ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

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

Rev. ed.

Federico Attili

Quaderni - Working Paper DSE N°1167

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

Federico Attili

Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy

E-mail: federico.attili2@unibo.it

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

Abstract

This paper considers a partitioned population and develops a decomposition of the Gini index in two components, which measure the within and the between groups inequality. Differently from the most widespread inequality measure decompositions, having a between component that compares the means of the groups, ours informs about the distance between their entire distributions. This makes the decomposition helpful in several frameworks, such as in the measurement of spatial concentration A Monte Carlo experiment supports the appropriateness of our components highlighting that they strongly correlate with two axiomatically derived benchmarks. The presentation of a case study concerning the income distribution in the Italian provinces concludes the work and stresses the informativeness of the proposed decomposition.

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

Non technical summary

In this paper we consider the Gini index and propose a within-between decomposition that nests considerable information on both inequality and its spatial distribution. The decomposition proposal deals with partitioned population and only consists of the two intra and inter groups inequality components.

All the well known decompositions of the Gini index possess three components, and the between component 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. The availability of a decomposition composed by only-two highly informative terms also provides relevant advantages in a descriptive context or a regression task, in terms of both interpretability and parsimony.

We demonstrate that both our components are highly informative because - as we support by employing a Monte Carlo procedure - they strongly and positively correlate with two benchmarks. The two benchmarks which have been considered 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 Italian municipality based *Income and principal Irpef variables* statistical data. In the Monte Carlo procedure, correlation values are calculated simulating from independent scenarios. In the real data analysis we complement the Monte Carlo results calculating correlations over time - our analysis ranges from 2000 to 2017 - and over different territorial aggregation. This strengthens the evidence on the informativeness of the components.

With the same data - focusing on the income distribution of Italian provinces - we highlight the advantages of the proposed between component share in assessing spatial distribution of inequality and we discuss 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

The presence of severe inequality and territorial disparities in several countries facilitates the emergence of challenging socioeconomic issues (see e.g. Rodríguez-Pose, 2018). In the last two decades an increasing effort has been paid to the study of between territories inequality and its socioeconomic effects. Driven by the attention in 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.

This paper develops a novel within-between decomposition of the Gini index and presents its numerous advantages. Focusing on geographical partitions, we argue that our proposal is well-suited to measure spatial concentration. In particular, it improves upon several decompositions currently employed by researchers for the same purpose, solving two critical issues that those methods may suffer of, and that we refer to as the *oversimplification* and the *overestimation* issue.

Decomposing an inequality index to measure spatial concentration is a meritorious idea introduced by Shorrocks and Wan (2005). They discuss the existing *subgroup* decompositions of inequality measures, and suggest to employ them in the spatial context. This means partitioning the population into geographical regions, and decomposing an inequality index to obtain a *within* and a *between* component measuring, respectively, the intra- and the inter-territories inequality. The idea is to assess spatial concentration by jointly considering the inequality index and its between component. In this manner, both inequality and its spatial distribution - the key features of spatial concentration - should be under control. In the conception of Shorrocks and Wan, and in accordance to the conventional literature about inequality decomposition, the between component has to compare the means of the regions, resulting to be zero if the means are the same. As Ebert (2010) effectively points out, restricting the attention to the first moment of the distributions constitute a serious *oversimplification* issue, indeed it could associate the same value of between inequality to alternative situations despite different skewness of the distributions. More recently, Rey and Smith (2013) have introduced a within-between decomposition of the Gini index arguing that it nests sufficient information on both inequality and its spatial distribution. In particular, they decompose the Gini index according to a matrix defining pairs of neighbours and non-neighbours. The differences among pairs of neighbours constitute the within component, while the others sum up to the between term. Evidently, their between component does not depend on the means of the groups but on the pairwise differences among non-neighbours, thus solving the oversimplification issue. However, it is not suited to deal with mutually exclusive groups, such as geographical partitions. When it is the case, their between component overestimates between groups inequality. As a clear example of the *overestimation* issue, their between component is positive even if the groups have the same distribution.

Our proposal overcomes both the discussed issues thanks to important properties. First, our between component is zero if and only if the groups have the same distribution. This property is a necessary condition to avoid both the oversimplification and the overestimation issues. In addition, both our within and between components are strongly correlated with two benchmarks, which are inspired by literature and axiomatically derived to measure within and between inequality. Hence, our between component measures between groups inequality as its benchmark does, namely looking at the distance between the entire distributions of the groups. Recommending to measure spatial concentration by the Gini index and our between component comes naturally at this point, and enhances the proposal of Shorrocks and Wan thanks to the advantages of our decomposition.

In fact, this paper considers income and draws inspiration from the spatial framework: the groups of the partition could be regions in a country, or countries in a confederation, and we often refer to the within and the between components as to the spatial components. Nonetheless, we stress that our decomposition is meaningful and convenient in several applications outside of this context. To give some examples, the focus may be in well-being, occupation or export rate, and groups could be defined by factors such as: gender, occupation, race, age, education level or other criteria for individuals; industrial sector or other relevant dimension for firms.

The paper unfolds as follows. Section 2 introduces the decomposition methodology, presenting the decomposition rationale and its formalisation for a population partitioned in equal-sized groups. Section 3 explores the properties and the advantages of our decomposition, and discusses the possibility of our two-component decomposition to exist. Section 4 generalises our proposal to the different-sized groups case. In Section 5 a Monte Carlo algorithm proves the informativeness of the components showing that they are strongly correlated with the two benchmarks. Section 6 shows that the same levels of correlation hold with real data, strengthening the Monte Carlo results. The same data also highlight the advantages of our decomposition in assessing spatial distribution of inequality, and more generally emphasise the interpretative benefits that a two-component decomposition ensures in empirical analysis. Section 7 gives conclusive remarks.

2 The decomposition proposal

Consider a population of N individuals. We denote by x_i the income of the generic individual $i = 1,\dots,N$ and by $\mu = \sum_1^N x_i/N$ the average income in the population. Among the many different formulations of the Gini index (see Ceriani and Verme, 2015 and Ceriani and Verme, 2012), we consider the following:

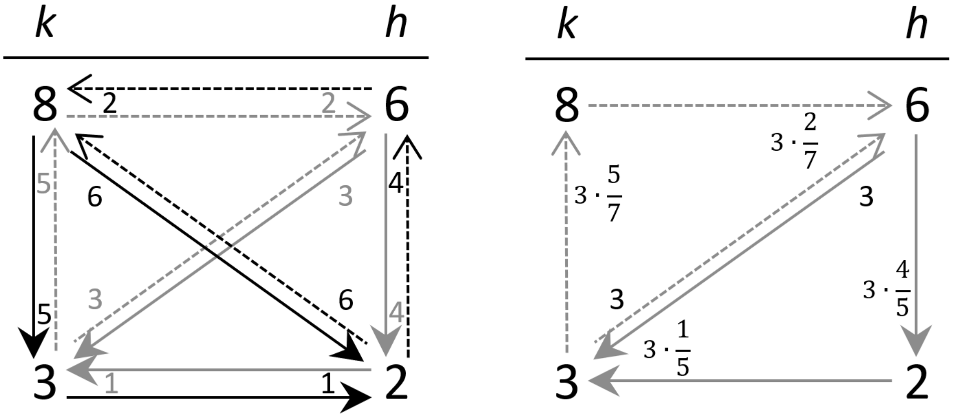
$$G = \frac{1}{2\mu N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| = \frac{g}{2\mu N^2} \tag{1}$$

The numerator g is the sum of all the pairwise absolute[1] differences between individual incomes. It is standardised by the factor $(2\mu N^2)^{-1}$, so that G is scale invariant and $G \in [0,1]$ if all the $x_i \geq 0$.

Consider the population as partitioned in K equal-sized groups; define n to be their size and x_i^k to be the i-th element in the group k ordered (descending) vector of incomes $\mathbf{x}^k = (x_1^k, \ldots, x_n^k)$. All the information concerning with the spatial distribution of inequality is within g, which can be written as:

$$g = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| = \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i^k - x_j^h| \tag{2}$$

A new insight. The assumption of equal-sized groups might appear as extremely simplistic but, as we show in Section 4, it is not a limit in the applicability of our proposal. Conversely, it provides an innovative insight into the structure of the Gini index. Look at Figure 1, which illustrates a two-group-two-individual situation. Figure 1a highlights all the pairwise differences between units, considered twice so that they constitute g if they are summed up. As the scheme suggests, with equal-sized groups we can distinguish three kind of differences: vertical, horizontal and diagonal ones. The vertical differences involve same-group pairs of elements. The horizontal differences involve same-rank pairs from different groups. The diagonal differences involve different-rank pairs from different groups.



(a) Pairwise differences composing g

(b) Decomposition in the non-trivial case

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

As an intuitive point of departure, we address the vertical and the horizontal differences to the within and the between component, respectively. The diagonal differences involve different-group pairs. Despite this, they partly reflect the vertical (same-group) differences and are not entirely addressable to the

[1] If not differently specified, in the remainder we always refer to absolute differences.

between groups inequality. For example, imagine to replace the values in the scheme so that the groups are identical: pose $x_1^h = x_1^k = 8$ and $x_2^h = x_2^k = 3$. The values of the diagonal differences - 5 - equal the vertical ones and should not contribute to the (absent) between groups inequality at all.

At this stage, the diagonal differences may be instinctively thought as the addenda of a residual term arising from the decomposition. However, one effective paradigm exists to disentangle from this *residual* two informative contributions to the within and the between components.

The decomposition paradigm. The two black diagonals in Figure 1a are decomposable with a straightforward strategy. For example, looking at the solid black diagonal line and moving along the legs of the solid black triangle in the scheme: the difference between the richest of group k and the poorest of group h is 6 since the former is 5 units richer than her group poorest individual, who is 1 unit richer than her counterpart in group h ($6 = 5 + 1$). A similar argument holds from the opposite point of view, which is looking at the dashed black diagonal line representing the difference between the poorest of group h and the richest of group k ($6 = 4 + 2$). The two black diagonal differences are predominantly due to and reflect the within inequality of the two groups. Accordingly, we suggest to split their contribution to g ($6 + 6 = 12$) assigning $5 + 4 = 9$ to the within component and $1 + 2 = 3$ to the between one.

This strategy is not viable in the other half of the diagonals, which are the focus of Figure 1b. Here, the three values involved in the path along the grey legs do not increase or decrease monotonically as for the black lines, namely the product between the horizontal and the vertical signed differences 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. However, it would be paradoxical to decrease the between component by the horizontal value, i.e. by 1 in the case of the solid grey lines[2].

As Figure 1b illustrates, to overcome this issue 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. From the other perspective, we propose to rescale both the vertical and the horizontal values to make the summation of the two contributions equal to the diagonal difference: the vertical and the horizontal values are divided by their sum and multiplied by the diagonal difference. Thanks to this solution we preserve both reasonable proportions[3] between the values added to the components and the Gini index compliance, i.e. the possibility to have a two-component decomposition. With this overall strategy, the contributions to within and between inequality from the diagonal

[2] To see the paradox, imagine to replace the poorest individual of group h with a poorer one. Subtracting $3 - (2 - \varepsilon) > 1$ would produce a lower value of the between component, though the intuition suggests that between inequality is now higher because the poor group is poorer.

[3] These proportions observe the black diagonals decomposition argument.

differences mimic to the utmost the vertical and the horizontal ones. This is the key that makes the two resulting components extremely informative.

Formalisation. We have just presented the intuition grounding the decomposition proposal. We now generalise this strategy, formalise the decomposition and deliver the two spatial components.

For each difference $|x_i^k - x_j^h| > 0$, we define $L_{ij}^{kh} = |x_i^k - x_j^k| + |x_j^k - x_j^h|$ and $c_{ij}^{kh} = (x_i^k - x_j^k)(x_j^k - x_j^h)$, where the equal-sized groups hypothesis guarantees the element x_j^k to exists. We can write[4]:

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

$$= \begin{cases} \left|x_i^k - x_j^k\right| + \left|x_j^k - x_j^h\right| & \text{if } c_{ij}^{kh} \geq 0 \\ \left|x_i^k - x_j^k\right| \frac{|x_i^k - x_j^h|}{L_{ij}^{kh}} + \left|x_j^k - x_j^h\right| \frac{|x_i^k - x_j^h|}{L_{ij}^{kh}} & \text{if } c_{ij}^{kh} < 0 \end{cases} \tag{3}$$

where to obtain the first equation we use $c_{ij}^{kh} \geq 0 \Rightarrow |x_i^k - x_j^h| = |x_i^k - x_j^k + x_j^k - x_j^h| = L_{ij}^{kh}$. The first equation fills all the trivial-case decompositions: the vertical differences ($k = h$), the horizontal differences ($i = j$) and the black diagonal differences kind ($k \neq h$, $i \neq j$ and $c_{ij}^{kh} \geq 0$). The second equation stands for situations as the two sketched by the grey diagonals ($k \neq h$, $i \neq j$ and $c_{ij}^{kh} < 0$). Following the *decomposition paradigm* we assign the first and the second addenda of the final expressions in eq. (3) to the within and the between components, respectively.

To simplify the notation in eq. (3) we define[5] $w_{ij}^{kh} = |x_i^k - x_j^h| / L_{ij}^{kh}$ and obtain:

$$\left|x_i^k - x_j^h\right| = \left|x_i^k - x_j^k\right| w_{ij}^{kh} + \left|x_j^k - x_j^h\right| w_{ij}^{kh} \tag{4}$$

As in eq. (3), $|x_i^k - x_j^k| w_{ij}^{kh}$ and $|x_j^k - x_j^h| w_{ij}^{kh}$ are interpretable, respectively, as the contributions from the difference $|x_i^k - x_j^h|$ to the within and the between groups inequality; and w_{ij}^{kh} as the vertical and the horizontal differences rescaling factor.

By definition $w_{ij}^{kh} \in [0,1]$, and $w_{ij}^{kh} = 1$ iff $c_{ij}^{kh} \geq 0$. The rescaling factor w_{ij}^{kh} can be lower than one due to the possibility of a proportional reduction of the horizontal and the vertical values: as in the grey diagonals of Figure 1, the difference $|x_i^k - x_j^h|$ can be less than $|x_i^k - x_j^k| + |x_j^k - x_j^h|$, so we have to rescale the two addenda by the factor $w_{ij}^{kh} \leq 1$ before assigning them to the spatial components. This ensures

[4] Notice that considering x_j^k or x_i^h in eq. (3) is not an issue because the Gini index counts each difference twice by inverting the indices of the summations.

[5] We set w_{ij}^{kh} to zero *a priori* if $|x_i^k - x_j^h| = 0$, i.e. when there is nothing to decompose.

that the proportion of the two contributions follows the ratio between the vertical and the horizontal differences. This is crucial for the informativeness of the components.

The Gini index decomposition follows by substituting eq. (4) into eq. (2). Denoting $\sum_{h=1}^{K} w_{ij}^{kh} = w_{ij}^{k}$ and $\sum_{i=1}^{n} w_{ij}^{kh} = w_{j}^{kh}$, we have:

$$g = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i^k - x_j^k| \, w_{ij}^k + \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{j=1}^{n} |x_j^k - x_j^h| \, w_j^{kh} = g_w + g_b \tag{5}$$

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 two terms. We propose to interpret G_w as the within component of inequality, because it depends on the contributions from the same-group pairwise differences, i.e. the vertical differences multiplied by the weights w_{ij}^k; and G_b as the between component of inequality, because it depends on the contributions from the same-rank pairwise differences, i.e. the horizontal differences multiplied by the weights w_j^{kh}.

Notice that the within and the between components involve, respectively, the weights w_{ij}^k and w_j^{kh}, which do not exclusively depend on the two individuals involved in the difference that they multiply. This feature allows each same-group (same-rank) difference to contribute to within (between) inequality according to how much it affects the diagonal ones. For example, if a vertical difference increases, and this enlarges some of the grey-like kind diagonal differences, then the related rescaling factors consistently increase and inflate the weight w_{ij}^k.

3 Properties

As it is well known, the Gini index does not observe important properties designed in the inequality decomposition literature. Thus, many researchers prefer to employ the decomposition of alternative inequality measures to obtain information about the contributions to inequality from within and between groups disparities. In this section, we firstly examine these properties and explain in what contexts we believe that they are not appropriate. Then, we present the properties and the advantages of our decomposition. They are relevant and strongly revive the motivations to employ the Gini index to decompose inequality in a within-between fashion.

The class and the properties of the *additively decomposable* and of the *path independent* inequality measures are presented in Bourguignon (1979), Shorrocks (1980), Shorrocks (1984) and Foster and Shneyerov (2000). The Gini index is not additively decomposable in the sense intended by these authors, which partition the population and require the index to be expressed as the summation of two terms: a weighted average of the inequality values within each group and a contribution arising from the variability in the means of the groups. Path independence is similar, but allows the second term to depend on a more general class of representative income functions. To decompose the Gini index obtaining two terms of this kind implies a residual component. Consequently, the Gini index is not *subgroup consistent*, neither. Indeed, even if the mean income in each group stays constant, an increase in the inequality within some groups can be accompanied, due to the residual term, by a decrease of the Gini index (Cowell, 1988). This is not allowed by subgroup consistency (Shorrocks and Wan, 2005, p. 63).

We highlight that these three central properties of the inequality decomposition literature have a common perspective: they conceive the between component as exclusively linked to the variability (among groups) of some representative income function, generally the mean. This is desirable only if there are reasons to believe that comparing the first moment of the distributions effectively informs about the distance of the groups, e.g. when the distributions have similar higher order moments or when the population is partitioned by non overlapping[6] groups. Differently, our proposal determines a between component that is explicitly dependent on the pairwise differences between individuals. Together with the *new insight* in the Gini structure, this allows the two-term decomposition to exist and to deliver a between component that carries precious information when the distributions of the groups overlap and their moments differ.

It is straightforward to verify that $w_{ij}^{kk} = 1$, $w_{jj}^{kh} = 1$ and $w_{ij}^{kh} \geq 0 \, \forall i, j, k, h$. This implies $w_{ij}^{k} \geq 1 \; \forall k, i, j$ and $w_{j}^{kh} \geq 1 \; \forall j, k, h$. Hence, the following properties hold:

$$G_w = 0 \iff |x_i^k - x_j^k| = 0 \;\; \forall \; i, j, k \tag{i}$$

$$G_b \; = 0 \iff |x_j^k - x_j^h| = 0 \;\; \forall j, k, h \tag{ii}$$

The first relation ensures that the within component is zero iff all the same-group differences are zero, i.e. all the individuals equal their group mean. The second condition guarantees that the between component is zero iff all the same-rank differences are zero, i.e. all the individuals equal their rank mean (the groups have the same distribution).

Properties (i)-(ii) are evidently symmetric. All the decompositions considered in the introduction have a within component that observes property (i), but only our between component observes property (ii).

[6] A set of distributions has no overlapping if the intervals where the distributions take values from have empty intersection.

The benefits for our between component are relevant: the sufficiency of property (ii) - groups with the same distribution imply that the between component is zero - solves the overestimation issue of the decomposition by Rey and Smith, while its necessity - the between component is zero only if groups have the same distribution - solves the oversimplification of the between components based on the means of the groups.

As stressed before, the weights w_{ij}^{k} and w_{j}^{kh} do not only depend on the two individuals involved in the difference that they multiply. Unfortunately, this compromises the mathematical tractability of the components and hinders the analytical derivation of additional properties. We exploit an alternative route to corroborate the appropriateness of the two components. Before presenting it, we stress that an index is an inequality measure if it observes a rigorous axiomatic approach (for an effective overview see Allison, 1978). We believe that a similar approach is desirable to evaluate the components of an inequality index decomposition. Coherently, in Section 5 we show that our components are strongly correlated with two benchmarks, which independently measure within and between inequality and are derived from literature following an axiomatic approach. Properties (i)-(ii) state that our components have a correct starting point - zero - and we demonstrate that they strongly correlate with the axiomatically derived benchmarks. This has a remarkable consequence: our components approximately provide all the information contained in the benchmarks and observe their axiomatically derived properties; correlation coefficients close to 1 imply that the extent of the approximation is negligible.

As a two-component decomposition our proposal has a further advantage: the inequality index and the two components are collinear. All the information about the inequality index and its decomposition can be provided by specifying the value of the index and its between (within) component. Relevant advantages in descriptive or regression tasks in terms of both interpretability and parsimony directly follow. To the best of our knowledge, our decomposition is the first to possess both the advantages of a two-component decomposition and a between component that properly informs about the distance between the entire distributions of the groups.

4 The different-sized groups extension

In this section we show that the equal-sized groups hypothesis is not binding. It was necessary to understand the decomposition arguments, but the proposal can be easily extended to cope with more general situations in which the K groups are different-sized. Denote the vector of the sizes with $\mathbf{n} = (n_1, \dots n_K)$, where $\sum_{k=1}^{K} n_k = N$. Eq. (2) has to be reformulated as:

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

and a way to guarantee the element x_j^k to exist is needed to employ the decomposition proposal. We propose two distinct solutions. The first allows to evaluate the two exact components but necessitates potentially unaffordable computations. The second drastically reduces computational requirements paying the cost of a negligible approximation.

The *exact* approach. It considers a new common size $n = mcm(\mathbf{n})$ and the resampling weights $p_k = n_k/n$, 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 $\forall\, m_k = 1, \ldots, p_k^{-1}$. Therefore $\forall\, (k,h) \in \{1, \ldots, K\} \times \{1, \ldots, K\}$ the following holds:

$$\sum_{i=1}^{n_k}\sum_{j=1}^{n_h} \left|x_i^k - x_j^h\right| = \sum_{i=1}^{n}\sum_{j=1}^{n} p_k p_h \left|y_i^k - y_j^h\right| \tag{6}$$

Proof.

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

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

□

We provide an example which should give the intuition of what we formally wrote. Imagine two groups composed, respectively, of two and three individuals, as the ones reported in the left rectangle of Figure 2. Replace them with those in the right rectangle. 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 within inequality point of view. In addition, the empirical cumulative distribution functions of the two groups are the same, before and after the replacement: also the distance between the two groups is unvaried. However, each couple difference in the left scheme appears in the

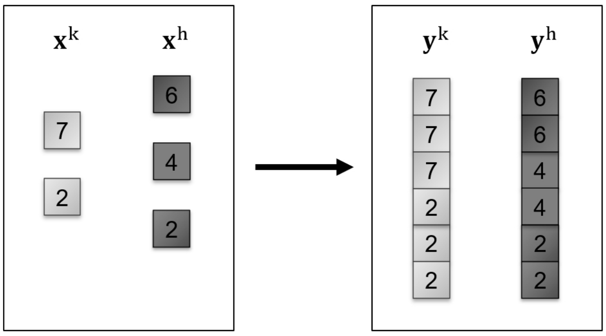


Figure 2: Exact approach: a two-group illustration

right scheme 9 times if the couple belongs to $\mathbf{y}^k$, 4 times if it belongs to $\mathbf{y}^h$ and 6 times if the two units belong to different groups. The p_k and p_h in eq. (6) adjust for this effect multiplying the three kinds of differences, respectively, by $1/9$, $1/4$ and $1/6$. In this way, equal-sized groups are obtained preserving the correspondence with the Gini index and with the original distributions of the groups.

The Gini index numerator can be decomposed with an analogous technique to the one employed deriving eq. (5). The following is obtained:

$$\begin{aligned} g=\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i-x_j\right| &= \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}\left|x_i^k-x_j^h\right| = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} p_k p_h \left|y_i^k-y_j^h\right| = \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i^k-y_j^k\right| w_{ij}^k + \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{j=1}^{n}\left|y_j^k-y_j^h\right| w_j^{kh} = g_w+g_b \end{aligned} \tag{7}$$

The only difference w.r.t. eq. (5) 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 previously defined weights and incorporate the information needed to preserve the original impact of each couple.

In most cases this approach requires an unaffordable computational effort because of the potentially-huge magnitude of the minimum common multiple. To reduce computational requirements, we present an alternative procedure that we refer to as *quantilisation*.

Quantilisation. We propose to consider a lower value of n and to calculate differently each $\mathbf{y}^k$: 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 constrains $n \geq n_k$, so it can be $p_k > 1$. The decomposition proposal has the same form of eq. (7) but G, G_w and G_b now incur in some approximation.

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 statistical software R. Given each vector $\mathbf{x}^k \in \mathscr{R}^{n_k}$, accordingly to this definition and in order to minimise the approximation of the quantilisation results, we suggest to interpolate linearly the vertices $\big((i-1)/(n_k-1), x_i^k\big)$ where $i = 1,\ldots,n_k$, and then to estimate the n quantiles - i.e. to determine the vector $\mathbf{y}^k$ - by the values associated to the probabilities

$$prob_j = \frac{j-1}{n-1} \quad j = 1,\ldots,n \tag{8}$$

on the resulting piecewise linear curve. As for the latter, we define $w_k = n_k / \sum_{k=1}^{K} n_k$ and advise the value:

$$n = \sum_{k=1}^{K} w_k n_k = \frac{\sum_{k=1}^{K} n_k^2}{\sum_{k=1}^{K} n_k} \tag{9}$$

which determines n as the average of the n_k, each weighted by its own share of population w_k.
The decisions proposed both for the quantile definition and for the value of n are motivated in the appendix. Here we only inform that, if they are employed, the approximation that the quantilisation procedure copes with is negligible. However, when $min(\mathbf{n})$ is high and a computational cost saving choice is required, it could be also acceptable - in terms of the magnitude of the approximation - to choose $n << min(\mathbf{n})$.

Before concluding this section, we suggest how to overcome the issue of the approximation in the absolute value of the components due to the quantilisation procedure. As we show in the appendix, the suggested strategy for the selection of quantiles ensures that the shares of the components for quantile-transformed data are consistent for the shares obtained by the exact approach. Hence, to obtain two consistent estimates of the exact components that sum up to the Gini index of the original data, it is sufficient to multiply the shares of the components (obtained by quantilisation) with the value of the index.

We finally observe that a generalised version of the quantilisation procedure may be wanted to cope with weighted data, i.e. to consider the weights of the observations. We suggest 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 apply a *standard* quantile function to vectors repeating the income of each unit as many times as the value of its weight states. As for the choice of *n*, we advise - but other choices that we do not discuss here can be suitable - to replace the w_k in eq. (9) with the cumulated relative weights of each group. This is how we proceed in Section 6.

In the next section a Monte Carlo experiment exhibits that each component of the introduced decomposition is strongly correlated with an axiomatically derived benchmark. In the Monte Carlo simulations each group k is replaced by n quantiles selected from $\mathbf{x}^k$; the value of n is determined by eq. (9); the Definition 7 from Hyndman and Fan (1996) and eq. (8) are employed to select the quantiles.

5 Correlation with benchmarks

Benchmarks. The two benchmarks which we consider are developed *ex ante* to measure within and between inequality. They are derived from literature and observe an axiomatic approach. Let G_k be the value of the Gini index in group k; μ_k its mean and n_k its dimension. The **within benchmark** is:

$$W_r = \sum_{k=1}^{K} \frac{n_k}{N} \frac{\mu_k}{\mu} G_k = \sum_{k=1}^{K} s_k G_k$$

where s_k is the share of income possessed by group k. Every G_k observes the axioms and the properties of the Gini index. Thus, the global properties of W_r incorporate them by a weighted mean, which in this case assigns a greater weight to the inequality of the groups possessing the biggest shares of income. As for the **between benchmark**, we employ the following index:

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

where Eb_{kh} is the diversity measure between two groups (k and h) proposed by Ebert (1984). Actually, Ebert proposes a general class of measures dependent on a parameter r. Here, Eb_{kh} is the measure corresponding to $r = 1$. A slight modification - we standardise the incomes dividing by their average μ - is introduced to observe the scale invariance criterion in addition to the other properties which the index already observes. The measure is defined as:

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

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 group k. A preceding proposal by Dagum (1980) had already developed a measure of *economic distance* between two income distributions, but it has been criticized by 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 entire distributions. B_r inherits these properties: it depends on them and on the aggregating consequences of the weighted mean. In this case, the weights are proportional to the share of couples in each group pair, in accordance with the weights of the "intercountry terms" calculated by Milanovic (2011) (p. 88-89).

Competitors. As anticipated, a Monte Carlo algorithm is employed to evaluate the extent of the correlation between the spatial components of our decomposition and the benchmarks. To outline the advantages of our proposal, we also compare the benchmarks with the components of the most widespread *subgroup* decompositions of the Gini index. Comprehensive outlines of these decompositions are provided in Giorgi (2011) and Radaelli (2010). This variety of decompositions originates by alternative formulations of the Gini index and by different approaches, but the resulting proposals are ascribable to two main representative strands.

Aiming for tractability, in presenting their characteristics we restrict the attention to the decompositions presented in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967). They both rely on a partitioned population and exhibit the two components measuring within and between groups inequality, plus a third term. We can write their general structure as:

$$G = G_w^{YL} + G_b^{YL} + R^{YL} \tag{10}$$

$$G = G_w^{BM} + G_b^{BM} + R^{BM} \tag{11}$$

where the apices YL and BM identify the two proposals.

The within components G_w^{YL} and G_w^{BM} measure the intra-territories inequality by two differently-weighted averages of the Gini index in each group. The within component from eq. (10) coincides with the measure that we selected as within benchmark:

$$G_w^{YL} = \sum_{k=1}^{K} \frac{n_k}{N} \frac{\mu_k}{\mu} G_k = \sum_{k=1}^{K} s_k G_k \tag{12}$$

where each weight s_k is immediately interpretable as the group k income share. Differently, in the weights of the within component from eq. (11) the population shares multiply the s_k, and G_w^{BM} reads:

$$G_w^{BM} = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 \frac{\mu_k}{\mu} G_k = \sum_{k=1}^{K} \frac{n_k}{N} s_k G_k \tag{13}$$

The distance among G_w^{YL} and G_w^{BM} only depends on the sizes of the population shares, which enter linearly in eq. (12) and quadratically in eq. (13). The choice to select G_w^{YL} as the within benchmark is due to the immediate interpretability of its weights.

As for the two between components, they also depend on the decomposition choice. The between component in eq. (10) is defined as:

$$G_b^{YL} = \frac{2}{\mu} Cov\left(\mu_k, \frac{\frac{1}{n_k}\sum_{i=1}^{n_k} R_{ik}}{N}\right) \tag{14}$$

where R_{ik} is the rank of the unit i from the group k in the overall population. The between component in eq. (11) reads:

$$G_b^{BM} = \frac{1}{2\mu} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{n_k n_h}{N^2} |\mu_k - \mu_h| \tag{15}$$

The two components in eq. (14)-(15) appear very different but, as the between component of every *subgroup* decompositions of inequality measures, they are both based on the means of the groups: the discussed oversimplification issue evidently appears in the results of this section.

As for the third components R^{YL} and R^{BM} - which are non-negative and disappear if the distributions of the groups do not overlap - they have initially remained uninterpreted and just considered as a residual. Thereafter, Yitzhaki and Lerman (1991), Yitzhaki (1994) and Dagum (1997) have proposed interesting interpretations - which are beyond the scope of this paper - in terms of overlapping, stratification and transvariation, making all the components in eq. (10)-(11) informative. However, there is no way to extract information on within and between inequality from R^{YL} and R^{BM}, as we do with our *residual* (the sum of the diagonal differences).

The Monte Carlo experiment. The algorithm works with three predetermined parameters: the number of groups, the parameter(s) of the distribution of **n** and the coefficient of variation between the averages of the groups ($CV[\mathbb{E}[\bar{x}^k]]$). It is schematised in Figure 3 and can be summarised as follows:

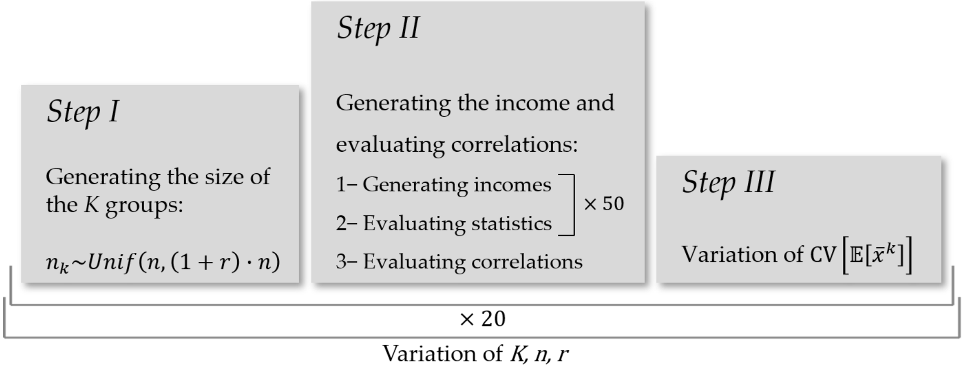

Step II
Generating the income and evaluating correlations:
1– Generating incomes
2– Evaluating statistics
× 50
3– Evaluating correlations
Step I
Generating the size of the K groups:
$n_k \sim Unif(n, (1+r) \cdot n)$
Step III
Variation of $CV[\mathbb{E}[\bar{x}^k]]$
× 20
Variation of K, n, r

Figure 3: The Monte Carlo algorithm

Step I. With K, (n,r) and $CV[\mathbb{E}[\bar{x}^k]]$ fixed, generate the vector **n**: each n_k is drawn from a uniform $[n, (1+r) \cdot n]$, where $100 \cdot r$ is the maximum percentage deviation from the minimum n.

Step II. Generate the incomes from lognormal distribution 50 times, each time evaluating all the involved indices. This allows to estimate the following two triples of correlation estimates:

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

where the superscripts A, YL and BM identify the three alternative decompositions.

Step III. Repeat *Step II* for different values of $CV[\mathbb{E}[\bar{x}^k]]$.

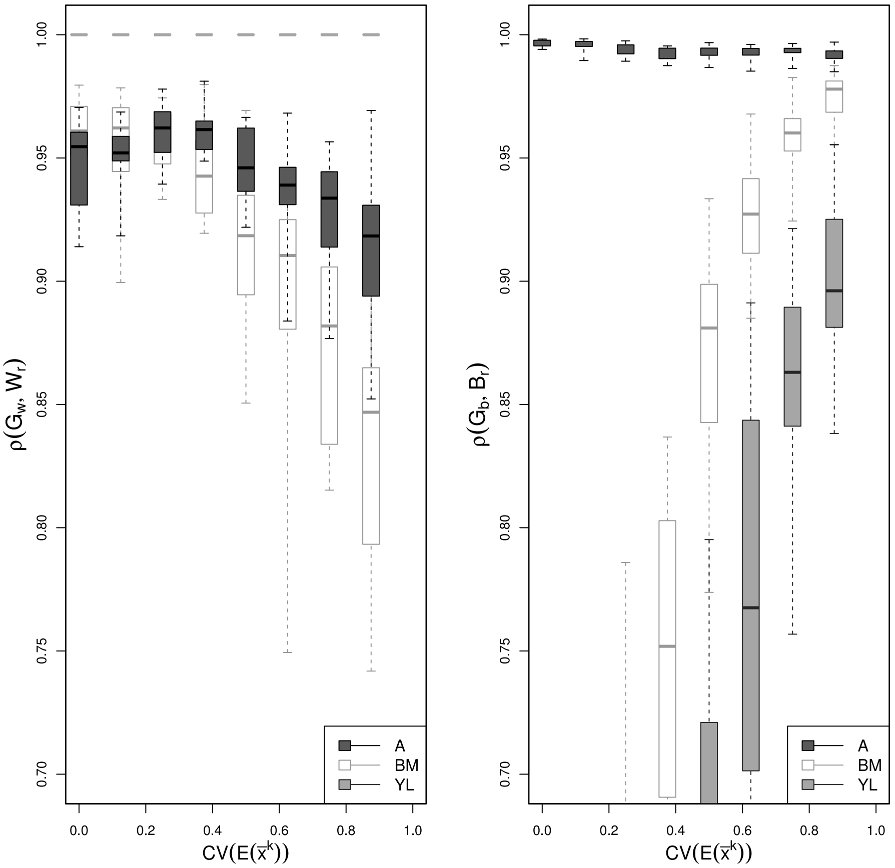
The algorithm runs *Step I-III* 20 times and delivers, for each value of $CV[\mathbb{E}[\bar{x}^k]]$, 20 replicates of the triples defined in *Step II*. More details about the income simulation procedure and its theoretical foundations can be found in the Appendix. Here we only stress that the parameters of the lognormal distribution are micro-founded. Indeed, as it is detailed in the second part of the Appendix, they are chosen sampling from the parameters estimated in Bandourian et al. (2002) using real data along different countries and periods. This should guarantee robust results with respect to real income distributions, as also confirmed by the results obtained in the next section using real data.

Look at the diagrams reported in Figure 4. Eight values of $CV[\mathbb{E}[\bar{x}^k]]$ are used to obtain the eight triples of boxplots. For each value of $CV[\mathbb{E}[\bar{x}^k]]$, the three kinds of boxplots in Figure 4a (4b) are built by the 20 replicates of the triples: black, grey 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 $\mathbf{n} \sim U\left([100, 500]^K\right)$.

The eight values of $CV[\mathbb{E}[\bar{x}^k]]$ allow to evaluate correlation in contexts characterised by increasing variability in the means of the groups. This highlights the advantages of our between component, which are striking in situations where the variability in the means of the groups is not large. As for the within component, G_w^{YL} presents by definition a perfect correlation with W_r. However, better than G_w^{BM}, G_w^A always reports extremely high correlation with W_r.

Results. The correlation values are studied by the same algorithm in multiple contexts by varying the number of groups and the distribution of $\mathbf{n}$. Table 1 reports the results for representative parameters. It summarises the 20 replicates produced to estimate each correlation distribution by their average and standard deviation $\left(\overline{\mu}, \overline{sd}\right)$. These pairs are evaluated for the eight values of $CV[\mathbb{E}[\bar{x}^k]]$ and are averaged pairwise determining four pairs $\left(\overline{\mu}, \overline{sd}\right)$. The pairs correspond to low, medium-low, medium-high and high levels of $CV[\mathbb{E}[\bar{x}^k]]$ and are available for both the within and the between components.

The results about our decomposition are remarkable. The correlations of the proposed components only marginally depend on the specification of the parameters. We only notify the most relevant variations in the table. Higher values of K negatively influence all the $\overline{\mu}$ related to our within component, but an increase of the values in $\mathbf{n}$ absorbs this small effect. The values of $\overline{\mu}$ also decrease for higher level of $CV[\mathbb{E}[\bar{x}^k]]$, while the values of $\overline{sd}$ tend to increase. However, all the $\overline{\mu}$ referred to our within component



(a) Within components

(b) Between components

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

are never below 0.92 and the maximum of the $\overline{sd}$ is $2.8 \cdot 10^{-2}$. As for our between component, its $\overline{\mu}$ slightly decreases and shows higher $\overline{sd}$ when the variability in $\mathbf{n}$ gets higher and the values in $\mathbf{n}$ and K are small.

Despite these details, Table 1 strengthens the conclusions drawn looking at Figure 4: the correlation estimates for our components always[7] maintain the described advantages.

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

Data. In this section we apply the proposed decomposition to the municipality based *Income and principal Irpef variables* statistical data, which are available among the Open-Source Data released by the Italian Ministry of Economy and Finance. They annually collect - our analysis ranges from 2000 to

[7] The only exception occurs when the variability in $\mathbf{n}$ is low: in this case the results for the within component from eq. (13) are enhanced. Indeed, it perfectly correlates with the within component from eq. (12) when the groups are equal-sized. However, the correlation rapidly decreases when the variability in $\mathbf{n}$ increases, so this aspect is negligible.

2017 - data from the tax declarations of the whole set of Italian taxpayers and report several variables on a municipality base; we consider the variable *total income*. For each municipality, the available information refers to 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[8] are available: the average income of each class with an attached weight given by the frequency in that class. We group them on provincial, regional and territorial (NUTS 1) base obtaining three different areal-unit partitions.

Validation. In the previous section, the correlation values were calculated simulating from independent scenarios. Here we complement the analysis evaluating the benchmarks and the components in each of the 18 years, and calculating their correlations over time. We apply this procedure to the whole dataset (Italy) and to five independent subsets identified by NUTS-1. For each subset, we group the data according to different administrative borders: provinces, regions and, exclusively for the analysis on the whole dataset, NUTS-1. The results reported in Table 2 confirm the very high correlations between the benchmarks and the proposed components. Despite the difference in the derivation, all the reported correlation estimates are definitely compatible with the findings in Table 1 and strengthen the consistency of the conclusions driven by the Monte Carlo experiment: the two components are appropriate to measure within and between inequality. But do they really show something interesting and new?

Provincial-based evidence. In Figure 5 we consider the provincial-based aggregation and investigate the consequences of the decomposition choice on the between component share trajectory, which is central in the strategy proposed by Shorrocks and Wan (2005) to assess spatial concentration. The three time series range in different intervals. To better underline their relative evolution, we rescaled them dividing by their own initial value. 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 years of the financial crisis, then the component from Bhattacharya and Mahalanobis (1967) moves more similarly to ours. However, they both vary irregularly during the first ten years and do not unambiguously capture the decreasing path followed by the introduced between component. This confirms that the variability in the means is not always able to exhaustively inform about the economic distance between groups and about spatial patterns in the income distribution. The introduced between component can better assess spatial concentration.

[8] They are less for the municipalities with classes containing less than four units.

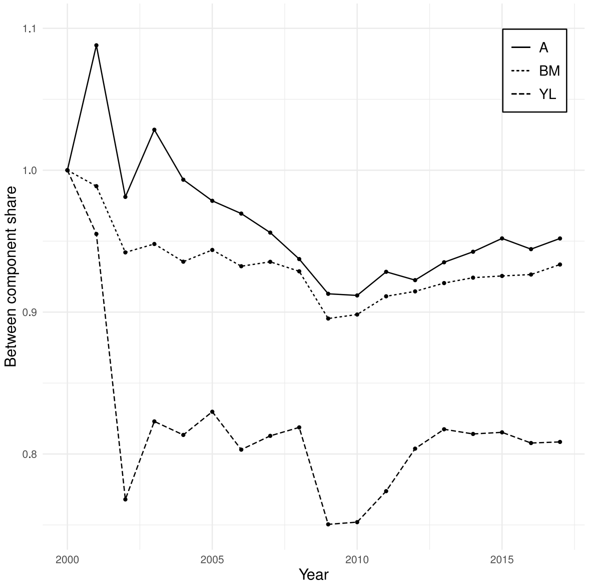


Figure 5: Time series of the shares over the Gini index of the between components from the three considered Gini index decompositions. In the legend A, YL and BM stand for the decomposition proposed in this work, in Yitzhaki and Lerman (1991) and in Bhattacharya and Mahalanobis (1967), respectively. Each series is rescaled dividing by its own initial value.

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 follow, but their reliability depends on both the appropriateness of the components and the index involved. As for the former, we have already justified 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 partitioned population.

We still consider the provincial-based aggregation. Thus, the Gini index measures the inequality in the municipal per-class income distribution; and the within and the between component shares represent the contributions of the within and between provinces differences to the overall inequality. The values of

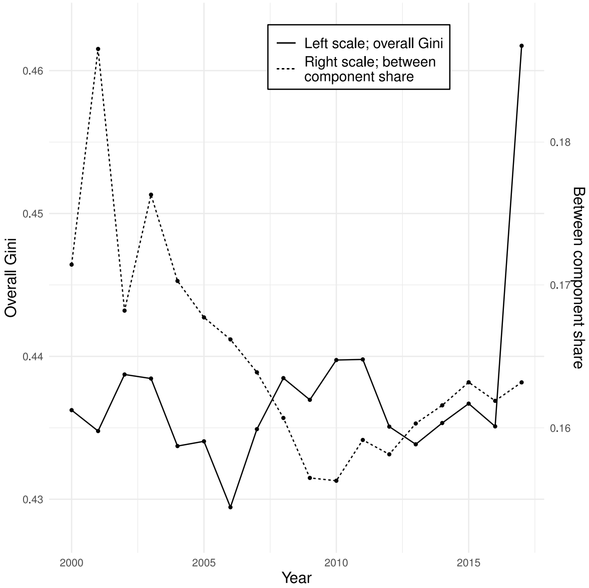


Figure 6: Gini index trajectory of the municipal per-classes income distribution - left scale. Time series of the proposed between component share over the Gini index - right scale.

the share of our between component on the Gini index are plotted in Figure 6 - right scale. The Gini index is reported too - left scale. The latter varies quite irregularly during the considered period, with a sudden increase in the last year. The former shows an initial marked decreasing trend followed by a light recovery; the converse holds for the within component share, which can be easily derived by a reflection and a one-unit-long vertical translation of the between share path.

It is also possible to effectively inform about the contributions that the two components provide to the Gini index percentage variation. The Gini index yearly percentage variations are reported in Figure 7, along with the contributions to them from the changes in the between component. Despite the between component has a minority share over the Gini index at this aggregation level (Figure 6), its influence to the Gini index path is relevant. In the first years of the period, the changes in the between inequality have mainly acted restraining the effects of the within inequality on the Gini index path. Conversely, the two components have affected the overall inequality in the same direction since 2010. As before, interesting conclusions on the within term can be evinced from the results about the between component.

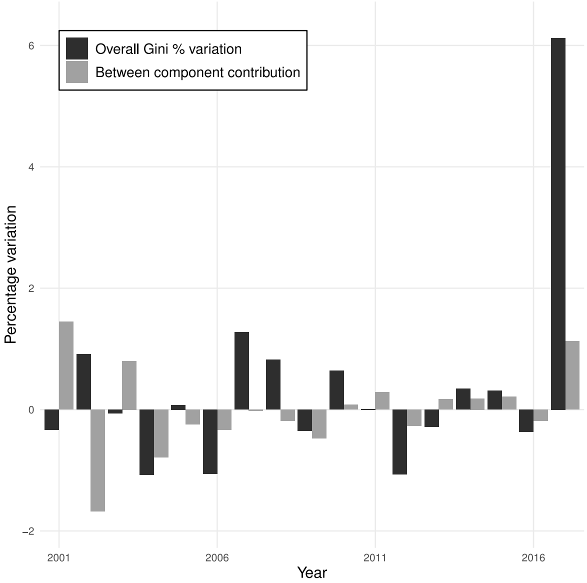


Figure 7: Gini index yearly percentage variation and between component change contribution.

7 Conclusions

A recent line of research convincingly considers the decomposition of inequality measures as a source to effectively evaluate spatial concentration.

The most widespread decompositions of the Gini index consider a partitioned population, and consist of three components: a residual term augments the sum of two (spatial) components measuring within and between groups inequality. The resulting between component is always such that groups with similar means present low levels of territorial inequality. However, when the distributions of the groups overlap, further characteristics of the distributions are relevant to assess between inequality. The decomposition developed in this paper fills this gap.

Exploiting a new insight into the Gini index, we extract crucial information from our residual term, disentangling two contributions to within and between inequality and delivering a decomposition that is exactly composed by the two spatial components. As far as we know, this is the first Gini index decomposition to deal with a partitioned population and to avoid the residual. In addition, its between component compares the entire distributions of the groups and not only their first moment. The importance of these aspects is highlighted in our empirical analysis: unlike the existing measures of between groups inequality, our decomposition is able to capture a decreasing path in the income inequality between Italian provinces.

As we have discussed, focusing on the geographical framework, the advantages of our proposal makes it well-suited to measure spatial concentration and can be helpful to design and to evaluate place-based policies. However, our Gini index decomposition can be applied to any kind of partition, providing two informative measures of within and between groups inequality. We strongly believe that the importance of the Gini index for within-between decomposition of inequality should be reconsidered in the light of the introduced decomposition and of the novelty of its characteristics, which open new opportunities in the analysis of inequality and related fields.

Acknowledgement

I would like to express my gratitude to Prof. G. Pellegrini and Prof. R. Zelli, who constantly advised me in the first steps of this work. My special gratitude is for Prof. M. Costa, who was essential in several aspects leading to the final version of this manuscript. I would also like to thank Prof. G. Cavaliere, Prof. G. Pignataro, Prof. Martyna Kobus, Dr. G. De Marzo, Dr. V.S. Bramante and all the involved friends for their useful comments.

Declarations

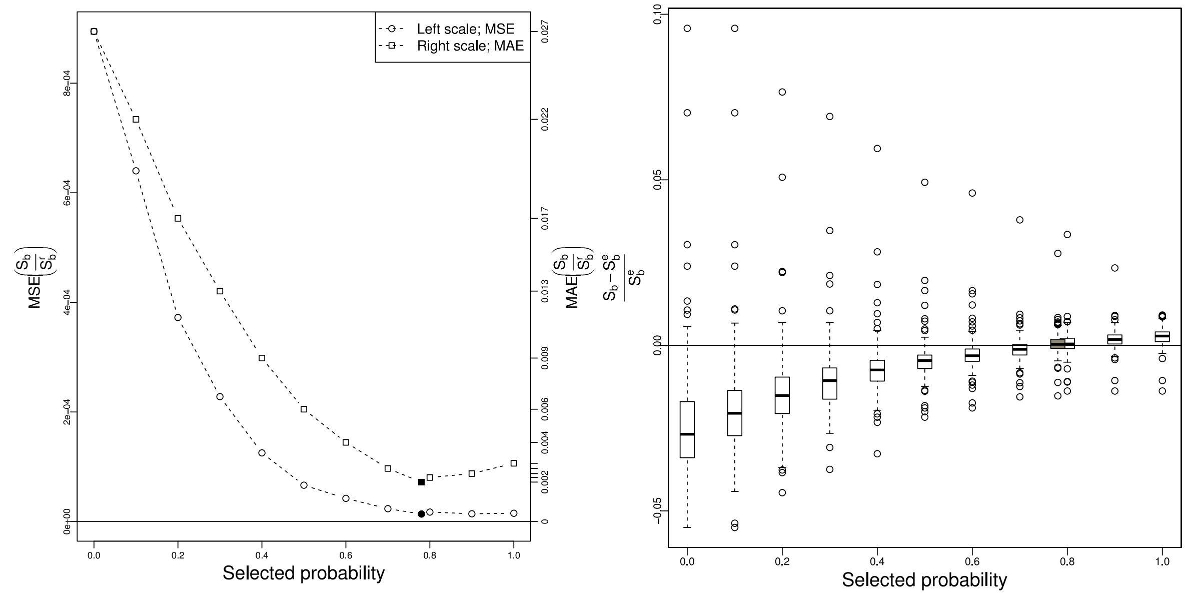
The author has no conflicts of interest to declare that are relevant to the content of this article.

Appendix

On the quantilisation procedure

The first section of this appendix is aimed at analysing the quantilisation procedure, namely to identify the optimal definition of quantile and the optimal value of n, to explain the reason behind their optimality and to quantify the magnitude of the approximation incurred.

Defining $\mathbf{w} = (w_1, \dots w_K)$ we can rewrite the suggested value of n as $n = \mathbf{w}\mathbf{n}^\intercal$. This expression determines n as the average of the n_k, each weighted by the share of population w_k. The performance of this value is firstly shown in Figure 8, where approximation is evaluated looking at the relative discrepancy between the two values of G_b/G obtained employing the exact and the quantilisation method. Precisely, define $S_b = G_b/G$ as the between component share obtained by the quantilisation method and $S_b^e = G_b^e/G^e$ as the same share obtained by the exact approach. The relative discrepancy is measured by the Mean and the Absolute Squared Error of S_b/S_b^e w.r.t. $1 = S_b^e/S_b^e$. They are obtained running 150 simulations and evaluating the empirical counterpart of $\text{MSE}\big(S_b/S_b^e\big) = \mathbb{E}[\big(S_b/S_b^e - 1\big)^2]$ and $\text{MAE}\big(S_b/S_b^e\big) = \mathbb{E}[\big|S_b/S_b^e - 1\big|]$.



(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 8: 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.

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. We compare alternative choices of n, which are the minimum and the maximum of $\mathbf{n}$, its deciles and the value obtained by eq. (9). The vector $\mathbf{n}$ is also drawn, but some constraints on its elements are imposed to ensure affordable values for $mcm(\mathbf{n})$. To be specific, the algorithm firstly specifies K ($= 5$, 10 or 20). Then it builds 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) are both included in $\mathbf{n}$. The other $K-2$ values are sampled with repetition from **mul**. With this choice the mcm cannot exceed the value 2160 and the computations are affordable. Figure 8 represents the results for $K = 20$, $min = 72$ and $max = 720$.

As shown in Figure 8a the proposed value of n - represented by the solid indicators - minimizes (or reach a value very close to the minimum of) the approximation that this method copes with, both for the MSE (left scale) and the MAE (right scale). This result is achieved thank to a vanished distortion and a variance reduction, as Figure 8b shows. We 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 between component share percentage approximation depends on the simulation parameters, as Table 3 points out. It reports the MAE[9] - multiplied by 10^2 to express the relative discrepancy in percentage points - of the between component share obtained by the described procedure for different choices of n, K and of the interval $[min, max]$.

Results are really encouraging. The values of the MAE are below the percentage point approximately in half of the analysed contexts and always when the suggested choice of n is employed. In addition, the dependence of results on the employed parameters - which is described just below - could further ensure a reduction in the approximation in many realistic contexts where the parameters are presumably more conducive.

For each choice of n, when the ratio $^{max}/_{min}$ decreases - i.e. if the variability in $\mathbf{n}$ decreases - the approximation reduces, too. If that ratio stays constant, the MAE informs about better performance for higher min and max. Results are enhanced when n is selected by eq. (9) and the number of groups is high. The described dependence of the MAE on the values of n, K and of the interval $[min, max]$ can be considered as a kind of consistency for our procedure.

Furthermore, the suggested choice of n almost always guarantees a 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. (9) 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. However, as supported

[9] We prefer it because of its interpretability as average absolute percentage error.

by the values in the third column of the table - which decrease when $min(\mathbf{n})$ increase - 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.

We now present the assessment procedure which leads to the selection of the quantile definition employed in the analysis. We evaluate the impact of several quantile definitions on the approximation that the quantilisation procedure copes with. Figure 9 compares the approximations achieved iterating the same procedure which generates Figure 8a using the nine different quantile definitions presented in Hyndman and Fan (1996). The Definition 7 essentially presents the lowest MSE (and MAE) for each choice of n and it ensures computational advantages because the MSE approaches 0 for smaller n.

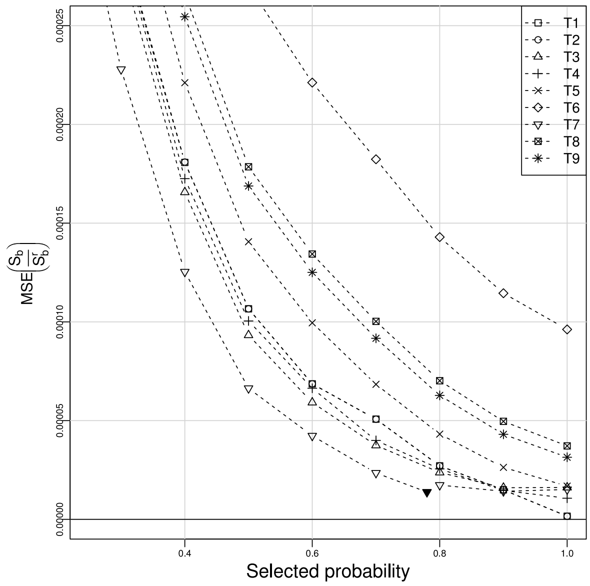


Figure 9: Between component share approximations - measured by the MSE and through the same procedure which produced Figure 8 - obtained employing the 9 quantile definitions presented in Hyndman and Fan (1996). Using the software R, each definition can be selected by the option *type* of the function *quantile()*. Here, T_j: $j = 1, \dots, 9$ stands for selecting the option $type = j$.

The better performance resulting from the definitions 1 and 2 when the selected probability is close to 1 are exceptions. Both the definitions rely on a stepwise cumulative probability function which selects the quantiles in the set of the values in the starting vector. Thus, if $p = 1$ and $max(\mathbf{n}) = mcm(\mathbf{n})$, the vector of quantiles corresponds to the $\mathbf{y}^k$ of the exact approach: no approximation is encountered. The approximation is negligible if $max(\mathbf{n})$ close to $mcm(\mathbf{n})$ and p approaching 1, i.e. n close to $max(\mathbf{n})$. In Figure 9 this is evident from $p = 0.8$. Nonetheless, in the vast majority of 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 $max(\mathbf{n})$ and the Definition 7 from Hyndman and Fan (1996) is definitely recommended.

Actually, the optimal performance associated to the suggested quantiles selection strategy should not come as a surprise. Its outstanding results have a twofold explanation. First, the performance of the proposed choice of n directly derives from its consistency with the exact-approach weighting system. This choice assigns greater weights w_k to the sizes of the most sized groups, which is desirable because these are the groups with the biggest associated values of p_k. It is reasonable to preserve their information choosing a large n and resampling the smaller groups taking their quantiles. But if many small groups are present, n is attracted towards their small size. Here the quantilisation of big groups and the related loss of information are preferred to the approximation which would be incurred resampling the many small groups. The second explanation for the optimal performance of the suggested strategy is the following. Eq. (8) selects the values $prob_j$ so to 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, when the discussed quantiles selection procedure is employed, then $min(\mathbf{x}^k)$ and $max(\mathbf{x}^k)$ are preserved for each n and k. Moreover, if $n_k = n \ \ \forall\ k$, then the vectors $\mathbf{x}^k$ are entirely preserved, too. Both this properties hold at the same time only employing the Definition 7 from Hyndman and Fan (1996) and the discussed choice of the values $prob_j$. They ensure robustness to the quantilisation procedure w.r.t. outliers and contribute to explain the negligible approximation which is incurred.

The income simulation algorithm

A Monte Carlo algorithm is employed to evaluate the approximation of the quantilisation procedure and to estimate the correlation between the two benchmarks and the components from the alternative decompositions. This section of the appendix provides with the theoretical foundations of the income simulation procedure which feeds both these algorithms.

The distribution of $\mathbf{n}$ is a K-variate uniform, where the number of groups K and the extremes of the distribution are determined *ex-ante*. A uniform distribution is 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 is set to 10^4. As for the maximum M, it is fixed to $5 \cdot 10^4$ in the simulations which generated the results of the first part of this appendix. Differently, in the analysis described in Section 5, M is varied to highlight the impact of the variability in the means of the groups on the investigated correlations. This is possible because a modification of M directly affects $CV[\mathbb{E}[\bar{x}^k]]$. For the uniform distribution $\mathbb{E}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right] = (M+m)/2$ and $\mathrm{Var}_u\left[\mathbb{E}\left[\bar{x}^k\right]\right] = (M-m)^2/12$, therefore 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, it only depends on the value of M. In Figure 4, the interval $\left[0, 1/\sqrt{3}\right]$ and the values of $CV[\mathbb{E}[\bar{x}^k]]$ are rescaled to the interval $[0,1]$ by a simple scale transformation. This is not an issue because $CV[\mathbb{E}[\bar{x}^k]]$ is not directly comparable with values of $\bar{x}^k$ coming from a non-uniform distribution.

The values of M are selected so that the coefficient of variation divides the interval in S equal parts.

Denote by $M^{(s)}$, $s = 1 \dots S$ the different values required for this scope. The values $M^{(s)}$ satisfy:

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

with $M^{(0)} = m$ and $c = 1/\left(\sqrt{3}S\right)$. With easy calculations the following holds:

$$M^{(s)} = \frac{m(cM^{(s-1)} + cm + 2M^{(s-1)})}{(2m - cM^{(s-1)} - cm)}$$

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

Once that all the parameters are fixed, the incomes of each group k are drawn from a lognormal distribution with expected value $\mathbb{E}\left[\bar{x}^k\right] \sim Unif(m,M)$. The last requirement is to define a meaningful way to determine the two parameters μ and σ^2 of the distribution. As it is well known, for a lognormal distribution the following holds:

$$\mathbb{E}\left[\bar{x}^k\right] = e^{\mu_k + \frac{\sigma_k^2}{2}} \tag{16}$$

This equation allows to design an effective way to split $\mathbb{E}\left[\bar{x}^k\right]$ in the two elements μ_k e σ_k which are required to draw from the distribution - in a manner that the lognormal is a plausible income distribution. Starting from eq. (16) 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:

$$\mu_k = \alpha_k \ln \mathbb{E}\left[\bar{x}^k\right] \tag{17}$$

$$\sigma_k^2 = 2\left(1 - \alpha_k\right) \ln \mathbb{E}\left[\bar{x}^k\right] \tag{18}$$

Their ratio is

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

At this point, we consider the 82 couples of lognormal parameters estimated in Bandourian et al. (2002) using 82 real income distributions from 23 countries over several years (from the end of sixties to the end of nineties). We evaluate the $c_i = \sigma_i^2/\mu_i,\ i = 1,\dots,82$ corresponding to each couple.

Verisimilar values for α_k can be obtained sampling a value of i for each group and posing $c_k = c_i$. Finally, solve the following equation:

$$c_i = c_k = \frac{2(1-\alpha_k)}{\alpha_k} \Longrightarrow \alpha_k = \frac{2}{c_i+2} \tag{19}$$

Therefore μ_k and σ_k^2 are determined - taking $\mathbb{E}[\bar{x}^k]$ as known - by eq. (17)-(19).

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 et al. (2002) do not appear to be influenced by the associated $\mathbb{E}[\bar{x}^k]$: a simple linear regression reports an approximately null coefficient ($5.6 \cdot 10^{-4}$) and a large p-value (0.65) for the regressor $\mathbb{E}[\bar{x}^k]$. Consequently, 82 possible proportions to split $\mathbb{E}[\bar{x}^k]$ in a likely way in the two addenda μ_k and $\sigma_k^2/2$ are available. We exploit them to simulate income.

References

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

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

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

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

Ceriani, Lidia and Paolo Verme (2012). “The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini”. In: *The Journal of Economic Inequality* 10.3, pp. 421–443.

— (2015). “Individual diversity and the Gini decomposition”. In: *Social Indicators Research* 121.3, pp. 637–646.

Cowell, Frank A (1988). “Inequality decomposition: three bad measures”. In: *Bulletin of Economic Research* 40.4, pp. 309–312.

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

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

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

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

Foster, James E and Artyom A Shneyerov (2000). “Path independent inequality measures”. In: *Journal of Economic Theory* 91.2, pp. 199–222.

Giorgi, Giovanni M (2011). “The Gini inequality index decomposition. An evolutionary study”. In: *The measurement of individual well-being and group inequalities: Essays in memory of ZM Berrebi*, pp. 185–218.

Harrell Jr, Frank E, with contributions from Charles Dupont, and many others. (2020). *Hmisc: Harrell Miscellaneous*. R package version 4.4-1. URL: `https://CRAN.R-project.org/package=Hmisc`.

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

Milanovic, Branko (2011). *Worlds apart: Measuring international and global inequality*. Princeton University Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

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

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

Rodríguez-Pose, Andrés (2018). “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, pp. 189–209.

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

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

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

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

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

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

			Within components								**Between components**							
			$K=3$				$K=30$				$K=3$				$K=30$			
Magnitude of $CV[\mathbb{E}[\bar{x}^k]]$			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 benchmarks and the components from three Gini index decompositions: A, YL and BM in the third column identify 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 Section 5 for different values of K, $\mathbf{n}$ and $CV[\mathbb{E}[\bar{x}^k]]$. In this table, the eight vectors of correlation replicates for each component are summarised by their average and standard deviation, which are averaged pairwise in four pairs ($\overline{\mu}$, $\overline{sd}$), corresponding to low, medium-low, medium-high and high levels of $CV[\mathbb{E}[\bar{x}^k]]$.

			$cor(G_w, W_r)$			$cor(G_b, B_r)$		
	Aggregation	K	A	YL	BM	A	YL	BM
Italy	Provinces	107	.968	1	.887	.998	.630	.768
	Regions	20	.988	1	.963	.993	.778	.880
	NUTS-1	5	.970	1	.987	.997	.677	.781
ITC	Provinces	25	.871	1	.868	.998	.121	.663
	Regions	4	.825	1	.846	.997	.170	.733
ITF	Provinces	24	.995	1	.994	.985	.640	.548
	Regions	6	.990	1	.992	.961	.650	.744
ITG	Provinces	14	.997	1	.981	.992	.562	.016
	Regions	2	.992	1	.986	.990	.422	.943
ITH	Provinces	22	.981	1	.983	.998	.779	.822
	Regions	4	.973	1	.988	.994	.552	.638
ITI	Provinces	22	.994	1	.937	.990	-.218	.734
	Regions	4	.974	1	.921	.967	.473	-.106

Table 2: Correlations over time between the components from the three considered decompositions and the discussed benchmarks. Alternative subsets of the data - Italy as a whole and the five independent NUTS-1 territories - are analysed separately to increase robustness of the results. Correlations are evaluated over different aggregations.

		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 lower than the suggested n.

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