Poverty Analysis based on Kernel Density Estimates from Grouped Data

Camelia Minoiu[1]

Revised version of job market paper[2]

December 15, 2006

Abstract. Kernel density estimation (KDE) has been prominently used to measure poverty from grouped data (representing mean incomes of a small number of population quantiles). In this paper I analyze the performance of this method. Using Monte Carlo simulations for plausible theoretical distributions and unit data from several household surveys, I compare KDE-based poverty estimates with their true and survey counterparts. It is shown that the technique gives rise to biases in poverty whose sign and magnitude vary with the smoothing parameter, the kernel, the number of data-points analyzed, and the poverty indicators used. I also demonstrate that KDE-based global poverty rates and headcounts are highly sensitive to the choice of smoothing parameter. Depending on the parameter, the estimated proportion of '$1/day poor' in 2000 varies by a factor of 1.8, while the estimated number of '$2/day poor' in 2000 varies by 287 million people. These findings give rise to concern about the validity and robustness of kernel density estimation in poverty analysis. However, they provide a framework for interpretation of existing results using this technique.

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

[1] Department of Economics and Institute for Social and Economic Research and Policy, Columbia University. Email: cm2036@columbia.edu Tel. +1 212 854 6385 Fax +1 212 854 7998.

[2] A full version of my job market paper is available on www.columbia.edu/~cm2036/kdepoverty.pdf. The full-length version also contains the acknowledgements, the full-length appendix, and a slightly more detailed discussion of the results.

I. Motivation

Several recent studies have employed nonparametric smoothing techniques, and in particular kernel density estimation methods (henceforth, 'KDE') on grouped data to obtain estimates of national, regional, and global poverty (see Sala-i-Martin 2002a, 2002b, and 2006; Ackland, Dowrick, and Freyens, 2004; Fuentes, 2005).[3] World poverty assessments require the use of grouped data (usually expressed as income averages for a small number of population quantiles) because unit data from representative household surveys is not available, or is difficult to obtain for all countries and years of interest (this is the case in particular for large countries such as China and India). To assess long-run trends in global poverty and inequality, researchers often rely on grouped data since household surveys are unavailable; however, published summary statistics exist.[4] The accuracy of poverty estimates and the plausibility of visual representations of income distributions crucially depend on the statistical method employed on this informationally limited data structure.

The goals of this study are twofold. First, I assess the appropriateness of kernel density estimation methods on grouped data for poverty analysis. Biases in poverty estimates are identified for a wide range of poverty indicators, poverty lines, parameters (e.g., bandwidths and kernels) and possible income distributions. In this way, I propose new ways of interpreting the national, regional, and global KDE-based poverty estimates that have been recently put forth in the literature. Second, I analyze the robustness of the procedure with the given choice of parameter (such as the smoothing parameter) in an exercise of global poverty estimation. The findings in this study can be used by applied researchers who wish to undertake smoothing techniques in order to analyze poverty and/or describe salient features of income distributions from grouped data.

The data structure considered represents income averages for a small number of population quantiles (usually five). Since analytical derivations of the properties of the KDE estimator are impossible in small samples, Monte Carlo simulations are needed. The following plausible theoretical income distributions are considered: Log-normal, Dagum, Generalized Beta II, and a notional multimodal distribution. Next, I 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, I assess the performance of KDE in global poverty analysis, using grouped income data for a large number of developing countries.

There are reasons to believe that the application of KDE to grouped data may give rise to biases in poverty estimates. However, the size of those biases (for different poverty indicators, plausible income distributions, and distinct poverty lines) is unknown and requires a study of this kind. The data structure analyzed in this study is informationally poorer than a large sample

[3] 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.

[4] See, for example, the study on the evolution of world inequality since 1820 by Bourguignon and Morrison (2002).

drawn from the underlying distribution. This is one reason why nonparametric density estimation methods are thought of as inappropriate on grouped data. However, a small number of average incomes are a richer source of information about the underlying distribution than are a small number of observations from that same distribution. I find that poverty is incorrectly estimated in a majority of cases, but is occasionally correctly estimated when biases in the density at different points along the support cancel out. The order of magnitude of the poverty headcount ratio biases identified in this study reaches 6-7 percentage points (for unimodal distributions) and 10-11 percentage points (for the multimodal distribution). Furthermore, the biases associated with Foster-Green-Thorbecke (FGT) indicators generally increase with the degree of distributional sensitivity.

One reason why KDE has been applied to poverty analysis is that, in contrast to parametric approaches, it does not require prior beliefs about the nature of the true data generating process. Furthermore, it is a convenient procedure because it reproduces the entire underlying density function from a small, manageable amount of data. Therefore, it is particularly useful for poverty analyses undertaken for multiple countries and years, since the analysis of unit data may be prohibitive in terms of time and manpower, and since unit data may be unavailable for numerous country-years. As mentioned, unit data from nationally representative household surveys for large countries such as China and India are not publicly available. The Chinese State Statistical Bureau publishes grouped data from underlying rural and urban household surveys in its China Statistical Yearbook. Similarly, summary statistics from Indian National Statistical Surveys are available for the period 1950-1994 through the World Bank Database on Poverty and Growth in India (1994). While China's survey data has not been made available to outside researchers, India's can be obtained from the National Sample Survey Organization (NSSO) provided that the research is relevant to national development and planning.[5] It is therefore important to know whether the statistical techniques undertaken on grouped data provide reliable estimates of the various features of interest of underlying distributions. As Reddy and Minoiu (2006) demonstrate, conclusions regarding the world's progress towards achieving the first Millennium Development Goal of halving severe poverty, crucially depend on China's and India's poverty reduction experiences.

Kernel density estimation is one of *two* methods that have been used most widely in poverty analysis from grouped data. The alternative approach is the parametric estimation of Lorenz curve functional forms.[6] Many Lorenz parametric forms have been proposed, and two can be readily implemented with the poverty calculation tool POVCAL developed by the World Bank (see, e.g., Chen and Ravallion (2005) for a poverty assessment in China from grouped data; Yotopoulos (1989) for an inequality, poverty, and affluence assessment exclusively based on grouped data from selected developed and developing countries[7]; and Chen and Ravallion (2002,

[5] 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.

[6] Maximum entropy density estimation (for densities from the exponential family) has recently been proposed by Wu and Perloff (2003) as an alternative technique for poverty analysis on grouped data. In an application to Chinese data, the authors found that the technique provided reliable estimates (Wu and Perloff, 2005). However, we are unaware of studies which assess the performance of maximum entropy density estimation methods on small samples of quantile means derived from a range of plausible income distributions.

[7] The Lorenz curve parameterization used by Yotopoulos (1989) is that of Kakwani and Podder (1976) as distinguished from those of POVCAL (Villasenor and Arnold, 1989 and Kakwani, 1980).

2004), Bhalla (2002), Pritchett (2006), and Kakwani and Son (2006) for global poverty analyses using POVCAL). Although we do not explicitly compare the estimates presented in the study with those that would be obtained through Lorenz curve estimation, it is noteworthy that the parameterizations embodied in POVCAL perform well in estimating poverty from grouped data when the underlying distributions are unimodal, but do less well in estimating poverty from distributions that are multimodal (Minoiu and Reddy, 2006).

In poverty analysis, KDE methods have been undertaken on datasets of five quantile means (of income) per country and per year.[8] For example, Sala-i-Martin (2002a, 2002b, 2006) uses five quantile means obtained from (actual or fitted) income shares and an estimate of each country's income for 138 countries. The author concludes that there have been substantial reductions in world income poverty (according to all indicators considered) over the past three decades. In particular, after applying KDE to this grouped data, the author arrives at the conclusion that the share of people with an income level lower than $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 thereby estimating world poverty. The first method (described as the "kernel of quintiles" method) consists of constructing a dataset in which each person's income level is taken to be the average income of the national population quintile to which that person belongs. Subsequently, kernel density estimates are obtained from the data. 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. In both methods, poverty is subsequently estimated from the KDE-based world distribution of income, while regional estimates are obtained from KDE-based (population-weighted) regional distributions of income. The assessment undertaken in the current study of the KDE technique on grouped data is more relevant to the "kernel of kernels" method proposed by Sala-i-Martin.

Sala-i-Martin's use of KDE in poverty analysis has been widely cited and his global poverty estimates have been debated.[9] A number of academic papers have subsequently used his proposed methodology. For example, Ackland et al (2004) use the same technique to investigate the sensitivity of regional and global poverty estimates to alternative approaches of constructing purchasing power parity (PPP) conversion factors. The authors use five income shares and an estimate of per capita income for 97 countries as well as the "kernel of kernels" method to obtain poverty estimates for countries, regions and the world. However, their final aim is to show that the choice of index number in the estimation of PPPs (in particular, Geary-Khamis versus EKS) greatly affects the resulting poverty estimates. In another study, Fuentes (2005) uses grouped

[8] For example, Chen and Ravallion (2004) use household surveys for some countries and grouped data for others (e.g., China) to arrive at the conclusion that the share of the developing world's population living under a consumption level of $1.08 per day has fallen between 1981 and 2001 from 40.4 percent to 21.1 percent. When grouped data is available, the authors use parametric Lorenz forms to estimate poverty.

[9] 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 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).

data (i.e., an unspecified number of income shares for population percentiles) to estimate the distribution of income, as well as inequality and poverty incidence in several countries.[10]

The remainder of this paper is organized as follows: in the next two sections, I discuss the nature of the data structure and the bias of the kernel density estimator on grouped data. Section IV contains a description of the methods used in this paper. In Section V, the results of the Monte Carlo analysis for plausible income distributions are presented. Section VI shows findings from a comparison of poverty estimates from household surveys with those from KDE on grouped data for three countries with varying levels of poverty. In Section VII, a sensitivity analysis of KDE-based global poverty estimates is discussed. Conclusions are drawn in section VIII.

II. The data structure

The situation facing a researcher who seeks to estimate poverty from grouped data can be described, in short, as follows. Information on a variable of interest (e.g., income, consumption, or total wealth) is collected through a nationally representative household survey. The survey however, is not available to the researcher in its entirety. Instead, she possesses average incomes of several population quantiles. Alternatively, the researcher possesses income shares computed from the survey. An estimate of total income[11] is then used to scale the income shares and obtain average incomes of several population quantiles.

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" (Mosteller, 1946) represent the sole source of information from which the researcher aims to recover features of the income distribution.[12]

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*

[10] 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).

[11] Estimates of per capita income (or consumption) can be drawn from representative household surveys or the national accounts. Large discrepancies have been documented between survey-based and national accounts-based estimates of per capita income and consumption. Deaton (2005), for example, concludes that choosing the latter estimate over the former in global assessments of economic performance, may lead to understating the rate of global poverty reduction, and overstating average growth rates.

[12] Mosteller (1946) coined the term "systematic statistics" to refer to linear functions of order statistics. The early literature following his paper focused on robust estimation from systematic statistics of location and scale parameters of the underlying distributions. In our context, since kernel density estimation is applied to quantile means instead of the actual income realizations used to compute these means, it is important to characterize the joint distribution of the quantile means.

in ascending order. The unit data from the survey represents independently and identically distributed draws from the unknown income distribution. The process of ordering the independent and identically distributed draws from the underlying distribution 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 subsequent operation of averaging the order statistics reduces the informational content of the original sample. However, the averages retain important information about the underlying distribution due to the prior *ordering* of the original observations.

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.[13] The limit behavior of trimmed means has been investigated, among others, by Stigler (1973, 1974) and Mason (1981). These authors 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.[14] A small number of quantile means are therefore informationally richer than a small sample from the underlying distribution (in particular because it carries more precise information about the location of underlying order statistics along the support), but informationally poorer than a large sample from the underlying distribution.

Finally, 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), may appear inappropriate due to the very nature and purpose of nonparametric statistics. The applied economist is encouraged to use nonparametric estimators in "exploratory data analysis, as a confirmatory tool, or as a supplement to the standard parametric fare" (Yatchew, 1998). 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) argues that "interpolation is only deemed reliable among close neighbour[ing] observations, and extrapolation outside the observed domain is considered entirely speculative". With these considerations in mind, I proceed to discuss the bias of the estimator.

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

[14] 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).

III. Bias of the estimator

The bias of the kernel density estimator used on grouped data is derived and compared to the survey-based kernel density estimator (which employs all the underlying data). Assume that the observed individual income levels are realizations of a collection of iid random variables $\{X_1, X_2,...X_N\}$ drawn from the unknown density $f(x)$ with positive support $[0,\infty)$. The order statistics of the data are given by $\{\tilde{X}_1 \leq \tilde{X}_2 \leq ... \leq \tilde{X}_N\}$. The sample is divided into J equal-sized groups, also known as population quantiles. Suppose, WLOG, that the number of observations within each group is M , such that $JM = N$. Averages across incomes within each quantile are then computed such that a collection of quantile means representing linear functions of order statistics, denoted by $\{\hat{u}_1, \hat{u}_2,..,\hat{u}_J\}$, is obtained.

The underlying data are not observed. Instead, the quantile means $\{\hat{u}_1, \hat{u}_2,..,\hat{u}_J\}$ are available. Each quantile mean denoted by $\widehat{u}_j$ is equal to:

$$\hat{u}_J = \frac{1}{M}\sum_{i=1}^{M} \tilde{X}_i^{(j)} \text{ , where } j = 1,...,J \qquad [1]$$

The kernel density estimator on quantile means is given by:

$$\hat{f}(x) = \frac{1}{Jh}\sum_{j=1}^{J} k(\frac{x-\hat{u}_j}{h}) \qquad [2]$$

The bias of the kernel density estimator at point of estimation x is given by:

$$Bias(\hat{f}(x)) = E(\hat{f}(x)) - f(x) = \frac{1}{J}\sum_j E(\frac{1}{h}k(\frac{x-\hat{u}_j}{h})) - f(x) = \frac{1}{Jh}\sum_j \int k(\frac{x-\hat{u}_j}{h}) g_j(\hat{u}_j) d\hat{u}_j - f(x)$$
$$[3]$$

where $g_j(\cdot)$ is a density probability function of the j^{th} quantile mean. Following Silverman's (1986) derivation of the bias of the regular kernel density estimator, a change of variable $\frac{x-\widehat{u}_j}{h} = t_j$ is performed. Using the symmetry of the kernel density, equation [4] obtains:

$$Bias(\hat{f}(x)) = \frac{1}{J}\sum_j \int k(t_j) g_j(x - ht_j) dt_j - f(x) \qquad [4]$$

As in Silverman (1986), h is treated as constant in these derivations.[15] For small t_j, a Taylor expansion series approximation of $g_j(x - ht_j)$ around $g_j(x)$ is undertaken:

[15] The bandwidth can be chosen according to some optimality criterion (e.g., minimization of the approximate mean integrated squared error) and thus be made a function of the size of the underlying data.

$$g_j(x - ht_j) \simeq g_j(x) + g_j'(x)ht_j + \frac{g_j''(x)}{2}h^2 t_j^2 + \frac{g_j'''(x)}{6}h^3 t_j^3 + \ldots \qquad [5]$$

The bias becomes:

$$Bias(\hat{f}(x)) = \frac{1}{J}\sum_j \int k(t_j)\left[g_j(x) + g_j{}'(x)ht_j + \frac{g_j{}''(x)}{2}h^2 t_j{}^2 \right] dt_j + .. - f(x)$$

$$= \frac{1}{J}\sum_j g_j(x) + \frac{h}{J}\sum_j g_j{}'(x)\int t_j k(t_j)dt_j + \frac{h^2}{2J}\sum_j g_j{}''(x)\int t_j{}^2 k(t_j)dt_j + \ldots - f(x) \qquad [6]$$

Note that if the kernel is symmetric about zero, the second term disappears and equation [6] becomes:

$$Bias(\hat{f}(x)) = \frac{1}{J}\sum_j g_j(x) + \frac{h^2}{2J}\sum_j g_j{}''(x)\int t_j{}^2 k(t_j)dt_j + \ldots - f(x) \qquad [7]$$

For purposes of comparison, the bias of the ‘standard’, survey-based kernel density estimator (Silverman, 1986) is given by:

$$Bias(\hat{f}_S(x)) = \frac{h^2}{2} f''(x)\int t^2 k(t)dt + \ldots \qquad [8]$$

where the higher order terms in *h* have been suppressed and $\int t^2 k(t)dt$ is a constant depending on the weighting function. The bias of the standard kernel density estimator itself depends on the true unknown density function $f(x)$ (as well as the derivatives of this function). Furthermore, the bias is an increasing function of the bandwidth. A larger bandwidth leads to a larger bias since the former implies that information about the mass at a given point of estimation is collected from observations distant from that point of estimation.

As expected, the grouped data-based bias is itself a function of the unknown probability density functions associated with the quantile means. If we let $\frac{1}{J}\sum_j g_j(x) = v(x)$ 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, 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, I use Monte Carlo simulations to determine the size of the bias in the grouped data-based estimator when the data represents a small number of trimmed (quantile) means.

A second issue concerns the bounded nature of the support of income distributions.[16] If kernel density estimation is applied to the raw unit data or the raw 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). Most studies, however, undertake a log-transformation of the income averages before estimating the density. This operation shifts the mass towards the center of the distribution, partially circumventing the boundary bias problem (at the left-hand tail). Following the practice in the literature, quantile means are log-transformed in this analysis.

IV. The bandwidth and kernels considered

Kernel density analysis is undertaken using software developed for this purpose entitled the "Kernel Density Estimation and Analysis Tool".[17] Quintile, decile, and ventile means of per capita income (or consumption) computed from representative household surveys are inputted into the software. The data sources for all household surveys and descriptions of the income and consumption variables are given in the Appendix. A variety of bandwidths and kernels are used. Populations are subsequently drawn from the estimated density. Poverty indicators and quantile means are computed from those populations.

In the Monte Carlo exercises, I present results based on three different bandwidths. The first three are the "rule-of-thumb" automatic bandwidths which seek to minimize the approximate mean integrated squared error as proposed by Silverman (1986, pp. 45-48): Bandwidth S1 is given by the formula $1.06\times\hat{\sigma}\times J^{-\frac{1}{5}}$ where $\hat{\sigma}$ denotes the standard deviation of the data, *J* refers to the number of observations analyzed and the constant 1.06 corresponds to the Gaussian kernel[18]. Bandwidth S2 is given by $0.79\times IQR\times J^{-\frac{1}{5}}$ where IQR is the inter-quartile range of the data and serves as a more robust estimator of the spread of the distribution. Finally, Bandwidth S3 is given by $0.9\times A\times J^{-\frac{1}{5}}$ and is the third optimal bandwidth (for the Gaussian kernel) where $A=\min(IQR/1.34,\hat{\sigma})$. Bandwidths S1 to S3 are derived based on an important assumption that the data are generated from the normal distribution. Bandwidth S1 tends to over-smooth the density and may lead to important features of the density (such as heavy skewness) to be concealed, whereas Bandwidth S2 employs a more robust estimator of the dispersion of the underlying distribution and leads to superior density estimates for long-tailed and heavily skewed distributions (but not so on heavily bimodal distributions).[19] Bandwidth S3 attempts to achieve a

[16] Although in household surveys, negative income levels are not uncommon (since individuals can be dissaving).

[17] The software will be made available as freeware. The documentation of the software is available online on: www.columbia.edu/~cm2036/documentation.pdf

[18] 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).

[19] Notably, the estimator of dispersion (or more precisely, of the rapidity of fluctuations in the density) plays the role of a proxy for the second derivative of the unknown true density *f* - a quantity which affects the size of the kernel

balanced amount of smoothing that will work reasonably well on both skewed and multimodal distributions (Silverman, 1986). Silverman's optimal bandwidths are classified as "first generation" bandwidths by Jones, Marron and Sheather (1996). We choose to focus on them because of their widespread use in applied work (due to their availability as default in statistical packages).

In the exercises in which unit data from nationally representative household surveys is used (Section VI), I also present results based on the Sheather and Jones (1991) bandwidth (S-J). This bandwidth 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 practically (in simulations for a wide range of density shapes). 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).

Finally, in the global poverty exercise (Section VII), I enlarge the set of bandwidths to include other rule-of-thumb values proposed in the literature. The aim is to cover as broad a range as possible of first and second generation bandwidths in order to determine the impact of the bandwidth choice on world poverty estimates. I consider, first, a variant of the plug-in estimator (Wand and Jones, 1995), as well as a variant of Silverman's 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 highest bandwidth consistent with a 'reasonable' amount of smoothing and is likely to result in even more smoothing than Silverman's S1 bandwidth. However, it is considered to be a good starting point for subjective choice of bandwidth (Jann, 2005).

Throughout the study, I also employ a bandwidth (labeled as 'hybrid') that corresponds to a variation of Silverman's S3 bandwidth in which the scale parameter is $\hat{\sigma}$ instead of $\min(IQR/1.34, \hat{\sigma})$, and has additional important features. It is kept constant *across kernels* despite the fact that the amount of smoothing it achieves is different for each kernel; hence, the resulting density estimates are not strictly comparable across kernels because each kernel-bandwidth pair corresponds to a different amount of smoothing. Furthermore, the hybrid bandwidth is kept constant *across datasets* (corresponding, for example, to countries) despite the fact that optimal smoothing parameters defined in the literature are data-driven. The main reason why the hybrid bandwidth is considered in this study is to be able to assess the claim that despite fixing the bandwidth, different kernels produce the same poverty estimates from any given set of quantile means (Sala-i-Martin, 2006). Additionally, I wish to determine whether bandwidths that may not minimize the integrated mean square error (one of the most often used optimality criterion for bandwidth choice) and are not data-driven, under- or outperform optimal bandwidths when applied to poverty assessment. In the Monte Carlo simulations and for the country studies, the value of the hybrid bandwidth is set at 0.39 for quintile data, 0.34 for decile data, and 0.296 for ventile data. Following Sala-i-Martin (2002a, 2002b, and 2006), it is computed assuming a standard deviation for the data (regardless of the dataset on which it is employed) of 0.6. These set values lead the hybrid bandwidth to generally be smaller than the optimal bandwidths (in the datasets used in this study) and will naturally lead to undersmoothing.

density estimator bias [as shown in Section III].

The following six weighting functions are employed: 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. Hence, a wide range of weighting functions is considered. Furthermore, no restrictions are effectively put on the support of the density.[20] The density is estimated at 100 equidistant points along the support. Samples of 5,000 observations are drawn from the estimated densities using a deterministic approach in which the proportion of persons in the drawn sample with a specific income is equal to the density estimated at that income (up to rounding). The incomes in the sample drawn are linearly interpolated so as to avoid the clumping inherent in this deterministic approach.[21]

V. Monte Carlo study

V.i. Theoretical distributions

A Monte Carlo analysis of the properties of the kernel density estimator on grouped data is undertaken using quantile means computed from four distributions: the Log-normal, Dagum, Generalized Beta II, and a notional multimodal distribution.[22] The parameters chosen for the first three distributions are those resulting from a parametric density estimation exercise undertaken by Bandourian, McDonald and Turley (2002) 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).[23]

[20] The estimated density support is given by $[x_{\min} - h, x_{\max} + h]$ where $x_{\min}$ and $x_{\max}$ are the bottom and top quantile means.

[21] There are several approaches that can be used to draw samples from kernel density estimates. The simplest is the deterministic approach in which the sample is constructed by requiring that, up to rounding, the proportion of persons in the population with a specific income should be equal to the density associated with that income. This is the common approach in the literature on world poverty estimation. A second approach is the approach described in the text, for which we show results in the paper. A third approach involves directly drawing from the density using an algorithm that constructs a random variable whose p.d.f. (or alternative, C.D.F.) is precisely that estimated by KDE from the grouped data. Since the methods for generating synthetic populations lead to the same results, we only show findings based on the deterministic generator with interpolation.

[22] From each distribution, 200 samples with 1000 observations each are drawn. Quantile means are then computed from the samples, a (natural) log-transformation is applied to them, and kernel density estimation is undertaken.

[23] In the study undertaken by Bandourian, McDonald and Turley (2002) to assess the performance of different distributions in estimating real income survey data (using 82 datasets from 23 countries) the best-fitting two parameter distribution is the Weibull. 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.

I 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. The two modes of this distribution are produced by the large mass at the mean incomes of India and China, and a third, lower mode corresponds to the high average income of the richest nations. True densities of the four (log-income) distributions are shown in the Appendix .

V.ii. Summary statistics, density estimates and diagrams

A first question is whether the kernel density estimator from grouped data performs well in describing the underlying distribution through summary statistics (e.g., means, medians, standard deviation, and quintile means). The findings are reported in Table 1.[24]

Across the four distributions, it is found that the mean is systematically *overestimated* (by at most one fifth), while the median is estimated fairly well for all distributions. The standard deviation is substantially *underestimated* for the Log-normal, Dagum distribution and Generalized Beta II distribution (by as much as 65 percent) and is *overestimated* for the multimodal distribution (by as much as 24 percent). Some regularities can be observed in the second half of the table. For instance, the ratio of estimated quintile means to their true counterparts are always lower than 1 for the bottom two quantiles, and always higher than 1 for the upper two quantiles, for every distribution. The average income of the poorest 40 percent of the population is systematically and substantially *understated* whereas that of the richest 40 percent of the population is systematically and substantially *overstated*. 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*).

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.

In Graphs 1 and 2, the ability of the grouped-data KDE to estimate the quantile means of the underlying distribution is analyzed. This is of intrinsic interest because this analysis will shed

[24] Table 1 shows results for input data representing quintile means, S3 bandwidth, and the Quartic kernel. The results are broadly similar for the other Silverman bandwidths and the other five kernels.

further light on the performance of the technique in estimating the tails and it will demonstrate the technique's sensitivity to the smoothing parameter. Graph 1 shows histograms of the estimated quantile means (from the 200 draws) obtained from kernel density estimates with the optimal Silverman 3 bandwidth. The figure depicts the downward bias in the average income of the poorest population quintile and the upward bias in the average income of the richest population quintile.[25] Graph 2 repeats the exercise for the hybrid bandwidth. As mentioned earlier, the hybrid bandwidth chosen in this study is generally smaller than Silverman's optimal bandwidths.[26] Naturally, the under-smoothing induced by the small bandwidth value yields a much better approximation of the observed quintile means. Graph 2 shows that the histograms of the 200 fitted quintiles are centered at the observed quintile values for *all* quintiles. In terms of fitting the moments of the data, the hybrid bandwidth chosen in this study outperforms the optimal bandwidths. It should be mentioned, however, that this is solely an artifact of the low value of the hybrid bandwidth relative to the optimal bandwidths. If the hybrid bandwidth were larger than those, then this conclusion would be reversed since the hybrid bandwidth would lead to an even higher degree of smoothing.

Actual versus fitted densities and the size of the density bias along the support (for Log-normal data) are plotted for various distributions and kernel-bandwidth pairs in Graphs 3 and 4.[27] 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 first conclusion from these diagrams is that the choice of kernel does not seem to matter in terms of the visual impression created by the density estimate. This is, however, not surprising given that canonical bandwidths are used to ensure that each kernel-bandwidth combination achieves the same amount of smoothing. The second conclusion is that there are distortions in the estimated density *at every point* along the income support (with the exception of two crossing points where the bias is zero). The estimated density is biased *upwards* in the tails of the distribution, and *downwards* in the middle of the distribution.

It is easiest to see the likely consequences of these density biases on poverty for the poverty headcount ratio. Suppose that the poverty line falls below or at the first crossing point. Then the poverty headcount ratio will be over-estimated. As the poverty line moves rightward on the support, the extent of over-estimation will fall until it becomes zero when the poverty line is such that the density over-estimation in the tail is perfectly offset by the under-estimation in the middle of the density. As the poverty line moves rightward, the headcount ratio will remain under-estimated until further tradeoffs are encountered.

[25] It is worthwhile mentioning here that the distributions of the (fitted) quantile means move in the expected direction (i.e., towards being centered on the expectation of the quintile) as the number of data points available for analysis increases (from quintile means to decile means and ventile means). However, it does not do so monotonically.

[26] For example, for the Generalized Beta II distribution, the hybrid bandwidth is approximately 1.5 times smaller than the Silverman 3 rule-of-thumb bandwidth in the case of quintile means.

[27] 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 200 draws (Graphs 3 and 4).

In contrast, it is more difficult to foresee the poverty biases associated with this technique in the case of the multimodal distribution (Graph 5). The left and right panels shows densities estimated from grouped data overlaid with the true density (left), and a plot of the difference between the estimated density and the true density (right). It is observed that the extent to which salient features of the underlying density are replicated by KDE critically depends on the choice of bandwidth and kernel. The Gaussian kernel in conjunction with Silverman's optimal bandwidth S3 produces a largely over-smoothed density that conceals the multiple modes of the distribution. In contrast, the (lower) 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 should be noted that visual illustrations of multimodal distributions obtained through density smoothing from grouped data might be misleading. Distortions should be expected in the resulting density estimates in both directions (over- and under-estimation) and along the entire support.

The four panels in Graph 6 reveal the pitfalls of using a non data-driven, fixed bandwidth for analysis, as opposed to data-driven bandwidths. In some datasets, the hybrid bandwidth might fall close to an optimal bandwidth. The first panel shows that this could be the case.[28] The two curves, corresponding to the S1 and hybrid bandwidths for Dagum, although different, show that the hybrid bandwidth tends to under-smooth. More importantly, the size of 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 second panel (Graph 6) shows the effect of changing the kernel and keeping the bandwidth fixed. It reveals the consequences on density diagrams of changing the amount of smoothing. The estimated density corresponding to the hybrid bandwidth is now concentrated at the quintile means, and is zero between the extreme modes and the central mode. Since the hybrid bandwidth is too small (and the kernel has finite support), there is no information from adjacent points because those points do not fall in the window of neighboring points in which the kernel density estimator seeks information. Hence, at those points of estimation, the estimated density is zero. The panels in this section lead to the following conclusions regarding the hybrid bandwidth: (a) the hybrid bandwidth might lead to the same level of smoothing as an optimal bandwidth, but it will do so only by chance; (b) otherwise, the hybrid bandwidth may lead to substantial under- or over-smoothing of the estimated curve (in conjunction with some kernels), which renders diagrams of that curve difficult to interpret; (c) the hybrid bandwidth can be used for some purposes (e.g., fitting the observed data well) but doing so might render it less appropriate for other purposes (e.g., producing accurate diagrams of the underlying density).

[28] In order to avoid the difficulties that arise in keeping the points of estimation fixed, these panels superimpose histograms of the true density with kernel density estimates from grouped data computed directly from the universes. Therefore, there is no Monte Carlo exercise involved in Graph 6.

V.iii. Poverty

Poverty estimates are reported for two different poverty lines in Tables 2 and 3. The poverty lines are set at the median multiplied by factors equal to 0.25 and 1.75, respectively. From the previous analysis, 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 located at income levels at which the biases in the estimated density do not cancel out (e.g., in the far left tail). 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 others that examine the level of inequality among the poor. The indicators considered are: the poverty headcount ratio, the poverty gap, the squared poverty gap, and the distributionally-sensitive FGT (3) and FGT (4) indices.

It is found that the poverty headcount ratio is over-estimated for the lower poverty line, and underestimated for the higher poverty line. For input data representing quintile means, the poverty headcount ratio is overestimated by a factor of 1.17 of its true counterpart (for the Log-normal distribution) corresponding to a bias of 2 percentage points (Table 2). For input data representing decile means, the biases rise up to a factor of 1.28 corresponding to an overestimation of 3.4 percentage points. The biases are slightly lower for ventile means. The FGT indicators of the depth of poverty (with parameter values between 1 and 4) are more substantially underestimated by quintile data than they are for decile and ventile data. The biases appear to rise with the distributional sensitivity of the FGT indicator. The situation is reversed for the higher 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) when data from the other distributions is utilized.

Table 4 summarizes the results for a wider range of poverty lines using all the distributions. This table focuses, however, only on the poverty headcount ratio (as it is the poverty indicator with the widest application). As before, it is observed that generally the extent of poverty is overestimated for lower poverty lines, is estimated correctly for poverty lines that are close to the population median (that is, in regions where the positive density biases cancel out with the negative density biases) and is underestimated for higher poverty lines. The behavior of biases associated with the multimodal distribution is somewhat different, with a pronounced underestimation for poverty lines equal to and higher than the median. In particular, at the median, the poverty headcount ratio is underestimated by almost 11 percentage points (because the estimated density "misses" the first mode of the distribution) whilst at ½ of the median, it is overestimated by almost 9 percentage points due to positive density biases at the left end of the support. The biases (expressed in percentage points) are plotted against the size of the poverty line in Graph 7.

The effect of alternative bandwidths and kernels is shown in Tables 5-6, where poverty estimates are computed for lower poverty lines and the FGT indicators with parameter values 0 to 2. 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. In Table 5, the degree of distortion inherent in choosing a non data-driven bandwidth for a distribution with multiple modes becomes apparent: there are substantial downward biases associated with this bandwidth for *all* of the poverty indicators considered. The Silverman bandwidths only occasionally do better, but still the magnitude of the biases associated with this technique is very large. In Table 6, we report the findings for different kernels (keeping the bandwidth fixed at the hybrid value and using quintile means). The effect of the kernel is, in some cases, substantial (again, the degree of smoothing achieved with a fixed bandwidth is different across different kernels). For example, for the Dagum distribution, the estimates of the poverty headcount ratio switch from being biased upwards by 11 percent (Gaussian kernel) to being biased downwards by 9 percent (Triweight kernel), as shown in Table 6.

It is difficult to describe the magnitude and sign of biases in poverty indicators through statements applicable across a wide range of possible income distributions and parameters of analysis. However, 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 unsurprising, given the nonparametric approach involved), as well as with the bandwidth, weighting function and number of quantile means available for analysis. For a range of unimodal distributions, the poverty headcount ratio is overestimated for lower poverty lines, is accurately estimated at poverty lines close to the population median, and is underestimated for higher poverty lines. Its biases are harder to predict in the case of multimodal distributions, where the positioning of the poverty line relative to the modes, and the extent of smoothing, determine the sign and size of the bias.

VI. 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. The $1/day and $2/day international poverty lines are used[29], along with a capability, nutritionally anchored poverty line developed by Reddy, Visaria and Asali, 2006). Results are presented in Tables 7 and 8.

For the $1/day poverty line, it is observed that 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 available for analysis (Table 7). 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 under-statement 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 the bias of

[29] 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).

poverty estimates does *not* vary monotonically with the number of quantile means analyzed.

Table 8 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, which explains the higher relative accuracy of the estimator). The choice of the bandwidth, however, appears to have 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). It is apparent that the biases for any given bandwidth vary 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.[30] All the estimates in Table 8 indicate that KDE on quintile means yields poverty estimates that are higher than their true counterparts. This can be explained, in light of the Monte Carlo evidence, by the relative position of the poverty lines vis-à-vis the survey median.

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 (Graph 8). 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 is associated with some degree of over-smoothing 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 kernel density estimate. Panels (2) and (3) for Nicaragua show the effect of changing the kernel in two environments: the first uses canonical bandwidths and the next keeps the bandwidth fixed across kernels (hybrid case). In the former case, the amount of smoothing remains unchanged across density estimates; in the latter, it changes. Panel 2 demonstrates that keeping the bandwidth fixed across kernels may 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, since the canonical Sheather-Jones bandwidth is used. Finally, the last panel proves yet again that the density estimator (in this case obtained with the Quartic kernel and the S1 optimal bandwidth) leads to positive density biases in the left tail of the distribution, negative biases in the center of the distribution, and positive biases in the right tail of the distribution. As seen previously, these distortions take place at every point along the density support and have important consequences for the estimation of poverty using alternative indicators. 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.

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

VII. Global poverty analysis

In this section, I assess the sensitivity of world poverty estimates to parameters of the kernel density estimation procedure. The focus is on the effect of the smoothing parameter on the poverty rates and headcounts of the developing world as a whole. 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.[31] Income averages for (five) population quintiles were subsequently obtained for each country using the per capita GDP (at PPP) from the World Development Indicators online database (2006). Finally, kernel density estimation was undertaken on each country's five income averages. The resulting density estimates were subsequently aggregated in a world KDE-based distribution of income.[32]

I consider the following data-dependent bandwidths that have been proposed in the literature: 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 (all of which were discussed in Section IV). To compare these results with global poverty rates proposed in the literature, I also report the results for the hybrid bandwidth (which is equal to 0.39 and happens to be the optimal S3 bandwidth for China, but is kept constant across countries). Countries other than China with similar dispersions of their data would also have similar S3 values associated with them, but for these other countries the hybrid bandwidth is unlikely to satisfy a common optimality criterion (except by chance). The results are reported for the Gaussian kernel. [33]

In Tables 9-10, the world poverty headcount ratio and the aggregate headcount for five international poverty lines, ranging between $1/day and $4/day, are reported.[34] The rates and headcounts presented in Table 9 demonstrate the lack of robustness of global poverty rates to changes in the value of the bandwidth even when the bandwidth is chosen according to an optimality criterion. In both years, the $1/day poverty rates are most sensitive to changes in the bandwidth (since the $1/day falls in the left tail of the regional income distributions, where poverty is likely to be severely overestimated and small changes in the bandwidth may significantly alter the estimated density). Furthermore, estimated poverty rates vary more across bandwidths for the lowest poverty lines considered. 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 1.6 when it is not (given that it is likely to result in substantial overestimation of the density in

[31] Income shares were selected for the years 1990, 2000 or the closest year in which they were available in the database.

[32] To compare our results to Sala-i-Martin's 2006 study, we also undertook the same analysis for the year 2000 in 134 developed and developing countries. We obtain similar poverty rates as those reported by the author. 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.

[33] The results were largely similar for the Epanechnikov kernel as long as we used the canonical bandwidth.

[34] Results are shown for poverty lines up to $4/day since the country distributions were scaled using the per capita GDP, which might have lead to an overestimation of the quantile means. For more on the use of survey versus National-Accounts based estimates of per capita income, see Deaton (2003, 2005). It is important to stress, therefore, that the global rates and headcounts presented here should not be interpreted as authoritative.

the left tail). The headcount ratios vary to a lesser degree as the poverty line increases and they are almost equal regardless of the bandwidth 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. In terms of numbers of poor people, this variation translates into a range of 162 to 278 million people. To put these numbers in perspective, under- or over-counting the "$2/day poor" by 278 million individuals (in 2000) would represent an error of 10 percent (based on the $2/day global headcount for 2001 of Chen and Ravallion, 2004). Similarly, under- or over-counting the "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).

How does this range of variation inform us on the trend in world poverty between 1990 and 2000? As reported in Table 11, the fall in the $1.5/day and $2/day poverty rates ranges between 7 percent and 18 percent (corresponding to the oversmoothed and S3 bandwidths). The number of people who were lifted from $1/day poverty between 1990 and 2000 ranges between 19 million and 38 million (oversmoothed and hybrid bandwidth), whereas the reduction in $1.5/day poverty ranges between 45 and 92 million (oversmoothed and S3). It should be noted that a reduction in the number of '$1.5/day poor' by 45 million is *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 '$1 /day poor' 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. Importantly, all estimates are consistent with a *reduction* in world poverty. While the underlying cause of this finding may be kernel density estimation itself, the use of National-Accounts based per capita income estimates, or the composition of the sample, it should be stressed that it is at odds with the World Bank's documented *increase* in the number of '$2/day poor' of 81 million over the same period (Chen and Ravallion, 2004).

VIII. Conclusions

Recent influential studies of national, regional, and global poverty employ kernel density estimation techniques on grouped data to analyze poverty and to describe features of the underlying income distributions (e.g., Sala-i-Martin (2002a, 2002b, 2006), Ackland et al (2004), and Fuentes (2005)). This method is used because of the lack of availability or the difficulty in obtaining access to unit data from representative household surveys for all countries and years of interest (as is the case with large countries such as China and India). Grouped data often takes the form of income averages for a small number of population quantiles, but are sometimes derived from income shares scaled with an estimate of per capita income. This data, despite its limited informational content, is also used to assess long-term trends (in poverty and inequality) since household unit data are no longer available, but summary statistics may have been published and are therefore accessible to researchers.

In this paper, I analyzed the performance of the kernel density smoothing technique in estimating income distributions from grouped data. There are reasons to believe that kernel density estimation techniques may be inappropriate for this data structure. Several income averages of population groups are less informative than a large sample drawn from the underlying distribution. However, they are more informative than a small sample drawn from the underlying distribution due to the nature of the grouping process. The income averages are a collection of trimmed means that carry useful information about the unit data from which they have been computed. Their joint density (of linear functions of order statistics) is prohibitively difficult to derive analytically, and the properties of the kernel density estimator on grouped data need be assessed through Monte Carlo simulations. In this study, I have considered several plausible income distributions (Log-normal, Dagum and Generalized Beta II), as well as a distribution with multiple modes (corresponding to the 2004 population-weighted world distribution of per capita income). Furthermore, unit and grouped data from three household surveys (Nicaragua, Tanzania and Vietnam) have been used to compare KDE-based poverty estimates from grouped data with their survey counterparts.

The biases resulting from the application of this technique depend on the smoothing parameter, the kernel, the number of data-points analyzed and, naturally, the data generating process. The average income of the poorer population quantiles is *overstated* by the technique, while the average income of the richer quantiles 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 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. In general, the bandwidth has an important effect on the accuracy of poverty estimates. Kernel density estimation on grouped data can also give rise to misleading diagrams of the underlying distributions as these too are sensitive to the choice of parameters.

To assess the robustness of kernel density estimation methods in global poverty analysis, KDE-based global poverty rates and headcounts have been computed using income shares for 94 developing countries. I show that kernel density-based headcount ratios for poverty lines such as \$1/day, \$1.5/day and \$2/day vary by a factor of 1.8 for a range of bandwidths that have been recommended and used in the literature. The difference between the highest and the lowest estimate of the number of '\$2/day' poor is 136 million in 1990, and 278 million in 2000.

The findings of this study give rise to serious concern about the validity and robustness of poverty analysis based on kernel density estimation on grouped data. The magnitude and direction of biases identified in this study can, however, serve as a framework of interpretation to researchers who wish to employ nonparametric smoothing techniques on grouped data (especially if the data can be assumed to be drawn from a unimodal distribution). The applied researcher should exercise caution in employing standard kernel density estimation methods on grouped data and should as a matter of routine assess the robustness of their results to changes in the bandwidth and kernel. Finally, given the empirical regularities uncovered in this study for plausible income distributions, it is possible that the performance of the standard kernel density estimator on grouped data could be improved upon, and these improvements are a subject for future research.

References

1. Ackland, R., Dowrick, S. and Freyens, B. (2004) “Measuring Global Poverty: Why PPP Methods Matter”, Paper presented at the Conference of the International Association for Research in Income and Wealth (August 26), Cork, Ireland.
2. Aziz, J. and Duenwald, C. (2001) “China's Provincial Growth Dynamics”, Economics Working Paper #0012004 at the Washington University Economics Department, and IMF Working Paper No. 01/3.
3. 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, GA.
4. Bandourian, R., McDonald, J.B. and Turley, R.S. (2002) “A Comparison of Parametric Models of Income Distribution across Countries and Over Time”, Department of Economics, Birgham Young University mimeograph.
5. Bhalla, S. (2002) Imagine There’s No Country: Poverty, Inequality, and Growth in the Era of Globalization, Institute for International Economics, Washington, DC.
6. Berry, A., Bourguignon, F. and Morrisson, Ch. (1983) “Changes in the World Distribution of Income between 1950 and 1977”, *Economic Journal*, Vol. 93, pp. 331-350.
7. Bianchi, M. (1997) “Testing for Convergence: Evidence from Non-Parametric Multimodality Tests”, *Journal of Applied Econometrics*, Vol. 12(4), pp. 393-409.
8. Bourguignon, F. and Morrisson, C. (2002) “Inequality Among World Citizens: 1820-1992”, *American Economic Review*, Vol. 92(4), pp 727-744.
9. Chen, S. and Ravallion, M. (2002) “How Did the World's Poorest Fare in the 1990s?”, *Review of Income and Wealth*, Vol. 47(3), pp 283-300.
10. Chen, S. and Ravallion, M. (2004) “How Have the World's Poorest Fared Since the Early 1980s?”, World Bank Development Research Group Working Paper no. 3341.
11. Chen, S. and Ravallion, M. (2005) “China’s (Uneven) Progress Against Poverty”, forthcoming, *Journal of Development Economics.*
12. Deaton, A. (2003) “How to Monitor Poverty for Millennium Development Goals”, *Journal of Human Development*, Vol. 4(3), pp. 353-378.
13. Deaton, A. (2005), “Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)”, *Review of Economics and Statistics*, Vol. 87(1), pp. 1-19.
14. 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.
15. Fuentes, R. (2005) “Poverty, Pro-Poor Growth and Simulated Inequality Reduction”, Human Development Report Office Occasional Paper No. 11.
16. Grosh, M.E. and Nafziger, E.W. (1986) “The Computation of World Income Distribution”, *Economic Development and Cultural Change*”, Vol. 35, pp. 347-359.
17. Jann, B. (2005) “Univariate Kernel Density Estimation”, Boston College Department of Economics, Statistical Software Component No. S 456410.
18. Jones, M.C., Marron J.S. and Sheather, J.S. (1996) “A Brief Survey of Bandwidth Selection for Density Estimation”, *Journal of the American Statistical Association*, Vol. 91, pp. 401-407.

19. Jones, C.I. (2002) "On the Evolution of the World Income Distribution", *Journal of Economic Perspectives*, Vol. 11(3), pp 19-36.
20. Kakwani, N.C. and Son, H.H. (2006) "New Global Poverty Counts", UNDP International Poverty Center Working Paper No. 29.
21. Kakwani, N.C. (1980) "On A Class of Poverty Measures", *Econometrica*, Vol. 48(2), pp. 437-446.
22. Kakwani, N.C. and Podder, N. (1976) "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations", *Econometrica*, Vol. 44(1), pp 137-149.
23. Korzeniewick, R.P. and Moran, T. (1997) "World-Economic Trends in the Distribution of Income, 1965-1992", *American Journal of Sociology*, Vol. 102, pp. 1000-1039.
24. Marron, J.S. and Nolan, D. (1988) "Canonical Kernels for Density Estimation", *Statistics and Probability Letters*, Vol. 7, pp 195-199.
25. Marron, J.S., and Ruppert, D. (1994) "Transformations to Reduce Boundary Bias in Kernel Density Estimation", *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 56(4), pp 653-671.
26. Mason, D.M. (1981) "Asymptotic Normality of Linear Combinations of Order Statistics with a Smooth Score Function", *The Annals of Statistics*, Vol. 9(4), pp. 899-908.
27. Milanovic, B. (2002) "True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone", *Economic Journal*, Vol. 112, pp 51-92.
28. Milanovic, B. (2005) Worlds Apart: Measuring International and Global Inequality, Princeton University Press.
29. Minoiu, C. and Reddy, S. (2006) "The Assessment of Poverty and Inequality through Parametric Estimation of Lorenz Curves: An Evaluation", Columbia University mimeograph.
30. Moore, D.S. (1968) "An Elementary Proof of Asymptotic Normality of Linear Functions of Order Statistics", *The Annals of Mathematical Statistics*, Vol. 39(1), pp. 263-265.
31. Mosteller, F. (1946) "On Some Useful *Inefficient* Statistics", *The Annals of Mathematical Statistics*, Vol. 17(4), pp. 377-408.
32. Pittau, M.G. (2005) "Fitting Regional Income Distributions in the European Union", *Oxford Bulletin of Economics and Statistics*, Vol. 67(2), pp. 135-161.
33. Pittau, M.G. and Zelli, R. (2006) "Empirical Evidence of Income Dynamics across EU Regions", *Journal of Applied Econometrics*, Vol. 21, pp. 605-628.
34. Pritchett, L. (2006) "Who Is *Not* Poor? Dreaming of a World Truly Free of Poverty", forthcoming, *The World Bank Research Observer*.
35. Quah, D.T. (1997) "Empirics for Growth and Distribution: Polarization, Stratification and Convergence Clubs", *Journal of Economic Growth*, Vol. 2(1), pp 27-59
36. Quah, D.T. (1996) "Twin Peaks: Growth and Convergence in Models of Distribution Dynamics", *Economic Journal*, Vol. 106, pp. 1045-1055.
37. Reddy, S. and Minoiu, C. (2006) "Has World Poverty *Really* Fallen?" United Nations Department of Economic and Social Affairs Working Paper No. 36.
38. Reddy, S., Visaria, S. and Asali, M. (2006) "Inter-country Comparisons of Poverty Based on a Capability Approach: An Empirical Exercise", United Nations Development Programme International Poverty Center Working Paper No. 27.
39. Reddy, S. and Pogge, T. (2006) "How *Not* To Count the Poor", forthcoming in Anand, S., P. Segal and J. Stiglitz, eds.: *Measuring Global Poverty*, Oxford University Press.

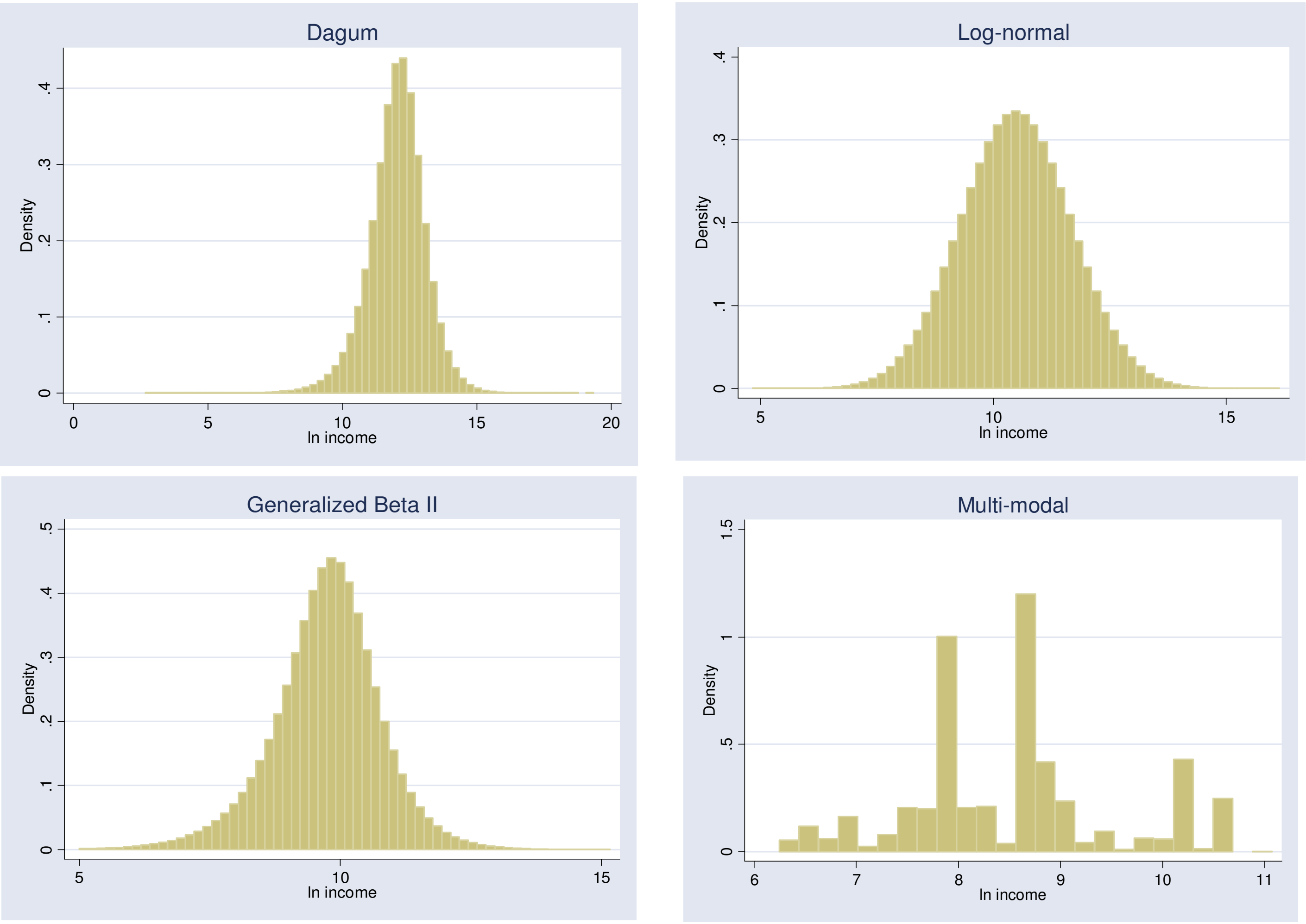
40. Sala-i-Martin, X. (2002a) “The world distribution of income (estimated from individual country distributions)”, National Bureau of Economic Research Working Paper No. 8933.
41. Sala-i-Martin, X. (2002b) “The ‘disturbing’ rise of world income inequality”, National Bureau of Economic Research Working Paper No. 8904.
42. Sala-i-Martin, X. (2006) “The World Distribution of Income: Falling Poverty and Convergence, Period”, *Quarterly Journal of Economics*, Volume 121, No. 2, pp. 351-397.
43. Sheather, S.J. and Jones, M.C. (1991) “A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation”, *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. 53(3), pp. 683-690.
44. Siddiqui, M. M. and Butler, C. (1969) “Asymptotic Joint Distribution of Linear Systematic Statistics from Multivariate Distributions”, *Journal of the American Statistical Association*, Vol. 64(325), pp. 300-305.
45. Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability 26, Chapman & Hall/CRC.
46. Stigler, S.M. (1974) “Linear Functions of Order Statistics with Smooth Weight Functions”, *The Annals of Statistics*, Vol. 2(4), pp. 676-693.
47. Stigler, S.M. (1973) “The Asymptotic Distribution of the Trimmed Mean”, *The Annals of Statistics*, Vol. 1(3), pp. 472-477.
48. UNU/WIDER (2005) World Inequality Database V. 2.0a (June).
49. Villasenor, J.A. and Arnold, B.C. (1989) “Elliptical Lorenz Curves”, *Journal of Econometrics*, Vol. 40, pp. 327-338.
50. Yatchew, A. (1998) “Nonparametric Regression Techniques in Economics”, *Journal of Economic Literature*, Vol. 36(2), pp. 669-721.
51. 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.
52. Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. Chapman and Hall: London.
53. World Development Indicators online (2006) The World Bank Group, Washington, DC.
54. World Bank (1994) A Database on Poverty and Growth in India, Poverty and Human Resources Division Policy, Research Department, The World Bank, Washington D.C.
55. Wu, X. and Perloff, J. (2003) “Calculation of Maximum Entropy Densities with Application to Income Distributions”, *Journal of Econometrics*, Vol. 115, pp. 347-354.
56. Wu, X. and Perloff, J. (2005) “China’s Income Distributions: 1985-2001”, *Review of Economics and Statistics*, Vol. 87, pp. 763-775.

Appendix

Data sources (Household surveys)

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).
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).
Tanzania: The 2000-01 Household Budget Survey contains information on per capita consumption for 22,176 households. Source: National Bureau of Statistics, Tanzania, 2002.

True distributions used in the Monte Carlo analysis

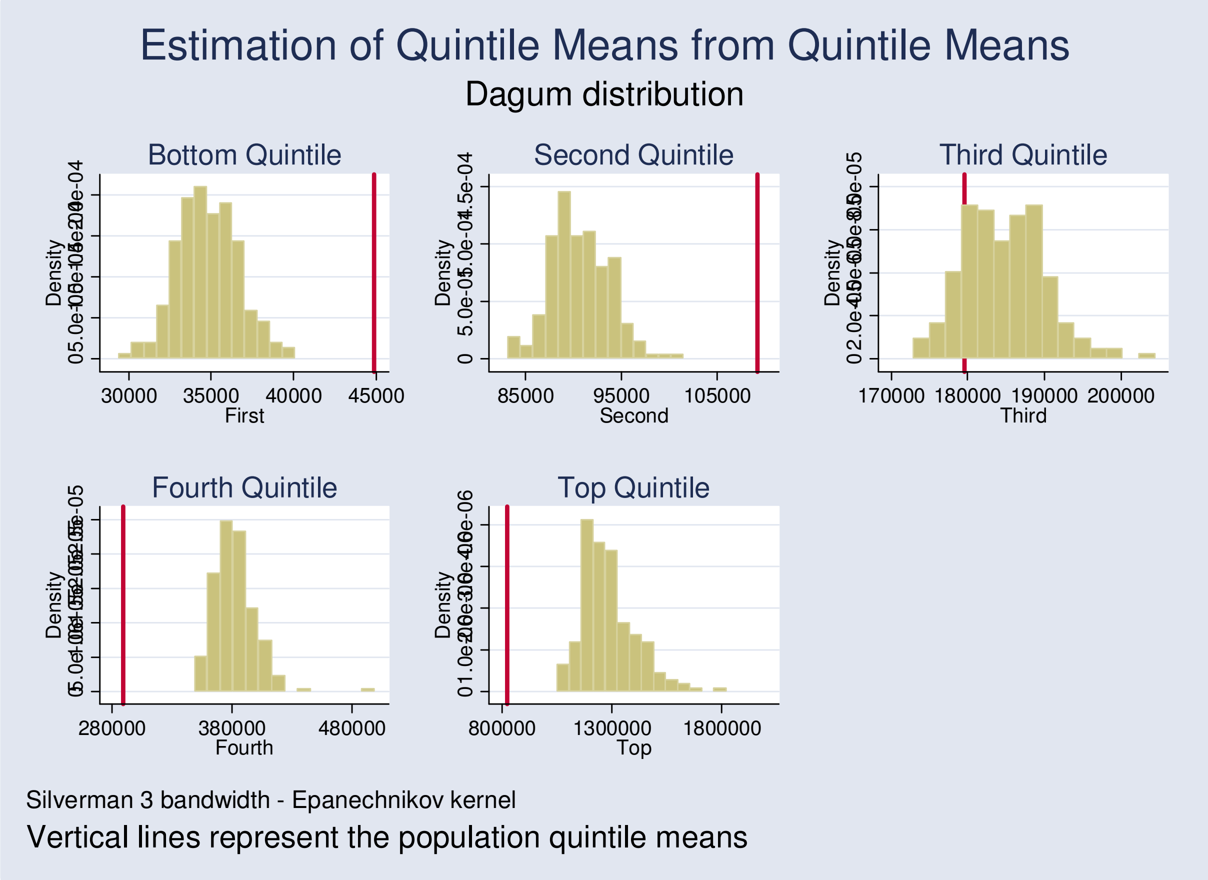


Monte Carlo findings (Summary statistics and diagrams)

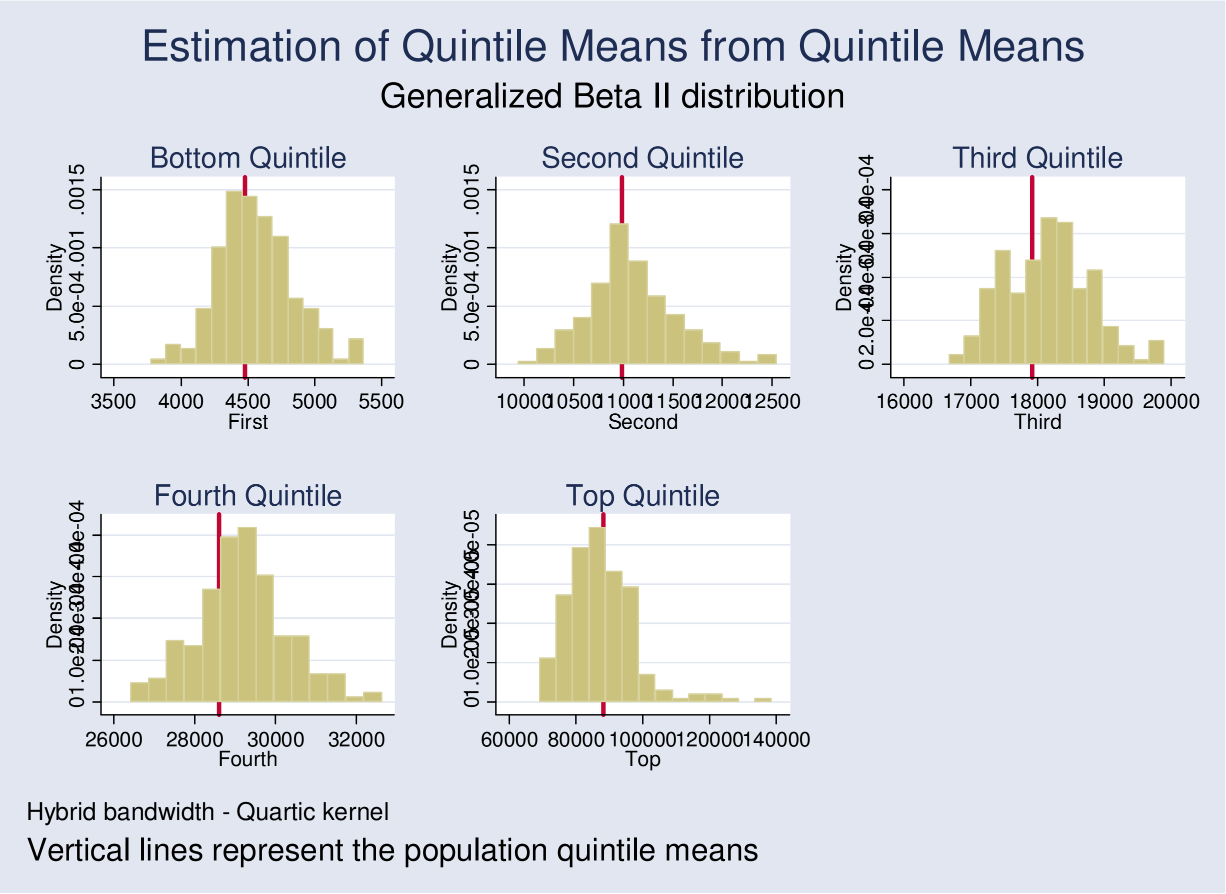
Table 1. Summary statistics. KDE parameters: Quartic kernel, S3 bandwidth. Input data: Quintile means. All figures represent the ratio between the estimated quantity and its true value.

	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

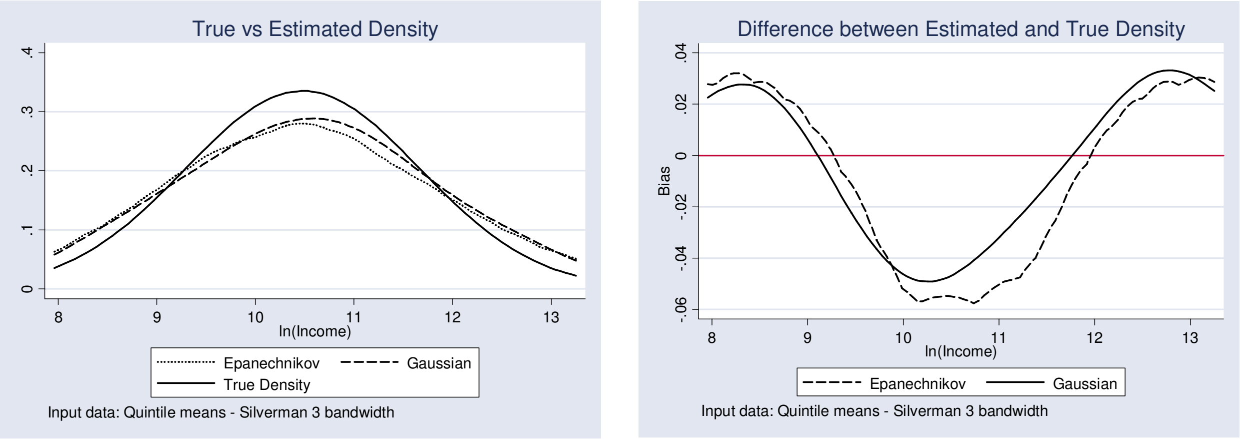
Graph 1. Histograms of 200 quintile means estimated using KDE on quintile means (Dagum distribution). KDE parameters: Epanechnikov kernel, Silverman 3 bandwidth.



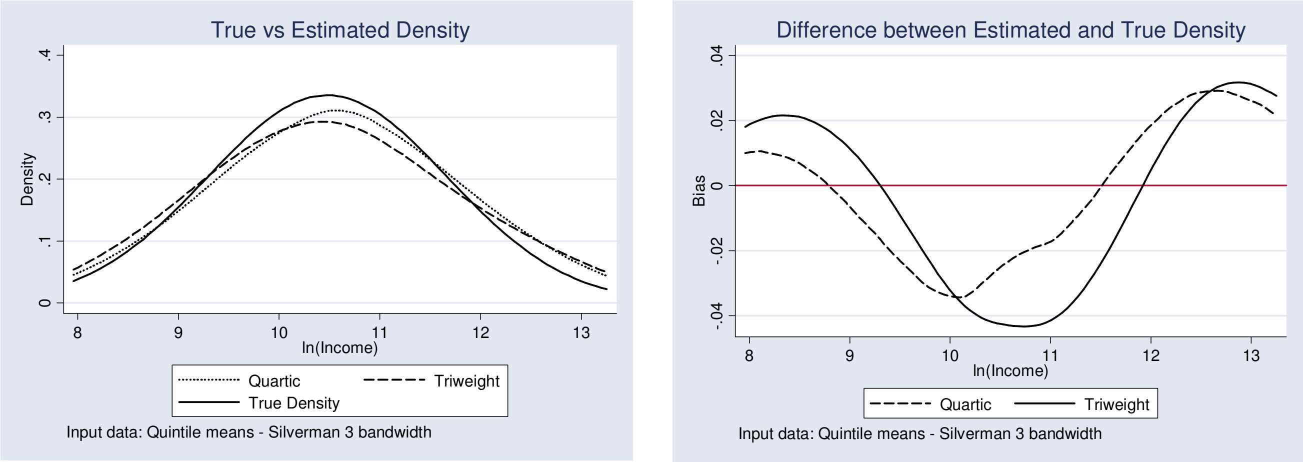
Graph 2. Histograms of 200 quintile means estimated using KDE on quintile means (Generalized Beta II distribution). KDE parameters: Quartic kernel, hybrid bandwidth.



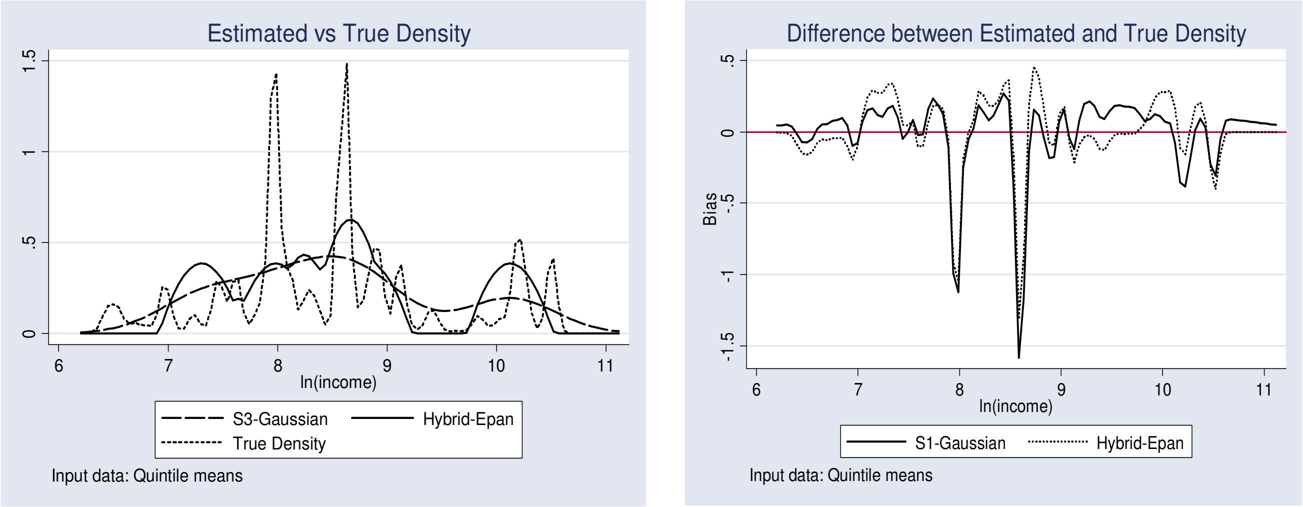
Graph 3. Bias of estimated density at fixed points of estimation. Log-normal distribution. Input: Quintile means. (KDE parameters: Gaussian and Epanechnikov kernels, Silverman 3 bandwidth)



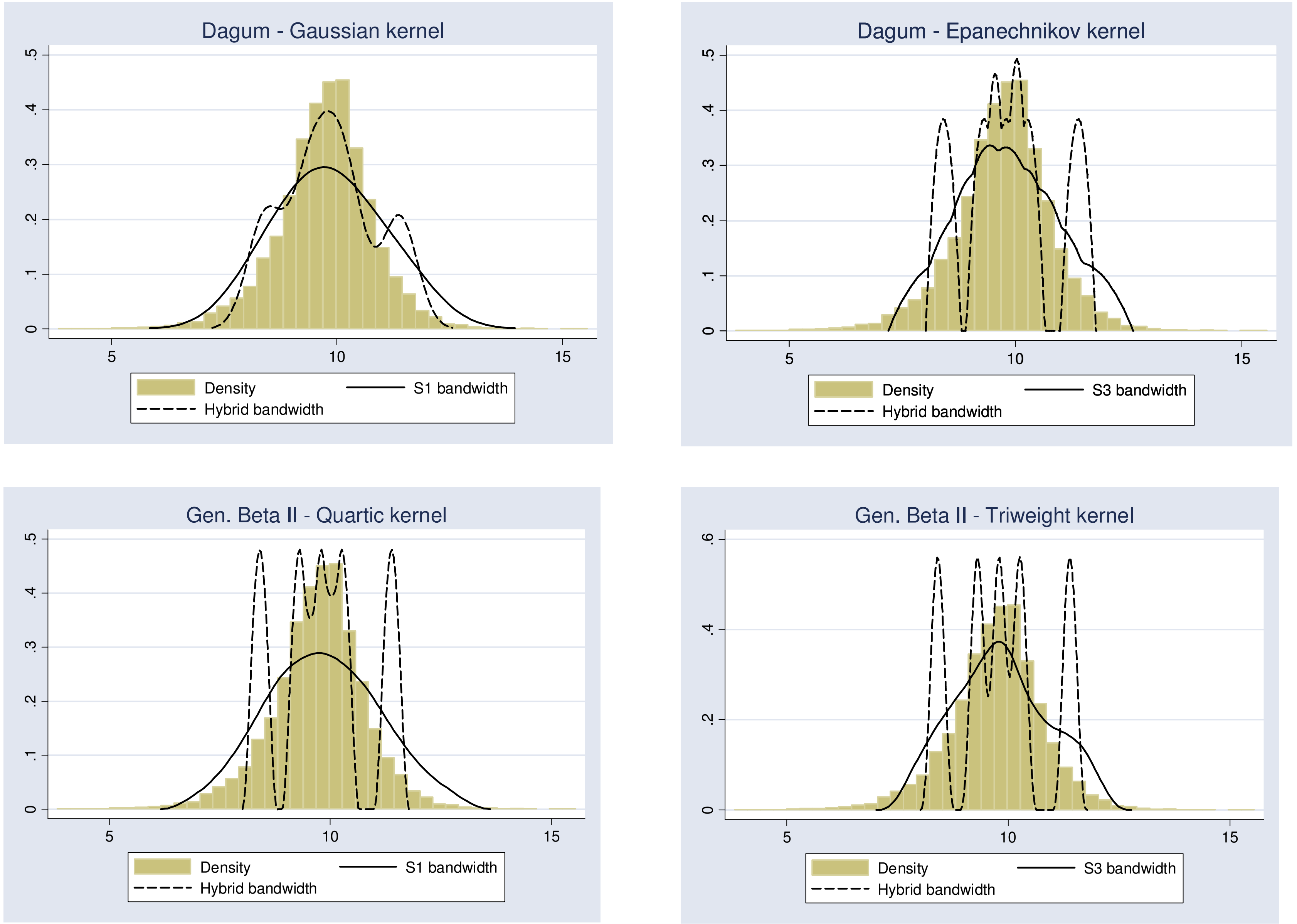
Graph 4. Bias of estimated density at fixed points of estimation. Log-normal distribution. Input: Quintile means. (KDE parameters: Quartic and Triweight kernels, Silverman 3 bandwidth)



Graph 5. Bias of estimated density at fixed points of estimation. Multimodal distribution. Input: Quintile means. (KDE parameters: Various kernels, various bandwidths)



Graph 6. Bias of estimated density. Dagum and Generalized Beta II Distributions. Input: Quintile means.



Monte Carlo findings (Poverty biases)

Table 2. Poverty line: **0.25** x median [KDE parameters: S3 bandwidth, Epanechnikov kernel]

				Input data:	
Poverty indicator:	**Distribution**	**True quantity**	**Quintiles**	**Deciles**	**Ventiles**
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

Table 3. Poverty line: **1.75** x median [KDE parameters: S3 bandwidth, Epanechnikov kernel]

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

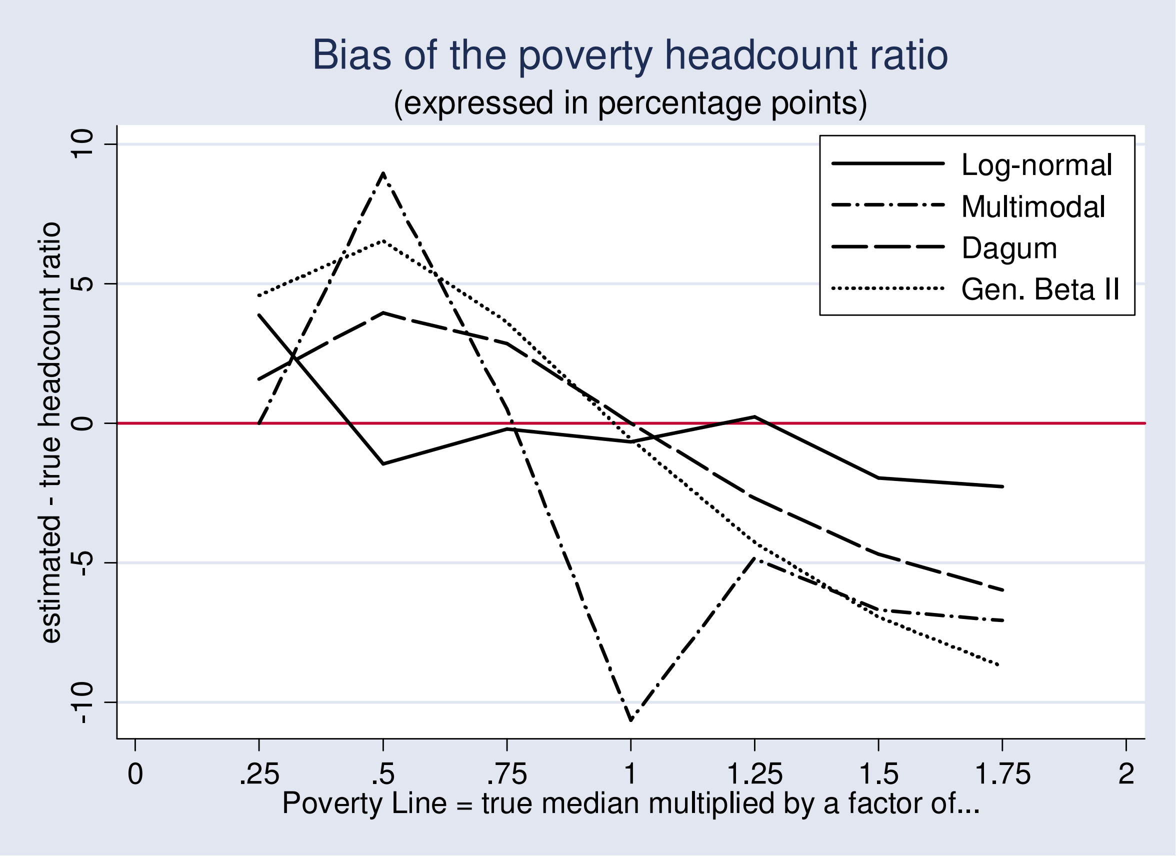
Note: Figures in the last three panels represent the ratio between the estimated quantity and its true counterpart.

Table 4. Summary of results: Bias in the poverty headcount ratio for various distributions and kernel-bandwidth pairs. Input data: Quintile means.

Poverty line is equal to the population median multiplied by:	**Log-normal**[35]	**Dagum**[36]	**Generalized Beta II**[37]	**Multimodal**[38]
0.25	1.33	1.17	1.49	1.00
0.50	0.95	1.17	1.28	1.47
0.75	0.99	1.08	1.10	1.01
1.00	0.99	1.00	0.99	0.84
1.25	1.00	0.96	0.93	0.93
1.50	0.96	0.93	0.90	0.92
1.75	0.97	0.92	0.88	0.93

Note: The bias is expressed as the ratio between fitted values and the true headcount ratio.

Graph 7. Summary of results: Bias in the poverty headcount ratio for various distributions and kernel-bandwidth pairs. Input data: Quintile means.



[35] KDE parameters: Hybrid bandwidth and Triangular kernel.

[36] KDE parameters: S2 bandwidth and Epanechnikov kernel.

[37] KDE parameters: S1 bandwidth and Quartic kernel.

[38] KDE parameters: S3 bandwidth and Triweight kernel.

Monte Carlo findings (Poverty biases)

Table 5. Poverty line: **0.25** x median [KDE parameters: Triweight kernel, input data: Quintile means]

			Bandwidth:			
Poverty indicator:	**Distribution**	**True quantity**	**S1**	**S2**	**S3**	**Hybrid**
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 6. Poverty line: **0.5** x median [KDE parameters: Hybrid bandwidth, input data: Quintile means]

			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.

Country studies (Poverty biases)

Table 7. KDE parameters: Epanechnikov kernel, Silverman 3 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

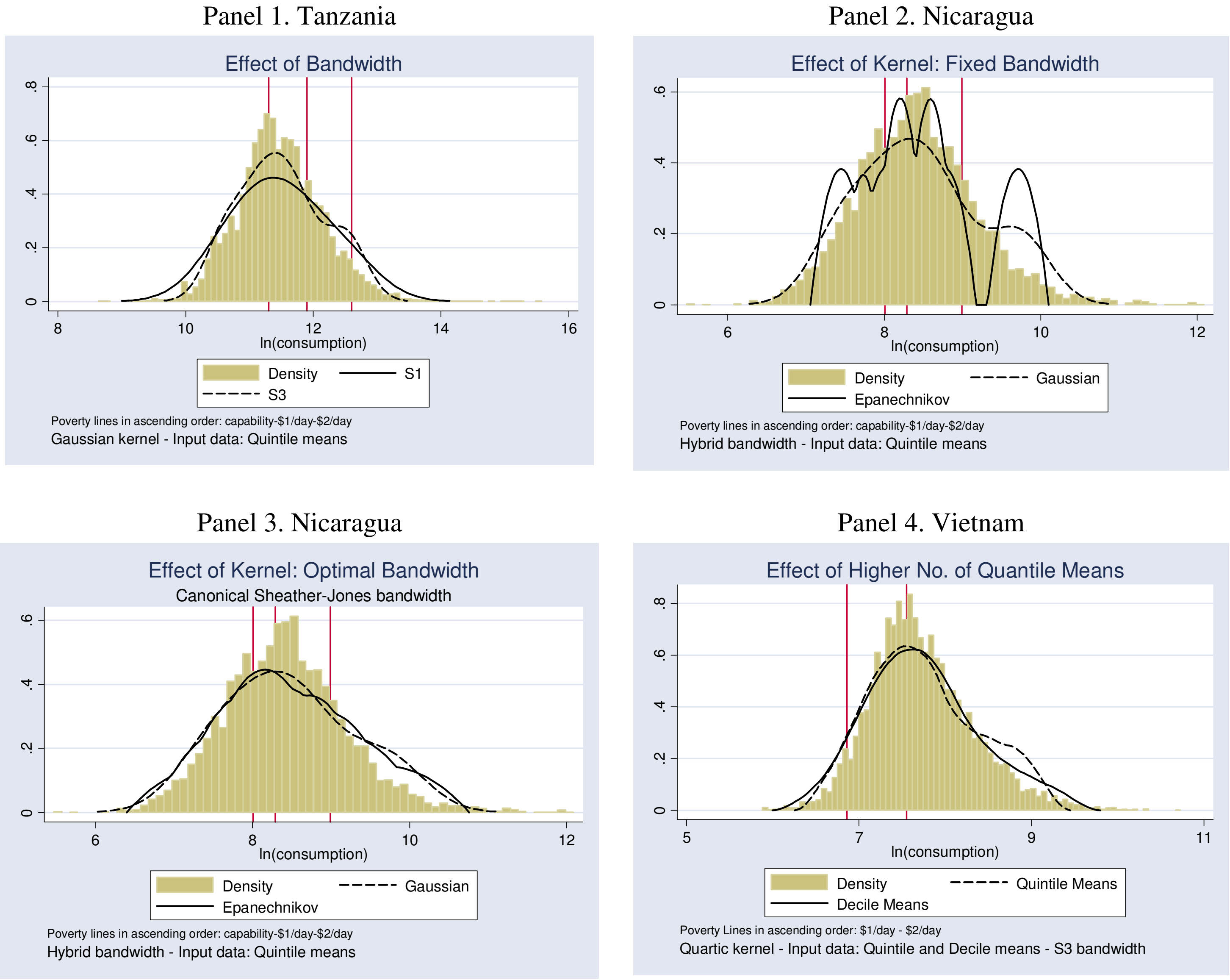
Note: Figures in the relevant panels represent the ratio between the estimated quantity and its survey counterpart.

Table 8. KDE parameters: Gaussian kernel, Quintile means. 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
FGT(4)	Vietnam	1.02	2.31	1.41	**1.21**	1.54	1.69
	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.

Graph 8. Survey-based and grouped data KDE-based density estimates.


Panel 1. Tanzania

Poverty lines in ascending order: capability-$1/day-$2/day
Gaussian kernel - Input data: Quintile means

Panel 2. Nicaragua

Poverty lines in ascending order: capability-$1/day-$2/day
Hybrid bandwidth - Input data: Quintile means

Panel 3. Nicaragua

Poverty lines in ascending order: capability-$1/day-$2/day
Hybrid bandwidth - Input data: Quintile means

Panel 4. Vietnam

Poverty Lines in ascending order: $1/day - $2/day
Quartic kernel - Input data: Quintile and Decile means - S3 bandwidth

Sensitivity analysis of global poverty

Table 9. Extent of global poverty. Indicator: the headcount ratio (%)

Bandwidth→	S3	Overs-smoothed	Variant of S3	Sheather-Jones	Direct plug-in	Hybrid	Ratio between 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 10. Extent of global poverty. Indicator: the aggregate poverty headcount (millions)

Bandwidth→	S3	Overs-smoothed	Variant of S3	Sheather-Jones	Direct plug-in	Hybrid	Difference between 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*

Table 11. Trend of world poverty between 1990 and 2000.

Bandwidth→	S3	Over-smoothed	Variant of S3	Sheather-Jones	Direct plug-in	Hybrid
% reduction in poverty rate, 1990-2000						
$1/day	-11%	-5%	-10%	-11%	-11%	-11%
$1.5/day	-16%	-7%	-18%	-14%	-14%	-16%
$2/day	-16%	-8%	-18%	-16%	-14%	-16%
$3/day	-13%	-7%	-13%	-13%	-10%	-13%
$4/day	-9%	-5%	-9%	-7%	-6%	-9%
No. of people lifted from poverty, 1990-2000 (millions)						
$1/day	33	19	25	34	38	33
$1.5/day	88	45	92	78	82	88
$2/day	157	81	179	166	141	157
$3/day	205	115	205	192	160	205
$4/day	175	96	183	146	128	175