.

Table 7: Mean absolute percentage error: squared coefficient of variation

Grouping Pattern	Unadjusted data			Adjusted data		
	LN	SM	GB	ALN	ASM	AGB
			1000 sample observations, 100 replications			
Quintile	26.03	6.93	13.14	5.86	9.64	14.53
Quintile-TB	36.89	5.94	11.26	3.92	8.69	11.91
Decile	31.97	6.74	16.17	3.80	8.29	12.72
			1000 sample observations, 200 replications			
Quintile	26.91	6.71	12.78	6.35	9.40	14.30
Quintile-TB	37.75	5.93	10.85	4.04	8.50	11.69
Decile	32.84	6.65	15.65	3.90	8.09	12.49
			2000 sample observations, 100 replications			
Quintile	21.61	8.90	15.07	3.68	11.37	15.76
Quintile-TB	31.90	6.51	12.90	2.48	10.03	12.81
Decile	27.24	7.55	17.15	2.48	9.56	13.36
			2000 sample observations, 200 replications			
Quintile	21.50	9.00	15.06	3.58	11.50	15.85
Quintile-TB	31.67	6.83	12.88	2.53	10.12	12.88
Decile	27.06	7.76	17.05	2.53	9.64	13.42

Note: LN = lognormal; SM = Singh-Maddala; GB = Generalized Beta.

Source: Authors' calculations.

The most surprising feature of Tables 5-7 is the fact that the Stage II adjustment to the synthetic sample does not always improve the accuracy of the estimate of inequality. Indeed, for the Singh-Maddala and Generalized Beta distributions, the adjustment raises the mean absolute percentage error in every case reported for the MLD index in Table 5, and in most cases recorded for the Theil coefficient in Table 6 and for the squared CV in Table 7. In contrast, the Stage II adjustment always improves the accuracy of the lognormal based estimates, quite dramatically in the case of the Theil coefficient values in Table 6.

The post adjustment deterioration in the predictive accuracy of the Singh-Maddala and Generalized Beta synthetic samples is unanticipated and not easy to comprehend, since the general tendency for an improvement during Stage II was documented earlier in Table 2. To explore the possible explanations, some specific sets of synthetic observations were examined, before and after adjustment. Figure 1 illustrates one (inevitably unrepresentative) sample obtained by applying the Singh-Maddala distribution to quintile groups. The graph—which plots the deviation of the synthetic Lorenz curve from the true Lorenz curve—shows the general improvement (and the exact quintile share match) resulting from the Stage II adjustment. But the ranking of the pre- and post-adjustment samples is less clear in the top quintile, precisely the place where the greatest inaccuracies occur. In particular, note that the top quintile Lorenz values for the adjusted sample always exceed the true values, but the deviations for the unadjusted sample can be negative or positive, allowing the possibility that the negative deviations offset the general tendency to underestimate inequality in the top quintile.[13]

Figure 2 examines the implications for estimates of the squared coefficient of variation by plotting the partial sum of the expression given in equation (8), in other words

$$\text{(9)} \qquad \text{partial squared CV} = \frac{1}{n}\sum_{i=1}^{k}\left\{\left(\frac{x_i}{\mu}\right)^2 - 1\right\}, \quad k = 1, \ldots, n.$$

The pattern is broadly similar to that depicted in Figure 1. The Stage II adjustment improves the accuracy of the partial squared CV for most of its range. But it is the errors in the uppermost tail that determine the accuracy of the overall squared CV value, and here the superiority of the post adjustment sample is not established unequivocally. Indeed, the fact that the Stage II adjustment magnifies the underestimate of the very richest incomes is the most likely explanation why the unadjusted Singh-Maddala sample is a better predictor of the squared CV. For lognormal based samples the adjustment always tends to improve the inequality estimates, so this anomaly does not arise, providing further grounds for favouring lognormal based samples, despite the well known deficiencies of the lognormal distribution as a representation of observed income distributions.

[13] The inequality bias occurs because the incomes of the super rich are underestimated, as is evident in the the sharp decline in the deviation from the true Lorenz curve at the very top of the distribution.

Figure 1: Deviation from true Lorenz curve

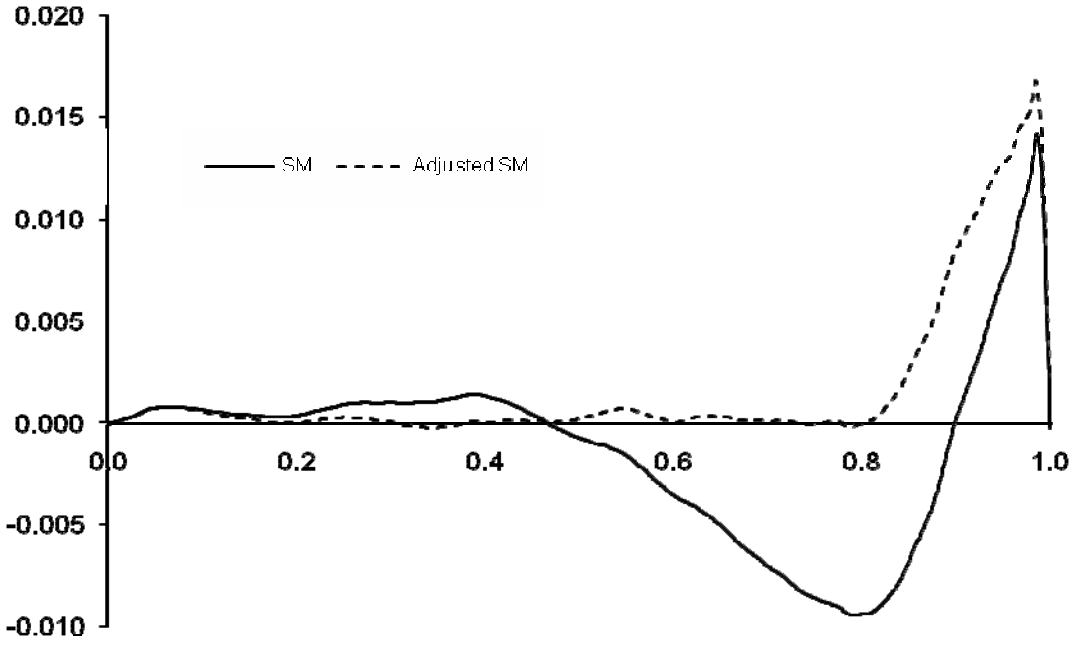
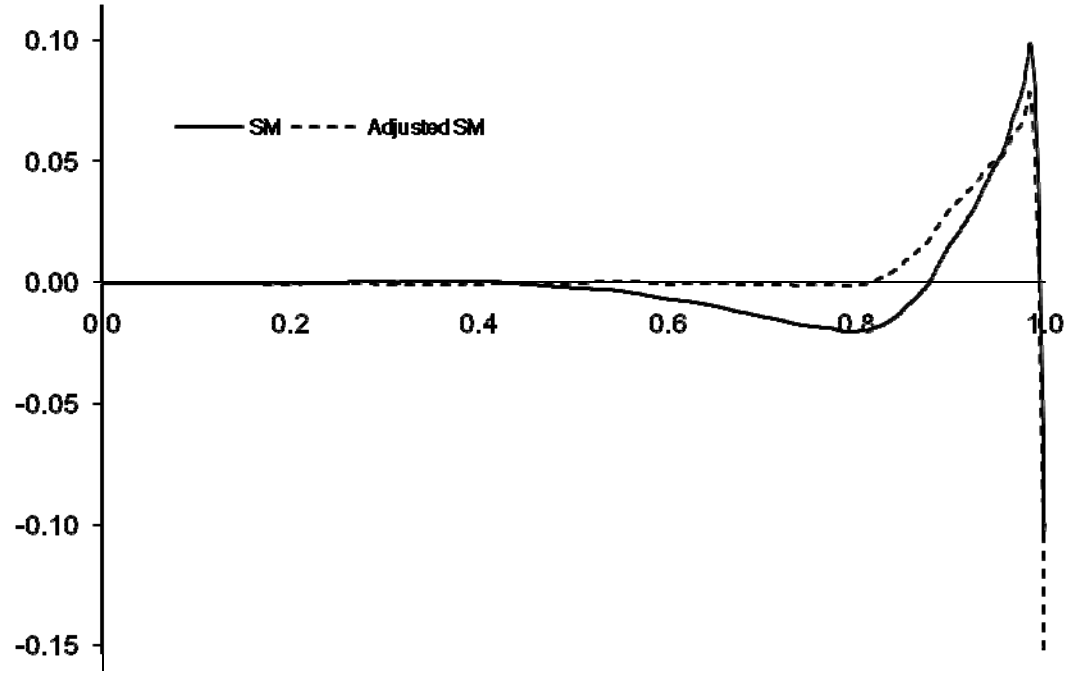


Figure 2: Deviation of partial squared coefficient of variation



4 Summary and conclusion

Despite the increasing availability of survey data, poverty and inequality analysts are often confronted with the need for individual income observations when only grouped data are either accessible or affordable. As a result, there is a continuing demand for algorithms than can generate synthetic samples of observations from grouped information, as demonstrated by the popularity of the POVCAL program offered by the World Bank.

This paper proposes an alternative method of reconstructing individual income observations from grouped distributional data. It involves two stages, first fitting a parametric Lorenz curve or distribution function to the grouped data, then adjusting the raw data generated by the fitted function. The procedure has two major virtues: it ensures that the characteristics of the synthetic sample exactly match the reported group values; and it is universally applicable in the sense of being able to handle any feasible pattern of grouped data. Our method also has the advantages of speed and simplicity.

Using individual income records drawn from the CPS data, our method was tested by comparing the true values of individual observations with their synthetic counterparts. The results clearly demonstrate the superiority of the final adjusted sample, the adjustment leading to a better match with individual incomes in the vast majority of cases. Relative to the raw data, the adjustment often reduces the average deviation by a factor of two or more.

Comparison between the true and generated values of inequality indices provides a second way of assessing our proposed algorithm. In this respect, results for the Gini coefficient are rather encouraging. The expected error never exceeds one per cent and falls below 0.1 per cent for lognormal generated observations constructed from decile share information, which translates into a confidence interval of around ±0.001 for estimates of a typical income Gini value (say, 0.4). Results for a selection of entropy indices are less good, but still satisfactory. Needless to say, the performance of the method improves as the grouping pattern becomes finer, changing from quintiles to quintiles plus the top and bottom deciles, and then to deciles.

As regards the parametric function used to generate the raw sample, the Beta and General Quadratic Lorenz functions employed in POVCAL and by Datt and Ravallion (1992) are deficient in one major respect: most of the synthetic samples contain one or more negative observations despite the fact that CPS income observations are always positive. Among the remaining candidates, the lognormal form is the clear winner in our test results after the Stage II adjustment has been implemented. Compared to samples derived from the Singh-Maddala or Generalized Beta distributions, the lognormal based observations are consistently closer to the true values, and ensure more accurate estimates of most inequality indices.

On the basis of our findings, we conclude that our proposed adjustment procedure, coupled with an initial lognormal fit and a sample size of at least 1000, is capable of reproducing individual data from grouped statistics with a high degree of accuracy. However we encourage others to subject our algorithm to further tests, using alternative sources of micro data (for example, the Luxembourg Income Study) and using alternative functional forms to generate the raw sample observations.

References

Aitchinson, J., and Brown, J.A.C., 1957. *The Lognormal Distribution with Special Reference to its Use in Economics*. Cambridge: Cambridge University Press.

Bandourian, R., McDonald, J.B., and Turvey, R.S., 2002. A comparison of parametric models of income distribution across countries and over time. *Luxembourg Income Study Working Paper* No. 305.

Bourguignon, F., and Morrison, C., 2002. Inequality among world citizens: 1820-1992. *American Economic Review*, 92: 727-44.

Capéau, B., and Decoster, A., 2004. The rise or fall of world inequality: a spurious controversy? *UNU-WIDER Discussion Paper* No. 2004/02.

Chotikapanich, D., Griffiths, W.E., and Rao, D.S.P., 2007. Estimating and combining national income distributions using limited data. *Journal of Business and Economics Statistics*, 25: 97-109.

Chotikapanich, D., Rao, D.S.P., and Tang, K.K., 2007. Estimating income inequality in China using grouped data and the Generalized Beta distribution. *Review of Income and Wealth*, 53: 127-47.

Chotikapanich, D., Valenzuela, R., and Rao, D.S.P., 1997. Global and regional inequality in the distribution of income: Estimation with limited and incomplete data. *Empirical Economics*, 22(4): 533-46.

Cowell, F.A., and Mehta, F., 1982. The estimation and interpolation of inequality measures. *Review of Economic Studies*, 49: 273-90.

D'Ambrosio, C., 1999. The Distribution of wages: A non-parametric decomposition. *Working Paper* No. 284, Universiti Bocconi and New York University.

Datt, G., 1998. Computational tools for poverty measurement and analysis. Available at: www.ifpri.org/divs/fcnd/dp/papers/dp50.pdf

Datt, G., and Ravallion, M., 1992. Growth and redistribution components of changes in poverty measures: A decomposition with application to Brazil and India in the 1908s. *Journal of Development Economics*, 38: 275-95.

Davies, J.B., and Shorrocks, A.F., 1989. Optimal grouping of income and wealth data, *Journal of Econometrics*, 42: 97-108.

Davies, J.B., Sandstrom, S., Shorrocks, A., and Wolff, E., 2007. Estimating the level and distribution of global household wealth. *UNU-WIDER Research Paper* No. 2007/77.

Davies, J.B., Sandstrom, S., Shorrocks, A., and Wolff, E., 2008. The world distribution of household wealth. *UNU-WIDER Discussion Paper* No. 2008/03.

Deaton, A., 1997. *The Analysis of Household Surveys*. Baltimore: Johns Hopkins University Press.

DiNardo, J., Fortin, N.M., and Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973-1993: a semiparametric approach. *Econometrica*, 64(5): 1001-44.

Dowrick, S., and Akmal, M., 2005. Contradictory trends in global income inequality: a tale of two biases. *Review of Income and Wealth*, 51: 201-29.

Gastwirth, J.L., 1972. The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics*, 54: 306-16.

Harrison, A., 1981. Earnings by size: a tale of two distributions. *Review of Economic Studies*, 48: 621-31.

Kakwani, N., 1976. On the estimation of income inequality measures from grouped observations. *Review of Economic Studies*, 43: 483-92.

Kakwani, N., 1980, On a class of poverty measures, *Econometrica*, 48: 437-46.

Kakwani, N.C., and Podder, N., 1973. On the estimation of lorenz curves from grouped observations, *International Economic Review*, 14: 278-92.

Kakwani, N., and Podder, N., 1976. Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. *Econometrica*, 44: 137-48.

Kloek, T., and Van Dijk, H.K., 1978. Efficient estimation of income distribution parameters. *Journal of Econometrics*, 8: 61-74.

Kolenikov, S., and Shorrocks, A.F., 2005. A decomposition analysis of regional poverty in Russia. *Review of Development Economics*, 9: 25-46

McDonald, J.B., 1984. Some generalized functions for the size distribution of income. *Econometrica*, 52: 647-63.

McDonald, J.B., and Ransom, M.J., 1979. Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47:1513-26.

McDonald, J.B., and Ransom, M.J., 1981. An analysis of the bounds for the Gini coefficient. *Journal of Econometrics*, 17: 177-88.

Milanovic, B., 2002. True world income distribution, 1988 and 1993: first calculation based on household surveys alone. *Economic Journal*, 112: 51-92.

Milanovic, B., 2005. *Worlds Apart: Measuring International and Global Inequality*. New Jersey: Princeton University Press.

Minoiu, C., and Reddy, S.G., 2006. The estimation of poverty and inequality from grouped data using parametric curve fitting: an evaluation of POVCAL, mimeo.

Sala-i-Martin, X., 2002. The world distribution of income. *NBER Working Paper* No. 8933.

Salem, A.B.Z., and Mount, T.D., 1974. A convenient descriptive model of income distribution: the gamma density, *Econometrica*, 42: 1115-27.

Schultz, T. P., 1998. Inequality in the distribution of personal income in the world: how it is changing and why. *Economic Growth Center Working Paper* No. 784, Yale University.

Sen, A.K., 1973, *On Economic Inequality*. Oxford: Clarendon.

Sen, A.K., 1976. Poverty: an ordinal approach to measurement. *Econometrica*, 44, 219-31.

Shorrocks, A.F., and Kolenikov, S., 2001. Poverty trends in Russia. Mimeo.

Singh, S.K., and Maddala, G.S., 1976, A function for the size distribution of incomes. *Econometrica*, 44: 963-70.

Villasenor, J., and Arnold, B.C., 1989. Elliptical Lorenz curves. *Journal of Econometrics*, 40: 327-38.