**M A TH E M ATI C AL E CO NO MI C S**

No. 14(21) 2018

# DECOMPOSITION OF THE GINI INDEX IN THE PRESENCE OF OBSERVATIONS WITH NEGATIVE VALUES

**Katarzyna Ostasiewicz, Achille Vernizzi**

**Abstract.** The problem of calculating the Gini index in the presence of negative inputs may be overcome by decomposing the quantity at stake into "positive" and "negative" parts, which, by definition, include only non-negative values. For such constructed sources, concentration indexes may be calculated and their influence on the overall inequality evaluated. In the paper we present the methodology of such an assessment and illustrate it with the example of Italian income data.

**Keywords:** inequality, the Gini index, negative inputs, decomposition.

**JEL Classification:** C18, D63

**DOI:** 10.15611/me.2018.14.04

## 1. Introduction

One of the aims of decomposing the overall inequality into different sources is to estimate the impact of the given source on the total inequality. The issue is very important in view of implementing the proper policy with desirable consequences. On the other hand, in the presence of negative observations, the Gini index loses its property of being scaled within the interval <0,1>, thus becoming more difficult to interpret. A common practice in such situations is either to neglect the negative observations or to turn them into zeros. Both these methods obviously lower inequality, however in most cases, if the share of negative observations is low, the error introduced is not high. On the other hand, if one wants to deal with the decomposition, there might appear some sources with many – or even most of – negative observations (e.g. taxes). In such a case the above-mentioned treatments, neglecting or turning into zeros, would on no account be acceptable. In such a situation, by

**Katarzyna Ostasiewicz**
Wrocław University of Economics
e-mail: katarzyna.ostasiewicz@ue.wroc.pl
ORCID: 0000-0002-0115-3696

**Achille Vernizzi**
University of Milan
e-mail: achille.vernizz@unimi.it
ORCID: 0000-0002-1641-5003

decomposing the overall quantity into "positive" and "negative" components, each consisting of non-negative values, one neither introduces non-acceptable errors nor loses the useful property of the Gini index of being scaled.

In this paper we present the methodology of decomposition of the Gini index proposed by Podder [1993]. Then we define the decomposition of sources into "positive" and "negative" parts and show how to interpret their impact on the overall inequality based both on Podder's formulae and the Lorenz curve. Finally, the methodology is illustrated with data from the Bank of Italy.

## 2. The Gini index

In general, the Gini index is not easily decomposable, moreover the rankings of incomes from particular sources are generally different. The most widely known decomposition of the Gini index, into within-group, between-group and residual part, was disseminated by C. Dagum (see e.g. [Dagum 1998]). Podder [1993] proposed a slightly different decomposition which allows to evaluate the impacts of each source on the overall inequality. A short derivation of the Podder decomposition is given as follows.

The Gini index may be expressed in terms of covariation between the variable $x$ on whose inequality is measured and the vector of normalized ranks, that is, ranks divided by $N$, such as to obtain ranks between $\frac{1}{N}$ and 1: $\frac{r}{N} = \left(\frac{1}{N}, \frac{2}{N}, \dots, 1\right)$:

$$G = \frac{2}{\mu}\mathrm{Cov}\left(x, \frac{r}{N}\right), \tag{1}$$

where $x = (x_1, \dots, x_N)$, $x_i \leq x_{i+1}$, and $\mu$ denotes the average value $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.

If each $x_i$ is the sum of $k$ sources or components, then, in order to get $x_i = \sum_{s=1}^{k} x_{is}$, (and $x = \sum_{s=1}^{k} x_s$), the $x_s$ parades have to be aligned according to the overall ranking of *x*: let us denote the sources aligned according to the *x* ranking as $x_s|x$ (*s*=1, 2, ….*k*).

Let us now define

$$C(x_s|x) \equiv \frac{2}{\mu_s}\mathrm{Cov}\left(x_s|x, \frac{r}{N}\right), \tag{2}$$

where $\mu_s$ is the average of the *s*th source ($s = 1,2, \dots, k$). (2) is the expression of the pseudo-Gini index (or of the concentration index), when the ordering

is given by *x*. Note that $C(x_s|x) \equiv G(x_s)$ only if the ranking of $x_s$ is the same as that of $x_s|x$.

Expression (1) can be re-written as the sum:

$$G = \sum_{s=1}^{k} \frac{\mu_s}{\mu} C(x_s|x). \tag{3}$$

Analogously to the Lorenz curve, a concentration curve can be drawn for each source: on the *x*-axis the percentile of the population up to the $r$th observation as usually reported, whilst the *y*-axis reports $\frac{1}{N\mu_s}\sum_{i=1}^{r} x_{si}|x_i$.

Note that the concentration curve does not need to be non-decreasing, like the Lorenz curve. The latter is a special case of the concentration curve – in the case of having $x_s$ ordered according to $r$.

Podder's idea of estimating the impact of the *s*-th source on the total inequality is based upon observing that:

$$G = \sum_{s=1}^{k} \frac{\mu_s}{\mu} C(x_s|x), \tag{4}$$

and using:

$$G = \sum_{s=1}^{k} \frac{\mu_s}{\mu} G, \tag{5}$$

one may write:

$$0 = \sum_{s=1}^{k} \frac{\mu_s}{\mu} [C(x_s|x) - G]. \tag{6}$$

As Podder explains, the $x_s$'s for which $C(x_s|x) > G$ are considered to increase the overall inequality, while the ones for which $C(x_s|x) < G$ are considered to decrease the overall inequality.

An important property of such an interpretation is that all increments and all decrements add up to zero. On the contrary, when comparing $G$ and $C_s$ to decide whether source $x_s$ decreases or increases the total inequality, it may appear that all sources increase inequality and none decreases – which seems unintuitive. For example, let us consider the following distribution: (0,2,4). If the two underlying sources are (0,2,0) and (0,0,4), the Gini indexes for both sources ($G_1 = G_2 = 0.67$) are greater than the Gini index for the total distribution ($G = 0.44$): thus both sources would be interpreted to increase the total inequality. However, according to Podder's analysis, the first of the sources decreases ($C_1 = 0$) while the second one increases ($C_2 = 0.67$) the total inequality, the two effects exactly compensating each other.

Podder [1993] lists other possible ways of estimating the influence of a given source on the total inequality, also suffering the weakness of all

influences not being added to zero (e.g. all increasing inequality or all decreasing). On the other hand, Podder's method also may produce some counterintuitive results. Let us say, we have the distribution: (0,0,0,1,1,1), for which $G = 0.5$, and let us assume that there are three different sources generating this distributions: (0,0,0,0,0,1); (0,0,0,0,1,0) and (0,0,0,1,0,0). For each source $G_s = 0.833$. Comparing $C_s$ ($s = 1,2,3$) and $G$ yields to the unreasonable conclusion that the total inequality is increased by the first source ($C_1 = 0.833$), remains unaltered by the second ($C_2 = 0.5$) and is decreased by the third one ($C_3 = 0.167$).Yet the role of all three sources is perfectly exchangeable, so this difference in influences has no justification (violating, in some sense, the anonymity rule).

However, the issues with Podder's interpretation emerge virtually only in some very sophisticated and artificial situations, and the method has been adopted by many authors and is also chosen in this paper.

It is employed here within the specific context of sources with negative inputs. In general, with such sources the issue arises that in the presence of negative inputs, the Gini index is no longer restricted to the interval $\langle 0,1\rangle$).Some methods of overcoming this difficulty have been proposed (see e.g. [Berrebi, Silber 1985; Raffinetti et al. 2015; Ostasiewicz, Vernizzi 2017]). Podder's methodology allows in a natural way to divide a source with both positive and negative inputs, and to deal with them as the composition of two different sources – separately estimating their contributions to the total inequality.

Let us have $k$ different sources, $x_s$, $s = 1, \dots, k$, $\sum_{i=1}^{k} x_s = x$, each of which may have (in general) in addition to positive and zero values, also some negative inputs. Then, let us split each of these sources into two series: the former including only the non-negative values, and the latter including only the original non-positive values, having turned the negative signs into positive:

$$x_s = x_{ps} - x_{ns}, \text{ for } s = 1, \dots, k. \tag{7}$$

Formula (6) may be adopted thus as:

$$0 = \sum_{s=1}^{k} \frac{\mu_{ps}}{\mu} \left[C\left(x_{ps}|x\right) - G\right] - \sum_{s=1}^{k} \frac{\mu_{ns}}{\mu} \left[C\left(x_{ns}|x\right) - G\right]. \tag{8}$$

Moreover, the geometrical interpretation of the Gini index as (twice) the area between the Lorenz curve and the line of equal distribution can be extended, accordingly, to the concentration index being equal to twice the area between the concentration curve and the line of equal distribution. Therefore,

the negative or positive inputs to the total inequality from different sources may be represented in a plot, as the differences between the Lorenz curve for the total quantity and the given concentration curves. Let us consider, for instance, the total quantity $x = (x_1, x_2, \dots, x_6) = (5,5,5,5,5,20)$ and one of the sources $x_s = (x_{s1}, x_{s2}, \dots, x_{s6}) = (0,1,5,0,9,5.4)$. As $G = 0.278$ and $C(x_s|x) = 0.278$, it turns out that this source has neither a negative nor positive impact on the total inequality. However, examining the Lorenz and the concentration curve presented in Figure 1, reveals a more detailed picture.

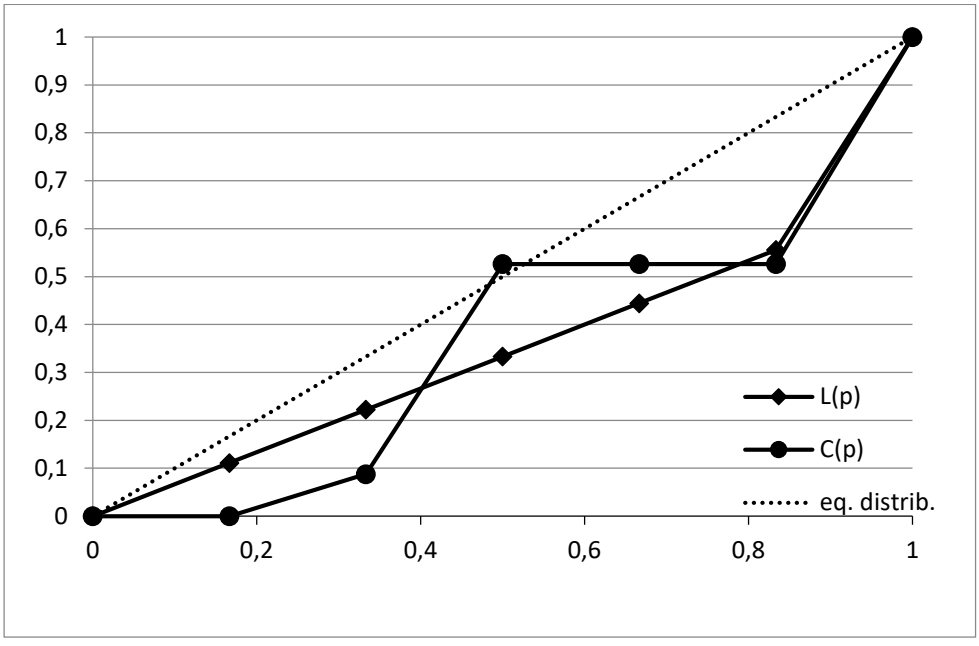


Fig. 1. Lorenz and concentration curves for
$x = (5,5,5,5,5,20)$ and for $x_s = (0,1,5,0,9,5.4)$

Source: own elaboration.

It may be noticed that although the overall impact of this source on the total inequality is null, it increases inequality in the lower part of the distribution and decreases it in the upper part – these two effects perfectly canceling each other. However, in many cases, e.g. considering taxes, we are particularly interested in the influence of some policies on some specific parts of the distribution. Thus, plotting the Lorenz curve versus the concentration curve of particular sources seems to be potentially of great importance.

## 3. The data

We now consider the data set collected by the Bank of Italy's 2016 Survey of Household Income and Wealth (SHIW). The SHIW began in the 1960s with the aim of gathering data on the incomes and savings of Italian households. The 2016 survey covered 7,421 households scattered over approximately 300 Italian municipalities. In particular, we consider households' incomes: the overall household income is split into six sources which are described in Table 1.

We have applied the so-called Carbonaro equivalence scale [Carbonaro 1985], which in its simplified version is $n_{eq} = n^{0.669}$, with $n$ being the components of each household. Table 1 reports the main summary statistics.

Table 1. Summary statistics of data

| Name | Description | Average per equivalent unit | Number of records with negative values |
|---|---|---|---|
| Y | Total | 18 465 | 10 |
| YL | dependent workers' wages | 6 230.67 | 0 |
| YTP | pensions | 6 578.21 | 0 |
| YTA | money transfers | 9.90 | 778 |
| YM | net self-employment income | 1 767.06 | 17 |
| YCA | capital gains (rentals etc.) | 3 883.89 | 0 |
| YCF | financial capital gains | -4.77 | 1 482 |

Source: own calculation based on [Banca d'Italia].

As we can see in Table 1, some sources reveal no negative values, others only a few, whilst financial capital gains and money transfers have many negative values. One of the sources, YCF, even has a negative average value. Thus omitting negative values while decomposing the overall quantity into these sources would not be acceptable.

The sources having negative and positive values are split into two parts, $x_s = x_{ps} - x_{ns}$. The positive part is labelled "$p$" in addition to the label of the source $s$ and consists of the following sequence of values: if the source assumes a non-negative value, it is just this value itself. If the source assumes a negative value, it is replaced by zero. The second part is called a "negative" part (and is labelled by an "$n$") and consists of the absolute values of negative inputs, while the positive values are replaced by zeros.

## 4. Results

Table 2 summarizes the results for the total inequality, decomposed into inputs from different sources.

It can be observed that the greatest negative impact comes from YL, while the greatest positive is due to YM(p).

Table 2. Shares of particular sources into total inequality as measured by the Gini index

| Name | Weight, $w_X$ | $C_{X/Y} - G_Y$ | $w_X(C_{X/Y} - G_Y)$ | Share (including sign) |
|---|---|---|---|---|
| YL | 0.33743 | -0.06190 | -0.02089 | -0.02089 |
| YTP | 0.35625 | -0.006419 | -0.00228 | -0.00228 |
| YTA(p) | 0.00875 | -0.43958 | -0.00385 | -0.00385 |
| YTA(n) | 0.00821 | -0.06702 | -0.00055 | 0.00055 |
| YM(p) | 0.09577 | 0.19609 | 0.01878 | 0.01878 |
| YM(n) | $7.248*10^{-5}$ | -0.11024 | $-7.99*10^{-6}$ | $7.99*10^{-6}$ |
| YCA | 0.21034 | 0.02759 | 0.00580 | 0.00580 |
| YCF(p) | 0.00513 | 0.31504 | 0.00162 | 0.00162 |
| YCF(n) | 0.00539 | -0.04795 | -0.00026 | 0.00026 |
| | | | Sum | 0 |

Source: own calculations based on [Banca d'Italia].

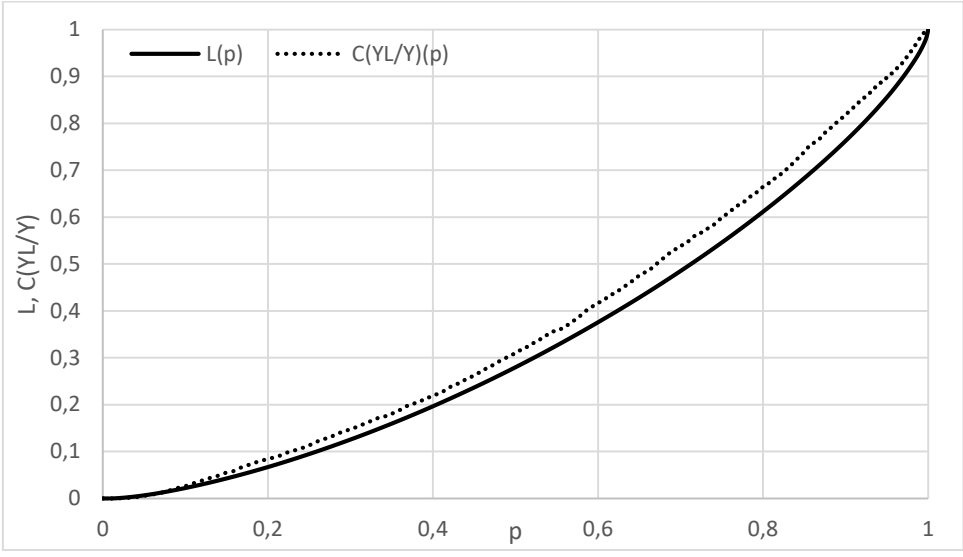


Fig. 2. Cumulative curve for YL (dependent workers' wages) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

Let us investigate the impact of each particular source in detail. Figure 2 and Figure 3b present $G_Y$ and $C_{X/Y}$. Note that for the Lorenz/concentration curves, the higher the curve the lower the inequality. However, if there is a negative part at stake where the absolute are the values taken into account, the lower concentration curve corresponds to lowering inequality. The doubled area between the Lorenz and the concentration curve is equal to the quantity in the last column of Table 2. Notice that for the Lorenz curve placed above the concentration curve, an area enters with the negative sign, and that for the "negative" parts of sources the total value has to change the sign.

It seems that dependent workers' wages contribute to lower concentration except at the very initial part of the distribution (of the overall incomes). The contrast to overall concentration is most remarkable at the highest percentiles, which means that at the final part of the distribution there is a strong counter-graduation of wages in respect of the overall income.

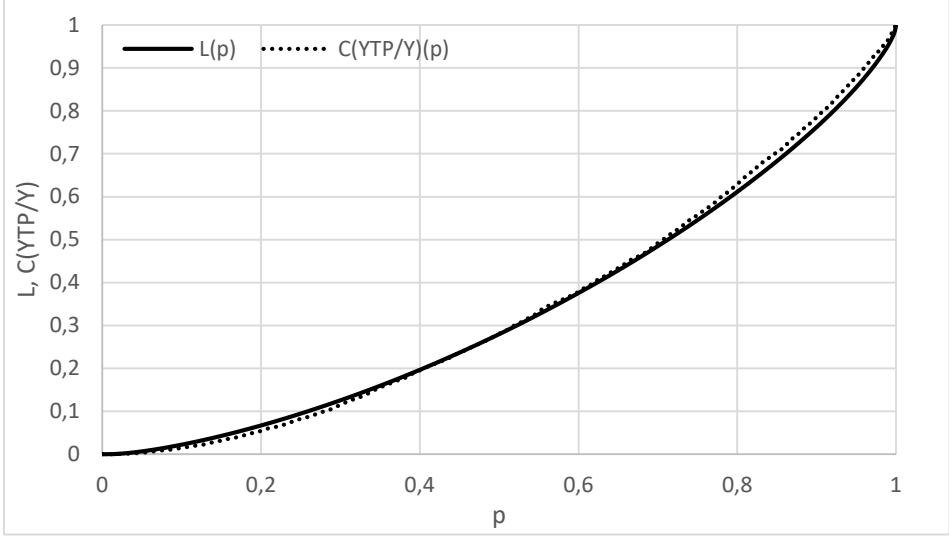


Fig. 3. Cumulative curve for YTP (pensions)
together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

Pensions contribute positively to inequality up to the median (in the overall income ranking). After the median, pensions reflect inequality. Analogously to wages, at the end of the pensions distribution they strongly reflect inequality, even if they do not show the strong counter-graduation which we noticed for wages.

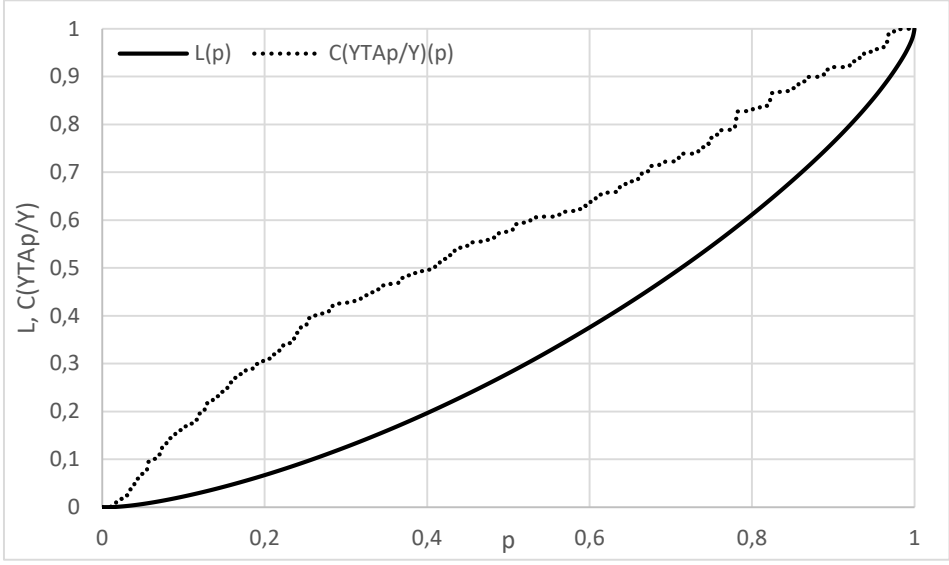


Fig. 4. Cumulative curve for YTA(p) (positive part of money transfers) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

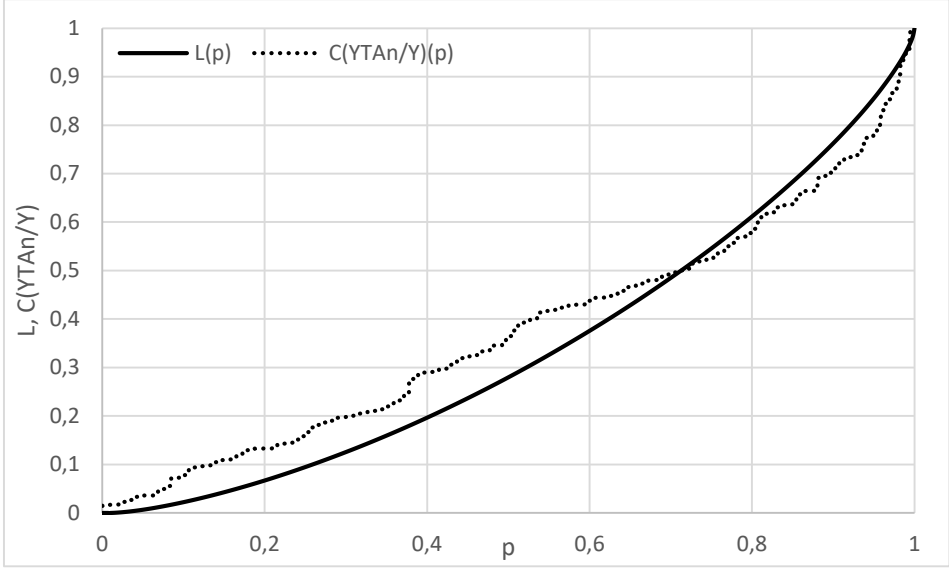


Fig. 5. Cumulative curve for YTA(n) (negative part of money transfers) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

The positive and negative transfers tell two different stories: they are plotted in Figure 4 and in Figure 5, respectively. Positive transfers (i.e. money received due to transfers) clearly contrast the overall inequality especially in the lower cumulated relative frequencies (the maximum distance between the two curves is around $p = 0.38$ (or between the twentieth and the fortieth percentile). On the other hand, negative transfers (bearing in mind that now a concentration curve closer to the line of equal distribution means increasing inequality), contribute to inequality nearly up to $p = 0.70$. This means that, although money transfers produce an overall decrease in total inequality ($-0.00385+0.00055<0$), they may be viewed as not being completely fair as the pattern of taking money favours the richest at the cost of those below the 7$^{th}$ decile.

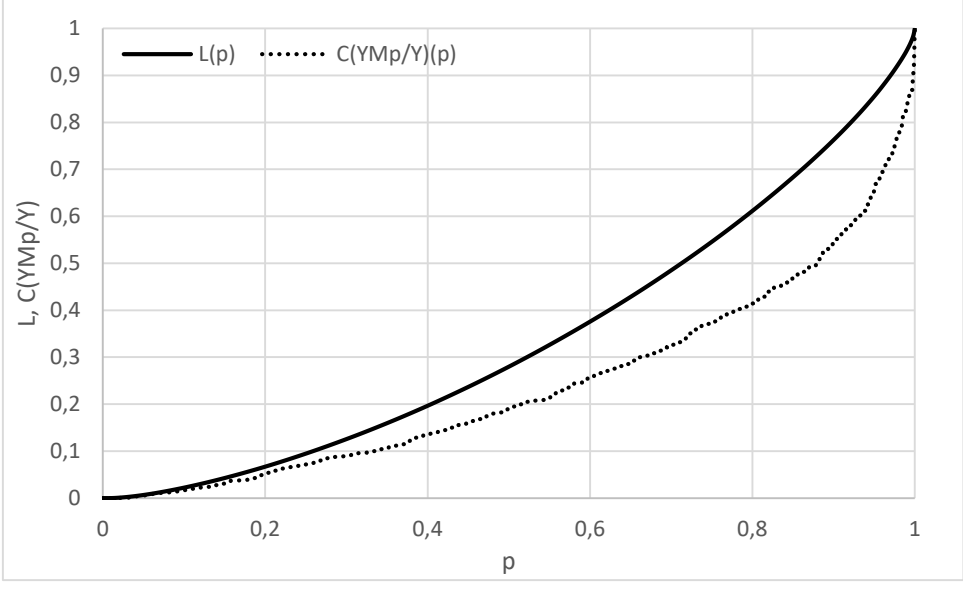


Fig. 6. Cumulative curve for YM(p) (positive part of net self-employment income) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

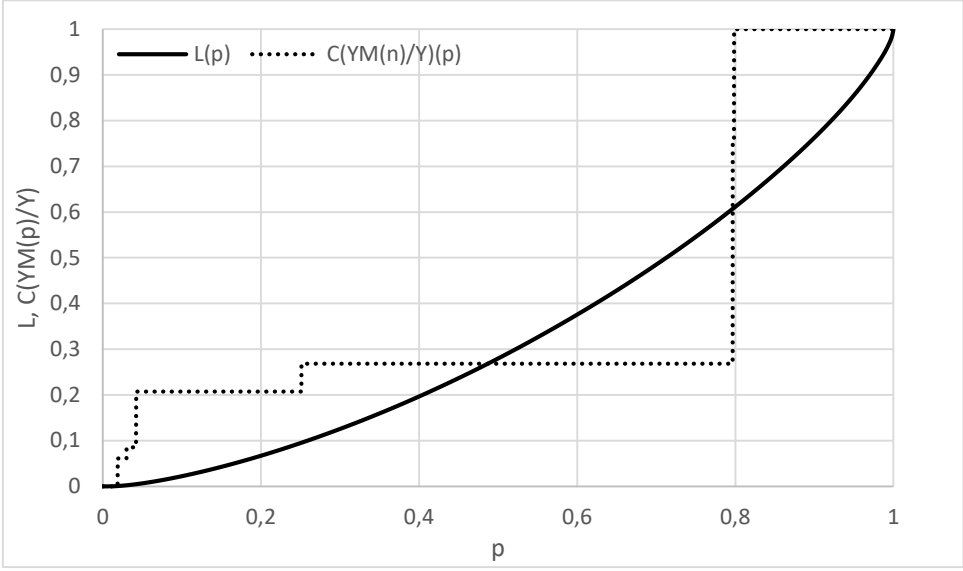


Fig. 7. Cumulative curve for YM(n) (negative part of net self-employment income) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

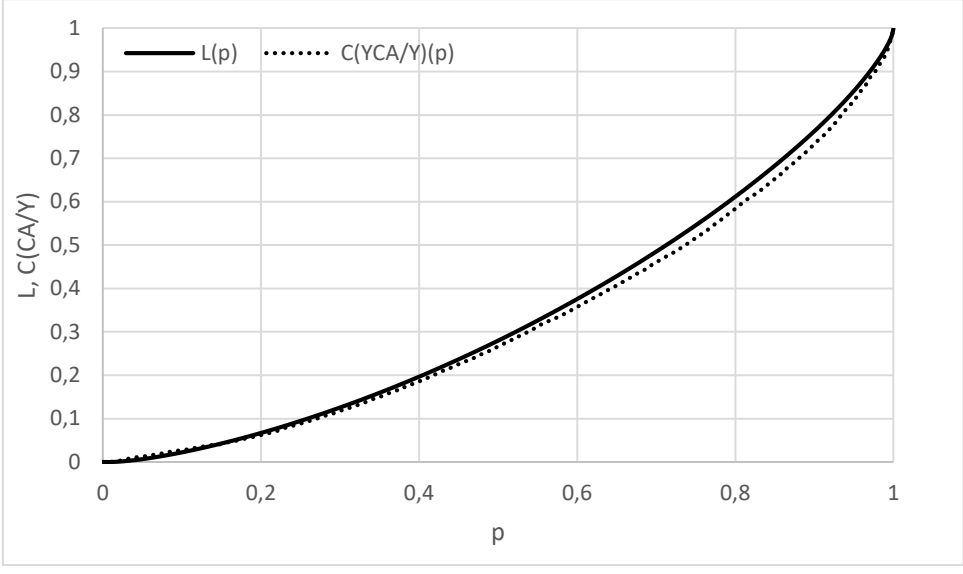


Fig. 8. Cumulative curve for YCA (capital gains) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

The non-positive values, although presented in Figure 5, are irrelevant here (only 17 inputs) and not worth considering. Where it concerns the non-negative values, they surely contribute to inequality: Figure 5 shows a strong increase in inequality, especially due to income units in the upper part of the distribution.

When considering capital gains (mainly rentals and virtual property house rentals) in Figure 8, they slightly decrease inequality for the poorest 15%. This is due to the fact that there are persons who just pay a rent or a virtual rent (depending on house ownership of where they live), often sharing it with other persons (mainly the spouse). For the higher part of the distribution of total income, capital gains have an apparent positive effect on inequality, increasing it.

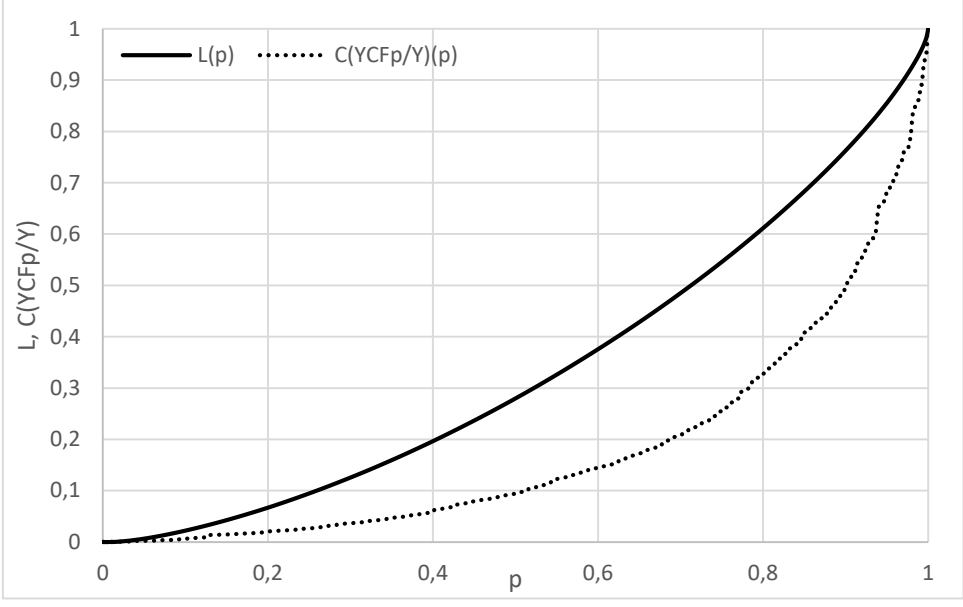


Fig. 9. Cumulative curve for YCF(p) (positive part of financial capital gains) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

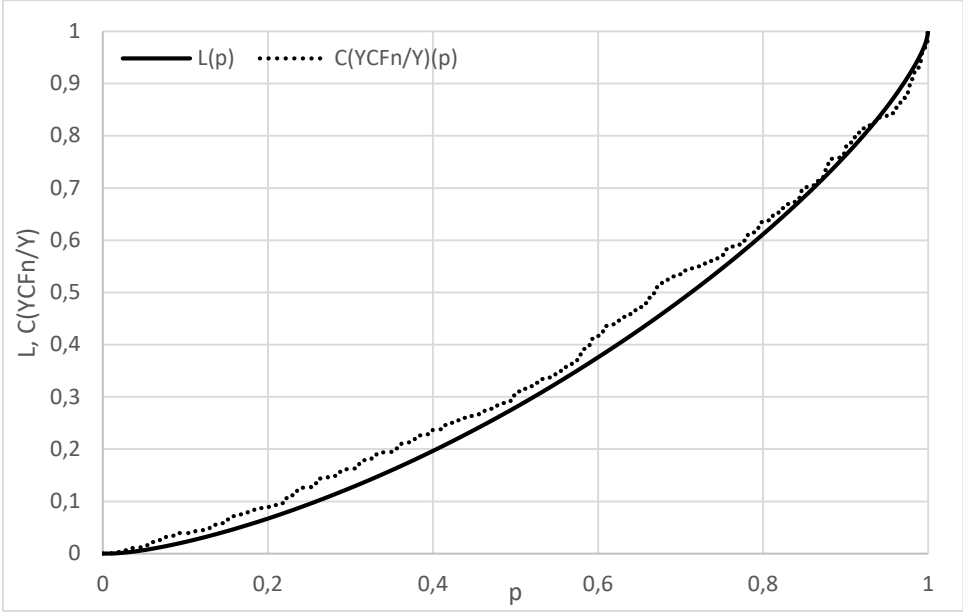


Fig. 10. Cumulative curve for YCF(n) (negative part of financial capital gains) together with the Lorenz curve for the total income

Source: own calculations based on [Banca d'Italia].

While capital gains coming from rentals have only positive values, financial gains have many negative inputs. Not surprisingly, positive gains increase inequality in the whole range of the distribution, as we can see in Figure 9. A more interesting picture comes from the results on negative gains. As may be seen in Figure 10, for almost all the range of total income the concentration curve lies above the Lorenz curve. Dealing with negative values, this means an increase of inequality for almost the whole range, apart from the top 5% of the distribution. The interpretation of the increasing gap between the Lorenz and the concentration curve, in Figure 9, might be explained by the observation that the richer the persons, the more skilled they are in investing their money thus obtaining greater profits. On the other hand, it is not necessary to suppose that poorer persons are worse investors, to explain the increase of inequality due to financial losses. If financial losses were distributed uniformly, they would increase inequality (the Gini index increases if we subtract the same positive value from each value of the distribution, that is, it decreases under a uniform translation to the lower values). The final crossing of the curves, in Figure 10, could be explained by the fact that persons with

overall high incomes could risk and lose more financial resources. However, the decrease of inequality due to the losses of the richest cannot overcome the total increase of inequality caused by financial losses in the whole population.

## 5. Conclusion

Calculating the Gini index for distributions with negative inputs, is somewhat troublesome, as the index is not more scaled within the range <0;1> and thus difficult to interpret. However, when decomposing overall incomes into sources, the problem of negative values can be overcome by splitting series having negative values into "positive" and "negative" parts. In doing so, we extend Podder's approach which originally considers just non-negative series, and we can yield further interesting pieces of information which concerns the contribution of each source to the overall inequality.

## Bibliography

Banca d'Italia. https://www.bancaditalia.it/statistiche/basi-dati/bds/index.html.

Berrebi Z.M., Silber J. (1985). *The Gini index and negative income: A comment*. Oxford Economic Papers. No. 37, pp. 525-526.

Carbonaro G. (1985). *Nota sulle scale di equivalenza*. [In:] *Presidenza del consigliodeiministry*. La povertà in Italia – Rapporto conclusivo della Commissione di studio istituita presso la Presidenza del Consiglio dei Ministri. Roma. Istituto Poligrafico e Zecca dello Stato, pp. 153-159.

Dagum C. (1998). *A New Approach to the Decomposition of the Gini Income Inequality Ratio*. [In:] *Income Inequality, Poverty, and Economic Welfare*. Physica-Verlag HD, pp. 47-63.

Ostasiewicz K., Vernizzi A. (2017). *Decomposition and normalization of absolute differences, when positive and negative values are considered: Applications to the Gini index*. Quantitative Methods in Economics. No. 18(3), pp. 472-491.

Podder N. (1993). *A new decomposition of the Gini index among groups and its interpretations with applications to Australia*. Sankhyā: The Indian Journal of Statistics, Series B, pp. 262-271.

Raffinetti E., Siletti E., Vernizzi A. (2015). *On the Gini index Normalization when Attributes with Negative Values Are Considered*. Statistical Methods & Applications. No. 24, pp. 507-521.