the review of income and wealth

Review of Income and Wealth
Series 60, Number 2, June 2014
DOI: 10.1111/roiw.12083

INTERPOLATING THE LORENZ CURVE: METHODS TO PRESERVE SHAPE AND REMAIN CONSISTENT WITH THE CONCENTRATION CURVES FOR COMPONENTS

BY MASATO OKAMOTO*

Statistics Bureau, Ministry of Internal Affairs and Communications

C^1-class interpolation methods that preserve monotonicity and convexity and are thus suitable for the estimation of the Lorenz curve from grouped data are not widely known. Instead, parametric models are usually applied for such estimation. Parametric models, however, have difficulty in accurately approximating every part of income/expenditure distributions. This paper proposes two types of C^1-class shape-preserving interpolation methods. One is a piecewise rational polynomial interpolation (proposed independently by Stineman and Delbourgo) that enables consistent interpolation of the concentration curves for income/expenditure components, attaining approximately the same accuracy as that of the existing methods when applied to decile-grouped data or to more detailed aggregation. Another is a Hybrid interpolation that employs pieces of curves derived from parametric models on end intervals. Empirical comparisons show that the Hybrid interpolation (with the assistance of parametric models for class-boundary estimation) outperforms the existing methods even when applied to quintile-grouped data without class boundaries.

JEL Codes: C14, D31

Keywords: income distribution, inequality, poverty, rational interpolation

1. INTRODUCTION

Although access to microdata is now no longer exceptional, there are still many cases in which income distributions[1] need to be estimated from grouped data; for instance, for the measurement of global income inequality and poverty (which requires data for many countries) or the estimation of income inequality and poverty in the past (for which microdata no longer exist). In fact, several researchers have recently estimated income distributions from grouped data, such as Chotikapanich *et al*. (2007) and Bresson (2009). The recent proposal by Shorrocks and Wan (2009) of a new method for estimating income distributions from grouped data (abbreviated hereafter as the SW-method) also supports the view that the need to make estimations based on available tabulated data still exists.

For estimating the Lorenz curve (LC) of income distribution from grouped data, parametric models such as the Beta Lorenz curve (β-LC) of Kakwani (1980a) and the General Quadratic Lorenz curve (GQ-LC) of Villaseñor and Arnold

Note: I would like to thank the anonymous referees and the editor Stephan Klasen for providing useful comments to substantially improve the first draft.

*Correspondence to: Masato Okamoto, Statistics Bureau, Ministry of Internal Affairs and Communications, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan (mokamoto2@soumu.go.jp).

[1] The proposed methods in this paper are applicable to expenditure distributions as well. For brevity, I use the term “income” instead of “income/expenditure” in some cases.

(1984, 1989) are frequently applied. The SW-method is applied to make adjustments to fitted parametric models. Nonetheless, it appears natural to also consider the application of interpolation methods. Gastwirth and Glauberman (1976) propose the application of Hermite's polynomial interpolation; in practice, however, Hermite's interpolation curves frequently fail to satisfy monotonicity and convexity, particularly on both end intervals. Simple linear interpolation, on the other hand, satisfies monotonicity and convexity, but the accuracy is much worse. Consequently, interpolation methods as a whole appear to have a very poor reputation (cf. Datt, 1998). That said, interpolation methods have a certain underlying attractiveness when considering that parametric models with few parameters generally have difficulty in accurately approximating every part of heterogeneous income distributions in societies consisting of various population groups in addition to the lack of decomposability into the concentration curves (CCs) for income components.

This paper discusses the application of piecewise rational polynomial interpolation of class C^1, which is identical to that proposed by Stineman (1980), when data points to be passed by the interpolation curve satisfy the condition of strict convexity. Although Stineman's interpolation was published in a non-academic journal, his method has been implemented as a function in MATLAB and R, and has managed to survive until today because of its simplicity and shape-preserving property that is essentially without restrictions. In the numerical analysis literature, Delbourgo (1989) later noted that a special case of the interpolation method of Delbourgo and Gregory (1985a) has a simple form and satisfies strict convexity. This special case is identical to Stineman's. This not entirely straightforward development might hinder widespread use of their method, despite its simplicity and good property. In this paper, this interpolation is called the Stineman–Delbourgo–Gregory interpolation, abbreviated as "the SDG interpolation" or simply SDG, and its generalization by Delbourgo and Gregory (1985a) is abbreviated as "the DG interpolation" or simply DG. With the goal of higher accuracy, this paper also discusses C^1-class interpolations that employ pieces of curves derived from popular parametric models such as the Pareto distribution and β-LC (Kakwani, 1980a) on end intervals and the SDG interpolant on intermediate intervals. This latter method is hereafter called "the Hybrid interpolation" or simply Hybrid.

In addition to being the simplest among rational interpolation methods, SDG has the advantages of convexity without restrictions and decomposability into interpolation curves of the CCs for income components by generalizing it to the DG interpolation. In the literature, Brown and Mazzarino (1984) have applied a rational interpolation method studied by Gregory and Delbourgo (1982) to interpolate the LC; this version of rational interpolation by Gregory and Delbourgo, however, is only assured of monotonicity as a sufficient condition. The Hybrid interpolation discussed in this paper is similar to that of Kakwani (1980b), which employs the Pareto interpolation curve on end intervals and Hermite's interpolant on intermediate intervals. The Hybrid interpolation in this paper differs from Kakwani's method in that the SDG interpolant always assures the estimated LC of monotonicity and convexity on intermediate intervals, and the use of the β-LC improves the accuracy of the estimation on the right-end interval. This paper also provides estimation methods for the derivatives of intermediate points using the β-LC or GQ-LC when class boundaries of grouped data are unavailable. These methods reduce the deterioration of accuracy to a minimum and maintain the superiority of the Hybrid interpolation over existing methods, even when applied to quintile-grouped data without class boundaries.

This paper is organized as follows. In Section 2, SDG is introduced along with estimation methods for derivatives at intermediate points and endpoints that are required to be estimated when not given. Section 3 discusses the mixed use of different types of interpolants, that is, the Hybrid method of substituting pieces of curves derived from some parametric models such as the Pareto, log-normal distribution, and/or β-LC for the SDG interpolant on end intervals. Section 4 considers methods of interpolating the CCs of income components consistently with the LC for overall income using DG. In Section 5, the two proposed types of interpolation are empirically compared with existing methods for LC estimation from grouped data using microdata for seven countries. In Section 6, DG is applied to ten sets of income/expenditure data classified according to types of sources/purchased-items for five countries to estimate the CCs for the components, and the results are compared with the piecewise linear and quadratic (composite Simpson) interpolation. The final section offers concluding remarks.

2. Interpolation of the Lorenz Curve by the Stineman–Delbourgo–Gregory Method

2.1. *Formula and Properties*

Suppose $F(x)$ is the cumulative distribution function (CDF) of a positive income variable with a finite expectation μ. Its LC, then, is represented as $LC(p) = \int_0^p F^{-1}(\pi)/\mu d\pi$ (cf. Gastwirth, 1971; Kleiber and Kotz, 2003). Consider that a division of the unit interval [0, 1] is given as $0 = p_1 < p_2 < \ldots < p_n = 1$. Let l_i denote $LC(p_i)$, $i = 1 \ldots n$, and let the interval between p_i and p_{i+1}, its width and the slope of the line passing through (p_i, l_i) and (p_{i+1}, l_{i+1}) be denoted as follows:

$$(1) \qquad I_i = [p_i, p_{i+1}], \quad h_i = p_{i+1} - p_i, \quad \Delta_i = (l_{i+1} - l_i)/h_i \quad \text{for} \quad i = 1, \cdots, n-1.$$

Furthermore, assuming that $F(x)$ is a strictly increasing continuous function, then the data points (in other words, the Lorenz coordinates) (p_i, l_i), $i = 1 \ldots n$, form a strictly convex set, i.e., Δ_i, $i = 1 \ldots n$, satisfying the following inequalities:

$$(2) \qquad \Delta_1 < \Delta_2 < \cdots < \Delta_{n-1}.$$

A function $L(p)$ on the closed interval [0, 1] is called an interpolation curve for $LC(p)$ when $L(p)$ passes through the given data points, i.e., $L(p_i) = l_i$, $i = 1 \ldots n$. We only consider continuously differentiable $L(p)$ which has derivatives equal to those of $LC(p)$ at the data points, i.e., $L(p)$ satisfies the following equalities:

$$(3) \qquad L^{(1)}(p_i) = d_i = LC^{(1)}(p_i) = F^{-1}(p_i)/\mu (\geq 0) \quad \text{for} \quad i = 1, \cdots, n.$$

Such interpolation is called a C^1-class interpolation. Note that $\{d_i\}$ satisfy the following inequalities:

$$(4)\qquad d_1 < \Delta_1 < \cdots < \Delta_{i-1} < d_i < \Delta_i < \cdots < \Delta_{n-1} < d_n.$$

The inequalities in (4) imply that the following quantities are strictly positive:

$$(5)\qquad A_i = d_{i+1} - \Delta_i, \quad B_i = \Delta_i - d_i \quad \text{for} \quad i = 1, \cdots, n.$$

Stineman (1980) and Delbourgo (1989) have proposed a C^1-class piecewise interpolation method that employs the following rational polynomial as an interpolant on each interval:

$$(6)\quad L_{SDG}(p) = h_i^{-1}[l_i(p_{i+1}-p) + l_{i+1}(p-p_i)] - \frac{A_iB_i(p_{i+1}-p)(p-p_i)}{A_i(p_{i+1}-p)+B_i(p-p_i)} \quad \text{for} \quad p \in I_i.$$

In the case of $A_i = B_i$, $L_{SDG}(p)$ is a piecewise quadratic-polynomial interpolation equivalent to Hermite's interpolation. The first and second derivatives of $L_{SDG}(p)$ are represented as follows:

$$(7)\ L_{SDG}^{(1)}(p) = \frac{A_i^2 d_i (p_{i+1}-p)^2 + 2A_iB_i\Delta_i(p_{i+1}-p)(p-p_i) + B_i^2 d_{i+1}(p-p_i)^2}{[A_i(p_{i+1}-p)+B_i(p-p_i)]^2} \quad \text{for} \quad p \in I_i,$$

$$(8)\qquad L_{SDG}^{(2)}(p) = \frac{2h_i^{-1}A_i^2B_i^2}{[A_i(p_{i+1}-p)+B_i(p-p_i)]^3} \quad \text{for} \quad p \in I_i.$$

Because $L_{SDG}^{(2)}(p) > 0$, $L_{SDG}(p)$ is strictly convex and, thus, strictly increasing on the whole [0, 1] interval when $d_i \geq 0$. As the equalities $L_{SDG}(0) = p_1 = 0$ and $L_{SDG}(1) = p_n = 1$ also hold true, $L_{SDG}(p)$ satisfies all required conditions for the LC (Thompson, 1976); thus, $L_{SDG}(p)$ has an accompanied CDF that can be regarded as an approximation of $F(x)$. In cases where the income variable has point masses, $F^{-1}(p)$ is not strictly increasing; its LC is not strictly convex; and there may be an interval on which equalities $d_i = \Delta_i = d_{i+1}$ hold. By applying a linear interpolant to such intervals rather than the SDG interpolant, the interpolation curve remains in the C^1-class and satisfies the conditions of the LC.[2] The advantages of SDG are simplicity and the desired property that it satisfies strict convexity essentially without restrictions, in contrast to the Hermite interpolation that fails to satisfy convexity in the cases $2A_i < B_i$ or $A_i > 2B_i$ (Gastwirth and Glauberman, 1976).[3]

[2] The SDG interpolation satisfies convexity even if $d_1 < 0$. Thus, SDG is also applicable to the case in which there are negative incomes. However, if negative incomes are not negligible, income data is inappropriate as a measure of the standard of living. Alternative data or variables should be considered in such cases.

[3] As shown by Delbourgo (1989), if the LC is fourth continuously differentiable and min $LC^{(2)} > 0$, the approximation error of $L_{SDG}(p)$ is in the order of $O(h_i^4)$ on interval I_i, the same order as the Hermite interpolation. However, because in many cases the LC needs to be interpolated from relatively coarsely grouped data such as quintile- or decile-grouped data, a higher order of accuracy does not necessarily imply higher accuracy in practice. Furthermore, it also should be noted that the LC cannot be generally assumed to be fourth continuously differentiable at both endpoints. As empirically shown in Section 5, SDG attains the same level of accuracy as the Hermite interpolation unless the latter method violates the conditions of the LC.

2.2. *Estimating the Derivatives of Intermediate Points*

The LC for an income distribution with a finite expectation μ has the continuous derivative $LC^{(1)}(p) = F^{-1}(p)/\mu$ over $0 \le p < 1$ when its CDF $F(x)$ is strictly increasing or its inverse CDF $F^{-1}(p)$ is continuous. Thus, the derivatives at the Lorenz coordinates corresponding to the data points in Section 2.1 can be calculated from grouped data when the class boundaries of the grouped data $F^{-1}(p_i)$, $i = 1, \ldots, n$, are specified.[4] However, when we need to estimate the LC from quintile- or decile-grouped data, class boundaries are not specified in many cases. For instance, the World Income Inequality Database (WIID) provided by UNU-WIDER (2008) does not contain class boundaries. To apply SDG in such cases, the derivatives at the data points need to be estimated. In this subsection, estimation methods for the derivatives at data points other than endpoints are discussed.

Delbourgo and Gregory (1985b) and Delbourgo (1989) propose estimation methods in the forms of arithmetic, geometric and harmonic means as follows:

$$\text{(9)} \qquad \begin{bmatrix} d_i = (h_i\Delta_{i-1} + h_{i-1}\Delta_i)/(h_{i-1} + h_i), \\ \log(d_i) = (h_i \log(\Delta_{i-1}) + h_{i-1}\log(\Delta_i))/(h_{i-1} + h_i), \\ 1/d_i = (h_{i-1} + h_i)/(h_i/\Delta_{i-1} + h_{i-1}/\Delta_i) \end{bmatrix}.$$

These methods are hereafter called the arithmetic, geometric, and harmonic means, respectively. The derivatives estimated in (9) satisfy the inequalities in (4) and approximate the true value in the order $O(h^2)$, where $h = \max\{h_{i-1}, h_i\}$.[5] However, as shown in Section 5, all methods in (9) fail to attain sufficient accuracy. Instead, we consider applying the β-LC of Kakwani (1980a), represented as $L_\beta(p) = p - \theta p^\gamma(1 - p)^\delta$, as well as the GQ-LC of Villaseñor and Arnold (1984, 1989), represented as $L_{gq}(p) = -\frac{1}{2}\left[bp + e + \sqrt{mp^2 + np + e^2}\right]$. Parameters θ, γ, δ of the β-LC passing through three successive data points, (p_{i-1}, l_{i-1}), (p_i, l_i) and (p_{i+1}, l_{i+1}), can be easily obtained by solving the following simultaneous linear equations:

$$\text{(10)} \qquad \log(p_k - l_k) = \log\theta + \gamma\log p_k + \delta\log(1 - p_k) \quad \text{for} \quad k = i-1, i, i+1.$$

The derivative at p_i of the fitted β-LC is calculated as shown in (11). $L_\beta^{(1)}(p_i)$ shall be used as d_i, rather than the estimates in (9). As for the estimation of d_2, d_{n-1}, i.e., the derivatives at the leftmost and rightmost intermediate data points p_2, p_{n-1}, the β-LCs fitted for estimating d_3, d_{n-2}, respectively, shall be used.

$$\text{(11)} \qquad L_\beta^{(1)}(p_i) = 1 - \theta p_i^\gamma(1 - p_i)^\delta[\gamma/p_i - \delta/(1 - p_i)].$$

[4] Normally, the average income μ is available or computable using grouped data.

[5] When the estimated derivatives are in order $O(h^2)$ of accuracy, SDG approximates the LC in order $O(h^3)$ in the case that the LC is fourth continuously differentiable (Delbourgo, 1989). However, the estimation methods with the same order of accuracy do not necessarily yield the same level of accuracy in practice, particularly when applied to coarsely grouped data.

The parameters of the GQ-LC passing through the three data points can also be easily obtained from the following simultaneous linear equations:

$$(12)\qquad l_k(1-l_k)=a(p_k^2-l_k)+bl_k(p_k-1)+c(p_k-l_k) \quad \text{for} \quad k=i-1,i,i+1,$$
$$e=-(a+b+c+1),\quad m=b^2-4a,\quad n=2be-4c,\quad r=\sqrt{n^2-4me^2}.$$

The derivative at p_i is calculated as follows:

$$(13)\qquad L_{gq}^{(1)}(p_i)=-b/2-(mp/2+n/4)\Big/\sqrt{mp^2+np+e^2}\,.$$

Although neither model necessarily satisfies monotonicity or convexity, and both of them may take a negative value near the left endpoint as noted in the literature,[6] the derivatives of the β-LCs at the data points always satisfy the inequalities in (4) in our empirical studies. As for the GQ-LC, the problems are not observed when applied to quintile-grouped data.[7]

2.3. *Estimating the Derivatives of Endpoints*

Even in the cases where the derivatives of intermediate points are given, the derivatives of endpoints typically need to be estimated. In this paper, the left and right endpoint derivatives d_1, d_n shall be estimated by the following methods in the forms of arithmetic, geometric, and harmonic means:

$$(14)\qquad \begin{bmatrix} d_1=2\Delta_1-d_2, \\ \log(d_1)=2\log(\Delta_1)-\log(d_2), \\ 1/d_1=2/\Delta_1-1/d_2, \end{bmatrix}, \quad \begin{bmatrix} d_n=2\Delta_{n-1}-d_{n-1}, \\ \log(d_n)=2\log(\Delta_{n-1})-\log(d_{n-1}), \\ 1/d_n=2/\Delta_{n-1}-1/d_{n-1} \end{bmatrix}.$$

Hereafter, these methods are called the arithmetic, geometric, and harmonic means, respectively. When the derivatives of intermediate points are not given, those derivatives shall first be estimated by the methods in Section 2.2. Then, the endpoint derivatives shall be estimated. Estimations I and II—presented by Delbourgo (1989) as arithmetic and harmonic means—are equivalent to the arithmetic and geometric means, respectively, in (14), when $h_1 = h_2$. The geometric

[6] The β-LC is criticized for this theoretical problem in the literature. Nevertheless, Cheong (2002) empirically shows that the β-LC is superior overall to other parametric models by using the U.S. income data of the CPS March Supplement. Our empirical comparisons using income/expenditure data from seven countries confirm that the β-LC as well as the GQ-LC (which is excluded in the study of Cheong) is superior to other parametric models, at least in some evaluation measures, for example the estimation accuracy of the Gini coefficient.

[7] The derivatives estimated by the procedures using the β-LC and GQ-LC have accuracy of order $O(h^2)$, the same order as those of the methods in (9).

mean always yields positive values, whereas the arithmetic and harmonic means may inappropriately yield negative values at the left and right endpoints, respectively. The harmonic mean may also be infinite. When the harmonic mean is negative or infinite, or when interpolation is applied to quintile-grouped data, the following inverse CES (constant elasticity of substitution) mean of order two shall be employed for the right endpoint's derivative as an intermediate between the geometric and harmonic means:

$$1/\sqrt{d_n} = 2/\sqrt{\Delta_{n-1}} - 1/\sqrt{d_{n-1}}. \tag{15}$$

Hereafter, estimation (15) is called the R-harmonic mean. It is not computable when the right-hand side is negative; in practice, however, it is always computable in the empirical studies in Section 5. As for the left endpoint's derivative, we also try the method of fixing it to zero. Note that poverty measures are not computable when the estimates of d_1 lie above the ratio of the poverty line to the average income μ. The zero derivatives have an advantage in this respect.

3. Interpolation of the Lorenz Curve by the Hybrid Method

The estimation methods for endpoint derivatives in Section 2.3 are required to fit the SDG interpolant to both end intervals. That said, when an income distribution follows a power law, its LC is represented as $1 - C(1 - p)^k$ $(C > 0, 0 < k < 1)$ around the right endpoint, which is not continuous differentiable at the right endpoint. Similarly, when the distribution has a Pareto tail on the lower end, the LC is represented as $Cp^k(C > 0, 1 < k < 2)$ around the left endpoint,[8] which is not twice continuous differentiable at the left endpoint. Thus, the SDG interpolation is expected to be ill-fitting at both end intervals. To make matters worse, inequality indices such as the Theil index and lower-tail-sensitive poverty indices such as the Squared Poverty Gap require higher accuracy on the end intervals. To overcome the limitation of SDG's accuracy, we consider adapting pieces of curves derived from parametric models to interpolants for the end intervals. First, the Pareto interpolation curves (P-ICs) are introduced, as follows:

$$\begin{aligned} L_{LP}(p) &= C_L p^{k_L} (C_L > 0, 1 < k_L), \\ L_{RP}(p) &= 1 - C_R (1-p)^{k_R} (C_R > 0, 0 < k_R < 1). \end{aligned} \tag{16}$$

Parameters C_L and k_L of $L_{LP}(p)$ are uniquely determined by the conditions that the P-IC should pass through (p_2, l_2) and that its derivative at p_2 should equal d_2. Similarly, parameters C_R and k_R of $L_{RP}(p)$ are uniquely determined by the conditions that the P-IC should pass through (p_{n-1}, l_{n-1}) and that its derivative

[8] Several researchers such as Champernowne (1953) observed that actual income distributions appear to follow the left power law as well as the (right) power law, i.e., the density of incomes at a low level x is proportional to x^κ, where $\kappa > 0$. Major parametric models with a good reputation for being well-fitting to empirical income distributions, such as the Dagum distribution and the Generalized Beta distribution of the 2nd kind, have the left and right Pareto tails.

at p_{n-1} should equal d_{n-1}.[9] Note that $k_L \geq 2$ is possible. Next, the log-normal interpolation curves (LN-ICs) are introduced:

$$(17) \qquad \begin{aligned} L_{LLN}(p) &= C_L \Phi\left(\Phi^{-1}(p) - \sigma_L\right)(C_L > 0, \sigma_L > 0), \\ L_{RLN}(p) &= 1 - C_R\left[1 - \Phi\left(\Phi^{-1}(p) - \sigma_R\right)\right](C_R > 0, \sigma_R > 0), \end{aligned}$$

where Φ denotes the CDF of the standard normal distribution. The parameters are uniquely determined by the same conditions as the P-ICs. Parameters σ_L and σ_R need to be determined by solving the following implicit equations:

$$(18) \qquad \begin{aligned} \frac{d_2}{l_2}\phi\left(\Phi^{-1}(p_2)\right) &= \frac{\phi\left(\Phi^{-1}(p_2) - \sigma_L\right)}{\Phi\left(\Phi^{-1}(p_2) - \sigma_L\right)}, \\ \frac{d_{n-1}}{(l_{n-1} - 1)}\phi\left(\Phi^{-1}(p_{n-1})\right) &= \frac{\phi\left(\Phi^{-1}(p_{n-1}) - \sigma_R\right)}{1 - \Phi\left(\Phi^{-1}(p_{n-1}) - \sigma_R\right)}, \end{aligned}$$

where ϕ denotes the probability density function of the standard normal distribution. With σ_L and σ_R thus determined and the conditions that $L_{LLN}(p)$ and $L_{RLN}(p)$ should pass through (p_2, l_2) and (p_{n-1}, l_{n-1}), respectively, parameters C_L, C_R can be easily obtained from (17). As shown by our empirical studies in Section 5, more flexible parametric models are desirable for the end interval interpolation. As for the right-end interval, a piece of the β-LC is considered as an additional interpolant β-IC.

$$(19) \qquad L_{R\beta m}(p) = p - \theta_R p^{\gamma_R}(1-p)^{\delta_R}.$$

To fix parameters θ_R, γ_R, δ_R, it appears natural to add the condition that the β-IC should pass through (p_{n-2}, l_{n-2}) to the condition that the β-IC should pass through (p_{n-1}, l_{n-1}) with its derivative equal to d_{n-1} at p_{n-1}. In this paper, however, because the β-LC does not necessarily need to pass through (p_{n-2}, l_{n-2}), an "average" over several intermediate points other than (p_{n-1}, l_{n-1}) is taken, as follows, aiming for a more stable estimation:

$$(20) \qquad \begin{aligned} \log \bar{l}_m &= \sum\nolimits_k h_k \log(p_k - l_k) = \log\theta_R + \gamma_R \sum\nolimits_k h_k \log p_k + \delta_R \sum\nolimits_k h_k \log(1 - p_k) \\ &= \log\theta_R + \gamma_R \log\tilde{p}_m + \delta_R \log\bar{p}_m, \end{aligned}$$

where $\bar{l}_m = \prod_k (p_k - l_k)^{h_k}$, $\tilde{p}_m = \prod_k p_k^{h_k}$, $\bar{p}_m = \prod_k (1 - p_k)^{h_k}$. Summation over k is taken in the range $\sum_{k=l}^{n-2} h_k < m$ in (20) for a given value $0 \leq m \leq 1$. If $h_{n-2} \geq m$, then (20) shall be replaced with the condition that the β-LC should pass through (p_{n-2}, l_{n-2}). For instance, $L_{R\beta 0}(p)$ is determined by the extra condition that the β-LC should

[9] The P-ICs can be regarded as generalizations of the LCs of the (left or right) Pareto distribution by introduction of an additional parameter C_L or C_R. The P-ICs are identical to the LCs of some statistical size distributions that follow the Pareto distribution on the respective intervals. The mean of the size distributions outside the respective intervals contributes to determining C_L, C_R. The shape of the size distributions outside the respective intervals does not make any contribution. The same is true for the LN-ICs.

pass through (p_{n-2}, l_{n-2}), while $L_{R\beta 1}(p)$ is determined by the extra condition (20) with summation over all intermediate points except (p_{n-1}, l_{n-1}). In the case of $L_{R\beta 0.4}(p)$, the summation is taken over four points from the 5th through 8th decile points for decile-grouped data and the 2nd and 3rd quintile points for quintile-grouped data. The parameters are obtained by solving simultaneous linear equations, as follows:

$$\text{(21)}\qquad \gamma_R = \frac{\left(\dfrac{1}{1-p_{n-1}}\log\dfrac{p_{n-1}-l_{n-1}}{\bar{l}_m}+\dfrac{1-d_{n-1}}{p_{n-1}-l_{n-1}}\log\dfrac{1-p_{n-1}}{\bar{p}_m}\right)}{\left(\dfrac{1}{p_{n-1}}\log\dfrac{1-p_{n-1}}{\bar{p}_m}+\dfrac{1}{1-p_{n-1}}\log\dfrac{p_{n-1}}{\tilde{p}_m}\right)},$$

$$\delta_R = \frac{\left(\dfrac{1}{p_{n-1}}\log\dfrac{p_{n-1}-l_{n-1}}{\bar{l}_m}-\dfrac{1-d_{n-1}}{p_{n-1}-l_{n-1}}\log\dfrac{p_{n-1}}{\tilde{p}_m}\right)}{\left(\dfrac{1}{p_{n-1}}\log\dfrac{1-p_{n-1}}{\bar{p}_m}+\dfrac{1}{1-p_{n-1}}\log\dfrac{p_{n-1}}{\tilde{p}_m}\right)},$$

$$\theta_R = \log(p_{n-1}-l_{n-1})-\gamma_R\log p_{n-1}-\delta_R\log(1-p_{n-1}).$$

$\delta_R \leq 1$ is a necessary condition for the convexity of the β-IC on [p_{n-1}, p_n(= 1)]. In our empirical studies in Section 5, $\delta_R > 1$ occurs in some cases when the intermediate point derivatives are estimated by the arithmetic or geometric mean. In contrast, the β-IC always satisfies $\delta_R \leq 1$ (as well as monotonicity and convexity) for $0 \leq m \leq 1$ in the cases when the intermediate point derivatives are available or estimated by the harmonic mean or β-LC.[10]

4. Consistent Interpolation of the Concentration Curves for Income/Expenditure Components by the Delbourgo–Gregory Method

Now, assume that the CDF $F(x)$ for the overall income is strictly increasing; the amount variable of income component s, denoted as X_s, has a finite expectation μ_s ($s = 1 \ldots K$); and its conditional expectation $g_s(x) = E(X_s|x)$ is continuous with respect to an income level x. Then, the concentration curve (CC) for component s is defined as $C^s(p) = \int_0^p g_s(F^{-1}(\pi))/\mu_s d\pi$ (Kakwani, 1977).

One of the advantages of polynomial interpolation methods such as the Hermite interpolation is consistency with the corresponding interpolation of the CCs for income components. Rational interpolation SDG also shares this advantage because SDG is a special case of the DG interpolation proposed by

[10] A flexible interpolant is also desirable for the interpolation on the left end interval. The β-LC is inappropriate for this purpose because it often fails to satisfy the required conditions on the lower end. An interpolant based on the parametric model of Ortega *et al.* (1991), represented as $\theta_R p^{\gamma_R}\left[1-(1-p)^{\delta_R}\right]$, does not improve the accuracy, although the interpolant empirically satisfies the required conditions (note that the parameters need to be allowed to take values outside of the range specified for the original Ortega model, i.e., $\theta_R > 0$, $\delta_R \geq 0$, $0 < \gamma_R \leq 1$. The parameters may even take negative values. Another Ortega-type interpolant represented as $p^{\gamma_R}\left[1-\theta_R(1-p)^{\delta_R}\right]$ fails to satisfy convexity in some cases. Although an exhaustive inquiry is not made, parametric models that impose strict restrictions on the ranges of parameter values appear to be inappropriate as interpolants, even if they are assured of satisfying the required conditions for the LC.

Delbourgo and Gregory (1985a). None of the parametric models exhibit this property.[11] Using the same notation as in Section 2, the DG interpolant on interval I_i is represented as follows:

$$(22)\qquad Q_i(\theta)=\frac{l_i(1-\theta)^3+(t_il_i+h_id_i)\theta(1-\theta)^2+(t_il_{i+1}-h_id_{i+1})\theta^2(1-\theta)+l_{i+1}\theta^3}{1+(t_i-3)\theta(1-\theta)},$$

where $0 \le \theta = (p - p_i)/h_i \le 1$, $p_i \le p \le p_{i+1}$. Conditions (2) and (4) are not necessarily required for the CC interpolation. The interpolant holds equalities $Q_i(\theta(p_i)) = l_i$, $Q_i(\theta(p_{i+1})) = l_{i+1}$, $dQ_i/dp|_{p_i}=d_i$, $dQ_i/dp|_{p_{i+1}}=d_{i+1}$. Thus, the DG interpolation, which applies the interpolant in (22) piecewise, belongs to the C^1-class. An extra parameter $t_i(>-1)$ is called the tension parameter. The larger the t_i, the nearer the interpolant is to the linear interpolant. DG is equivalent to Hermite's interpolation when $t_i = 3$ whereas it is equivalent to SDG when t_i is set as in (23).

$$(23)\qquad t_i=1+A_i/B_i+B_i/A_i\ (\ge 3 \text{ if } A_i, B_i>0).$$

According to Delbourgo and Gregory (1985a), when d_i, $d_{i+1} \ge 0$, $\Delta_i > 0$, DG is increasing on interval I_i if the tension parameter satisfies the following inequality:

$$(24)\qquad t_i \ge (d_i+d_{i+1})/\Delta_i.$$

Under assumption (4), DG is assured of convexity on I_i if and only if the tension parameter satisfies the following condition:

$$(25)\qquad t_i \ge 1+\max\{A_i/B_i, B_i/A_i\}.$$

When the derivative d_i^s of the CC at data point $(p_i, l_i^s(=C^s(p_i)))$ is available for $i = 1 \ldots n$, the DG interpolant $Q_i^s(\theta(p))=C_{DG}^s(p)$, obtained by substituting l_i^s, d_i^s for l_i, d_i in (22) with the same tension parameter as that for the overall income, satisfies the following equality:[12]

$$(26)\qquad L_{SDG}(p_{I_i}(\theta))=\sum\nolimits_{s=1}^{K}w_sC_{DG}^s(p_{I_i}(\theta))=\sum\nolimits_{s=1}^{K}w_sQ_i^s(\theta)\quad\text{for}\quad 0\le\theta\le 1,$$

where $p_{I_i}(\theta)=p_i(1-\theta)+p_{i+1}\theta$, w_s denotes the money share of component s. DG is expected to generally be more accurate than the linear interpolation; however, because the common tension parameter determined by the SDG interpolation of the LC for overall income needs to be used for all components, the incidence

[11] The Bernstein polynomial fitted by Ryu and Slottje (1996) to the LC might be exceptional. The number of parameters, however, appears to be too high if it is regarded as a parametric model. Furthermore, their empirical study shows that their method yields relatively large errors when applied to decile-grouped data.

[12] In practice, we usually know only about a sample estimate of l_i^s. We do not distinguish the notation for sample estimates from that for the true value because there appears to be no fear of confusion.

of inappropriate cases cannot be eliminated completely. Our empirical studies in Section 6 indicate that such cases may occur when the empirical CCs to be interpolated cannot be regarded as smooth due to excessively minute component classification and/or income class breakdowns. Excessively minute components should be collapsed before applying the interpolation.

The derivatives of the CCs at the data points are usually unavailable, unlike those of the LC for overall income. Even if we know the amounts of income components earned by individuals whose overall incomes correspond to class boundaries, it is not appropriate to use the amounts relative to the respective average amount μ_s as the intermediate point derivatives. Some averages should be taken around the class boundaries to compute the derivatives, but such processed data are usually unavailable. Therefore, those derivatives need to be estimated from grouped data. A two-stage procedure is employed here. First, a tentative estimate $\hat{d}_i^s$ shall be calculated using the arithmetic, geometric, or harmonic mean formulas in (9) for the intermediate points and (14) for the right endpoints.[13] Next, the tentative estimates shall be adjusted proportionally to make the final estimates consistent with the derivative of the LC for overall income, as follows:[14]

$$(27) \qquad d_i^s = d_i \cdot \hat{d}_i^s \Big/ \sum_{j=1}^{K} w_j \hat{d}_i^j.$$

Regarding the left endpoint, all derivatives including that for overall income shall be set to zero. Note that the estimated derivatives based on the geometric or harmonic mean depend on component classification; however, our empirical studies show that inconsistencies among classifications are sufficiently small. The arithmetic mean basically yields the classification-free estimates, except for the cases wherein the exceptional treatment in footnote 13 is applied.

The adjustment in (27) sometimes makes the accuracy slightly worse. Usually, a specific component such as the wages of the household heads has a dominant share in the overall income. In such cases, it may be reasonable to use the estimate for the largest component at the first stage without adjustment and make adjustments among the rest of the components. That said, as a whole, this modification shows no particular improvement and reduces the accuracy in some cases. For this reason, modifications to (27) are not taken up in this paper.

The integral of the interpolated CC for component s on interval I_i can be calculated by the following formula:

[13] Because the CCs are not necessarily monotonic when the components allow entries of both positive and negative amounts, particular consideration should be made for the geometric and harmonic means. In the case that both Δ_{i-1} and Δ_i are negative, the calculation of $\hat{d}_i^s$ shall be the geometric and harmonic means of the absolute values multiplied by −1, while in the case that Δ_{i-1} has an opposite sign to Δ_i, the calculation shall be replaced by the arithmetic mean. Probably in most cases, the components contain only (or almost only) non-negative values, and the deduction components only allow entries of non-positive values. As for such components, it may be better that $\hat{d}_i^s$ is set to zero when either Δ_{i-1} or Δ_i is zero, as illustrated in Section 6.

[14] The final estimates also have an accuracy of $O(h^2)$. Thus, similarly to SDG, the accuracy of the DG interpolation is in the order of $O(h^3)$ under the assumption of fourth continuous differentiability of the original curve (Delbourgo and Gregory, 1985a).

(28)

$$S_i^s = \begin{cases} h_i \dfrac{l_i^s + l_{i+1}^s}{2} - h_i^2 \dfrac{d_{i+1}^s - d_i^s}{2(t_i - 3)} \left[1 - 2 \dfrac{1}{\sqrt{(t_i - 3)(t_i + 1)}} \log \left(\dfrac{\sqrt{t_i + 1} + \sqrt{t_i - 3}}{\sqrt{t_i + 1} - \sqrt{t_i - 3}} \right) \right] & \text{if } t_i > 3 \\ h_i \dfrac{l_i^s + l_{i+1}^s}{2} - h_i^2 \dfrac{d_{i+1}^s - d_i^s}{12} & \text{if } t_i = 3 \end{cases}.$$

Thus, the quasi-Gini coefficient of component s is calculated as $1 - 2\sum_i S_i^s$.

5. Empirical Comparisons of Estimation Methods of the Lorenz Curve

5.1. *Data*

The WIID (UNU-WIDER, 2008) contains quantile-grouped data and the Gini coefficients for many countries. Both Minoiu and Reddy (2007) and Shorrocks and Wan (2009), however, have noted that the Gini coefficients estimated from the grouped data in some cases differ inexplicably from those in the database. Therefore, they used grouped data aggregated from microdata by their own calculations for their empirical studies. Our empirical studies should also be conducted using available microdata for the additional reason that the interpolation methods proposed in this paper attain higher accuracy in the cases in which class boundaries are given (which is not the case for the WIID).

With care to maintain a fairly even spread over the globe, freely downloadable microdata are chosen from the website of the Living Standards Measurement Study (LSMS) for Bulgaria, Cote d'Ivoire, China (Hebei and Liaoning Provinces), Peru, and Timor-Leste. In addition, microdata from the Survey of Italian Household Income and Wealth (SHIW) and the U.S. Survey of Consumer Finances (SCF) are chosen. For the U.S., SCF data are taken instead of public use data from the March Current Population Survey (CPS) because of differences in data anonymization procedures. Public use data from the SCF exclude the top 400 wealthiest people in the *Forbes* list and sample households owning property larger than the minimum of the top 400 (cf. Kennickell and Lane, 2007); nonetheless, the effect of this cutoff appears much smaller in comparison with that of the top-coding procedure in CPS. As shown in Table 1, the survey years vary among the seven countries, ranging from 1985 to 2006. Because per capita amounts are used for the measurement of global economic inequality and poverty (cf. Chen and Ravallion, 2001; Milanovic, 2002, 2005), size distributions of per capita consumption are used in our empirical comparisons for Cote d'Ivoire, China, Peru, and Timor-Leste; those of per capita gross income are used for Bulgaria and the U.S.; those of per capita net disposable income are used for Italy. To make the evaluations stable, a number of subsample sets are generated from the original samples and used for the comparisons. As for the five LSMS countries, 50 sets of subsamples are generated by a single-stage cluster sampling with replacement and aggregated into quintile, decile, and ventile groups. Although the actual sampling procedures taken for conducting the surveys employed techniques of stratification and, in some cases, three-stage samplings, a single-stage cluster sampling without stratification is employed here for simplicity. It can be said that our procedure

TABLE 1

Description of the Income/Expenditure Data

Country	Source	Year	Household$^{\#}$ Count	Cluster Count	Variable (per capita)	Inequality Measure			
						Gini	MLD	Theil	CV
Bulgaria	LSMS	2003	3016	743	Income	0.47870	0.43161	0.41679	1.21476
China*	LSMS	1995/97	787	30	Expenditure	0.27498	0.12683	0.13827	0.61516
Cote d'Ivoire	LSMS	1985	1588	100	Expenditure	0.40693	0.28621	0.28658	0.87557
Italy	SHIW	2006	7762	–	Income	0.34933	0.22275	0.23160	0.97347
Peru	LSMS	1994	3623	364	Expenditure	0.47617	0.40546	0.42864	1.27022
Timor-Leste	LSMS	2001	1800	300	Expenditure	0.40321	0.27319	0.39884	2.26691
U.S.	SCF	2004	4498	–	Income	0.53810	0.54587	0.70507	3.05635
Average						0.41820	0.32742	0.37226	1.46749

Notes: *Hebei and Liaoning Provinces. $^{\#}$Households reporting zero or negative amounts are excluded.

somewhat reflects the actual sampling procedures because the clusters are formed based on the actual sampling units. The total numbers of the clusters are listed in Table 1. Since Italian microdata do not contain the sampling unit codes, a simple random sampling is employed to generate 50 sets of subsamples. U.S. microdata contain 999 sets of replicate weights for sampling variance estimation and 5 sets of plausible values of income for imputation variance estimation. Thus, 50 sets of replicate weights among the 999 sets and the 5 sets of plausible values are used as 250(= 50 · 5) sets of subsamples.[15] To be exact, our studies are based on simulation results rather than empirical evidence; however, because the simulation results are considered to reflect the actual situations closely, our studies are termed "empirical" studies in this paper.

5.2. *Evaluation Methods*

The accuracy of various estimation methods for the LC shall be assessed by comparing the square root of the mean squared errors (RMSE) of the derived CDF along with the RMSEs of the derived function related to the poverty gap over the whole population, over the lowest group or over the lower 60 percent group (but excluding the lowest group) and the absolute errors of the inequality indices relative to the respective inequality values (RAE). Three popular indices, i.e., the Gini index, the Mean Log Deviation, and the Theil index (abbreviated as Gini, MLD, and Theil, hereafter), are used for the evaluations. Assuming that N households in a subsample are arranged in ascending order of per capita amount, the RMSE of the derived CDF is defined as follows:

$$RMSE = \sqrt{\sum_{i=1}^{k} \omega_i m_i (H(y_i) - p_i)^2 \Big/ \sum_{i=1}^{k} \omega_i m_i}, \tag{29}$$

where $H(y) := L^{(1)^{-1}}(y/\mu)$ denotes the derived CDF and ω_i denotes the weight for aggregation assigned to household i in which m_i persons are living together. $p_i = \sum_{j \le i} \omega_j m_j \big/ \sum_{j=1}^{N} \omega_j m_j$ is the cumulative population share up to household i, and y_i is per capita income of household i. $H(y_i)$ and p_i correspond to the estimated and empirical poverty rates, respectively, when the poverty line is set to y_i. Thus, the RMSE in (29) indicates the accuracy of the poverty rate estimation. The RMSE of the derived function related to the poverty gap is defined as follows:

$$RMSE = \sqrt{\sum_{i=1}^{k} \omega_i m_i (PG(y_i) - pg_i)^2 \Big/ \sum_{i=1}^{k} \omega_i m_i}, \tag{30}$$

where $PG(y) := H(y) - \dfrac{L(H(y))}{y/\mu}$, and $pg_i = p_i - \dfrac{LC(p_i)}{y_i/\mu}$ $\left(LC(p_i) = \sum_{j \le i} \omega_j m_j y_j \big/ \sum_{j=1}^{N} \omega_j m_j y_j\right)$. $PG(y_i)$ and pg_i correspond to the estimated and empirical poverty gap, respectively, when the poverty line is set to y_i. Thus, the RMSE in (30) indicates the accuracy of the poverty gap estimation. The summations in the

[15] The proposed methods are evaluated by averaging results for the respective sets of subsamples. The number of subsample sets is determined to evaluate the proposed methods without being affected by specific choices of subsamples within reasonable computational time.

numerators and denominators in (29) and (30) are taken over the whole population, over the lowest group, or over the lower 60 percent group (but excluding the lowest group).

To save space, we mainly present the overall aggregates of the individual estimation errors, supplementing them with abbreviated notes for the results of individual countries. The aggregations are taken in the form of the RMSE using the overall averages and the variations among averages of individual countries, as follows:

$$\text{(31)} \qquad \sqrt{\left(\overline{ARMSE}\right)^2 + \frac{1}{K-1}\sum\nolimits_k \left(ARMSE_k - \overline{ARMSE}\right)^2},$$
$$\overline{ARMSE} = \frac{1}{K}\sum\nolimits_k ARMSE_k,\ ARMSE_k = \sqrt{\sum\nolimits_j w_{k,j} RMSE_{k,j}^2},$$

where K denotes the number of countries to be aggregated; $RMSE_{k,j}$ denotes the RMSE for subsample j of country k calculated by the formulas in (29) or (30); and weight $w_{k,j}$ is 1/250 for the U.S. and 1/50 for the other countries. Regarding the inequality indices, because estimation errors tend to be larger along with the levels of index values, the relative absolute errors (RAE) to the index values are aggregated in the form of the RMSE, as follows:

$$\text{(32)} \qquad \sqrt{\overline{ARAE}^2 + \frac{1}{K-1}\sum\nolimits_k \left(ARAE_k - \overline{ARAE}\right)^2},$$
$$\overline{ARAE} = \frac{1}{K}\sum\nolimits_k ARAE_k,\ ARAE_k = \sum\nolimits_j w_{k,j} \left|\hat{\hat{I}}_{k,j} - \hat{I}_{k,j}\right| \Big/ I_k \cdot \bar{I},$$

where $I_k, \hat{I}_{k,j}, \hat{\hat{I}}_{k,j}$ denote the inequality index values estimated from the original sample, subsample j, and interpolation applied to grouped data tabulated from subsample j of country k. $\bar{I} = K^{-1}\sum_{k=1}^{K} I_k$ is a simple average of the inequality index estimates from the original samples over all countries (see Table 1). Multiplying by $\bar{I}$, we intend to make the magnitude of errors intuitively comprehensible. It is also possible to aggregate the estimation errors in the form of the RMSE for individual countries, similarly to (29). Because the results are similar to those for ARAE in (32), they are omitted here to save space.

In Appendix 1, formulas are listed for computing major poverty and inequality indices from the estimated LC by analytic means. In our empirical comparisons, we mainly make approximations of the index calculations by generating discrete distributions from the estimated LC, except for the Gini index. The approximations are made as follows: taking a sufficiently large number J (5,000,000 for Timor-Leste and the U.S., 1,000,000 for the other five countries), the interval [0, 1] shall be evenly divided into J subintervals. Then, the inequality indices are approximated using derivatives $L^{(1)}(p_j)$ at the midpoints $p_j = (j + 0.5) / J \quad (j = 1 \dots J)$ on subintervals. One of the reasons for the approximation is that some Hybrid interpolations do not allow analytic means. Another essential reason for the approximation is that the SW-method, our main competitor, requires making approximations by generating discrete distributions. As for

Theil, the approximations tend to yield estimates that are more or less biased downward. Thus, we also present the actual estimation errors for estimation methods that allow for analytic calculations, and we give upper bound estimates of actual interpolation errors for the Hybrid methods employing the β-IC on the right-end interval. The estimation methods for the LC generally produce such accurate estimates of Gini that errors coming from the approximate calculations are not ignorable for comparisons when the P-IC or β-IC is employed as the interpolant on the right-end interval. Thus, analytic calculations are made in principle. Although the LN-IC does not allow analytic calculations, the approximation errors are sufficiently small because the log-normal distributions have light tails. Estimates of Gini by the SW-method also appear sufficiently accurate, taking into consideration that differences between the analytic calculations and the approximations are sufficiently small among the fitted parametric models to which the SW-method is applied.

As mentioned briefly in the next subsection, accurate estimation of the coefficient of variation (CV) appears to go beyond the ability of any estimation method from grouped data. Thus, results for the CV are omitted in most tables.

5.3. *Comparisons among the Hermite, SDG, and Hybrid Interpolations*

The Hermite interpolation frequently yields ineligible curves as the LCs, violating monotonicity or convexity particularly on both end intervals. Here, the most trouble-free Hermite interpolation with the endpoint derivatives estimated by the geometric mean (the second and fifth formulas in (14)), denoted as g-H-g, is compared with the corresponding SDG interpolation (g-SDG-g), the more accurate SDG with a zero derivative at the left endpoint and an estimated derivative by the harmonic mean at the right endpoint (the bottom formula in (14)), denoted as z-SDG-h. Because g-H-g yields ineligible curves from all subsamples of the U.S., the comparisons are made for the other six countries.

The RMSEs of estimates from decile-grouped data are presented in Table 2. The RMSEs of estimated CDFs ($H(y)$) and functions related to the poverty gap ($PG(y)$) are presented in columns "H" and "PG," respectively. The Hermite interpolation g-H-g is slightly more accurate than the corresponding SDG interpolation g-SDG-g; however, it can be said that both have nearly the same level of accuracy. When limiting the evaluation to the accuracy of the interpolations on intermediate intervals, g-SDG-g is just slightly better than g-H-g. As SDG holds monotonicity and convexity without restrictions, it has wider choices of estimation methods for both endpoint derivatives. A more appropriate choice, z-SDG-h, makes SDG superior to g-H-g, except for $H(y)$ and $PG(y)$ at the lowest decile D1.[16] The same is true for estimates from quintile-grouped data if replacing z-SDG-h with z-SDG-rh, which employs the R-harmonic mean in (15) for estimation of the right endpoint's derivative. Replacement of the SDG interpolant on both end intervals by pieces of curves derived from suitable parametric models (P-SDG-β0.4) further improves accuracy, except for the case of CV. Because P-SDG-β0.4

[16] Decile groups are denoted as D1, D2, . . . , D10, from the lowest to the highest. Similarly, quintile groups are denoted as Q1, Q2, . . . , Q5, and ventile groups are denoted as V1, V2, . . . , V20, from the lowest to the highest.

TABLE 2

Comparison among the Hybrid, SDG, and Hermite Interpolations (Decile Groups)

	Interpolation	Gini	MLD	Theil		CV		H (10^{-5})		PG (10^{-5})	
	Method	Exact	Approx.	Exact	Approx.	Exact	Approx.	Overall	D1–D6	Overall	D1–D6
RMSE* over 4 countries	P-SDG-β0.4	0.00043	0.00491	0.00943$^{\#}$	0.00888	0.82313$^{\#\#}$	0.18876	287	259	39	33
	g-SDG-g	0.00233	0.01384	0.03190	0.03190	0.23203	0.23203	506	377	129	95
	z-SDG-h	0.00094	0.00717	0.01724	0.01660	0.13061	0.13061	398	350	88	81
	g-H-g	0.00209	0.01341	0.03058	0.03058	0.22564	0.22564	482	370	119	92
RMSE** over 6 countries	P-SDG-β0.4	0.00050	0.00440	0.02024$^{\#}$	0.02004		0.19562	276	248	42	30
	g-SDG-g	0.00266	0.01383	0.04906	0.04906	0.36986	0.36986	527	377	140	88
	z-SDG-h	0.00088	0.00629	0.02859	0.02869	0.27904	0.27904	416	404	102	107
	g-H-g	0.00239	0.01331	0.04772	0.04772	0.36519	0.36519	498	371	129	86

Notes: *Aggregation over Bulgaria, China, Cote d'Ivoire, and Peru. **Aggregation over six countries other than the U.S. $^{\#}$Upper bounds of the exact RAEs are used for the calculation. $^{\#\#}$Infinite values are excluded (2 and 5 cases for China and Peru).

yields infinite CVs for Italy and Timor-Leste, the RMSEs over the other four countries are also presented in Table 2, which reveals that the CV estimates by P-SDG-β0.4 have very large errors. Other interpolations also generally yield inaccurate estimates for CV.

In Tables 3a and 3b, the RMSEs over all seven countries are compared for five types of Hybrid methods and two types of pure SDG methods for decile-/quintile-grouped data. P-SDG-P corresponds to the interpolation method proposed by Kakwani (1980b), although the SDG interpolant is employed instead of Hermite's as the interpolant on intermediate intervals. Among the five types of Hybrids, LN-SDG-β0.4 and P-SDG-β0.4[17] show the best performance. When applied to decile- or ventile-grouped data, P-SDG-β0.4 yields smaller errors than LN-SDG-β0.4 for MLD and $PG(y)$, and larger errors for $H(y)$ but the differences are very small except for MLD. Then again, when applied to quintile-grouped data, LN-SDG-β0.4 yields smaller errors for all measures except for MLD, for which the differences of the RMSEs are very small.

For individual countries, estimates by LN-SDG-β0.4 are close to those by P-SDG-β0.4 for every country. P-SDG-β0.4 attains smaller errors in five countries for MLD when applied to decile-grouped data, whereas LN-SDG-β0.4 attains smaller errors in five countries for $H(y)$ and $PG(y)$. When applied to quintile-grouped data, clear differences are not found in terms of the number of countries for which the smallest errors are attained for MLD, whereas LN-SDG-β0.4 attains smaller errors in all seven countries for $H(y)$ and $PG(y)$.

5.4. *Comparisons among Estimation Methods for Intermediate Points Derivatives*

Among the estimation methods for intermediate point derivatives, the β-LC in (10) and (11), the GQ-LC in (12) and (13), and the arithmetic, geometric, and harmonic means in (9) are compared for the SDG and Hybrid interpolations. The Hybrid interpolation employing the P-IC at the left end and β-IC with $m = 0.4$ at the right end are denoted by P-SDG$_{\beta}$-β0.4, P-SDG$_{gq}$-β0.4, P-SDG$_{a}$-β0.4, P-SDG$_{g}$-β0.4 and P-SDG$_{h}$-β0.4 according to the applied estimation method for the intermediate point derivatives. P-SDG.-β0.4 is used as the general notation for those methods. The corresponding notation is used for LN-SDG-β0.4, z-SDG-h, etc. The RMSEs of the estimates from decile- and quintile-grouped data are presented in Tables 4a and 4b. P-SDG.-β0.4 and LN-SDG.-β0.4 are taken up here because both types are considered the most appropriate among the various Hybrid methods for estimation from decile- and quintile-grouped data, respectively, as shown in the previous subsection.[18] The respective smallest RMSEs among the estimation methods for intermediate point derivatives are in bold in the tables for the SDG and Hybrid methods.

[17] When the β-IC is employed at the right end, an appropriate value shall be set as m. The optimal value of m depends on evaluation measures and minuteness of grouped data as shown in Appendix 2. It also depends on the countries to be studied. If taking into consideration that the interpolations may be applied to various grouped data in various countries for different purposes, an intermediate value of approximately 0.4 appears to be appropriate. It should be noted, however, that too much attention need not be paid to this issue because the accuracy is generally insensitive to a choice of value for m.

[18] Although the choice between P-SDG.-β0.4 and LN-SDG.-β0.4 is not easy when interpolating decile-grouped data, the former is chosen here because of its relatively clear superiority for MLD. The differences in the other evaluation measures are very small.

TABLE 3a

Comparison among Various Hybrid and SDG Interpolations (Decile Groups)

Interpolation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	D1	D2–D6	Overall	D1	D2–D6
LN-SDG-LN	0.00126	0.00767	0.03802	0.03803	329	**262**	258	82	80	10
P-SDG-P	0.00059	0.00432	0.02234	0.02252	283	279	258	49	**72**	10
P-SDG-LN	0.00124	0.00653	0.03789	0.03790	331	279	258	82	**72**	10
LN-SDG-β0.4	0.00048	0.00492	0.02181#	0.01967	**276**	**262**	258	43	80	10
P-SDG-β0.4	**0.00047**	**0.00409**	0.02181#	**0.01966**	278	279	258	**41**	**72**	10
z-SDG-h*	0.00100	0.00587	0.03479	0.03487	411	792	258	108	254	10
z-SDG-rh	0.00166	0.00478	0.04508	0.04515	448	792	258	122	254	10

Notes: The minimum RMSEs among all interpolations are in bold. *The R-harmonic mean is used for estimation of the right endpoint's derivatives in all 250 iterations for the U.S. #Upper bounds of the exact RAEs are used for the calculation.

TABLE 3b

Comparison among Various Hybrid and SDG Interpolations (Quintile Groups)

Interpolation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	Q1	Q2–Q3	Overall	Q1	Q2–Q3
LN-SDG-LN	0.00291	0.01001	0.04946	0.04947	517	**394**	338	170	**114**	21
P-SDG-P	0.00210	0.00624	0.03156	0.03121	499	563	338	153	209	21
P-SDG-LN	0.00280	0.00673	0.04895	0.04896	547	563	338	187	209	21
LN-SDG-β0.4	**0.00113**	0.00509	**0.02365**#	**0.02281**	**394**	**394**	338	**89**	**114**	21
P-SDG-β0.4	0.00116	**0.00506**	0.02376#	0.02290	433	563	338	119	209	21
z-SDG-h*	0.00499	0.02330	0.06249	0.06230	1209	1402	338	442	559	21
z-SDG-rh	0.00306	0.01186	0.05492	0.05529	866	1402	338	305	559	21

Notes: The minimum RMSEs among all interpolations are in bold. *The R-harmonic mean is used for estimation of the right endpoint's derivatives in 10/50 cases for Peru and in all 250 iterations for the U.S. #Upper bounds of the exact RAEs are used for the calculation.

TABLE 4a

Comparison among Estimation Methods for Intermediate Data Points' Derivatives (Decile Groups)

Interpolation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	D1	D2–D6	Overall	D1	D2–D6
P-SDG-β0.4	0.00047	0.00409	0.02181#	0.01966	278	279	258	41	72	10
z-SDG-h*	0.00100	0.00587	0.03479	0.03487	411	792	258	108	254	10
z-SDG-rh	0.00166	0.00478	0.04508	0.04515	448	792	258	122	254	10
P-SDG$_{\beta}$-β0.4	**0.00055**	**0.00444**	**0.02190#**	**0.01968**	**339**	**313**	**319**	**47**	**80**	**15**
z-SDG$_{\beta}$-h*	**0.00120**	0.00654	**0.03746**	**0.03736**	**480**	819	**319**	**123**	261	**15**
P-SDG$_{a}$-β0.4	0.00514	0.01386	0.06578#	0.06578	957	365	353	269	87	16
z-SDG$_{a}$-h	0.00465	0.00738	0.06359	0.06372	938	681	353	259	194	16
P-SDG$_{g}$-β0.4	0.00386	0.01329	0.05935#	0.05935	816	465	387	221	129	19
z-SDG$_{g}$-h	0.00340	0.00698	0.05740	0.05749	790	**656**	387	201	159	19
P-SDG$_{h}$-β0.4	0.00259	0.01195	0.04783#	0.04783	677	587	454	164	184	23
z-SDG$_{h}$-h	0.00210	**0.00595**	0.04347	0.04347	647	669	454	124	**147**	23

Notes: The respective minimum RMSEs among estimation methods for intermediate data point derivatives are in bold for the SDG and Hybrid interpolations. *,#See the footnotes denoted by the same symbols below Table 3a.

TABLE 4b

Comparison among Estimation Methods for Intermediate Data Points' Derivatives (Quintile Groups)

Interpolation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	Q1	Q2–Q3	Overall	Q1	Q2–Q3
LN-SDG-β0.4	0.00113	0.00509	0.02365#	0.02281	394	394	338	89	114	21
z-SDG-h*	0.00499	0.02330	0.06249	0.06230	1209	1402	338	442	559	21
z-SDG-rh	0.00306	0.01186	0.05429	0.05229	866	1402	338	305	559	21
LN-SDG$_{\beta}$-β0.4	0.00209	**0.00464**	0.03296#	0.03111	534	**431**	**430**	142	**122**	**32**
z-SDG$_{\beta}$-rh	**0.00270**	0.01287	**0.05385**	**0.05354**	899	1433	**430**	301	562	**32**
LN-SDG$_{gq}$-β0.4	**0.00168**	0.00467	**0.02832#**	**0.02686**	**506**	**431**	462	**122**	123	39
z-SDG$_{gq}$-rh	0.00288	0.01131	0.05391	0.05427	**880**	1353	462	**287**	516	39
LN-SDG$_{g}$-β0.4	0.01112	0.02460	0.08302#	0.08302	1746	808	761	547	275	70
z-SDG$_{g}$-h	0.00732	**0.00763**	0.07056	0.07065	1527	**1218**	761	423	357	70
LN-SDG$_{h}$-β0.4	0.00643	0.02057	0.06297#	0.06297	1303	1129	987	378	432	85
z-SDG$_{h}$-h**	0.00439	0.00963	0.06388	0.06373	1289	1272	987	314	**304**	85

Notes: The respective minimum RMSEs among estimation methods for intermediate data point derivatives are in bold for the SDG and Hybrid interpolations. **The R-harmonic mean is used for estimation of the right endpoint's derivatives in 57/250 cases for the U.S. *,#See the footnotes denoted by the same symbols below Table 3b.

When P-SDG$_{\cdot}$-β0.4 is applied to decile-grouped data, P-SDG$_{\beta}$-β0.4 attains the smallest errors.[19] When LN-SDG$_{\cdot}$-β0.4 is applied to quintile-grouped data, LN-SDG$_{gq}$-β0.4 is superior to LN-SDG$_{\beta}$-β0.4, except for MLD and the lower part of $H(y)$ and $PG(y)$ (but the differences are relatively small). Both LN-SDG$_{gq}$-β0.4 and LN-SDG$_{\beta}$-β0.4 outperform LN-SDG$_{a}$-β0.4, LN-SDG$_{g}$-β0.4, and LN-SDG$_{h}$-β0.4.

When z-SDG$_{\cdot}$-h is applied to decile-grouped data, z-SDG$_{\beta}$-h attains the smallest errors for Gini, Theil, and the overall $H(y)$ and $PG(y)$, whereas z-SDG$_{\beta}$-h is inferior to z-SDG$_{h}$-h for MLD, which is sensitive to lower-tail dispersion, and z-SDG$_{\beta}$-h is also inferior to z-SDG$_{a}$-h, z-SDG$_{g}$-h, and z-SDG$_{h}$-h for $H(y)$ and $PG(y)$ at the lowest decile D1. Even in the cases where the intermediate point derivatives are known, the accuracy of $H(y)$ and $PG(y)$ at D1 is worse than that of z-SDG$_{a}$-h, z-SDG$_{g}$-h, and z-SDG$_{h}$-h. For the applications to quintile-grouped data, the R-harmonic mean (z-SDG$_{\cdot}$-rh) gives more accurate estimates than the harmonic mean (z-SDG$_{\cdot}$-h) as an estimation method for the right endpoint's derivative in the cases in which the intermediate point derivatives are known or estimated using the GQ-LC or β-LC; however, the converse is true in the cases of the other estimation methods for the intermediate point derivatives. Results for z-SDG$_{gq}$-rh and z-SDG$_{\beta}$-rh are close to each other. Similarly to the results for z-SDG$_{\beta}$-h applied to decile-grouped data, z-SDG$_{gq}$-rh and z-SDG$_{\beta}$-rh attains smaller errors than z-SDG$_{g}$-h and z-SDG$_{h}$-h for Gini, Theil, and the overall $H(y)$ and $PG(y)$, whereas both (and also the SDG interpolation with actual intermediate point derivatives) are inferior to z-SDG$_{g}$-h and z-SDG$_{h}$-h for MLD, $H(y)$ and $PG(y)$ at the lowest quintile Q1.[20]

The above results indicate that pure SDG interpolations suffer from poor fit at the lowest group, even if the intermediate point derivatives are given. For this reason, an exceptional treatment is introduced for the cases in which the intermediate point derivatives are known or estimated by the GQ-LC or β-LC, as follows: the derivatives at the leftmost intermediate point are replaced with an estimate by the harmonic mean. The respective procedures, denoted by z-$_{h}$SDG-h and so on, succeed in lowering the errors below the level of z-SDG$_{h}$-h for MLD and to the same level as z-SDG$_{h}$-h for the lower-end part of $H(y)$ and $PG(y)$ (see Tables 5a and 5b). When z-SDG$_{\cdot}$-h is applied to ventile-grouped data, z-SDG-h and z-SDG$_{\beta}$-h generally improve the accuracy relative to the other methods, resulting in the smallest errors for MLD, whereas the accuracy at the lower-end part of $H(y)$ and $PG(y)$ is still inferior to z-SDG$_{h}$-h. Thus, the modification of the leftmost intermediate point's derivative still has good effects, although the effects are smaller in comparison with those for decile-grouped data. It should be noted, however, that the exceptional treatment has a side effect that makes the accuracy worse at the second-lowest group, i.e., Q2, D2, or V2. The problem of poor fit at the lowest decile can be visually observed in the upper-left panel of Figure 1, which

[19] Results for the GQ-LC are not presented because it yields ineligible curves for the LC in some cases when applied to decile-grouped data.

[20] Results for the arithmetic mean are omitted because LN-SDG$_{a}$-β0.4 yields ineligible curves for the LC in some cases when applied to quintile-grouped data. Results for z-SDG$_{a}$-h that are omitted together are inferior to those for z-SDG$_{g}$-h.

TABLE 5a

Comparison with Parametric Methods and Those Adjusted by the Shorrocks–Wan Method (Decile Groups)

Estimation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	D1	D2–D6	Overall	D1	D2–D6
P-SDG-β0.4	**0.00047**	**0.00409**	*0.02181*[#]	**0.01966**	**278**	**279**	**258**	**41**	**72**	**10**
P-SDG$_{\beta}$-β0.4	*0.00055*	*0.00444*	**0.02090**[#]	*0.01968*	*339*	*313*	*319*	*47*	*80*	*15*
z-SDG-h*	0.00100	0.00587	0.03479	0.03487	411	792	**258**	108	254	**10**
z-$_{\text{h}}$SDG-h*	0.00105	0.00455	0.03532	0.03532	446	669	402	88	147	26
z-SDG$_{\beta}$-h*	0.00120	0.00654	0.03476	0.03736	480	819	*319*	123	261	*15*
z-$_{\text{h}}$SDG$_{\beta}$-h*	0.00124	0.00518	0.03779	0.03778	504	669	440	103	147	27
β-LC	0.00078				705	480	732	324	919	141
GQ-LC	0.00072				626	730	658	152	360	107
Rasche	0.00109	0.00704		0.03288	785	566	837	245	305	227
SM	0.00103	0.00710	0.02848	0.02872	838	604	920	263	359	250
Dagum	0.00099	0.00864	0.03215	0.03216	902	642	967	306	397	296
GB2	0.00099	0.00670	0.02842	0.02872	672	449	717	180	265	161
SW-Rasche	0.00109	0.00541		0.03082	378	359	324	82	89	17
SW-LN	0.00235	0.01014		0.04971	487	327	322	142	100	16
SW-SM	0.00092	0.00495		0.02722	370	372	324	74	92	17
SW-Dagum	0.00099	0.00513		0.02877	382	351	324	79	90	17
SW-GB2	0.00089	0.00589		0.02759	360	335	322	67	88	17

Notes: The minimum and second minimum RMSEs among all interpolations are in bold and italics, respectively. *,#See the footnotes denoted by the same symbols below Table 3a.

TABLE 5b

Comparison with Parametric Methods and Those Adjusted by the Shorrocks–Wan Method (Quintile Groups)

Estimation	Gini	MLD	Theil		H (10^{-5})			PG (10^{-5})		
Method	Exact	Approx.	Exact	Approx.	Overall	Q1	Q2–Q3	Overall	Q1	Q2–Q3
LN-SDG-β0.4	**0.00113**	*0.00509*	**0.02365**[#]	**0.02281**	**394**	**394**	**338**	**89**	**114**	**21**
LN-SDG$_{\beta}$-β0.4	0.00209	**0.00464**	0.03296[#]	0.03111	534	431	430	142	*122*	32
LN-SDG$_{gq}$-β0.4	0.00168	0.00467	*0.02382*[#]	*0.02686*	*506*	*431*	462	*122*	123	39
z-SDG-rh	0.00306	0.01186	0.05492	0.05529	866	1402	**338**	305	559	**21**
z-$_{h}$SDG-rh	0.00342	0.00627	0.05677	0.05664	1006	1272	975	231	304	105
z-SDG$_{\beta}$-rh	0.00270	0.01287	0.05055	0.05354	899	1433	430	301	562	32
z-$_{h}$SDG$_{\beta}$-rh	0.00305	0.00643	0.05088	0.05487	1029	1272	1014	225	304	110
z-SDG$_{gq}$-rh	0.00288	0.01131	0.05391	0.05427	880	1353	462	287	516	39
z-$_{h}$SDG$_{gq}$-rh	0.00319	0.00628	0.05557	0.05543	1035	1272	1031	229	304	115
β-LC	0.00242				650	547	571	291	551	64
GQ-LC	*0.00125*				635	718	605	165	282	60
Rasche	0.00247	0.00806		0.03799	727	539	699	229	260	135
SM	0.00244	0.00885	0.03487	0.03490	789	573	781	250	310	152
Dagum	0.00270	0.00998	0.03876	0.03857	823	569	799	276	342	170
GB2	0.00175	0.00580	0.03099	0.03128	613	475	604	175	223	101
SW-Rasche	0.00242	0.00622		0.03710	583	507	411	170	159	30
SW-LN	0.00485	0.01396		0.05978	752	455	421	257	140	32
SW-SM	0.00233	0.00603		0.03394	608	545	414	173	176	30
SW-Dagum	0.00255	0.00660		0.03682	628	522	413	185	181	30
SW-GB2	0.00169	0.00519		0.03094	518	456	409	136	156	30

Notes: The minimum and second minimum RMSEs among all interpolations are in bold and italics, respectively. [#]See the footnote denoted by the same symbol below Table 3b.

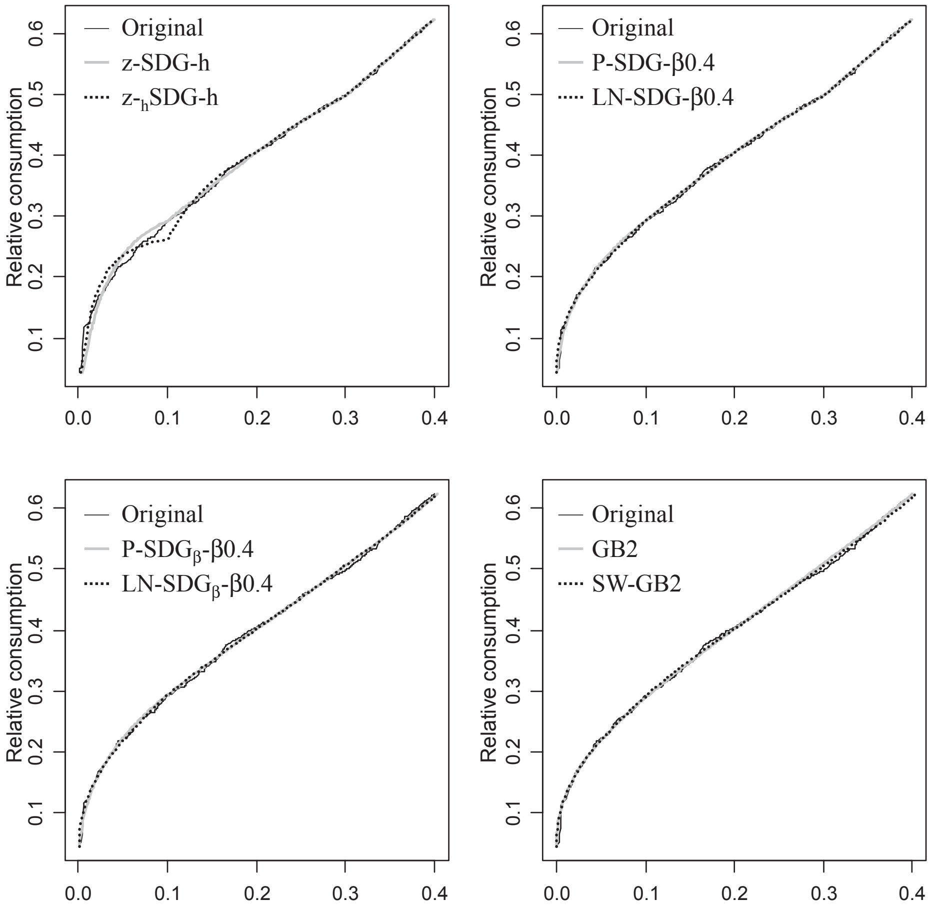

Figure 1. Examples of Derived Inverse CDFs for Per Capita Consumption; Cote d'Ivoire (Decile Groups)

charts the derivative of the LCs (inverse CDFs) estimated from the interpolations when applied to decile-grouped data for Cote d'Ivoire.

Looking at the results for individual countries, the β-LC (P-SDG$_β$-β0.4) outperforms other estimation methods for the intermediate point derivatives in most cases (precisely, in all seven countries or six out of the seven countries) when applied to decile-grouped data. As for the estimation from quintile-grouped data, LN-SDG$_{gq}$-β0.4 is superior to LN-SDG$_β$-β0.4 in most cases for the Gini and the overall $H(y)$ and $PG(y)$, whereas the superiority/inferiority is not clear for the other evaluation measures in terms of the number of countries in which smaller errors are attained. Both LN-SDG$_{gq}$-β0.4 and LN-SDG$_β$-β0.4 outperform LN-SDG$_g$-β0.4 and LN-SDG$_h$-β0.4 in most cases (except for Theil in the case of LN-SDG$_β$-β0.4). In comparisons among the pure SDG methods, z-SDG$_β$-h attains the smallest errors in a majority of countries for Gini, Theil, and $H(y)$ at the intermediate and highest deciles (D2–D10) when applied to decile-grouped data. Clear superiority cannot be observed for *PG*(*y*) at D2–D10. As for the estimation from quintile-grouped data, z-SDG$_{gq}$-rh is superior to z-SDG$_{\beta}$-rh in all countries for MLD, the overall *PG*(*y*) and the lowest part of *H*(*y*) and *PG*(*y*). z-SDG$_{gq}$-rh also attains smaller errors in five countries for the overall *H*(*y*), whereas z-SDG$_{\beta}$-rh yields better estimates in five and six countries for Gini and Theil, respectively. z-SDG$_{gq}$-rh outperforms z-SDG$_{g}$-h and z-SDG$_{h}$-h in a majority of countries for Gini, *H*(*y*) and *PG*(*y*) at the intermediate and highest quintiles (Q2–Q5).

5.5. *Comparisons with Existing Parametric Models and Those Adjusted by the Shorrocks–Wan Method*

At the end of this section, the proposed interpolation methods are compared with other types of estimation methods for the LC or size distribution of income, including parametric models and combinations of parametric models with the SW-adjustment method.

As parametric models of the LC, the β-LC (Kakwani, 1980a), GQ-LC (Villaseñor and Arnold, 1984, 1989), and Rasche model (Rasche *et al.*, 1980)[21] are used here, as well as the Log-Normal (LN), Singh–Maddala (SM) (Singh and Maddala, 1976), Dagum (Dagum, 1977), and Generalized Beta of the 2nd kind model (GB2) (McDonald, 1984) as parametric models of size distributions.

Both β-LC and GQ-LC frequently violate the conditions for the LC due to the occurrence of negative values near the left endpoint. However, since both models tend to be more accurate than the complete parametric models in terms of some evaluation measures such as accuracy of the Gini estimation, the RMSEs of Gini, *H*(*y*) and *PG*(*y*) are calculated for comparisons even in the cases that the estimated LC does not satisfy the conditions. Both models are fitted using simple regression methods in the same way as POVCAL (Datt, 1998; Chen *et al.*, 2001). Parameter σ for LN corresponding to the standard deviation of log-transformed incomes is estimated in the same way as Shorrocks and Wan (2009), as follows:

$$\sigma = \frac{1}{n-2}\sum_{i=2}^{n-1}(\Phi^{-1}(p_i) - \Phi^{-1}(l_i)). \tag{33}$$

Because LN is clearly inferior to other models, direct comparisons regarding goodness-of-fit are omitted, but the results after making adjustments by the SW-method are presented. The rest of the models are fitted by the following least squares method, attaching importance to the accuracy of the LC estimation:

$$\operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=2}^{n-1}[\log(p_i - l_i) - \log(p_i - L(p_i|\boldsymbol{\theta}))]^2, \tag{34}$$

where $L(p_i|\boldsymbol{\theta})$ denotes the estimated LC when the parameters of the respective model are set as $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots\}$. Goodness-of-fit around both endpoints is given

[21] The Ortega model (Ortega *et al.*, 1991) is omitted here because it is inferior to Rasche in most of the evaluation measures. Sarabia's unified model of Rasche and Ortega (Sarabia *et al.*, 1999) is also excluded because the unified model results in either Rasche or Ortega when applying the fitting procedure described in this subsection, and the estimates do not show any particular improvement in accuracy compared with Rasche. The Sarabia model $L_S(p|\delta, \gamma, \beta) = p^{\gamma}[1 - (1 - p)^{\delta}]^{\beta}$, where $\gamma \geq 0$, $\beta \geq 1$, is equivalent to Rasche when $\gamma = 0$ and to Ortega when $\beta = 1$.

greater importance in (34). The following simpler method is also conceivable; however, results are similar in practice:

$$(35) \qquad \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=2}^{n-1} (l_i - L(p_i|\boldsymbol{\theta}))^2.$$

The maximum likelihood estimation is not applied because there are frequently cases in which the parameters are unable to be estimated or are extraordinarily large, particularly when fitted to GB2. Furthermore, the estimation results are worse than those of the least squares method. A few inappropriate cases also occur when GB2 is fitted using (34). In such cases, SM or Dagum is substituted for GB2 according to the goodness-of-fit in terms of the least square errors in (34). In the GB2 fitting procedure, parameter p, q is regarded as extraordinarily large when either p or q exceeds $\log(10^2) \fallingdotseq 4.6$, $\log(10^6) \fallingdotseq 13.8$, respectively. Those upper bounds are determined based on the ranges of the parameter values obtained by Bandourian *et al.* (2002) by fitting GB2 to 23 countries. Note that the probability density function of GB2 in (36) is equivalent to SM's when $p = 1$ and to Dagum's when $q = 1$.

$$(36) \qquad \mathrm{GB2}(y|a, b, p, q) = \frac{ay^{ap-1}}{b^{ap}B(p, q)\left(1+(y/b)^a\right)^{p+q}}, \quad a, b, p, q > 0.$$

Shorrocks and Wan (2009) propose to adjust the outcome of the fitted parametric models by a two-step procedure outlined as follows. First, taking a sufficiently large integer J, let interval [0,1] be evenly divided into J subintervals and p_j-quantile of the fitted model be computed for each subinterval, where $p_j = (j + 0.5)/J$, $j = 1 \dots J$. The quantiles are grouped piecewise as specified, and linear transformation is applied to each group, maintaining the order of values. Then, the adjusted values are regrouped piecewise in a different way, and linear transformations are applied again to make the averages by income classes identical to those in the grouped data to be fitted. Shorrocks and Wan recommend $J = 1000$ because of the small computational burden coupled with practically sufficient precision, while they mention that higher precision is obtained by taking a larger J. In this paper, for comparisons with other methods, it is natural to make use of the subdivisions for the approximate calculation of inequality indices, i.e., $J = 5{,}000{,}000$ for Timor-Leste and the U.S. and $J = 1{,}000{,}000$ for the other five countries.

RMSEs of the various estimation methods are listed in Tables 5a and 5b. SW-Rasche, SW-GB2, etc., denote estimations by the respective models with adjustment by the SW-method. When those estimation methods are applied to decile-grouped data (Table 5a), a Hybrid interpolation P-SDG-β0.4 outperforms the existing parametric methods and those adjusted by the SW-method even if the intermediate point derivatives have to be estimated by the β-LC (P-SDG$_\beta$-β0.4). Comparisons with the best among the existing methods reveal that P-SDG-β0.4 attains 15–40 percent smaller errors, and P-SDG$_\beta$-β0.4 attains 10–30 percent smaller errors, except for $H(y)$, for which 1–5 percent smaller errors are attained. Although pure SDG interpolations z-SDG-h/z-$_\mathrm{h}$SDG-h and z-SDG$_\beta$-h/z-$_\mathrm{h}$SDG$_\beta$-h yield larger errors for Theil and the lowest part of $H(y)$ and $PG(y)$, they can stand comparison with the existing estimation methods on the whole. In sharp contrast to the interpolation methods, the parametric models generally suffer from ill-fitting at intermediate parts (D2–D9). RMSEs of $H(y)$ are more than two times larger, and those of $PG(y)$ are more than ten times larger at D2–D9. The SW-method succeeds in substantially reducing those large RMSEs yielded by the parametric models at D2–D9 to the levels slightly above those of P-SDG$_{\beta}$-β0.4. Among the parametric models and those adjusted by the SW-method, SW-GB2 performs the best on the whole. SW-LN is clearly inferior to the other combination of parametric models with the SW-method, in contrast to the results drawn by Shorrocks and Wan (2009). The conflicting conclusions might be caused by the data they used (March CPS) and/or the fitting procedures for parametric models (which are not specified in their study except for LN), although no definite reasons are known. The omitted comparison results for ventile-grouped data are similar to those for decile-grouped data, but one thing to be noted is that SDG improves the accuracy more substantially than the other methods.

When applied to quintile-grouped data (Table 5b), a Hybrid interpolation LN-SDG-β0.4 with known intermediate point derivatives outperforms the existing parametric methods and those adjusted by the SW-method. LN-SDG-β0.4 attains 15–35 percent smaller errors except for MLD, for which SW-GB2 attains approximately the same level of accuracy. The Hybrid interpolation with the intermediate point derivatives estimated by the GQ-LC also attains better estimations except for Gini and the intermediate part of $H(y)$ and $PG(y)$. As for Gini, a parametric model GQ-LC yields the second smallest errors below those of LN-SDG$_{gq}$-β0.4. SW-GB2 yields the second smallest errors for the intermediate part of $H(y)$ and $PG(y)$ but the differences from those of LN-SDG$_{gq}$-β0.4 (and z-SDG$_{gq}$-rh) are small. Thus, it can be said that LN-SDG$_{gq}$-β0.4 as well as LN-SDG-β0.4 are more accurate in comparison with the parametric models and their variants derived by the SW-adjustment on the whole. The pure SDG interpolations z-SDG$_{\cdot}$-rh and z-$_{h}$SDG$_{\cdot}$-rh are inferior to others except for the intermediate part of $H(y)$ and $PG(y)$ when applied to quintile-grouped data. As with applications to decile-grouped data, SW-GB2 is the best on the whole among the parametric models and those adjusted by the SW-method. The parametric models cannot avoid larger RMSEs at the intermediate part of $H(y)$ and $PG(y)$ although the relative differences from those of the interpolations are smaller compared to the cases of decile-grouped data.

Comparisons between the Hybrid interpolations and SW-GB2 in individual countries reveal that P-SDG-β0.4 attains smaller errors in most cases except for the lowest part of $PG(y)$, for which P-SDG-β0.4 attains smaller errors in five countries, when applied to decile-grouped data. P-SDG$_{\beta}$-β0.4 also attains smaller errors than SW-GB2 for MLD, Theil, and Gini in most cases and for every part of $PG(y)$ in five countries. As for $H(y)$, P-SDG$_{\beta}$-β0.4 attains smaller errors in only approximately half of the countries. When applied to quintile-grouped data, LN-SDG-β0.4 attains smaller errors for Gini, every part of $H(y)$, the intermediate and highest parts of $PG(y)$ in most cases and the lowest part of $PG(y)$ in five countries. As for MLD and Theil, LN-SDG-β0.4 attains smaller errors in only approximately half of the countries. Comparisons between LN-SDG$_{gq}$-β0.4 and

SW-GB2 reveal that both are approximately equal in terms of the number of countries for which smaller errors are attained, except for the intermediate and highest parts of $H(y)$ and $PG(y)$. SW-GB2 has the majority for the intermediate part of $H(y)$ and $PG(y)$, whereas LN-SDG$_{gq}$-β0.4 has the majority for the highest part of $H(y)$ and $PG(y)$. Thus, the clear superiority of LN-SDG$_{gq}$-β0.4 over SW-GB2 cannot be observed in terms of the number of the countries gained.

Charts of inverse CDFs derived from the estimated LCs in Figure 1 illustrate the goodness-of-fit of selected estimation methods when applied to decile-grouped data for Cote d'Ivoire. GB2 does not fit well at around the 30th percentile, whereas P-SDG-β0.4 and LN-SDG-β0.4 (with known intermediate point derivatives) fit well in every part. When intermediate point derivatives are estimated by the β-IC, the goodness-of-fit of the Hybrid interpolations get worse at around the 30th percentile but still maintain superiority over GB2. Although the SW-adjustment improves the goodness-of-fit of GB2 at around the 30th percentile, the superiority of the Hybrid interpolations cannot be overcome. The inverse CDF estimated by a pure SDG interpolation z-SDG-h is identical to those of P-SDG-β04 and LN-SDG-β0.4 at the 10th through the 90th percentile, so it is accurate on the intermediate interval; however, it suffers from poor fit at below the 10th percentile. If the leftmost intermediate point's derivative is replaced with an estimate by the harmonic mean, the goodness-of-fit of the pure SDG interpolation (z-$_h$SDG-h) is improved at the lower-end, but the exceptional treatment makes it rather worse at around the 10th percentile.

6. Empirical Comparisons among Interpolation Methods of the Concentration Curve

6.1. *Data and Evaluation Methods*

As shown in Section 5, the SDG interpolation can estimate the LC with approximately the same accuracy as the existing methods when applied to decile-grouped or more detailed grouped data, although it is inferior to the Hybrid interpolation. Its advantage over the Hybrid interpolation and the existing methods is consistent decomposability into the CCs for income components using the DG interpolation in (20) as its generalization.

From the sample survey data for the seven countries used in Section 5, we take 10 sets of income/expenditure data classified according to types of sources/purchased-items for five countries: for Bulgaria, incomes are classified into 6 and 27 categories and expenditures are classified into 7 categories; for Cote d'Ivoire, incomes are classified into 10 categories and expenditures into 6 categories; for Italy, incomes are classified into 4 and 16 categories and expenditures into 3 categories; for Peru, expenditures are classified into 9 categories; for Timor-Leste, expenditures are classified into 6 categories. Fifty sets of subsamples are generated from those classified data in the same manner as those in the previous section and aggregated into decile- and ventile-groups of per capita overall income/expenditure to which DG is applied. The detailed income categories are subdivisions of the broader categories in the Bulgarian and Italian data. Among the 10 income categories for Cote d'Ivoire, there are two deduction categories:

"depreciation of farm equipment" and "non-farm capital asset depreciation." Among the 16 Italian income categories, there are two deduction categories: "alimony and gifts paid" and "interest payable." Some of the categories are excessively minute; however, the original categories are used here without aggregation because such minutely categorized data illustrate what happens to the estimated CCs in such cases. Consumption data for China are excluded because of the small sample size and unclear classification system. Classified income data from the U.S. SCF are also excluded because of inconsistency with the total income data.

Assuming that N households in a subsample are arranged in ascending order of per capita income/expenditure, the accuracy of DG is assessed by the magnitude of estimation errors of the CCs in terms of the RMSE as follows:

$$(37) \qquad RMSE = \sqrt{\sum\nolimits_{i=1}^{k} \omega_i m_i \left(C_{DG}^{s}(p_i) - cc_i\right)^2 \Big/ \sum\nolimits_{i=1}^{k} \omega_i m_i},$$

where $C_{DG}^{s}(p)$ denotes the estimated CC for component s defined in (26), and $cc_i = \sum_{j\le i} \omega_j m_j \dfrac{y_j^s}{\mu_s} \Big/ \sum_{j=1}^{N} \omega_j m_j$ denotes the corresponding empirical CC at p_i (y_j^s is the per capita amount of component s in household j, and $\mu_s = \sum_{j=1}^{N} \omega_j m_j y_j^s \big/ \sum_{j=1}^{N} \omega_j m_j$). The estimation errors relative to the overall per capita income/expenditure are also evaluated as follows:

$$(38) \qquad RMSE = \sqrt{\sum\nolimits_{i=1}^{k} \omega_i m_i \left(\frac{C_{DG}^{s}(p_i)\mu_s - cc_i \cdot \mu_s}{y_i} \right)^2 \Big/ \sum\nolimits_{i=1}^{k} \omega_i m_i}.$$

The measure in (38) (termed the RMSE of the "relative CC," hereafter) is introduced as a comparable measure to the RMSE of the poverty gap in (30) in a loose sense, putting importance on accuracy at lower income levels.

The estimation errors are aggregated over categories and subsamples in the form of the RMSE, as follows:

$$(39) \qquad \sqrt{\overline{ARMSE}^2 + \frac{1}{K-1} \sum\nolimits_{k} \left(ARMSE_k - \overline{ARMSE} \right)^2},$$

$$\overline{ARMSE} = \frac{1}{K} \sum\nolimits_{k} ARMSE_k,\ ARMSE_k = \sqrt{\sum\nolimits_{j,s} w_{k,j}^{s} \left(RMSE_{k,j}^{s} \right)^2},$$

where $RMSE_{k,j}^{s}$ denotes the RMSE for component s in a subsample j of classified data k calculated in (37) or (38); $w_{k,j}^{s}$ is a reciprocal of the number of categories when taking simple averages, or the share of component s in terms of money when taking weighted averages for the aggregation of RMSEs in (37); as for the aggregation of RMSEs in (38), $w_{k,j}^{s}$ is set to a reciprocal of the number of categories. When deduction categories exist, the money share of components shall be replaced with its absolute value.

The accuracy of DG is also assessed by the absolute estimation errors of the Quasi-Gini indices (Q-Gini) relative to the Gini indices for the overall per capita income/expenditure with aggregation over categories and subsamples in the form of the RMSE, as follows:

$$\text{(40)}\qquad \sqrt{\overline{ARAE}^2 + \frac{1}{K-1}\sum\nolimits_k \left(ARAE_k - \overline{ARAE}\right)^2},$$
$$\overline{ARAE} = \frac{1}{K}\sum\nolimits_k ARAE_k,\ ARAE_k = \sum\nolimits_{j,s} w^s_{k,j}\left|\hat{\hat{I}}^s_{k,j} - \hat{I}^s_{k,j}\right| \Big/ I_k \cdot \bar{I},$$

where $\hat{I}^s_{k,j}$ denotes the Q-Gini for component s directly calculated from subsample j of classified data k; $\hat{\hat{I}}^s_{k,j}$ denotes the corresponding Q-Gini derived from interpolation applied to grouped data aggregated from the respective subsample; and I_k denotes Gini for the overall per capita income/expenditure distribution in the respective country directly calculated from the original sample. The reference value $\bar{I}\,(=0.41920)$ is a simple average of I_k over all ten datasets from five countries. The weighted averages attach importance to contributions to Gini for the overall income/expenditure. So far, the only competitor to DG as a consistent CC estimator is the piecewise linear interpolation method (termed "Linear," hereafter) in (41), which belongs to the C^0-class.

$$\text{(41)}\qquad \frac{p_{i+1}-x}{p_{i+1}-p_i}l^s_{i+1} + \frac{x-p_i}{p_{i+1}-p_i}l^s_i \quad \mathit{for} \quad x \in [p_i, p_{i+1}],\, i = 1\ldots n-1.$$

The composite Simpson rule (termed "Simpson," hereafter) in (42) is also regarded as a kind of piecewise quadratic interpolation method sequentially connecting three data points $\left(p_{2i}, l^s_{2i}\right)$, $\left(p_{2i+1}, l^s_{2i+1}\right)$, $\left(p_{2i+2}, l^s_{2i+2}\right)$ for $i = 0\ldots\frac{n}{2}-1$ by quadratic polynomials. Simpson enables a consistent CC interpolation and generally gives better estimates of Gini and Q-Gini in comparison with Linear,[22] although it does not necessarily satisfy monotonicity and convexity of the estimated LC for the overall income derived from the CC interpolations for components.

$$\text{(42)}\qquad \sum\nolimits_{k=0}^{2} \frac{\prod_{j\in\{1,2,3\}\backslash k}\left(x - p_{2i+j}\right)}{\prod_{j\in\{1,2,3\}\backslash k}\left(p_{2i+k} - p_{2i+j}\right)} l^s_{2i+k} \quad \text{for} \quad x \in [p_{2i}, p_{2i+2}],\, i = 0\ldots\frac{n}{2}-1.$$

Note that the number of income classes must be even, such as decile- and ventile-groupings, to apply Simpson.

6.2. *Empirical Comparisons among Estimation Methods for Data Points Derivatives and between DG and Other Interpolation Methods*

Table 6 shows the level of interpolation accuracy among ten classified datasets in terms of the RMSE defined as (37)–(40) when applied to decile-grouped

[22] The RMSE among seven countries for Gini estimated by the composite Simpson rule is 0.00434 when applied to decile-grouped data. It is worse than those of most of the interpolation methods in Table 5a, at approximately 0.001 or less, but better than the 0.00744 of the LN model and the 0.01139 of the linear interpolation. The composite Simpson rule is used for the calculation of the Gini index from decile-grouped data officially tabulated from the National Survey of Family Income and Expenditures in Japan.

TABLE 6

Comparison among Various Interpolation Methods of Concentration Curves for Income/Expenditure Components (Decile Groups)

Interpolation Method		Quasi-Gini		CC (10^{-5})						CC Relative to Income/ Consumption Level (10^{-5})		
				Simple Avg.			Weighted Avg.					
LC	CC	Simple Avg.	Weighted Avg.	Overall	D1	D2–D6	Overall	D1	D2–D6	Overall	D1	D2–D6
z-SDG-h	DG_g	0.00622	0.00276	2262	1754	899	946	258	283	204	336	**154**
	DG_a	0.00684	0.00345	2356	1758	908	1108	261	284	223	337	155
z-$_h$SDG-h	DG_g	**0.00621**	**0.00274**	**2261**	**1748**	**899**	**945**	244	**283**	**200**	**313**	154
	DG_a	0.00683	0.00342	2355	1750	908	1107	247	284	219	313	155
z-SDG$_\beta$-h	DG_g	0.00624	0.00280	2267	1754	899	953	259	283	206	341	154
	DG_a	0.00687	0.00348	2362	1759	909	1112	263	285	225	342	155
z-$_h$SDG$_\beta$-h	DG_g	0.00624	0.00277	2267	**1748**	899	953	244	283	201	**313**	154
	DG_a	0.00686	0.00344	2361	1750	908	1112	247	284	220	313	155
Simpson		0.00813	0.00392	2919	2082	1008	1610	**219**	299	310	459	163
Linear		0.01405	0.00986	3251	1796	916	1999	293	301	455	936	204

Note: The minimum RMSEs among all interpolations are in bold.

data. z-SDG-h and z-SDG$_\beta$-h are applied for interpolating the LCs of the overall per capita income/expenditure. The latter is a choice when the intermediate point derivatives are not given. In addition, z-$_h$SDG-h and z-$_h$SDG$_\beta$-h, the modified methods with replacement of the leftmost intermediate point's derivative by an estimate of the harmonic mean, are applied. The two-stage procedure in Section 4 is employed for the DG interpolation of the CCs for components. At the first stage, the arithmetic and geometric means are employed for the estimation of intermediate point derivatives and the right endpoint's derivative (see (9) and (14)). DG$_g$ and DG$_a$ denote DG with those estimation procedures for the data point derivatives, respectively.[23] The left endpoint's derivative shall be set to zero. Results for the harmonic mean are omitted because large estimation errors occur frequently. Because all categories in this study can be regarded as (almost) non-negative or non-positive, the intermediate point's derivative shall be set to zero when the slope is flat on either side of the intermediate point, as explained in Section 4. Note that subsamples generated by the cluster sampling with replacement from the original samples appear to make the accuracy of DG worse than the original samples when the categories have only a small number of non-zero records. This phenomenon forces DG to lower the accuracy relative to Simpson and Linear in some cases. For this reason, supplementary notes are given when the comparisons with Linear or Simpson differ from those based on the original samples.

Comparisons between z-SDG-h and z-$_h$SDG-h as well as between z-SDG$_\beta$-h and z-$_h$SDG$_\beta$-h reveal that replacing the leftmost intermediate point's derivative with the harmonic mean has only slight effects. Differences between z-$_\cdot$SDG-h and z-$_\cdot$SDG$_\beta$-h are also small. Thus, the common results among the four methods are described below.

DG$_g$ attains the smallest errors for the Q-Gini, CCs, and relative CCs, except for the weighted-average of the RMSEs of the CCs at the lowest decile, for which Simpson attains the smallest errors. In contrast to its best performance in terms of the weighted-average version, Simpson suffers from the worst performance in terms of the simple-averaged RMSE of the CCs at the lowest decile. On the whole, DG$_g$ outperforms DG$_a$, and both yield more accurate estimates than Simpson and Linear. DG$_\cdot$ maintains its superiority when applied to ventile-grouped data on the whole; however, some points should be noted: Linear yields smaller RMSEs than those of DG$_\cdot$ at the lowest ventile, but, when applied to the original samples, its RMSEs are at the same level as those of DG$_\cdot$ in terms of the simple-average version and larger than them in terms of the weighted-average version. Similar to the application to decile-grouped data, Simpson attains the smallest errors in terms of the weighted-average version at the lowest ventile, but its RMSEs are slightly larger than those of DG$_\cdot$ when applied to the original samples.

Looking at results for each set of the classified data, DG$_g$ attains smaller errors than those of DG$_a$ for the Q-Gini, overall CCs and relative CCs in a majority of the classified datasets. However, DG$_g$ suffers from larger errors than those of DG$_a$ for the lowest part of the CCs and relative CCs in a majority of the

[23] If the leftmost intermediate point's derivative is replaced with that of the harmonic mean for the CC estimation, the accuracy of the estimation at the lowest income class is slightly improved; however, because the effect is so small, the results are omitted here.

classified datasets. DG. attains smaller errors than those of Simpson in a majority of the classified datasets except for the weighted average of the RMSEs of the CCs and relative CCs at the lowest group (however, Simpson loses the majority when applied to ventile-grouped data aggregated from the original samples). All evaluation measures reveal that Linear yields larger errors than those of DG. in a majority of the classified datasets.

Using broad and detailed income categories for Bulgaria and Italy, the magnitude of inconsistencies between different levels of classification can be verified for DG_g estimates. The discrepancies tend to be larger along with the size of the respective broader categories, ranging from 10^{-3}–10^{-5} for the Q-Gini or 10^{-4}–10^{-5} in terms of contributions to the overall Gini. Thus, the discrepancies are regarded as sufficiently small. If priority should be given to independence from component classification, DG_a is an appropriate choice.

Some examples of the CC interpolations are given in Figure 2. The empirical CC of "net income from self-employment and entrepreneurial income" among the four Italian income categories (on the bottom right panel) changes its slope so

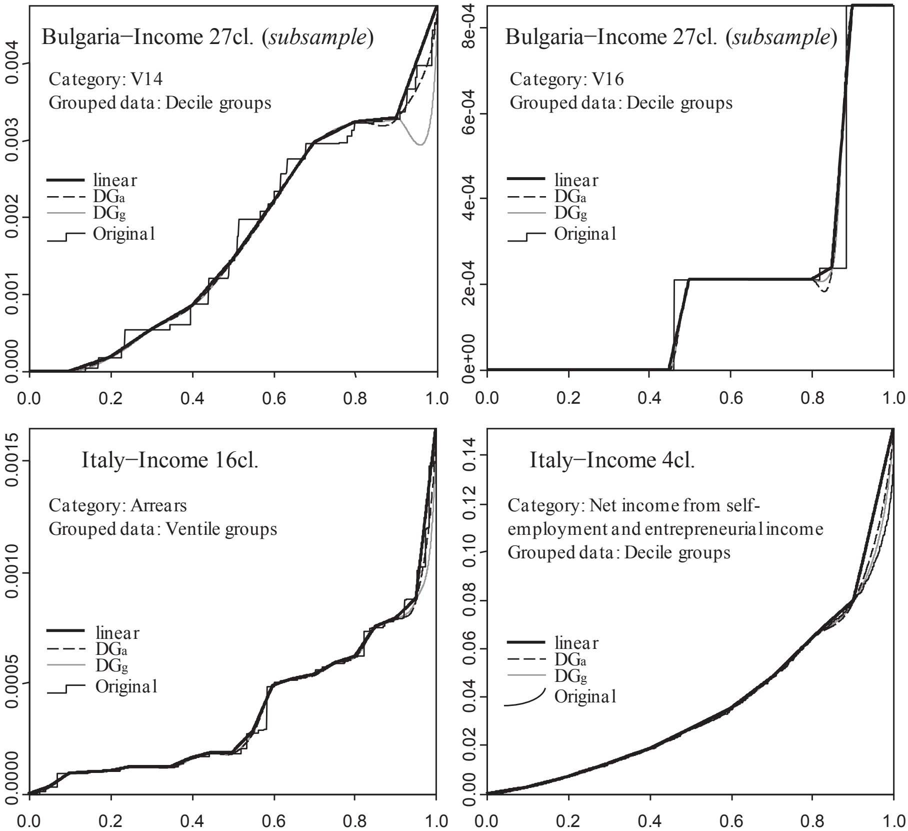

Figure 2. Examples of Concentration Curves for Income Components

Note: The concentration curves are multiplied by the respective money shares.

steeply on the right end that DG_g cannot fit the empirical CC perfectly, although it yields a better estimate compared to DG_a and Linear. Like this Italian example, DG_g fits to changes in slope on the right end generally better than DG_a, Simpson, and Linear in many cases. The other three examples in Figure 2 illustrate the results of interpolating the CCs when the corresponding empirical CCs cannot be regarded as smooth. The estimated CC of "V14" among the 27 Bulgarian income categories (on the top left panel) is apparently inappropriate for employing DG_g because the interpolation curve has a hollow on interval [0.9, 1]. DG_a also yields the CC estimate with a smaller hollow area on [0.8, 0.9]. In both cases, the tension parameter t_i in (23) determined by SDG for interpolating the LC of the overall income does not satisfy the monotonicity condition in (24) on the respective intervals. The estimated CCs of "V16" from the same data (on the top right panel) also have hollows on [0.8, 0.9] when DG_a and DG_g are applied. The latter's hollow is much smaller for this category. If the particular treatment is not employed in the case of a flat slope on either side of the intermediate data points, the CCs estimated by DG_a and DG_g are more wildly shaped. The inappropriate interpolations for "V14" and "V16" (although such problems do not occur when applied to grouped data directly aggregated from the original data) indicate that care should be taken to avoid excessively minute classification when employing DG. In those Bulgarian examples, the appropriateness of the estimated CCs can be examined by checking monotonicity, whereas the estimated CC of "Arrears" among the 16 Italian income categories satisfies monotonicity; nevertheless, Linear gives a better estimate than DG_a and DG_g. Like the two Bulgarian examples, the empirical CC of "Arrears" cannot be regarded as smooth.

7. Conclusions

This paper proposes two types of interpolation methods for the Lorenz curve. Both types preserve monotonicity and convexity essentially without restriction. One type is the SDG interpolation, the piecewise rational polynomial interpolation proposed by Stineman (1980) and Delbourgo (1989), and the other is the Hybrid interpolation, employing pieces of curves derived from parametric models as interpolants on both end intervals and the SDG interpolant on intermediate intervals. The pure SDG interpolation has the advantage of consistent decomposability into the CCs for income components. Use of the beta/GQ Lorenz curve is proposed for estimation of the Lorenz coordinates' derivatives (or class boundaries) when not given. Empirical comparisons using survey data for seven countries indicate that SDG attains approximately the same level of accuracy as the existing methods for inequality index estimation if excluding the upper-tail-sensitive measures such as the Theil index; moreover, SDG substantially reduces estimation errors of the LCs at intermediate intervals in comparison with the existing parametric models to levels slightly better than those of the Shorrocks–Wan method (2009) when applied to decile-grouped data or more minute aggregation (even if class boundaries are unavailable). The Hybrid interpolation attains higher accuracy than the existing methods even when applied to quintile-grouped data without class boundaries.

The proposals in this paper are not simple applications of the rational interpolation studied in the field of numerical analysis. The accumulation of research on parametric models for the LC greatly contributes to the adaptation of the interpolation technique to LC estimation. In particular, it may be appropriate to call the proposed methods the model-assisted interpolation methods in the cases in which the Lorenz coordinates' derivatives need to be estimated by the β-LC/GQ-LC. The excellent properties of the SDG interpolation and the enhancement and/or retention of its accuracy by utilizing the existing parametric models are expected to be useful for the analysis of economic inequality and poverty under situations in which access to microdata is still limited.

References

Bandourian, R., J. B. McDonald, and R. S. Turley, "A Comparison of Parametric Models of Income Distribution Across Countries and Over Time," Luxembourg Income Study Working Paper 305, 2002.

Bresson, F., "On the Estimation of Growth and Inequality Elasticities of Poverty with Grouped Data," *Review of Income and Wealth*, 55, 266–302, 2009.

Brown, J. A. C. and G. Mazzarino, "Drawing the Lorenz Curve and Calculating the Gini Concentration Index from Grouped Data by Computer," *Oxford Bulletin of Economics and Statistics*, 46, 273–8, 1984.

Champernowne, D. G., "A Model of Income Distribution," *Economic Journal*, 63, 318–51, 1953.

Chen, S. and M. Ravallion, "How Did the World's Poorest Fare in the 1990s?" *Review of Income and Wealth*, 47, 283–300, 2001.

Chen, S., G. Datt, and M. Ravallion, *POVCAL, A Program for Calculating Poverty Measures for Grouped Data*, World Bank, 2001 (http://go.worldbank.org/YMRH2NT5V0).

Cheong, K. S., "A Comparison of Alternative Functional Forms for Parametric Estimation of the Lorenz Curve," *Applied Economics Letters*, 9, 171–6, 2002.

Chotikapanich, D., D. S. P. Rao, and K. K. Tang, "Estimating Income Inequality in China Using Grouped Data and the Generalized Beta Distribution," *Review of Income and Wealth*, 53, 127–47, 2007.

Dagum, C., "A New Model for Personal Income Distribution: Specification and Estimation," *Economie Appliquée*, 30, 413–37, 1977.

Datt, G., "Computational Tools for Poverty Measurement and Analysis," FCND Discussion Paper 50, International Food Policy Research Institute, 1998.

Delbourgo, R., "Shape Preserving Interpolation to Convex Data by Rational Functions with Quadratic Numerator and Linear Denominator," *IMA Journal of Numerical Analysis*, 9, 123–36, 1989.

Delbourgo, R. and J. A. Gregory, "Shape Preserving Piecewise Rational Interpolation," *SIAM Journal on Scientific and Statistical Computing*, 6, 967–76, 1985a.

———, "The Determination of Derivative Parameters for a Monotonic Rational Quadratic Interpolant," *IMA Journal of Numeric Analysis*, 5, 397–406, 1985b.

Gastwirth, J. L., "A General Definition of the Lorenz Curve," *Econometrica*, 39, 1037–9, 1971.

Gastwirth, J. L. and M. Glauberman, "The Interpolation of the Lorenz Curve and the Gini Index from Grouped Data," *Econometrica*, 44, 479–83, 1976.

Gregory, J. A. and R. Delbourgo, "Piecewise Rational Quadratic Interpolation to Monotonic Data," *IMA Journal of Numerical Analysis*, 2, 123–30, 1982.

Kakwani, N. C., "Applications of Lorenz Curves in Economic Analysis," *Econometrica*, 45, 719–28, 1977.

———, "On a Class of Poverty Measures," *Econometrica*, 48, 437–46, 1980a.

———, *Income Inequality and Poverty: Methods of Estimation and Policy Applications*, Oxford University Press, New York, 1980b.

Kennickell, A. and J. Lane, *Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances*, The Federal Reserve Board, 2007 (http://www.federalreserve.gov/pubs/oss/oss2/method.html).

Kleiber, C. and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, John Wiley, Hoboken, NJ, 2003.

McDonald, J. B., "Some Generalized Functions for the Size Distribution of Income," *Econometrica*, 52, 647–63, 1984.

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

———, *Worlds Apart: Global and International Inequality 1950–2000*, Princeton University Press, Princeton, NJ, 2005.

Minoiu, C. and S. Reddy, "The Assessment of Poverty and Inequality Through Parametric Estimation of Lorenz Curves," ISERP Working Paper 07-02, Columbia University, 2007.

Ortega, P., G. Martín, A. Fernández, M. Ladoux, and A. García, "A New Functional Form for Estimating Lorenz Curves," *Review of Income and Wealth*, 37, 447–52, 1991.

Rasche, R. H., J. Gaffney, A. Y. C. Koo, and N. Obst, "Functional Forms for Estimating the Lorenz Curve," *Econometrica*, 48, 1061–2, 1980.

Ryu, H. K. and D. J. Slottje, "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve," *Journal of Econometrics*, 72, 251–74, 1996.

Sarabia, J. M., E. Castillo, and D. J. Slottje, "An Ordered Family of Lorenz Curves," *Journal of Econometrics*, 91, 43–60, 1999.

Shorrocks, A. and G. Wan, "Ungrouping Income Distributions: Synthesising Samples for Inequality and Poverty Analysis," in K. Basu and R. Kanbur (eds), *Arguments for a Better World: Essays in Honor of Amartya Sen: Ethics, Welfare, and Measurement*, Oxford University Press, New York, 414–34, 2009.

Singh, S. K. and G. S. Maddala, "A Function for the Size Distribution of Incomes," *Econometrica*, 44, 963–70, 1976.

Stineman, R. W., "A Consistently Well-Behaved Method of Interpolation," *Creative Computing*, 6, 54–7, 1980.

Thompson, W. A., Jr., "Fisherman's Luck," *Biometrics*, 32, 265–71, 1976.

UNU-WIDER, *World Income Inequality Database V. 2.0c*, May 2008 (http://www.wider.unu.edu/research/Database/en_GB/wiid/).

Villaseñor, J. A. and B. C. Arnold, "The General Quadratic Lorenz Curve," Colegio de Postgraduados Technical Report, 1984.

———, "Elliptical Lorenz Curves," *Journal of Econometrics*, 40, 327–38, 1989.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix 1: Formulas for Calculating Poverty/Inequality Indices from the Interpolated Lorenz Curve

Appendix 2: RMSEs of P-SDG-β/LN-SDG-β with Various Values of *m* for the β-IC