WP/08/183

IMF Working Paper

Kernel Density Estimation Based on Grouped Data: The Case of Poverty Assessment

Camelia Minoiu and Sanjay G. Reddy

INTERNATIONAL MONETARY FUND

 WP/08/183

IMF Working Paper

African Department

Kernel Density Estimation Based on Grouped Data: The Case of Poverty Assessment

Prepared by Camelia Minoiu and Sanjay G. Reddy[1]

Authorized for distribution by Arend Kouwenaar

July 2008

Abstract

We analyze the performance of kernel density methods applied to grouped data to estimate poverty (as applied in Sala-i-Martin, 2006, *QJE*). Using Monte Carlo simulations and household surveys, we find that the technique gives rise to biases in poverty estimates, the sign and magnitude of which vary with the bandwidth, the kernel, the number of datapoints, and across poverty lines. Depending on the chosen bandwidth, the $1/day poverty rate in 2000 varies by a factor of 1.8, while the $2/day headcount in 2000 varies by 287 million people. Our findings challenge the validity and robustness of poverty estimates derived through kernel density estimation on grouped data.

JEL Classification Numbers: I32, D31, C14, C15

Keywords: kernel density estimation, income distribution, grouped data, poverty

Author's E-Mail Address: CMinoiu@imf.org, sr793@columbia.edu

---

[1] IMF African Department and Department of Economics, Barnard College, Columbia University, respectively. This project was supported financially by the United Nations Development Programme's Bureau of Development Policy (with the assistance of Terry McKinley). The authors are grateful to Sudhir Anand, Abdelkrim Araar, Andrew Berg, Judith Clarke, Arend Kouwenaar, Cristian Pop-Eleches, Ronald Findlay, Marc Henry, Patrick Imam, 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), the University of British Columbia (CEA 2008), and the IMF Research Department brown bag seminar for their helpful comments.

Contents

Appendix Figures

Appendix Tables

I. MOTIVATION

Several recent studies have employed nonparametric smoothing techniques, and in particular kernel density estimation (henceforth, 'KDE') on grouped data to obtain poverty estimates (Sala-i-Martin 2002a, 2002b, 2004, 2006; Ackland, Dowrick, and Freyens, forthcoming; Fuentes, 2005).[1] World poverty and inequality assessments—especially over longer time-horizons—require the use of grouped data (usually expressed as income averages for a small number of population quantiles) because representative household surveys are not available, or are difficult to obtain or analyze (Shorrocks and Wan, 2008). However, the accuracy of poverty estimates and visual representations of income distributions extrapolated from this informationally limited data structure crucially depends on the statistical method employed.

The goals of this study are twofold. First, we assess the appropriateness of kernel density estimation methods on grouped data for poverty analysis. Biases in poverty estimates associated with the method are identified for a wide range of poverty indicators, poverty lines, parameters (e.g., bandwidths and kernels), and income distributions. Second, we analyze the robustness of KDE-based global poverty estimates to the choice of bandwidth, and provide a framework for the interpretation of these estimates. Our analysis is also relevant to researchers undertaking poverty and social impact analyses (PSIA), especially those aiming to cover multiple years or groups of countries. Full household surveys are often unavailable for low-income or post-conflict countries, but grouped data may have been published, making the use of smoothing techniques attractive. Assessing the distributional effects of PRGF program measures may, for instance, require the use of such data.[2]

This study is a response to the influential work of Sala-i-Martin (2002a, 2002b, 2004, 2006) who uses KDE on grouped data to estimate national, regional, and global poverty. The author finds that there have been substantial reductions in world income poverty (according to all indicators and poverty lines considered) between 1970 and 2000. We demonstrate that Sala-i-Martin's figures lack robustness to the choice of underlying parameters. In particular, a number of optimal bandwidth values[3] are consistent with an estimated share of '$1/day poor' in 2000 which is higher than Sala-i-Martin's estimate by a factor of 1.8. Similarly, alternative

---

[1] Other studies (e.g., Berry et al (1983), Grosh and Nafziger (1986), Korzeniewick and Morran (1997), Bhalla (2002), Bourguignon and Morrison (2002), Milanovic (2002, 2005)) have also used grouped data representing average incomes of population quantiles to estimate national, regional, and world inequality. Grouped data has been used to illustrate of the shape of regional and world income distributions, too.

[2] Income distributions estimated from grouped data may be integrated into partial and general equilibrium models to assess the welfare impact of policies (see, e.g.,Essama-Nssah *et al*, 2002; Essama-Nssah, 2005a, 2005b; Coady, 2005).

[3] Throughout the paper, the term 'optimal' is used to describe smoothing parameter values which maximize the approximate integrated mean square error. This optimality criterion refers to the 'global' goodness-of-fit of the estimated density function. It should be noted that alternative optimality criteria may be defined for the problem at hand, for example, a high goodness-of-fit of the estimated income density upto a given poverty line or simply a good estimate of the poverty headcount ratio or other poverty measure. Different bandwidth values may result from the application of alternative optimality criteria.

bandwidths gives rise $2/day poverty headcounts in 2000 higher than Sala-i-Martin's estimate by 287 million people. This suggests that the authors may have underestimated the $2/day headcount by as much as 50 percent due to this methodological choice alone. The magnitude of the possible errors associated with KDE-based global poverty analysis, of which these are two examples, gives rise to serious concern about the validity and robustness of the technique in poverty analysis.

The grouped data we consider consists of income averages for a small number of population quantiles (usually five). Since analytical derivations of the properties of the KDE estimator in small samples are prohibitively difficult or impossible, we undertake Monte Carlo simulations for a range of plausible income distributions. The following distributions are considered: Log-normal (two parameters), Dagum (three parameters), Generalized Beta II (four parameters), and a notional multimodal distribution. We also use three nationally representative household surveys (Nicaragua, Tanzania and Vietnam) to compare KDE-based poverty estimates obtained from grouped data with those obtained directly from unit data. Finally, we assess the performance of KDE in *global* poverty analysis, using grouped data for a large number of developing countries.

There are reasons to believe that the application of KDE in this data environment to estimate poverty may give rise to biases. However, the size of the possible biases (for distinct poverty indicators and poverty lines, and for various income distributions) is unknown *ex ante* and requires a study of this kind. The data structure on which we focus—five income averages for five population quantiles—is informationally poorer than a large sample drawn from the underlying distribution, rendering nonparametric density estimation methods inappropriate. However, a small number of *average* incomes is a richer source of information about the underlying distribution than would be a small random sample from that distribution. We find that poverty is misestimated in a majority of cases, but occasionally the biases at different income levels cancel out so that good estimates are obtained. The order of magnitude of the biases in poverty headcount ratio estimates identified in this study reaches 6-7 percentage points (for the unimodal distributions considered) and 10-11 percentage points (for the multimodal distribution considered). Furthermore, the biases in estimates of the Foster-Green-Thorbecke (FGT) poverty measures generally increase with the degree of distributional sensitivity.

Why have researchers used KDE to estimate poverty from grouped data? First, the method draws its appeal from the fact that unlike parametric approaches, it does not require prior beliefs about the functional form of the underlying distribution. Second, it is convenient to use because it reproduces the entire income distribution from a manageable amount of data. It is thus particularly useful when the analysis is regional or global in scope: not only are household surveys not available for numerous country-years, but their analysis could be prohibitive in terms of time and manpower. Third, unit data from many nationally representative household surveys (including for China and India) are not readily publicly available.[4] In this environment of data paucity marked by an increasing interest in estimates

[4] The Chinese State Statistical Bureau only publishes grouped data from underlying rural and urban household surveys in its China Statistical Yearbook. Similarly, grouped data from Indian National Statistical Surveys are

(continued…)

of long-term trends in global poverty, it is essential to determine whether the statistical techniques employed are indeed reliable.

Kernel density estimation is one of *two* methods that have been used extensively in poverty analysis from grouped data. The alternative approach is parametric estimation of the Lorenz curve (applied in studies including Yotopoulos, 1989; Chen and Ravallion, 2002, 2004, 2007; Bhalla, 2002; Pritchett, 2006; and Kakwani and Son, 2006).[5] Many Lorenz parametric forms have been proposed, of which two are easily implemented using the computational tools POVCAL and SimSIP developed by the World Bank. Although we do not explicitly compare KDE-based estimates with their parametric analogues, it is noteworthy that the Lorenz curve parameterizations embodied in POVCAL and SimSIP perform well on grouped data for unimodal distributions, but less so in the case of multimodal distributions (Minoiu and Reddy, forthcoming).

In poverty analysis, KDE methods have been undertaken on datasets as small as five income averages corresponding to five population quantiles per country and per year. For example, Sala-i-Martin (2002a, 2002b, 2004, 2006) uses five such datapoints for each of 138 countries to fit income distributions and estimate the long-term trend in income poverty. The author concludes that substantial reductions in world poverty have been recorded over the past three decades. In particular, after applying KDE to grouped data, the author finds that the share of people with an income level below $1.50 per day in the world's population has fallen from 20.2 percent to 7 percent between 1970 and 2000. The author proposes two methods for constructing a world income distribution from individual country distributions. The first method (described as the *kernel of quintiles* method) consists of constructing a dataset in which each person's income level is the average income of the national population quintile to which that person belongs. Subsequently, kernel density estimation is applied to this the dataset. The second method (described as the *kernel of kernels* method) consists of first estimating each country's income density from quintile means, and integrating the individual country densities into a population-weighted world income density. We conduct our analysis in a manner directly comparable to the *kernel of kernels* method proposed by Sala-i-Martin, but the two approaches yield similar results.

Sala-i-Martin's KDE-based poverty estimates have given rise to heated debates on the extent and trend of global poverty.[6] Despite the uncertainties concerning the applicability of KDE

---

readily but India's unit data can only be obtained from the National Sample Survey Organization (NSSO) provided that the research is deemed relevant to India's national development and planning by its statistical authorities. Such unit data is moreover only available for recent years. For this study, a formal request for unit data on consumption was submitted to the NSSO, but it was rejected on the grounds that the project was not relevant to national development and planning.

[5] Maximum entropy density estimation (for densities from the exponential family) has recently been proposed by Wu and Perloff (2007) as an alternative technique for poverty analysis on grouped data. An application to Chinese data demonstrated that the estimates are reliable (Wu and Perloff, 2005). However, the method has not been the subject of systematic assessment.

[6] See, for example, articles in The Economist ("More or less equal?", March 11, 2004 and "Pessimistic on poverty?", April 7, 2004), NBER Digest ("Economic growth is reducing global poverty", October 2002), The

(continued…)

methods to grouped data, a number of studies have subsequently used his methodology. For example, Ackland, Dowrick, and Freyens (2007) use *kernel of kernels* method on quintile means to investigate the sensitivity of global poverty estimates to alternative purchasing power parity (PPP) conversion factors. However, their analysis is subject to the biases inherent in the 'KDE on grouped data' approach, which render the level or trend estimates of poverty produced difficult to interpret. A similar application of KDE to grouped data is presented by Fuentes (2005), who uses an unspecified number of income averages to estimate inequality and poverty in several countries.[7]

The remainder of this paper is organized as follows: in the next two sections, we discuss the nature of the data and the bias of the kernel density estimator on grouped data. Section III contains a description of our methods. Section IV discusses the results of the Monte Carlo analysis. Section V presents findings from a comparison of poverty estimates from household surveys with those from KDE on grouped data for three countries with different levels of poverty. In Section VI, we present a sensitivity analysis of KDE-based global poverty estimates to changes in the bandwidth. Conclusions are drawn in Section VII.

II. THE DATA STRUCTURE AND THE BIAS OF THE ESTIMATOR

The situation facing a researcher who seeks to estimate poverty from grouped data can be described as follows. Information on a variable of interest (e.g., income, consumption, or total wealth) is collected through a nationally representative household survey. While the survey is not available in its entirety, the researcher possesses average incomes of several population quantiles.[8] One way of representing the data is as a collection of linear functions of order statistics: the order statistics represent the income levels of individuals in the nationally representative household survey arranged in ascending order. The averages of incomes of population quantiles are linear functions of order statistics. These "systematic statistics" (a term coined by Mosteller in 1946) represent the sole source of information from which the researcher aims to recover features of the income distribution.

The process of grouping the data can be described as follows: income information for a *large* number of individuals is transformed into *summary* income information for a *small* number of equally-sized *groups* of individuals after those individuals' income levels have been *arranged in ascending order*. The unit data from the survey represents independently and

---

Financial Times ("Location, location, location", September 24, 2002), The National Center for Policy Analysis Daily Policy Digest ("World poverty rate has fallen", June 11, 2002), and The New York Times ("Good news about poverty", November 27, 2004).

[7] Other studies that do not make use of distributional information within population groups, but still employ KDE to estimate or illustrate income distributions, include Dhonghe (2005), Aziz and Duenwald (2001), Milanovic (2002, 2005), Bourguignon and Morrison (2002), Bianchi (1997), Jones (2002), Quah (1996, 1997), Pittau (2005), and Pittau and Zelli (2006).

[8] Alternatively, income shares are available from the survey. An estimate of total income (drawn from the national accounts or surveys) is then used to scale them and obtain average incomes of several population quantiles.

identically distributed draws from the unknown income distribution. The process of ordering the independent and identically distributed draws from the underlying distribution, and of collecting them into groups, generates a complex correlation structure among the order statistics. The correlation structure would be inconsequential for the properties of the kernel density estimator if all the underlying observations were available to the researcher. However, this is not the case. The order statistics contain less information than the original sample. Nevertheless, the averages retain important information about the underlying distribution due to the *ordering* of the original observations.[9]

Each quantile mean available for KDE is a trimmed mean obtained by discarding a number of order statistics. Four of the quintile means, for example, are asymmetrically trimmed means, whereas the central one (corresponding to the middle twenty percent of the population) is a symmetrically trimmed mean. Symmetrically trimmed sums are robust estimators of location (to heavy-tailed distributions and outliers). Furthermore, if the data are drawn from a symmetric distribution, they are unbiased estimators for the mean of that distribution.[10] Stigler (1973, 1974) and Mason (1981) have shown that trimmed means are asymptotically normally distributed under mild conditions on the weighting function for the ordered observations and an arbitrary data generating process for the unordered observations.[11] A small number of quantile means are therefore informationally richer than a small sample from the underlying distribution (in particular because they carry more precise information about the location of underlying order statistics along the support), but informationally poorer than a large sample from the underlying distribution.

It should be noted that nonparametric approaches to estimating the density from small datasets (comprised of draws from the underlying density or, as is the case here, quantile means), are also inappropriate due to the very nature and purpose of nonparametric statistics. The statistical literature encourages the use nonparametric estimators in "exploratory data analysis, as a confirmatory tool, or as a supplement to the standard parametric fare" (Yatchew, 1998, p. 672). The purpose of nonparametric techniques is to provide means of uncovering patterns in the data using information from a wealth of (nearby) observations. Yatchew (1998, p. 715) argues that "interpolation is only deemed reliable among close neighbour[ing] observations, and extrapolation outside the observed domain is considered entirely speculative".

---

[9] As evidenced by the early literature following Mosteller (1946) which focused on robust estimation of location and scale parameters of the underlying distributions from order statistics.

[10] This is relevant in the context of income distributions, since Log-transformed incomes are distributed normally (hence, symmetrically) if incomes are distributed Log-normally.

[11] A necessary and sufficient condition for this result to hold is that the sample is trimmed at sample percentiles such that the corresponding population percentiles are uniquely defined (Stigler, 1973). Similarly, Moore (1968) and Siddiqui and Butler (1969) have shown that linear functions of order statistics are asymptotically normally distributed (under the condition that the weighting function which gives rise to the linear functions of order statistics is differentiable, its first derivative is continuous and of bounded variation except at finitely many jumps. This condition is trivially fulfilled by the weighting function giving rise to the quantile means).

With these considerations in mind, we derived the bias of the kernel density estimator of an the unknown density from grouped data data[12], given below:

$$Bias(\hat{f}(x)) = \frac{1}{J}\sum_j g_j(x) + \frac{h^2}{2J}\sum_j g_j''(x)\int t_j^2 k(t_j)dt_j + \ldots - f(x) \tag{1}$$

where *j=1,...,J* represents the number of grouped datapoints, *h* represents the bandwidth, k(.) is the weighting function (kernel), $\int t^2 k(t)dt$ is therefore a constant depending on the weighting function, and $g_j(\cdot)$ is the density probability function of the jth quantile mean. Higher order terms in *h* arising from a Taylor approximation have been omitted for simplicity. For purposes of comparison, the bias of the 'standard', survey-based kernel density estimator (Silverman, 1986) is given by:

$$Bias(\hat{f}_S(x)) = \frac{h^2}{2} f''(x)\int t^2 k(t)dt + \ldots \tag{2}$$

where the higher order terms in *h* have also been suppressed. As expected, the grouped data-based bias is itself a function of the unknown probability density functions associated with the quantile means. Letting

$$\tfrac{1}{J}\textstyle\sum_j g_j(x) = v(x) \tag{3}$$

then the grouped data-based estimator will have the same bias as the survey-based estimator if *v(x) = f(x)*. As the number of observations underlying each trimmed mean increases ($J \to \infty$), it is known that $g_j(\cdot)$ becomes a normal distribution. However, an evaluation of *v(x)* requires an analytical expression for the density (and its derivatives) of a summation of *J* normally distributed trimmed means that possess a complex correlation structure. Since the analytical derivation is prohibitively difficult, and since it may be unreasonable to invoke asymptotic results in the context of grouped data computed from household surveys, we use Monte Carlo simulations to determine the 'small-sample bias' for the estimator.

An issue related to the estimation of income distributions concerns the bounded nature of their support (at zero).[13] If kernel density estimation is applied to the unit data or the quantile means, a downward boundary bias may arise at income levels close to or at the boundary. The boundary bias may, in turn, affect estimates of poverty and lead to distorted visual illustrations of income distributions. This is due to the fact that the mass close to and at zero (or, more generally, at the left boundary) is underestimated, in expectation, by as much as 50 percent (Marron and Ruppert, 1994). The aforementioned studies undertake a log-

[12] Full derivations can be found in Minoiu (2007, section 3).

[13] Although in household surveys, negative income levels are not uncommon.

transformation of the income averages before estimating the density. Since the transformation increases the mass towards the left hand side of the distribution, the boundary bias problem is partially circumvented. Boundary bias correction methods are available and should be applied for bounded support distributions. In this case, the log-transformation enables the (standard) kernel density estimator to perform better by increasing the number of observations in the area in which the density would otherwise be underestimated. We too apply a log transformation to the quantile means so as to demonstrate that even when the boundary bias is diminished by this transformation, application of KDE to grouped data remains problematic.

III. THE BANDWIDTH AND KERNELS CONSIDERED

We undertake kernel density analysis using software developed for this purpose, in the following sequence.[14] First, quintile, decile, and ventile income (or consumption) averages are computed from large populations (drawn from theoretical distributions) and representative household surveys. Second, a variety of bandwidths and kernels (described below) are used to obtain kernel density estimates. Third, samples are drawn from the fitted densities. In the final step, we compare summary statistics and poverty measures of the simulated samples with the same quantities from the original data.

In the Monte Carlo exercise, we use three first generation, rule-of-thumb bandwidths proposed by Silverman (1986, pp. 45-48). These are optimal in the sense that they seek to minimize the approximate mean integrated squared error. An important feature shared by Silverman's bandwidths is that they assume a normal distribution for the data. Other properties are presented below in summary form:

Bandwidth	**Formula**[15]	**Performance**
Silverman 1 (S1)	$1.06 \times \hat{\sigma} \times J^{-\frac{1}{5}}$, $\hat{\sigma} =$ standard deviation, J = # obs.; C = 1.06 for	- tends to oversmooth the density - performs poorly on heavily skewed distributions
Silverman 2 (S2)	$0.79 \times IQR \times J^{-\frac{1}{5}}$, IQR is the inter-quartile range	- leads to superior density estimates for long-tailed and heavily skewed distributions - does not do well on bimodal distributions
Silverman 3 (S3)	$0.9 \times A \times J^{-\frac{1}{5}}$, $A = \min(IQR/1.34, \hat{\sigma})$.	- achieves a more balanced amount of smoothing - works reasonably well on both

[14] The software ("Kernel Density Estimation and Analysis Tool") is available from the authors upon request.

[15] We use canonical bandwidths for all kernels so that all estimates are comparable across different kernels. The canonical bandwidths ensure that each bandwidth-kernel combination leads to the same amount of smoothing (or tradeoff between bias and variance) represented by the approximate value of the integrated mean squared error (Marron and Nolan, 1988). The constants shown in the table correspond to the Gaussian kernel.

	skewed and multimodal distributions (Silverman, 1986)

We also consider (in Section V) a bandwidth proposed by Sheather and Jones (1991) (labeled S-J) which is entirely data-driven and has been shown to outperform other rule-of-thumb bandwidths both theoretically (by achieving a smaller value of mean integrated squared errors) and in simulations. It is considered to be the best second generation plug-in estimator and is recommended as a benchmark for good performance (Jones, Marron, and Sheather, 1996).

A number of additional rule-of-thumb bandwidths proposed in the literature are added to the analysis (in Section VI) in order to cover as broad a range as possible of bandwidths in assessing the sensitivity of global poverty estimates to this parameter. We consider a variant of the plug-in estimator (Wand and Jones, 1995), as well as a variant of the S3 bandwidth in which the scale parameter is $\hat{\sigma}$ instead of $\min(IQR/1.34, \hat{\sigma})$. Results are also presented for the "oversmoothed bandwidth" (representing the upper bound to the integrated mean square error minimizer). It is the largest bandwidth consistent with a 'reasonable' amount of smoothing and thus likely to result in more smoothing than Silverman's S1 bandwidth. The oversmoothed bandwidth is thought of as a good starting point for subjective choice of bandwidth (Jann, 2005).

We also employ a hybrid bandwidth (which is Silverman's S3 bandwidth with scale parameter $\hat{\sigma}$ instead of $\min(IQR/1.34, \hat{\sigma})$) that is not data-driven in two major ways. First, the hybrid bandwidth takes the same value *across kernels* despite the fact that the amount of smoothing it achieves varies across kernels. This implies that the density estimates are not comparable across kernels given that each kernel-bandwidth pair corresponds to a different amount of smoothing. Second, the hybrid bandwidth is constant *across datasets* (e.g., countries) despite the fact that the standard deviation of each dataset (its scale parameter) will be different. Although the hybrid bandwidth does not possess theoretical underpinnings, we include it in the analysis to model the procedure of Sala-i-Martin (2006) who claims that when the fixed hybrid bandwidth is employed, different kernels give rise to the same national and global poverty estimates (Sala-i-Martin, 2006). Additionally, we wish to determine whether such a bandwidth (which is clearly not optimal in the sense defined in the literature) may outperform bandwidths that would conventionally be deemed optimal, when applied in poverty analysis. Following Sala-i-Martin (2002a, 2002b, 2004, 2006), the hybrid bandwidth is computed assuming a standard deviation of 0.6 (regardless of the country or dataset on which it is employed). The value of the hybrid bandwidth is therefore set at 0.39 for quintile data, 0.34 for decile data, and 0.296 for ventile data. These values exogenous to the data lead the hybrid bandwidth to generally be smaller than the optimal bandwidths for the datasets considered (and will thus lead to 'undersmoothing').

A wide range of weighting functions is considered (Gaussian, Epanechnikov, Quartic, Triweight, Triangular, and Uniform). It has been shown that the mean integrated squared error is minimized for the Epanechnikov kernel, but that asymptotically, the choice of kernel is inconsequential for achieving the minimum mean integrated squared error (Silverman, 1986). Since this analysis, however, is based on a small number of quantile means, there is

no *a priori* reason to discard any kernels. Each density is estimated at 100 equidistant points. Large samples (of 5000 observations) are simulated from the fitted density such that the share of observations with a specific income level is equal to the corresponding density estimate (upto rounding).[16]

IV. MONTE CARLO STUDY[17]

A. Theoretical Distributions

The Monte Carlo exercises uses data from four distributions: Log-normal, Dagum, Generalized Beta II, and a notional multimodal distribution (Figure 1).[18] We use parameter values for the Log-normal distribution that have been fit to Russian 1995 income data. The parameter values for the Dagum and Generalized Beta II distributions have been fit to Mexican 1996 income data. The multimodal distribution is the population-weighted 2004 world distribution of income, in which individuals of each country are assigned the per capita GDP of that country. Two modes of this distribution correspond to the mean incomes of India and China, and a third, lower mode corresponds to the mean income of the richest nations.

B. Summary Statistics, Density Estimates and Diagrams

The first question we seek to answer is how the grouped data-based kernel density estimator performs in describing the underlying distribution through summary statistics (e.g., means, medians, standard deviation, and quintile means). The findings are reported in Table 1.[19]

---

[16] Samples can be drawn from the kernel density estimate in several ways. One is to construct the sample so that the share of observations with a certain income level is equal to the estimated density at that point (up to rounding). A second approach goes one step further, by also linearly interpolating the incomes so that clumping of incomes (for example, at points of high density) does not take place. A third approach involves drawing observations from the density by constructing a random variable whose p.d.f. (or alternatively, C.D.F.) is exactly the kernel density estimate from grouped data. The third is that described in the text. The various approaches lead essentially to the same results, and we report here results based on the second.

[17] Sections IV-VI are based on Minoiu (2007).

[18] From each distribution, 200 samples with 1000 observations each are simulated. The parameters chosen for the first three distributions are those resulting from a parametric density estimation exercise undertaken by Bandourian, McDonald and Turley (2003) on 82 household surveys from 23 countries. The authors show, for example, that the Dagum distribution provides the best fit to unit income data in the class of three parameter distributions, and the Generalized Beta II distribution is the best performing distribution in the class of four parameter distributions. In the family of two-parameter distributions, the Log-normal distribution is chosen due to its wide usage in the literature on income distributions (see, for example, the estimation of country income distributions by Babones, 2003). In Bandourian, McDonald and Turley (2003), the best-fitting two parameter distribution to income datasets is the Weibull distribution. A Monte Carlo exercise has also been undertaken using data from the Weibull distribution, but the results were largely similar to those for the other three distributions (log-normal, Dagum and Generalized Beta II) and are therefore not reported.

[19] All tables in the paper show results based on KDE from five datapoints, unless otherwise specified.

Across the four distributions, we find that the mean is systematically *overestimated* (by at most one fifth), while the median is estimated fairly well for all distributions, the latter finding being consistent with the well-known fact that trimmed means are good estimators of location. In contrast, the standard deviation is substantially underestimated for the three unimodal distributions and is overestimated for the multimodal distribution. Some regularities can be observed in relation to the estimation of quantile means themselves. For every distribution considered, the ratios between the estimated quintile means and the true ones are always sub-unity for the poorest two population groups, and always higher than 1 for the richest two population groups. The average income of the poorest 40 percent of the population is systematically and substantially underestimated whereas that of the richest 40 percent of the population is systematically and substantially overestimated. It is only the average income of the middle 20 income quantile that is precisely estimated for all distributions other than the multimodal distribution (for which it is understated due to an inability of the density estimate to properly capture the first mode of the distribution).[20]

These findings (especially those for the middle of the distribution) are not surprising given the robust nature of trimmed means for estimating the location of underlying densities. It is observed, however, that using kernel density methods on grouped data generates important distortions precisely in the tails of the distributions. The systematic misestimation of the (average) incomes of the poorer and of the richer in a country will have an important effect on the values of poverty indicators, and will depend on the location of the poverty line along the density support. Although the density estimator assigns densities to income levels in the tails around the observed quantile means, it does so by drawing information primarily from the extreme quantile means. It thus faces a real difficulty in estimating the density at income levels far to the left (or right) of the extreme quantile means, and therefore the bandwidth plays a crucial role in allowing the weighting functions to "stretch" so as to produce nonzero densities at these far-off income levels.

Actual versus fitted densities and the size of the density bias along the support (for Log-normal data) are plotted two kernel-bandwidth pairs in Figure 2.[21] The first panel overlays the estimated densities fitted from grouped data on the true density, while the second panel plots the bias in the density estimate (expressed as difference between the average density estimate and its true counterpart). The most important conclusion based on these diagrams is that

[20] Histograms of the estimated quantile means for different kernel-bandwidth pairs are reported in Minoiu (2007, Graphs 1 and 2), demonstrating that a smaller bandwidth (in our setting, the lowest bandwidth considered happened to be the hybrid bandwidth) leads to a better fit of the observed quantile means. More specifically, an under-smoothed estimated density (resulting from a 'too-low' bandwidth) centers the mass on the observed datapoints, which are the quantile means themselves. The results concerning this comparison between optimal and the hybrid bandwidth are not reported here for brevity, but will be discussed again in the next two sections.

[21] We choose to do so rather than describe the performance of the estimator with statistics such as the Sum of Squared Errors or the Sum of Estimated Errors as these might miss important variation in the biases along the support. Furthermore, the points of estimation are kept fixed across draws to enable computation of the bias at each income level on the support. At every point of estimation, the densities are averaged across the draws (Figure 2).

KDE on grouped data gives rise to distortions in the estimated density *at every point* along the income support (with the exception of two crossing points where the bias is zero). Notably, the estimated density is biased *upwards* in the tails of the distribution. This is consistent with the previous finding that the average income of the poorest is underestimated, since too much mass is assigned to lower income levels. Similarly, the positive bias in the density at the right end of the distribution is consistent with the previous finding that too much mass is assigned to the higher income levels, inflating the average income of the richest. Furthermore, the diagrams demonstrate that the grouped data-based density is biased *downwards* in the middle of the distribution, which is also consistent with the finding that the average income of the middle 20 percent of the population is roughly well by the procedure, since the positive bias from the far left tail and the negative bias from the center of the distribution tend to cancel out. The second conclusion arising from these diagrams is that the choice of kernel is not consequential for the visual impression created by the density estimate. However, it should be noted that this is simply an artifact of our using canonical bandwidths to ensure that each kernel-bandwidth combination achieves the same amount of smoothing. As will be shown subsequently, dropping the canonical bandwidths (in favor of, say, the hybrid bandwidth), yields strikingly different density estimates depending on the kernel.

How do these biases affect poverty estimates? It is easiest to see the consequences of the density biases when poverty is measured with the headcount ratio. Consider a thought experiment in which the poverty line is first assumed to be below or at the first crossing point, and is then assumed to take increasing values, moving rightwards on the income support. In the first instance, the poverty headcount ratio is overestimated as long as the density is biased upwards. As the poverty line moves rightward on the support, the extent of overestimation will decrease until it becomes zero at the point at which the overestimation of the density in the tail cancels out the under-estimation of the density in the middle of the support. As the poverty line continues to be shifted rightward, the headcount ratio will remain underestimated (although to a decreasing degree) until it reaches 100 percent.

The poverty biases associated with KDE on grouped data for the multimodal distribution (Figure 3) demonstrate that the extent to which salient features of the underlying density are replicated by the estimator critically depends on the choice of parameters. The Gaussian-S3 pair produces a largely over-smoothed density which conceals the multiple modes of the distribution. In contrast, the (smaller) hybrid bandwidth is better able to reveal the modes of the data, although these modes are located at the quintile means instead of their true location. It can be concluded that visual illustrations of multimodal distributions obtained through density smoothing can be misleading in such a sparse data environment. As with unimodal distributions, distortions need be expected in the resulting density estimates in both directions (over- and under-estimation) and along the entire support.

Figure 4 reveals the pitfalls of the (constant) hybrid bandwidth relative to data-driven, optimal alternatives. In some datasets, the hybrid bandwidth is close to an optimal bandwidth by chance. The first panel shows how this could be the case. The two curves, corresponding to the S1 and hybrid bandwidths for Dagum data, although different, show that the hybrid bandwidth tends to under-smooth simply because it is smaller in value than S1. More

importantly, the bandwidth greatly influences the lowest income level at which the estimator *can* estimate nonzero density. Should a poverty line fall between the minimum income levels at which each of the two curves has nonzero density, then the hybrid bandwidth will yield zero poverty level (by any indicator), whereas S1 would yield positive values for poverty indicators. The first and second panels, taken together, show the effect of changing the kernel and keeping the bandwidth fixed (at the hybrid value). It is straightforward to see that the changing the amount of smoothing will be very consequential for the quality of the resulting diagram. For the Epanechnikov-hybrid bandwidth pair, density is concentrated at the quintile means, and is zero between the extreme modes and the central mode. These areas of zero density arise since the hybrid bandwidth is too small and the kernel has finite support, so there is no information from adjacent points to use. This is because no observations are available for estimation in the window where the estimator looks for 'neighbors'.

We may conclude that (a) first, the hybrid bandwidth can lead to the same level of smoothing as an optimal bandwidth, but it will do so only by chance; (b) in all other cases, the hybrid bandwidth leads to nontrivial bias in the estimated curve. This renders KDE-based diagrams difficult to interpret. This also renders the non data-driven hybrid bandwidth inappropriate for poverty analysis from grouped data.

C. Poverty Estimates

Poverty estimates are reported for a low and high poverty line (representing the population median multiplied by factors of 0.25 and 1.75, respectively) in Table 2. Given the results from the previous sections, we anticipate that the share of poor will be fairly well estimated for poverty lines located close to the center of the distribution, and less well estimated for poverty lines elsewhere. We consider, however, a range of poverty indicators, some of which take account of the depth of poverty (measured as the distance between the income of the poor and the poverty line) and report results for the poverty headcount ratio, the poverty gap, the squared poverty gap, and the more distributionally-sensitive FGT (3) and FGT (4) indices.

Table 2 demonstrates that the poverty headcount ratio is systematically overestimated for the low poverty line, and underestimated for the high poverty line. For datasets of quintile means, the share of poor is overestimated by a factor of 1.17 of its true counterpart (Log-normal distribution) or 2 percentage points. The biases rise up to a factor of 1.28 or 3.4 percentage points for datasets of decile means. The biases are slightly lower for ventile means.[22] The FGT indicators of the depth of poverty are more substantially underestimated by quintile data than they are for decile and ventile data. The biases also appear to rise with the distributional sensitivity of the FGT indicator. The situation is reversed for the high

[22] Based on the Monte Carlo simulations we undertook, we concluded that the lack of monotonicity in the size of the biases associated with poverty estimates with the numbers of observations is only apparent in small datasets, and monotonicity is restored after approximately 25-30 quantile means. However, it is worth mentioning that 25 to 30 quantile means are rarely, if ever, available to researchers in lieu of unit data. Reassuringly, the 'global' performance of the estimator – assessed using the Sum of Squared Errors and the Sum of Absolute Errors monotonically improves with the number of available datapoints (results not reported).

poverty line. In particular, the poverty headcount ratio is now underestimated by almost 9 percent (or 7 percentage points) in the case of multimodal data. It is underestimated by between 5 and 7 percent (or approximately 5 percentage points) with data from the other distributions.

Figure 5 illustrates the bias in the headcount ratio – the poverty indicator with the widest application - for a wider range of poverty lines and all the distributions considered. As before, it is observed that the extent of poverty is broadly overestimated for lower poverty lines, is estimated relatively accurately for poverty lines near the population median (that is, in regions where the density biases cancel out), and is underestimated for higher poverty lines. When data are drawn from the multimodal distribution, a pronounced underestimation of poverty is observed for poverty lines around or higher than the median (upto 11 percentage points at the median since the density estimate "misses" the first mode of the distribution). For lower poverty lines, the positive bias is of at most 9 percentage points (at 0.5 x the median). The size of the bias and the way in which it varies with the data generating process and the location of the poverty line, is often large enough to give rise to concern about the appropriateness of the method in poverty analysis.

How the biases vary with the KDE parameter (bandwidth, kernel) is shown in Tables 3-4. The bandwidth has a substantial effect on the estimated poverty headcount ratio in the case of the multimodal distribution: while S1 leads to an upward bias of 70 percent, the hybrid bandwidth leads to a downward bias of 50 percent. Furthermore, the degree of distortion which arises when using a hybrid bandwidth on data from a multimodal distribution apparent: there are substantial downward biases associated with this bandwidth for *all* of the poverty indicators considered (Table 3). The Silverman bandwidths only occasionally do better, however the magnitude of the biases is large. The biases arising from different bandwidth-kernel pairs which do not attain the same amount of smoothing are in some cases, substantial (Table 4). For example, for the Dagum distribution, the estimates of the poverty headcount ratio are either biased upwards by 11 percent (Gaussian kernel) or biased downwards by 9 percent (Triweight kernel).

To conclude, we find it difficult to describe the magnitude and sign of biases in poverty indicators through statements applicable across many distributions and parameters. The Monte Carlo simulations demonstrate that the biases are often substantial, and that they vary with the nature of the data generating process (which is in the very nature of the non-parametric estimator), as well as with the bandwidth, weighting function and number of quantile means available for analysis. For the unimodal distributions considered, the poverty headcount ratio is overestimated for lower poverty lines, accurately estimated at poverty lines close to the population median, and underestimated for higher poverty lines. It may be much more difficult to determine the pattern of biases for distributions suspect of multiple models, since the positioning of the poverty line relative to the modes, and the extent of smoothing, is likely to determine the sign and size of the error.

V. COUNTRY STUDIES

In this section, grouped-data KDE-based and survey-based poverty estimates are presented using nationally representative household data for three countries with varying levels of poverty: Tanzania, Nicaragua, and Vietnam.[23] The $1/day and $2/day international poverty lines are considered[24], along with a capability, nutritionally anchored poverty line developed by Reddy, Visaria and Asali (forthcoming).

For the $1/day poverty line, the headcount ratio is overstated by a factor of at most 1.6 and understated by a factor of at most 0.94 regardless of the number of quantile means used (Table 5). For the $2/day poverty line, the headcount ratio is, in contrast, understated by at most 8 percent (e.g., the Nicaraguan $2/day poverty headcount ratio of 79.03 percent is understated by approximately 6 percentage points when the input data are quintile means). The degree of over- or underestimation of the poverty headcount ratio is lower for the higher poverty line. Similarly, the poverty gap ratio is overestimated (by a factor of maximum 1.75) for the least poor country (Vietnam), is less misestimated for Nicaragua, and is occasionally underestimated for the poorest country (Tanzania). It is noteworthy that, as in the case of Monte Carlo simulations, the bias of poverty estimates does *not* vary monotonically with the number of quantile means analyzed.

Table 6 contains poverty estimates for different bandwidths (using the capability poverty line, which falls closer to the median of the surveys than do the $1/day and $2/day poverty lines, hence the higher relative accuracy of the estimator). The choice of the bandwidth, however, has a substantial impact on estimated poverty. In particular, the poverty headcount ratio is overestimated by 12 percent (S1 bandwidth, Nicaragua) or by 5 percent (S3 and hybrid bandwidth, Nicaragua). The distributionally-sensitive FGT (3) is overestimated by a factor of 2 using S1 and by one fifth using S3 (Vietnam). Holding the bandwidth the same, the biases vary both across countries and across poverty indicators. In each case considered, we have highlighted in bold face the best performing optimal bandwidth, which appears to be S3 in the majority of cases.[25] All the figures in Table 6 indicate positive biases associated with the technique on quintile means. This can be explained, in light of the Monte Carlo evidence, by the relative position of the capability poverty line vis-à-vis the survey median.

---

[23] The datasources are as follows. For Vietnam: The 1998 Vietnam Living Standards Survey (VLSS) contains information on per capita expenditure of households at current prices for 22,510 individuals. Source: World Bank Living Standards Measurement Study (LSMS), Development Economics Research Group (DECRG). For Nicaragua: The 1997-98 Living Standards and Measurement Survey contains information on per capita consumption for 18,383 individuals. Source: World Bank Living Standards Measurement Study (LSMS), Development Economics Research Group (DECRG). For Tanzania: The 2000-01 Household Budget Survey contains information on per capita consumption for 22,176 households. Source: National Bureau of Statistics, Tanzania, 2002. The data for all three household surveys are weighted to take account of survey design.

[24] We do not here discuss the conceptualization of the poverty lines, as we only use them for expository purposes. However, an assessment of the money-metric approach to setting poverty lines can be found in Reddy and Pogge (2006).

[25] Biases vary less across kernels (when we use canonical bandwidths) and we do not report the results here.

Diagrams of kernel density estimates from grouped data are presented for varying numbers of quantile means, bandwidths, and weighting functions in order to determine whether KDE-based visual representations of the underlying log-consumption distributions can accurately replicate features of that distribution (Figure 6). The first panel super-imposes kernel density estimates from grouped data for different bandwidths (for a fixed kernel and quintile means). It is apparent, in this example, that the S1 bandwidth gives rise to some oversmoothing of the density. The density biases in the left tail of the distribution are also evident. The S3 bandwidth reveals the beginning of a mode in the right tail. However, this is entirely the artifact of using quintile means as input data. There is no such mode in the underlying survey data, as shown by its survey-based histogram. Panels (2) and (3) for Nicaragua show the effect of changing the kernel in two environments: the first uses canonical bandwidths (ensuring that the amount of smoothing is kept constant across density estimates) and the second utilizes the hybrid bandwidth (implying that the amount of smoothing changes across density estimates). Panel 2 demonstrates that keeping the bandwidth fixed across kernels can lead to extremely distorted visual representations of the underlying density. This is naturally not the case in Panel 3, where the effect of the kernel is smaller on the estimated density. Finally, the last panel proves yet again that the density estimator (Quartic kernel - S1 bandwidth) leads to positive density biases in the left and right tails of the distribution, and negative biases in the center of the distribution. These distortions take place at every point along the density support and have important consequences for poverty estimation. Furthermore, the kernel density estimate on decile means is more biased locally in the left tail of the density than the estimate on quintile means. However, the estimate on decile means is less biased globally than the estimate on quintile means.

VI. Global Poverty

In this section, we assess the sensitivity of world poverty estimates to parameters of the kernel density estimation procedure. The focus is on the effect of changing the bandwidth on the estimated share and number of poor in the developing world. Income shares for 94 developing countries covering 94 percent of the world's population in 1990 were obtained from the UNU/WIDER World Inequality database V. 2.0a (2005) for the years 1990 and 2000 (or closest available year). These were scaled using the per capita GDP (at PPP) from the WDI (2006) to obtain quintile income averages, and kernel density estimation was undertaken for each country. We then aggregated the fitted country distributions to obtain the world distribution of income.

To compare the results to Sala-i-Martin's studies, we also undertook the same analysis for the year 2000 in a larger sample comprising 134 developed and developing countries. Similar poverty rates to those reported by the author were obtained. For example, the world poverty headcount ratio computed in this study for the $1.5/day poverty line in the year 2000 (using the Gaussian kernel and a similar value for the bandwidth) is 405 million, while the author's is 398.4 million (Sala-i-Martin, 2006). Consequently, the range of global poverty estimates associated with various bandwidths presented in this section are directly comparable to those reported by Sala-i-Martin (2002a, 2002b, 2004, 2006). To preview our results, the analysis

leads us to conclude that Sala-i-Martin's choice of bandwidth is likely to have led in the direction of underestimating global poverty.[26]

We consider the following data-dependent bandwidths (described in Section 3): Silverman's rule-of-thumb bandwidth (S3) and a variant, the "oversmoothed bandwidth", the Sheather-Jones plug-in estimator, and the direct plug-in estimator. To enable a direct comparison of our global poverty estimate with those proposed by Sala-i-Martin (2006), we also report the results for the hybrid bandwidth. We focus on the Gaussian kernel, although the results are similar with the Epanechnikov kernel and canonical bandwidth. Poverty is estimated using the headcount ratio and the aggregate headcount using five international poverty lines, ranging between $1/day and $4/day. It is important to stress that the global rates and headcounts presented here should not be interpreted as authoritative due to numerous uncertainties concerning crucial elements of the analysis (Reddy and Pogge, forthcoming).

The results (Tables 7 and 8) demonstrate the remarkable lack of robustness of global poverty rates to changes in the value of the bandwidth even when optimal data-driven bandwidths are considered. In both 1990 and 2000, the poverty rates for the $1/day poverty line are the most sensitive to the bandwidth; this is because the $1/day poverty line is likely to fall in the left tail of the income distributions, where poverty is typically overestimated, and small changes in the bandwidth have a large impact. Furthermore, the extent of poverty shows the highest variation across bandwidths for the lowest poverty lines considered. In particular, for the $1/day poverty line, the poverty headcount ratio varies by a factor of 1.8 when the oversmoothed bandwidth is considered and by a factor of 1.6 when it is excluded. The headcount ratios vary to a lesser degree for higher poverty lines and they are almost equal across bandwidths for the $3/day and $4/day poverty lines in 1990. However, the poverty rates are even more sensitive to the choice of optimal bandwidth in the year 2000. For the $1/day, $1.5/day and $2/day poverty lines, they vary by a factor between 1.4 and 1.8.

This variation in headcount ratios translates into a variation in the number of poor people between 162 and 278 million. To put these numbers in perspective, under- or overestimating the '$2/day poor' by 278 million individuals (in 2000) would represent an error of 50 percent (based on the $2/day global headcount for 2000 by Sala-i-Martin, 2006). Similarly, under- or overestimating the number of '1.5/day poor' by 180 million individuals (in 1990) would represent an error of 36 percent (based on the $1.5/day global headcount in the same year of Sala-i-Martin, 2006). Since Sala-i-Martin's analysis exclusively focuses on the hybrid

[26] It is noteworthy that our analysis only concerns one assumption underlying Sala-i-Martin's estimates of global poverty. There may be other reasons still why the true extent and trend of world poverty may be misestimated by the author, such as the choice of per capita income estimates (Deaton, 2005), or that of poverty lines (Reddy and Pogge, forthcoming, and Nye, Reddy and Pogge, 2002). More generally, the analysis presented in this section (and in the paper) does not enable us to make unambiguous statements concerning the bias associated with the application (in a manner appropriate for unit data) of KDE methods to grouped data on the estimated level and trend of world poverty. Based on the results presented in this study, such statements could only be made under stringent assumptions, such as that the true income distributions of various countries are unimodal and do not change shape over time, and that the poverty lines considered stand in a certain relation to the median of the true income distributions.

bandwidth, it is apparent that the global poverty estimates he reports may substantially underestimate the true level of poverty in the world as a result of this methodological choice.

How does this range of variation inform us on the trend in world poverty between 1990 and 2000? In results not reported (see Minoiu, 2007, Table 11), the fall in the $1.5/day and $2/day poverty rates ranges between 7 percent and 18 percent across the bandwidths considered. The number of people who were lifted from $1/day poverty between 1990 and 2000 ranges between 19 million and 38 million, whereas the reduction in $1.5/day poverty ranges between 45 and 92 million. It should be noted that a reduction in the number of '$1.5/day poor' by 45 million is only *one half* of that documented by Sala-i-Martin (2006). Similarly, a reduction in the number of '$1 /day poor' of 25 million is only *one fifth* of Chen and Ravallion's (2004) documented fall of 129.5 million (between 1990 and 2001). It can thus be concluded that the range of variation associated with kernel density estimates based on different bandwidths may lead us to reach more pessimistic conclusions about the trend in world poverty since 1990. Nonetheless, it must be pointed out that all estimates are consistent with a *reduction* in world poverty. The underlying cause of this finding may be the use of kernel density estimation as such, the composition of the sample, or other methodological choices. This finding it is at odds with Chen and Ravallion's (2004) reported *increase* in the number of '$2/day poor' of 81 million over the same period.

VII. CONCLUSIONS

Recent influential poverty studies employ kernel density estimation methods on grouped data to analyze poverty and to describe features of the income distributions (e.g., Sala-i-Martin, 2002a, 2002b, 2004, 2006). This method is often used because of the lack of availability or the difficulty in obtaining access to unit data from representative household surveys for countries and years of interest (including from large countries such as China and India). In this paper, we analyze the performance of the kernel density estimator in poverty analysis from grouped data. We use both Monte Carlo simulations and nationally representative household surveys to compare KDE-based poverty estimates with their counterparts from the original data.

We find that the biases resulting from the application of this technique depend on the bandwidth, the kernel, the number of data-points analyzed and the data generating process. The average income of the poorer population groups is *overstated* by the technique, while the average income of the richer groups is *understated* for the range of unimodal distributions considered here. This often leads to *overestimation* of the poverty headcount ratio for lower poverty lines, and *underestimation* of the poverty headcount ratio for higher poverty lines. The ambiguity of these results illustrates the point that, whereas the existence of biases in poverty estimates derived from the application of KDE to grouped data is not in itself surprising, the direction and magnitude of the biases could not easily have been predicted in the absence of an exercise of the kind we have undertaken. The biases associated with poverty indicators are substantial: for the poverty rate, they can reach 6-7 percentage points in unimodal distributions and 10-11 percentage points in the multimodal distribution considered. Kernel density estimation on grouped data can also give rise to misleading visual representations of the income distributions as these too are sensitive to the choice of parameters.

A sensitivity analysis for global poverty estimates reveals that KDE-based headcount ratios vary by a factor of 1.8 for a range of bandwidths that have been recommended and used in the literature. Similarly, the number of '$2/day poor' varies by 136 million in 1990 and by 278 million in 2000 depending on this parameter. This analysis demonstrates that the global poverty estimates recently presented by Sala-i-Martin (2002a, 2002b, 2004, 2006) are highly sensitive to the bandwidth and that the author's choices of how to implement kernel density estimation is likely to have led in the direction of underestimating global poverty.

The findings of this study give rise to serious concern about the validity and robustness of poverty analysis based on kernel density estimation from grouped data. They also raise questions about recently published national, regional, and global poverty estimates. The applied researcher should exercise caution in employing standard kernel density estimation methods on grouped data. Alternative estimation methods such as parametric Lorenz curve interpolation should be applied whenever possible.

REFERENCES

Ackland, R., Dowrick, S. and B. Freyens, forthcoming, "Measuring Global Poverty: Why PPP Methods Matter," *Economic Journal*.

Aziz, J. and C. Duenwald, 2001 "China's Provincial Growth Dynamics," IMF Working Paper No. 01/3 (Washington: International Monetary Fund).

Babones, S.J., 2003, "One World or Two? A Snapshot of the Global Income Distribution", paper presented at the American Sociological Association 98th Annual Meeting (August 16−19), Atlanta, Georgia.

Bandourian, R., McDonald, J.B. and R.S. Turley, 2003 "A Comparison of Parametric Models of Income Distribution across Countries and Over Time," *Revista Estadística*, Vol. 55, pp. 164−165.

Bhalla, S., 2002, *Imagine There's No Country: Poverty, Inequality, and Growth in the Era of Globalization*, IIE (Washington: Institute for International Economics).

Berry, A., Bourguignon, F. and C. Morrisson, 1983, "Changes in the World Distribution of Income between 1950 and 1977," *Economic Journal*, Vol. 93, pp. 331−350.

Bianchi, M., 1997, "Testing for Convergence: Evidence from Non-Parametric Multimodality Tests," *Journal of Applied Econometrics*, Vol. 12, No. 4, pp. 393−409.

Bourguignon, F. and C. Morrisson, 2002, "Inequality Among World Citizens: 1820−1992," *American Economic Review*, Vol. 92, No. 4, pp. 727−744.

Chen, S. and M. Ravallion, 2002 "How Did the World's Poorest Fare in the 1990s?," *Review of Income and Wealth*, Vol. 47, No. 3, pp. 283−300.

_____and_____, 2004, "How Have the World's Poorest Fared Since the Early 1980s?," World Bank Development Research Group Working Paper No. 3341 (Washington: The World Bank).

_____and_____, 2007, "China's (Uneven) Progress Against Poverty," *Journal of Development Economics*, Vol. 82, No. 1, pp. 1−42.

Coady, D., 2005, "The Distributional Impacts of Indirect Tax and Public Pricing Reforms: A Review of Methods and Empirical Evidence," IMF Fiscal Affairs Department PSIA Review Paper (Washington: International Monetary Fund).

Deaton, A., 2003, "How to Monitor Poverty for Millennium Development Goals," *Journal of Human Development*, Vol. 4, No. 3, pp. 353−378.

_____, 2005, "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)," *Review of Economics and Statistics*, Vol. 87, No. 1, pp. 1−19.

Dhongde, S., 2005, "Spatial Decomposition of Poverty in India", in Kanbur, R., Venables, A., Wan, G. (eds.), *Spatial Disparities in Human Development: Perspectives from Asia*, United Nations University Press.

Essama-Nssah, B., Pereira da Silva, L.A. and I. Samake, 2002, "A Poverty Analysis Macroeconomic Simulator (PAMS) Linking Household Surveys with Macro-Models," World Bank Policy Research Working Paper No. 2888 (Washington: The World Bank).

_____, 2005a, "Simulating the Poverty Impact of Macroeconomic Shocks and Policies," World Bank Policy Research Working Paper No. 3788 (Washington: The World Bank).

_____, 2005b, "Inequality and Poverty Simulations within the Lorenz Framework," Unpublished manuscript, World Bank Poverty Reduction Group (Washington: The World Bank).

Fuentes, R., 2005, "Poverty, Pro-Poor Growth and Simulated Inequality Reduction," UNDP Human Development Report Office Occasional Paper No. 11 (New York: United Nations Development Programme).

Grosh, M.E. and E.W. Nafziger, 1986, "The Computation of World Income Distribution," *Economic Development and Cultural Change*, Vol. 35, pp. 347−359.

Jann, B., 2005, "Univariate Kernel Density Estimation", Boston College Department of Economics, Statistical Software Component No. S 456410.

Jones, M.C., Marron J.S. and J.S. Sheather, 1996, "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, Vol. 91, pp. 401−407.

Jones, C.I., 2002, "On the Evolution of the World Income Distribution," *Journal of Economic Perspectives*, Vol. 11, No. 3, pp 19−36.

Kakwani, N.C. and H.H. Son, 2006, "New Global Poverty Counts," UNDP International Poverty Center Working Paper No. 29 (Brasilia: United Nations Development Programme).

Marron, J.S. and D. Nolan, 1988, "Canonical Kernels for Density Estimation," *Statistics and Probability Letters*, Vol. 7, pp 195−199.

Marron, J.S., and D. Ruppert, 1994, "Transformations to Reduce Boundary Bias in Kernel Density Estimation," *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 56, No. 4, pp 653−671.

Mason, D.M., 1981, "Asymptotic Normality of Linear Combinations of Order Statistics with a Smooth Score Function," *The Annals of Statistics*, Vol. 9, No. 4, pp. 899−908.

Milanovic, B., 2002, "True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone," *Economic Journal*, Vol. 112, pp 51−92.

_____, 2005, *Worlds Apart: Measuring International and Global Inequality*, Princeton University Press.

Minoiu, C., 2007, "Poverty Analysis based on Kernel Density Estimates from Grouped Data", Columbia University Institute of Social and Economic Research and Policy Working Paper No. 2007−07 (New York: Columbia University).

_____and S. Reddy, forthcoming, "Estimating Poverty and Inequality from Grouped Data: How Well Do Parametric Methods Perform?," *Journal of Income Distribution*.

Moore, D.S., 1968, "An Elementary Proof of Asymptotic Normality of Linear Functions of Order Statistics," *The Annals of Mathematical Statistics*, Vol. 39, No. 1, pp. 263−265.

Mosteller, F., 1946, "On Some Useful *Inefficient* Statistics," *The Annals of Mathematical Statistics*, Vol. 17, No. 4, pp. 377−408.

Nye, H., Reddy, S.G. and T. Pogge, 2002, "What is Poverty?," *The New York Review of Books*, Vol. 49, No. 18, November 21.

Pittau, M.G., 2005, "Fitting Regional Income Distributions in the European Union," *Oxford Bulletin of Economics and Statistics*, Vol. 67, No. 2, pp. 135−161.

_____and R. Zelli, 2006, "Empirical Evidence of Income Dynamics across EU Regions," *Journal of Applied Econometrics*, Vol. 21, pp. 605−628.

Pritchett, L., 2006, "Who Is *Not* Poor? Dreaming of a World Truly Free of Poverty," *The World Bank Research Observer*, Vol. 21, No. 1, pp. 1−23.

Quah, D.T., 1996, "Twin Peaks: Growth and Convergence in Models of Distribution Dynamics," *Economic Journal*, Vol. 106, pp. 1045−1055.

Quah, D.T., 1997, "Empirics for Growth and Distribution: Polarization, Stratification and Convergence Clubs," *Journal of Economic Growth*, Vol. 2, No. 1, pp 27−59.

Reddy, S.G. and C. Minoiu, 2007, "Has World Poverty *Really* Fallen?," *The Review of Income and Wealth*, Vol. 53, No. 3, pp. 466−484.

_____, Visaria, S. and M. Asali, forthcoming, "Inter-country Comparisons of Poverty Based on a Capability Approach: An Empirical Exercise," in Basu, K. and R. Kanbur, eds.: *Arguments for a Better World: Essays in Honor of Amartya Sen*, Oxford University Press.

_____ and T. Pogge, forthcoming, "How *Not* To Count the Poor," in Anand, S., P. Segal and J. Stiglitz, eds.: *Measuring Global Poverty*, Oxford University Press.

Sala-i-Martin, X., 2002a, "The World Distribution of Income (Estimated from Individual Country Distributions)," NBER Working Paper No. 8933 (Boston: National Bureau of Economic Research).

_____, 2002b, "The 'Disturbing' Rise of World Income Inequality," NBER Working Paper No. 8904 (Boston: National Bureau of Economic Research).

_____, 2004, "The World Distribution of Income: Falling Poverty and …Convergence, Period," Unpublished manuscript, Columbia University, Department of Economics.

_____, 2006, "The World Distribution of Income: Falling Poverty and …Convergence, Period," *Quarterly Journal of Economics*, Vol. 121, No. 2, pp. 351−397.

Sheather, S.J. and M.C. Jones, 1991, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 53, No. 3, pp. 683−690.

Siddiqui, M. M. and C. Butler, 1969, "Asymptotic Joint Distribution of Linear Systematic Statistics from Multivariate Distributions," *Journal of the American Statistical Association*, Vol. 64, No. 325, pp. 300−305.

Silverman, B.W., 1986, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability 26, Chapman & Hall/CRC.

Shorrocks, A. and G. Wan, 2008, "Ungrouping Income Distributions: Synthesising Samples for Inequality and Poverty Analysis," UNU World Institute for Development Economics Research, Research Paper No. 16 (Helsinki: United Nations University).

Stigler, S.M., 1973, "The Asymptotic Distribution of the Trimmed Mean," *The Annals of Statistics*, Vol. 1, No. 3, pp. 472−477.

_____, 1974, "Linear Functions of Order Statistics with Smooth Weight Functions," *The Annals of Statistics*, Vol. 2, No. 4, pp. 676−693.

UNU/WIDER (2005) World Inequality Database V. 2.0a (June). Available on: http://www.wider.unu.edu/research/Database/en_GB/database/

Yatchew, A., 1998, "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, Vol. 36, No. 2, pp. 669−721.

Yotopoulos, P. A., 1989, "Distributions of Real Income: Within Countries and by World Income Classes", *Review of Income and Wealth*, Vol. 35, pp. 357−376.

Wand, M.P. and M.C. Jones, 1995, *Kernel Smoothing*. Chapman and Hall: London.

World Development Indicators Online Database (2006) (Washington: The World Bank).

World Bank, 1994, A Database on Poverty and Growth in India, World Bank Research Department and Poverty and Human Resources Division Policy (Washington: The World Bank).

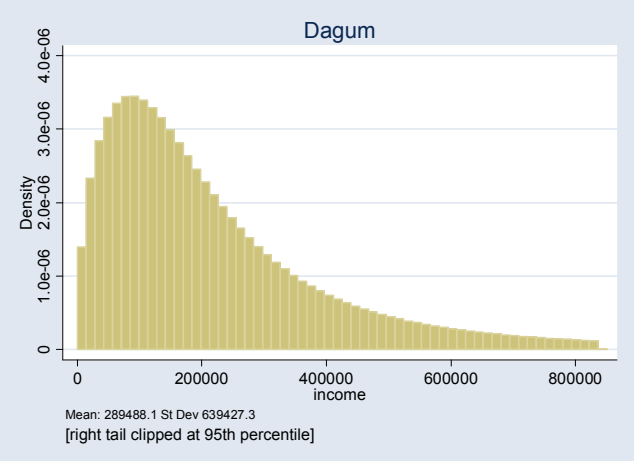
Wu, X. and J. Perloff, 2007, "GMM Estimation of a Maximum Entropy Distribution with Interval Data," *Journal of Econometrics*, Vol. 138, Issue 2, pp. 532−546.

_____and_____, 2005, "China's Income Distributions: 1985−2001," *Review of Economics and Statistics*, Vol. 87, pp. 763−775.

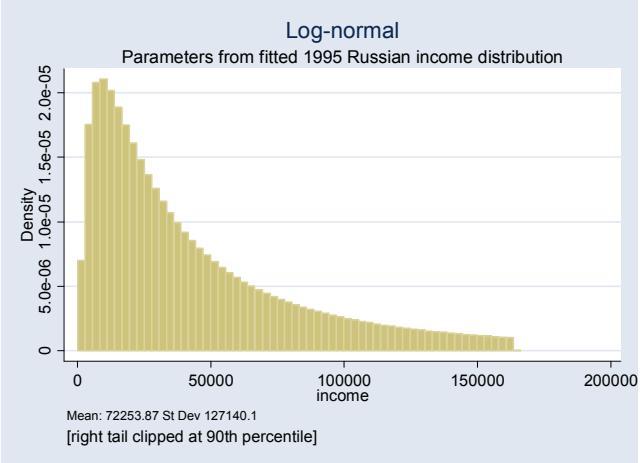
APPENDIX

Figure 1. Distributions used in Monte Carlo analysis

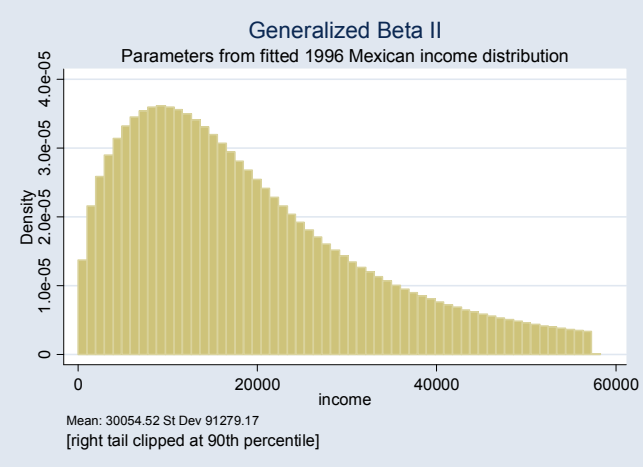
Panel 1. Dagum distribution



Panel 2. Log-normal distribution



Panel 3.
Generalized Beta II distribution



Panel 4.
Multimodal distribution representing the 2004 population-weighted world distribution of income

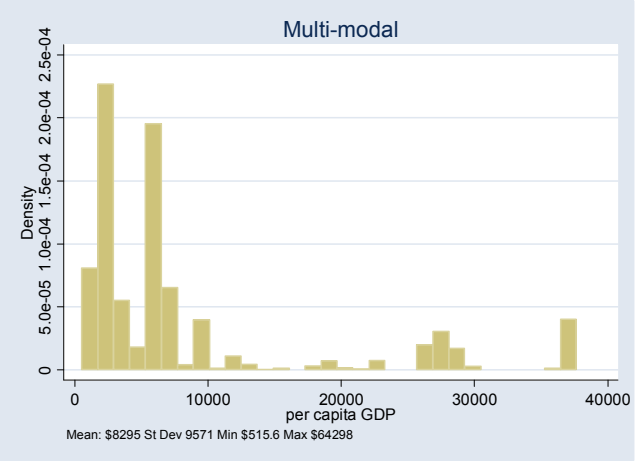


Table 1. Summary statistics from KDE-based sample

	Summary statistics			**Quintile means**				
Distribution↓	Mean	Median	St. Dev.	Bottom	Second	Third	Fourth	Top
Log-normal	1.12	1.03	0.89	0.93	0.94	1.03	1.20	1.25
Dagum	1.11	0.98	0.59	0.98	0.92	1.01	1.17	1.14
GB 2	1.13	1.02	0.45	0.99	0.92	1.01	1.19	1.12
Multimodal	1.17	0.91	1.24	0.68	0.87	0.92	1.11	1.04

Note: All figures represent the ratio between the estimated quantity and its true value.
Parameters: Quartic kernel, S3 bandwidth. Input data: Quintile means (The results are broadly similar for the other Silverman bandwidths and the other five kernels.)

Figure 2. Bias of KDE-based density (log-normal distribution)

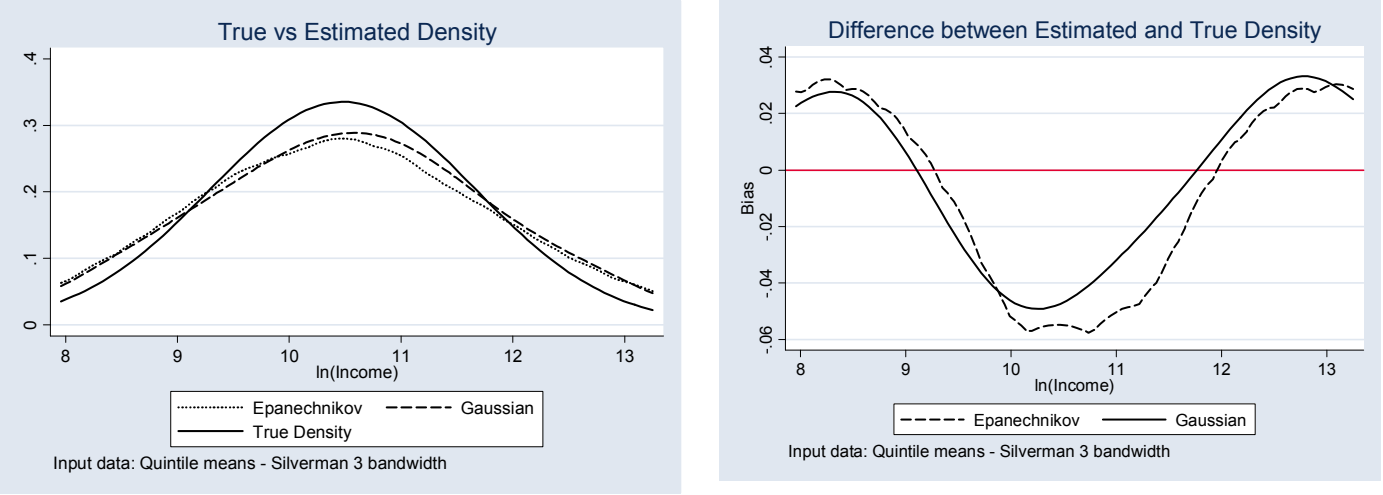


Figure 3. Bias of estimated density (multimodal distribution)

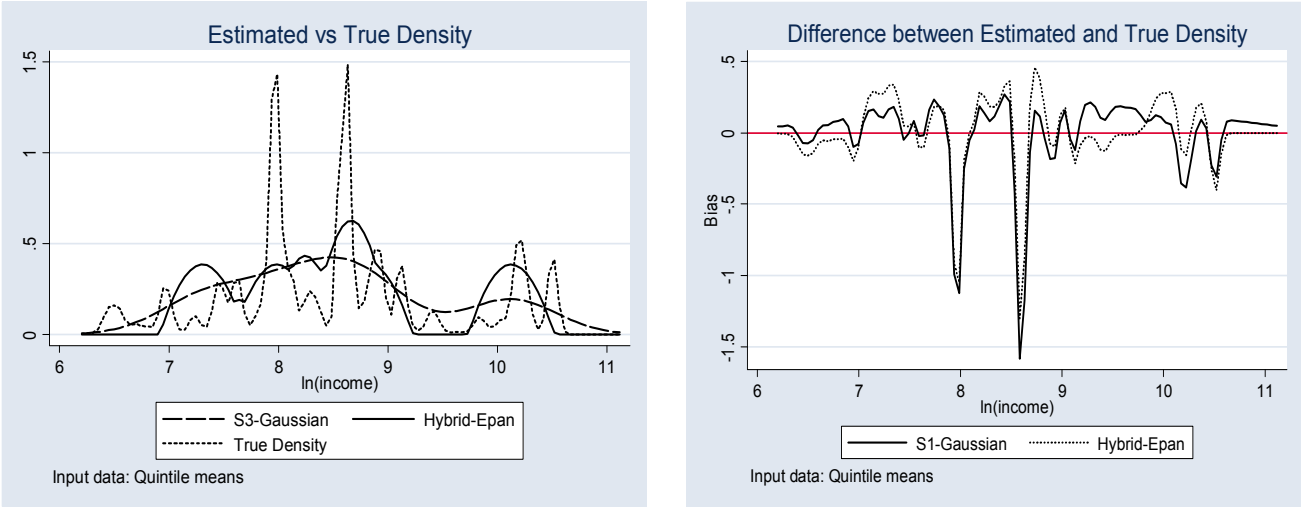


Figure 4. Bias of estimated density (Dagum distribution)

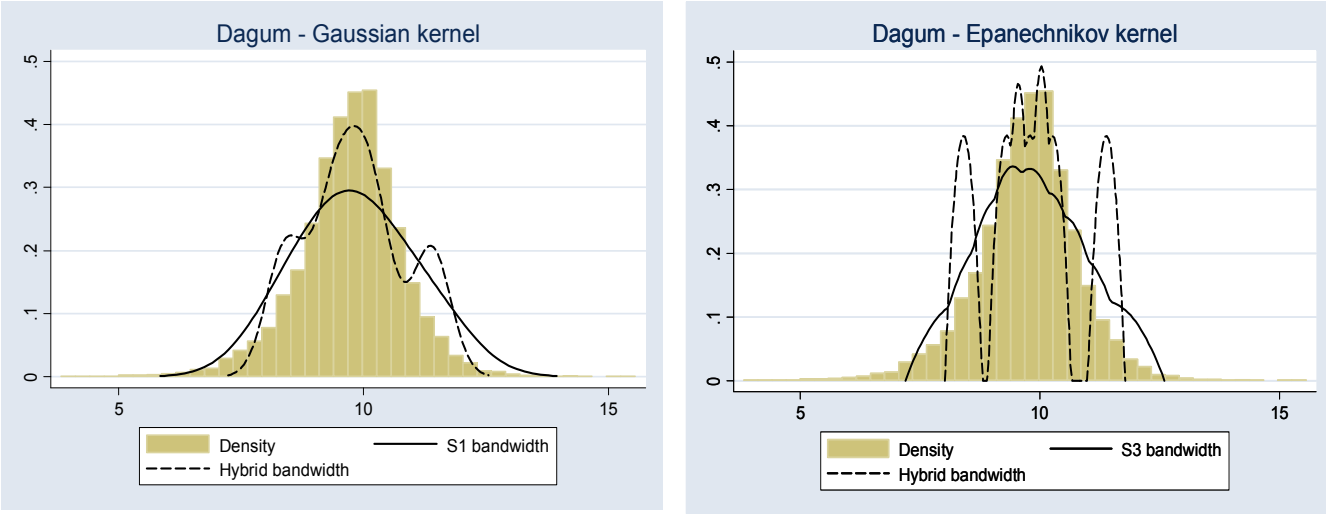


Table 2. Bias of poverty measures (low and high poverty lines; multiple poverty indicators)

Poverty indicator:	Distribution	True quantity	Input data: Quintiles	Deciles	Ventiles
	Poverty line →			**LOW**	
Poverty headcount ratio (%)	Log-normal	12.10	1.17	1.28	1.23
	Dagum	9.43	1.09	1.26	1.21
	Gen. Beta II	9.45	1.07	1.24	1.18
	Multimodal	8.14	1.00	1.18	1.15
Poverty gap ratio	Log-normal	4.57	1.14	1.40	1.34
	Dagum	3.93	0.77	1.13	1.19
	Gen. Beta II	4.02	0.73	1.10	1.15
	Multimodal	2.93	0.63	1.04	1.08
Squared poverty gap	Log-normal	2.49	1.02	1.40	1.35
	Dagum	2.30	0.52	0.98	1.12
	Gen. Beta II	2.40	0.48	0.93	1.07
	Multimodal	1.22	0.48	1.09	1.18
FGT(3)	Log-normal	1.54	0.92	1.42	1.37
	Dagum	1.56	0.35	0.83	1.03
	Gen. Beta II	1.65	0.31	0.78	0.98
	Multimodal	0.56	0.38	1.17	1.34
FGT(4)	Log-normal	1.03	0.82	1.41	1.37
	Dagum	1.15	0.23	0.69	0.94
	Gen. Beta II	1.23	0.20	0.64	0.89
	Multimodal	0.27	0.30	1.29	1.54
	Poverty line →			**HIGH**	
Poverty headcount ratio (%)	Log-normal	68.02	0.95	0.96	0.97
	Dagum	73.63	0.93	0.95	0.97
	Gen. Beta II	73.97	0.93	0.95	0.96
	Multimodal	81.83	0.91	0.91	0.94
Poverty gap ratio	Log-normal	40.77	0.99	1.01	1.01
	Dagum	40.71	0.99	1.01	1.01
	Gen. Beta II	40.75	0.99	1.01	1.04
	Multimodal	45.79	0.96	0.97	0.98
Squared poverty gap	Log-normal	28.95	1.02	1.05	1.04
	Dagum	27.40	1.02	1.05	1.04
	Gen. Beta II	27.35	1.02	1.05	1.04
	Multimodal	30.18	1.00	1.01	1.01
FGT(3)	Log-normal	22.16	1.04	1.08	1.07
	Dagum	20.17	1.04	1.08	1.07
	Gen. Beta II	20.11	1.04	1.08	1.06
	Multimodal	21.54	1.02	1.05	1.04

FGT(4)	Log-normal	17.71	1.06	1.11	1.09
	Dagum	15.68	1.05	1.11	1.09
	Gen. Beta II	15.63	1.04	1.10	1.08
	Multimodal	16.19	1.04	1.07	1.07

Note: Figures in the last three panels represent the ratio between the estimated quantity and its true counterpart.
Parameters: S3 bandwidth, Epanechnikov kernel

Figure 5. Bias in the poverty headcount ratio versus location of poverty line

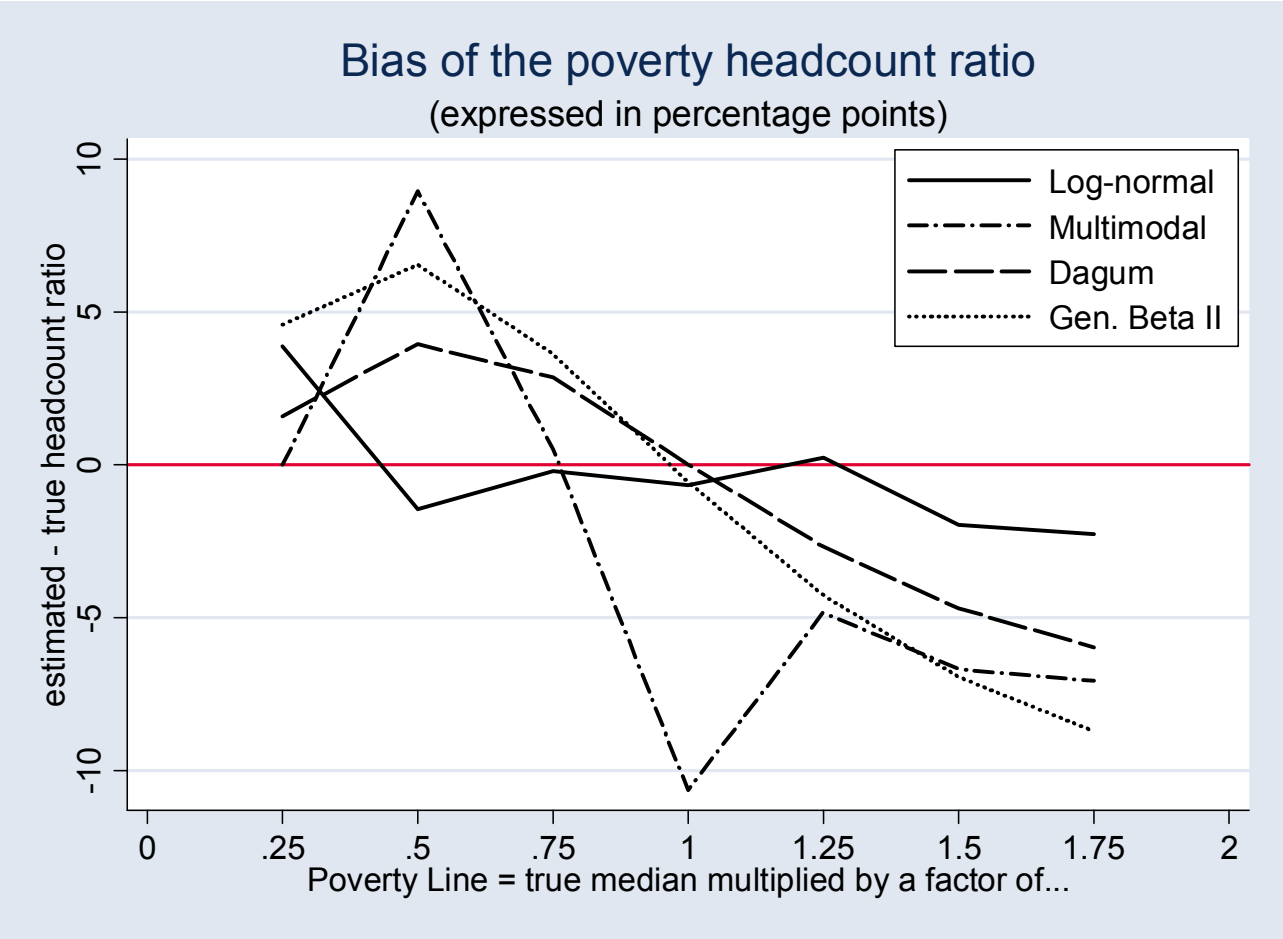


Table 3. Bias of poverty measures (Triweight kernel, Poverty line: 0.25 x median)

			Bandwidth:			
Poverty indicator:	Distribution	True quantity	S1	S2	S3	Hybrid (Sala-i-Martin)
Poverty headcount ratio (%)	Log-normal	12.10	1.45	1.24	1.17	1.40
	Dagum	9.43	1.48	1.17	1.10	1.04
	Gen. Beta II	9.45	1.48	1.14	1.08	1.00
	Multimodal	8.14	1.70	1.10	1.00	0.50
Poverty gap ratio	Log-normal	4.57	1.71	1.28	1.12	0.63
	Dagum	3.93	1.37	0.88	0.75	0.26
	Gen. Beta II	4.02	1.34	0.84	0.71	0.24
	Multimodal	2.93	1.67	0.76	0.62	0.10
Squared poverty gap	Log-normal	2.49	1.82	1.21	1.00	0.25
	Dagum	2.30	1.22	0.65	0.50	0.07
	Gen. Beta II	2.40	1.18	0.60	0.46	0.06
	Multimodal	1.22	1.95	0.65	0.47	0.03

Note: Figures in the last four panels represent the ratio between the estimated quantity and its true counterpart.

Table 4. Bias of poverty measures (Hybrid bandwidth, Poverty line: 0.5 x median)

			Kernel:					
Poverty indicator:	Distribution	True quantity	Gaussian	Uniform	Epan.	Quartic	Tri-weight	Tri-angular
Poverty headcount ratio (%)	Log-normal	27.91	0.98	1.01	0.97	0.95	0.93	0.95
	Dagum	23.57	1.04	1.05	0.98	0.93	0.91	0.95
	Gen. Beta II	23.30	1.11	1.05	0.98	0.94	0.91	0.95
	Multimodal	19.19	1.39	1.43	1.40	1.36	1.33	1.36
Poverty gap ratio	Log-normal	12.37	0.91	1.00	0.97	0.96	0.95	0.97
	Dagum	10.15	0.89	1.00	0.98	0.98	0.98	0.98
	Gen. Beta II	10.12	1.04	0.98	0.98	0.97	0.97	0.98
	Multimodal	8.71	0.99	1.05	1.05	1.04	1.03	1.04
Squared poverty gap	Log-normal	7.39	0.78	0.91	0.89	0.89	0.88	0.89
	Dagum	6.04	0.70	0.83	0.82	0.82	0.82	0.82
	Gen. Beta II	6.09	0.89	0.80	0.81	0.81	0.80	0.81
	Multimodal	4.89	0.74	0.79	0.80	0.79	0.78	0.80

Note: Figures in the last six panels represent the ratio between the estimated quantity and its true counterpart.

Table 5. Bias of poverty measures (Epanechnikov kernel, S3 bandwidth)

		Poverty line: $1/day				Poverty line: $2/day			
Indicator	Country	Survey estimate	Quintiles	Deciles	Ventiles	Survey estimate	Quintiles	Deciles	Ventiles
Poverty headcount ratio (%)	Vietnam	5.20	1.34	1.59	1.47	35.69	1.04	1.04	1.03
	Nicaragua	44.62	1.02	1.02	1.02	79.03	0.92	0.95	0.96
	Tanzania	75.39	0.94	0.96	0.96	94.75	0.97	0.97	0.98
Poverty gap ratio	Vietnam	0.89	1.18	1.75	1.64	9.07	1.16	1.23	1.18
	Nicaragua	16.59	1.08	1.12	1.10	40.93	0.98	0.99	1.00
	Tanzania	34.67	0.99	1.00	1.01	61.40	0.96	0.97	0.98
Squared poverty gap	Vietnam	0.26	0.87	1.65	1.61	3.35	1.21	1.37	1.30
	Nicaragua	8.24	1.11	1.20	1.17	25.27	1.01	1.04	1.03
	Tanzania	19.39	1.03	1.06	1.05	43.47	0.97	0.99	0.99
FGT(3)	Vietnam	0.10	0.57	1.43	1.46	1.49	1.22	1.47	1.39
	Nicaragua	4.66	1.13	1.28	1.24	16.96	1.04	1.08	1.07
	Tanzania	11.94	1.06	1.11	1.09	32.18	0.99	1.01	1.01
FGT(4)	Vietnam	0.04	0.36	1.18	1.28	2.49	1.18	1.53	1.45
	Nicaragua	2.85	1.13	1.34	1.30	11.98	1.07	1.12	1.10
	Tanzania	7.82	1.09	1.16	1.13	24.55	1.00	1.03	1.02

Note: Figures in the relevant panels represent the ratio between the estimated quantity and its survey counterpart.

Table 6. Bias of poverty measures (Gaussian kernel, Poverty line: Capability)

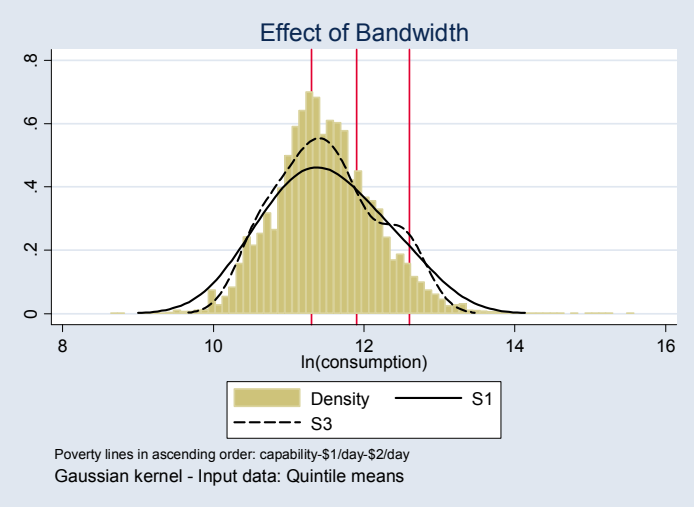
			Bandwidth:				
Indicator	Country	Survey estimate	S1	S2	S3	S-J	Hybrid
Poverty headcount ratio (%)	Vietnam	41.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	Nicaragua	30.61	1.12	1.06	**1.05**	1.07	**1.05**
	Tanzania	40.13	1.04	1.03	**1.02**	1.03	1.03
Poverty gap ratio	Vietnam	11.39	1.33	1.17	**1.12**	1.19	1.22
	Nicaragua	9.69	1.42	1.19	**1.13**	1.23	1.15
	Tanzania	12.62	1.29	1.15	**1.11**	1.18	1.16
Squared poverty gap	Vietnam	4.38	1.65	1.29	**1.19**	1.34	1.41
	Nicaragua	4.33	1.71	1.26	**1.16**	1.33	1.18
	Tanzania	5.61	1.50	1.22	**1.14**	1.26	1.24
FGT(3)	Vietnam	2.00	1.98	1.37	**1.22**	1.45	1.56
	Nicaragua	2.26	1.98	1.30	**1.15**	1.40	1.18
	Tanzania	2.91	1.69	1.25	**1.14**	1.32	1.29

	Vietnam	1.02	2.31	1.41	**1.21**	1.54	1.69
FGT(4)	Nicaragua	1.30	2.25	1.32	**1.12**	1.46	1.17
	Tanzania	1.65	1.87	1.26	**1.11**	1.36	1.31

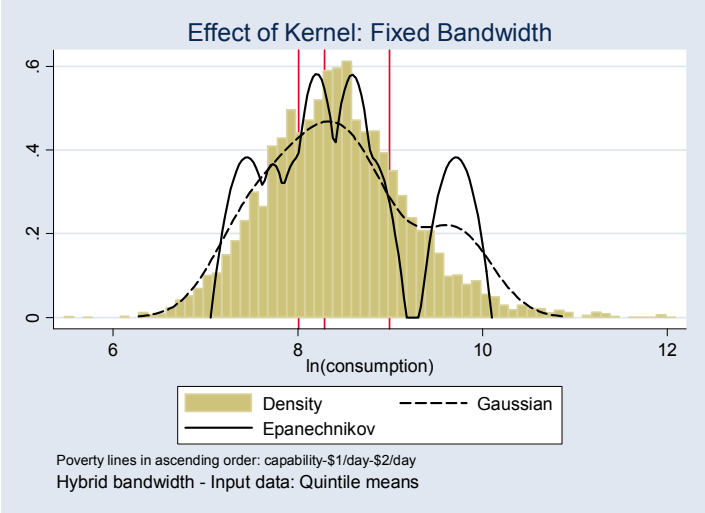
Note: Figures in the last five panels represent the ratio between the estimated quantity and its survey counterpart.

Figure 6. Survey-based and grouped data KDE-based density estimates

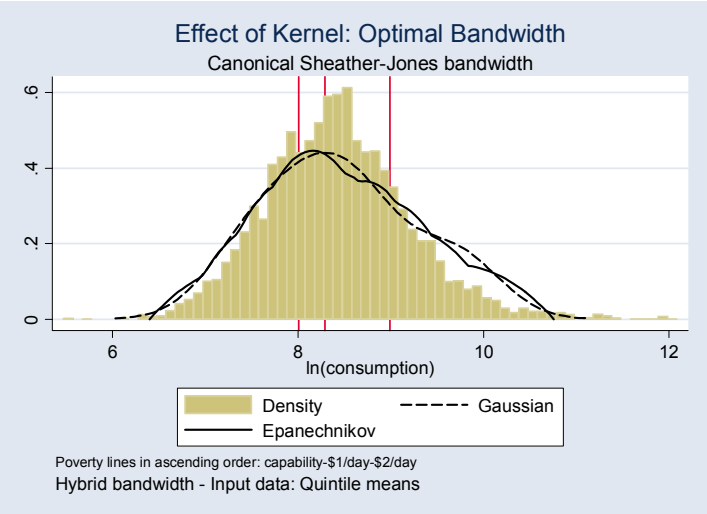
Panel 1. Tanzania



Panel 2. Nicaragua



Panel 3. Nicaragua



Panel 4. Vietnam

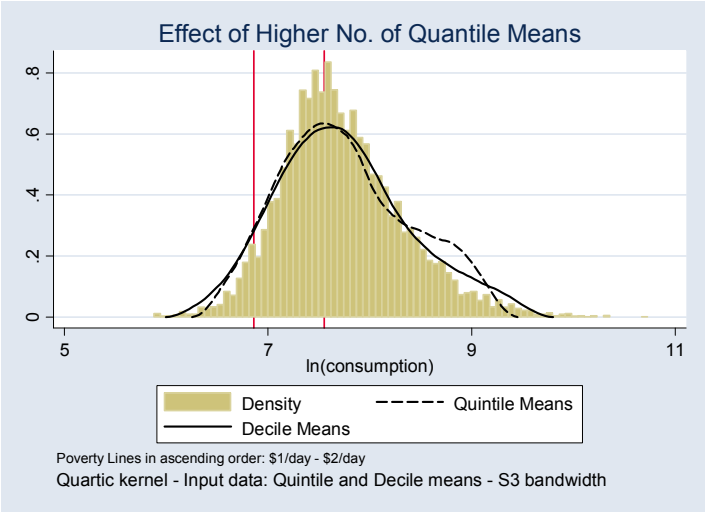


Table 7. Global poverty rates (% poor)

Bandwidth→	S3	Overs-smoothed	Variant of S3	Sheather-Jones	Direct plug-in	Hybrid (Sala-i-Martin)	Ratio b/w highest and lowest estimate	Percentage point diff. b/w highest and lowest estimate
			Year: 1990					
$1/day	7.2	9.5	6.4	7.5	8.4	5.3	*1.8*	*4.2*
$1.5/day	13.4	16.2	12.8	13.9	14.9	11.7	*1.4*	*4.5*
$2/day	24.5	26.8	24.2	25.2	25.8	23.4	*1.1*	*3.4*
$3/day	38.1	38.7	37.8	38.0	38.3	37.1	*1.0*	*1.6*
$4/day	49.8	49.4	50.3	49.9	49.6	49.6	*1.0*	*0.9*
			Year: 2000					
$1/day	5.3	7.5	4.8	5.6	6.2	4.2	*1.8*	*3.3*
$1.5/day	9.4	12.6	8.9	10.0	10.7	6.9	*1.8*	*5.7*
$2/day	17.2	20.7	16.5	17.7	18.7	15.0	*1.4*	*5.7*
$3/day	27.7	30.0	27.4	27.9	28.8	25.7	*1.2*	*4.3*
$4/day	38.1	39.4	38.3	38.8	38.9	37.1	*1.1*	*2.3*

Table 8. Global poverty counts (millions)

Bandwidth→	S3	Overs-smoothed	Variant of S3	Sheather-Jones	Direct plug-in	Hybrid (Sala-i-Martin)	Difference b/w highest and lowest estimate
			Year 1990				(millions)
$1/day	289	381	257	303	338	213	*168*
$1.5/day	540	651	518	559	599	471	*180*
$2/day	987	1079	975	1016	1040	943	*136*
$3/day	1536	1560	1524	1533	1544	1496	*64*
$4/day	2008	1989	2026	2012	1998	2001	*37*
			Year 2000				
$1/day	256	362	232	269	300	200	*162*
$1.5/day	452	606	426	481	517	333	*273*
$2/day	830	998	796	850	899	720	*278*
$3/day	1331	1445	1319	1341	1384	1235	*210*
$4/day	1833	1893	1843	1866	1870	1784	*109*