AStA Advances in Statistical Analysis (2023) 107:73–100
https://doi.org/10.1007/s10182-021-00397-6

ORIGINAL PAPER

# Decompositions by sources and by subpopulations of the Pietra index: two applications to professional football teams in Italy

Francesco Porro[1] · Mariangela Zenga[2]

Received: 6 May 2020 / Accepted: 23 February 2021 / Published online: 19 April 2021

**Abstract**
In this paper two innovative procedures for the decomposition of the Pietra index are proposed. The first one allows the decomposition by sources, while the second one provides the decomposition by subpopulations. As special case of the latter procedure, the "classical" decomposition in two components (*within* and *between*) can be easily obtained. A remarkable feature of both the proposed procedures is that they permit the assessment of the contribution to the Pietra index at the smallest possible level: each source for the first one and each subpopulation for the second one. To highlight the usefulness of these procedures, two applications are provided regarding Italian professional football (soccer) teams.

**Keywords** Pietra index · Income inequality · Decomposition by sources · Decomposition by subpopulations

## 1 Introduction

Introduced by Pietra (1915) as one of the first inequality measures, the Pietra index has more than a century of history, albeit that a quite similar measure was proposed a few years before (see Bresciani Turroni 1907; Hasan and Malik 2019). In the literature, the Pietra index has been "rediscovered" many times with different names: it coincides with the index proposed by Ricci (1916), with the Hoover index (Hoover 1936), and with the Schutz coefficient (Schutz 1951), and some papers refer to it as the Robin Hood index (see for example Koolman and van Doorslaer 2004; Wilkinson and Symon 2000; Kennedy et al. 1996).

Historically, the popularity of this index decreased as the Pigou–Dalton transfer principle became popular. It is well known that the Pietra index satisfies only

✉ Francesco Porro
fporro@uniss.it

[1] Università degli Studi di Sassari, Sassari, Italy

[2] Università degli Studi di Milano-Bicocca, Milan, Italy

the *weak* version of that transfer principle, given that it is not sensible to transfers between units with values on the same side of the mean (see for example Castagnoli and Muliere 1990; Frosini 2012).

Nevertheless, the Pietra index continues to be used in many different fields, as an indicator of heterogeneity, given that it can be regarded as a measure of the distance between the situation at stake and the egalitarian one, where all the units have the same amount. The Pietra index appears in many papers in the fields of public health (Wilkinson and Symon 2000; Theodorakis et al. 2006; Mobaraki et al. 2013; Koolman and van Doorslaer 2004; Zafari and Ekin 2019; Mantzavinis et al. 2002; De Maio 2007; Johnston and Wilkinson 2001) and medicine (see among others Gravelle and Sutton 1998; Kennedy et al. 1996; Beck et al. 2013). It has been also used in some sociological analyses (Shi et al. 2003; Kennedy et al. 1998; Rogerson and Plane 2013; Shumway and Otterstrom 2001; Ray and Singer 1973; Alker and Russett 1964; Khanal 2011). Finally and unsurprisingly, the Pietra index has been considered in many economic studies (see among others Hasan and Malik 2019; Khosravi Tanak et al. 2015; Moothathua 1989; Davydov and Zitikis 2005; Sarabia and Jorda 2014; Koolman and van Doorslaer 2004; Sarabia 2008; Eliazar and Sokolov 2010; Hustopecky and Vlachy 1978; Habib 2012; Huang and Leung 2009; Frosini 2012).

Another reason for the broad notoriety and longevity of the Pietra index is likely its very intuitive interpretation: it represents the portion of the total amount to be redistributed from the owners with more than the mean to the others to obtain the egalitarian situation. Moreover, the Pietra index has been shown to have an immediate and fundamental interpretation within renewal processes and continuous-time random walks, infinite-server queueing and shot-noise processes, and even in the field of financial derivatives (Eliazar and Sokolov 2010). In the more general context of infinite populations, dealing with random variables, the Pietra index can be used as heterogeneity index in more cases than can the relative standard deviation (RSD), since it can be calculated also whether the variance is not finite. In the literature, many aspects related to inequality (and heterogeneity) analysis have been investigated, mainly because of the important applicative implications. Among the others, decompositions play an important role. The two more general kinds of decomposition are by subpopulations (or by subgroups) and by sources (or by factors). The first one is performed when the population is divided into, say $k$, exhaustive and disjoint subpopulations. The question to be answered is how the subpopulations contribute to the value of the index, with the ideal answer being an evaluation of the contribution of each single subpopulation. Unfortunately, many decomposition procedures in the literature do not achieve this goal, given that they stop at a “higher” level of detail: usually, inspired by the well-known classical variance decomposition, they identify a *within* component (related to the inequality into the subpopulations) and a *between* component (depending on the inequality across the subpopulations). Moreover, many of them also require a third residual part (sometimes called the Transvariation component) that rescales the sum into the interval [0, 1], for example as in Dagum (1997).

The second type of decomposition arises when the investigated variable is the sum of other (say $c$) variables, called sources. In this framework, the research question is how to assess the contribution of each source. For two recent decompositions of the Pietra index and further references on the topic, see Habib (2012) and Frosini (2012).

Regarding the procedures for decomposing inequality indexes, a number of previous studies followed the approach proposed by Shorrocks (1982, 1984). In those two papers, some constricting hypotheses about the decomposition procedure are assumed, by forcing an analogy with the variance decomposition. The result is a very restricted class of decomposable inequality measures that does not contain many widely used ones such as the Gini coefficient and the Bonferroni index, to name but two. To overcome this issue, the methods proposed herein follow a different approach. The aim of this paper is to introduce two innovative procedures for decomposing the Pietra index, starting directly from its definition. The first is by sources, and the second is by subpopulations. As mentioned above, the relevant advantages of these two decompositions are that the first allows to assess the contribution of each source, while the second allows to evaluate the contribution of each subpopulation. For these reasons, these two procedures are very innovative and far more informative than many others proposed previously for the Pietra index. By using a different aggregation, the decomposition by subpopulations leads easily to the classical one into *within* and *between* components. The proposed methods are completed with two applications related to Italian professional football (soccer) teams. The decomposition by sources is applied to a dataset from the balance sheets of the teams in the top Italian league (Serie A), while the decomposition by subpopulations is illustrated by analyzing the values of all the players of the teams, grouped in all three professional leagues (Serie A, B, and C). The purpose of these applications is to visualize (albeit nonexaustively) and interpret the most popular sport in Italy.

This paper is organized as follows. In Sect. 2, the Pietra index is defined and some of its features are presented. In Sect. 3, a new equivalent formula for the Pietra index is given, one that is useful for the remainder of the paper. In Sects. 4 and 5, the two proposed decomposition procedures are detailed. Section 6 is devoted to the two applications to Italian football teams with real datasets, and the paper concludes in Sect. 7 with some final remarks. Appendix A provides an example in which the two proposed decompositions for the Pietra index are computed. For the datasets used in the applications and the R code to replicate the analyses, see the provided supplementary material.

## 2 The Pietra index

Let $Y$ be a non-negative statistical variable on a population of size $N$. Let $y_1 < y_2 < \cdots < y_r$ denote the distinct values assumed by $Y$ with frequencies $n_{1\cdot}, n_{2\cdot}, \ldots, n_{r\cdot}$. Obviously, it holds that $\sum_{h=1}^{r} n_{h\cdot} = N$. Let $M(Y) = \frac{T(Y)}{N} = \sum_{h=1}^{r} y_h \cdot \frac{n_{h\cdot}}{N}$ be the arithmetic mean of $Y$ in the whole population, and let $T(Y)$ denote the sum of the values of $Y$. Let

$$S_{M(Y)} = \sum_{h=1}^{r} |y_h - M(Y)| \cdot \frac{n_{h\cdot}}{N}$$

be the mean absolute deviation (MAD) of $Y$ from $M(Y)$. The index proposed by Pietra (1915) for the variable $Y$ is given by

$$\mathcal{P}_Y = \frac{S_{M(Y)}}{2M(Y)} = \frac{\sum_{h=1}^{r} |y_h - M(Y)| n_{h\cdot}}{2 \sum_{h=1}^{r} y_h n_{h\cdot}}.$$

It is worth to remark that in that paper:

(a) the Lorenz curve $L(p)$, proposed by Lorenz (1905) and defined as the piecewise linear curve, starting from the origin and interpolating the $r$ points $L_h$

$$L_h = \left( \frac{n_{h\cdot}}{N}; \sum_{i=1}^{h} \frac{y_i n_{i\cdot}}{T(Y)} \right), \qquad h = 1, \dots r$$

is formalized for the continuous case;

(b) it is shown that

$$\begin{aligned} \mathcal{P} &= \max_{p \in [0,1]} \left[ p - L(p) \right] \\ &= \tilde{p} - L(\tilde{p}), \end{aligned} \tag{1}$$

where $\tilde{p}$ is the quantile corresponding to the mean of $Y$, meaning that $y_{(\tilde{p})} = M(Y)$. Expression (1) can be seen as representing the maximum distance between the Lorenz curve of the variable at stake and the Lorenz curve of the egalitarian situation, therefore the Pietra index can be considered as "the Lorenzian counterpart of the Kolmogorov–Smirnov statistic, which quantifies the distance between two probability laws as the $L_\infty$ distance between their cumulative distribution functions" (Eliazar and Sokolov 2010).

The interested reader may care to know that De Capitani (2013a, 2013b) provided an English translation of the paper by Pietra (1915).

The Pietra index $\mathcal{P}$ has the following properties.

1. The minimum of $\mathcal{P}$ occurs in the case of perfect equality, namely $r = 1, y_1 = M(Y)$, and $n_{1\cdot} = N$. In a such case:

$$\mathcal{P} = 0.$$

2. The maximum of $\mathcal{P}$ occurs if $r = 2$, $y_1 = 0$, $y_2 = T(Y)$, $n_{1\cdot} = N - 1$, and $n_{2\cdot} = 1$, resulting in

$$\mathcal{P} = \frac{N-1}{N}.$$

3. $\mathcal{P}$ decreases in case of positive translations: if $c > 0$ and $W = Y + c$, then
$$\mathcal{P}_W < \mathcal{P}_Y.$$
4. $\mathcal{P}$ is invariant to positive scale transformations: if $c > 0$ and $W = c \cdot Y$, then
$$\mathcal{P}_W = \mathcal{P}_Y.$$
5. As mentioned in Sect. 1, $\mathcal{P}$ satisfies the weak principle of transfers but not the strong principle of transfers. This means that it is sensible to transfer between two units only if the corresponding values are one lower and the other higher than the mean. If the two values are both lower (or higher) than the mean, the Pietra index does not change. For more details on this point, see Castagnoli and Muliere (1990), and Frosini (2012).
6. $\mathcal{P}$ satisfies the population replication principle, since both $S_{M(Y)}$ and $M(Y)$ are invariant to population replication.

The interpretation of the Pietra index is very interesting and immediate: it is the share of the total amount $T(Y)$ that should be properly redistributed from the units possessing more than the mean $M(Y)$ toward the units possessing less than or equal to $M(Y)$, in order to achieve the situation of the perfect equality, where all the units have the same amount (absence of inequality). In fact, it holds that:

$$\mathcal{P} \cdot T(Y) = \sum_{\{y_h \le M(Y)\}} [M(Y) - y_h] n_{h\cdot} = \sum_{\{y_h > M(Y)\}} [y_h - M(Y)] n_{h\cdot}$$

## 3 An alternative useful expression for the Pietra index

At each $y_h$, $h \in \{1, 2, \dots, r\}$ the whole population can split into two non-overlapping groups:

– a lower group corresponding to $\{Y \le y_h\}$, including the first $P_{h\cdot} = \sum_{t=1}^{h} n_{t\cdot}$ units with total amount $Q_{h\cdot}(Y) = \sum_{t=1}^{h} y_t n_{t\cdot}$;
– an upper group corresponding to $\{Y > y_h\}$ that contains the remaining $N - P_{h\cdot}$ units, with amount $T(Y) - Q_{h\cdot}(Y)$.

Let
$$\bar{h} = \max\{h : y_h \le M(Y)\},$$
and let

$$\bar{M}_{\bar{h}\cdot}(Y) = \frac{Q_{\bar{h}\cdot}(Y)}{P_{\bar{h}\cdot}},$$

be the arithmetic mean of the lower group corresponding to $\left\{Y \leq y_{\bar{h}}\right\}$. Now, it follows that

$$\begin{aligned}\mathcal{P} =& \frac{1}{T(Y)} \cdot \sum_{h=1}^{\bar{h}} [M(Y) - y_h] \cdot n_{h\cdot} \\ =& \frac{M(Y)P_{\bar{h}\cdot} - Q_{\bar{h}\cdot}(Y)}{NM(Y)} \\ =& \frac{M(Y)P_{\bar{h}\cdot} - \bar{M}_{\bar{h}\cdot}(Y)P_{\bar{h}\cdot}}{NM(Y)} \\ =& \frac{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)}{NM(Y)} \cdot P_{\bar{h}\cdot} \\ =& V_{\bar{h}}(Y)p_{\bar{h}\cdot},\end{aligned} \tag{2}$$

where

$$V_{\bar{h}}(Y) = \frac{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)}{M(Y)}$$

is the relative variation of the lower mean $\bar{M}_{\bar{h}\cdot}(Y)$ with respect to the mean $M(Y)$, and $p_{\bar{h}\cdot} = \frac{P_{\bar{h}\cdot}}{N}$ is the cumulative relative frequency of the lower group with $\left\{Y \leq y_{\bar{h}}\right\}$. It is worth remarking that the quantity $V_{\bar{h}}(Y)$ is the Bonferroni pointwise measure of inequality at $\bar{h}$: for details see Zenga (2013), and Zenga and Valli (2016). It is also important to note that the formula (2) shows the Pietra index as the product of two factors: the first one, namely $V_{\bar{h}}(Y)$, is the economic distance between the lower mean $\bar{M}_{\bar{h}\cdot}(Y)$ and the total mean $M(Y)$; the second one, namely $p_{\bar{h}\cdot}$, is the relative weight associated to the units with amount less than or equal to the total mean $M(Y)$.

The Sect. 4 will show why this expression for the Pietra index is very suitable for the proposed decompositions by sources and by subpopulations.

## 4 Decomposition by sources

Let the variable $Y$ be the sum of $c$ variables $X_1, X_2, \dots, X_c$ that represent the sources. Using the same notation as that in the previous sections, let

$$Q_{\bar{h}\cdot}(X_j), \quad j = 1, 2, \dots, c$$

be the sum of the values assumed by the source $X_j$ on the $P_{\bar{h}\cdot}$ units belonging to the lower group $\{Y \leq y_{\bar{h}}\}$, let

$$\bar{M}_{\bar{h}\cdot}(X_j) = \frac{Q_{\bar{h}\cdot}(X_j)}{P_{\bar{h}\cdot}} \quad j = 1, 2, \dots, c$$

be the arithmetic mean of $X_j$ in the lower group, and let $M(X_j) = \frac{T(X_j)}{N}$ be the mean of $X_j$ in the whole population. As $Y = \sum_{j=1}^{c} X_j$, it follows that

$$M(Y) = \sum_{j=1}^{c} M(X_j) \text{ and } \bar{M}_{\bar{h}\cdot}(Y) = \sum_{j=1}^{c} \bar{M}_{\bar{h}\cdot}(X_j), \tag{3}$$

therefore the Pietra index is given by

$$\begin{aligned}\mathcal{P} =& \frac{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)}{M(Y)} \cdot \frac{P_{\bar{h}\cdot}}{N} \\ =& \sum_{j=1}^{c} \frac{M(X_j) - \bar{M}_{\bar{h}\cdot}(X_j)}{M(Y)} \cdot p_{\bar{h}\cdot} \\ =& \sum_{j=1}^{c} W_{\bar{h}\cdot}(X_j) p_{\bar{h}\cdot},\end{aligned}$$

where $W_{\bar{h}\cdot}(X_j)p_{\bar{h}\cdot}$ is the contribution of the source $X_j$ to $\mathcal{P}$, and indeed it is the relative difference $\frac{M(X_j) - \bar{M}_{\bar{h}\cdot}(X_j)}{M(Y)}$ times the relative frequency $\frac{P_{\bar{h}\cdot}}{N}$ of the lower group $\{Y \leq y_{\bar{h}}\}$. The relative contribution of the source $X_j$ to the value of the Pietra index is given by the ratio

$$\omega_{\bar{h}\cdot}(X_j) = \frac{W_{\bar{h}\cdot}(X_j)p_{\bar{h}\cdot}}{\mathcal{P}} = \frac{M(X_j) - \bar{M}_{\bar{h}\cdot}(X_j)}{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)} \tag{4}$$

with obviously $\sum_{j=1}^{c} \omega_{\bar{h}\cdot}(X_j) = 1$. By comparing these contributions and the shares

$$\gamma(X_j) = \frac{M(X_j)}{M(Y)} = \frac{T(X_j)}{T(Y)} \quad j = 1, 2, \dots, c \tag{5}$$

it is possible to understand whether a given source $X_j$ has an exacerbating or a mitigating impact on inequality (or heterogeneity) in the distribution of $Y$. In more detail, the quantity $\omega_{\bar{h}\cdot}(X_j) - \gamma(X_j)$ being positive means that the source $X_j$ plays an "increasing" role in terms of inequality (or heterogeneity), whereas it being negative means that the source $X_j$ "decreases" the inequality (or heterogeneity).

In real situations, the sources can also have negative values. When a variable assumes negative values, the use of any inequality index requires great attention; for example, see De Battisti et al. (2019) and Manero (2017) for further details about the Gini index. However, in the proposed decomposition of the Pietra index, the sources can assume also negative values, and in such a case attention is required only for the interpretation of the quantities $\omega_{\bar{h}\cdot}(X_j)$ and $\gamma(X_j)$ defined in (4) and (5), respectively. Finally, it is also worth remarking that the relative contribution of $X_j$ to the Pietra index is equal to those of the Gini, Bonferroni, and Zenga-2007 pointwise inequality measures at the cumulative relative frequency $p = p_{\bar{h}\cdot}$. For more details on this point, see Zenga (2013), Zenga and Valli (2017), and Pasquazzi and Zenga (2018).

## 5 Decomposition by subpopulations

The procedure for decomposing the Pietra index by subpopulations is based on the bivariate distribution of the $N$ units split into $k$ disjoint subpopulations (with $k \geq 2$). Such a distribution is reported in Table 1, where $n_{hl}$ denotes the frequency of the value $y_h$ (with $h = 1, 2, \ldots, r$) in subpopulation $l$ (with $l = 1, 2, \ldots, k$), and

$$n_{\cdot l} = \sum_{h=1}^{r} n_{hl}$$

is the size of the subpopulation $l$. Obviously:

$$\sum_{h=1}^{r} \sum_{l=1}^{k} n_{hl} = \sum_{l=1}^{k} \sum_{h=1}^{r} n_{hl} = \sum_{h=1}^{r} n_{h\cdot} = \sum_{l=1}^{k} n_{\cdot l} = N.$$

For the distribution $\{(y_h, n_{hl}),\ h = 1, 2, \cdots, r\}$ of subpopulation $l$, the analogs of the quantities $P_{h\cdot}$ and $Q_{h\cdot}$ are

$$P_{hl} = \sum_{t=1}^{h} n_{tl}$$

**Table 1** Bivariate distribution of the variable $Y$, according to the $k$ subpopulations

| $Y$ | Subpopulations | | | | Total |
|---|---|---|---|---|---|
| | 1 | 2 | $\cdots$ | $k$ | |
| $y_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1k}$ | $n_{1.}$ |
| $y_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2k}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| $y_r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rk}$ | $n_{r.}$ |
| *Total* | $n_{.1}$ | $n_{.2}$ | $\cdots$ | $n_{.k}$ | $N$ |

which is the cumulative frequency of $y_h$ in subpopulation $l$, and

$$Q_{hl}(Y) = \sum_{t=1}^{h} y_t n_{tl}$$

which is the sum of the values of the lower group $\{Y \leq y_h\}$ in subpopulation $l$. Moreover,

$$M_l(Y) = \frac{Q_{rl}(Y)}{n_{\cdot l}} = \frac{T_l(Y)}{n_{\cdot l}},$$

where $T_l(Y)$ is the sum of the $n_{.l}$ values of $Y$ in subpopulation $l$ and $M_l(Y)$ is the corresponding arithmetic mean. The lower mean $\bar{M}_{hl}(Y)$ of the variable $Y$ evaluated at $h$ in subpopulation $l$ can be defined as

$$\bar{M}_{hl}(Y) = \begin{cases} \min\{y_h \ : \ n_{hl} > 0\} & \text{if } P_{hl} = 0, \\ \frac{Q_{hl}(Y)}{P_{hl}} & \text{if } P_{hl} > 0. \end{cases}$$

In this definition of $\bar{M}_{hl}(Y)$ the lower mean of $P_{hl} = 0$ is prolonged by continuity, in analogy to the continuous case. Finally, the two ratios

$$\frac{n_{\cdot l}}{N} = \frac{\sum_{h=1}^{r} n_{hl}}{N} \qquad \text{and} \qquad p(l|h) = \frac{P_{hl}}{P_{h\cdot}}$$

can be defined: they are the relative frequencies of subpopulation $l$ in the whole population and in the lower group $\{Y \leq y_h\}$, respectively.

Among the mean $M(Y)$ and the means of the $k$ subpopulations $M_g(Y)$, it holds that

$$M(Y) = \sum_{g=1}^{k} M_g(Y) \cdot \frac{n_{\cdot g}}{N}, \tag{6}$$

and similarly, among the lower mean $\bar{M}_{h\cdot}(Y)$ and the $k$ lower means $\bar{M}_{hl\cdot}(Y)$, it holds that

$$\bar{M}_{h\cdot}(Y) = \sum_{l=1}^{k} \bar{M}_{hl}(Y) \cdot p(l|h). \tag{7}$$

By using (7) in expression (2) of the Pietra index, and by recalling that for any $h \in \{1, 2, \ldots r\}$ it holds that $\sum_{l=1}^{k} p(l|h) = 1$, it follows that

$$\begin{aligned}
\mathcal{P} =& \frac{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)}{M(Y)} \cdot p_{\bar{h}\cdot} \\
=& \frac{\sum_{l=1}^{k} \left[ M(Y)p(l|\bar{h}) - \bar{M}_{\bar{h}l}(Y)p(l|\bar{h}) \right]}{M(Y)} \cdot p_{\bar{h}\cdot} \\
=& \sum_{l=1}^{k} \frac{M(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot p(l|\bar{h}) \cdot p_{\bar{h}\cdot} \\
=& \sum_{l=1}^{k} V_{\bar{h}l\cdot}(Y) \cdot p_{\bar{h}\cdot},
\end{aligned}$$

where

$$V_{\bar{h}l\cdot}(Y) \cdot p_{\bar{h}\cdot} = \frac{M(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot p(l|\bar{h}) \cdot p_{\bar{h}\cdot} \tag{8}$$

can be interpreted as the contribution of subpopulation $l$ to the Pietra index, for $l \in \{1, 2, \dots k\}$. Using this decomposition procedure, it is thus possible to assess the contribution to the total value of $\mathcal{P}$ related to each single subpopulation, provided by (8). This result is important because, as already noted, other decomposition methods proposed in the literature cannot reach this goal.

For comparison, it is interesting to evaluate the relative contribution of subpopulation $l$ to the Pietra index, given by

$$\frac{V_{\bar{h}l\cdot}(Y) \cdot p_{\bar{h}\cdot}}{\mathcal{P}} = \frac{M(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)} \cdot p(l|\bar{h}).$$

As seen, this contribution depends on the ratio of two economic distances between the total mean of $Y$ and a lower mean (in the numerator related to subpopulation $l$, in the denominator related to the whole population), and on the relative frequency $p(l|\bar{h})$ of subpopulation $l$ in the lower group $\{Y \leq y_{\bar{h}}\}$.

Now, from (6) and the fact that

$$\sum_{g=1}^{k} \frac{n_{\cdot g}}{N} = 1,$$

it follows that

$$\begin{aligned}
\mathcal{P} &= \sum_{l=1}^{k} V_{\bar{h}l\cdot}(Y)\cdot p_{\bar{h}\cdot},\\
&= \sum_{l=1}^{k} \frac{M(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot}\\
&= \sum_{l=1}^{k}\sum_{g=1}^{k}\left[\frac{M_g(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\right]\cdot\frac{n_{\cdot g}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot}\\
&= \sum_{l=1}^{k}\sum_{g=1}^{k} V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot},
\end{aligned} \tag{9}$$

where

$$V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot} = \frac{M_g(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\cdot\frac{n_{\cdot g}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot} \tag{10}$$

is the contribution to $V_{\bar{h}l\cdot}$ related to the comparison between the lower mean $\bar{M}_{\bar{h}l}(Y)$ of subpopulation $l$ and the mean $M_g(Y)$ of subpopulation $g$.

In other words, (9) shows how the Pietra index can be split into a $k\times k$ matrix, according to the partition induced by the $k$ subpopulations.

The decomposition in (9) allows to further decompose the contribution of each subpopulation to the Pietra index into two quantities: the first one based on the comparison of means in the same subpopulation (which can be considered as *within* part), the second ones based on the comparison of means related to different subpopulations (which can be considered as *between* part). In effect

$$V_{\bar{h}l\cdot}(Y)\cdot p_{\bar{h}\cdot} = V_{\bar{h}ll}(Y)\cdot p_{\bar{h}\cdot} + \sum_{\{g:g\neq l\}}^{k} V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot},$$

with the *within* part of the contribution (due to subpopulation $l$) to the Pietra index given by

$$V_{\bar{h}ll}(Y)\cdot p_{\bar{h}\cdot} = \left[\frac{M_l(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\right]\cdot\frac{n_{\cdot l}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot}$$

and the *between* part of the contribution (due to subpopulation $l$) to the Pietra index equal to

$$\sum_{\{g:g\neq l\}}^{k} V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot} = \sum_{\{g:g\neq l\}}^{k}\left[\frac{M_g(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\right]\cdot\frac{n_{\cdot g}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot}.$$

At this point, it is clear the meaning of the two ratios

$$\frac{V_{\bar{h}ll}(Y)\cdot p_{\bar{h}\cdot}}{V_{\bar{h}l\cdot}(Y)\cdot p_{\bar{h}\cdot}} \qquad \text{and} \qquad \frac{\sum_{\{g:g\neq l\}}^{k} V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot}}{V_{\bar{h}l\cdot}(Y)\cdot p_{\bar{h}\cdot}},$$

which describe the weights of the *within* and *between* parts in the contribution of subpopulation $l$ to the Pietra index.

### 5.1 "Classical" decomposition into *within* and *between* components

From the decomposition represented in (9) it is also possible to reach the well-known decomposition into the *within* and *between* components. This is obtained by splitting the value of the Pietra index into two parts: the former ($\mathcal{P}_W$) based on mean comparisons of the same subpopulation, and the latter ($\mathcal{P}_B$) depending on comparison among means of different subpopulations:

$$\begin{aligned}
\mathcal{P} &= \sum_{l=1}^{k}\sum_{g=1}^{k}\left[\frac{M_g(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\right]\cdot\frac{n_{\cdot g}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot} \\
&= \sum_{l=1}^{k}\frac{M_l(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\cdot\frac{n_{\cdot l}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot} \\
&\quad + \sum_{l=1}^{k}\sum_{\{g:g\neq l\}}^{k}\frac{M_g(Y)-\bar{M}_{\bar{h}l}(Y)}{M(Y)}\cdot\frac{n_{\cdot g}}{N}\cdot p(l|\bar{h})\cdot p_{\bar{h}\cdot} \\
&= \sum_{l=1}^{k} V_{\bar{h}ll}(Y)\cdot p_{\bar{h}\cdot} + \sum_{l=1}^{k}\sum_{\{g:g\neq l\}}^{k} V_{\bar{h}lg}(Y)\cdot p_{\bar{h}\cdot} \\
&= \mathcal{P}_W + \mathcal{P}_B.
\end{aligned}$$

### 5.2 Comparison with other decompositions by subpopulations

As mentioned in Sect. 1, the literature contains two quite recent decompositions of the Pietra index by subpopulations. The first one, proposed by Frosini (2012), splits the Pietra index $\mathcal{P}$ into the sum of two terms, namely

$$\mathcal{P} = P_W^F + P_B^F.$$

In this decomposition:

- $P_W^F$ is the *within* component, defined as

$$P_W^F = \sum_{l=1}^{k}\frac{T_l(Y)}{T(Y)}\mathcal{P}_l, \tag{11}$$

  where $\mathcal{P}_l$ is the Pietra index of $Y$ in subpopulation $l$,

$$\mathcal{P}_l = \frac{\sum_{h=1}^{r} |y_h - M_l(Y)| \cdot n_{hl}}{2M_l(Y)n_{.l}} \qquad l = 1, \dots, k;$$

- $P_B^F$ is the *between* component, and it is the sum of the two quantities $P_{Bt}$ and $P_{Bm}$:

  (i) $P_{Bt}$ is the mixture effect and is equal to

$$\begin{aligned} P_{Bt} =& \frac{1}{T(Y)} \left[ \sum_{M_l(Y)>M(Y)} \sum_{M(Y)<y_h \le M_l(Y)} [y_h - M_l(Y)] n_{hl} + \right. \\ & \left. - \sum_{M_l(Y)<M(Y)} \sum_{M_l(Y)<y_h \le M(Y)} [y_h - M_l(Y)] n_{hl} \right], \end{aligned}$$

  (ii) $P_{Bm}$ is the mean effect, given by

$$\begin{aligned} P_{Bm} =& \frac{1}{2T(Y)} \sum_{l=1}^{k} P_{Bm}^l \\ =& \frac{1}{2T(Y)} \sum_{l=1}^{k} \left[ \sum_{h=\bar{h}+1}^{r} n_{hl} - \left( \sum_{h=1}^{\bar{h}} n_{hl} - K_l \right) \right] [M_l(Y) - M(Y)], \end{aligned}$$

  where

$$K_l = \begin{cases} n_{h_0 l} & \text{if } \exists\ h_0 \in \{1, \dots, r\} \text{ such that } y_{h_0} = M(Y) \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, in Frosini (2012), the contribution $D_l$ to the Pietra index $\mathcal{P}$ due to subpopulation $l$, can be computed as follows:

$$D_l = \begin{cases} \frac{T_l(Y)}{T(Y)} \mathcal{P}_l - \frac{1}{T(Y)} \sum_{M_l(Y)<y_h \le M(Y)} [y_h - M_l(Y)] n_{hl} - \frac{P_{Bm}^l}{2T(Y)} & \text{if } M_l(Y) < M(Y) \\ \frac{T_l(Y)}{T(Y)} \mathcal{P}_l & \text{if } M_l(Y) = M(Y) \\ \frac{T_l(Y)}{T(Y)} \mathcal{P}_l + \frac{1}{T(Y)} \sum_{M(Y)<y_h \le M_l(Y)} [y_h - M_l(Y)] n_{hl} + \frac{P_{Bm}^l}{2T(Y)} & \text{if } M_l(Y) > M(Y). \end{cases}$$

The procedure proposed by Frosini (2012) can be useful, even if the *between* component $P_B^F$ requires caution in the interpretation, given that it is the sum of two quantities related to different effects and its meaning is therefore neither very intuitive nor immediate. The possibility to assess the contribution due to each subpopulation is surely a worthy characteristic of this procedure.

The second of the aforementioned decompositions was proposed by Habib (2012). Starting from the definitions of

- the overall variation

$$d_{hl} = \begin{cases} [y_h - M(Y)]n_{hl} & \text{if } y_h > M(Y) \\ 0 & \text{otherwise;} \end{cases}$$

- the *within* variation

$$w_{hl} = \begin{cases} [y_h - M_l(Y)]n_{hl} & \text{if } y_h > M_l(Y) \\ 0 & \text{otherwise;} \end{cases}$$

- the *between* variation

$$z_l = \begin{cases} M_l(Y) - M(Y) & \text{if } M_l(Y) > M(Y) \\ 0 & \text{otherwise,} \end{cases}$$

the following decomposition of the Pietra index $\mathcal{P}$ is obtained

$$\mathcal{P} = \tilde{P}_W + \tilde{P}_B + \tilde{P}_E,$$

where:

- $\tilde{P}_W$ is the *within* component, defined as

$$\tilde{P}_W = \sum_{l=1}^{k} \frac{n_{\cdot l}}{N} \cdot \frac{M_l(Y)}{M(Y)} \cdot \mathcal{P}_l; \tag{12}$$

- $\tilde{P}_B$ is the *between* component

$$\tilde{P}_B = \frac{\sum_{l=1}^{k} z_l}{\sum_{l=1}^{k} M_l(Y)} \cdot \frac{\sum_{l=1}^{k} M_l(Y)}{k \cdot M(Y)} \cdot \sum_{l=1}^{k} \frac{n_{\cdot l}}{N} \cdot \frac{z_l \cdot k}{\sum_{l=1}^{k} z_l};$$

- $\tilde{P}_E$ is the (error) remaining term

$$\tilde{P}_E = \frac{1}{NM(Y)} \sum_{l=1}^{k} \sum_{h=1}^{r} d_{hl} - \sum_{l=1}^{k} \frac{n_{\cdot l}}{N} \left[ \frac{\sum_{h=1}^{r} w_{hl}}{n_{\cdot l} M(Y)} + \frac{z_l}{M(Y)} \right].$$

In this decomposition, it is easy to see that the *within* component $\tilde{P}_W$ defined in (12) coincides with the corresponding $P_W^F$ defined in (11), given that by definition it holds that

$$\frac{T_l(Y)}{T(Y)} = \frac{n_{\cdot l}}{N} \cdot \frac{M_l(Y)}{M(Y)}.$$

As in the previous procedure, the interpretation of the *between* component $\tilde{P}_B$ is not very immediate, and the presence of the quantity $\tilde{P}_E$, required to rescale the sum of $\tilde{P}_W$ and $\tilde{P}_B$ into the interval [0, 1], does not facilitate that task. To simplify the decomposition, Habib (2012) proposes removing the third term $\tilde{P}_E$ by dividing it into two parts to be summed to $\tilde{P}_W$ and $\tilde{P}_B$, arguing that "the separation of the error term to within-groups and between-groups errors could be based on a proportionate of each error to the total error" (Habib 2012). In other words, he proposes splitting $\tilde{P}_E$ into the sum of $\tilde{P}_{E_W}$ and $\tilde{P}_{E_B}$, to obtain the so-called perfect decomposition

$$\mathcal{P} = P_W^H + P_B^H,$$

where

$$P_W^H = \tilde{P}_W + \tilde{P}_{E_W} \qquad \text{and} \qquad P_B^H = \tilde{P}_B + \tilde{P}_{E_B}.$$

However, dividing $\tilde{P}_E$ into the sum of $\tilde{P}_{E_W}$ and $\tilde{P}_{E_B}$ with no objective rule for determining how the splitting must be performed seems very arbitrary, and it makes the interpretations of $P_W^H$ and $P_W^B$ less intuitive and more problematic.

## 6 Applications to two actual sport datasets

In this section, two applications of the proposed decomposition procedures are presented, both regarding professional football teams in Italy. The first one deals with their balance-sheet data, while the second one deals with the market values of all their players.

### 6.1 Decomposition by sources

Serie A is the most important football league in Italy, with $N = 20$ professional teams. As in other European countries, the correctness of the balance sheets of the football teams has become increasingly important in recent years, leading to burgeoning studies on this topic; for example, see PwC (2018) and KPMG (2019). Within this framework, the following analysis is provided. For each Serie A team, the following five balance-sheet variables are considered:

- $Y$ = Total assets;
- $X_1$ = Cash;
- $X_2$ = Total accounts receivable;
- $X_3$ = Inventories, short-term investments and other current assets;
- $X_4$ = Net property (tangible and intangible assets), investments and advances, other assets.

All the variables are in millions of Euros and refer to fiscal year 2018. The data are from https://www.analisiaziendale.it and are repeated in the provided supplementary material. The Total assets is the investigated variable, while $X_j$, (with $j = 1, 2, 3, 4$) are the four sources, given that $Y = \sum_{j=1}^{4} X_j$. Table 2 provides some descriptive statistics about all the variables, and Fig. 1 shows their boxplots.

Clearly, Net property, investments and advances, other assets ($X_4$) is the most relevant source: its mean and mean absolute deviation (MAD) are the highest and the most comparable to those of Total assets ($Y$).

**Table 2** Descriptive statistics of the variables considered for the decomposition by sources

| | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| Min | 6.281 | 0.037 | 1.752 | 0.010 | 2.452 |
| Max | 772.669 | 110.694 | 125.934 | 21.483 | 619.283 |
| Median | 160.730 | 3.000 | 51.054 | 2.044 | 95.309 |
| Mean | 233.701 | 11.992 | 56.237 | 3.995 | 161.477 |
| Mean abs. dev. (MAD) | 167.517 | 14.186 | 35.587 | 3.5420 | 128.439 |
| Pietra Index $\mathcal{P}$ | 0.3584 | 0.5915 | 0.3164 | 0.4433 | 0.3977 |

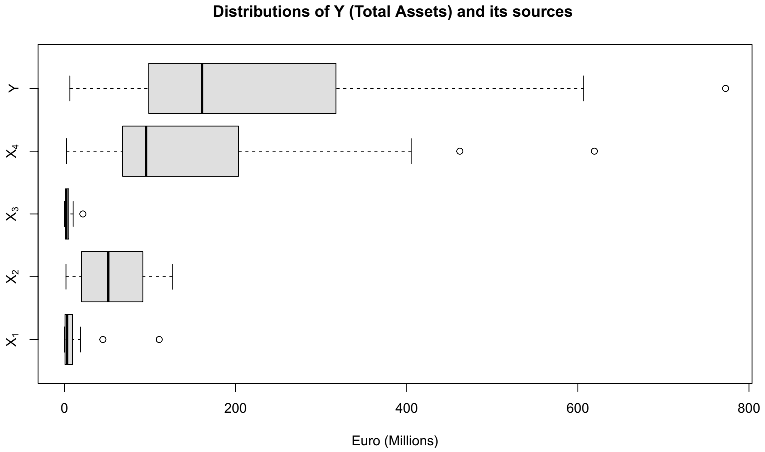

**Fig. 1** Boxplots of the variables considered for the decomposition by sources

**Table 3** The contributions of the sources to the Pietra index $\mathcal{P}$

| Source | | $W_{\bar{h}\cdot}(X_j)p_{\bar{h}\cdot}$ | $\omega_{\bar{h}\cdot}(X_j)$ | $\gamma(X_j)$ | $\omega_{\bar{h}\cdot}(X_j)-\gamma(X_j)$ |
|---|---|---|---|---|---|
| Cash | $(X_1)$ | 0.0228 | 0.0636 | 0.0513 | 0.0123 |
| Total accounts receivable | $(X_2)$ | 0.0660 | 0.1842 | 0.2406 | -0.0564 |
| Inventories, short-term invest. and other current assets | $(X_3)$ | 0.0048 | 0.0134 | 0.0171 | -0.0036 |
| Net property, invest. and advances, other assets | $(X_4)$ | 0.2648 | 0.7388 | 0.6910 | 0.0478 |

The value of the Pietra index of $Y$ is 0.3584, denoting a medium level of heterogeneity. The four contributions obtained by the proposed decomposition by sources are stored in Table 3.

As expected, the highest contribution is due to the source Net property, investments and advances, other assets ($X_4$), which represents the most part of $\mathcal{P}$ (73.88%). Total accounts receivable ($X_2$) follows with 18.42%, then Cash ($X_1$) with 6.36%. The source

Inventories, short-term investments and other current assets ($X_3$) is the last one, with a negligible contribution of 1.34%.

The comparison of the differences reported in the last column of Table 3 shows that the sources Cash and Net property, investments and advances, other assets exacerbate the heterogeneity of Total assets, while the other two play a mitigating role. As a final remark, it can be argued that the heterogeneity in the distribution of $Y$ is due mainly to the source Net property, investments and advances, other assets, given that this variable includes intangible assets, to which a percentage of the values of the team players is allocated: as analyzed in the following application, these values differ considerably also among teams in the same league.

## 6.2 Decomposition by subpopulations

This second application concerns not just Serie A teams but all Italian professional football teams in the three existing leagues, namely Serie A, B, and C, with $n_{\cdot 1} = 20$, $n_{\cdot 2} = 19$, and $n_{\cdot 3} = 59$ teams, respectively. In fact, Serie C is divided territorially into three more subgroups, but in this application they are considered all together. For each team, the variable $Y =$ "Value (in millions of Euro) of the team players in November 2018", which is the sum of the market values of all the team players, is investigated. The data are available online at https://www.transfermarkt.it. The three leagues are the three subpopulations, and the purpose of this analysis is to investigate how the heterogeneity of $Y$ is split across the three leagues and to assess the heterogeneity levels between the leagues and within each one. Table 4 provides some statistical indicators for the variable $Y$ regarding the three subpopulations and the whole population. The calculations give $\bar{h} = \max\{h \ : \ y_h \leq M(Y)\} = 80$, and therefore $p_{\bar{h}\cdot} = \frac{80}{98} = 0.8163$. The lower means $\bar{M}_{\bar{h}l}$ and the relative frequencies $p(l|\bar{h})$ needed for the calculations are also given in Table 4. The complete dataset can be found in the provided supplementary material.

**Table 4** Descriptive statistics of the variable $Y$ for the three subpopulations and for the whole population

| | Subpopulations | | | |
|---|---|---|---|---|
| | Serie A | Serie B | Serie C | Total |
| | $l = 1$ | $l = 2$ | $l = 3$ | |
| Size $n_{\cdot l}$ | 20 | 19 | 59 | 98 |
| Min | 38.05 | 8.38 | 0.675 | 0.675 |
| Max | 797.8 | 37.38 | 8.9 | 797.8 |
| Mean $M_l(Y)$ | 234.986 | 17.556 | 4.017 | 53.778 |
| Median | 146.765 | 15.85 | 3.8 | 5.155 |
| Mean abs. dev. (MAD) | 171.023 | 5.442 | 1.039 | 74.483 |
| Pietra index $\mathcal{P}$ | 0.3639 | 0.1550 | 0.1293 | 0.6925 |
| $\bar{M}_{\bar{h}l}$ | 40.825 | 17.556 | 4.017 | - |
| $p(l\|\bar{h})$ | 0.025 | 0.2375 | 0.7375 | 1 |

An overview of the data shows that these three subpopulations are almost non-overlapping, given that all the values of teams in Serie A are higher than those of all the teams in the other two subpopulations and only one team in Serie C has a value higher than that of the poorest team in Serie B. The three subpopulations are therefore very different, as shown by the orders of magnitude of all the location indexes and by the variability indicator, the mean absolute deviation (MAD). It is also interesting to note that the values of $Y$ for all the teams in Serie B and C are lower than the total mean $M(Y) = 53.778$.

Direct computation shows that for the whole population of the 98 professional football teams, the mean absolute deviation is $S_{M(Y)} = 74.483$, therefore the Pietra index is given by

$$\mathcal{P} = \frac{74.483}{2 \cdot 53.778} = 0.6925.$$

The proposed procedure allows to decompose the Pietra index as

$$\mathcal{P} = \sum_{l=1}^{3} \sum_{g=1}^{3} V_{\bar{h}lg}(Y) \cdot p_{\bar{h}\cdot}.$$

The values $V_{\bar{h}lg}(Y) \cdot p_{\bar{h}\cdot}$ are the entries of the 3 × 3 decomposition-by-subpopulations matrix, reported in Table 5. Interestingly, note that in this application, the two values

$$V_{\bar{h}22}(Y) \cdot p_{\bar{h}\cdot} = V_{\bar{h}33}(Y) \cdot p_{\bar{h}\cdot} = 0,$$

since

$$V_{\bar{h}22}(Y) = V_{\bar{h}33}(Y) = 0.$$

This can be proved by replacing data in Table 4 by their definitions, given by

$$\begin{aligned} V_{\bar{h}22}(Y) =& M_2(Y) - \bar{M}_{80,2} = 17.556 - 17.556 = 0, \\ V_{\bar{h}33}(Y) =& M_3(Y) - \bar{M}_{80,3} = 4.017 - 4.017 = 0. \end{aligned}$$

To interpret the obtained values, the quantity

**Table 5** The 3 × 3 decomposition by subpopulations matrix, and the contributions of each subpopulation split in *within* and *between* parts

| | Serie A | Serie B | Serie C | |
|---|---|---|---|---|
| | $l = 1$ | $l = 2$ | $l = 3$ | |
| Serie A $g = 1$ | 0.0150 | 0.1599 | 0.5277 | |
| Serie B $g = 2$ | −0.0017 | 0 | 0.0294 | |
| Serie C $g = 3$ | −0.0084 | −0.0294 | 0 | |
| $V_{80l\cdot}(Y) \cdot p_{80\cdot}$ | 0.0049 (0.71%) | 0.1305 (18.84%) | 0.5571 (80.45%) | 0.6925 (100%) |
| *Within* part | 0.0150 | 0 | 0 | 0.0150 (2.2%) |
| *Between* part | −0.0101 | 0.1305 | 0.5571 | 0.6775 (97.8%) |

$$V_{\bar{h}31}(Y) \cdot p_{\bar{h}\cdot} = \frac{M_1(Y) - \bar{M}_{\bar{h}3}(Y)}{M(Y)} \cdot \frac{n_{\cdot 1}}{N} \cdot p(3|\bar{h}) \cdot p_{\bar{h}\cdot} = 0.5277$$

shows that the average value of the teams in Serie A is much greater than the lower mean of the teams in Serie C, and therefore it allows to assess a relevant "economic distance" between the two subpopulations Serie A and C.

A minor distance is registered between Serie A and Serie B, since

$$V_{\bar{h}21}(Y) \cdot p_{\bar{h}\cdot} = \frac{M_1(Y) - \bar{M}_{\bar{h}2}(Y)}{M(Y)} \cdot \frac{n_{\cdot 1}}{N} \cdot p(2|\bar{h}) \cdot p_{\bar{h}\cdot} = 0.1599,$$

and a very low one is registered between Serie B and Serie C given that

$$V_{\bar{h}32}(Y) \cdot p_{\bar{h}\cdot} = \frac{M_2(Y) - \bar{M}_{\bar{h}3}(Y)}{M(Y)} \cdot \frac{n_{\cdot 2}}{N} \cdot p(3|\bar{h}) \cdot p_{\bar{h}\cdot} = 0.0294.$$

The value

$$V_{\bar{h}12}(Y) \cdot p_{\bar{h}\cdot} = \frac{M_2(Y) - \bar{M}_{\bar{h}1}(Y)}{M(Y)} \cdot \frac{n_{\cdot 2}}{N} \cdot p(1|\bar{h}) \cdot p_{\bar{h}\cdot} = -0.0017$$

and the other negative ones ($V_{\bar{h}13}(Y) \cdot p_{\bar{h}\cdot} = -0.0084$; $V_{\bar{h}23}(Y) \cdot p_{\bar{h}\cdot} = -0.0294$) show that the average value of the teams in Serie B is less than the lower mean of the teams in Serie A, and the average value of the teams in Serie C is less than both the lower means of the other two leagues.

The aggregated values in the last rows of Table 5 show that the subpopulation with the largest contribution to the Pietra index is Serie C (with 80.45%), and the one with the smallest contribution is Serie A (with 0.71%). A reasonable cause for this can be identified also in the very high weight of Serie C in the whole population (since $p(3|\bar{h}) = 73.75\%$). Table 5 also provides the *within* and *between* parts of the contribution of each subpopulation. In this application, given that the quantities $V_{\bar{h}22}(Y)$ and $V_{\bar{h}33}(Y)$ are zero, Serie B and C do not contribute to the *within* component of the Pietra index, and the heterogeneity due to those is carried into the *between* component. The *within* part of Serie A coincides with the *within* component of $\mathcal{P}$.

As special case, also the decomposition of the Pietra index into the *within* and *between* components can be obtained. The former is the sum of the entries in the main diagonal of Table 5, while the latter is the sum of all the remaining ones:

$$\mathcal{P}_W = 0.0150 \quad \text{and} \quad \mathcal{P}_B = 0.6775.$$

From these, it is interesting to evaluate the two ratios

$$\frac{\mathcal{P}_W}{\mathcal{P}} = \frac{0.0150}{0.6925} = 0.022 \quad \text{and} \quad \frac{\mathcal{P}_B}{\mathcal{P}} = \frac{0.6775}{0.6925} = 0.978,$$

which show that in this application the *between* component of the Pietra index is far more relevant than the *within* one. The *between* component represents the 97.8% of the total, while the *within* one represents only 2.2%. This result also shows that the disparity across the three leagues counts for much more than that into the leagues.

Now, the other two decomposition procedures reported in the previous sections, are applied and the results compared with those already obtained. Applying the decomposition proposed by Frosini (2012) to the examined dataset gives the following values:

$$P_W^F = 0.3401 \quad \text{and} \quad P_B^F = 0.3524,$$

with

$$P_{Bt} = -0.2664 \quad \text{and} \quad P_{Bm} = 0.6188.$$

The computation of the two ratios

$$\frac{P_W^F}{\mathcal{P}} = \frac{0.3401}{0.6925} = 0.491 \quad \text{and} \quad \frac{P_B^F}{\mathcal{P}} = \frac{0.3524}{0.6925} = 0.509$$

allows to argue that in this decomposition, the two components (*within* and *between*) are quite balanced, since they represent 49.1% and 50.9% of the total, respectively. This result is obtained even if the subpopulations are almost non-overlapping and have very different means, as already remarked.

Table 6 summarizes the contributions of each subpopulation according to the procedure proposed by Frosini (2012).

Unlike with the previous decomposition, here the subpopulation with the lowest contribution to the Pietra index is Serie B (9.43%). This conclusion is motivated neither by the order of magnitude of the values of $Y$ in that league nor by the relative weight of that league in the whole population, which is very close to that of Serie A.

Computating the decomposition proposed by Habib (2012) gives the following values:

$$\tilde{P}_W = 0.3401, \quad \tilde{P}_B = 0.6876 \quad \text{and} \quad \tilde{P}_E = -0.3352,$$

where the *within* component $\tilde{P}_W$ coincides with $P_W^F$, as already remarked. The presence of the third term $\tilde{P}_E$ makes the interpretation of the results not very intuitive. Then, following the same approach used in the application in Habib (2012), $\tilde{P}_E$ can be divided equally into the two errors $\tilde{P}_{E_W}$ and $\tilde{P}_{E_B}$, giving

**Table 6** The contributions of the three subpopulations in the decomposition proposed by Frosini (2012)

| | Serie A | Serie B | Serie C | |
|---|---|---|---|---|
| | $l = 1$ | $l = 2$ | $l = 3$ | |
| $D_l$ | 0.3487 (50.35%) | 0.0653 (9.43%) | 0.2785 (40.22%) | 0.6925 (100%) |

$$P_W^H = 0.1725 \quad \text{and} \quad P_B^H = 0.5200.$$

The relative weights of these two components are therefore

$$\frac{P_W^H}{\mathcal{P}} = \frac{0.1725}{0.6925} = 0.249 \quad \text{and} \quad \frac{P_B^H}{\mathcal{P}} = \frac{0.52}{0.6925} = 0.751.$$

These values show that it is the *between* component that is most relevant, given that it represents 75.1% of the total, while the *within* component represents the remaining 24.9%. It is worth recalling that the lack of a rule governing how to split $\tilde{P}_E$ into $P_{E_W}$ and $P_{E_B}$ is a non-negligible point of weakness of this procedure.

## 7 Conclusions and final remarks

In this paper, two innovative procedures for decomposing the Pietra index are introduced based on an alternative expression for this "evergreen" measure. The decomposition by sources allows to obtain the contribution related to each source. The decomposition by subpopulations allows to assess how each subpopulation contributes to the value of the index, also by assessing the *within* and *between* parts in each contribution. By using a different aggregation, it is also easy to obtain the classical decomposition of the Pietra index into the *within* and *between* components. Because of their very fine decomposition levels, these two procedures are very innovative and provide researchers with more information than do many others available in the literature for the Pietra index. The two presented applications add interesting details about Italian professional football teams: the first one shows how the heterogeneity of the Total assets in Serie A teams can be split among its sources, while the second one highlights how much of the disparity in team values is due to each of the three considered leagues.

## Appendix: Computation and decompositions of the Pietra index

This appendix exemplifies calculating the Pietra index and the proposed decompositions. To show the flexibility of the proposed procedures, these example data are deliberately very different from those used in the applications described in the main paper. Consider the dataset in Table 7: there are $N = 20$ units from $k = 3$ different subpopulations, and the value of the variable $Y$ is given by the sum of $c = 3$ sources ($X_1$, $X_2$, and $X_3$). The last three columns describe to which subpopulation each unit belongs.

First, consider the distribution of the variable $Y$ as given in Table 8. Using the same notation as that in the main paper, we have $r = 6$ and $N = 20$. Straightforward calculations give.

From straightforward calculations, it derives that

$$M(Y) = \frac{T(Y)}{N} = \frac{600}{20} = 30, \quad \text{and} \quad S_{M(Y)} = \frac{380}{20} = 19,$$

meaning that the Pietra index is:

**Table 7** Original data matrix

| $i$ | Sources | | | $Y$ | Subpopulations | | |
|---|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | | 1 | 2 | 3 |
| 1 | 6 | 4 | 0 | 10 | 0 | 1 | 0 |
| 2 | 6 | 3 | 1 | 10 | 0 | 0 | 1 |
| 3 | 8 | 1 | 1 | 10 | 0 | 0 | 1 |
| 4 | 8 | 7 | 0 | 15 | 0 | 1 | 0 |
| 5 | 2 | 13 | 0 | 15 | 0 | 0 | 1 |
| 6 | 4 | 8 | 3 | 15 | 1 | 0 | 0 |
| 7 | 4 | 10 | 1 | 15 | 1 | 0 | 0 |
| 8 | 8 | 6 | 1 | 15 | 1 | 0 | 0 |
| 9 | 9 | 6 | 0 | 15 | 0 | 0 | 1 |
| 10 | 13 | 8 | 1 | 22 | 0 | 0 | 1 |
| 11 | 11 | 10 | 1 | 22 | 1 | 0 | 0 |
| 12 | 15 | 5 | 2 | 22 | 1 | 0 | 0 |
| 13 | 12 | 10 | 0 | 22 | 1 | 0 | 0 |
| 14 | 13 | 7 | 2 | 22 | 1 | 0 | 0 |
| 15 | 20 | 0 | 16 | 36 | 0 | 1 | 0 |
| 16 | 16 | 0 | 20 | 36 | 1 | 0 | 0 |
| 17 | 14 | 0 | 22 | 36 | 1 | 0 | 0 |
| 18 | 40 | 10 | 16 | 66 | 0 | 1 | 0 |
| 19 | 8 | 22 | 36 | 66 | 0 | 0 | 1 |
| 20 | 59 | 30 | 41 | 130 | 1 | 0 | 0 |
| Total | 276 | 160 | 164 | 600 | 10 | 4 | 6 |

**Table 8** Frequency distribution of the variable $Y$

| $y_h$ | 10 | 15 | 22 | 36 | 66 | 130 | Total |
|---|---|---|---|---|---|---|---|
| $n_{h\cdot}$ | 3 | 6 | 5 | 3 | 2 | 1 | 20 |

$$\mathcal{P} = \frac{19}{2 \cdot 30} = 0.3166.$$

The set $\left\{y_h \,:\, y_h \le M(Y)\right\}$ has cardinality 3:

$$\left\{y_h \,:\, y_h \le M(Y)\right\} = \left\{y_1, y_2, y_3\right\} = \{10, 15, 22\},$$

and $\bar{h} = \max\left\{h \,:\, y_h \le M(Y)\right\} = 3$. Thus,

$$P_{\bar{h}\cdot} = P_{3\cdot} = \sum_{h=1}^{3} n_{h\cdot} = 3 + 6 + 5 = 14,$$

and

**Table 9** Calculations and coordinates of the Lorenz curve

| $h$ | $y_h$ | $n_{h\cdot}$ | $y_h n_{h\cdot}$ | $P_{h\cdot}$ | $Q_{h\cdot}$ | $p_{h\cdot}$ | $L(p_{h\cdot})$ | $p_{h\cdot} - L(p_{h\cdot})$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 3 | 30 | 3 | 30 | 0.15 | 0.05 | 0.10 |
| 2 | 15 | 6 | 90 | 9 | 120 | 0.45 | 0.20 | 0.25 |
| 3 | 22 | 5 | 110 | 14 | 230 | 0.70 | 0.3833 | 0.3166 |
| 4 | 36 | 3 | 108 | 17 | 338 | 0.85 | 0.563 | 0.2865 |
| 5 | 66 | 2 | 132 | 19 | 470 | 0.95 | 0.783 | 0.166 |
| 6 | 130 | 1 | 130 | 20 | 600 | 1 | 1 | 0 |

$$Q_{\bar{h}\cdot}(Y) = Q_{3\cdot}(Y) = \sum_{h=1}^{3} y_h n_{h\cdot} = 30 + 90 + 110 = 230$$

are the cumulative frequency and the cumulative sum of values corresponding to $\bar{h}$, respectively. In the row identified by $h = 3$, Table 9 reports the cumulative relative frequency

$$p_{\bar{h}\cdot} = p_{3\cdot} = \frac{14}{20} = 0.70,$$

the corresponding value of the Lorenz curve

$$L\left(p_{3\cdot}\right) = \frac{230}{600} = 0.3833,$$

and the difference

$$p_{3\cdot} - L\left(p_{3\cdot}\right) = 0.3166,$$

thereby confirming the property reported in Sect. 2, that $\mathcal{P} = \tilde{p} - L(\tilde{p})$, where $\tilde{p}$ is such that $y_{(\tilde{p})} = M(Y)$. In Fig. 2 the Lorenz curve of the variable $Y$ is shown and the segment of extremes $(p_{3\cdot}, L(p_{3\cdot}))$, whose length is the value of the Pietra index, is highlighted.

### Decomposition by sources

The proposed decomposition of the Pietra index is now applied to the distribution of $Y$, by considering the $c = 3$ sources. For them, it follows from Table 7 that

$$T(X_1) = 276, \quad T(X_2) = 160, \quad T(X_3) = 164,$$

and

$$Q_{3\cdot}(X_1) = 119, \quad Q_{3\cdot}(X_2) = 98, \quad Q_{3\cdot}(X_3) = 13.$$

Then:

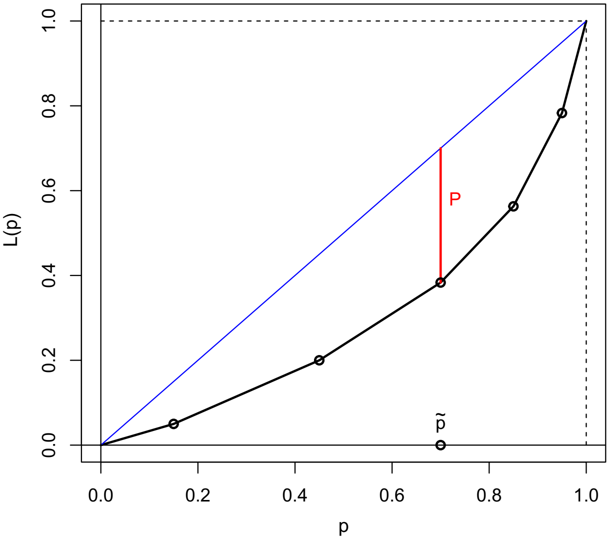

**Fig. 2** The Lorenz curve for the variable *Y*, and the geometrical representation of the Pietra index $\mathcal{P}$

$$M(X_1) = \frac{276}{20} = 13.8, \quad M(X_2) = \frac{160}{20} = 8, \quad M(X_3) = \frac{164}{20} = 8.2,$$
$$\bar{M}_{3\cdot}(X_1) = \frac{119}{14} = 8.5, \quad \bar{M}_{3\cdot}(X_2) = \frac{98}{14} = 7, \quad \bar{M}_{3\cdot}(X_3) = \frac{13}{14} = 0.9285.$$

The calculations for the decomposition by sources are reported in Table 10. The fifth column contains the contributions of the sources (0.1236, 0.0233, and 0.1697) to the Pietra index, and the sixth one contains $\omega_{\bar{h}\cdot}(X_j)$.

Comparing the values in the last column, shows that the first two sources ($X_1$ and $X_2$) exacerbate the inequality (since the differences $\omega_{3\cdot}(X_j) - \gamma(X_j)$ are negative), while $X_3$ has a mitigating impact because the difference $\omega_{3\cdot}(X_3) - \gamma(X_3)$ is positive.

## Decomposition by subpopulations

The proposed decomposition of the Pietra index by subpopulation is now applied to the considered distribution. The data regarding the subpopulations of Table 7 are summarized in Table 11.

The means of *Y* in the three subpopulations are

**Table 10** Calculations for the decomposition by sources

| $j$ | $M(X_j) - \bar{M}_{3\cdot(X_j)}$ | $W_{3\cdot}(X_j)$ | $p_{3\cdot}$ | $W_{3\cdot}(X_j)p_{3\cdot}$ | $\omega_{3\cdot}(X_j)$ | $\gamma(X_j)$ | $\omega_{3\cdot}(X_j) - \gamma(X_j)$ |
|---|---|---|---|---|---|---|---|
| 1 | 5.30 | 0.1766 | 0.70 | 0.1236 | 0.390 | 0.460 | −0.070 |
| 2 | 1 | 0.0333 | 0.70 | 0.0233 | 0.074 | 0.267 | −0.193 |
| 3 | 7.2715 | 0.2424 | 0.70 | 0.1697 | 0.536 | 0.273 | 0.263 |

**Table 11** Bivariate distribution of the variable $Y$, according to the $k = 3$ subpopulations

| $Y$ | Subpopulations | | | Total |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 10 | 0 | 1 | 2 | 3 |
| 15 | 3 | 1 | 2 | 6 |
| 22 | 4 | 0 | 1 | 5 |
| 36 | 2 | 1 | 0 | 3 |
| 66 | 0 | 1 | 1 | 2 |
| 130 | 1 | 0 | 0 | 1 |
| Total | 10 | 4 | 6 | 20 |

$$M_1(Y) = 33.5, \quad M_2(Y) = 31.75, \quad M_3(Y) = 23,$$

and the lower means $\bar{M}_{\bar{h}l}(Y)$ at $\bar{h} = 3$ are

$$\bar{M}_{31}(Y) = \frac{133}{7} = 19, \quad \bar{M}_{32}(Y) = \frac{25}{2} = 12.5, \quad \bar{M}_{33}(Y) = \frac{72}{5} = 14.4.$$

The values of $p(l|\bar{h})$ are given by

$$p(1|3) = \frac{7}{14} = 0.5, \quad p(2|3) = \frac{2}{14} = 0.1429, \quad p(3|3) = \frac{5}{14} = 0.3571,$$

and the relative frequencies of the subpopulations are:

$$\frac{n_{\cdot 1}}{N} = \frac{10}{20} = 0.5, \quad \frac{n_{\cdot 2}}{N} = \frac{4}{20} = 0.2, \quad \frac{n_{\cdot 3}}{N} = \frac{6}{20} = 0.3.$$

As seen before, $p_{\bar{h}\cdot} = p_{3\cdot} = 0.7$. The 3 × 3 decomposition-by-subpopulations matrix with entries given by the quantities $V_{3l_g}(Y) \cdot p_{3\cdot}$ is given in Table 12.

From that, the *within* and *between* components of the Pietra index are obtained as

$$\mathcal{P}_W = 0.1188 \quad \text{and} \quad \mathcal{P}_B = 0.1978,$$

which represent 37.52% and 62.48% of $\mathcal{P}$, respectively.

**Table 12** The 3 × 3 decomposition-by-subpopulations matrix, and the contributions of each subpopulation split in *within* and *between* parts

| | $l = 1$ | $l = 2$ | $l = 3$ | |
|---|---|---|---|---|
| $g = 1$ | 0.0845 | 0.0350 | 0.0796 | |
| $g = 2$ | 0.0298 | 0.0128 | 0.0289 | |
| $g = 3$ | 0.0140 | 0.0105 | 0.0215 | |
| $V_{3l\cdot}(Y) \cdot p_{3\cdot}$ | 0.1283 (40.52%) | 0.0583 (18.41%) | 0.1300 (41.07%) | 0.3166 (100%) |
| *Within* part | 0.0845 | *0.0128* | 0.0215 | 0.1188 (37.52%) |
| *Between* part | 0.0438 | 0.0455 | 0.1085 | 0.1978 (62.48%) |

**Supplementary Information** The online version supplementary material available at https://doi.org/10.1007/s10182-021-00397-6.

**Funding** Open access funding provided by Università degli Studi di Sassari within the CRUI-CARE Agreement.

## References

Alker, H., Russett, B.: On measuring inequality. Behav. Sci. **9**, 207–218 (1964)

Beck, A.F., Moncrief, T., Huang, B., Simmons, J.M., Sauers, H., Chen, C., Kahn, R.S.: Inequalities in neighborhood child asthma admission rates and underlying community characteristics in one US county. J. Pediatr. **163**(2), 574–580 (2013)

Bresciani Turroni, C.: Sull'interpretazione e comparazione di seriazioni di redditi o di patrimoni. Giornale degli Economisti **34**, 13–47 (1907)

Castagnoli, E., Muliere, P.: A note on inequality measures and the Pigou–Dalton principle of transfers. In: Dagum, C., Zenga, M. (eds.) Income and Wealth Distribution, Inequality and Poverty, pp. 171–182. Springer, Heidelberg (1990)

Dagum, C.: A new approach to the decomposition of the Gini income inequality ratio. Empir. Econ. **22**, 515–531 (1997)

Davydov, Y., Zitikis, R.: An index of monotonicity and its estimation: a step beyond econometric applications of the Gini index. Metron Int. J. Stat. **63**(3), 351–372 (2005)

De Battisti, F., Porro, F., Vernizzi, A.: The Gini coefficient and the case of negative values. Electron. J. Appl. Stat. Anal. **12**, 85–107 (2019)

De Capitani, L.: On the relations between variability indices–part I. Statistica & Applicazioni **11**(1), 7–22 (2013)

De Capitani, L.: On the relations between variability indices–part II. Statistica & Applicazioni **11**(2), 123–132 (2013)

De Maio, F.G.: Income inequality measures. J. Epidemiol. Community Health **61**(10), 849–852 (2007)

Eliazar, I.I., Sokolov, I.M.: Measuring statistical heterogeneity: the Pietra index. Phys. A **389**, 117–125 (2010)

Frosini, B.V.: Approximation and decomposition of Gini, Pietra–Ricci and Theil inequality measures. Empir. Econ. **43**, 175–197 (2012)

Gravelle, H., Sutton, M.: Trends in geographical inequalities in provision of general practitioners in England and Wales. Lancet **352**(9144), 1910 (1998)

Habib, E.B.: On the decomposition of the Schutz coefficient: an exact approach with an application. Electron. J. Appl. Stat. Anal. **5**, 187–198 (2012)

Hasan, M.U., Malik, S.: Income inequality measurement in Pakistan: empirical evidence from Punjab province. Pak. J. Soc. Sci. **39**(2), 391–401 (2019)

Hoover, E.M.: The measurement of industrial localization. Rev. Econ. Stat. **18**(4), 162–171 (1936)

Huang, Y., Leung, Y.: Measuring regional inequality: a comparison of coefficient of variation and Hoover concentration index. Open Geogr. J. **2**, 25–34 (2009)

Hustopecky, J., Vlachy, J.: Identifying a set of inequality measures for science studies. Scientometrics **1**(1), 85–98 (1978)

Johnston, G., Wilkinson, D.: Increasingly inequitable distribution of general practitioners in Australia, 1986–96. Aust. N. Zeal. J. Public Health **25**(1), 66–70 (2001)

Kennedy, B.P., Kawachi, I., Prothrow-Stith, D.: Income distribution and mortality: cross sectional ecological study of the Robin Hood index in the United States. Br. Med. J. **312**, 1004–1007 (1996)

Kennedy, B.P., Kawachi, I., Prothrow-Stith, D., Lochner, K., Gupta, V.: Social capital, income inequality, and firearm violent crime. Soc. Sci. Med. **47**(1), 7–17 (1998)

Khanal, B.: Is community forestry decreasing the inequality among its users? Study on impact of community forestry on income distribution among different users groups in Nepal. Int. J. Soc. For. **4**(2), 139–152 (2011)

Khosravi Tanak, A., Mohtashami Borzadaran, G., Ahmadi, J.: Entropy maximization under the constraints on the generalized Gini index and its application in modeling income distributions. Phys. A Stat. Mech. Appl. **438**, 657–666 (2015)

Koolman, X., van Doorslaer, E.: On the interpretation of a concentration index of inequality. Health Econ. **13**, 649–656 (2004)

KPMG.: 'Football clubs' valuation: the European elite 2019. Online Report (2019)

Lorenz, M.O.: Methods of measuring the concentration of wealth. Publ. Am. Stat. Assoc. **9**(70), 209–219 (1905)

Manero, A.: The limitations of negative incomes in the Gini coefficient decomposition by source. Appl. Econ. Lett. **24**(14), 977–981 (2017)

Mantzavinis, G.D., Theodorakis, P.N., Dimoliatis, I.: Robin Hood index under discussion. Aust. N. Zeal. J. Public Health **26**(1), 79–80 (2002)

Mobaraki, H., Hassani, A., Kashkalani, T., Khalilnejad, R., Ehsani Chimeh, E.: Equality in distribution of human resources: the case of Iran's ministry of health and medical education. Iran. J. Public Health **42**, 161–165 (2013)

Moothathua, T.S.K.: On unbiased estimation of Gini index and Yntema–Pietra index of lognormal distribution and their variances. Commun. Stat. Theory Methods **18**(2), 661–672 (1989)

Pasquazzi, L., Zenga, M.: Components of Gini, Bonferroni, and Zenga inequality indexes for EU income data. J. Off. Stat. **34**(1), 149–180 (2018)

Pietra, G.: Delle relazioni tra gli indici di variabilitá. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti **LXXIV**(2) (1915)

PwC.: Accounting for typical transactions in the football industry. Online Report (2018)

Ray, J., Singer, J.: Measuring the concentration of power in the international system. Sociol. Methods Res. **1**(4), 403–437 (1973)

Ricci, U.: L' indice di variabilitá e la curva dei redditi. Giornale degli Economisti e Rivista di Statistica **53**, 177–228 (1916)

Rogerson, P.A., Plane, D.A.: The Hoover index of population concentration and the demographic components of change: an article in memory of Andy Isserman. Int. Reg. Sci. Rev. **36**(1), 97–114 (2013)

Sarabia, J.M.: Parametric Lorenz curves: models and applications. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves. Social Exclusion and Well-Being. Economic Studies in Equality, vol. 5, pp. 167–190. Springer, York (2008)

Sarabia, J.M., Jorda, V.: Explicit expressions of the Pietra index for the generalized function for the size distribution of income. Phys. A **416**, 582–595 (2014)

Schutz, R.: On the measurement of income inequality. Am. Econ. Rev. **41**, 107–122 (1951)

Shi, L., Macinko, J., Starfield, B., Wulu, J., Regan, J., Politzer, R.: The relationship between primary care, income inequality and mortality in US states, 1980–1995. J. Am. Board Family Pract. **16**, 412–422 (2003)

Shorrocks, A.F.: Inequality decomposition by factor components. Econometrica **50**, 193–211 (1982)

Shorrocks, A.F.: Inequality decomposition by population subgroups. Econometrica **52**, 1369–1385 (1984)

Shumway, J., Otterstrom, S.: Spatial patterns of migration and income change in the mountain west: the dominance of service-based, amenity-rich counties. Prof. Geogr. **53**(4), 492–502 (2001)

Theodorakis, P.N., Mantzavinis, G.D., Rrumbullaku, L., Lionis, C., Trell, E.: Measuring health inequalities in Albania: a focus on the distribution of general practitioners. Hum. Resour. Health **4**(5), 1–9 (2006)

Wilkinson, D., Symon, B.: Inequitable distribution of general practitioners in Australia: estimating need through the Robin Hood index. Aust. N. Zeal. J. Public Health **24**(1), 71–75 (2000)

Zafari, B., Ekin, T.: Topic modelling for medical prescription fraud and abuse detection. J. R. Stat. Soc. Ser. C (Applied Statistics) **68**(3), 751–769 (2019)

Zenga, M.: Decomposition by sources of the Gini, Bonferroni and Zenga inequality indexes. Statistica & Applicazioni **11**(2), 133–161 (2013)

Zenga, M., Valli, I.: On the decomposition by subpopulations of the point and synthetic Bonferroni inequality measures. Statistica & Applicazioni **14**(1), 3–28 (2016)

Zenga, M., Valli, I.: Joint decomposition by subpopulations and sources of the point and synthetic Bonferroni inequality measures. Statistica & Applicazioni **15**(2), 83–120 (2017)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.