Attili, Federico

Working Paper

Within-between decomposition of the Gini index: A novel proposal

Quaderni - Working Paper DSE, No. 1153

Provided in Cooperation with:
University of Bologna, Department of Economics

Suggested Citation: Attili, Federico (2020) : Within-between decomposition of the Gini index: A novel proposal, Quaderni - Working Paper DSE, No. 1153, Alma Mater Studiorum - Università di Bologna, Dipartimento di Scienze Economiche (DSE), Bologna, https://doi.org/10.6092/unibo/amsacta/6497

This Version is available at:
https://hdl.handle.net/10419/245894

ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Within-between decomposition of the Gini index: a novel proposal

Federico Attili

Quaderni - Working Paper DSE N°1153

Within-between decomposition of the Gini index: a novel proposal

Federico Attili*

July 23, 2020

Abstract

The socioeconomic impact of spatial concentration has been receiving an increasing attention during the last two decades. Consequently, the necessity of effective measures of this phenomenon has increased too. This paper considers a population partitioned by subgroups and develops a decomposition of the Gini index in two components, which measure the within and the between group inequality and are also particularly effective to quantify spatial concentration. Indeed, they possess a crucial property which overcomes important issues that may arise using any Gini index decomposition in the spatial context, following a recent approach. In addition, the availability of an only-two highly informative components decomposition provides in numerous applications and several frameworks further significant advantages in the determination of the contributions to global inequality of the intra and the inter groups differences. The ability of the components to capture these phenomena is supported by a parametric bootstrap procedure. This highlights extremely high correlations between the components and two axiomatically derived benchmarks. The presentation of a case study concerning the income distribution in the Italian provinces concludes the works, the informativeness and the interpretative advantages of the proposed decomposition appear evidently.

Keywords Inequality decomposition • Gini index • Regional inequality • Spatial concentration

JEL classification: D31; D63; O15; R10; R12

*Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy.
E-mail: federico.attili2@unibo.it

Non technical summary

In the last two decades an increasing attention has been paid to the study of between territories inequality and its socioeconomic impact. Driven by this cutting-edge attention on territorial inequality and, more in general, on the effects of spatial concentration - which accounts for both inequality and its spatial distribution - a particular new emphasis has to be also devoted to the ability of effectively measuring these phenomena.

Several measures of spatial concentration have been proposed in literature. They can be very useful but they also introduce some kind of arbitrariness. An alternative solution might be to exploit the information nested in the most used inequality measure, following a recent approach. We specifically consider the Gini index and propose a within-between decomposition of the index that nests considerable information on both the inequality and its spatial distribution.
The decomposition proposal deals with partitioned population and only consists of the two intra and inter groups inequality. The availability of an only-two highly informative components provides relevant advantages in a descriptive context or a regression task in terms of both interpretability and parsimony. To our knowledge, this is the first attempt of decomposition of the Gini index dealing with a partitioned population which comes up with just two components. We also notice that all the well known subgroup decompositions of the Gini index possess three components and a between component that only accounts for the variability of the means of the groups. This may be an oversimplification when the interest lies on the overall distance between distributions. Thanks to a crucial property, the between component from our proposal solves this potential drawback and further issues that may arise employing any Gini index decomposition in the spatial context.

We demonstrate that both the components are highly informative thanks to their properties and because - as we support by employing a parametric bootstrap procedure - they strongly and positively correlate with their relative benchmark. The two benchmarks which have been considered are developed *ex ante* to reach their objective - to measure the within and the between inequality - without constraining their sum to a preassigned formula, such as in our case with reference to the Gini index. They are derived from literature and follow an axiomatic approach.
Hence, the introduced components approximately provide all the information contained in the benchmarks and observe their axiomatically derived properties, despite the fact that they are derived with a decomposition boundary - and not independently as for the benchmarks.

In addition, we prove that the same levels of correlation also hold in a real data analysis. We apply the proposed decomposition to the municipality based *Income and principal Irpef variables* statistical data, which are available on the Open-Source Data released by the Department of Finance of the Italian Ministry of Economy and Finance. In the bootstrap procedure the correlation values are calculated simulating from independent scenarios. In the real data analysis we complement the assessment calculating correlations over time - our analysis ranges from 2000 to 2017 - and over different territorial aggregation. This strengthens the already mentioned results on the informativeness of the components.
With the same data - focusing on the income distribution of the Italian provinces - we also highlight the advantages of the proposed between component share in assessing the spatial distribution of inequality and the interpretative benefits that a two-component decomposition ensures in empirical contexts.

In fact, the decomposition is inspired by the spatial framework. Nonetheless, several applications of our decomposition are also meaningful and convenient outside of this context: groups could be defined by several factors such as gender, education level, occupation, race, age, or other criteria.

1 Introduction

Inequality between territories of the same country has been diffusely considered as just ancillary in the analysis of the inequality for a long period. But the presence in several countries of many territories which have been left behind is leading to increasing socioeconomic problems and facilitates populism, as Rodríguez-Pose (2018) pointed out. An increasing attention has been recently paid to the study of place-based policies (see e.g. Neumark and Simpson, 2015; Kline and Moretti, 2014). Driven by this cutting-edge attention on the territorial inequality and, more in general, on the effects of spatial concentration - it accounts for both inequality and its spatial distribution - a particular new emphasis has to be also devoted to the ability of measuring these phenomena.

The inequality indexes such as the Gini or the Theil index (only to mention the most used) are invariant to permutations of the units in the space. This can lead to the same inequality value associated to drastically different spatial patterns of the variable of interest - in this paper we will refer to income - as Arbia (2001) and Arbia and Piras (2009) effectively clarified. This means that the classical inequality measures can not be used to assess spatial concentration because they only partially capture this phenomenon. A symmetric argument holds for the spatial autocorrelation measures (see e.g. Moran, 1950; Ord and Getis, 1995), which only deal with the income distribution along space but do not capture concentration.

Several measures of spatial concentration have been proposed in literature, see among the others Ellison and Glaeser (1997), Campante and Do (2007), Bickenbach and Bode (2008) and Arbia and Piras (2009). They can be very useful but they introduce some kind of arbitrariness: in Ellison and Glaeser (1997) and Bickenbach and Bode (2008) the choice of a reference distribution is required; the proposal introduced in Campante and Do (2007) defines a centered index of spatial concentration and is appropriate when there is a place that can be considered central w.r.t. the others; in Arbia and Piras (2009) a class of measures of spatial concentration which depend on the choice of a spatial autocorrelation statistic is developed.
An alternative solution might be to combine inequality and spatial autocorrelation measures, as firstly suggested in Arbia (2001). The introduced proposal was criticized in Bickenbach and Bode (2008) which considers it as an ad hoc measure, but a similar approach was also adopted in A. Shorrocks and Wan (2005). Here, the authors suggest to exploit a subgroup decomposition of an inequality index in the spatial context. This means partitioning the population into geographical regions and evaluating a within component that measures the intra-territories inequality by a weighted average of the inequality index in each region; and a between term that measures the variation in the means of the regions. The idea of Shorrocks and Wan consists in assessing the distributional impact of spatial factors by jointly considering the overall inequality and the share of the between component. In this manner both the inequality and its spatial distribution - the key features of spatial concentration - should be under control. Rey and Smith (2013) also pursued a similar idea. They specifically consider the Gini index and propose a within-between decomposition of the index arguing that it potentially nests sufficient information on both the inequality and its spatial distribution. The main diversities between the two approaches are that the decomposition in Rey and Smith is not a subgroup decomposition and does not require a partition of the population. Rey and Smith have decomposed the index according to a matrix which defines pairs of neighbours and non-neighbours. The differences between the former constitute the within component while the latter sum up to the between term. More on the properties of the two strategies will be discussed in the following section.

In this paper we follow these recent approaches and introduce a novel proposal to decompose the Gini index. It deals with partitioned population and only consists of two components - within and between - which result to be highly informative on the intra and the inter groups inequality. To our knowledge, this is the first attempt of decomposition of the Gini index dealing with a partitioned population which comes up with just two components. We will show that its properties guarantee important benefits and makes it also particularly effective to measure spatial concen-

tration, in the spirit of Arbia (2001), A. Shorrocks and Wan (2005) and Rey and Smith (2013). In fact, the work is inspired by the spatial framework that is clearly considered throughout the work - the groups could be the regions or the sub-regions in a country, or the countries in a confederation - and we often refer to any within-between decomposition as to a spatial decomposition; and to the within and the between terms as to its spatial components. Nonetheless, several applications of our decomposition are also meaningful and convenient outside of this context: groups could be defined by several factors such as gender, education level, occupation, race, age, or other criteria.

In Section 2 we present the main features of the most common decompositions of the Gini index. Then we discuss the possibility of a two-component decomposition dealing with partitioned population to exist; and its potential advantages. Section 3 introduces our decomposition methodology when the groups are equal-sized and explains its ratio; a crucial property of the obtained components is also derived. Section 4 generalises the procedure to the different-sized groups case. In presenting Section 5 we stress how a rigorous axiomatic approach was followed to accept an index as an inequality measure (see Allison, 1978 for an effective overview). We believe the same approach should also be pursued when evaluating the components of an inequality index decomposition. Thus Section 5 shows, by employing a parametric bootstrap procedure, that the components strongly and positively correlate with two related benchmarks. They are inspired by literature and axiomatically derived to measure the within and the between inequality. Section 6 proves that the same levels of correlation also hold in a real data analysis, strengthening the preceding section results on the informativeness of the components; it also highlights the advantages of the proposed between component share in assessing the spatial distribution of inequality and the interpretative benefits that a two-component decomposition ensures in empirical contexts. Section 7 gives conclusive remarks.

2 Decomposing the Gini index

A very comprehensive outline of the most common subgroup Gini index decompositions (Bhattacharya and Mahalanobis, 1967; Rao, 1969; Pyatt, 1976; Mookherjee and A. Shorrocks, 1982; Yitzhaki and Lerman, 1991; Dagum, 1997) is provided in Radaelli (2010). These proposals are developed by different approaches but the resulting components - except the ones derived in Yitzhaki and Lerman (1991) - coincide. They all rely on a population of N individuals which is partitioned in K groups; and exhibit three components: the two spatial components and a third term.

The within components measure the intra-territories inequality by a weighted average of the Gini index in each group, where the weights vary with the considered decomposition. Let G_k be the value of the Gini index in group $k = 1,\dots,K$; μ_k and n_k its mean and dimension; μ the overall mean. In the within component from Yitzhaki and Lerman (1991):

$$G_w^{YL} = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)\left(\frac{\mu_k}{\mu}\right) G_k = \sum_{k=1}^{K} s_k G_k \tag{1}$$

each weight s_k is immediately interpretable as the income share of the group k. In the other proposals - we will refer to them considering their first specification from Bhattacharya and Mahalanobis (1967) - the income shares are multiplied again by the population shares:

$$G_w^{BM} = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2\left(\frac{\mu_k}{\mu}\right) G_k = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right) s_k G_k \tag{2}$$

The between component also depends on the decomposition choice, but every proposal presents a measure which substantially just compares the means of the groups. As for the two decompositions we are considering, they are:

$$G_b^{YL} = \frac{2}{\mu} Cov\left(\mu_k, \frac{\frac{1}{n_k}\sum_{i=1}^{n_k} R_{ik}}{N}\right) \tag{3}$$

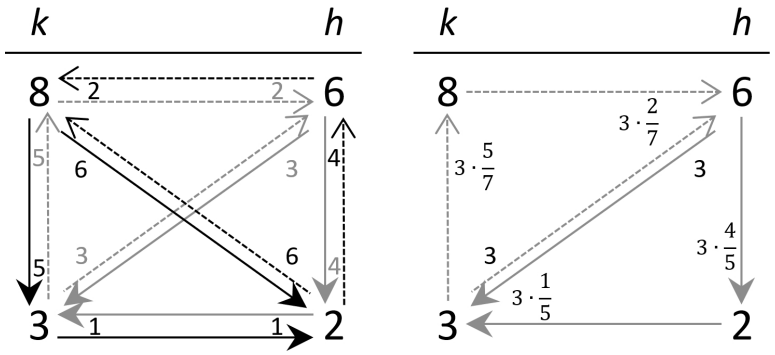
where R_{ik} is the rank of the unit i from the group k in the overall population; and

$$G_b^{BM} = \frac{1}{2\mu}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{n_k n_h}{N^2}|\mu_k - \mu_h| \tag{4}$$

The two components in eq. (3)-(4) appear very different in their form but they account accordingly for low levels of territorial inequality when the groups have similar means - they are zero if the means are identical - and vice versa, regardless of the potential different dispersion within each group. This could be an oversimplification when the overall distance between the distributions is relevant, as Ebert (2010) effectively pointed out.
The third term of the decompositions - which disappears if the group distributions do not overlap - has been left uninterpreted in the first attempts of decomposition and considered just as the residual which guarantees the identity. Then several interesting interpretations in terms of groups overlapping or stratification have been proposed (Yitzhaki and Lerman, 1991, Yitzhaki, 1994) and all the three components became potentially useful and informative.
Nonetheless, it is incontestable that every two-component decomposition possesses an important advantage: all the information of the inequality index and its decomposition can be provided specifying only the overall inequality and its between component. Indeed, the collinearity between the inequality index and the two spatial components allows to provide all the information contained in the decomposition by delivering only two of these three values, with relevant advantages in a descriptive context or a regression task in terms of both interpretability and parsimony. The number and the specification of the components may become relevant also when measuring spatial concentration using an inequality index and the share of its between component, as will become clear in Section 6.
These arguments and the potential oversimplification generated by a between component relying only on the means of the groups encouraged us to look for a novel solution. As we will show, this possesses both the advantage of a two-component decomposition and a between component which informs about the differences between the overall group distributions.
Actually, the spatial Gini index decomposition presented in Rey and Smith (2013) is also composed only by the two spatial components and its between component directly relies on the individual values - not only on the means. However it is only appropriate when the population groups do not form a partition. When they do, a relevant drawback arises: the between component is positive even if the groups have the same distribution, hence it may tend to overestimate the between group inequality in cases of groups characterised by similar distributions.

As we have anticipated, in this paper we introduce the first two-component decomposition of the Gini index dealing with a partitioned population. This could appear in contrast with the well known results presented in Bourguignon (1979), A. F. Shorrocks (1980) and A. F. Shorrocks (1984) where the class and the properties of the measures which are additively decomposable are derived. The Gini index is not additevely decomposable in the sense intended by these works, which is partitioning the population into geographical regions and expressing the index as a weighted average of the inequality values within each group, plus some contribution arising from the variation in the means of the regions. Such a decomposition on the Gini index requires a third component, as we explained above.
Differently, our proposal determines a between component which does not rely on the means of the groups but directly depends on the incomes of individuals. This allows a two terms decomposition to exist and to perform better than the other decomposition proposals when the groups overlap[1]. In addition, as we will show in the following section, the between term is null if and only

[1] If the groups do not overlap the third component of the subgroups decompositions disappears and the variation of the



(a) Couple differences composing g (b) Decomposition in the non-trivial case

Figure 1: A two-group-two-individual illustration

if the groups have the same distribution. The sufficiency of this condition solves the drawback of the decomposition in Rey and Smith (2013) while its necessity solves the oversimplification of a between component substantially relying just on the means of the groups.

3 The decomposition proposal

Keep on considering a population of N individuals. Let x_i be the income of the generic individual $i = 1,\dots,N$. The Gini index is defined as follows:

$$G = \frac{1}{2\mu N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right| = \frac{g}{2\mu N^2}$$

where $\mu = \frac{1}{N}\sum_1^N x_i$ is still the average income. The index accounts for the sum g of all the differences between individual incomes, averaging and normalizing by the factor $\left(2\mu N^2\right)^{-1}$, so that G is scale invariant and $G \in [0,1]$ if all the $x_i \geq 0$. Consider now the population partitioned in K equal-sized groups, i.e. $n_k = n \;\; \forall\; k = 1,\dots,K$ and define x_i^k to be the i-th element in the ordered vector $\mathbf{x}^k = (x_1^k,\dots,x_n^k)$ with $x_i^k \geq x_{i+1}^k \;\; \forall\; i = 1,\dots,n-1$. All the information concerning with the spatial distribution of inequality is within g, which can be reformulated as

$$g = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i^k - x_j^h\right| \tag{5}$$

The assumption of equal-sized groups might be thought as extremely simplistic but, as we will show in Section 4, it is only apparently limiting in the applicability of the proposal. Conversely, it provides a deeply innovative insight into the structure of the Gini index.
Look at the Figure 1 which shows a two-group-two-individual situation. Figure 1a highlights the absolute differences between all the couples of units, considered twice so that they constitute g if they are summed up. As the scheme suggests we distinguish three kind of differences: vertical, horizontal and diagonal ones. The vertical differences compare same-group pairs of elements. The horizontal differences compare same-rank pairs from different groups. The diagonal differences compare different-rank pairs from different groups.

As a point of departure, we would like a decomposition which assigns the vertical and the horizontal absolute differences to the within and the between component, respectively.

means of the groups fully characterises the between group inequality. Our proposals is useless in such cases.

Despite the fact that the diagonal differences involve different-group pairs, they partly reflect the vertical differences - i.e. the within inequality - and can not be totally addressed to the between group inequality. Consider a situation with two identical groups: the possibly positive values of the diagonal differences equal the vertical ones and can not be addressed to the absent between group inequality.
In Lambert and Aronson (1993) the residual term is described as "at once a between groups and a within groups effect: it measures a between groups phenomenon, overlapping, that is generated by inequality within groups" (p. 1224). The sum of the diagonal differences especially accommodates this characterisation and can be thought as a kind of residual term of this decomposition strategy. In fact, this term is unambiguously dependent on the interaction of the intra and the inter groups differences but it does not disappear when the groups do not overlap. Thus it can not be interpreted as an overlapping component.
This is not an issue for our purpose because with the equal-sized group representation a quite intuitive and effective paradigm to decompose the diagonal differences exists and generates two contributes to the spatial components.
Sometimes, an extremely reasonable strategy is viable to decompose the diagonal lines and to disentangle the within contribution from the different-group differences. We invite the reader to inspect the diagonal differences and to move along the legs of the triangles depicted in the scheme in Figure 1a. Look at the solid black diagonal line as an example: the absolute difference between the richest of the group k and the poorest of the group h is 6. The former is 5 units richer than his group poorest individual, who is 1 unit richer than his counterpart in group h. A similar argument holds from the opposite point of view, looking at the difference between the poorest of group h and the richest of group k - dashed black diagonal line. The diagonal absolute differences between the two individuals considered in the example are predominantly due to the within inequality in the two groups and reflect it. Accordingly, we suggest to split the global contribution to g of the couple composed by the richest of the group k and the poorest of the group h ($6+6=12$) assigning $5+4=9$ to the within component and $1+2=3$ to the between one.
This straightforward strategy is not directly feasible in situations - such as those represented by the gray lines - in which the three values involved in the path along the legs do not increase or decrease monotonically as in the black lines situations, namely when the product between the horizontal and the vertical difference is negative. In such cases we should subtract the horizontal value from the vertical one to obtain the value of the difference along the diagonal, with paradoxical effects on interpretation if we subtract the horizontal value - 1 - from the between component. As an example, imagine to replace the poorest individual of group h with a poorer one. We would obtain a lower value of the between inequality (we subtract a value $3-(2-\varepsilon)>1$) even though the intuition suggests the between inequality is now higher - the poor group is poorer - and we should sum up a positive value to it.
To overcome this problem, we propose to reduce both the vertical and the horizontal values proportionally to make their sum equal to the diagonal difference. Hence, both the vertical and the horizontal values are divided by their sum and multiplied by the diagonal difference - look at the Figure 1b for an illustration. In other words we suggest to split each diagonal difference proportionally to the vertical and the horizontal ones and to assign these two (positive) values to the within and the between component, respectively. Thanks to this solution we preserve both reasonable proportions between the values added to the spatial components - these proportions observe the black lines decomposition argument - and the Gini index compliance, i.e. the possibility to have a two-components decomposition. With this overall strategy the contributions of the diagonal differences to the within and between inequality mimic to the utmost the vertical and horizontal differences. This is the key that makes the two components highly informative.
Until now, we have just presented the intuition at the base of the decomposition. We will now generalise this strategy, formalise the decomposition proposal and deliver the two spatial components. At the end of this section, the preannounced property of the two components is derived.
For each difference $|x_i^k - x_j^h| > 0$, we define the sum of the (absolute) legs $L_{ij}^{kh} = |x_i^k - x_j^k| + |x_j^k - x_j^h|$

and their product $c = (x_i^k - x_j^k)(x_j^k - x_j^h)$. We can write:

$$\left|x_i^k - x_j^h\right| = \left|x_i^k - x_j^h\right| \frac{\left|x_i^k - x_j^k\right| + \left|x_j^k - x_j^h\right|}{\left|x_i^k - x_j^k\right| + \left|x_j^k - x_j^h\right|} = \left|x_i^k - x_j^h\right| \frac{\left|x_i^k - x_j^k\right|}{L_{ij}^{kh}} + \left|x_i^k - x_j^h\right| \frac{\left|x_j^k - x_j^h\right|}{L_{ij}^{kh}} =$$

$$= \begin{cases} \left|x_i^k - x_j^k\right| + \left|x_j^k - x_j^h\right| & \text{if } c \geq 0 \\ \left|x_i^k - x_j^k\right| \frac{|x_i^k - x_j^h|}{L_{ij}^{kh}} + \left|x_j^k - x_j^h\right| \frac{|x_i^k - x_j^h|}{L_{ij}^{kh}} & \text{if } c < 0 \end{cases} \tag{6}$$

where to obtain the first equation we exploited the fact that $c \geq 0$ implies $|x_i^k - x_j^h| = |x_i^k - x_j^k + x_j^k - x_j^h| = L_{ij}^{kh}$.

We propose to assign the first and the second addends of eq. (6) to the within and the between inequality components, respectively. The first equation fills all the trivial-case decompositions (the vertical differences when $k = h$; the horizontal differences when $i = j$ and the black diagonal differences kind when $k \neq h$ and $i \neq j$) while the second stands for analogous situations to the two sketched by the gray lines.

We stress here the importance of the equal-sized groups hypothesis which guarantees that each unit has a counterpart in each other group, i.e. given the couple (x_i^k, x_j^h), the element x_j^k always exists. Notice that considering x_j^k or x_i^h in the decomposition is not an issue, since the Gini index counts each difference twice by inverting the indices of the summations.

To simplify the notation in eq. (6) we define[2] $w_{ij}^{kh} = \frac{|x_i^k - x_j^h|}{L_{ij}^{kh}}$ and obtain:

$$\left|x_i^k - x_j^h\right| = \left|x_i^k - x_j^k\right| w_{ij}^{kh} + \left|x_j^k - x_j^h\right| w_{ij}^{kh} \tag{7}$$

It is always $w_{ij}^{kh} \in [0,1]$: $w_{ij}^{kh} \geq 0$ holds by definition and $w_{ij}^{kh} = 1$ iff $c \geq 0$. Hence, $|x_i^k - x_j^k| w_{ij}^{kh}$ and $|x_j^k - x_j^h| w_{ij}^{kh}$ can be considered, respectively, as the contributions of the difference $|x_i^k - x_j^h|$ to the within and the between group inequality; and w_{ij}^{kh} the vertical and the horizontal differences rescaling factor. This factor can be less than one because of the possibility of a proportional reduction of the horizontal and the vertical values before assigning them to the spatial components: as we showed in the numerical example, the difference $|x_i^k - x_j^h|$ can be less than $|x_i^k - x_j^k| + |x_j^k - x_j^h|$ (it happens iff $c < 0$), so we have to rescale the two addenda by the factor $w_{ij}^{kh} \leq 1$. This value decreases, by definition, according to the ratio between the diagonal difference and the summation between the horizontal and the vertical ones. This means that when two individuals are similar the contribution of their couple to the within and the between components are accordingly small, even if the two vertical and horizontal differences involved in the decomposition are greater. The latters account, as desired, only in the proportion between the two spatial contributions, while their absolute values depend on the starting level of the inequality in the couple.

The Gini index decomposition proposal can be now derived just by substituting eq. (7) into eq. (5). Denoting $\sum_{h=1}^{K} w_{ij}^{kh} = w_{ij}^{k}$ and $\sum_{i=1}^{n} w_{ij}^{kh} = w_{j}^{kh}$, we have:

$$g = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{K} \left|x_i^k - x_j^k\right| w_{ij}^k + \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{j=1}^{n} \left|x_j^k - x_j^h\right| w_j^{kh} = g_w + g_b \tag{8}$$

and

$$G = G_w + G_b = \frac{g_w}{2\mu N^2} + \frac{g_b}{2\mu N^2}$$

The Gini index appears composed by only two terms. We propose to interpret the first as the

[2] We set w_{ij}^{kh} to zero *a priori* if $|x_i^k - x_j^h| = 0$

within component of the inequality, because it depends on the summation of the contribution of all the couples of units belonging to the same group, i.e. the couple differences multiplied by the relative weights w_{ij}^k; and the second as the between component of inequality, because it depends on the summation of the contribution of all the couples of same-rank units in different groups, i.e. the couple differences multiplied by the relative weights w_j^{kh}.
Note that the within and the between component involve, respectively, the weights w_{ij}^k and w_j^{kh} which do not exclusively depend on the differences within or between groups. This is not a drawback but a necessary property which allows each contribute of the intra-group (same-rank units) differences to the within (between) inequality to depend on how it affects the different-rank units differences, i.e. the diagonal ones. Indeed w_{ij}^k (w_j^{kh}) grows when the diagonal differences are enlarged by the vertical (horizontal) ones.

Given that $w_{ij}^{kk} = 1$, $w_{jj}^{kh} = 1$ and $w_{ij}^{kh} \geq 0 \; \forall i,j,k,h$ imply $w_{ij}^k \geq 1 \; \forall k,i,j$ and $w_j^{kh} \geq 1 \; \forall j,k,h$, the crucial property - that we have mentioned before - of the two spatial components is ensured:

$$G_w = 0 \iff |x_i^k - x_j^k| = 0 \quad \forall \; i,j,k$$
$$G_b \; = 0 \iff |x_j^k - x_j^h| = 0 \quad \forall \; j,k,h$$

The first relation ensures the within component to be zero iff all the intra-group differences are zero, i.e. all the individuals equal their group mean. Actually, all the decompositions we have considered until now possess this property. The second relation is symmetric to the first and guarantees the between component to be zero iff all the same-rank differences are zero, i.e. all the groups have the same distribution. This is a very reasonable condition too - at the end of the previous section we have already stressed the relevant advantages it ensures - but it is satisfied just by our component.
As stressed before, the structure of the weights w_{ij}^k and w_j^{kh} depends on many units and not only on the two involved in the difference they multiply. Unfortunately, this compromises the analytical tractability of the expressions we obtained for the two components and hinders the derivation of additional properties. An alternative approach to corroborate the appropriateness of the two components is followed. Section 5 will show that each component strongly and positively correlates with a related benchmark. The two benchmarks which have been considered are developed *ex ante* to reach their objective - to measure the within and the between inequality - without constraining their sum to a preassigned formula, such as in our case with reference to the Gini index. They are derived from literature and follow an axiomatic approach.
We have essentially demonstrated that the two components have a correct starting point - zero - and we will demonstrate that they move in the right direction and proportion with reference to different scenarios, because they strongly correlate with their relative benchmark. These two properties have a remarkable consequence: the components approximately provide all the information contained in the related benchmarks and observe their axiomatically derived properties. In addition, the extent of the approximations is small because it inversely depends on the (high) correlation values.

In the following section we show that the equal sized groups hypothesis is not binding. It is necessary to understand the decomposition arguments but the proposal can be easily generalised to cope with a more realistic situation in which groups have different sizes.

4 Different-sized groups

Consider now the K groups having a vector of sizes $\mathbf{n} = (n_1, \ldots n_K)$ where $\sum_{k=1}^K n_k = N$. We have:

$$g = \sum_{i=1}^{N} \sum_{j=1}^{N} \left| x_i - x_j \right| = \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \left| x_i^k - x_j^h \right|$$

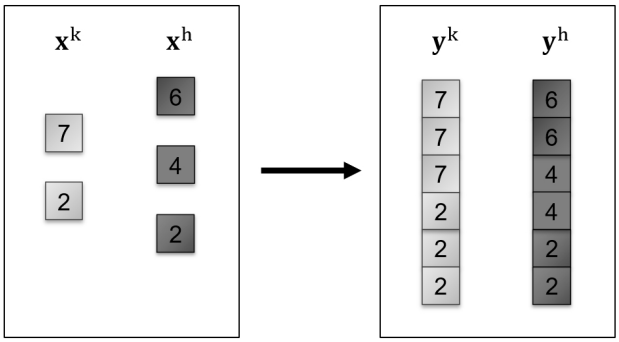


Figure 3: Exact approach: a two-group illustration

and a way to guarantee the element x_j^k to exist is needed to employ the decomposition proposal.

We can proceed by two distinct approaches. The first allows to evaluate with no approximation the two components but necessitates potentially unaffordable computations. The second drastically reduces computational requirements paying the cost of a negligible approximation.

The *exact* approach. It exploits the principle of population. It considers a new common size $n = mcm(\mathbf{n})$ and the resampling weights $p_k = \frac{n_k}{n}$ so to build the vectors $y^k = (y_1^k, \ldots, y_n^k) = (\underbrace{x_1^k \ldots x_1^k}_{p_k^{-1}}, \ldots \underbrace{x_{n_k}^k \ldots x_{n_k}^k}_{p_k^{-1}})$. Defining $l_{m_k}^i = p_k^{-1}(i-1) + m_k$, by construction we have $x_i^k = y_{l_{m_k}^i}^k$, $\forall\, i = 1, \ldots n_k$ and $\forall\, m_k = 1, \ldots, p_k^{-1}$. Hence $\forall\, (k,h) \in \{1, \ldots, K\} \times \{1, \ldots, K\}$ the following holds:

$$\sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \left| x_i^k - x_j^h \right| = \sum_{i=1}^{n} \sum_{j=1}^{n} p_k p_h \left| y_i^k - y_j^h \right| \tag{9}$$

Proof.

$$\sum_{i=1}^{n} \sum_{j=1}^{n} p_k p_h \left| y_i^k - y_j^h \right| = \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \sum_{m_k=1}^{p_k^{-1}} \sum_{m_h=1}^{p_h^{-1}} p_k p_h \left| y_{l_{m_k}^i}^k - y_{l_{m_h}^j}^h \right| = \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \sum_{m_k=1}^{p_k^{-1}} \sum_{m_h=1}^{p_h^{-1}} p_k p_h \left| x_i^k - x_j^h \right| =$$
$$= \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} p_k p_h p_k^{-1} p_h^{-1} \left| x_i^k - x_j^h \right| = \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \left| x_i^k - x_j^h \right|$$

□

Substantially, imagine to have two groups composed, respectively, of two and three individuals, as the ones reported in the left rectangle of Figure 3, and replace them with those in the right part. By the principle of population $\mathbf{x}^k$ and $\mathbf{y}^k$ (as well as $\mathbf{x}^h$ and $\mathbf{y}^h$) are identical from the point of view of their internal inequality, and also the comparison between the two groups should remain unvaried. Notice that each difference in the left scheme appears in the right scheme nine times if the couple belongs to $\mathbf{y}^k$, four times if it belongs to $\mathbf{y}^h$ and six times if the units belong to different groups. In order to respect the principle of population we have only to adjust for this effect. This is what eq. (9) means.

The Gini index numerator can be now decomposed with an analogous technique to the one employed deriving eq. (8). The following is obtained:

$$\begin{aligned} g &= \sum_{i=1}^{N} \sum_{j=1}^{N} \left| x_i - x_j \right| = \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \left| x_i^k - x_j^h \right| = \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} p_k p_h \left| y_i^k - y_j^h \right| = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{K} \left| y_i^k - y_j^k \right| w_{ij}^k + \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{j=1}^{n} \left| y_j^k - y_j^h \right| w_j^{kh} = g_w + g_b \end{aligned} \tag{10}$$

The only difference w.r.t. eq. (8) is in the new weights $w_{ij}^k = \sum_{h=1}^K p_k p_h w_{ij}^{kh}$ and $w_j^{kh} = \sum_{i=1}^n p_k p_h w_{ij}^{kh}$. They are the general case of the previous defined weights and incorporate the information needed to preserve the original impact of each couple.

As we said, in most cases this approach requires an unaffordable computational effort because of the potentially huge magnitude of the minimum common multiple. Thus, we present an alternative procedure which drastically reduces computational requirements paying the minimal cost of a negligible approximation.

Quantilisation. We propose to replace the vectors $\mathbf{y}^k$ with others K vectors: for each group we select a vector composed by n quantiles from the income vector of the group. As for the resampling weights, their calculation is the same employed in the *exact* approach, but now nothing ensures $n \geq n_k$ so it can be $p_k > 1$. The decomposition proposal has the same form of eq. (10) but G, G_w and G_b now incur in some approximation.

In order to employ this method there are the definition of quantile and the value of n to be selected.

As for the former, we advise the definition 7 reported in Hyndman and Fan (1996), which is also the default definition adopted by the *quantile*() function in the software R. Given each vector $\mathbf{x}^k \in \mathscr{R}^{n_k}$, accordingly to this definition and in order to alter as little as possible the *quantilisation* results w.r.t. the *exact* calculation, we suggest to interpolate linearly the vertices $\left(\frac{i-1}{n_k-1}, x_i^k\right)$ where $i = 1,\ldots,n_k$, and then to estimate the n quantiles by determining the values associated to the probabilities

$$prob_j = \frac{j-1}{n-1} \quad j = 1,\ldots,n \tag{11}$$

on the resulting piecewise linear curve.

As for the latter, we define $w_k = \frac{n_k}{\sum_{k=1}^K n_k}$ and recommend the value

$$n = \sum_{k=1}^K w_k n_k = \frac{\sum_{k=1}^K n_k^2}{\sum_{k=1}^K n_k} \tag{12}$$

This expression determines n as the average of the n_k, each weighted by its own share of population w_k.

A formal assessment which legitimises the decisions proposed both for the quantile definition and for the value of n is provided in the appendix. Here we only inform about the negligibility of the approximation which the *quantilisation* procedure cope with when the advised definitions are employed.

Actually, the optimal performance associated to these two decisions should not come as a surprise. The outstanding results of the proposed choice of n derives directly from the consistency with the resampling weights system. This choice assigns greater weights w_k to the sizes of the most sized groups, which is desirable because these are the groups the biggest values of p_k are associated to. It is reasonable to preserve their information choosing a large n and resampling the smaller groups. But if many small groups are present n is attracted towards their small size. Here the *quantilisation* reduction of big groups and the related loss of information are preferred to pay the approximation cost of the *quantilisation* resampling for many small groups. Nonetheless, notice that it could be also acceptable to choose a value $n << min(\mathbf{n})$ if $min(\mathbf{n})$ is high and a computational cost saving choice is required.

With the choice determined by eq. (11) the values p_j partition the interval $[0,1]$ in $n-1$ equal parts, with 0 and 1 two of the n vertices of the partition. It is straightforward to verify that, also employing the quantile selection procedure discussed, $min(\mathbf{x}^k)$ and $max(\mathbf{x}^k)$ are preserved for each n and k. Moreover, if $n = n_k$ for any k the vectors $\mathbf{x}^k$ are entirely preserved, too. Both this properties hold at the same time only employing the definition 7 and the discussed choice of the values p_j. They ensure robustness to the *quantilisation* procedure w.r.t. outliers and contribute to

explain the negligible approximation incurred.

We finally observe that a generalised version of the quantilisation procedure may be wanted to increase its appropriateness in coping with weighted data, i.e. to consider the weights of the observations. We suggest - but other choices we do not discuss here can be also suitable - to employ functions which account for the sample weights in returning quantiles (as the R function *wtd.quantile()*) in place of a *standard* quantile function (as the R function *quantile()*). This is equivalent to applying a *stantard* quantile function to vectors with the income of each unit in the starting vector repeated as many times as the value of its weight states. We consequently advise to consider the weights summation in each group as the group weight in eq. (12), in place of the defined w_k. This is the way we will proceed in Section 6.

In the analysis which follows each group k will be replaced by the n quantiles selected from $\mathbf{x}^k$. The value of n will be determined by eq. (12). The definition 7 and eq. (11) will be employed to select the n quantiles.

In the next section a parametric bootstrap will exhibit that each component of the introduced decomposition strongly and positively correlates with the related benchmark. These results will be compared to the one achieved by employing the components from eq. (1)-(4) in place of the introduced ones. In Section 6 the reliability of the simulation process will be strengthened showing very similar results obtained using real data.

5 Correlation with benchmarks

The two benchmarks which we consider are developed *ex ante* to measure the within and the between inequality, without constraining their sum to a preassigned formula as it happens for the Gini index decompositions. They are derived from literature and observe an axiomatic approach.

The following equation defines the reference measure of the global intra-group inequality:

$$W_r = \sum_{k=1}^{K} s_k G_k$$

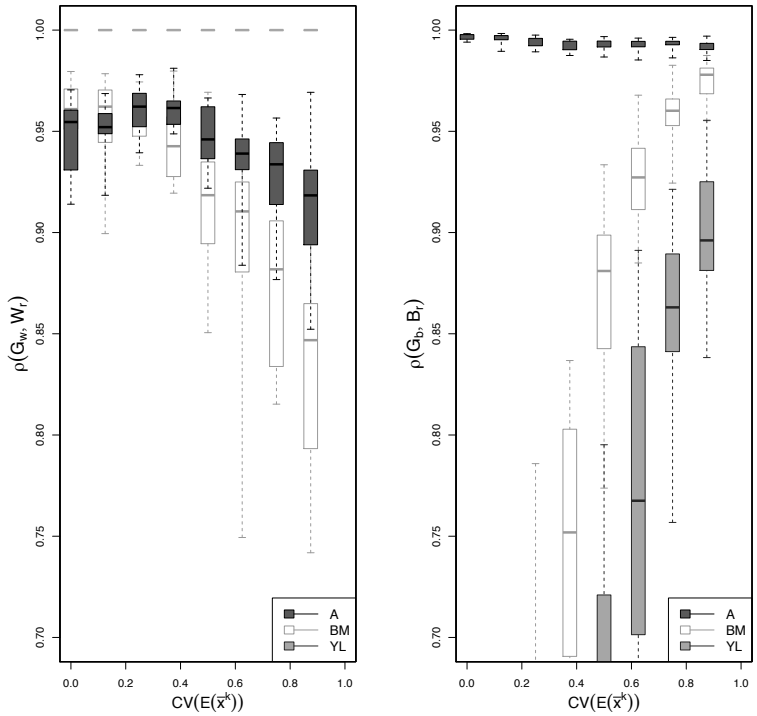
where G_k is the Gini index in group k and s_k is the share of income possessed by group k. Hence, the **within benchmark** is defined as a weighted average in which the Gini of each group is weighted by its own share of income. Every G_k incorporate the axiomatic approach observed by the Gini index and the related properties. Thus, the global properties of W_r depends on them and on the aggregating consequences of the weighted mean, which in this case assigns a greater weight to the inequality values of the groups which possess the biggest shares of income. This measure coincides with the within component reported in (1). We selected this weighting choice due to its immediate interpretability, as discussed in Section 2.

As for the **between benchmark**, the following index has been selected as the reference measure of the global inter groups inequality:

$$B_r = \sum_{k=1}^{K} \sum_{h=1}^{K} \left(\frac{n_k n_h}{N^2} \right) Eb_{kh}$$

where Eb_{kh} is the diversity measure between the two groups k and h proposed in Ebert (1984). He proposes a class of measures dependent on a parameter r. Here Eb_{kh} is the measure corresponding to $r = 1$, which employs the absolute difference as a distance. A slight modification - the incomes are standardized dividing by the average of the population μ - is required to guarantee the scale invariance criterion to be respected in addition to the other properties the index already observes. The measure is defined as:

$$Eb_{kh} = \frac{1}{m\mu} \sum_{i=1}^{m} \left| x_i^k - x_i^h \right|$$



(a) **Within components.** (b) **Between components.**

Figure 4: Correlations between the intra and the inter groups inequality benchmarks and the related spatial components from the three Gini index decompositions: A, YL and BM in the legend stand for the decomposition proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967), respectively. The figures report the values obtained simulating with $K = 30$ and $\mathbf{n} \sim U\left([100,500]^K\right)$ for different values of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$

where $m = min(n_k, n_h)$ and x_i^k is the i-th of the m quantiles selected from the income vector of the group k. A preceding proposal by Dagum (1980) had already developed a measure of *economic distance* but it has been criticized by A. F. Shorrocks (1982) because of its asymmetric nature. Ebert proposal, instead, presents all the properties of a distance and observes a more general axiomatic approach. In addition, it perfectly reflects our idea that a measure of inequality between groups has to compare their overall distributions. B_r inherits all these properties and its value depends on them and on the aggregating consequences of the weighted mean, which in this case assigns greater weights to the inequalities of the couples selected from the groups which possess the biggest shares of couples.

In order to evaluate the extent of the correlation between the components of the three considered decompositions and the related benchmarks, a parametric bootstrap algorithm has been employed. It considers three parameters: the number of groups, the distribution of $\mathbf{n}$ and the expected coefficient of variation between the averages of the groups $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$.
The algorithm fixes these parameters and generates incomes from a lognormal distribution. Properly, the third parameter is the vector composed by the minimum and the maximum values of the uniform distribution the expected value of the lognormal distribution is drawn from. These two values straightforwardly determine the value of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$. More details about the income simulation procedure and its theoretical foundations can be found in the Appendix. The algorithm evaluates 50 times all the involved indices to estimate the two triples of correlations between the

two spatial components of the three involved decompositions and the related benchmarks. It runs this procedure 20 times and obtains 20 replicates of the following couple of triples of correlation estimates:

$$\begin{bmatrix} cor_i\left(G_w^A, W_r\right) \\ cor_i\left(G_w^{YL}, W_r\right) \\ cor_i\left(G_w^{BM}, W_r\right) \end{bmatrix}; \quad \begin{bmatrix} cor_i\left(G_b^A, B_r\right) \\ cor_i\left(G_b^{YL}, B_r\right) \\ cor_i\left(G_b^{BM}, B_r\right) \end{bmatrix}$$

with $i = 1, \ldots, 20$. Following the notation introduced in Section 2, the superscripts A, YL and BM identify the decomposition proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967), respectively.
Each of the 20 replicates of triples constitute one triple of boxplots in the pertaining diagram reported in Figure 4. The value of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ is varied to obtain all the eight triples of boxplots. Hence, for each considered value of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$, the three kinds of boxplots in Figure 4a (4b) represent the replicates of the triples: black, gray and white boxplots describe the distribution of the correlation that the within (between) benchmark has, respectively, with the within (between) components proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967). Figure 4 reports the values obtained simulating with $K = 30$ and $\mathbf{n} \sim U\left([100, 500]^K\right)$.
We have varied the value of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ to evaluate correlations in contexts characterised by different degree of homogeneity in the vector of the means of the groups. This allows us to notice the evident advantages that the proposed between component has w.r.t. the between components derived from the alternative decompositions - which only account for the variability of the means of the groups - in situations characterised by low and medium levels of variability in the vector of the means of the groups. As for the within component, the one obtained using the decomposition in Yitzhaki and Lerman (1991) presents by definition a perfect correlation with W_r. However, better than G_w^{BM}, G_w^A always reports extremely high correlations with W_r.

Furthermore, the correlation values are studied by the same algorithm in multiple contexts by also varying the number of groups and the distribution of $\mathbf{n}$. A summary of the results for some representative choices of the parameters is reported in Table 1. It also contains the results already represented in Figure 4. In this table the 20 replicates produced to estimate each correlation distribution are summarised by their averages and standard deviations. These values - obtained from all the eight different levels of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ - are averaged pairwise determining four values for $\overline{\mu}$ and $\overline{sd}$. They correspond to low, medium-low, medium-high and high levels of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$.
The results are extremely encouraging. The correlation values of the proposed decomposition only marginally depend on the parameters specification. They just show some negligible variations both in the mean and the standard deviation. We notify the most perceptible. Higher K values negatively influence the within average correlations, but an increase in the values in $\mathbf{n}$ seems to absorb this small effect. The within average correlations also decrease for higher level of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$, while their standard deviations tend to increase. However, the averages are never behind 0.92 and the maximum standard deviation is $2.8 \cdot 10^{-2}$. As for the between correlation, it slightly decreases and shows higher standard deviations when the variability in $\mathbf{n}$ gets higher and both the values in $\mathbf{n}$ and K are small.
Despite all these details, the values of the correlations for the components of the introduced decomposition reported in Table 1 always maintain analogous advantages to those explained describing Figure 4. The sole exception occurs when the variability in $\mathbf{n}$ is small: in this case the results for the within component from eq. (2) are enhanced and turn out to be comparable with the decomposition introduced with this work. Indeed, the former perfectly correlates with the within component from eq. (1) when the groups are equal-sized - it can be easily proved. However, the correlation decreases rapidly as soon as the size variability increases, so this aspect remains a negligible issue. Summarising, Table 1 strengthens the conclusions drawn looking at Figure 4.

			Within components								Between components							
			$K=3$				$K=30$				$K=3$				$K=30$			
Magnitude of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$			L	L-M	M-H	H	L	L-M	M-H	H	L	L-M	M-H	H	L	L-M	M-H	H
$\mathbf{n}\sim U\left([10,20]^K\right)$	$\overline{\mu}$	A	0.972	0.966	0.964	0.961	0.943	0.942	0.944	0.941	0.988	0.986	0.984	0.978	0.990	0.989	0.983	0.977
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.386	0.465	0.704	0.807	0.117	0.313	0.468	0.668
		BM	0.981	0.977	0.968	0.957	0.963	0.960	0.949	0.932	0.771	0.840	0.903	0.952	0.829	0.881	0.908	0.944
	$\overline{sd}$	A	0.005	0.013	0.010	0.012	0.018	0.016	0.015	0.018	0.006	0.007	0.008	0.006	0.003	0.003	0.005	0.007
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.199	0.214	0.131	0.085	0.197	0.215	0.202	0.100
		BM	0.013	0.017	0.024	0.031	0.013	0.012	0.018	0.029	0.073	0.050	0.053	0.019	0.060	0.050	0.037	0.015
$\mathbf{n}\sim U\left([10,50]^K\right)$	$\overline{\mu}$	A	0.974	0.970	0.959	0.943	0.946	0.945	0.936	0.921	0.964	0.966	0.965	0.972	0.984	0.982	0.976	0.972
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.238	0.445	0.692	0.851	0.033	0.234	0.475	0.734
		BM	0.939	0.923	0.899	0.841	0.937	0.925	0.894	0.857	0.683	0.766	0.881	0.943	0.737	0.796	0.891	0.937
	$\overline{sd}$	A	0.009	0.010	0.012	0.020	0.012	0.018	0.020	0.028	0.028	0.031	0.030	0.025	0.006	0.006	0.008	0.009
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.249	0.180	0.137	0.066	0.162	0.241	0.232	0.138
		BM	0.034	0.043	0.052	0.080	0.024	0.033	0.049	0.047	0.124	0.093	0.057	0.041	0.080	0.069	0.031	0.019
$\mathbf{n}\sim U\left([100,200]^K\right)$	$\overline{\mu}$	A	0.968	0.962	0.949	0.938	0.947	0.953	0.940	0.921	0.996	0.995	0.994	0.992	0.995	0.994	0.993	0.992
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.019	0.525	0.797	0.912	-0.120	0.369	0.694	0.858
		BM	0.989	0.985	0.974	0.953	0.979	0.971	0.952	0.932	0.425	0.748	0.926	0.974	0.519	0.718	0.906	0.963
	$\overline{sd}$	A	0.011	0.012	0.014	0.019	0.017	0.013	0.016	0.020	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.003
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.265	0.173	0.072	0.029	0.202	0.175	0.114	0.049
		BM	0.011	0.011	0.020	0.044	0.009	0.012	0.017	0.017	0.155	0.087	0.033	0.009	0.123	0.082	0.027	0.012
$\mathbf{n}\sim U\left([100,500]^K\right)$	$\overline{\mu}$	A	0.970	0.966	0.952	0.930	0.950	0.960	0.942	0.921	0.996	0.995	0.994	0.993	0.996	0.993	0.993	0.992
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.040	0.437	0.812	0.908	-0.086	0.357	0.716	0.879
		BM	0.980	0.968	0.937	0.892	0.958	0.950	0.903	0.853	0.336	0.700	0.924	0.974	0.401	0.661	0.896	0.966
	$\overline{sd}$	A	0.009	0.014	0.011	0.017	0.015	0.010	0.017	0.025	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.003
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.233	0.207	0.071	0.034	0.208	0.162	0.075	0.038
		BM	0.013	0.022	0.039	0.073	0.017	0.015	0.041	0.041	0.184	0.109	0.032	0.012	0.125	0.092	0.034	0.011

Table 1: Correlations between the intra and inter groups inequality benchmarks and the related spatial components from three Gini index decompositions: the A, YL and BM in the third column stand for the decomposition proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967), respectively. The correlations are evaluated by the algorithm presented in this section for different values of K, $\mathbf{n}$ and $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$. In this table the vectors of replicates produced to estimate each correlation distribution from the eight different levels of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ are grouped in four classes, corresponding to low, medium-low, medium-high and high levels of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$. $\overline{\mu}$ and $\overline{sd}$ summarise the correlation distribution in such classes.

6 Validation on real data and the Italian provincial-based evidence

In this section we apply the proposed decomposition to the municipality based *Income and principal Irpef variables* statistical data, which are available on the Open-Source Data released by the Department of Finance of the Italian Ministry of Economy and Finance. It annually collects - our analysis ranges from 2000 to 2017 - data from the tax declarations related to the whole set of Italian taxpayers and reports several variables on a municipality base. The income information are detailed by source but we focused our attention only on the variable *total income*. For each municipality, the available information about total income is referred to the following eight classes: $(-\infty, 0]$, $(0, 10000]$, $(10000, 15000]$, $(15000, 26000]$, $(26000, 55000]$, $(55000, 75000]$, $(75000, 120000]$, $(120000, \infty)$; the frequency of the taxpayers and the total amount possessed in each class are provided. Hence, up to eight observations for each municipality[3] - the average income of each class with an attached weight given by its frequency - were available and we grouped them on provincial, regional and territorial (NUTS 1) base obtaining three different areal-unit partitions. We notice that this kind of data clearly provides much more information than only considering the per capita income in each group.

In the previous section the correlation values were calculated simulating from independent scenarios. Here we complement the analysis evaluating the components from the three considered decompositions and the discussed benchmarks in each of the 18 years, and calculating their correlations over time. Despite the differences in the derivation, the results reported in Table 2

		$cor(G_w, W_r)$			$cor(G_b, B_r)$		
Aggregation level	K	A	YL	BM	A	YL	BM
Provinces	107	.968	1	.887	.998	.630	.768
Regions	20	.988	1	.963	.993	.778	.880
Territorials	5	.970	1	.987	.997	.677	.781

Table 2: Correlations between the components from the three considered decompositions and the discussed benchmarks over time. The results are evaluated over three different aggregation levels.

confirm the very high correlations between the benchmarks and the proposed components. All the reported correlation values are definitely compatible with the findings in Table 1 and strengthen the consistency of the conclusions driven by the simulation analysis: the two components are appropriate to measure the within and the between inequality.
In Figure 6 we investigate the consequences of the decomposition choice on the between component share trajectory, with consideration to the provincial based aggregation. The three time series range in different intervals. Hence, we have normalised them to their starting values to better underline their relative evolution. The between component share from the proposed decomposition presents an initial marked decreasing path followed by an inversion started during the years in which the financial crisis affected Italy. The trajectories from the other decompositions appear quite similar in their shapes until the financial crisis years, then the component from Bhattacharya and Mahalanobis (1967) moves more similarly to ours. However, they have both varied irregularly during the first ten years and did not unambigously capture the decreasing path followed by the introduced between component. This suggests that the variation of the group means may not be always able to exhaustively inform about the overall differences between groups and the spatial patterns in the income distribution, as guessed in Section 2. Instead, the introduced between component consider additional information to better assess this phenomenon.

We now present further advantages from our proposal that arise in this simple descriptive context. These are general traits which are discernible whenever a two-component decomposition is employed. Indeed, every two-component decomposition of an inequality index allows for the considerations which will follow, but their reliability depends on both the appropriateness of the components and the index involved. As for the former, we have already appropriately justified

[3] They are less if the municipality possesses some classes containing less than four units

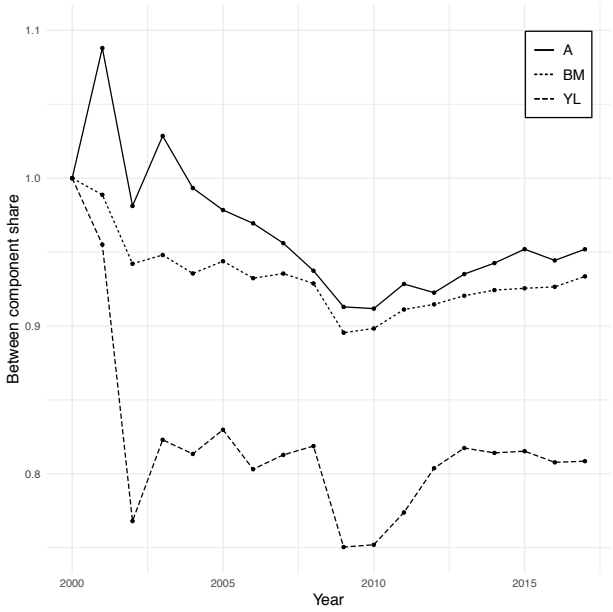


Figure 6: Time series of the share of the between components from three different Gini index decompositions - A, YL and BM stand for the decomposition proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967), respectively - over the Gini index. The series are normalised to their initial values.

both the components. As for the index to decompose, the Gini index is the most used inequality measure; and we have now the opportunity to decompose it in two components while considering a population partitioned by subgroups.

We refer again to the data already presented and we still consider the provincial based aggregation. Thus, the Gini index measures the inequality in the municipal per-class income distribution; and the between and the within component share represent the contributions, respectively, of the inter and intra-provinces diversity to the overall inequality. The values of the share of the proposed between component on the Gini index are plotted in Figure 7 - right scale. The path of the within component share can be easily derived by a reflection and a one unit long vertical translation, because the two components complement to the overall Gini. The Gini index is reported too, but on the left scale. The share of the components seems to be independent from the overall value of the index. Indeed, the latter varies quite irregularly during the considered period, while S_b shows an initial marked decreasing trend followed by a light recovery; the converse holds for the within component share. As it is evident, we can deliver a lot of information about the inequality and its subgroup distribution by a simple plot.

It is also possible to inform on the contributions which the two components provide to the Gini index percentage variation. The Gini index yearly percentage variations are reported in Figure 8 along with the relative contributions of the change in the between component. Despite, as shown in Figure 7, for this aggregation level the between component has a minority share over the Gini index, its contribution on the Gini index path has been relevant. In the first years of the period under consideration the changes in the between inequality have mainly acted restraining the effects of the variations of the within inequality on the Gini index path. Conversely, the two spatial components have affected the overall inequality in the same direction since 2010.

From the results shown for the between component, conclusions on the within inequality can also be easily evinced. This is the main advantage of having a two highly informative components decomposition in empirical contexts.

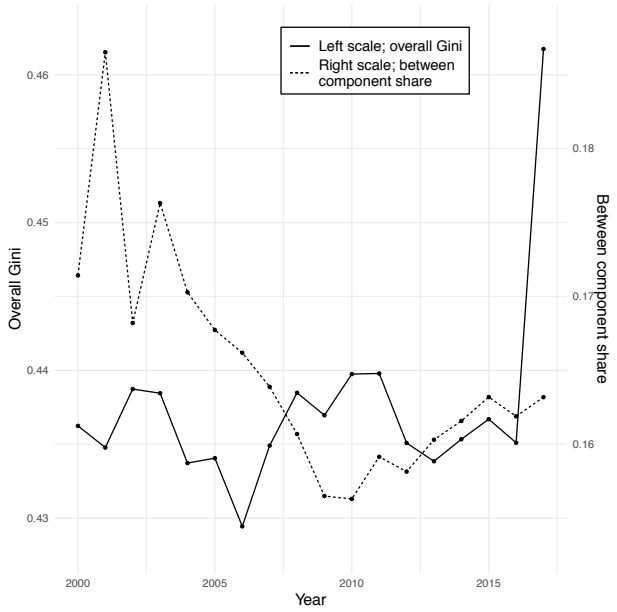


Figure 7: Time series of the share of the proposed between component over the Gini index - left scale. Gini index trajectory of the municipal per-classes income distribution - right scale.

7 Conclusions

Many spatial concentration measures have been developed in the literature. A recent line of research convincingly considers the Gini index as a source from which to derive an effective spatial concentration measure. A certain spatial decomposition of the inequality index has to be employed for this purpose.
The most common spatial Gini index decomposition have been considered in this work. They consist of three components: a residual term augment the sum of the two spatial components, which measure the within and the between group inequality. When the groups do not overlap this term generates some drawbacks w.r.t. our aim of assessing spatial concentration, as we explained in Section 2, and represents a cost in a descriptive context or a regression task in terms of both interpretability and parsimony. Furthermore, the between components of the considered decompositions are designed so that groups with similar means present low levels of territorial inequality regardless of the potential different dispersion within each group. This is an oversimplification that should be avoided if the distance in the overall group distributions is of interest and the groups overlap.

In order to overcome these two potential drawbacks, in Section 3 and Section 4 we have presented a Gini index decomposition for a population partitioned in groups. It is only composed by the two spatial components. As far as we know this is the first Gini index spatial decomposition proposal for partitioned populations to avoid the residual term. In addition, the between component relies directly on the individual incomes, hence the overall group income distributions are considered while accounting for territorial inequality.
We have demonstrated that the two components have a correct starting point and move in the right direction and proportion with reference to different scenarios. Indeed, they strongly and positively correlate with two related benchmarks developed to measure the within and the between inequality following an axiomatic approach. In other words, the components approximately provide all the information contained in the benchmarks and observe their axiomatically derived properties, despite the fact they are derived with a decomposition boundary - and not independently as for the benchmarks. Contrariwise, the decomposition boundary strengthens the informativeness of the components. It ensures their collinearity with the Gini index and consequently provides rele-

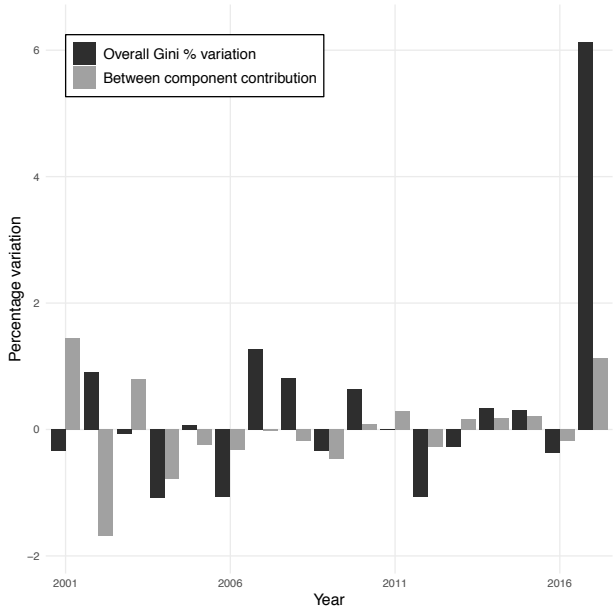


Figure 8: Gini index yearly percentage variation and contribution of the change of the between component.

vant advantages in a descriptive context or a regression task in terms of both interpretability and parsimony.

The correlation values we refer to were calculated in Section 5 by simulations. Many contexts have been considered by varying the simulation parameters and results appear to be robust. In Section 6 we strengthened the reliability of the simulation algorithm and of the correlation estimates by considering the municipality based *Income and principal Irpef variables* statistical data available on the Open-Source Data released by the Department of Finance of the Italian Ministry of Economy and Finance. In addition, an illustration of the advantages ensured by the proposed decomposition - in terms of informativeness and interpretability in a spatial empirical context - is provided.

References

Allison, Paul D. “Measures of inequality”. In: *American sociological review* (1978), pp. 865–880.

Arbia, Giuseppe. “The role of spatial effects in the empirical analysis of regional concentration”. In: *Journal of Geographical systems* 3.3 (2001), pp. 271–281.

Arbia, Giuseppe and Gianfranco Piras. “A new class of spatial concentration measures”. In: *Computational Statistics & Data Analysis* 53.12 (2009), pp. 4471–4481.

Bandourian, Ripsy, James McDonald, and Robert S Turley. “A comparison of parametric models of income distribution across countries and over time”. In: *Luxembourg income study working paper* (2002).

Bhattacharya, Nath and B Mahalanobis. “Regional disparities in household consumption in India”. In: *Journal of the American Statistical Association* 62.317 (1967), pp. 143–161.

Bickenbach, Frank and Eckhardt Bode. “Disproportionality measures of concentration, specialization, and localization”. In: *International Regional Science Review* 31.4 (2008), pp. 359–388.

Bourguignon, Francois. “Decomposable income inequality measures”. In: *Econometrica: Journal of the Econometric Society* (1979), pp. 901–920.

Campante, Filipe R and Quoc-Anh Do. “Inequality, Redistribution, and Population”. In: (2007).

Dagum, Camilo. “A new approach to the decomposition of the Gini income inequality ratio”. In: *Empirical Economics* (1997), pp. 515–531.

— “Inequality measures between income distributions with applications”. In: *Econometrica (pre-1986)* 48.7 (1980), p. 1791.

Ebert, Udo. “Measures of distance between income distributions”. In: *Journal of Economic Theory* 32.2 (1984), pp. 266–274.

— “The decomposition of inequality reconsidered: Weakly decomposable measures”. In: *Mathematical Social Sciences* 60.2 (2010), pp. 94–103.

Ellison, Glenn and Edward L Glaeser. “Geographic concentration in US manufacturing industries: a dartboard approach”. In: *Journal of political economy* 105.5 (1997), pp. 889–927.

Hyndman, Rob J and Yanan Fan. “Sample quantiles in statistica packages”. In: *The American Statistician* 50.4 (1996), pp. 361–365.

Kline, Patrick and Enrico Moretti. “People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs”. In: *Annual Review of Economics* 6.1 (2014), pp. 629–662.

Lambert, Peter J and J Richard Aronson. “Inequality decomposition analysis and the Gini coefficient revisited”. In: *The Economic Journal* 103.420 (1993), pp. 1221–1227.

Mookherjee, Dilip and Anthony Shorrocks. “A decomposition analysis of the trend in UK income inequality”. In: *The Economic Journal* 92.368 (1982), pp. 886–902.

Moran, Patrick AP. “Notes on continuous stochastic phenomena”. In: *Biometrika* 37.1/2 (1950), pp. 17–23.

Neumark, David and Helen Simpson. “Place-based policies”. In: *Handbook of regional and urban economics*. Vol. 5. Elsevier, 2015, pp. 1197–1287.

Ord, J Keith and Arthur Getis. “Local spatial autocorrelation statistics: distributional issues and an application”. In: *Geographical analysis* 27.4 (1995), pp. 286–306.

Pyatt, Graham. “On the interpretation and disaggregation of Gini coefficients”. In: *The Economic Journal* 86.342 (1976), pp. 243–255.

Radaelli, Paolo. “On the decomposition by subgroups of the Gini index and Zenga’s uniformity and inequality indexes”. In: *International Statistical Review* 78.1 (2010), pp. 81–101.

Rao, VM. “Two decompositions of concentration ratio”. In: *Journal of the Royal Statistical Society. Series A (General)* 132.3 (1969), pp. 418–425.

Rey, Sergio J and Richard J Smith. “A spatial decomposition of the Gini coefficient”. In: *Letters in Spatial and Resource Sciences* 6.2 (2013), pp. 55–70.

Rodríguez-Pose, Andrés. “The revenge of the places that don’t matter (and what to do about it)”. In: *Cambridge Journal of Regions, Economy and Society* 11.1 (2018), pp. 189–209.

Shorrocks, Anthony F. “Inequality decomposition by population subgroups”. In: *Econometrica: Journal of the Econometric Society* (1984), pp. 1369–1385.

— “On the distance between income distributions”. In: *Econometrica: Journal of the Econometric Society* (1982), pp. 1337–1339.

— “The class of additively decomposable inequality measures”. In: *Econometrica: Journal of the Econometric Society* (1980), pp. 613–625.

Shorrocks, Anthony and Guanghua Wan. “Spatial decomposition of inequality”. In: *Journal of Economic Geography* 5.1 (2005), pp. 59–81.

Yitzhaki, Shlomo. “Economic distance and overlapping of distributions”. In: *Journal of Econometrics* 61.1 (1994), pp. 147–159.

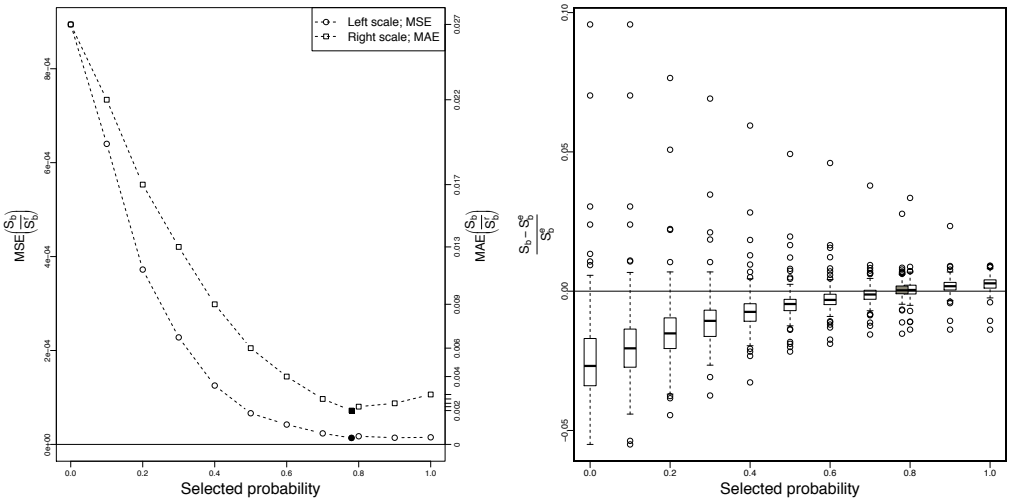
Yitzhaki, Shlomo and Robert I Lerman. “Income stratification and income inequality”. In: *Review of income and wealth* 37.3 (1991), pp. 313–329.

Appendix

On the approximation due to the *quantilisation* procedure

The first section of this appendix is due in order to assess the *quantilisation* procedure, quantify the magnitude of the approximation incurred and identify the optimal definition of quantile and the value of n to be selected.

Recall that, defining $\mathbf{w} = (w_1, \dots w_K)$, we suggested the value $n = \mathbf{w}\mathbf{n}^\intercal$.



(a) MSE (left scale) and MAE (right scale) of the 150 relative differences in S_b for different choices of n.

(b) Boxplots of the 150 relative differences in S_b for different choices of n.

Figure 9: Between component share relative approximation for different choices of n. The approximation is evaluated considering the 150 values of the relative difference in the between share obtained by the *quantilisation* procedure w.r.t. the one obtainable employing the exact approach.

This expression determines n as the average of the n_k, each weighted by the related share of population w_k. The performance of this value are firstly shown in Figure 9, where approximation is evaluated looking at the relative discrepancy between the two values of $\frac{G_b}{G}$ obtained employing the *exact* or the *quantilisation* method. More precisely, let $S_b = \frac{G_b}{G}$ the between component share obtained by the *quantilisation* method and $S_b^e = \frac{G_b^e}{G^e}$ the same share in the *exact* approach. The relative discrepancy is measured by the Mean and the Absolute Squared Errors of $\frac{S_b}{S_b^e}$ w.r.t. $1 = \frac{S_b^e}{S_b^e}$ obtained running 150 simulations - $\text{MSE}\left(\frac{S_b}{S_b^e}\right) = \mathbb{E}\left[\left(\frac{S_b}{S_b^e} - 1\right)^2\right]$ and $\text{MAE}\left(\frac{S_b}{S_b^e}\right) = \mathbb{E}\left[\left|\frac{S_b}{S_b^e} - 1\right|\right]$. In other words in each simulation, for different choices of n, we looked at the relative difference of the between share obtained by the *quantilisation* procedure w.r.t. the one obtainable employing the exact approach. The results are summarised by averaging the square or the absolute errors.
The simulation procedure flows as follow. In each running lognormal-distributed incomes with a vector of sizes $\mathbf{n}$ are drawn as described in the second section of this Appendix. The different choices of n are its minimum value, its deciles and the value obtained by eq. (12). The vector $\mathbf{n}$ is also drawn and we imposed some constraint on its elements to ensure affordable values for the *mcm* when employing the *exact* approach. To be specific, we firstly specified K (= 5, 10 or 20). Then we built a vector **mul** composed by the divisors of $2^4 3^3 5$ belonging to an interval $[min, max]$. The *min* (36 or 72) and the *max* (360 or 720) were both included in $\mathbf{n}$. The other

$K-2$ values were sampled with repetition from **mul**. With this choice the *mcm* can not exceed the value 2160 and the computations are affordable. Figure 9 represents the results for $K = 20$, $min = 72$ and $max = 720$. As shown in Figure 9a the proposed value of n, represented by the solid point, always minimizes (or reach a value very close to the minimum of) the approximation this method copes with, both for the MSE (left scale) and the MAE (right scale). The minimum approximation by the proposed value of n is achieved both thank to a vanished distortion and the variance reduction, as Figure 9b shows. We also stress the irrelevance of the approximation when that value of n is employed: the correspondent MAE measures for S_b a mean absolute percentage error of the 0.22%.
Obviously, the magnitude of the percentage between component share approximation due to the

		Probability associated to the deciles											$n = \mathbf{wn}^\intercal$	
K	$[min, max]$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	n	$\bar{p}$
5	[36,360]	6.12	4.88	4.14	3.20	2.53	2.17	1.42	1.20	1.00	0.92	0.90	0.94	0.78
	[36,720]	8.28	6.41	5.45	4.13	3.26	2.79	1.71	1.44	1.05	0.91	0.86	0.91	0.83
	[72,360]	2.66	2.19	1.93	1.52	1.18	1.01	0.76	0.69	0.61	0.57	0.58	0.60	0.73
	[72,720]	4.07	3.24	2.79	2.14	1.70	1.45	1.00	0.86	0.69	0.61	0.59	0.64	0.79
10	[36,360]	4.72	3.90	2.97	2.23	1.61	1.14	0.86	0.65	0.55	0.56	0.60	0.53	0.77
	[36,720]	5.44	4.42	3.35	2.56	1.94	1.43	0.99	0.69	0.52	0.45	0.47	0.45	0.84
	[72,360]	2.11	1.76	1.41	1.11	0.87	0.65	0.49	0.38	0.34	0.33	0.35	0.35	0.71
	[72,720]	3.17	2.66	2.13	1.63	1.20	0.85	0.61	0.47	0.38	0.32	0.32	0.34	0.78
20	[36,360]	3.72	2.94	2.26	1.77	1.26	0.83	0.54	0.37	0.31	0.36	0.42	0.30	0.75
	[36,720]	4.85	3.81	2.93	2.20	1.55	1.08	0.67	0.43	0.29	0.25	0.30	0.24	0.82
	[72,360]	1.78	1.53	1.17	0.89	0.64	0.44	0.29	0.22	0.21	0.23	0.26	0.21	0.69
	[72,720]	2.68	2.20	1.66	1.26	0.89	0.61	0.43	0.29	0.24	0.26	0.32	0.22	0.78

Table 3: Percentage between component share approximation generated by the *quantilisation* procedure. It is evaluated by the algorithm described in this section for different choices of n, K and of the interval $[min, max]$. The approximation is measured by the MAE. The last column represents the average fraction of elements in the vector $\mathbf{n}$ which are less than the suggested n.

quantilisation procedure depends on the simulation parameters, as Table 3 points out. It reports the MAE - already multiplied by 10^2 - of the between component share obtained by the procedure designed in this section for different choices of n, K and of the interval $[min, max]$. We have chosen it because of its interpretability as average absolute percentage error.
Results are really encouraging. The values of the MAE are below the percentage point approximately in the half of the analysed contexts and always when the suggested choice of n is employed. In addition, the dependence on the employed parameters - which is described just below - could further ensure a reduction in the approximation in many realistic contexts where presumably the parameters are more conducive.
The way the MAE reacts to choices of n, K and of the interval $[min, max]$ should not come as a surprise. For each choice of n, when the ratio $\frac{max}{min}$ stays constant, the MAE informs about better performances for higher *min* and *max*. If that ratio increases - i.e. if the variability of $\mathbf{n}$ increases - the approximation raises, too. Results are also enhanced when the number of groups increases.
As desired, the MAE reaches a quasi-minimum value when n is selected by eq. (12), especially when K is high. Furthermore, the suggested choice guarantees an almost always relevant reduction in the computational cost which the procedure would incur in choosing $n = max(\mathbf{n})$. This reduction is not negligible in our simulations: $\bar{p}$ is the average of the probabilities corresponding to the values of n selected by eq. (12) in the 150 simulations. It is reported in the last column of the table. Its values range from 0.69 to 0.84 and a clear dependence from the distribution of $\mathbf{n}$ is highlighted in the table.
In addition we point out - as supported by the values in the third column of the table, which decrease when $min(\mathbf{n})$ increase - that it could be also acceptable to choose a value $n << min(\mathbf{n})$ if $min(\mathbf{n})$ is high and a computational cost saving choice is required.

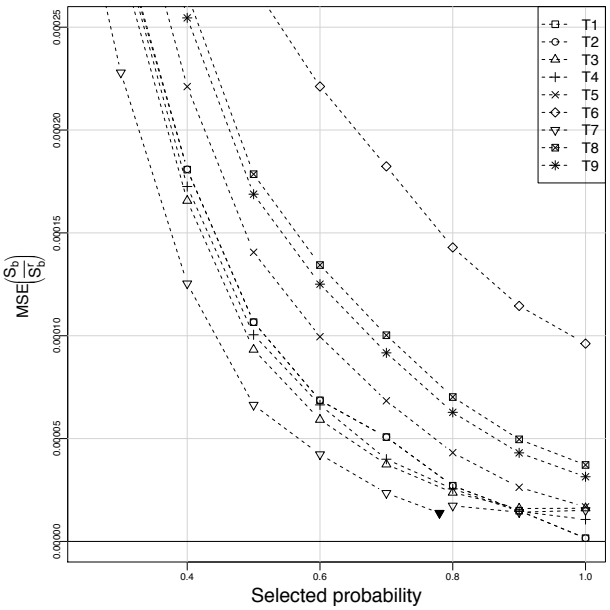


Figure 11: Between component share approximations - measured by the MSE and through the same procedure which produced Figure 9 - obtained employing the 9 quantile definitions presented in Hyndman and Fan (1996). In software R, each definition can be selected by the option *type* of the function *quantile()*. Here, T_j: $j = 1,\ldots,9$ stands for selecting the option $type = j$.

We now present the assessment procedure which leads to the selection of the quantile definition employed in the analysis. We evaluated the impact of several quantile definitions on the approximation the *quantilisation* procedure copes with. Figure 11 shows differences between results achieved iterating the same procedure which generates Figure 9a using the 9 different quantile definitions presented in Hyndman and Fan (1996). Definition 7, which is also the default definition adopted by the *quantile*() function in the software R, presents both the lowest MSE and MAE substantially for each choice of n and ensures computational advantages because it approaches 0 for smaller n.
The better performances resulting from the definitions 1 and 2 when the selected probability approaches 1 are exceptions. Both the definitions rely on a stepwise cumulative probability function which estimates the quantiles only allowing for the values present in the starting vector. Thus, if $p = 1$ and $max(\mathbf{n}) = mcm(\mathbf{n})$ the vector of the quantiles corresponds to the $\mathbf{y}^k$ of the *exact* approach and no approximation is encountered. Positive effects are also encountered for $max(\mathbf{n})$ close to $mcm(\mathbf{n})$ and p approaching 1, i.e. n close to $max(\mathbf{n})$. In Figure 11 they are evident from $p = 0.8$.
Nonetheless, the advantage in terms of approximation is quite negligible. Most importantly, in the vast majority of the real applications the vector $\mathbf{n}$ is much more variable than the bounded vectors used in these simulations. Hence $mcm(\mathbf{n})$ is generally far from $max(\mathbf{n})$. In conclusion, the definition 7 is definitely recommended unless $max(\mathbf{n})$ is close to $mcm(\mathbf{n})$ and the computational burden is not an issue.

The income simulation algorithm

A parametric bootstrap algorithm has been employed to evaluate both the extent of the approximation of the *quantilisation* procedure and the correlations between the two components of the introduced decomposition and the related benchmarks. This section of the appendix provides with the theoretical foundations of the income simulation procedure which feeds both the algorithms.

The distribution of $\mathbf{n}$ is a K-variate uniform. The number of groups K and the extremes of the distribution were determined *ex-ante*.

The uniform distribution was also exploited to draw the expected average income of each group: $\mathbb{E}\left[\bar{x}^k\right] \sim Unif(m,M)$.

The minimum m of this distribution was set to 10^4. As for the maximum M, it was fixed to $5 \cdot 10^4$ in the simulations which generate the results in Section 4. In Section 5 it was varied to highlight the impact of the variability in the vector of the means of the groups on the values of interest, i.e. on the correlations between the two components of the introduced decomposition and the related benchmarks.
Indeed, a modification of M directly affects $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$. For the uniform distribution $\mathbb{E}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right] = \frac{1}{2}(M+m)$ and $\mathrm{Var}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right] = \frac{1}{12}(M-m)^2$, the coefficient of variation of $\mathbb{E}\left[\bar{x}^k\right]$ is:

$$CV\left[\mathbb{E}\left[\bar{x}^k\right]\right] = \frac{\sqrt{\mathbb{V}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right]}}{\mathbb{E}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right]} = \frac{1}{\sqrt{3}}\frac{(M-m)}{(M+m)} \in \left[0, \frac{1}{\sqrt{3}}\right]$$

and, with m fixed, only depends on the value of M. In Figure 4 the interval $\left[0, \frac{1}{\sqrt{3}}\right]$ and the values of $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ are normalised to $[0,1]$ by a simple scale transformation. Notice that the values $CV\left[\mathbb{E}\left[\bar{x}^k\right]\right]$ are not comparable with the ones dealing with a $\mathbb{E}\left[\bar{x}^k\right]$ from an other distribution, which can be potentially unbounded. Any comparison would be meaningless, hence the scale transformation is not an issue.
Denote by $M^{(s)}$, $s = 1 \dots 8$ the $S = 8$ different values of the *maxima* required to produce the eight values of the coefficient of variation. The *maxima* were selected so that the values of the coefficient of variation divided the interval in eight equal parts. Hence, the values $M^{(s)}$ satisfy:

$$\frac{M^{(s)} - m}{M^{(s)} + m} - \frac{M^{(s-1)} - m}{M^{(s-1)} + m} = c$$

with $M^{(1)} = m$ and $c = \frac{1}{\sqrt{3}S}$. The following holds:

$$\begin{aligned}
&(M^{(s)} - m)(M^{(s-1)} + m) - (M^{(s-1)} - m)(M^{(s)} + m) = c(M^{(s)} + m)(M^{(s-1)} + m) \Rightarrow \\
&\Rightarrow 2mM^{(s)} - 2mM^{(s-1)} = cM^{(s)}M^{(s-1)} + cmM^{(s-1)} + cmM^{(s-1)} + cm^2 \Rightarrow \\
&\Rightarrow 2mM^{(s)} - cM^{(s)}M^{(s-1)} - cmM^{(s)} = cmM^{(s-1)} + cm^2 + 2mM^{(s-1)} \Rightarrow \\
&\Rightarrow M^{(s)}(2m - cM^{(s-1)} - cm) = m(cM^{(s-1)} + cm + 2M^{(s-1)}) \Rightarrow \\
&\Rightarrow M^{(s)} = \frac{m(cM^{(s-1)} + cm + 2M^{(s-1)})}{(2m - cM^{(s-1)} - cm)}
\end{aligned}$$

and the $M^{(s)}$ can be calculated iteratively.

Fixed all the parameters, the incomes of each group k are drawn from a log-normal distribution with $\mathbb{E}\left[\bar{x}^k\right]$ as expected value. The last requirement is to define a reasonable way to accordingly determine the two parameters μ and σ^2 of the distribution.
As it is well known, for a log-normal distribution the following holds:

$$\mathbb{E}\left[\bar{x}^k\right] = e^{\mu_k + \frac{\sigma_k^2}{2}} \tag{13}$$

This equation allows to design an effective way to split $\mathbb{E}\left[\bar{x}^k\right]$ in the two seminal elements μ_k e σ_k - required to draw from the distribution - in a manner that the log-normal distribution is likely to be an income distribution. Starting from eq. (13) it is possible to write:

$$\ln \mathbb{E}\left[\bar{x}^k\right] = \mu_k + \frac{\sigma_k^2}{2}$$

and to split linearly $\ln \mathbb{E}\left[\bar{x}^k\right]$ in μ_k and σ_k^2:

$$\mu_k = \alpha_k \ln \mathbb{E}\left[\bar{x}^k\right] \tag{14}$$

$$\sigma_k^2 = 2\,(1-\alpha_k) \ln \mathbb{E}\left[\bar{x}^k\right] \tag{15}$$

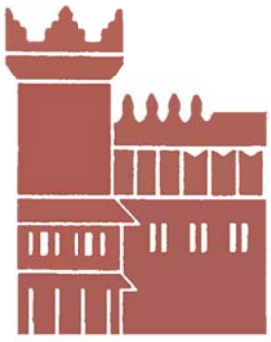
Their ratio is:

$$\frac{\sigma_k^2}{\mu_k} = \frac{2\,(1-\alpha_k) \ln \mathbb{E}\left[\bar{x}^k\right]}{\alpha_k \ln \mathbb{E}\left[\bar{x}^k\right]} = \frac{2(1-\alpha_k)}{\alpha_k}$$

At this point, the 82 couples of log-normal parameters estimated in Bandourian, McDonald, and Turley (2002) using real data along different countries and periods have been considered. For each couple the correspondent $c_i = \frac{\sigma_i^2}{\mu_i}$, $i = 1, \ldots, 82$ has been evaluated.
Verisimilar values for α_k can be obtained sampling a value of i for each group and using the correspondent c_i to solve the following equation:

$$c_i = \frac{\sigma_k^2}{\mu_k} = \frac{2(1-\alpha_k)}{\alpha_k} \implies \alpha_k = \frac{2}{c_i + 2} \tag{16}$$

Thus μ_k and σ_k^2 were determined - taking $\mathbb{E}[\bar{x}^k]$ as known - from the equations (14)-(16).
The appropriateness of the last step - i.e. sampling a value of i for each group and using the correspondent c_i - is justified by the fact that the 82 values of α in Bandourian, McDonald, and Turley (2002) seems not to be influenced by the associated $\mathbb{E}[\bar{x}^k]$. A simple linear regression reports for the regressor $\mathbb{E}[\bar{x}^k]$ a very low coefficient and a large p-value (0.65).
Substantially, 82 possible proportions to split $\mathbb{E}[\bar{x}^k]$ in a likely way in its two addends μ_k and $\frac{\sigma_k^2}{2}$ are available.



Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it