mathematics

MDPI

Article

Estimating Gini Coefficient from Grouped Data Based on Shape-Preserving Cubic Hermite Interpolation of Lorenz Curve

Songpu Shang [1] and Songhao Shang [2,*]

[1] School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou 450045, China; shangsongpu@ncwu.edu.cn
[2] State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Beijing 100084, China
* Correspondence: shangsh@tsinghua.edu.cn

Abstract: The Lorenz curve and Gini coefficient are widely used to describe inequalities in many fields, but accurate estimation of the Gini coefficient is still difficult for grouped data with fewer groups. We proposed a shape-preserving cubic Hermite interpolation method to approximate the Lorenz curve by maximizing or minimizing the strain energy or curvature variation energy of the interpolation curve, and a method to estimate the Gini coefficient directly from the coefficients of the interpolation curve. This interpolation method can preserve the essential requirements of the Lorenz curve, i.e., non-negativity, monotonicity, and convexity, and can estimate the derivatives at intermediate points and endpoints at the same time. These methods were tested with 16 grouped quintiles or unequally spaced datasets, and the results were compared with the true Gini coefficients calculated with all census data and results estimated with other methods. Results indicate that the maximum strain energy interpolation method generally performs the best among different methods, which is applicable to both equally and unequally spaced grouped datasets with higher precision, especially for grouped data with fewer groups.

Keywords: Gini coefficient; inequality; Lorenz curve; shape-preserving interpolation; cubic Hermite interpolation

Citation: Shang, S.; Shang, S. Estimating Gini Coefficient from Grouped Data Based on Shape-Preserving Cubic Hermite Interpolation of Lorenz Curve. *Mathematics* **2021**, *9*, 2551. https://doi.org/10.3390/math9202551

Academic Editors: Theodore E. Simos and Charampos Tsitouras

Received: 6 September 2021
Accepted: 8 October 2021
Published: 12 October 2021

1. Introduction

First proposed by Corrado Gini in 1912 [1], the Gini coefficient or Gini index has been widely used in describing inequalities in various fields, such as income/wealth [2], meteorology [3], ecology [4], hydrology [5], water resources [6], and the environment [7]. However, accurate estimation of the Gini coefficient has been a continuing topic of investigation, especially for grouped data [2,8–10].

The Gini coefficient is an important index that is related with the Lorenz curve (Figure 1), which was developed by Lorenz in 1905 [11] and shows the cumulative share of income or another variable under consideration ($y \in [0, 1]$) from different sections of population or another variable ($p \in [0, 1]$). The Lorenz curve is a non-negative, monotonic increasing, and convex curve [12,13]. In Figure 1, the straight diagonal line $y = p$ represents perfect equality in income or other distribution, while a Lorenz curve $y = L(p)$ generally lies beneath the line of perfect equality. The area between the line $y = p$ and curve $y = L(p)$, S_A, represents the inequality in income or other distribution. The greater the S_A, the greater the inequality in the distribution. Line segments $y = 0$ and $p = 1$ ($p, y \in [0, 1]$) with the greatest $S_A = 1/2$ represent another extreme distribution, the line of absolute inequality. The Gini coefficient, a scalar measurement of inequality, is defined as $2S_A$ for a Lorenz curve $y = L(p)$, which varies from 0 (representing perfect equality) to 1 (representing absolute inequality).

Besides estimation methods of the Gini coefficient directly from statistical data or its probability distribution, many estimation methods are based on the approximation of the corresponding Lorenz curve using curve fitting or interpolation methods, especially for grouped data.

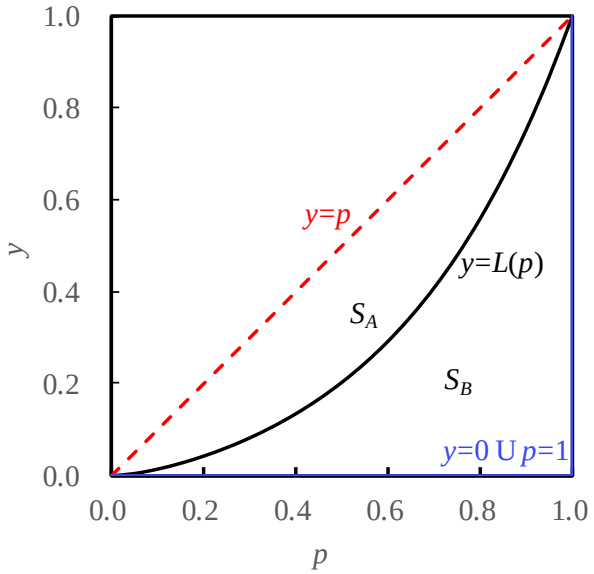


Figure 1. Sketch of the Lorenz curve showing the cumulative share of income or another variable under consideration ($y \in [0, 1]$) from different sections of population or another variable ($p \in [0, 1]$).

The curving fitting method fits the data with functions that meet the requirements of the Lorenz curve, i.e., non-negativity, monotonicity, and convexity [12,13]. However, an appropriate function is usually selected from a group of possible functions for specified data, and the selected function is generally not flexible enough to depict the complex variation of actual data globally [13]. Moreover, the fitted Lorenz curves can only represent the global trend and generally do not pass through all the data points, which is a common characteristic of the curve fitting method.

Contrary to the curve fitting method, the interpolation method constructs an interpolation curve that passes all data points. For grouped data, the simplest method to estimate the Gini coefficient is the trapezoidal rule, which approximates the Lorenz curve with piecewise linear interpolation. However, the trapezoidal rule always underestimates the Gini coefficient, and is generally taken as the lower limit of the Gini coefficient [14]. To improve the estimation accuracy of the Gini coefficient, higher order numerical integration methods that approximate the Lorenz curve with piecewise polynomial interpolation, such as Simpson's and Romberg's rules, can be used [14]. However, these numerical integration methods are generally applicable to equally spaced data except the trapezoidal rule. Furthermore, widely used Lagrangian, Hermite, and other interpolation curves do not necessarily preserve the non-negativity, monotonicity and convexity of the Lorenz curve [8]. Therefore, monotonicity and convexity should be considered in interpolating the Lorenz curve.

Another problem in interpolating the Lorenz curve with Hermite or other similar interpolators is the estimating of derivatives at intermediate points and endpoints [9], which has significant influence on the estimated Gini coefficient and should be considered with care in the interpolation.

The main objective of the present study is to develop a shape-preserving cubic Hermite interpolation method to approximate the Lorenz curve for estimating the Gini coefficient directly from the interpolation coefficients, where the derivatives of the Lorenz curve at intermediate points and endpoints are optimized by maximizing or minimizing the strain energy or curvature variation energy of the interpolation curve subject to non-negativity, monotonicity and convexity conditions. The applicability of this method was tested with 16 grouped datasets.

2. Materials and Methods

2.1. Conditions of Shape-Preserving Cubic Hermite Interpolation for the Lorenz Curve

Suppose we have $n + 1$ points in the Lorenz curve, (p_i, y_i), $i = 0, 1, \ldots, n$, where $0 = p_0 < p_1 < \ldots < p_n = 1$ is the cumulative fractions of the population or other variable of interest, and $0 = y_0 < y_1 < \ldots < y_n = 1$ is the cumulative fractions of income or another variable (Figure 2). The length of interval $I_i = [p_i, p_{i+1}]$, h_i, and slope of the line passing through (p_i, y_i) and (p_{i+1}, y_{i+1}), δ_i, are denoted as:

$$h_i = p_{i+1} - p_i,\ \delta_i = (y_{i+1} - y_i)/h_i,\ i = 0,\ 1,\ \ldots,\ n-1 \tag{1}$$

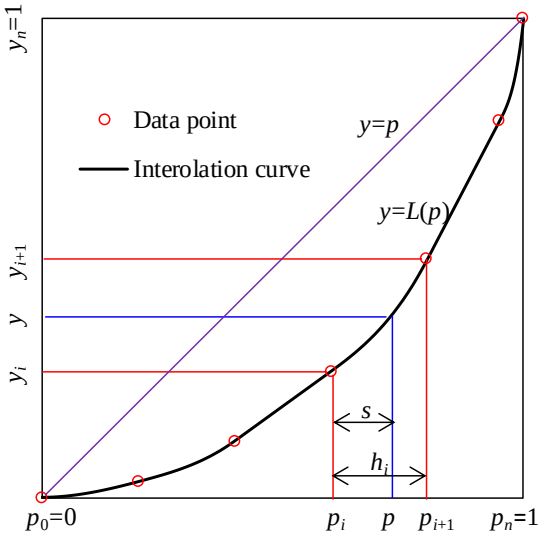


Figure 2. Sketch of data points and interpolated Lorenz curve.

Since the Lorenz curve is generally a convex curve, data points (p_i, y_i), $i = 0, 1, \ldots, n$, form a strictly convex set, i.e.,

$$0 < \delta_0 < \delta_1 < \cdots < \delta_{n-1} \tag{2}$$

A continuously differentiable function, $L(p)$, to approximate the Lorenz curve should pass all the interpolation points (p_i, y_i), $i = 0, 1, \ldots, n$, and has the same derivative as the Lorenz curve at each interpolation point (Figure 2). These conditions can be expressed as:

$$L(p_i) = y_i,\ i = 0,\ 1,\ \ldots,\ n \tag{3}$$

$$L^{(1)}(p_i) = d_i,\ i = 0,\ 1,\ \ldots,\ n \tag{4}$$

where d_i is the derivative of the Lorenz curve at p_i, $i = 0, 1, \ldots, n$. Generally, d_i, $i = 0, 1, \ldots, n$, are not known from statistics, and their estimation is essential for the interpolation and will be discussed later. The interpolation curve should have the same properties as the Lorenz curve, including non-negativity, monotonicity, and convexity.

A piecewise cubic Hermite interpolation curve that satisfies conditions (3) and (4) is [15,16]:

$$y = L(p) = L_i(s),\ s = p - p_i,\ p_i \le p \le p_{i+1},\ i\ = 0,\ 1,\ \ldots,\ n-1 \tag{5}$$

$$L_i(s) = y_{i+1}\frac{3h_is^2 - 2s^3}{h_i^3} + y_i\frac{h_i^3 - 3h_is^2 + 2s^3}{h_i^3} + d_{i+1}\frac{s^2(s-h_i)}{h_i^2} + d_i\frac{s(s-h_i)^2}{h_i^2},\ i = \ 0,\ 1,\ \ldots,\ n-1 \tag{6}$$

The first and second derivatives of the interpolation curve $L(p)$ are

$$L_i^{(1)}(s) = y_{i+1}\frac{6h_is - 6s^2}{h_i^3} + y_i\frac{-6h_is + 6s^2}{h_i^3} + d_{i+1}\frac{3s^2 - 2sh_i}{h_i^2} + d_i\frac{3s^2 - 4sh_i + h_i^2}{h_i^2},\ i = \ 0,\ 1,\ \ldots,\ n-1 \tag{7}$$

$$L_i^{(2)}(s) = \frac{2}{h_i}(3\delta_i - d_{i+1} - 2d_i) + \frac{6s}{h_i^2}(-2\delta_i + d_{i+1} + d_i),\ i = 0, 1, \ldots, n-1 \tag{8}$$

Following [17], the necessary and sufficient condition for the convexity of the interpolated Lorenz curve can be determined. The convexity of the Lorenz curve requires that $L_i^{(2)}(s) > 0$, $i = 0, 1, \ldots, n-1$, which is equivalent to $L_i^{(2)}(0) > 0$ and $L_i^{(2)}(h_i) > 0$, $i = 0, 1, \ldots, n-1$, because $L_i^{(2)}(s)$ is a linear function of s. From Equation (8), the convexity of the interpolation curve can be preserved if and only if

$$3\delta_i - 2d_i - d_{i+1} > 0, -3\delta_i + d_i + 2d_{i+1} > 0,\ i = 0, 1, \ldots, n-1 \tag{9a}$$

or

$$(3\delta_i - d_i)/2 < d_{i+1} < 3\delta_i - 2d_i,\ i = 0, 1, \ldots, n-1 \tag{9b}$$

If the convexity condition (9) is satisfied, $L^{(1)}(p)$ is a strictly monotonic increasing function. Therefore, the monotonic condition of $L(p)$, $L^{(1)}(p) > 0$, is equivalent to

$$L_0^{(1)}(0) = d_0 > 0 \tag{10}$$

In this case, the monotonicity of $L(p)$ is satisfied. Meanwhile, the non-negativity of mboxemphL (p) is also valid because $L(0) = 0$.

In summary, if the convexity condition (9) and monotonicity condition (10) are satisfied, $L(p)$ have the properties of non-negativity, monotonicity, and convexity, and can be used as an approximation of the Lorenz curve.

2.2. Construction of the Shape-Preserving Cubic Hermite Interpolation for the Lorenz Curve

Cubic spline, $S(p)$, is the most widely used cubic Hermite interpolants, which has continuous derivatives to order two and minimizes some energy functions, such as the widely used strain energy [18] and curvature variation energy [19]. These two energy functions can be approximated as [20]

$$E_s = \int_0^1 [S^{(2)}(p)]^2 \mathrm{d}p \tag{11}$$

$$E_c = \int_0^1 [S^{(3)}(p)]^2 \mathrm{d}p \tag{12}$$

where E_s and E_c are approximated forms of strain energy and curvature variation energy, respectively.

However, $S(p)$ constructed from the energy minimization does not necessarily preserve the properties of non-negativity, monotonicity or convexity. To construct a shape-preserving cubic Hermite interpolation for the Lorenz curve, we determined the derivatives d_i, $i = 0, 1, \ldots, n$, by minimizing the energy function (11) or (12) subject to conditions (9) and (10).

For $L(p)$, the approximated strain energy is

$$E_s = \int_0^1 [L^{(2)}(p)]^2 \mathrm{d}p = \sum_{i=0}^{n-1} \int_0^{h_i} [L_i^{(2)}(s)]^2 \mathrm{d}s = \sum_{i=0}^{n-1} \left\{ L_i^{(2)}(s) L_i^{(1)}(s) \Big|_0^{h_i} - \int_0^{h_i} L_i^{(3)}(s) L_i^{(1)}(s) \mathrm{d}s \right\} \tag{13}$$

Because $L_i^{(3)}(s) = 6(d_i + d_{i+1} - 2\delta_i)/h_i^2$ is a constant in the interval $I_i = [p_i, p_{i+1}]$, $i = 0, 1, \ldots, n-1$, the approximated strain energy can be deduced to be

$$E_s = \int_0^1 [L^{(2)}(p)]^2 \mathrm{d}p = \sum_{i=0}^{n-1} \frac{4}{h_i} \left[d_i^2 + d_i d_{i+1} + d_{i+1}^2 - 3\delta_i d_i - 3\delta_i d_{i+1} + 3\delta_i^2 \right] \tag{14}$$

Meanwhile, the approximated curvature variation energy is

$$E_c = \sum_{i=0}^{n-1} 36(d_i + d_{i+1} - 2\delta_i)^2 / h_i^4 \tag{15}$$

Therefore, the derivatives d_i, $i = 0, 1, \ldots, n$, can be determined from the following quadratic programming model:

$$\text{min. } E_s' = \sum_{i=0}^{n-1} \frac{1}{h_i}\left[d_i^2 + d_i d_{i+1} + d_{i+1}^2 - 3\delta_i d_i - 3\delta_i d_{i+1}\right] \tag{16}$$

or

$$\text{min. } E_c' = \sum_{i=0}^{n-1} (d_i + d_{i+1} - 2\delta_i)^2 / h_i^4 \tag{17}$$

subject to linear constraints (8) and (9). Compared with Equations (14) and (15), constants and items that have no influence on the optimal solution of the quadratic programming model are omitted in Equations (16) and (17).

Generally, the minimization of energy functions results in a straight and smooth spline. However, the straightness and smoothness are not intrinsic properties of a Lorenz curve. Following [21], we also used an alternative criterion to define the constrained Lorenz curve, i.e., maximizing the strain energy or curvature variation energy functions using Equation (18) or (19) subject to linear constraints (9) and (10).

$$\text{max. } E_s'' = \sum_{i=0}^{n-1} \frac{1}{h_i}\left[d_i^2 + d_i d_{i+1} + d_{i+1}^2 - 3\delta_i d_i - 3\delta_i d_{i+1}\right] \tag{18}$$

$$\text{max. } E_c' = \sum_{i=0}^{n-1} (d_i + d_{i+1} - 2\delta_i)^2 / h_i^4 \tag{19}$$

In contrast to the straight and smooth spline resulted from (16) or (17), spline resulted from (18) or (19) will contain relatively sharp curvatures or curvature variations. These two types of splines represent the most and least smooth interpolation curves that meet the requirements of the Lorenz curve, and will be compared to find which is more appropriate to approximate the Lorenz curve.

Now we have four optimization models with the objective functions of (16)–(19) subject to constraints (9) and (10). Due to diversity in empirical points (p_i, y_i), $i = 0, 1, \ldots, n$, and the estimated δ_i, $i = 0, 1, \ldots, n-1$, it is difficult to obtain the optimal solution analytically using the Kuhn-Tucker condition for nonlinear programming [22]. Because the feasibility region bounded by the linear constraints (9) and (10) are a convex set and the second order items in objective function (16) and (17) are positive definite and positive semi-definite, respectively, the minimizations of strain energy and curvature variation energy are both convex programming that have a unique optimal solution.

Moreover, from inequalities (2), (9), and (10), we have

$$0 < d_0 < \delta_0 < d_1 < \delta_1 < \cdots < d_{n-1} < \delta_{n-1} < d_n < 3\delta_{n-1} - 2d_{n-1} \tag{20}$$

Therefore, all decision variables (d_i, $i = 0, 1, \ldots, n$), and corresponding objective functions (18) and (19) are finite with lower and upper limits. Consequently, objective functions (18) and (19) have maximum in the feasibility region.

To solve the above quadratic programming models, several algorithms and optimization tools can be used [22], among which the Microsoft Excel Solver was chosen because of its wide availability and easy applicability [23].

2.3. Estimating Gini Coefficient from the Interpolated Lorenz Curve

The Gini coefficient, G, can be estimated directly from the coefficients of the interpolated Lorenz curve (Figure 2) using the following formula.

$$G = 1 - 2\int_0^1 L(p)\mathrm{d}p = 1 - 2\sum_{i=0}^{n-1}\int_0^{h_i} L_i(s)\mathrm{d}s = 1 - \sum_{i=0}^{n-1}\left[h_i(y_i + y_{i+1}) - h_i^2(d_{i+1} - d_i)/6\right] \quad (21)$$

Using trapezoidal rule that approximates the Lorenz curve with piecewise linear interpolation, the estimated Gini coefficient, G_T, is usually taken as its lower bound, which is

$$G_T = 1 - \sum_{i=0}^{n-1} h_i(y_i + y_{i+1}) \quad (22)$$

From Equations (21) and (22), G can be estimated from

$$G = G_T + \sum_{i=0}^{n-1} h_i^2(d_{i+1} - d_i)/6 = G_T + (h_{n-1}^2 d_n - h_0^2 d_0)/6 + \sum_{i=1}^{n-1} d_i(h_{i-1}^2 - h_i^2)/6 \quad (23)$$

Because of the convexity of the interpolated Lorenz curve, its first derivative, d (p), is a monotonic increasing function. Therefore, G estimated with Equation (23) is always greater than its lower bound of G_T. Meanwhile, for a grouped dataset with known y_i, $i = 0, 1, \ldots, n$, and h_i, $i = 0, 1, \ldots, n-1$, G depends only on the estimated derivatives of the interpolated Lorenz curve, especially the right part of the curve that has significantly greater derivative. Because the derivative at the left endpoint is small and the influence of derivatives at intermediate points can be partly (for unequally spaced data) or completely (for equally spaced data) counterbalanced for successive intervals from Equation (23), accurate estimation of derivatives at intermediate points and endpoints is crucial for accurate estimation of G, especially the derivative at the right endpoint.

When the grouped data is equally spaced with equal interval lengths of h, Equation (23) can be further simplified to

$$G = G_T + h^2(d_n - d_0)/6 \quad (24)$$

This equation further illustrates the importance of accurately estimating the derivatives at the endpoints, especially the right endpoint.

2.4. Data Used to Test the Method

To test the applicability of the interpolated Lorenz curve in estimating the Gini coefficient and to find whether minimizing or maximizing the strain energy or curvature variation energy is more preferable, we used 16 grouped datasets from published references to estimate their Gini coefficients, and compared the results with the "true" values estimated from all census data and estimates using other methods. These datasets include quintiles [24,25], quintiles plus the 95th percentile [10], and an unequally spaced dataset [26].

3. Results

The interpolated Lorenz curves for the grouped quintiles of US income census data in 2000 [24] by minimizing or maximizing the approximated strain energy are shown in Figure 3a, which shows that the former (Min. Es) is smoother than the latter (Max. Es). Meanwhile, negative and positive differences between these two interpolation curves occur alternatively in adjacent intervals, and the maximum absolute value of the differences in these intervals tend to increase with the cumulative population fraction (p). The maximum difference of 0.075 occurs at p = 0.92 in the last interval [0.8, 1]. Gini coefficients (G) estimated with these two interpolated Lorenz curves are 0.417 and 0.432, respectively; while G estimated with the interpolated Lorenz curves (not shown in Figure 3a for clarity) by minimizing (Min. Ec) and maximizing (Max. Ec) the approximated curvature variation

energy are 0.420 and 0.419, respectively. Compared with the estimates using the quintile rule of 0.422 [24], the estimated *G* corresponding to Max. Es interpolation of 0.432 is very close to the "true" value of 0.433 calculated from all the census data (Figure 3b), and is superior to other methods.

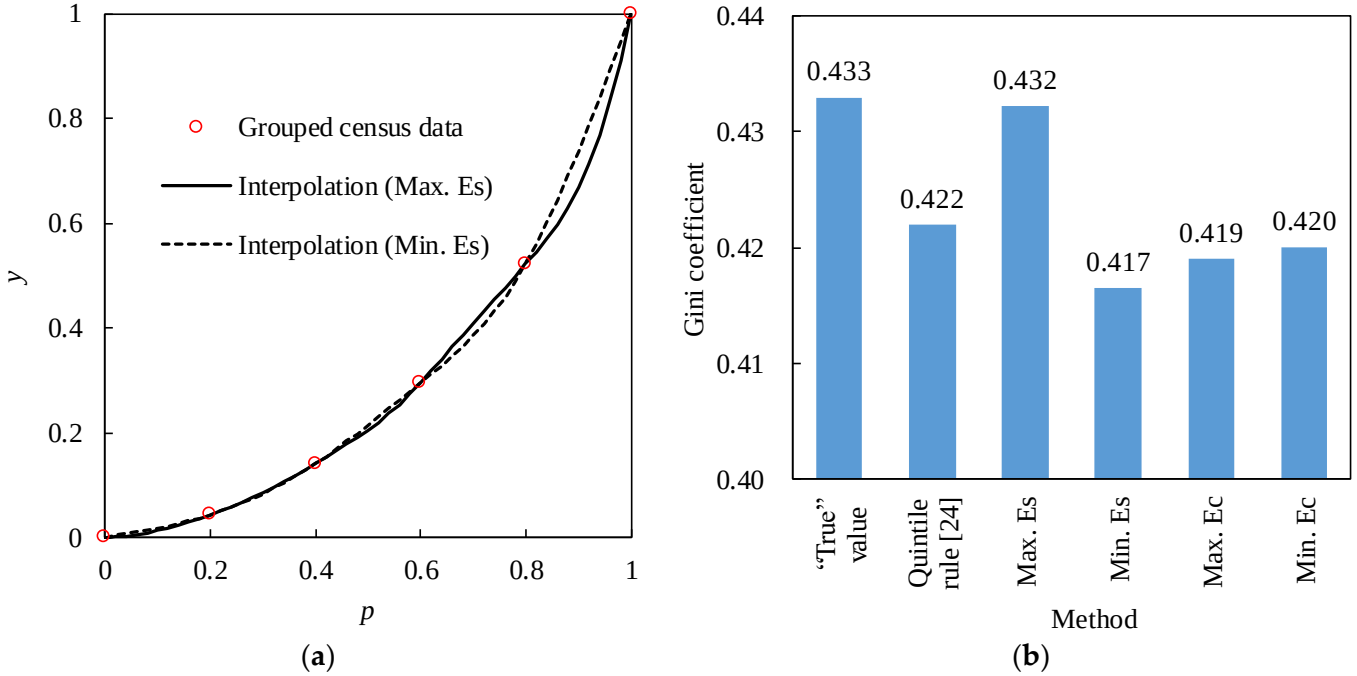


Figure 3. (**a**) Grouped quintiles of US income census data in 2000 [24] and interpolated Lorenz curves based on maximization (Max. Es) or minimization (Min. Es) of approximated strain energy; (**b**) Comparison of Gini coefficients estimated with different methods.

Using income quintiles of the United States in five-year intervals from 1947 to 2002 [20], the Gini coefficients were estimated with the present interpolation method and compared with those with Z-gradient and trapezoid rules [25] (Figure 4). For comparison purposes, all the *G* values were rounded to two decimals (or integers for 100G) as in [25]. From Figure 4, absolute errors between the actual values of 100G calculated from all census data and 100G estimated with the Z-gradient rule, trapezoid rule, Min. Es interpolation, and Max. Es interpolation range from 1–3, 2–3, 0–2, and 0–1 with the average values of 1.8, 2.8, 1.3, and 0.3, respectively. The average absolute error of *G* estimated with the Max. Es interpolation is only 12% to 27% of those with the other three methods. *G* is generally underestimated by the Z-gradient and trapezoid rules and the Min. Es interpolation method, while *G* estimates using the Max. Es interpolation method is the closest to the actual value.

The interpolated Lorenz curves for the grouped quintiles plus the 95th percentile of US income census data in 2010 [10] are shown in Figure 5a. Similar to Figure 3a, Figure 5a also shows that the interpolated curve by Min. Es interpolation is smoother than that by Max. Es interpolation, and their maximum differences in successive intervals tend to increase with p_i for $p_i < 0.95$, which reach the peak of 0.043 when $p_i = 0.89$. However, the difference becomes smaller for $p_i > 0.95$ due to the added data at $p_i = 0.95$ and the shorter interval. *G* estimated with these two interpolated Lorenz curves are 0.470 and 0.469, respectively, which are the same or very close to the "true" value of 0.470. *G* estimated with Max. Ec and Min Ec also give satisfactory results of 0.467 and 0.468, respectively, which are slightly poorer than *G* estimated with Max. Es and Min Es. Among six methods used in [10], four methods also gave very good *G* estimates of 0.467 to 0.470, while two other methods resulted in poorer estimates compared to the aforementioned results (Figure 5b).

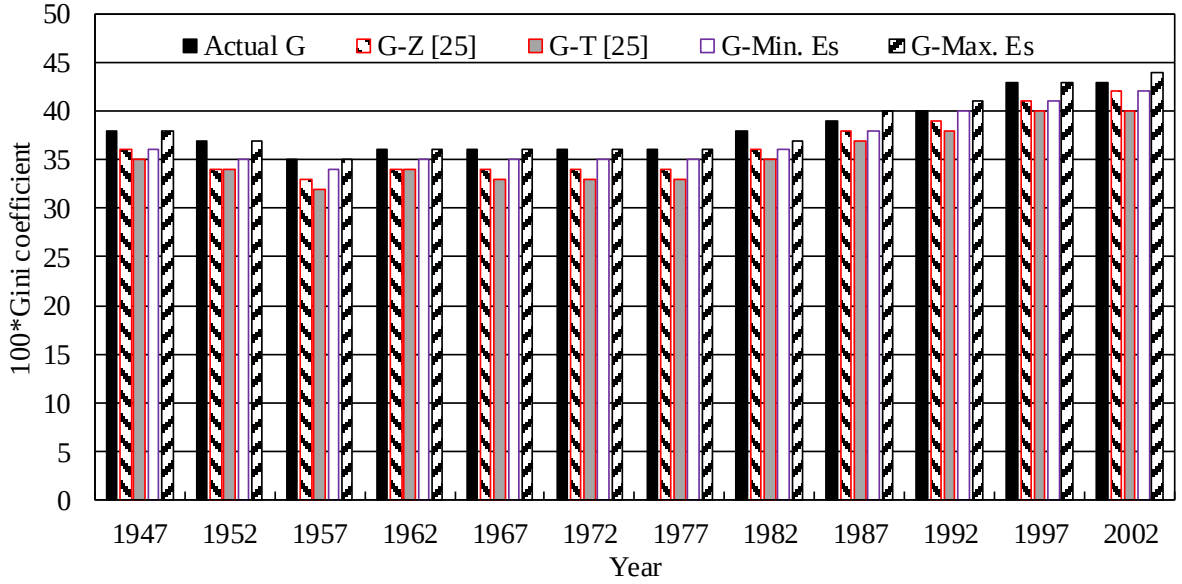
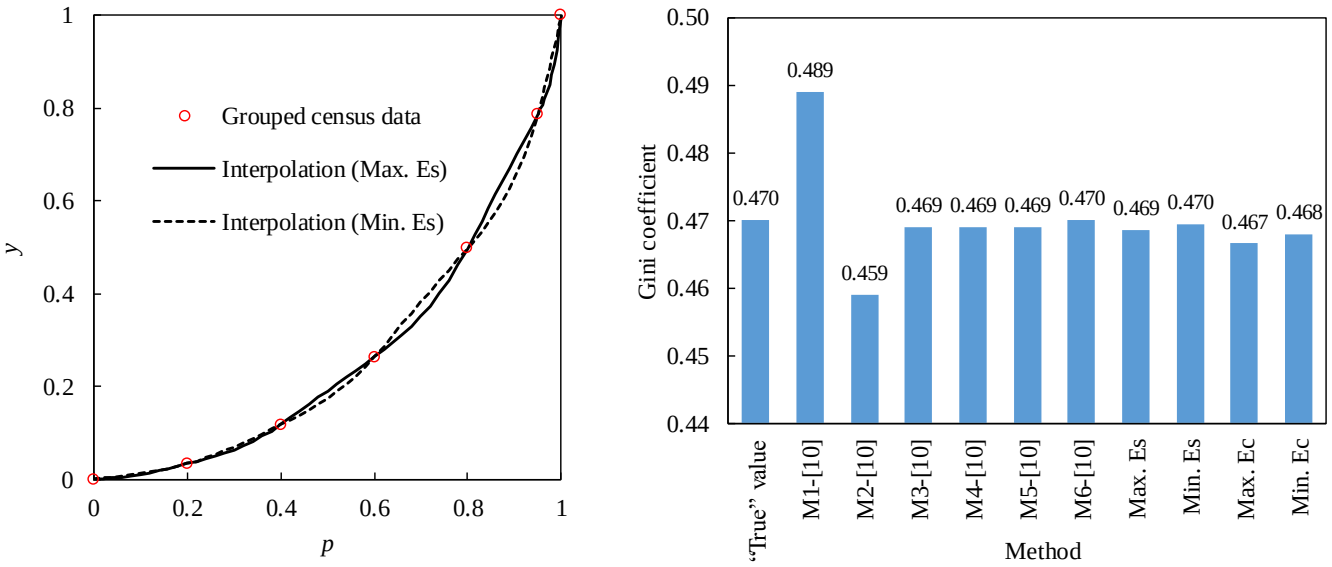


Figure 4. Comparison of Gini coefficients estimated with different methods for income quintiles of the United States in five-year intervals from 1947 to 2002. G-Z and G-T are estimates of Gini coefficient using Z-gradient and trapezoid rules [25], respectively.

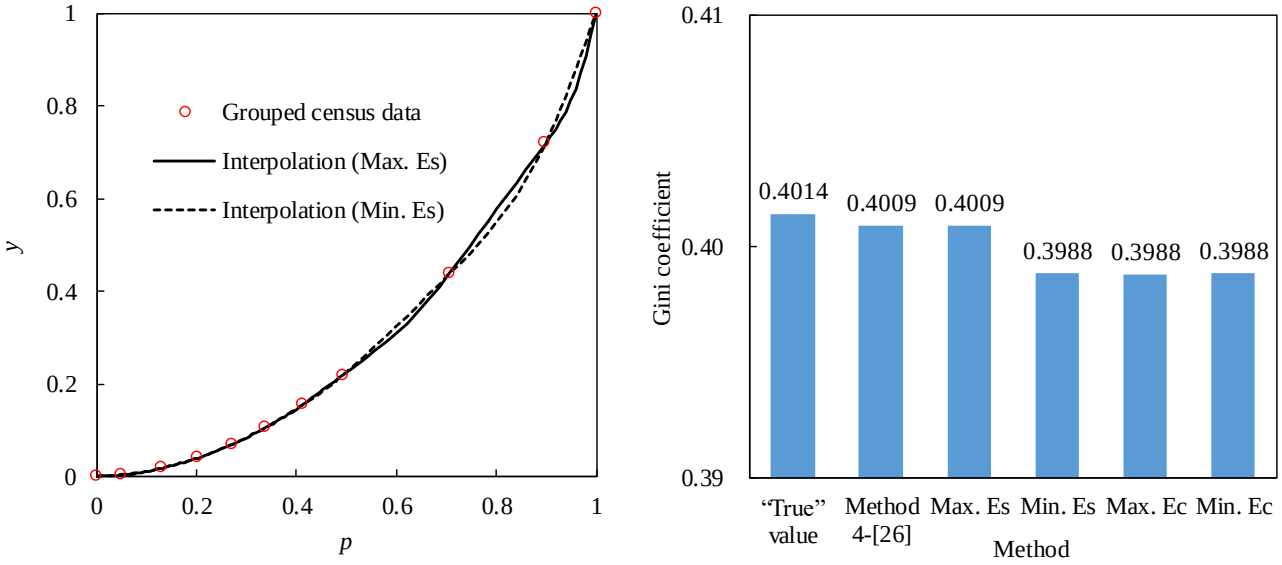


(**a**) (**b**)

Figure 5. (**a**) Grouped quintiles plus the 95th percentile of US income census data in 2010 [10] and interpolated Lorenz curves based on maximization (Max. Es) or minimization (Min. Es) of approximated strain energy; (**b**) Comparison of Gini coefficients estimated with different methods, where M1-[10] to M6-[10] are six methods used in [10].

The interpolated Lorenz curves for the income data of 10 unequally spaced groups [26] are shown in Figure 6a. For $p_i < 0.492$ that has interval lengths less than 0.08, the differences between two interpolated Lorenz curve with Max. Es and Min. Es are very small. However, the difference becomes greater with the longer interval when $p_i > 0.492$, and reaches its peak of 0.042 at $p_i = 0.959$. Gini coefficients (G) estimated with these two interpolated Lorenz curves are 0.3988 and 0.4009 (Figure 6b), respectively. G is slightly underestimated using the Min. Es interpolation method. Meanwhile, the estimated G based on Max. Es is the same as the estimates using the method 4 in [26], which are both very close to the "true"

value of 0.4014 calculated from all the census data. Values of G estimated with Max. Ec and Min. Ec are close to that estimated with Min Es., but they are all slightly poorer than G estimated with Max. Es.



(**a**) (**b**)

Figure 6. (**a**) Income data of 10 unequally spaced groups [10] and interpolated Lorenz curves based on maximization (Max. Es) or minimization (Min. Es) of approximated strain energy; (**b**) Comparison of Gini coefficients estimated with different methods.

4. Discussion and Conclusions

The Gini coefficient is widely used in describing inequalities in many fields, but its accurate estimation is still difficult for grouped data with fewer groups. We proposed a shape-preserving cubic Hermite interpolation method to approximate the Lorenz curve by maximizing or minimizing the approximated strain energy or curvature variation energy of the interpolated curve, which can then be used to estimate the Gini coefficient directly from the interpolation coefficients. Case studies with 16 grouped quintiles or unequally spaced datasets (Figures 3–6) showed that the maximum strain energy interpolation method generally performs the best among different methods compared with the "true" Gini coefficient calculated with all census data.

The proposed shape-preserving cubic Hermite interpolation method for the Lorenz curve has several advantages. First, the interpolated curves pass all data points, which is preferable to fitted curves that are generally not flexible enough to depict the complex variation of actual data globally and do not pass all data points [13]. Second, the interpolated curve preserves the essential requirements of the Lorenz curve, i.e., non-negativity, monotonicity, and convexity [17]. Third, derivatives at intermediate points and endpoints are optimized at the same time by maximizing or minimizing the energy functions subject to non-negativity, monotonicity, and convexity conditions (9) and (10), which is much simpler than some other interpolation methods that determine derivatives at intermediate points and endpoints with different methods [9]. Because accurate estimation of derivatives at intermediate points and endpoints, especially the derivative at the right endpoint, is crucial for accurate estimation of G, the simultaneous estimation of derivatives at intermediate points and endpoints may be a possible reason for the higher precision of the estimated G. Fourth, the method is applicable to both equally and unequally spaced grouped datasets with higher precision than other methods, especially for datasets with fewer groups (Figures 3–5). The estimated Gini coefficients using the maximizing strain

energy rule are better than or close to other methods for most of the case studies.

The Lorenz curve generated from the minimizing strain/curvature variation energies are smoother than those from the maximizing strain/curvature variation energies. These two types of interpolated Lorenz curves represent the most and least smooth interpolation curves that meet the requirements of the Lorenz curve, and their differences tend to increase with the population fraction and interval length (Figures 3, 5 and 6). The Lorenz curve interpolated from the maximizing strain energy generally contains relatively sharp curvatures that may better reflect the distribution of income or other variables under consideration within each group, and results in better estimation of derivatives at intermediate points and endpoints, which is the possible reason for better estimates for the Gini coefficient using the maximizing strain energy rule compared with other energy rules.

Author Contributions: Conceptualization, S.S. (Songpu Shang) and S.S. (Songhao Shang); methodology, S.S. (Songpu Shang); validation, S.S. (Songhao Shang); formal analysis, S.S. (Songpu Shang); investigation, S.S. (Songhao Shang); writing—original draft preparation, S.S. (Songpu Shang) and S.S. (Songhao Shang); writing—review and editing, S.S. (Songhao Shang); visualization, S.S. (Songhao Shang); funding acquisition, S.S. (Songhao Shang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 51839006.

Data Availability Statement: The data presented in this study are available in references cited in Section 3.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ceriani, L.; Verme, P. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econ. Inequal.* **2012**, *10*, 421–443. [CrossRef]
2. Furman, E.; Kye, Y.; Su, J. Computing the Gini index: A note. *Econ. Lett.* **2019**, *185*, 108753. [CrossRef]
3. Guo, E.; Wang, Y.; Jirigala, B.; Jin, E. Spatiotemporal variations of precipitation concentration and their potential links to drought in mainland China. *J. Clean. Prod.* **2020**, *267*, 122004. [CrossRef]
4. Damgaard, C.; Weiner, J. Describing inequality in plant size or fecundity. *Ecology* **2000**, *81*, 1139–1142. [CrossRef]
5. Masaki, Y.; Hanasaki, N.; Takahashi, K.; Hijioka, Y. Global-scale analysis on future changes in flow regimes using Gini and Lorenz asymmetry coefficients. *Water Resour. Res.* **2014**, *50*, 4054–4078. [CrossRef]
6. Qi, H.W.; Shang, S.H.; Li, J. Quantitative evaluation on spatial heterogeneity of water resources in China. *J. Hydroelectr. Eng.* **2020**, *39*, 28–38. [CrossRef]
7. Soares, T.C.; Fernandes, E.A.; Toyoshima, S.H. The CO_2 emission Gini index and the environmental efficiency: An analysis for 60 leading world economies. *EconomiA* **2018**, *19*, 266–277. [CrossRef]
8. Gastwirth, J.L.; Glauberman, M. The interpolation of the Lorenz curve and Gini index from grouped data. *Econometrica* **1976**, *44*, 479–483. [CrossRef]
9. Okamoto, M. Interpolating the Lorenz Curve: Methods to Preserve Shape and Remain Consistent with the Concentration Curves for Components. *Rev. Income Wealth* **2014**, *60*, 349–384. [CrossRef]
10. Lyon, M.; Cheung, L.C.; Gastwirth, J.L. The advantages of using group means in estimating the Lorenz curve and Gini index from grouped data. *Am. Stat.* **2016**, *70*, 25–32. [CrossRef]
11. Lorenz, M.O. Methods of measuring the concentration of wealth. *Publ. Am. Stat. Assoc.* **1905**, *9*, 209–219. [CrossRef]
12. Sarabia, J.M.; Castillo, E.; Slottje, D.J. An ordered family of Lorenz curves. *J. Econ.* **1999**, *91*, 43–60. [CrossRef]
13. Wang, Z.; Ng, Y.-K.; Smyth, R. A general method for creating Lorenz curves. *Rev. Income Wealth* **2011**, *57*, 561–582. [CrossRef]
14. Darkwah, K.A.; Nortey, E.N.N.; Lotsi, A. Estimation of the Gini coefficient for the lognormal distribution of income using the Lorenz curve. *SpringerPlus* **2016**, *5*, 1196. [CrossRef]
15. Fritsch, F.N.; Carlson, R.E. Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **1980**, *17*, 238–246. [CrossRef]
16. Moler, C. *Numerical Computing with MATLAB*; The MathWorks, Inc.: Natick, MA, USA, 2004.
17. Schrag, H.; Kramer, W. A simple necessary and sufficient condition for the convexity of interpolated Lorenz curves. *Statistica* **1993**, *53*, 167–170.
18. Jaklič, G.; Žagar, E. Planar cubic G^1 interpolatory splines with small strain energy. *J. Comput. Appl. Math.* **2011**, *235*, 2758–2765. [CrossRef]

19. Jaklič, G.; Žagar, E. Curvature variation minimizing cubic Hermite interpolants. *Appl. Math. Comput.* **2011**, *218*, 3918–3924. [CrossRef]
20. Li, J.C. Constructing planar C^1 cubic Hermite interpolation curves via approximate energy minimization. *J. Math. Res. Appl.* **2019**, *39*, 433–440. [CrossRef]
21. Durrans, S.R.; Burian, S.J.; Nix, S.J.; Hajji, A.; Pitt, R.E.; Fan, C.-Y.; Field, R. Polynomial-based disaggregation of hourly rainfall for continuous hydrologic simulation. *J. Am. Water Resour. Assoc.* **1999**, *35*, 1213–1221. [CrossRef]
22. Winston, W.L. *Operations Research: Applications and Algorithms*, 4th ed.; Thomson Brooks/Cole: Belmont, CA, USA, 2004.
23. Winston, W.L. *Microsoft Excel 2016 Data Analysis and Business Modeling*, 5th ed.; Microsoft Press: Redmond, WA, USA, 2016.
24. Gerber, L. A quintile rule for the Gini coefficient. *Math. Mag.* **2007**, *80*, 133–135. [CrossRef]
25. Golden, J. A simple geometric approach to approximating the Gini coefficient. *J. Econ. Educ.* **2008**, *39*, 68–77. [CrossRef]
26. Gastwirth, J.L. The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* **1972**, *54*, 306–316. [CrossRef]