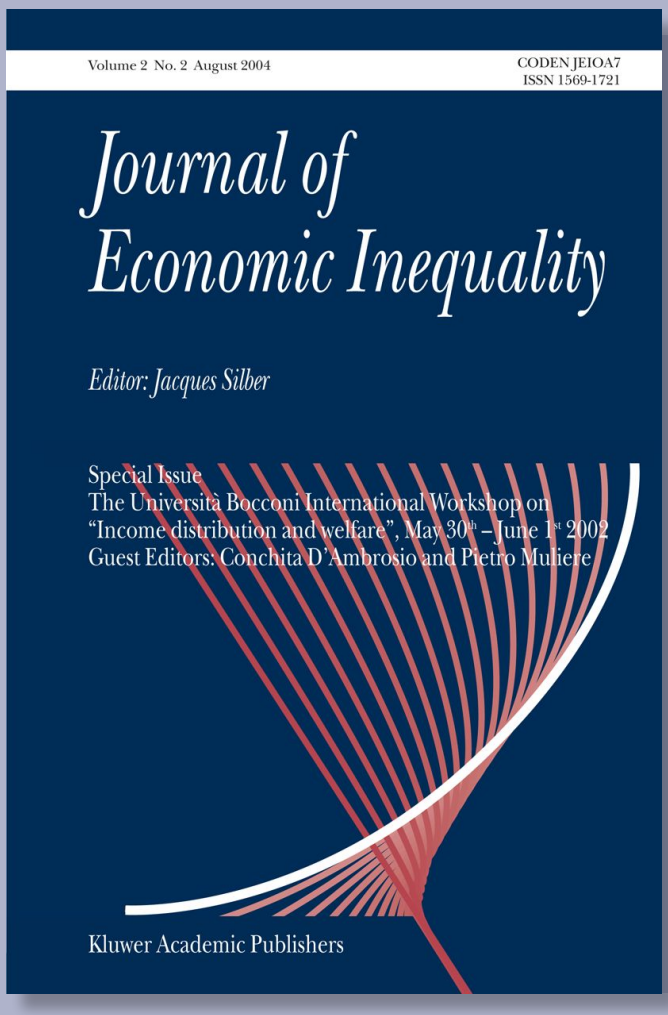
Kernel density estimation on grouped data: the case of poverty assessment

Camelia Minoiu & Sanjay G. Reddy

The Journal of Economic Inequality

ISSN 1569-1721

J Econ Inequal
DOI 10.1007/s10888-012-9220-9



Springer

J Econ Inequal
DOI 10.1007/s10888-012-9220-9

Kernel density estimation on grouped data: the case of poverty assessment

Camelia Minoiu · Sanjay G. Reddy

Received: 7 November 2010 / Accepted: 5 March 2012

Abstract Grouped data have been widely used to analyze the global income distribution because individual records from nationally representative household surveys are often unavailable. In this paper we evaluate the performance of nonparametric density smoothing techniques, in particular kernel density estimation, in estimating poverty from grouped data. Using Monte Carlo simulations, we show that kernel density estimation gives rise to nontrivial biases in estimated poverty levels that depend on the bandwidth, kernel, poverty indicator, size of the dataset, and data generating process. Furthermore, the empirical bias in the poverty headcount ratio critically depends on the poverty line. We also undertake a sensitivity analysis of global poverty estimates to changes in the bandwidth and show that they vary widely with it. A comparison of kernel density estimation with parametric estimation of the Lorenz curve, also applied to grouped data, suggests that the latter fares better and should be the preferred approach.

Keywords Kernel density estimation · Lorenz curve · Grouped data · Income distribution · Global poverty

JEL Classifications I32 · D31 · C14 · C15

Electronic supplementary material The online version of this article (doi:10.1007/s10888-012-9220-9) contains supplementary material, which is available to authorized users.

C. Minoiu (✉)
International Monetary Fund, IMF Institute, 700 19th St. NW, Washington, DC 20431, USA
e-mail: CMinoiu@imf.org

S. G. Reddy
Department of Economics, The New School for Social Research, 6 East 16th Street, New York, NY 10003, USA
e-mail: reddys1@newschool.edu

Published online: 01 April 2012

1 Introduction

Recent studies have used nonparametric smoothing techniques, and in particular kernel density estimation (KDE) on grouped data to obtain poverty estimates and to describe the global income distribution.[1] Grouped data—also referred to as tabulated data—typically take the form of income averages for a small number of population quantiles (quantile means). Grouped data have been popular because individual records from household surveys are often unavailable or are difficult to obtain or analyze, especially for multiple country-years.[2] Furthermore, large cross-country datasets such as the UNU-WIDER World Income Inequality Database and the World Bank's Povcalnet now offer a large amount of distributional information in grouped form. Despite recent efforts to estimate features of the global income distribution from grouped data alone, the relative performance of various methods in this setting remains unstudied.

Kernel density estimation is one of several alternatives for estimating income distributions from grouped data. Also popular are parametric approaches such as the estimation of a functional form for the Lorenz curve or income distribution density function.[3] In a study of these methods, Minoiu and Reddy [30] show that commonly-used parameterizations of the Lorenz curve such as the General Quadratic (GQ) and Beta models, respectively developed by Villasenor and Arnold [39] and Kakwani [18], perform well in estimating poverty and inequality from grouped data. We use these two parameterizations of the Lorenz curve in the analysis to provide a benchmark for our nonparametric results. This allows us to examine the performance of nonparametric kernel density estimation compared to parametric approaches.

We begin by examining the performance of the nonparametric approach in estimating the income distribution and poverty from grouped data. We report biases in poverty estimates for several plausible income distributions and a wide range of poverty indicators, poverty lines, bandwidths, and kernels. Our method is a Monte Carlo simulation study which allows us to compare the poverty estimates obtained from grouped data with their population counterparts. We find that KDE gives rise to nontrivial biases in estimated poverty levels that depend on the bandwidth, kernel, poverty indicator, poverty line, size of the dataset, and data generating process. For all income distributions considered, the average income of the poorest quantiles is generally underestimated, while that of the richest is overestimated. In turn, this leads to a systematic overestimation of the poverty headcount ratio for lower poverty lines, and opposite biases for higher ones. The poverty headcount ratio is statistically close to its theoretical counterpart only when the poverty line is close to

[1] See [2, 14, 35, 44] for analyses of global or national income distributions and poverty.

[2] See, e.g., [5] for an analysis of the long-run global income distribution, and [26–28] for estimates of global inequality based on grouped data.

[3] See, for instance, [6, 33]. Flexible functional forms for the income density from the exponential family and the Generalized Beta distribution also provide accurate estimates, as shown in [8, 42]. Lorenz curve estimation through the World Bank's POVCAL and SimSIP computational tools is widespread.

the population median (so that the true headcount ratio is around 50%). In contrast to these results, the parametric approach of estimating the Lorenz curve from grouped data appears consistently to fare better. Across all distributions, poverty lines, and poverty indicators considered, the empirical biases tend to be of smaller, often negligible magnitude.

We also assess the sensitivity of global poverty estimates obtained with the kernel density estimator to the choice of bandwidth. The bandwidth is a key parameter in nonparametric methods which controls the smoothness of the estimated density. Larger bandwidths are associated with smoother densities. Using grouped data from the World Bank's Povcalnet database for a large number of countries, we find that the estimated level of global poverty in 1995 and 2005 varies markedly with the choice of bandwidth. In contrast, the estimated trend of poverty reduction over the period is robust across bandwidths. Taken together, our findings suggest that researchers who employ nonparametric methods to analyze poverty should assess the robustness of their results to alternative parameter choices as a matter of routine, especially when using grouped data rather than individual records. Furthermore, preference for the parametric approach may be warranted due to its superior performance for a wide range of income distributions.

It will perhaps be unsurprising that applying nonparametric methods on sparse data gives rise to biases in the estimated income distribution and poverty measures. The purpose of nonparametric estimators is to provide means of uncovering patterns using information from a wealth of observations and they therefore work best on large samples. The statistical literature advises that they should be used in "exploratory data analysis, as a confirmatory tool, or as a supplement to the standard parametric fare" [43, p. 672].[4] Although their application to grouped data is almost sure to generate biases, the sign and magnitudes of these biases—for distinct poverty indicators and poverty lines, and for various income distributions—are unknown *ex ante*. Our goal is to document these biases for a range of plausible income distributions and to inform readers of possible caveats when applying these techniques to grouped data. Current debates on the extent and trend of world poverty underscore the importance of assessing the performance of alternative statistical methods.

The remainder of the paper is organized as follows. We discuss the problem of estimating poverty from grouped data using the kernel density estimator in Section 2. Section 3 describes the Monte Carlo design and the bandwidths and kernels considered. Section 4 presents the results of our Monte Carlo simulations for both the parametric and KDE approach. In Section 5 we discuss the sensitivity of global poverty estimates to changes in the bandwidth. Conclusions are deferred to Section 6. The results presented in the paper are accompanied by a series of robustness checks available in a Supplementary Online Appendix.[5]

[4] Yatchew [43, p. 715] further argues that "interpolation is only deemed reliable among close neighbour[ing] observations, and extrapolation outside the observed domain is considered entirely speculative."

[5] The Supplementary Online Appendix may be downloaded from www.camelia-minoiu.com/kde-online.pdf

2 The problem

We begin by defining the poverty indicators on which we focus in the analysis, as well as the data and the kernel density estimator.

2.1 Obtaining an estimate of poverty

Poverty is usually estimated using individual records from a household survey that collects information on a variable of interest such as income or consumption. Denoting the individual incomes in a survey of N individuals as $\{X_1, X_2, ..., X_N\}$ and z as the poverty line, the most popular poverty measures come from the FGT family generally written as $P_\alpha = \frac{1}{N} \sum_{X_i \leq Z}^{N} \left(\frac{z - X_i}{z}\right)^\alpha$, where α captures the degree of distributional sensitivity. The higher is the α, the more weight is placed on the income shortfalls from the poverty line experienced by the poorest individuals. For $\alpha = 0$ we obtain the poverty headcount ratio (the proportion of the population that is poor). Values $\alpha = 1$ and $\alpha = 2$ yield the poverty gap and the squared poverty gap. Here we consider values for α ranging between 0 and 4.

2.2 Grouped data vs. individual records

Suppose that individual records from the survey itself are unavailable but the researcher has access to grouped data. Grouped data are income averages for a number of population groups (for instance, quintiles, deciles and ventiles, corresponding respectively to five, ten, and twenty population groups). In what follows, we focus on deciles (rather than quintiles) because they have become increasingly available in recent years. We also briefly discuss the properties of quantile means as linear functions of order statistics and robust estimators of location to provide a rationale for our empirical findings. Decile means are obtained by first ordering the original income observations in ascending order to obtain order statistics $\{X_1 \leq X_2 \leq ... \leq X_N\}$, then dividing the sample into $J = 10$ groups of equal size M, and finally calculating income averages for each group j as $\left(u_j = \frac{1}{M} \sum_{i=1}^{M} X_i^{(j)}\right)$.

It should be noted that a dataset of decile means $\{u_1, u_2, ..., u_{10}\}$contains more information about the underlying distribution than does a dataset of ten random observations from that distribution. This is because grouping the data means transforming the individual records into summary information about the underlying distribution. Thus, quantile means in general, and decile means in particular, retain important information about the underlying distribution due to the ordering of the original observations. One way to see this is to think of them as *trimmed means*—that is, averages obtained over observations remaining after certain percentages of the lowest and the highest scores have been discarded. Trimmed means can further be *symmetric* or *asymmetric*. For example, the lowest decile mean, which is the average of incomes left after discarding the top 90% of observations, is an asymmetrically trimmed mean. In contrast, the middle quintile mean, obtained by averaging over

the incomes remaining after discarding the bottom and the top 40% of observations, is a symmetrically trimmed mean.

Following Mosteller's [31] seminal work on order statistics, it has been shown that *symmetrically* trimmed means are robust estimators of location. Furthermore, they are unbiased estimators for the population mean when the data are drawn from a symmetric distribution. This may be relevant to the case of income distributions because if the income distribution is log-normal, then log-incomes are distributed symmetrically. Since the log-normal distribution is traditionally considered as a good model for real-world income distributions, we might expect that the middle quantile means will yield relatively accurate estimates of the location of the underlying distribution. Indeed, we find below that the population median tends to be well estimated for a range of plausible log-income distributions and especially for the (symmetrical) normal distribution. This is the case irrespective of the bandwidth and kernel used in the estimation.

2.3 The kernel density estimator on grouped data

Assuming that the individual records in the household survey are *i.i.d.* draws from an unknown density $f(x)$ with positive support $[0, \infty)$, the kernel density estimator of $f(x)$ computed on the grouped data is given by: $\hat{f}(x)_{\text{grouped\_data}} = \frac{1}{Jh} \sum_{j=1}^{J} k\left(\frac{x-u_j}{h}\right)$, where h is the bandwidth and $k(\cdot)$ is the weighting function or kernel. Following the derivation of bias employed by Silverman [37], the bias of the grouped-data estimator at$x$ can be shown to be:

$$Bias\,\hat{f}(x)_{\text{grouped\_data}} \cong \frac{1}{J}\sum_{j=1}^{J} g_j(x) + \frac{h^2}{2J}\left(\int t^2 k(t)\,dt\right)\sum_{j=1}^{J} g_j{}''(x) - f(x)$$

where $g_j(\cdot)$ is the unknown probability density function of the j^{th} quantile mean, $\int t^2 k(t)\,dt$ is a constant depending on the kernel, and (small) higher-order terms in h arising from a Taylor approximation have been omitted for simplicity. Notably, the bias of the kernel density estimator is a function of the unknown data generating process $f(x)$—a key feature of nonparametric estimators. Evaluating the bias exactly requires analytical expressions for the density of random variables underpinned by trimmed means, which are intractable to arrive at, inter alia because the ordering of the original observations induces a complex correlation structure among the quantile means. We therefore resort to Monte Carlo simulations to quantify the bias of the estimator.

A second issue arising in the estimation is that the support of an income distribution generally has a left hand-side bound of zero. Whether KDE is applied to individual records or grouped data, the density close to the boundary will have a substantial downward boundary bias (as documented by Marron and Ruppert [24]), which will affect poverty estimates and distort visual illustrations of the income distribution. To mitigate the boundary bias, we follow the standard practice of log-transforming the income averages before estimating the density.

3 Set-up of Monte Carlo simulations

3.1 Monte Carlo design

In the Monte Carlo analysis, we undertake kernel density analysis on grouped data in three steps, as follows:[6]

First, we generate independent random samples of 10,000 observations from three theoretical distributions: the two-parameter log-normal, the three-parameter Dagum, and the four-parameter Generalized Beta 2 distribution (GB2). (We refer below to the number of such samples generated as the number of replications.) The distributions are parameterized with values reported in [4] from the fitting of these distributions to survey data from Russia (1995), Poland (1992), and Mexico (1996). Inspecting which distribution fits best a wide range of household surveys, the authors conclude that the Dagum distribution provides the best fit to survey data in the class of three-parameter distributions, while the GB2 distribution is the best performing four-parameter distribution. We add to the analysis the two-parameter log-normal distribution because it is widely used in the income distribution literature and has been often argued to provide a good fit to real-world income data [22]. The population distributions we use for incomes and log-incomes (shown in Fig. 1) have diverse distributional shapes, with the Gini coefficient of inequality ranging from 0.36 for the Dagum to 0.6 for the log-normal.

In the second step, we calculate quantile means from each sample and perform KDE upon these means to estimate the income distribution. For the majority of the results presented here we work with decile means, but also consider quintile and ventile means to investigate the link between the number of data points and the accuracy of the estimator. For the parametric analysis, in this step we estimate the Lorenz curve from the grouped data respectively using the GQ and Beta functional forms, the parameters of which are estimated by means of regression.[7]

In the third step, we compare summary statistics and poverty estimates from the fitted densities, either through the nonparametric or parametric approach, with the population values from the underlying (theoretical) distributions. Population values were obtained analytically. For the Dagum distribution, formulas for the density and quantile functions were obtained from [13, 20], while formulas for the GB2 distribution were taken from [25].

To ensure that our Monte Carlo estimates are accurate in the sense that the sample quantile means are statistically 'close' to the corresponding population values, we calculate the minimum number of replications (for each theoretical distribution and quantile mean) that ensure a small percentage error in each sample estimate vis-à-vis the population value (see [11, 12, 32]). The minimum number of replications

[6]For the Monte Carlo simulations, we use specially-designed software *KDETool* (available from the authors upon request) and the *DASP* STATA package [1].

[7]See [7, 9] for a detailed discussion of parametric models for the Lorenz curve.

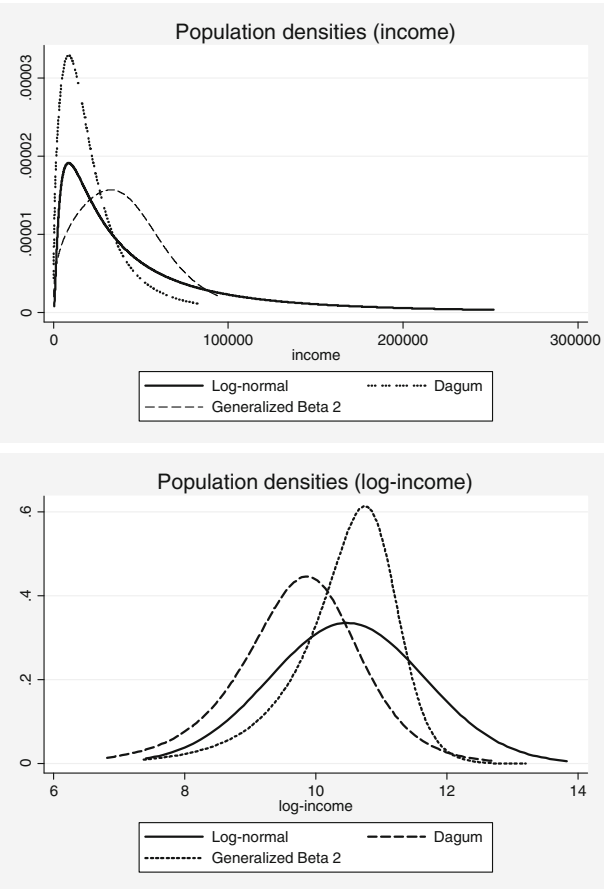


Fig. 1 Population densities for Monte Carlo simulations

Notes: The population densities have been parameterized according to income distributions that perform well in fitting survey data (see Section 3.1 for details). In the first panel the curves have been clipped at the 90th percentile for better visibility.

needed to assert with 95% confidence that our sample quantile means are off by at most 1 percent from the population means is 1,842. Therefore, we fix the number of replications at 2,000.[8]

[8]The minimum number of replications that ensures a given percentage error (equal to *100 x RE*) is given by $R = \left(\frac{(Z_{\alpha/2})/s}{\left(\frac{RE}{1+RE}\right)\bar{x}} \right)^2$, where RE is the relative error, s is the sample standard deviation which approximates the population standard deviation, $\bar{x}$ is the sample estimate of the theoretical x (i.e., a quantile mean), and $Z_{\alpha/2}$ is the standard normal distribution evaluated at $(\alpha/2)$, with α being the level of confidence with which we can assert that the difference between $\bar{x}$ and x exceeds the specified relative error amount RE: $P(|\bar{x} - x|) > RE|x|) = \alpha$.

3.2 Bandwidths and kernels

For the nonparametric estimator we consider eight data-driven bandwidths. First, we use the first generation, rule-of-thumb optimal bandwidths proposed by Deheuvels [10] and described in [37]. These bandwidths—labeled S1 to S4—are optimal in the sense that they seek to minimize the approximate mean integrated squared error. In doing so, they also make the assumption that the underlying distribution is normal, which means that they work best when estimating Gaussian-like distributions. The S1 bandwidth, which is given by $\left(1.06 \times \sigma \times J^{-1/5}\right)$ where σ is the standard deviation of the data and J is number of quantiles, tends to over-smooth the density and performs poorly on heavily skewed distributions. The S2 bandwidth is calculated by replacing the standard deviation with the interquartile range (IQR) as a measure of dispersion and is given by $\left(0.79 \times IQR \times J^{-1/5}\right)$. The S2 bandwidth performs better than S1 on long-tailed and heavily-skewed distributions, but does not do well on multimodal distributions. Two other variants proposed by Silverman [37] are the S3 bandwidth given by $\left(0.9 \times A \times J^{-1/5}\right)$ where $A = \min\left(IQR/1.34, \sigma\right)$ and the S4 bandwidth which is calculated by replacing A with σ in the above formula. The S3 and S4 bandwidths achieve a more balanced degree of smoothing than S1 and S2 and have been deemed to work well on skewed and multimodal distributions. In addition to S1–S4, we also include some results based on the 'over-smoothed' bandwidth, which is the largest bandwidth associated with a "reasonable" degree of smoothing [16]. The over-smoothed bandwidth is given by $\left(1.14 \times \sigma \times J^{-1/5}\right)$.

Second, we consider second-generation bandwidths such as the Sheather-Jones [36] and the direct plug-in (DPI) bandwidths [40]. Second-generation bandwidths tend to outperform Silverman's rule-of-thumb bandwidths theoretically and in simulations. The Sheather-Jones bandwidth has been recommended as a benchmark of good performance in simulation-based studies [17]. The DPI bandwidths have also been shown to perform very well in simulations, but are less appropriate for multimodal densities. The key difference between them is that the DPI bandwidth requires choosing a starting bandwidth (typically, a rule-of-thumb one), estimating the density, and obtaining an estimate of data dispersion. Then the density is re-estimated. The process can be repeated several times. We only consider the DPI-1 and DPI-2 bandwidths which involve respectively one and two re-estimations of the density.

The constants shown in the bandwidth formulas above correspond to the Gaussian kernel. For all other kernels we employ canonical bandwidths, which means that the constants are kernel-specific. Canonical bandwidths ensure that each bandwidth-kernel combination achieves the same approximate value of the integrated mean square error, so they lead to the same amount of smoothing [23]. This renders the results comparable across different kernels for a given bandwidth.

In regards to kernel functions, we consider the Gaussian, Epanechnikov, Quartic, Triweight and Triangle kernels. (For a comprehensive treatment of kernels, see [21]). Although the mean integrated squared error is minimized for the Epanechnikov kernel, all kernels have similar asymptotic performance [37]. Since our analysis is based on a small number of data points, we have no reason *ex ante* to discard any particular kernel. While most of the results shown in the paper refer to the commonly-used Gaussian and Epanechnikov kernels, they are robust to using alternate kernels and are included in the Supplementary Appendix.

4 Results of Monte Carlo simulations

Here we present the results of the Monte Carlo study, comparing estimated quantities such as summary statistics and poverty indicators based on the parametric and nonparametric approaches with their population values. Unless otherwise noted, the results are based on deciles and are expressed as a ratio of the estimate of interest, averaged over the 2,000 replications, and the corresponding population value. Values greater than one indicate that the technique leads to an overestimation of the true value, and values less than one indicate the opposite. To gauge the statistical significance of our results, we use boldface in the tables to highlight cases in which the population value lies *inside* the 95% confidence interval around the average estimate. All figures not in boldface therefore indicate that the estimator does not produce estimates within a 95% confidence interval of the population value.

4.1 Basic features of KDE-based densities

We first compare summary statistics (means, medians, standard deviation, and decile means) from the simulated samples with their population values. The results are shown in Table 1 (Panel A) for the Epanechnikov and the Gaussian kernels and four bandwidths (S4, S2, Sheather-Jones, and DPI-1).[9] The two kernels were chosen illustratively due to their widespread use in empirical studies. Panel A depicts some striking results, as we see that almost all estimated quantities are biased. Focusing on the first three columns, we find that the mean is systematically overestimated, especially for the Epanechnikov kernel. However, the size of the bias varies with both the underlying distribution and the bandwidth. The standard deviation is also consistently overestimated, which suggests that the KDE-fitted density is too smooth compared to the true one. In contrast, the median is well estimated through KDE, and so are the middle decile means (Q4 to Q6, but especially Q5). This is consistent with the statistical literature which has shown that symmetrically trimmed means—in our case, the middle decile means—are good indicators of the location of the true distribution [3, 38]. However, comparing the results solely across the last ten columns (Q1 to Q10) in Panel A, we find that the average income of the poorest population deciles tends to be underestimated whereas the average income of the richest deciles tends to be overestimated. In other words, the poor appear poorer and the rich appear richer. (This is especially the case for the log-normal and Dagum distributions, for reasons explained below.) The estimator systematically generates distortions in the tails of the distributions, which are crucial to estimating poverty.

Panel B presents the same results based on parametric estimation of the Lorenz curve. While the mean and median of the underlying distributions are well estimated, the parametric approach also tends to overestimate the standard deviation. Nevertheless, the decile means are generally more accurately estimated with this approach; the GQ functional form fares particularly well in estimating features of the log-normal distribution, whereas the Beta model works best for

[9] Results based on the remaining bandwidths are shown in Table A1 in the Supplementary Appendix.

Table 1 Summary statistics

Kernel-bandwidth pair	Underlying distribution	Mean	Median	St. Dev.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Panel A. Kernel density estimation														
Epanechnikov, S4	Log-normal	1.45	**1.00**	1.61	0.68	0.75	0.82	0.90	**0.98**	1.08	1.19	0.72	1.54	1.83
	Dagum	1.28	**0.98**	1.45	0.80	0.76	0.81	0.87	**0.95**	**1.05**	1.17	1.31	1.51	1.51
	GB2	1.17	0.94	1.51	0.92	0.82	0.83	0.87	0.92	**0.99**	1.07	1.19	1.34	1.57
Epanechnikov, S2	Log-normal	1.35	**1.01**	1.41	0.74	0.79	0.86	0.92	**0.99**	1.07	1.16	0.69	1.46	1.65
	Dagum	1.17	**0.99**	1.21	0.88	0.82	0.86	0.91	**0.97**	1.04	1.12	1.23	1.38	1.30
	GB2	1.11	0.95	1.31	**0.98**	0.86	0.88	0.91	0.94	**0.99**	1.05	1.13	1.23	1.38
Epanechnikov, Sheather-Jones	Log-normal	1.59	**1.01**	1.86	0.60	0.69	0.78	0.88	**0.98**	1.10	1.24	0.76	1.68	2.08
	Dagum	1.28	**0.98**	1.47	0.80	0.76	0.81	0.87	0.95	1.05	1.17	1.31	1.51	1.52
	GB2	1.14	0.94	1.43	0.95	0.84	0.85	0.88	0.93	**0.99**	1.06	1.16	1.29	1.49
Epanechnikov, DPI-1	Log-normal	1.24	**1.01**	1.19	0.82	0.84	0.89	0.94	**1.00**	1.05	1.12	0.66	1.38	1.44
	Dagum	1.11	**0.99**	1.09	0.95	0.86	0.90	0.94	**0.98**	1.03	1.09	1.18	1.31	1.19
	GB2	1.07	0.96	1.19	**1.02**	0.90	0.92	0.93	0.95	**0.99**	1.04	1.10	1.17	1.27
Gaussian, S4	Log-normal	1.19	**1.01**	1.01	**0.86**	0.85	0.91	**0.98**	1.06	1.16	1.28	0.78	1.80	1.53
	Dagum	1.11	**0.98**	1.00	0.96	0.84	0.88	**0.94**	**1.01**	1.10	1.22	1.39	1.69	1.31
	GB2	1.08	0.94	**1.20**	1.08	0.88	0.88	0.91	**0.96**	**1.03**	1.11	1.22	1.40	1.42
Gaussian, S2	Log-normal	1.14	**1.01**	0.93	0.91	0.88	0.93	**0.99**	1.06	1.14	1.24	0.75	1.69	1.41
	Dagum	1.05	**0.99**	0.89	1.04	0.89	0.92	**0.97**	**1.02**	1.09	1.17	1.30	1.53	1.16
	GB2	1.05	0.95	1.09	1.13	0.93	0.93	0.95	**0.98**	**1.02**	1.08	1.17	1.29	1.27
Gaussian, Sheather-Jones	Log-normal	1.27	**1.01**	**1.11**	0.78	0.80	0.88	**0.97**	1.07	1.19	1.34	0.84	1.99	1.73
	Dagum	1.11	**0.98**	**1.01**	0.96	0.83	0.87	0.94	**1.01**	1.10	1.22	1.39	1.68	1.31
	GB2	1.07	0.94	1.15	1.10	0.90	0.90	0.93	**0.97**	**1.02**	1.10	1.20	1.35	1.36
Gaussian, DPI-1	Log-normal	1.23	**1.01**	1.05	0.82	0.82	0.90	**0.98**	1.07	1.17	1.31	0.81	1.89	1.62
	Dagum	1.09	**0.99**	**0.97**	**0.99**	0.85	0.89	0.94	**1.01**	1.10	1.20	1.36	1.63	1.26
	GB2	1.06	0.94	1.13	1.11	0.91	0.91	0.93	**0.97**	**1.02**	1.09	1.19	1.33	1.33

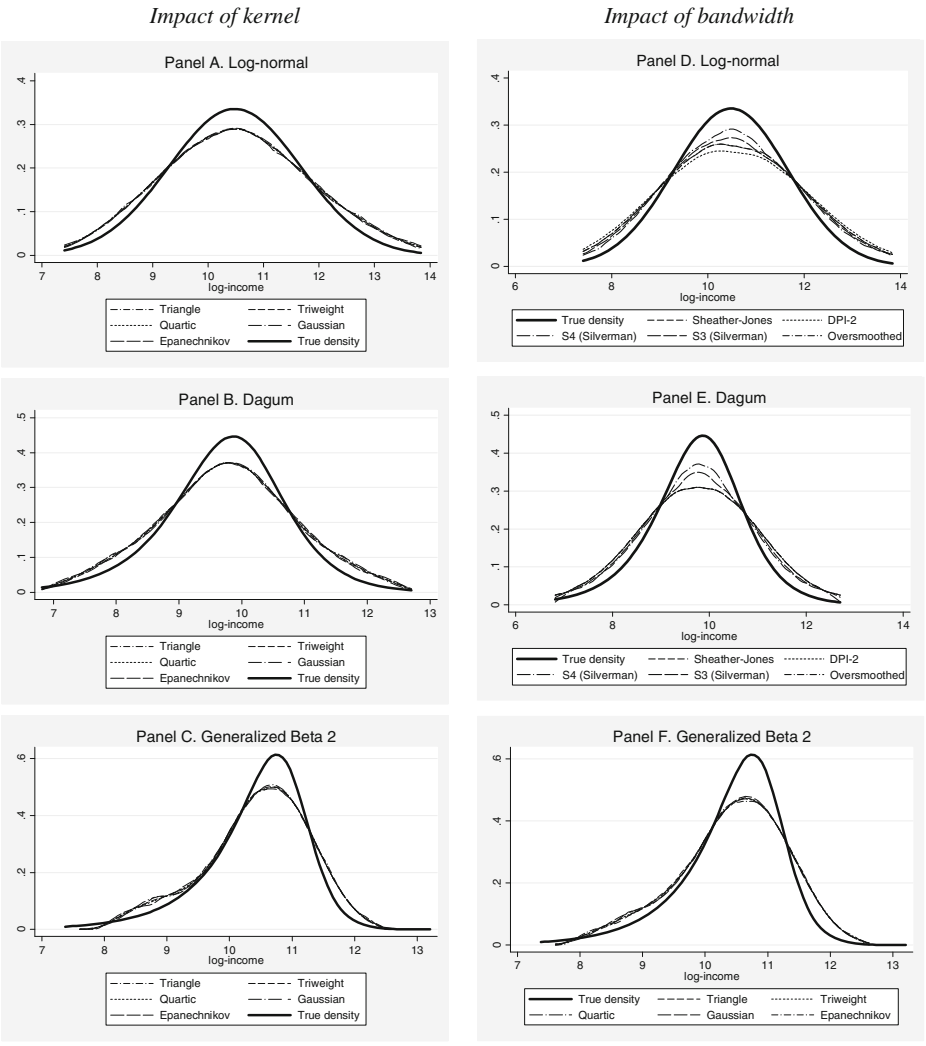
Table 1 (continued)

Kernel-bandwidth pair	Underlying distribution	Mean	Median	St. Dev.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Panel B. Lorenz Curve parametric estimation														
GQ	Log-normal	**1.00**	**1.01**	0.93	1.09	**0.97**	**0.98**	**0.99**	**1.00**	**1.01**	**1.00**	0.94	**0.99**	**1.00**
	Dagum	**0.97**	**1.00**	1.24	1.11	**0.99**	**1.01**	**1.03**	1.05	1.07	1.08	1.11	1.22	1.29
	GB2	**0.99**	0.88	1.08	1.61	1.08	0.95	0.90	0.88	0.89	0.91	0.94	**0.99**	1.14
Beta	Log-normal	**1.00**	1.08	1.12	1.14	0.89	**0.96**	**1.03**	1.07	1.08	1.06	0.94	**0.97**	**0.98**
	Dagum	**0.97**	**1.02**	1.09	1.06	**1.01**	**1.00**	**1.01**	**1.02**	**1.03**	**1.03**	**1.02**	**1.00**	0.91
	GB2	**1.00**	**0.99**	0.94	1.05	1.05	**1.00**	**0.99**	**0.99**	**0.99**	**1.00**	**1.00**	**0.99**	**0.99**

The figures represent the ratio between the KDE-based estimate from decile means (averaged over 2,000 replications) and the population value (Panel A); and the parametric estimate (averaged over 2,000 replications) and the population value (Panel B). Q1–Q10 are the decile means, from lowest (Q1) to highest (Q10). Estimates are based on decile means. Figures in boldface represent cases in which the population value lies inside the 95 percent confidence interval around the average estimate.

the Dagum and GB2 distributions. It is noteworthy to find evidence that specific functional forms may work better for parametric estimation of different income distributions.

Average KDE-based densities are compared to their population counterparts in Fig. 2 for each distribution. In Panels A-C we assess the impact of different kernels on the fitted density holding the bandwidth fixed. The choice of kernel does not appear visually consequential. In Panels D-F we assess the impact of different bandwidths on the fitted density holding the kernel fixed. The choice of bandwidth appears more important than that of kernel. Figure 2 also depicts a series of consistent patterns



Note: Estimates are based on decile means. Panels A-C show the impact of changing the kernel on the density estimates and correspond to the S3 bandwidth. Panels D-F show the impact of changing the bandwidth on the density estimate and correspond to the Epanechnikov kernel.

Fig. 2 Visual illustrations

across distributions.[10] First, the fitted densities are centered correctly vis-à-vis the population densities. Second, the bias in the fitted densities varies with the income level. The estimated densities are biased upwards in the left tail (e.g., for the log-normal and Dagum distribution), which is consistent with the mean income of the poorest deciles being underestimated.[11] The density estimates are biased downwards in the middle of the distribution, and the bias turns positive again in the right tail of the distribution, hence the overestimated average income of the richest.

These patterns give us a preview of the performance of the estimator in poverty analysis. With the mean income of the poorest being systematically underestimated, we expect the poverty headcount ratio to be overestimated for relatively low poverty lines. As the bias of the density turns from positive to negative on the log-income axis, there comes an income level (or poverty line) for which the biases will cancel out. We expect the poverty headcount ratio to be relatively well estimated when the poverty lines are close to the center of the distribution. Beyond that, toward the right, the KDE-based density continues to be underestimated, which suggests that for higher poverty lines the headcount ratio may be biased downwards.

4.2 Visualizing the empirical biases across poverty lines

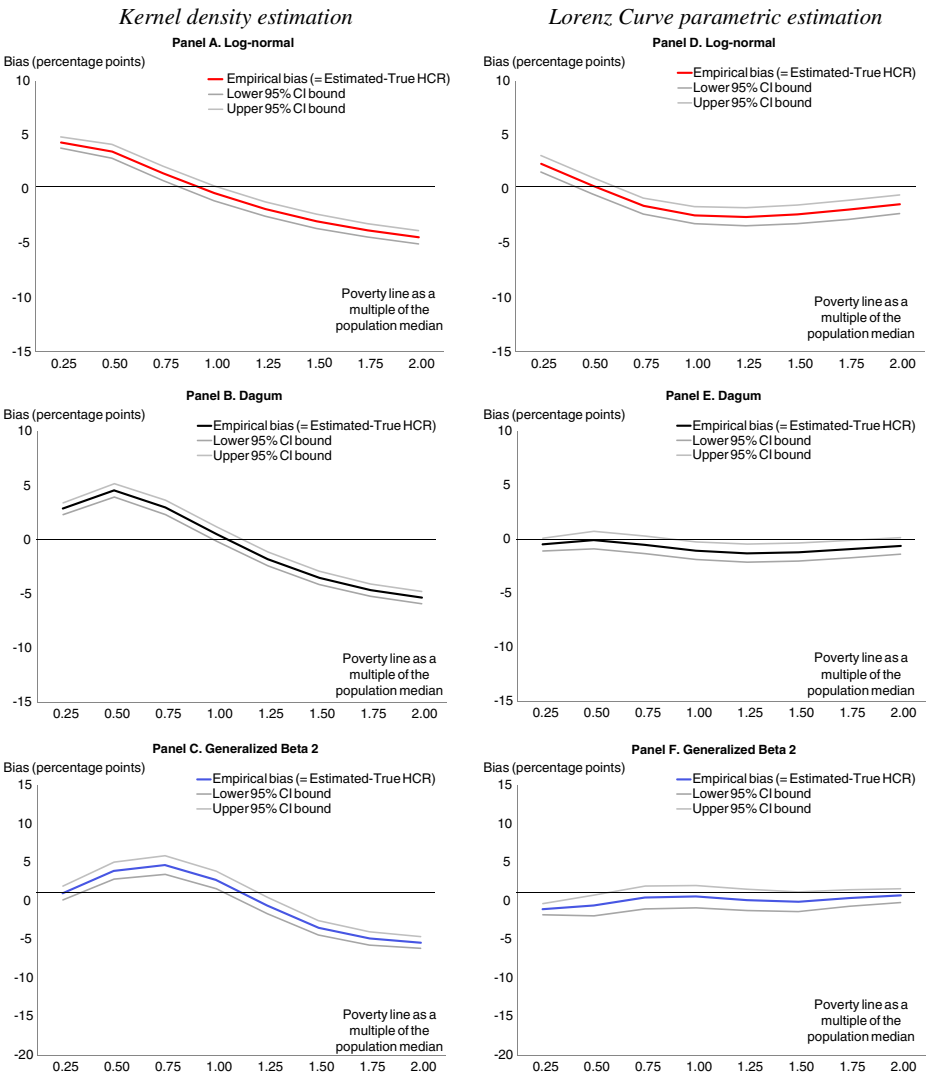
To describe how biases in the fitted densities relate to the accuracy of estimating poverty, we focus on the widely-used poverty headcount ratio (HCR). We use poverty lines ranging from a low poverty line which represents only a fraction of the population median (0.25 $\times$ median) to a high poverty line which is twice as large as the population median (2 $\times$ median). Figure 3 depicts the empirical bias of the estimated HCR as a function of the poverty line, along with a 95% confidence interval. The bias is calculated as the difference between the average estimated HCR and the population HCR, and is expressed in percentage points. Panels A–C in Fig. 3, each corresponding to a different distribution, show that KDE on grouped data leads the HCR to be overestimated for lower poverty lines, relatively well estimated for poverty lines close to median income, and underestimated for higher poverty lines. The HCR bias is positive and high for low poverty lines (at about 5 percentage points), diminishes as the poverty line approaches the population median, and becomes negative for higher-than-median poverty lines (up to about -5 percentage points). Although Fig. 3 plots the empirical bias for a particular kernel-bandwidth combination (the Gaussian kernel and DPI-2 bandwidth), the patterns are illustrative for other kernel-bandwidth pairs.[12]

Overall, it appears that as long as the underlying income distribution is unimodal and resembles one of the theoretical distributions considered here, kernel smoothing methods are likely to lead poverty to be overestimated in richer countries, which have

[10] These are robust to considering alternate bandwidths and kernels (see Figs. A1–A2 in the Supplementary Appendix).

[11] An exception here is the GB2 distribution, for which the fitted density is first biased downwards and then biased upwards in the left tail. The crossing of the estimated and true log-income densities explains why the degree of misestimation of the decile means does not vary monotonically with the decile rank.

[12] See Fig. A3 and Table A2 in the Supplementary Appendix.



Note: The figures show the empirical bias in the poverty headcount ratio (i.e., the difference between the estimated and the population headcount ratio, averaged over 2,000 replications), in percentage points. Panels A-C refer to the KDE-based poverty headcount ratio (Gaussian kernel, DPI-2 bandwidth); Panels D-F refer to the parametric approach (Beta model). Estimates are based on decile means. The poverty lines are expressed as multiples of the population median, ranging from a quarter of the median (0.25) to twice (2) the median. The grey lines represent the lower and upper bound of 95 percent confidence intervals.

Fig. 3 How does the bias vary across poverty lines?

national poverty lines that are low relative to median income, and underestimated in many poorer countries, where national poverty lines may be higher than median income. It is difficult to say what would happen in a regional or global poverty analysis when the sample contains both poor and rich countries, as some of the biases may cancel out. The dominating effect will likely depend both on the share of poor countries and their relative sizes.

These findings stand in contrast with the results from the parametric approach (Fig. 3, Panels D–F). For instance, when the Beta functional form is used, the empirical bias is smaller than for the nonparametric estimates discussed above, for all distributions. Consistent with the relatively good fit for the decile means afforded by the Beta model for the Lorenz curve, the bias is zero or negligible across all poverty lines in the case of the Dagum and GB2 distributions (Panels E–F).[13] Conditional on income data being generated by the distributions chosen here, the parametric approach appears to consistently outperform the nonparametric approach for estimating poverty from grouped data.

4.3 Empirical biases across bandwidths and poverty lines

In Table 2 we show how the empirical bias in the HCR varies across bandwidths (keeping the kernel constant) and poverty lines. The results are based on the Epanechnikov kernel and are representative for other kernels such as the Gaussian and Triweight.[14] From the first column to the last, the bandwidths are arranged roughly in ascending order of size, with larger bandwidths bringing about a higher degree of smoothing. Comparing across rows allows us to gauge the impact of the poverty line. For the lower-than-median poverty lines, the biases are positive and statistically significant, and become larger as the amount of smoothing increases. For the higher-than-median poverty lines, the biases are negative, statistically significant, and also increase with the bandwidth. It is only when the poverty line is equal to the population median that the nonparametric estimator accurately measures the HCR.

While the sign of the bias appears robust across income distributions, its size varies for any given bandwidth. For instance, for the lowest poverty lines ($0.25 \times$ median), the Sheather-Jones bandwidth leads the HCR to be overestimated by 38% for the log-normal distribution, 41% for the Dagum distribution, and 26% for the GB2 distribution. Similarly, for the highest poverty line ($2 \times$ median), the same bandwidth leads the HCR to be underestimated by 6% for the log-normal, 8% for the Dagum, and 9% for the GB2 distribution. This suggests that that KDE will lead to biases of different magnitude depending on the data generating process, so our results cannot be generalized. Nevertheless, they provide robust information about the *sign* of the bias as a function of the location of the poverty line relative to the population median. The researcher using this technique may expect a positive, negative, or zero bias depending on her prior belief about how the poverty line in a specific country compares with the true median.

4.4 Empirical biases across kernels and poverty lines

To examine the biases in the KDE-based HCR across selected kernels, we fix the degree of smoothing (by fixing the bandwidth to be S4). The results are reported in

[13] The results for the GQ model are similar (see Figure A4 in the Supplementary Appendix).

[14] See Table A3 in the Supplementary Appendix.

Table 2 How does the bias vary across bandwidths?

Underlying distribution	Poverty line (as multiple of population median)	Bandwidth							
		S4	S3	S2	Sheather-Jones	DPI-1	DPI-2	S1	Oversmoothed
Log-normal	0.25	1.27	1.27	1.36	1.38	1.44	1.44	1.48	1.54
	0.50	1.10	1.27	1.36	1.38	1.44	1.44	1.48	1.54
	0.75	**1.03**	**1.03**	1.04	1.04	1.05	1.05	1.05	1.06
	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	1.25	**0.98**	**0.98**	**0.97**	**0.97**	**0.97**	**0.97**	0.96	0.96
	1.50	**0.97**	**0.97**	0.96	0.96	0.95	0.95	0.94	0.94
	1.75	0.96	0.96	0.95	0.95	0.94	0.94	0.93	0.93
	2.00	0.96	0.96	0.94	0.94	0.93	0.93	0.93	0.92
Dagum	0.25	1.25	1.25	1.33	1.41	1.47	1.47	1.42	1.47
	0.50	1.15	1.25	1.33	1.41	1.47	1.47	1.42	1.47
	0.75	1.06	1.06	1.08	1.09	1.11	1.11	1.10	1.10
	1.00	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**
	1.25	**0.98**	**0.98**	**0.97**	0.96	0.96	0.96	0.96	0.96
	1.50	0.96	0.96	0.95	0.94	0.93	0.93	0.94	0.93
	1.75	0.95	0.95	0.93	0.92	0.92	0.92	0.92	0.92
	2.00	0.94	0.94	0.93	0.92	0.91	0.91	0.92	0.91
Generalized Beta 2	0.25	1.18	1.19	1.22	1.26	1.26	1.30	1.23	1.25
	0.50	1.14	1.19	1.22	1.26	1.26	1.30	1.23	1.25
	0.75	1.10	1.10	1.12	1.15	1.16	1.17	1.15	1.15
	1.00	1.04	1.04	1.05	1.05	1.05	1.05	1.05	1.05
	1.25	**0.99**	**0.99**	**0.98**	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**
	1.50	0.95	0.95	0.94	0.93	0.93	0.92	0.94	0.93
	1.75	0.94	0.94	0.93	0.92	0.92	0.91	0.92	0.92
	2.00	0.94	0.94	0.93	0.91	0.91	0.90	0.92	0.91

The figures represent the ratio between the grouped data KDE-based poverty headcount ratio (averaged over 2,000 replications) and the population value. Estimates are based on decile means and the Epanechnikov kernel. Figures in boldface represent cases in which the population value lies inside the 95 percent confidence interval around the average estimate.

Table 3 for multiple poverty lines.[15] Comparing across rows, we notice that KDE-based poverty estimates are relatively insensitive to the choice of kernel, which is consistent with Fig. 2 (Panels A–C). Comparing across columns, the degree of variation in the empirical bias corresponding to each poverty line seems robust across kernels. Put differently, the pattern identified above—that the HCR tends to be overestimated for lower poverty lines and underestimated for higher poverty lines—holds up across kernels.

4.5 Empirical biases across poverty indicators

We also explore whether the biases discussed above are confined to the HCR or afflict other poverty indicators as well. We consider indicators that take account of the depth of poverty, measured as the distance between the income of the poor

[15] For alternate bandwidths, see Table A4 in the Supplementary Appendix.

Table 3 How does the bias vary across kernels?

Underlying distribution	Kernel	Poverty line (as multiple of population median)							
		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Log-normal	Epanechnikov	1.27	1.10	**1.03**	**1.00**	**0.98**	**0.97**	0.96	0.96
	Gaussian	1.18	1.07	**1.02**	**1.00**	**0.98**	**0.97**	**0.97**	**0.97**
	Quartic	1.27	1.10	**1.03**	**1.00**	**0.98**	**0.97**	0.96	0.96
	Triweight	1.28	1.10	**1.03**	**1.00**	**0.98**	**0.97**	0.96	0.96
Dagum	Epanechnikov	1.25	1.15	1.06	**1.01**	**0.98**	0.96	0.95	0.94
	Gaussian	1.14	1.11	1.05	**1.01**	**0.98**	**0.97**	0.96	0.96
	Quartic	1.25	1.14	1.06	**1.01**	**0.98**	0.96	0.95	0.95
	Triweight	1.26	1.14	1.06	**1.01**	**0.98**	0.96	0.95	0.95
Generalized Beta 2	Epanechnikov	1.18	1.14	1.10	1.04	**0.99**	0.95	0.94	0.94
	Gaussian	1.05	1.10	1.09	1.04	**1.00**	**0.97**	0.96	0.96
	Quartic	1.19	1.14	1.10	1.04	**0.99**	0.95	0.95	0.94
	Triweight	1.19	1.14	1.10	1.04	**0.99**	0.96	0.95	0.94

The figures represent the ratio between the grouped data KDE-based poverty headcount ratio (averaged over 2,000 replications) and the population value. Estimates are based on decile means and the S4 bandwidth. Figures in boldface represent cases in which the population value lies inside the 95 percent confidence interval around the average estimate.

and the poverty line, such as the poverty gap ratio FGT(1), the squared poverty gap FGT(2), and the still more distributionally-sensitive FGT(3) and FGT(4). (For definitions see Section 2.1). The results are depicted in Table 4 for two representative kernel-bandwidth pairs, namely Epanechnikov-S4 and Epanechnikov-DPI-2 (Panels A–B). In Panel C we report for comparison the results based on the estimation of the Lorenz curve using the Beta model. We find that the extent of poverty according to FGT indicators other than the HCR is mostly overestimated by the nonparametric approach, with the biases diminishing with higher poverty lines. By contrast, the parametric estimator consistently leads to remarkably accurate results, with the exception of the lowest two poverty lines.[16]

There are also differences in the size of KDE-generated biases for different underlying income distributions (Table 4, Panels A–B). Data drawn from the log-normal distribution is associated with the highest positive biases. Data from the GB2 distributions are on average associated with lower positive biases and also with negative biases for both the lowest poverty line and for higher poverty lines. The crossing of the fitted and population densities in the left tail of this distribution gives rise to a non-monotonic relationship between the bias in the estimate of the HCR and the poverty line, for both the Dagum and the GB2 distributions. The biases associated with the parametric estimator (Panel C) are smaller, and do not vary systematically across the distributions considered. The results in Table 4, both for the parametric and nonparametric approach, underscore the sensitivity of both parametric and nonparametric methods to the data generating process, and echo our earlier findings on how they fared in recovering summary statistics (Table 1).

[16]The latter finding is not surprising considering the relatively poor performance of the Beta model for estimating the cumulative shares of the lower population groups, for which it sometimes generates negative fitted Lorenz curve estimates, as discussed by Kakwani [19] and empirically documented by Minoiu and Reddy [30].

Table 4 How does the bias vary across poverty indicators?

Underlying distribution	Poverty indicator	Poverty line (as multiple of population median)							
		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Panel A. Kernel-bandwidth pair: (Epanechnikov, S4)									
Log-normal	Poverty headcount ratio	1.27	1.10	**1.03**	**1.00**	**0.98**	**0.97**	0.96	0.96
	Poverty gap ratio	1.36	1.20	1.12	1.07	1.04	1.02	**1.01**	**1.00**
	Squared poverty gap	1.40	1.26	1.18	1.13	1.09	1.07	1.05	1.04
	FGT3	1.40	1.31	1.22	1.17	1.13	1.10	1.08	1.07
	FGT4	1.39	1.34	1.26	1.20	1.16	1.13	1.11	1.09
Dagum	Poverty headcount ratio	1.25	1.15	1.06	**1.01**	**0.98**	0.96	0.95	0.94
	Poverty gap ratio	1.11	1.18	1.13	1.08	1.05	**1.03**	**1.01**	**1.00**
	Squared poverty gap	0.95	1.16	1.15	1.12	1.09	1.07	1.05	**1.03**
	FGT3	0.80	1.11	1.15	1.14	1.12	1.10	1.08	1.06
	FGT4	0.66	1.05	1.13	1.14	1.13	1.12	1.10	1.09
Generalized Beta 2	Poverty headcount ratio	1.18	1.14	1.10	1.04	**0.99**	0.95	0.94	0.94
	Poverty gap ratio	0.94	1.12	1.12	1.10	1.06	**1.03**	**1.01**	**0.99**
	Squared poverty gap	0.71	1.05	1.10	1.11	1.09	1.07	1.05	**1.03**
	FGT3	0.52	0.96	1.07	1.10	1.10	1.09	1.07	1.06
	FGT4	0.38	0.86	1.02	1.07	1.09	1.09	1.09	1.08
Panel B. Kernel-bandwidth pair: (Epanechnikov, DPI-2)									
Log-normal	Poverty headcount ratio	1.54	1.18	1.06	**1.00**	0.96	0.94	0.93	0.92
	Poverty gap ratio	1.88	1.41	1.23	1.14	1.09	1.05	**1.02**	**1.00**
	Squared poverty gap	3.52	2.38	1.97	1.74	1.60	1.51	1.44	1.38
	FGT3	2.37	1.75	1.50	1.36	1.27	1.21	1.17	1.13
	FGT3	2.56	1.89	1.61	1.45	1.35	1.28	1.23	1.19
Dagum	Poverty headcount ratio	1.47	1.26	1.10	**1.01**	0.96	0.93	0.92	0.91
	Poverty gap ratio	1.42	1.37	1.25	1.16	1.09	1.05	**1.02**	**1.00**
	Squared poverty gap	1.32	1.39	1.32	1.25	1.18	1.13	1.10	1.07
	FGT3	1.21	1.39	1.36	1.30	1.24	1.20	1.16	1.12
	FGT4	1.09	1.36	1.37	1.34	1.29	1.24	1.21	1.17
Generalized Beta 2	Poverty headcount ratio	1.25	1.25	1.15	1.05	**0.98**	0.93	0.92	0.91
	Poverty gap ratio	1.06	1.23	1.21	1.16	1.10	1.05	**1.02**	0.99
	Squared poverty gap	1.77	2.05	2.02	1.92	1.80	1.69	1.60	1.53
	FGT3	0.71	1.08	1.18	1.19	1.18	1.16	1.13	1.10
	FGT4	0.57	1.00	1.14	1.18	1.19	1.17	1.16	1.14

Comparing across kernel-bandwidth pairs (Panel A vs. B), we note that the biases are unambiguously larger in Panel B. This is because the DPI-2 bandwidth leads to more smoothing than the S4 bandwidth. This pattern is robust to using alternate kernels.[17]

4.6 Empirical biases across quintiles, deciles, and ventiles

In our assessment of the performance of parametric and nonparametric methods in estimating quantile means, we have focused on deciles. How the techniques fare on datasets of quintiles or ventiles is also of interest. Previous studies of the global income distribution mostly focused on quintiles, while future studies are likely to use

[17] See Table A5 in the Supplementary Appendix.

Table 4 (continued)

Underlying distribution	Poverty indicator	Poverty line (as multiple of population median)							
		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Panel C. Lorenz Curve parametric estimation: Beta									
Log-normal	Poverty headcount ratio	1.20	**1.01**	0.96	0.95	0.96	0.96	**0.97**	**0.98**
	Poverty gap ratio	0.93	1.05	**1.01**	**0.99**	**0.98**	**0.97**	**0.97**	**0.97**
	Squared poverty gap	0.63	**1.01**	**1.02**	**1.01**	**0.99**	**0.98**	**0.98**	**0.98**
	FGT3	0.40	0.93	**1.01**	**1.01**	**1.00**	**1.00**	**0.99**	**0.98**
	FGT4	0.24	0.84	**0.98**	**1.00**	**1.01**	**1.00**	**1.00**	**0.99**
Dagum	Poverty headcount ratio	0.95	**1.00**	**0.99**	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**
	Poverty gap ratio	0.92	**0.98**	**0.99**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
	Squared poverty gap	0.93	0.96	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
	FGT3	**0.98**	0.95	**0.97**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
	FGT4	1.05	0.95	0.96	**0.97**	**0.98**	**0.98**	**0.98**	**0.98**
Generalized Beta 2	Poverty headcount ratio	0.87	**0.97**	**1.01**	**1.01**	**1.00**	**1.00**	**1.00**	**1.01**
	Poverty gap ratio	0.98	0.94	**0.98**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
	Squared poverty gap	1.17	0.96	0.96	**0.98**	**0.99**	**0.99**	**1.00**	**1.00**
	FGT3	1.40	**1.01**	**0.97**	**0.97**	**0.98**	**0.98**	**0.99**	**0.99**
	FGT4	1.66	1.09	**0.99**	**0.97**	**0.97**	**0.98**	**0.98**	**0.99**

The figures represent the ratio between the grouped data KDE-based poverty indicator and the population value (Panels A, B) and between the parametric estimate of the poverty indicator and the population value (Panel C). Poverty indicators are averaged over 2,000 replications. Estimates are based on deciles means. The poverty lines are expressed as multiples of the population median, ranging from a quarter of the median (0.25) to twice (2) the median. Figures in boldface represent cases in which the population value lies inside the 95 percent confidence interval around the average estimate.

more data points, as existing databases such as the UNU-WIDER World Income Inequality database and the World Bank's Global Income Inequality database now contain tabulations for both deciles and ventiles.

Table 5 shows how the errors in the estimated HCR vary across quintile means, decile means, and ventile means. For the nonparametric estimator, we focus on the Epanechnikov and Gaussian kernels and the S2 bandwidth to illustrate a general pattern (Panel A).[18] Comparing across rows, for the higher-than-median poverty lines more data points appear to improve the estimation. By contrast, for the lower poverty lines, perhaps surprisingly, there is no monotonic relationship between the number of quantile means and the bias. In results not reported here, we ran simulations on increasingly larger numbers of quantile means, and did not see monotonicity restored until a threshold of approximately 2530 quantile means was reached. However, 25 to 30 quantile means are rarely, if ever, available to researchers. The results for the parametric approach are shown in Panel B. Although it delivers consistently superior estimates of poverty (for all but the lowest poverty lines), neither the GQ nor the Beta functional forms yield a monotonic relationship between the number of data points and the empirical bias. This is in keeping with earlier evaluations of the performance of these methods on other plausible income distributions (see [30]).

[18]For alternate kernel-bandwidth combinations, see Table A6 in the Supplementary Appendix.

Table 5 How does the bias vary with the number of data points?

Kernel bandwidth pair	Underlying distribution	Number of data points	Poverty line (as multiple of population median) 0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Panel A. Kernel density estimation										
Epanechnikov, S2	Log-normal	Quintiles	1.23	1.09	**1.03**	**0.99**	0.96	0.94	0.93	0.93
		Deciles	1.36	1.13	1.04	**1.00**	**0.97**	0.96	0.95	0.94
		Ventiles	1.30	1.10	**1.03**	**1.00**	**0.98**	**0.97**	0.96	0.96
	Dagum	Quintiles	1.15	1.16	1.08	**1.00**	0.95	0.93	0.92	0.92
		Deciles	1.33	1.20	1.08	**1.01**	**0.97**	0.95	0.93	0.93
		Ventiles	1.27	1.15	1.06	**1.01**	**0.98**	0.96	0.95	0.95
	Generalized Beta 2	Quintiles	0.99	1.19	1.11	**1.03**	**0.97**	0.95	0.94	0.94
		Deciles	1.56	1.38	1.16	**1.01**	0.91	0.85	0.83	0.82
		Ventiles	1.17	1.14	1.10	1.04	**0.99**	0.96	0.95	0.94
Gaussian, S2	Log-normal	Quintiles	1.04	1.04	**1.01**	**0.98**	**0.97**	0.96	0.95	0.95
		Deciles	1.25	1.09	**1.03**	**0.99**	**0.98**	0.96	0.96	0.95
		Ventiles	1.24	1.09	**1.03**	**1.00**	**0.98**	**0.97**	**0.97**	0.96
	Dagum	Quintiles	0.90	1.10	1.04	**1.00**	**0.97**	0.95	0.94	0.94
		Deciles	1.22	1.15	1.06	**1.01**	**0.98**	0.96	0.95	0.94
		Ventiles	1.21	1.13	1.06	**1.01**	**0.98**	**0.97**	0.96	0.96
	Generalized Beta 2	Quintiles	0.64	1.11	1.08	1.04	**0.99**	**0.97**	0.96	0.96
		Deciles	1.10	1.15	1.11	1.05	**0.99**	0.96	0.95	0.95
		Ventiles	1.10	1.13	1.09	1.04	**0.99**	0.96	0.96	0.95
Panel B. Lorenz curve parametric estimation										
GQ	Log-normal	Quintiles	1.11	**1.01**	**0.97**	0.96	**0.97**	**0.98**	**0.98**	**0.99**
		Deciles	1.04	**1.01**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		Ventiles	1.04	**1.01**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	Dagum	Quintiles	1.05	**1.03**	**1.00**	**0.99**	**0.99**	**1.00**	**1.00**	**1.00**
		Deciles	1.06	1.04	**1.01**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		Ventiles	1.06	1.05	**1.02**	**1.01**	**1.00**	**1.00**	**1.00**	**1.00**
	Generalized Beta 2	Quintiles	0.41	**0.99**	1.16	1.14	1.08	1.04	**1.01**	**1.00**
		Deciles	0.43	**1.01**	1.17	1.15	1.08	1.04	**1.01**	**1.00**
		Ventiles	0.45	1.03	1.18	1.15	1.09	1.04	**1.01**	**1.00**
Beta	Log-normal	Quintiles	1.11	**1.01**	**0.97**	0.96	**0.97**	**0.98**	**0.98**	**0.99**
		Deciles	1.20	**1.01**	0.96	0.95	0.96	0.96	**0.97**	**0.98**
		Ventiles	1.27	**1.02**	0.96	0.95	0.95	0.96	0.96	**0.97**
	Dagum	Quintiles	0.93	**0.98**	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**	**1.00**
		Deciles	0.95	**1.00**	**0.99**	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**
		Ventiles	**0.98**	**1.01**	**0.99**	**0.98**	**0.98**	**0.98**	**0.98**	**0.99**
	Generalized Beta 2	Quintiles	0.89	0.96	**1.00**	**1.01**	**1.00**	**1.00**	**1.01**	**1.01**
		Deciles	0.87	**0.97**	**1.01**	**1.01**	**1.00**	**1.00**	**1.00**	**1.01**
		Ventiles	0.85	**0.98**	**1.02**	**1.01**	**1.00**	**1.00**	**1.00**	**1.01**

The figures represent the ratio between the grouped data KDE-based poverty headcount ratio and the population value (Panel A); and between the parametric estimate of the headcount ratio and the population value (Panel B). Poverty headcount ratios are averaged over 2,000 replications. The poverty lines are expressed as multiples of the population median, ranging from a quarter of the median (0.25) to twice (2) the median. Figures in boldface represent cases in which the population value lies inside the 95 percent confidence interval around the average estimate.

The main lesson we draw from the Monte Carlo study is that it is difficult to summarize the KDE-generated biases in poverty indicators with statements that hold true across distributions or parameters. The simulations show that the errors

depend on the bandwidth, kernel, and the size of the dataset—all of which are choice variables for the researcher. Moreover, they depend on the unknown data generating process. Biases associated with the parametric estimation of the Lorenz curve similarly vary with the underlying distribution, but are consistently of lower magnitude than when nonparametric estimation is used.

In the case of the unimodal distributions considered here, we have uncovered several robust patterns. The KDE-based HCR tends to be overestimated for lower poverty lines, roughly well estimated around the population median, and underestimated for higher poverty lines. It may be more difficult to find consistent patterns for distributions with multiple modes. Minoiu and Reddy [29] analyze a multimodal distribution and find that the positioning of the poverty line relative to the modes and the extent of smoothing achieved by a kernel-bandwidth combination play a more important role in determining the sign and size of the errors than with unimodal distributions. Similarly, Minoiu and Reddy [30] document a worse performance for the parametric approach when employed on multi-peaked rather than unimodal distributions.

Given the variety of income distributions likely present in real-world data, these results may not apply more generally. Nonetheless, they lead us to caution against the use of nonparametric density estimators for grouped data suspected to come from unimodal distributions, for which parametric approaches already in widespread use fare better. For grouped data likely to come from multimodal distributions, a thorough sensitivity analysis to assess the effect of alternate estimation methods and parameter choices is desirable before drawing firm conclusions.

5 Sensitivity analysis of global poverty estimates

We wrap up our assessment of the performance of grouped-data KDE methods for poverty analysis by undertaking a sensitivity analysis of global poverty estimates to changes in the bandwidth. Consumption shares by decile for 65 developing countries covering 70% of the world's population in 1995 were assembled from the Povcalnet website [41] for the years 1995 and 2005 (or closest available year).[19] These were scaled by total household consumption from the Penn World Tables Mark 6.3 [15] to obtain decile means. We applied KDE to each country's decile means and aggregated the estimated individual country distributions into a global distribution of consumption. Global poverty is measured with the HCR and the absolute headcount (AHC) for five international poverty lines that range between \$1/day and \$2.5/day (all expressed in 2005 PPPs).[20]

[19] See Table A7 in the Supplementary Appendix for the list of countries and available distributional data.

[20] See [6] for definitions of the poverty lines.

The sensitivity analysis was undertaken for the Quartic kernel and a range of bandwidths.[21] In addition to the data-driven bandwidths discussed in Section 3.2, we also consider here a "hybrid" bandwidth that has been used in previous studies of the global income distribution. This bandwidth is computed as is Silverman's S4 bandwidth, which assumes an underlying Gaussian distribution for log-incomes, but is kept constant across countries. Following [35], the hybrid bandwidth for deciles is computed assuming a standard deviation of 0.6—the value for China. Note that the hybrid bandwidth is 'optimal' for China, but may not be optimal (in a statistical sense) for other countries unless their distribution is close to China's.

The results are summarized in Table 6, where the bandwidths are arranged from left to right in ascending order of smoothing (i.e., from the smallest to the largest). There are significant variations in estimated poverty levels across bandwidths (Panels A–B). According to the standard $1.25/day poverty line, in 1995 the estimated global HCR ranges between 7.8% and 12.6%, while for 2005 it varies from 2.2% to 6.1%. As shown in the last two columns, which report the range of variation in the HCR and AHC across bandwidths, the degree of sensitivity to the choice of bandwidth is highest for the lowest poverty line. The extent of variation across bandwidths is significant for the AHC as well. In 1995, the estimated $1/day AHC ranges from 174 to 366 million (a factor of two), while in 2005 the estimated AHC for the same poverty line varies between 62 and 170 million (a factor of almost three).

Turning to the trend of global poverty, we find that the estimated number of people lifted from '$1/day poverty' over 1995–2005 is 112 million based on the hybrid bandwidth and between 150 and 196 for other bandwidths (Panels C–E). The range of variation across bandwidths in the estimated reduction in the AHC is from 80% (84 million individuals) for the lowest poverty line to 30% (116.3 million individuals) for the highest poverty line. Importantly, all estimates are consistent with a reduction in world poverty over 1995–2005 regardless of the bandwidth or poverty line considered.

Our sensitivity analysis illustrates two points. The first is that the range of variation in estimated poverty levels across all bandwidths considered is sizable. For this reason, undertaking sensitivity analyses to parameter choices such as the bandwidth is an important step when nonparametrically estimating the income distribution from grouped data. The second is that the range of variation is significantly lower among data-driven bandwidths. This underlines the potential drawbacks of non data-driven bandwidths, such as the hybrid bandwidth, which can be appealing due to their simplicity but may smooth "too much" or "too little" on some datasets. In our context, the hybrid bandwidth is optimal for the Chinese dataset but is likely unfit for the other countries. It systematically produces the lowest poverty estimates and is thus something of an outlier among the estimators.

Finally, we note that the global poverty estimates presented in this section should not be interpreted as authoritative due to numerous deficiencies in the methods employed to produce them (discussed, for instance, in [34]). Although we do not know how close these estimates are to the 'true' level of global poverty, this sensitivity exercise can help gauge the uncertainties surrounding KDE-based global poverty figures.

[21] The results are robust to using the Gaussian and Epanechnikov kernels (see Tables A8–A9 in the Supplementary Appendix).

Table 6 Estimating global poverty using KDE on grouped data (1995–2005)

Bandwidth	Hybrid	S3	S4	Sheather-Jones	DPI-1	DPI-2	Over-smoothed	Ratio between highest and lowest value in columns 1–7	Pp diff. between highest and lowest estimate in columns 1–7
Poverty line	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1995	Panel A. Poverty headcount ratio (% poor)								
$1.00	4.1	6.2	7.1	7.9	8.1	8.5	8.7	2.1	4.6
$1.25	7.8	10.4	11.0	12.0	12.3	12.3	12.6	1.6	4.7
$1.45	11.6	13.9	14.4	15.7	15.5	16.1	16.3	1.4	4.8
$2.00	21.1	23.6	24.2	24.7	25.1	25.2	25.2	1.2	4.0
$2.50	30.6	32.0	32.2	32.8	32.7	33.3	33.3	1.1	2.7
2005									
$1.00	1.3	2.4	2.8	3.5	3.2	3.5	3.6	2.8	2.3
$1.25	2.2	4.2	5.0	5.7	5.5	5.8	6.1	2.7	3.9
$1.45	3.6	6.2	7.0	8.3	7.9	8.1	8.1	2.3	4.6
$2.00	10.6	13.0	13.8	14.9	14.7	15.2	15.1	1.4	4.6
$2.50	17.1	19.7	20.8	21.5	21.1	22.0	21.9	1.3	4.9
1995	Panel B. Poverty headcount (millions)								(millions)
$1.00	174	261	297	330	339	357	366	2.1	192
$1.25	330	438	462	506	517	517	529	1.6	199
$1.45	485	586	603	658	650	675	686	1.4	201
$2.00	887	990	1016	1038	1056	1057	1057	1.2	170
$2.50	1285	1346	1354	1377	1373	1399	1397	1.1	115
2005									(millions)
$1.00	62	111	132	164	150	162	170	2.8	108
$1.25	106	196	234	270	258	271	288	2.7	182
$1.45	170	290	327	387	369	379	379	2.3	217
$2.00	497	609	648	697	689	711	709	1.4	214
$2.50	800	925	973	1009	988	1030	1027	1.3	230

Table 6 (continued)

Bandwidth	Hybrid	S3	S4	Sheather-Jones	DPI-1	DPI-2	Over-smoothed	Ratio between highest and lowest value in columns 1–7	Pp diff. between highest and lowest estimate in columns 1–7
Poverty line	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1995–2005	Panel C. Reduction in the poverty headcount ratio (pp)								
$1.00	2.8	3.9	4.3	4.4	4.9	5.0	5.1	1.8	2.3
$1.25	5.6	6.2	6.0	6.3	6.8	6.5	6.5	1.2	1.2
$1.45	7.9	7.8	7.4	7.4	7.6	8.0	8.3	1.1	0.9
$2.00	10.5	10.6	10.4	9.9	10.4	10.0	10.0	1.1	0.7
$2.50	13.5	12.3	11.5	11.3	11.6	11.3	11.4	1.2	2.3
1995–2005	Panel D. Reduction in the poverty headcount ratio (%)								
$1.00	68	62	60	56	60	59	58	1.2	12.6
$1.25	71	60	55	52	55	53	51	1.4	20.1
$1.45	69	56	51	47	49	50	51	1.4	21.3
$2.00	50	45	43	40	42	40	40	1.3	10.1
$2.50	44	38	36	34	36	34	34	1.3	10.1
1995–2005	Panel E. Reduction in the poverty headcount (millions)								(millions)
$1.00	−112	−150	−165	−166	−189	−195	−196	1.8	84.0
$1.25	−224	−242	−228	−236	−259	−246	−241	1.2	35.1
$1.45	−315	−295	−276	−271	−281	−295	−307	1.2	44.0
$2.00	−390	−381	−368	−341	−367	−346	−348	1.1	48.7
$2.50	−484	−421	−381	−368	−386	−369	−370	1.3	116.3

Estimates are based on decile means from the World Bank's Povcalnet database; and KDE with the Quartic kernel. See Section 5 for a description of the data and the definition of the hybrid bandwidth.

6 Concluding remarks

Kernel density estimation has gained popularity in recent years as an attractive alternative to parametric methods for estimating the income distribution. Its advantage is that it does not require (potentially restrictive) distributional assumptions concerning the underlying density function. However, this technique has been used primarily to analyze the global income distribution from grouped data rather than individual records. In this paper, we assessed the performance of kernel density estimation in recovering features of the income distribution from grouped data, focusing on poverty measures. We also examined its performance relative to a widely-used parametric approach which consists of estimating the Lorenz curve from grouped data using two alternate functional forms. Our goal has been to document the sign and size of biases associated with the application of nonparametric methods in poverty analysis, and to raise awareness of their potential caveats.

We found that kernel density estimation gives rise to nontrivial biases that depend on the bandwidth, kernel, poverty indicator, poverty line, size of the dataset, and data generating process. Using Monte Carlo simulations on data drawn from several unimodal distributions, we showed that the average income of the poorest individuals is generally understated by the technique, while that of the richest individuals is overstated: the poor seem to be poorer, and the rich appear richer. This translates into a systematic overestimation of the poverty headcount ratio for lower poverty lines and its underestimation for higher poverty lines. Poverty estimates based on the nonparametric method are reliable only when the poverty line is close to the population median. The further is the poverty line from the population median, the more the poverty headcount rate tends to be misestimated. We also undertook a sensitivity analysis of global poverty estimates to changes in the bandwidth, a key parameter in kernel density estimation, and found that the choice of bandwidth had a marked impact on global poverty statistics.

Taken together, our results suggest that the advantage of nonparametric estimation—its freedom from distributional assumptions—comes at a cost. We note, however, that our study does not represent an indictment of either kernel density estimation nor of the use of grouped data in income distribution analysis as such. The weakness of the kernel density estimator in this particular setting originates from the *combination* of the estimator and the nature of the data, and is also influenced by the specific distribution and poverty line. Kernel density estimation works well on large datasets (i.e., when individual records are available). Similarly, grouped data can be useful in conjunction with parametric methods for the Lorenz curve or income density. The methods for the parametric estimation of the Lorenz curve considered here consistently yield empirical biases of lower, often negligible, magnitude compared to their nonparametric counterparts. On this basis, we view them as the preferred approach. Nevertheless, there is some evidence that each of the parametric approaches considered works best on a different income distribution. Whether the applied researcher has a preference for parametric or nonparametric methods, our results underscore the need for a thorough sensitivity analysis to estimation method or to parameter assumptions whenever grouped data are involved.

Acknowledgements We are grateful for financial support from the United Nations Development Programme's Bureau of Development Policy. We thank the Editor-in-Chief, the Associate Editor, and two anonymous referees for their helpful suggestions, as well as Sudhir Anand, Andrew

Berg, Judith Clarke, Cristian Pop-Eleches, Ronald Findlay, Marc Henry, Tümer Kapan, Stephan Klasen, Branko Milanovic, Paul Segal, Joseph Stiglitz, Eric Verhoogen, and seminar participants at Columbia University, Cornell University (NEUDC 2006), Izmir University of Economics, DIW Berlin (ECINEQ 2007), Central European University (EEA-ESEM 2007), University of British Columbia (CEA 2008), and the IMF Research Department for discussions and comments. Catherine Choi, Pedro Ledesma III, and Anjuli Muttoo are acknowledged for their research assistance and Nicolas Kruchten, David Dekoning, and Sergey Kivalov for their help with software development. The views expressed in this paper are those of the authors and do not necessarily reflect those of the IMF or IMF policy. All remaining errors are our own.

References

1. Abdelkrim, A., Duclos, J.-Y.: DASP: distributive analysis stata package. PEP, World Bank, UNDP and Université Laval. Available online at http://dasp.ecn.ulaval.ca/ (2007)
2. Ackland, R., Dowrick, S., Freyens, B.: Measuring global poverty: why PPP methods matter. The Australian Demographic and Social Research Initiative Unpublished Manuscript, Australian National University, Canberra (2008)
3. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J, Rogers, W.H., Tukey, J.W.: Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton (1972)
4. Bandourian, R., McDonald, J.B., Turley, R.S.: A comparison of parametric models of income distribution across countries and over time. Rev. Estad. **55**, 164–165 (2003)
5. Bourguignon, F., Morrisson, C.: Inequality among world citizens: 1820–1992. Am. Econ. Rev. **92**(4), 727–744 (2002)
6. Chen, S., Ravallion, M.: The developing world is poorer than we thought, but no less successful in the fight against poverty. Q. J. Econ. **125**(4), 1577–1625 (2010)
7. Chen, S., Datt, G., Ravallion, M.: POVCAL, A Program for Calculating Poverty Measures for Grouped Data. World Bank Poverty and Human Resource Division, The World Bank Group, Washington (2001)
8. Chotikapanich, D., Griffiths, W.E., Rao, D.S.P.: Estimating and combining national income distributions using limited data. J. Bus. Econ. Stat. **25**, 97–109 (2007)
9. Datt, G.: Computational tools for poverty measurement and analysis. IFPRI Food Consumption and Nutrition Division Discussion Paper No. 50, International Food Policy Research Institute, Washington (1998)
10. Deheuvels, P.: Estimation nonparamétrique de la densité par histogrammes généralisés. Rev. Statist. Appl. **25**, 5–42 (1977)
11. Diaz-Emparanza, I.: Selecting the number of replications in a simulation study. Unpublished Manuscript, University of the Basque Country Department of Economics and Statistics. Available online at http://129.3.20.41/eps/em/papers/9612/9612006.pdf (1996)
12. Diaz-Emparanza, I.: Is a small Monte Carlo analysis a good analysis? Checking the size, power, and consistency of a simulation-based test. Stat. Pap. **43**(4), 567–577 (2002)
13. Domma, F., Perri, P.: Some developments on the log-dagum distribution. Stat. Methods Appl. **18**(2), 205–220 (2009)
14. Fuentes, R.: Poverty, pro-poor growth and simulated inequality reduction. UNDP Human Development Report Office Occasional Paper No. 11, United Nations Development Programme, New York (2005)
15. Heston, A., Summers, R., Aten, B.: Penn world table version 6.3, center for international comparisons of production, income, and prices at the University of Pennsylvania. Available online at http://pwt.econ.upenn.edu/php_site/pwt_index.php (2009)
16. Jann, B.: Univariate kernel density estimation. Statistical Software Component No. S456410, Department of Economics, Boston College, Boston (2007)
17. Jones, M.C., Marron, J.S., Sheather, J.S.: A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. **91**, 401–407 (1996)
18. Kakwani, N.C.: On a class of poverty measures. Econometrica **48**(2), 437–446 (1980)
19. Kakwani, N.C.: Functional forms for estimating the lorenz curve: a reply. Econometrica **44**(1), 1063–1064 (1980)
20. Kleiber, C.: A guide to the dagum distributions. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves, Economic Studies in Inequality, Social Exclusion, and Well-Being. Springer New York (2008)

21. Li, Q., Racine, J.S.: Nonparametric Econometrics: Theory and Practice. Princeton University Press (2006)
22. Lopez, J.H., Serven, L.: A normal relationship? Poverty, growth and inequality. World Bank Policy Research Working Paper No. 3814, The World Bank, Washington (2006)
23. Marron, J.S., Nolan, D.: Canonical kernels for density estimation. Stat. Probab. Lett. **7**, 195–199 (1988)
24. Marron, J.S., Ruppert, D.: Transformations to reduce boundary bias in kernel density estimation. J. R. Stat. Soc., Ser. B (Methodological) **56**(4), 653–671 (1994)
25. McDonald, J.B.: Some generalized functions for the size distribution of income. Econometrica **52**, 647–663 (1984)
26. Milanovic, B.: Global inequality recalculated: the effect of new 2005 PPP estimates on global inequality. J. Econ. Inequal. (forthcoming)
27. Milanovic, B.: True world income distribution, 1988 and 1993: first calculation based on household surveys alone. Econ. J. **112**, 51–92 (2002)
28. Milanovic, B.: Worlds Apart: Measuring International and Global Inequality. Princeton University Press (2005)
29. Minoiu, C., Reddy, S.: Kernel density estimation from grouped data: the case of poverty assessment. IMF Working Paper No. 183, International Monetary Fund, Washington (2008)
30. Minoiu, C., Reddy S.: Estimating poverty and inequality from grouped data: how well do parametric methods perform? J. Income Distrib. **18**(2), 160–179 (2009)
31. Mosteller, F.: On some useful inefficient statistics. Ann. Math. Stat. **17**(4), 377–408 (1946)
32. Ott, R.L., Longnecker, M.: An Introduction to Statistical Methods and Data Analysis, 6th Edition. Duxbury Press (2008)
33. Pinkovskiy, M., Sala-i-Martin, X.: Parametric distributions of the world distribution of income. NBER Working Paper No. 15433, The National Bureau for Economic Research, Cambridge (2009)
34. Reddy, S., Pogge, T.: How not to count the poor. In: Anand, S., Segal P., Stiglitz, J. (eds.) Debates on the Measurement of Global Poverty, Oxford University Press (2010)
35. Sala-i-Martin, X.: The world distribution of income: falling poverty and … convergence, period. Q. J. Econ. **121**(2), 351–397 (2006)
36. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B (Methodological), **53**(3), 683–690 (1991)
37. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, vol. 26, Chapman & Hall/CRC (1986)
38. Stigler, S.M.: Linear functions of order statistics with smooth weight functions. Ann. Stat. **2**(4), 676–693 (1974)
39. Villasenor, J.A., Arnold, B.C.: Elliptical lorenz curves. J. Econom. **40**, 327–338 (1989)
40. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)
41. World Bank: Povcalnet online database. The World Bank Group, Washington. Available online at http://go.worldbank.org/7X6J3S7K90 (2010)
42. Wu, X., Perloff J.: GMM estimation of a maximum entropy distribution with interval data. J. Econom. **138**(2), 532–546 (2007)
43. Yatchew, A.: Nonparametric regression techniques in economics. J. Econ. Lit. **36**(2), 669–721 (1998)
44. Zhang, Y., Wan, G.: Globalization and the urban poor in China. UNU-WIDER Working Paper No. 2006/42, United Nations University, World Institute for Development Economics Research, Helsinki (2006)