Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n1p85

The Gini coefficient and the case of negative values

By De Battisti, Porro, Vernizzi

Published: 26 April 2019

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 01, April 2019, 85-107
DOI: 10.1285/i20705948v12n1p85

The Gini coefficient and the case of negative values

Francesca De Battisti[a], Francesco Porro*[b], and Achille Vernizzi[a]

[a] *Università degli Studi di Milano*
[b] *Università degli Studi di Milano-Bicocca*

Published: 26 April 2019

When calculating the Gini coefficient for distributions which include negative values, the Gini coefficient can be greater than one, which does not make evident its interpretation. In order to avoid this awkward result, common practice is either replacing the negative values with zeros, or simply dropping out units with negative values. We show how these practices can neglect significant variability shares and make comparisons unreliable. The literature also presents some corrections or normalizations which restrict the modified Gini coefficient into the range [0-1]: unluckily these solutions are not free of deficiencies. When negative values are included, the Gini coefficient is no longer a concentration index, and it has to be interpreted just as relative measure of variability, taking account of its maximum inside each particular situation. Our findings and suggestions are illustrated by an empirical analysis, based on the Survey of Household Income and Wealth, released by Banca d'Italia.

keywords: Gini coefficient, negative values, concentration, variability

1 Introduction

While labour earnings are always non-negative, a business may lose money in any year. Hence, income data, financial assets (such as capital gains) and money transfers typically can include also negative values. The same happens to tax systems which admit negative income taxes, that can originate, for example, from child allowances. In the literature, the issue of negative values arises also in wealth distributions as described for example

*Corresponding author:francesco.porro1@unimib.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

in Amiel et al. (1996) and in Jenkins and Jantti (2005). The most used measure of income inequality is the Gini coefficient of concentration. In Gastwirth (1975) the author, addressing an issue originally proposed in Budd (1970), found a lower and an upper bound for the Gini coefficient. His approach fits well also in presence of negative values. When a distribution includes negative values, as Castellano (1937) remarked, the Lorenz curve lies below the X-axis (here we suppose that the mean of the variable is positive) and the Gini coefficient can assume values greater than one, as it is also observed by Hagerbaumer (1937), Pyatt et al. (1980), and Lambert and Yitzhaki (2013). In order to avoid these issues, the most common and simple practices are either to eliminate the observations with negative values or convert them into zero, with the latter also suggested by a very important international organization (OECD, 2015). These two methods have two important drawbacks: first, a significant proportion of information is neglected due to the elimination of negative values or to the setting them to zero; second, ignoring negative values can lead to unreliable comparisons among different distributions. In order to restrict the Gini coefficient to the interval $[0, 1]$, Chen et al. (1982, 1985) modify the normalizing factor by adding a component that depends on the distribution of negative values and on the smallest positive values, which are enough to compensate for the former. Chen et al.'s method was subsequently completed by Berrebi and Silber (1985), by providing a correct expression for the general case when the compensation does not take place exactly in correspondence to a particular unit. Chen et al.'s correction has the advantage of decreasing the modified Gini coefficient, whatever egalitarian transfer occurs. The major drawback of this procedure is that it does not refer to a theoretical extreme situation, and so it is an ad hoc procedure. Another drawback is that the CTR-BS index can have some unreasonable behaviours in particular situations. In the paper Raffinetti et al. (2015), the authors provide a deep investigation about this issue and suggest a normalization that keeps into account the potential maximum inequality, stating appropriate conditions for the application of their normalization. In presence of negative values, the issue of the normalization of the Gini coefficient, through the identification of an upper bound for the inequality index, is quite delicate. As pointed out by Cowell and Van Kerm (2015) (in the note 13, pag 701), the extreme situation which should be corresponding to the upper bound of the inequality index is debatable. If we allow the units to enter into debt with the others and we do not fix any restrictions to the transfers, it is not possible to identify the "maximum" situation of inequality, since we can do an unitary transfer from the poorest to the richest unit, and we can repeat such transfer, at least theoretically, indefinitely.

van de Ven (2001) defined the distribution of "perfect inequality" in presence of negative values. This extreme situation basically is the case when the maximum proportion q of the population owns the minimum value of the variable at stake, while the remaining proportion $1 - q$ has the maximum one, keeping fixed the mean of the distribution. A worthy remark is that the extreme situation proposed by van de Ven (2001) is the same one described by Castellano (1937).

Starting from the observation of Frosini (1984) at pag 376 that "...the concentration, unlike variability, can only be related to non-negative variables", perhaps we should give up the demand that the Gini coefficient is a concentration index in presence of negative

values, and we have to interpret it by making use of other complementary measures. The Raffinetti et al.'s index, which by itself is not a solution to the problem, could integrate the information provided by the standard Gini coefficient.

The purpose of the present paper is to review the existing methods for managing the Gini coefficient in presence of negative values, and to provide some general guidelines suitable for a such situation. The paper is organized as follows. The next section provides the basic notations and some initial settings; sections 3 and 4 illustrate in detail the most adopted procedures: the one which replaces the negative values with zeros, and the one which drops out units with negative values, respectively. The CTR-BS index and its implications are thoroughly analysed in section 5, whilst section 6 reconsiders the Raffinetti et al.'s index, suggesting how it could be properly used. Section 7 considers how the indexes reported in this article respect the Pigou-Dalton principle. Section 8 provides an empirical application to data from the Survey of Household Income and Wealth (SHIW) and illustrates how the indexes considered in the article work. Section 9 concludes.

2 Some notations and the initial settings

Let X be a statistical variable, which can assume both negative and positive values. Let

$$(x_1, x_2, \ldots, x_N, x_{N+1}, \ldots, x_M)$$

be the values of X in non-decreasing order. We suppose, without loss of generality, that the first N values $(x_1, \ldots, x_N)$ are negative, while the remaining $(x_{N+1}, \ldots, x_M)$ values are non-negative. We assume that the sum T_a of the non-negative values:

$$T_a = \sum_{i=N+1}^{M} x_i$$

is higher than the sum T_n of the absolute negative values:

$$T_n = \sum_{i=1}^{N} |x_i|,$$

meaning that:

$$T_a - T_n = \sum_{i=N+1}^{M} x_i - \sum_{i=1}^{N} |x_i| > 0.$$

Let G be the Gini coefficient, defined by

$$G = \frac{S}{2(M-1)(T_a - T_n)}, \tag{1}$$

where S denotes the sum of absolute differences:

$$S = \sum_{i=1}^{M} \sum_{j=1}^{M} |x_i - x_j|.$$

Remark 1 *In the following, we will calculate the Gini coefficient by the formula (1), even if, whenever the number of units M is large enough, G can also be approximated by*

$$G = \frac{\Delta_R}{2\mu}, \tag{2}$$

where

$$\Delta_R = \frac{S}{M^2} \qquad \textit{and} \qquad \mu = \frac{T_a - T_n}{M}$$

are the Gini mean difference and the mean of the distribution, respectively. We decided to use expression (1) because it implies that G ranges in the interval $[0, 1]$. Moreover, formula (1) is the original definition of G, provided by Corrado Gini (1914). Unfortunately, it is well-known in the literature that using such definition G does not satisfy the Daltons's principle of population. To overcome this issue, G should be defined as stated in (2): this can be performed, by a simple replacement of $(M - 1)$ by M in (1). As M is large enough, the difference between the two definitions is negligible; moreover, if one uses the expression (2) as definition of G, most of the considerations in the rest of the paper still hold true and most of the formulae can be easily adapted.

Remark 2 *When the variable X has positive or null values, the Gini coefficient can be interpreted as a measure of variability with respect to the maximum of variability. If the variable assumes also negative values (but it still have positive mean), the Gini coefficient becomes a relative variability measure with respect to the mean of the variable.*

Now, if we split the support of X into two groups, the former containing the negative values and the latter the non-negative ones, we can write the sum S of the absolute differences as

$$\begin{aligned} S &= \sum_{i=1}^{M}\sum_{j=1}^{M} |x_i - x_j| \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} |x_i - x_j| + 2\sum_{i=1}^{N}\sum_{j=N+1}^{M} |x_i - x_j| + \sum_{i=N+1}^{M}\sum_{j=N+1}^{M} |x_i - x_j| \\ &= S_n + 2[NT_a + (M - N)T_n] + S_a, \end{aligned} \tag{3}$$

where

$$S_n = \sum_{i=1}^{N}\sum_{j=1}^{N} |x_i - x_j|$$

is the Within component for the group of negative values, while

$$S_a = \sum_{i=N+1}^{M}\sum_{j=N+1}^{M} |x_i - x_j|$$

is the Within component for the group of non-negative values, and finally

$$[NT_a + (M-N)T_n] = \sum_{i=1}^{N}\sum_{j=N+1}^{M}|x_i - x_j| = \sum_{i=N+1}^{M}\sum_{j=1}^{N}(x_i - x_j)$$

is the Between-group component (see Dagum (1997) for further details). It is worth to note that formula (3) can be seen as a special case of the decomposition originally proposed by Yntema (1933) and used by Gastwirth (1975), where only two groups are considered. By using the formula (3), the expression of the Gini coefficient becomes:

$$G = \frac{S_n + 2[NT_a + (M-N)T_n] + S_a}{2(M-1)(T_a - T_n)}. \tag{4}$$

An important reason making the Gini coefficient so famous is its easy interpretation: when dealing with non-negative values it can be seen as the ratio of the concentration area and the area corresponding to the situation with maximum inequality, it follows that its value can represent the percentage of inequality with respect to the maximum possible. However, when negative values are considered, this interpretation is no longer adequate. In Figure 1, the Lorenz curve for a variable that assumes also negative values is drawn.

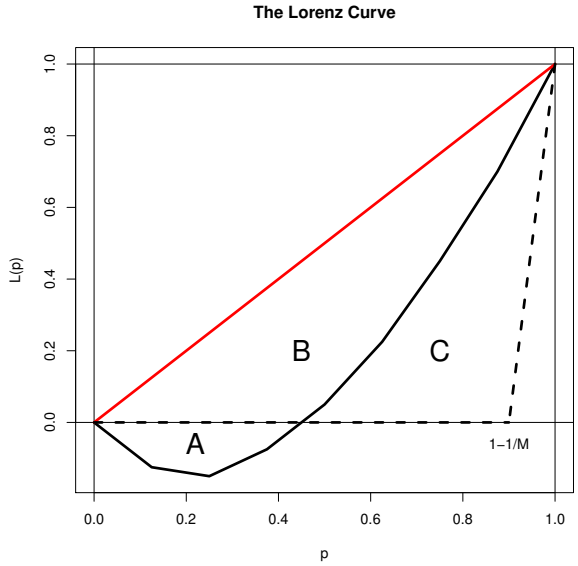

Figure 1: The Lorenz curve for a distribution with negative values.

The Gini coefficient is now the ratio between the sum of areas A and B, at the numerator, and the sum of areas B and C at the denominator, that is:

$$G = \frac{A+B}{B+C}. \tag{5}$$

It is easy to understand that the Gini coefficient assumes values greater than 1, whenever the area denoted by A is greater than the one denoted by C. However, when there are negative values of the variable at stake, even if the Gini coefficient is less than 1, it is no longer a normalized index, because area A, which is added in the numerator, does not appear in the denominator. Furthermore, in these circumstances, we can no longer talk of inequality. Indeed, as clearly underlined also in Frosini (1985), inequality is a characteristic related only to non-negative variables.

For these reasons, the case of negative values is delicate and it needs to be carefully managed and interpreted. Nevertheless, in such a situation the Gini coefficient remains a relative measure of variability with respect to the mean value of the variable: it can still be interpreted as the Gini mean difference, related to twice the mean of the distribution at stake.

3 Turning to zero all the negative values

The first method of our review is widely used in applications, and its aim is to change the data, in order to return to the "standard" situation with no negative values. It consists in turning into zero all the negative values. In this case, the Gini coefficient becomes the index G_{za}:

$$G_{za} = \frac{2NT_a + S_a}{2(M-1)T_a}. \tag{6}$$

In the formula (6), the quantity $2NT_a$ represents the differences between the first N values ($i = 1, 2, ..., N$) which are set equal to zero, and the values that maintain their original non-negative values ($i = N+1, N+2, ..., M$). By a comparison between formulae (4) and (6), it should be also noted that

$$\begin{aligned} G_{za} &= \frac{2NT_a + S_a}{2(M-1)T_a} \\ &\leq \frac{S_n + 2[NT_a + (M-N)T_n] + S_a}{2(M-1)T_a} \\ &\leq \frac{S_n + 2[NT_a + (M-N)T_n] + S_a}{2(M-1)(T_a - T_n)} \\ &\leq G, \end{aligned} \tag{7}$$

since $S_n + (M-N)T_a$ and $2(M-1)T_n$ are both non negative quantities.

Indeed, the index G_{za} coincides with the Gini coefficient that results after a redistribution by subtracting the quantity $x_i T_n/T_a$ from each positive x_i and by transferring the total amount T_n, so obtained, to the units with negative values: in this way, the final result is that all the negative x_i are set to zero and all the positive values are multiplied by the quantity $(T_a - T_n)/T_a$. Obviously this implies that formula (7) holds. A graphical proof of this result is shown in Figure 2, where the Lorenz curves both before and after the transformation of negative values into zero are drawn. Before the transformation, the numerator of the Gini coefficient is the sum of three areas: $B_1 + B_2 + A$. After the

transformation the concentration area reduces to B_1. Being the denominator the same, $B_1 + B_2 + C$, the reduction in the Gini coefficient is immediately perceived.

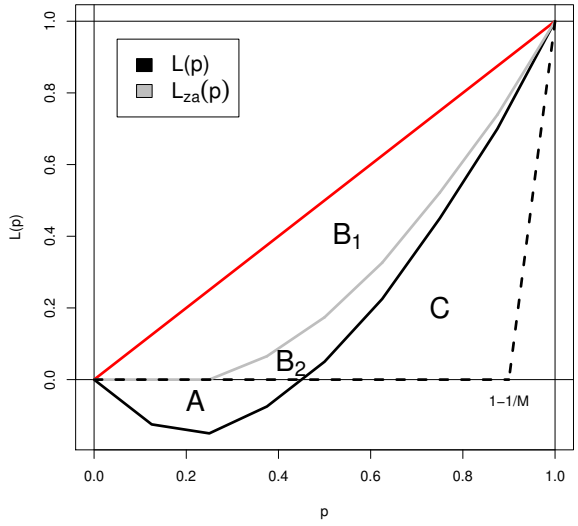

Figure 2: The Lorenz curve before -$L(p)$- and after -$L_{za}(p)$- the turning to zero of the negative values.

This method, suggested by important international organizations like OECD (see OECD (2015) for further details), has the important advantage of being very simple to apply, but it leads to a loss of information about the negative values after their modification.

4 Discarding all the negative values

Even if this second method is unsophisticated, actually it is very used and if the number of negative values of the variable X is limited, it is a forceful cheap solution. It consists basically in the discarding of the values which cause troubles, that is, the negative ones: in this way, we return to a situation with no negative values, where the "standard" Gini coefficient can be used without problems. Using our setting, if the negative values are erased, the Gini coefficient becomes:

$$G_a = \frac{S_a}{2(M - N - 1)T_a}. \tag{8}$$

Analytically it can be proved that $G_a \leq G_{za}$: in fact, let's consider G_{za}, as it is represented per expression (6), and let's re-write it as

$$G_{za} = \frac{2NT_a + S_a}{2(M - N - 1)T_a + 2NT_a},$$

having in mind G_a as per expression (8), as $S_a \leq 2(M-N-1)T_a$, we can conclude that $G_a \leq G_{za}$. From which it follows immediately, that a fortiori, $G_a \leq G$. Figure 3 shows the curve $L_{za}(p)$ and the curve $L_a(p)$ - obtained by the discarding of the negative values- for a variable with negative values. Referring to such Figure 3, it is easy to observe that

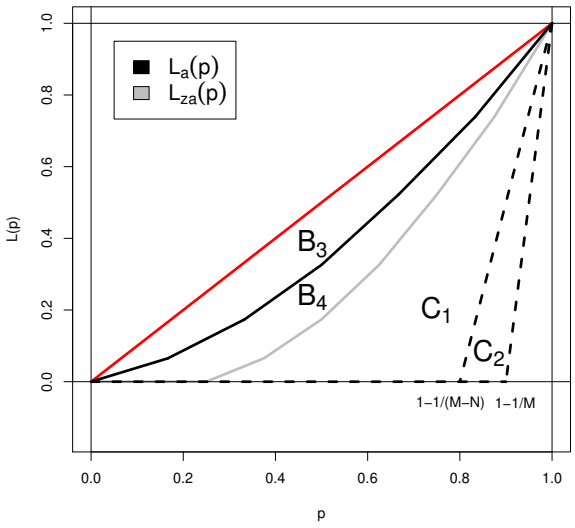

Figure 3: The Lorenz curve after the discarding of the negative values $L_a(p)$ and $L_{za}(p)$.

the concentration area associated to $L_a(p)$, B_3, is smaller than the concentration area associated to L_{za}, which is B_3+B_4. However, if we keep into account that, in calculating the coordinates of L_a, N units have been discarded, the maximum concentration area associated to G_a, $B_3+B_4+C_1$, is smaller than $B_3+B_4+C_1+C_2$, which is the maximum concentration area associated to G_{za}.

The major advantage of this method is that it is very easy to apply. If the negative values are negligible, it can result quite satisfactory. Its main drawback is instead the loss of information related to the negative values (which is greater than the one in the previous).

5 The CTR-BS correction of the Gini coefficient

Beyond the two previous methods, in the literature some corrections for the Gini coefficient have been proposed, in order to obtain an index with range $[0, 1]$. In the present section we approach the first normalization we will consider in this paper: it was initially proposed by Chen et al. (1982) and further completed by Berrebi and Silber (1985), where the authors highlighted a defect of the first proposal and suggested an adjustment to overcome it. The modification of the Gini coefficient suggested by these authors (henceforth denoted by G_{CTR-BS}), on the one hand, allows the preservation of the

whole variability in S and, on the other hand, restricts their modified Gini coefficient within the range $[0, 1]$.

This correction is obtained by taking into account the Lorenz curve below the X-axis in Figure 1: this portion of the curve involves the negative values and the smallest positive values which are enough to compensate the former, in order to ensure that their overall sum is zero. After that, the modified Gini coefficient is calculated as the ratio between the sum of areas A and B, at the numerator, and the sum of areas A, B and C at the denominator -see Figure 1-, that is to say $[A + (1 - 1/M)(1/2)]$:

$$G_{CTR-BS} = \frac{A+B}{A+B+C} = \frac{A+B}{A+\frac{1}{2}\left(1-\frac{1}{M}\right)}. \tag{9}$$

It is easy to see that G_{CTR-BS} can be also evaluated as:

$$G_{CTR-BS} = 1 - \frac{C}{A+B+C} \tag{10}$$

or alternatively as

$$G_{CTR-BS} = \frac{A+B}{B+C}\left[\frac{B+C}{A+B+C}\right] = \frac{A+B}{B+C}\cdot\left[1-\frac{A}{A+\frac{1}{2}\left(1-\frac{1}{M}\right)}\right] \tag{11}$$

where the ratio $\frac{A+B}{B+C}$ denotes the "standard" Gini coefficient.

By construction, the coefficient G_{CTR-BS} always lies between 0 and 1, since for example in (10) the ratio $\frac{C}{A+B+C}$ is clearly in the interval $[0, 1]$.

Moreover, as the authors argue (Chen et al., 1982), if there are no negative values, meaning that $A = 0$, the formula (11) clearly shows that G_{CTR-BS} coincides with the "standard" Gini coefficient.[1]

For a deeper investigation of the CTR-BS correction, the following decomposition of S is useful.

Suppose that we can identify the value $k \in \{1, \ldots, M-2\}$ such that:

$$\sum_{i=1}^{k} x_i \leq 0 \qquad \text{and} \qquad \sum_{i=1}^{k+1} x_i > 0. \tag{12}$$

In a such case, we can consider then the "fake" or "artificial" (in the sense that it is not a real possible distribution, since not all the frequencies are integer numbers) distribution of the variable X:

$Values$	x_1	$\ldots$	x_N	x_{N+1}	$\ldots$	x_k	$(x_{k+1})_1$	$(x_{k+1})_2$	$\ldots$	x_M
$Freq.$	1	$\ldots$	1	1	$\ldots$	1	η	$1-\eta$	$\ldots$	1

where:

[1] This correction differs from the original one, since it is slightly adjusted, according to the considered case of a discrete variable. In Chen et al. (1982) and Berrebi and Silber (1985), in the denominator of (9) $C + B$ is replaced by $1/2$, i.e. the asymptotic approximation of this sum.

- $x_1, \ldots, x_N$ are the negative values of X;
- $x_{N+1}, \ldots, x_M$ are the positive values of X;
- all the values of X, but x_{k+1}, have frequency equal to 1;
- the value x_{k+1} appears twice because its frequency (that is equal to 1) is split into the two "weights" η and $1-\eta$. In other words, in this distribution, the symbol $(x_{k+1})_1$ denotes the value x_{k+1} with weight η, and $(x_{k+1})_2$ denotes the value x_{k+1} with weight $1-\eta$.

The weight η is calculated by:

$$\eta = \frac{|\sum_{i=1}^{k} x_i|}{x_{k+1}} = -\frac{\sum_{i=1}^{k} x_i}{x_{k+1}} \tag{13}$$

so that:

$$\sum_{i=1}^{k} x_i + \eta x_{k+1} = 0$$

and

$$(1-\eta)x_{k+1} + \sum_{i=k+2}^{M} x_i = T_a - T_n.$$

Remark 3 *If $\sum_{i=1}^{k} x_i = 0$ it follows that $\eta = 0$. In a such case there exists a value x_k which exactly compensates the sum of the previous ones.*

If now we define:

$$S_0 = \sum_{i=1}^{k}\sum_{j=1}^{k} |x_i - x_j| + 2\sum_{i=1}^{k}(x_{k+1} - x_i)\eta$$

which is the sum of absolute differences within the subset:

$$\Omega_1 = \{x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2} \ldots, x_k, (x_{k+1})_1\}, \tag{14}$$

and

$$S_u = \sum_{i=k+2}^{M}\sum_{j=k+2}^{M} |x_i - x_j| + 2\sum_{i=k+2}^{M}(x_i - x_{k+1})(1-\eta)$$

which is the sum of absolute differences within the subset:

$$\Omega_2 = \{(x_{k+1})_2, x_{k+2}, \ldots, x_M\}, \tag{15}$$

we can decompose S as

$$S = S_0 + 2S_{0,u} + S_u, \tag{16}$$

where $S_{0,u}$ is the Between-group component, since it represents the sum of the absolute differences among the elements of the two subsets Ω_1 and Ω_2: as the elements of set Ω_1

are never greater than those in the set Ω_2, all these differences are non-negative and the modulus can be avoided:

$$S_{0,u} = \sum_{i=1}^{k}\sum_{j=k+2}^{M}(x_j - x_i) + \sum_{i=1}^{k}(x_{k+1} - x_i)(1-\eta) + \sum_{j=k+2}^{M}(x_j - x_{k+1})\eta.$$

If we rearrange, we have:

$$\begin{aligned}
S_{0,u} &= k\sum_{j=k+2}^{M}x_j - (M-k-1)\sum_{i=1}^{k}x_i + kx_{k+1}(1-\eta) - \sum_{i=1}^{k}x_i(1-\eta) + \\
&\qquad + \sum_{j=k+2}^{M}x_j\eta - (M-k-1)x_{k+1}\eta \\
&= k\left[x_{k+1}(1-\eta) + \sum_{j=k+2}^{M}x_j\right] + \left[-(M-k-1)\left(\sum_{i=1}^{k}x_i + \eta x_{k+1}\right)\right] \\
&\qquad + \left[-\sum_{i=1}^{k}x_i(1-\eta) + \sum_{j=k+2}^{M}x_j\eta\right] \\
&= k(T_a - T_n) + (M-k-1)\cdot 0 + \left[-\left(\sum_{i=1}^{k}x_i + \eta x_{k+1}\right) + \right. \\
&\qquad \left. + \eta\left(\sum_{i=1}^{k}x_i + x_{k+1} + \sum_{j=k+2}^{M}x_j\right)\right] \\
&= k(T_a - T_n) + (T_a - T_n)\eta \\
&= (k+\eta)(T_a - T_n).
\end{aligned}$$

We then obtain:

$$S = S_0 + 2(k+\eta)(T_a - T_n) + Su. \tag{17}$$

Now we prove that the area between the X-axis and the Lorenz curve for values in Ω_1, denoted by A in (9), (10), and (11), is given by:

$$A = \frac{S_0}{4M(T_a - T_n)}.$$

By using the trapeziums, it holds that

$$\begin{aligned}
A &= -\frac{1}{2M(T_a - T_n)}[x_1 + (x_1 + x_2 + x_1) + \cdots + (x_1 + x_2 + \cdots + x_k)\eta] \\
&= -\frac{1}{2M(T_a - T_n)}\sum_{i=1}^{k}[2(k-i) + 1 + \eta]x_i \\
&= \frac{-2\sum_{i=1}^{k}[2(k-i) + 1 + \eta]x_i}{4M(T_a - T_n)}
\end{aligned} \tag{18}$$

Now, recalling formula (13), we can obtain:

$$\begin{aligned} S_0 &= 2\sum_{i=1}^{k}\sum_{j=1}^{k}|x_i - x_j| + 2\sum_{i=1}^{k}(x_{k+1} - x_i)\eta \\ &= 2\left[\sum_{i=1}^{k}(2i-k-1)x_i - \sum_{i=1}^{k}x_i\eta + kx_{k+1}\eta\right] \\ &= 2\left[\sum_{i=1}^{k}(2i-k-1-\eta)x_i - k\sum_{i=1}^{k}x_i\right] \\ &= 2\sum_{i=1}^{k}(2i-2k-1-\eta)x_i \\ &= -2\sum_{i=1}^{k}[2(k-i)+1+\eta]x_i, \end{aligned}$$

that is exactly the numerator of (18). It follows therefore that

$$A = \frac{S_0}{4M(T_a - T_n)}. \tag{19}$$

Then, whenever $k \in \{1, \ldots, M-2\}$, by formulae (11), (1), (19), and (17), it follows that the modified Gini coefficient is:

$$\begin{aligned} G_{CTR-BS} &= \frac{A+B}{B+C}\cdot\left[1 - \frac{A}{A + \frac{1}{2}\left(1 - \frac{1}{M}\right)}\right] \\ &= \frac{S}{2(M-1)(T_a - T_n)}\left[1 - \frac{\frac{S_0}{4M(T_a-T_n)}}{\frac{S_0}{4M(T_a-T_n)} + \frac{1}{2}\left(1 - \frac{1}{M}\right)}\right] \\ &= \frac{S_0 + 2(k+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n)}\left[1 - \frac{S_0}{4M(T_a - T_n)}\cdot\frac{4M(T_a - T_n)}{S_0 + 2(M-1)(T_a - T_n)}\right] \\ &= \frac{S_0 + 2(k+\eta)(T_a - T_n) + S_u}{S_0 + 2(M-1)(T_a - T_n)}. \end{aligned}$$

To investigate the case $k = M-1$, we have to approach the extreme values of G_{CTR-BS}. The formula (10) shows that the G_{CTR-BS} can assume value 1 only if $C = 0$. In a such case, it is interesting consider two different situations:

- if also $A = 0$, the values of the variable at stake are non-negative, G_{CTR-BS} coincides with the standard Gini G, therefore it is equal to 1 if one value is non-zero, and all the remaining ones are zero;
- if $A \neq 0$, meaning that there are negative values, G_{CTR-BS} takes on value 1 if $k = M-1$. We can consider then two subcases:
  a) $\sum_{i=1}^{k} x_i = \sum_{i=1}^{M-1} x_i = 0$ (case $\eta = 0$);

b) $\sum_{i=1}^{k} x_i = \sum_{i=1}^{M-1} x_i < 0$, and $\sum_{i=1}^{k+1} x_i = \sum_{i=1}^{M} x_i > 0$ (case $\eta \neq 0$.)

Remark 4 *Indeed, the case b) is not considered neither in Chen et al. (1982), nor in Chen et al. (1985), since it differs from case a) only in the discrete framework, proposed in this paper. In the continuous case, it coincides with the case a).*

Let's see the case b) more in details. If the value of k is the largest admissible one, $(M-1)$, and the biggest value x_M is the first one which compensates the sum of all the negative ones, we have that

$$\sum_{i=1}^{M-1} x_i < 0 \qquad \text{and} \qquad \sum_{i=1}^{M} x_i > 0,$$

and also $S_u = 0$. In a such extreme case, the normalization area is the sum of the areas denoted by $A + B$ in the right panel of Figure 4, since the Lorenz curve lies beneath the X-axis also for some values greater than $1 - 1/M$, because it can be proved that the Lorenz curve crosses the X-axis in the point

$$P = \left(1 - \frac{(1-\eta)}{M}, 0\right),$$

where η is the same value defined in formula (13). If $\eta = 0$, the case a) holds and the corresponding Lorenz curve is drawn in the left panel of Figure 4.

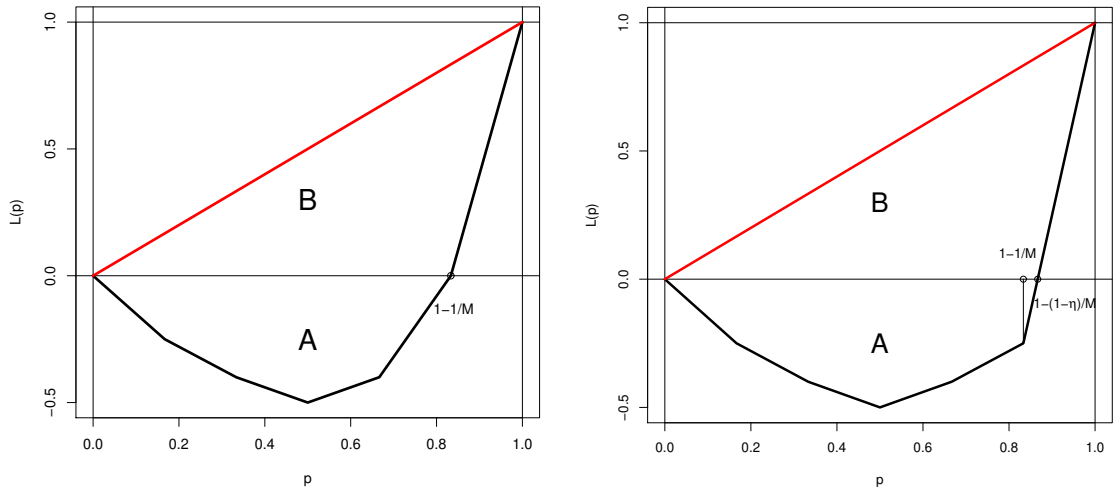

Figure 4: The Lorenz curve in the two extreme situations for the CTR-BS normalization: case a) (left panel), case b) (right panel).

Scenario	Values x_i						G	G_{CTR-BS}
(a)	-2	-2	-2	-2	-2	30	1.6	1
(b)	-10	0	0	0	0	30	2	1
(c)	-7	-3	0	0	0	30	1.94	1
(d)	-5	-3	-2	1	8	21	1.66	1
(e)	-5	-3	-2	1	9	20	1.64	1

Table 1: Five scenarios where the index G_{CTR-BS} assumes its maximum admissible value, which in the present paper is 1.

Taking in account all the above considerations, the (modification to the discrete case of) G_{CTR-BS} index is given by:

$$G_{CTR-BS} = \begin{cases} \frac{S_0+2(k+\eta)(T_a-T_n)+S_u}{S_0+2(M-1)(T_a-T_n)} & if\ k = 1, ..., M-2 \\ \\ 1 & if\ k = M-1. \end{cases}$$

Remark 5 *The two expressions of G_{CTR-BS} assume the same value in the special case with $k = M-1$ and $\eta = 0$.*

From the expression of the index G_{CTR-BS} it is easy to see that it is not suitable to distinguish among all these different extreme situations. Table 1 reports five scenarios, where the index G_{CTR-BS} takes on its maximum value, even if the scenarios are related to five very different situations: in all of them $k = M-1$, but in the first four ones $\eta \neq 0$ -case b) and right panel of Figure 4-, while in the last one $\eta = 0$ -case a) and left panel of Figure 4-.

6 The Raffinetti et al. normalization of Gini coefficient

In Raffinetti et al. (2015) a different normalization of the Gini coefficient has been proposed. By using our notation, the authors in their paper state that, given the sum of non negative values T_a, the sum of absolute negative values T_n, and the total number of units M: " ... if the attribute distribution presents negative values, it seems reasonable to identify as the maximum inequality scenario the situation where the total negative attribute amount $-T_n$ is assigned to one unit and the total positive T_a to another unit, while all the other $M-2$ units have a zero amount of attribute. As a consequence, the proposed reference distribution becomes $(-T_n, 0, \cdots, T_a)$. In such a scenario, as one can easily verify, the mean difference S yields the value:

$$S = 2(M-1)(T_a + T_n)." \tag{20}$$

In other words, the authors consider the following extreme situation where the variable X has the distribution:

$Values$	$x_1 = -T_n$	$x_2 = 0$	$x_3 = T_a$
$Freq.$	1	$M-2$	1

and they propose to modify the Gini coefficient, dividing S by the value in formula (20) assumed in the reference distribution; in a such way they define the Raffinetti et al.'s G_P as:[2]

$$G_P = \frac{S}{2(M-1)(T_a+T_n)}.$$

It is worth noting that G_P can be evaluated from the Gini coefficient by the following transformation:

$$G_P = G \cdot \left[\frac{T_a - T_n}{T_a + T_n}\right]. \tag{21}$$

Even if Raffinetti et al. (2015) do not show that the maximum for S is (20), it is possible to prove it by the following procedure. It is well-known that:

$$\begin{aligned} S &= 2\sum_{i=1}^{M}(2i-M-1)x_i \\ &= 2\sum_{i=1}^{M}(2i-1)x_i - 2M(T_a - T_n). \end{aligned} \tag{22}$$

Now, it is easy to see that the two inequalities hold (remembering that x_i is negative for $i = 1, \ldots, N$ and non-negative for $i = N+1, \ldots, M$):

$$\sum_{i=1}^{N}(2i-1)x_i \leq \sum_{i=1}^{N} x_i;$$

$$\sum_{i=N+1}^{M}(2i-1)x_i \leq \sum_{i=N+1}^{M}(2M-1)x_i.$$

The formula (22) then provides:

$$\begin{aligned} S &= 2\left[\sum_{i=1}^{N}(2i-1)x_i + \sum_{i=N+1}^{M}(2i-1)x_i\right] - 2M(T_a - T_n) \\ &\leq 2\left[\sum_{i=1}^{N} x_i + \sum_{i=N+1}^{M}(2M-1)x_i\right] - 2M(T_a - T_n) \\ &\leq 2\left[-T_n + (2M-1)T_a\right] - 2M(T_a - T_n) \\ &\leq 2T_n(M-1) + 2T_a(M-1) \\ &\leq 2(M-1)(T_a + T_n). \end{aligned} \tag{23}$$

[2]In the author's notation, the subscript P in G_P stands for "Polarization". In this framework we preferred to omit the word "Polarization", because it can create misunderstanding, since in the literature such word is used to describe a different phenomenon, largely studied.

Scenario	Values x_i						G	G_P
(a)	-2	-2	-2	-2	-2	30	1.6	0.8
(b)	-10	0	0	0	0	30	2	1
(c)	-7	-3	0	0	0	30	1.94	0.97
(d)	-5	-3	-2	1	8	21	1.66	0.83
(e)	-5	-3	-2	1	9	20	1.64	0.82

Table 2: The values of the index G_P for the five scenarios proposed in Table 1

Formula (23) implies that the maximum value of the Gini coefficient is $\frac{T_a+T_n}{T_a-T_n}$, for fixed values of T_a, T_n, and M: for this reason the index G_P can be seen as the normalization of the Gini coefficient with respect to its maximum value. Consequently G_P ranges in the interval $[0, 1]$.

To highlight the differences between G_P, G and G_{CTR-BS}, the values of G_P for the five scenarios in Table 1 are reported in Table 2.

7 Compensative redistributions

In the literature it is commonly accepted that, whenever a transfer occurs from a richer unit to a poorer one, and this transfer does not change the ranks of the units, a reasonable inequality index should decrease its value. This is in accordance to the Pigou-Dalton principle. Compensation processes are quite common in the analysis over time of economic and financial variables: it is not difficult to find examples of variables -that can assume both positive and negative values- which modify their distribution between two consecutive (or even not consecutive) observations, because of a redistribution of the total quantity.

In this section, we analyse the behaviour of the indexes presented in the previous sections, after a compensation. We do not investigate the changes of the index G_a, because it modifies the number of units and the sum of the values, since it discards the negative values of the variable at stake. For this reason the two situations -before and after the compensation- cannot be considered comparable.

Among all the possible compensations, we identify a particular one, where the negative values are replaced by zero, due to a transfer from the smallest positive values. More in detail, let's consider an egalitarian transfer, achieved at the expense of the units with smaller positive values. This compensation acts inside the subset $\Omega_1 = \{x_1, x_2, \ldots, x_N, x_{N+1}, \ldots, x_k, (x_{k+1})_1\}$ -already defined in (14)- and it modifies all the values in the subset Ω_1 into zero. It is useful to recall that the value of η is obtained by the relationship:

$$\sum_{i=1}^{k} x_i + \eta x_{k+1} = 0.$$

The subset $\Omega_2 = \{(x_{k+1})_2, x_{k+2}, \ldots, x_M\}$ -already defined in (15)- does not change.

We label a such redistribution the "minimal compensation", since the positive values involved are the smallest ones. In the pursue we shall label the before-compensation indexes by B, and the after-compensation indexes by A. The first important remark is that if we consider the Gini coefficient, it holds that

$$G^B \geq G^A.$$

After the compensation, it is not difficult to see that $G^A_{za}, G^A_{CTR-BS}, G^A_P$ coincide with G^A and they are all equal to the Gini coefficient evaluated on the (non-negative) values

$$G^A = \frac{2(k+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n)}, \tag{24}$$

where T_a, T_n, k, and η are defined as usual and refer to the situation before the compensation.

The index G_{za} is not included in this analysis, because it is not difficult to see that it does not satisfy the transfer principle. Starting from any distribution with negative values, the value of the index G_{za} is the Gini coefficient evaluated on a new distribution, where the negative values are replaced by zero and the positive ones remain the same. As mentioned in Section 3, this modification of the original distribution can be seen as the result of an egalitarian transfer. The index G_{za} applied to the new modified distribution (with positive values and zeros) coincides with the Gini coefficient and therefore it has the same values assumed from the original distribution (with also negative values). In a such case, G_{za} is not sensible to the distribution modification, and therefore it does not satisfy the Pigou-Dalton principle. By definition, the index G_P satisfies the transfer principle in the case of the minimal compensation if

$$G^A_P \leq G^B_P,$$

meaning that, by using the formula (21),

$$G^A \leq G^B \left(\frac{T_a - T_n}{T_a + T_n} \right),$$

where T_a and T_n refer to the situation before the compensation. It follows that the index G_P satisfies the Pigou-Dalton principle in the case of the minimal compensation, whenever it holds that

$$\frac{G^A}{G^B} \leq \frac{T_a - T_n}{T_a + T_n}.$$

Let's see what happens to the index G_{CTR-BS} if the minimal compensation occurs. It always holds that

$$G^B_{CTR-BS} \geq G^A,$$

meaning that, by the definitions (9) and (5) of the two indexes:

$$\frac{A+B}{A+B+C} \geq \frac{B}{B+C},$$

where A, B, and C are the areas already considered in Figure 1. This last relationship can be proved, by considering a positive real number a and the following function ϕ_a:

$$\begin{array}{rccc} \phi_a: & \mathbb{R}^+ & \to & \mathbb{R}^+ \\ & x & \mapsto & 1+\frac{a}{x}. \end{array}$$

It is easy to see that ϕ_a is a monotonically non-increasing function, i.e.

$$\text{if } x_1 \leq x_2, \text{ then } \phi_a(x_1) \geq \phi_a(x_2).$$

By choosing $a = A$, $x_1 = B$, and $x_2 = B + C$, since $B \leq B + C$ we have:

$$\begin{array}{rcl} \phi_a(B) & \geq & \phi_a(B+C) \\ 1+\dfrac{A}{B} & \geq & 1+\dfrac{A}{B+C} \\ \dfrac{A+B}{B} & \geq & \dfrac{A+B+C}{B+C} \\ \dfrac{A+B}{A+B+C} & \geq & \dfrac{B}{B+C}. \end{array}$$

It follows that in the case of the minimal compensation G_{CTR-BS} satisfies the transfer principle: obviously from this result we cannot state that this index satisfies such principle in any other situation.

8 An application to real data

In this section, we want to illustrate the methods introduced in the previous sections through an empirical analysis performed on income data collected in 1987 and in 2014 by the Survey of Household Income and Wealth (SHIW) released by Banca d'Italia (2015). We selected these two years in order to better illustrate the issue related to the presence of negative values. Even if they are more recent, the data of year 2016 have not been used in our analysis since in that year for the variable Financial Capital Gains (YCF) it happens that $T_a - T_n < 0$, and therefore the Chen et al.'s correction cannot be performed.

The SHIW survey began in the 1960s with the aim of gathering data on the incomes and on the savings of Italian households. Since its inception, the scope of the survey has grown and includes wealth and other aspects of households' economic and financial behaviour. The variable Total Income (Y) is the sum of six main income sources:

1. Earned Income, including income employment (YL);
2. Self-employment (YM);
3. Pensions (YTP);
4. Transfers (YTA) -which consist in many kinds of pensions and other government benefits;
5. Income from Real Estate Property (YCA);
6. Financial Capital Gains (YCF).

Sample Characteristics	1987	2014
M: sample size	7328	8156
N: number of households with negative value	340	700
Number of households with null value ($x_i = 0$)	1163	1329
T_a: total amount of positive values	8922383.25	3094826.61
T_n: total amount of absolute negative values	702443.14	1811122.53
Value of k satisfying the relationships in (12)	4376	7996
x_{min}: minimum value	-56376.88705	-19603.43467
x_{max}: maximum value	71506.19816	79133.89668
Median	378.479	43.013
IQR: interquartile range	1236.264	214.207
g_1: Fisher's sample coefficient of skewness	6.065	13.086
g_2: Fisher's sample coefficient of kurtosis	117.763	391.825

Table 3: The Financial Capital Gain (YCF) data for 1987 and 2014.

In this application, we take into consideration the last source: the Household Financial Capital Gain (YCF) of the two years 1987 and 2014. The surveys cover 7328 households (for 1987) and 8156 (for 2014). Table 3 summarizes the main characteristics of the two samples: in our analysis the unit coincides with the household. The monetary values of 1987 have been converted into euro. The index g_1 and g_2 are defined by:

$$\begin{aligned} g_1(X) &= \frac{M}{(M-1)(M-2)} \sum_{i=1}^{M} \left(\frac{x_i - \bar{x}}{s} \right)^3 \\ g_2(X) &= \frac{(M+1)M}{(M-1)(M-2)(M-3)} \sum_{i=1}^{M} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(M-1)^2}{(M-2)(M-3)}. \end{aligned}$$

The two years are quite different in terms of the percentage of households with negative values: 4.6% in 1987 and 8.6% in 2014. Moreover they have different ratios T_n/T_a: 7.9% in 1987 and 58.5% in 2014. The percentage of units with no financial capital gain remains almost stable in the two surveys: 15.9% and 16.3%, respectively. If now we apply the first procedure presented in the previous sections, by turning the negative values to zero, we consider the part $2NT_a + S_a$ -see formula (6)- of S; while if we apply the second method, by removing the negative values, S reduces to S_a -see formula (8)-. This means that in the former case we use 90.44% of the overall variability in 1987 and only 60.24% in 2014. In the latter case, the two percentages decrease nearly six percentage points, to 84.69% in 1987 and to 54.12% in 2014. The Table 4 summarizes some relevant statistics (like the percentages of variability taken into account, etc...) and the values of the indexes presented in the previous sections.

Statistics	1987	2014
$[2NT_a + S_a]/S$	90.44%	60.24%
S_a/S	84.69%	54.12%
G	0.8761	3.3799
Upper bound for G	1.1709	3.8217
G_a	0.7169	0.8300
G_{za}	0.7300	0.8446
G_{CTR-BS}	0.8178	0.9967
G_P	0.7483	0.8844
G^A	0.7959	0.9887
$S_0/4M(T_a - T_n)$	0.0356	1.1954

Table 4: Statistics for the Financial Capital Gain (YCF) data of 1987 and 2014.

In 1987, the Gini coefficient is equal to 0.8761; G_a is 16 percentage points lower (0.7169) and G_{za} is more than 14 percentage points lower than G. The differences between such three indexes are much more significant in 2014: in this year the Gini coefficient is 3.3799, G_a is 0.83, and G_{za} is equal to 0.8446. In the examples considered here, the most significant difference is definitely between G and G_{za}, rather than between G_{za} and G_a.

It is interesting to observe what would happen after a redistribution which compensates the negative values by an egalitarian redistribution from the lowest positive values, that is after a minimal compensation. G^A would become 0.7959 in 1987 and 0.9887 in 2014: in 1987 nearly 5 percent points greater than G_{za} and more than 8 percent points greater than G_a. In 2014 the differences would be much higher: more than 14 percent points for G_{za} and nearly 16 percent points for G_a. The G_P indexes are

$$G_P^{1987} = 0.7483 \qquad \text{and} \qquad G_P^{2014} = 0.8844.$$

As expected G_P is lower in 1987 and 2014 than the corresponding G^A after the minimal compensation. However we think that the proper use of G_P is to integrate G, and not to replace it. The standard Gini coefficient can not longer be considered a concentration index, both in 1987 and 2014: for the presence of negative values, it is just a relative measure of variability with respect to the mean value. It informs us that, if we measure the variability by the ratio between the Gini mean difference and twice the mean of the financial capital gain, the variability of the income source at stake has increased 3.86 times from 1987 to 2014.

Moreover, as the ratio $(T_a + T_n)/(T_a - T_n)$ is 1.1709 in 1987 and 3.8217 in 2014, G_P informs that the Gini coefficient (which assesses the relative variability) reached nearly

the 75% of its potential maximum in 1987, and more than the 88% of its potential maximum in 2014.

The G_{CTR-BS} indexes are

$$G^{1987}_{CTR-BS} = 0.8178 \qquad \text{and} \qquad G^{2014}_{CTR-BS} = 0.9967.$$

As expected, the G_{CTR-BS} index is greater than G^A: however if we consider the difference in variability expressed by the difference between G and G^A, which have the same denominator, the G_{CTR-BS} index reveals very little of the variability reduction due to the minimal compensation. In fact, in 1987 the difference between G and G^A is $0.8761 - 0.7959 = 0,0802$, whilst the difference between G_{CTR-BS} and G^A is $0.8178 - 0.7959 = 0,0219$. In 2014 the undervaluation is even more evident, as the difference between G and G^A is 3.3799-0.9887= 2,3912, whilst the difference between G_{CTR-BS} and G^A is just $0.9967 - 0.9887 = 0,0080$. Moreover, as the area below the X axis, $\frac{S_0}{4M(T_a - T_n)}$, in 2014 and 1987 is different, being 0.0356 in 1987 and 1.1954 in 2014, the G_{CTR-BS} index in 1987 is not immediately comparable with the G_{CTR-BS} index in 2014.

We would conclude that, when dealing with negative values, by the use of G and G_P together one can better evaluate the differences among distributions, as well as the effects of redistributions pertaining to the same population.

9 Final remarks

The purpose of this research was to indicate a valid operating procedure to manage the issue of the inequality when a distribution includes negative values. Generally, in overall income distributions only a few units present negative values. However, when we disaggregate overall income distributions into their sources, units presenting negative values can no longer be considered a negligible phenomenon. Another situation where several units with negative values can be observed is given by tax systems, which introduce family allowances through the form of negative income taxes. In this article we have shown that when a distribution includes negative values, neither dropping units with negative values nor transforming these values to zero are suitable practices. This should not be done if we do not want both to exclude a part of the variability that can be considerable and to bias comparisons among distributions, related either to different populations or to the same population in different periods. Even if the Chen et al. (1982) coefficient appears a feasible procedure that preserves the whole variability, it presents some limits due to unreasonable behaviours in some circumstances, as stressed by Raffinetti et al. (2015) and remarked in Section 5. This ad hoc procedure is not recommended to compare different situations.

From the results of our research, we suggest some general guidelines to deal with negative values, based on empirical experience. When the negative values represent less than the 1% of the total observations of the dataset, we suggest to use G_a, since the loss of information due to the discarding of negative values is balanced out by the ease of calculation. If the proportion of negative values is higher (between 1% and 5%), then

we recommend the use of G_{za}, as OECD suggests, since its bias with respect to G is smaller than the corresponding one of G_a. In all the other cases we propose the adoption of G_P, obtained by dividing the Gini coefficient by its upper bound. This normalized index measures the percentage of the potential maximum variability, in a given situation; it could compare any new distribution, obtained by a redistribution from the previous one which keeps constant the ratio between the overall negative amount and the overall positive amount. Moreover, G_P can be used to compare the normalized variability in the cases where the mentioned ratio is the same.

Finally, it is important to remark that when dealing with negative values the standard Gini coefficient G is no longer a concentration measure: it can still be computed for comparing different distributions, but it can be interpreted just as a relative measure of variability with respect to the mean value.

Acknowledgement

The authors would like to thank two anonymous reviewers, whose comments contributed to improve the paper, and also Professor V.B. Frosini, for his helpful suggestions. Usual disclaimers apply.

References

Amiel Y., Cowell F.A., and Polovin A. (1996). Inequality among the kibbutzim. *Economica*, 63.

Banca d'Italia (2015). Survey on Household Income and Wealth in 2014. *Supplements to the Statistical Bulletin - Sample Surveys, XXV (64). Available at http://www.bancaditalia.it.*

Berrebi, Z.M., and Silber, J. (1985). The Gini coefficient and negative income: a comment. *Oxford Economic Papers*, 37(3).

Budd, E.C. (1970). Postwar changes in the size distribution of income in the United States. *American Economic Review*, 60.

Castellano, V. (1937). Sugli indici relativi di variabilità e sulla concentrazione dei redditi con segno. *Metron*, 13.

Chen, C., Tsaur, T., and Rhai, T. (1982). The Gini coefficient and negative income. *Oxford Economic Papers*, 34(3).

Chen, C., Tsaur, T., and Rhai, T. (1985). The Gini coefficient and negative income: replay. *Oxford Economic Papers*, 37(3).

Cowell, F.A. and Van Kerm, P. (2015). Wealth inequality: a survey. *Journal of Economic Surveys*, 29(4).

Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio. *Empirical Economic*, 22.

Frosini, B.V. (1984). Concentration, dispersion and spread: an insight into their relationship. *Statistica*, 44.

Frosini, B.V. (1985). Types and properties of inequality measures. *Working Paper of Università Cattolica del Sacro Cuore. Istituto di Statistica.*

Gastwirth, J.L. (1975). The estimation of a family of measures of economic inequality. *Journal of Econometrics*, 3(1).

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. Anno Accademico 1913-1914, Tomo LXXIII-Parte Seconda.*

Gini, C. (1930). Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione. *Metron*, 8.

Jenkins, S.P. and Jantti, M. (2005). Methods for summarizing and comparing wealth distributions. *ISER Working Paper 2005-05. Colchester: University of Essex, Institute for Social and Economic Research (2005).*

Hagerbaumer, J. B. (1937). The Gini concentration ratio and the minor concentration ratio: a two parameter index of inequality. *Review of Economics and Statistics*, 59.

Lambert, P. J. and Yitzhaki, S. (2013). The inconsistency between measurement and policy instruments in family income taxation. *FinanzArchiv: Public Finance Analysis*, 69.

OECD (2015). Terms of Reference, OECD Project on the distribution of household incomes *2015/16 Collection, Available at https://www.oecd.org/els/soc/IDD-ToR.pdf.*

Pyatt, G., Chen, C. and Fei, J. (1980). The distribution of income by factor components. *The Quaterly Journal of Economics*, 94.

Raffinetti, E., Siletti, E. and Vernizzi, A. (2015). On the Gini coefficient normalization when attributes with negative values are considered. *Statistical Methods and Applications*, 24.

van de Ven, J. (2001). Distributional limits and the Gini coefficient. *Research Paper 776, Department of economics, University of Melbourne, Melbourne, Australia.*

Yntema, D. (1933). Measures of the inequality in the personal distribution of wealth or income. *Journal of American Statistical Association*, 28.