An Iterative Approach to Stratification: Poverty at Regional Level in Italy

Social Indicators Research

An Iterative Approach to Stratification: Poverty at Regional Level in Italy

--Manuscript Draft--

Manuscript Number:	SOCI-D-19-01432R1
Full Title:	An Iterative Approach to Stratification: Poverty at Regional Level in Italy
Article Type:	Original Research
Keywords:	Poverty; Log-normal mixture; Clustering; Stratification
Corresponding Author:	Francesca Mariani Universita Politecnica delle Marche Ancona, ITALY
Corresponding Author's Institution:	Universita Politecnica delle Marche
Order of Authors:	Francesca Mariani
	Mariateresa Ciommi
	Francesco Maria Chelli
	Maria Cristina Recchioni
Funding Information:	
Abstract:	We develop an iterative procedure to generate stratification in non-overlapping classes of a population based on a one-dimensional variable, namely, the Italian income. The procedure works under the assumption that the income distribution is generated by a log-normal mixture of homogeneous income groups. The number of income groups is not determined a priori but rather endogenously by the iterative procedure that stops when the solution to a specific second-degree polynomial equation does not exist. We apply the approach to study the heterogeneity of Italian incomes and the poor population at regional level in the years 2005, 2010, and 2015. The cross-regional comparisons show differences in inequality and stratification dynamics while comparisons over time show the evolution of the classes.

Title Page containing all authors contact details

An Iterative Approach to Stratification: Poverty at Regional Level in Italy

Francesca Mariani*

Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche

Piazza Martelli 8, 60121 Ancona, Italy

Tel. +39 071 2207243, Fax +39 071 2207102, E-mail: f.mariani@univpm.it

Mariateresa Ciommi

Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche

Piazza Martelli 8, 60121 Ancona, Italy

Tel. +39 071 2207089, Fax +39 071 2207102, E-mail: m.ciommi@univpm.it

Francesco M. Chelli

Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche

Piazza Martelli 8, 60121 Ancona, Italy

Tel. +39 071 2207054, Fax +39 071 2207102, E-mail: f.chelli@univpm.it

Maria Cristina Recchioni

Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche

Piazza Martelli 8, 60121 Ancona (AN), Italy

Tel. +39 071 2207066 , Fax +39 071 2207102, E-mail: m.c.recchioni@univpm.it

*corresponding author

An Iterative Approach to Stratification: Poverty at Regional Level in Italy

Abstract

We develop an iterative procedure to generate stratification in non-overlapping classes of a population based on a one-dimensional variable, namely, the Italian income. The procedure works under the assumption that the income distribution is generated by a log-normal mixture of homogeneous income groups. The number of income groups is not determined a priori but rather endogenously by the iterative procedure that stops when the solution to a specific second-degree polynomial equation does not exist. We apply the approach to study the heterogeneity of Italian incomes and the poor population at regional level in the years 2005, 2010, and 2015. The cross-regional comparisons show differences in inequality and stratification dynamics while comparisons over time show the evolution of the classes.

Keywords: Poverty, Log-normal mixture, Clustering, Stratification

1 Introduction

In the social sciences, the terms *inequality* and *stratification* are used from different perspectives. Inequality refers to the extent of the disparities between individuals or groups in the population. Stratification refers to the division of individuals into hierarchical layers or strata. When the stratification hierarchy is based on income, the term *layer* is usually replaced by *class* (see Mann, 1984, Zhou and Wodtke, 2019). This provides the context of this paper. The scope of the stratification is to divide the population into classes that are as homogeneous as possible with classes and heterogeneous between classes. To measure the degree of stratification, we consider the measure of similarity within and between classes. An initial attempt in this direction was made by Yitzhaki and Lerman (1991), who developed a stratification index for gauging the extent to which classes overlap. The idea underlying this approach is that the smaller the extent of overlap, the larger the degree of stratification. Specifically, when classes occupy non-overlapping ranges, we are in the presence of perfect stratification and the Yitzhaki and Lerman index is equal to one. In his paper, Liao (2006) proposes a stratification method based on clustering analysis and provides an index to measure the degree of stratification for non-overlapping classes.

Stratification can be a useful tool for measuring an aspect of poverty which is completely overlooked by inequality, that is, the influence of class membership on an individual's perception of poverty. In fact, despite an unchanged economic situation, the individual's perception of poverty can increase or decrease if his/her class is, respectively, impoverished or enriched relative to the others. Specifically, stratification of the poor population could be used to determine pockets of poverty requiring tailored assistance programs. This finding advocates accounting for stratification in any poverty-reducing policy. Studying this aspect, which is often neglected or underestimated, is a prominent issue as highlighted in Schotte et al. (2018), where a social stratification scheme that differentiates between transient and chronic poverty is proposed. This differentiation is performed in a dynamical perspective, separating the poor population into two classes: chronic poor people with high risk of remaining in poverty, and transient poor people who have above-average chances of moving out of poverty. Moreover, income stratification

is also used for measuring a population's vulnerability to poverty (see López-Calva and Ortiz-Juarez, 2014) and evaluating between-class and within-class inequality (see Jedrzejczak, 2014).

Developing effective methods for income stratification is not a trivial methodological exercise. The approaches to income stratification can be divided into two broad branches: absolute and relative; both are based on the positions occupied by the individuals in the income distribution (see Anikin et al. (2016) for a review of the principal methods of income stratification). Like with absolute and relative poverty (Foster, 1998) or absolute and relative inequality (Niño-Zarazúa et al., 2017), in the absolute approach to stratification, the income thresholds separating the classes are fixed income values, whereas in the relative approach, the income thresholds are determined by considering the standard of living of the population, for example, as fractions of mean or median income or as percentiles of the income distribution. The latter is adopted by Bellettini and Berti Cerioni (2007) who, using data on income distribution collected by the United Nations Development Program (World Income Inequality Database, WIID) for 22 OECD countries in the period 1960–1995, define three classes according to quintiles. Quintiles are also used by Dynan et al. (2004) and Feenberg and Poterba (2000) to identify the rich class and by Profeta (2007) to identify three broad classes in Italy. The main drawback of this absolute approach is that the relative size of each class remains constant over time. With regard to the relative approach, we cite, among others, Pressman (2007) and Peichl et al. (2010), who analyze wealth and poverty in Europe, and, finally, the definition of "lower income class", "middle income class" and "upper income class" created by OECD (2019). All these papers express the values of the income thresholds separating the classes as assigned percentages of the national income median. There have also been some attempts to define economic classes without fixing thresholds. For instance, Medeiros (2006) defines an individual (or a group of individuals) as "rich" if his/her income is sufficient to eliminate poverty. Note that all the above-mentioned absolute or relative methods for stratification suffer from the limitation of fixing the boundaries of the classes "a priori".

Here we study the income stratification of Italian regions using a relative approach based on an iterative clustering technique. The main advantage of our method is that the income thresholds are determined without any a priori choice. The analysis of the Italian income at regional level allows us to capture regional disparities and inequalities at sub level. Moreover, the regional level gives a disaggregate picture of the economic situation in Italy that might also help policy makers to implement more efficient policies in a comparative perspective. The importance of poverty measures on a sub-national level is attested to by the increasing interest in this topic (see, for example, Biggeri et al., 2018). The idea proposed here is to look within the income distribution and identify classes using iterative applications of log-normal mixture models.

To this end, we propose a two-stage procedure. In the first stage, we implement an iterative procedure that groups the incomes into non-overlapping classes. At each iteration, the procedure determines one class by splitting the subset of the still ungrouped incomes into two disjoint sets. The splitting is done, first, by approximating the distribution of the incomes in the considered subset by a bivariate log-normal mixture and, then, by determining the class as the set of incomes below the income where a "significant" change in the mixture distribution occurs. We call this income the "change point" of the mixture. The procedure ends when no further change point is detected. The iterative procedure associates a miss-identification error to each class (see Pittau and Zelli, 2014). This conclude the first stage of the procedure. It must be pointed out that, unlike the approaches discussed above, stratification is obtained without fixing any "a prior" values (relative or absolute) for the thresholds. Due to its characteristics, the first stage of the proposed procedure can be classified as an "iterative" top-down hierarchical clustering method based on a mixture model (for a review of papers on model-based clustering, we refer to Stahl and Sallis, 2012, and McNicholas, 2016). In the second stage, we focus on the poor population, concentrating on the stratification below the poverty line. Specifically, setting the poverty line at 60% of the median income of Italy,

the stratification of the poor population consists of the classes, obtained in the first stage, including incomes which are below the poverty line. When the poverty line does not coincide with any threshold, an additional class just below the poverty line is included in the stratification. This class includes incomes between the largest threshold below the poverty line and the poverty line itself.

Our paper is, therefore, related to several strands of the literature. First, it naturally relates to the literature on income stratification. Second, it relates to the economic literature on unidimensional poverty measurements since we focus on poor classes. Third, the paper is related to the theoretical statistics literature on mixture models in clustering analysis (see, for example, McLachlan and Peel, 2000, and Frühwirth-Schnatter, 2006). The use of the mixture to analyze the presence of classes with different income distributions within the overall income distribution has been widely investigated by Pittau et al. (2010) and by Pittau and Zelli (2014).

The proposed procedure is used to study the stratification of the poor population in Italian regions. To this end, we use Italian data from the survey of the European Union Statistics on Income and Living Conditions (EU-SILC) at regional level (NUTS2). We focus on three different years: 2005, 2010, and 2015. This choice of years is motivated by our interest in capturing and evaluating the Italian situation through a monetary variable, namely disposable income, before, during and after the economic crises generated by Lehman Brothers bankruptcy and, later, the sovereign debt crisis. The analysis is conducted at the individual level, using the equivalised disposable income, that is, the total income of a household, after tax and other deductions, divided by the number of household members converted into equalised adults using the so-called "modified OECD equivalence scale".[1] To complete the analysis, we analyze the differences and similarities among regional poor classes over time by looking at different indices to simultaneously capture inequality and stratification. Among others, we use the stratification index by Liao (2006), the Gini index, and some information about the income stratification. The analysis over the three years reveals the effects of the financial crisis of 2008, which later became a profound economic crisis. Similar effects were observed in the stratification of the U.S. income distribution in the study by Zhou and Wodtke (2019). As discussed in Section 3, the poor classes in Italian regions were affected in a different way with a more marked effect in Northern Italy. Specifically, some negative signals for the income condition of the poorest class is registered from 2005 to 2015 in Veneto, Liguria, Marche, Molise, and Calabria, while a slight positive one is registered in Campania, Basilicata and Sicilia. In the remaining regions, the puzzle of information is more difficult to untangle.

The remainder of this paper is organized as follows. In Section 2 we provide a description of the iterative clustering model-based procedure. In Section 3 an empirical analysis of the poverty stratification in the Italian regions is carried out and the results are commented on. In Section 4 some conclusions are drawn. In Appendix A a detailed explanation of the procedure presented in Section 2 is presented. In Appendix B a brief description of the Expectation Maximization algorithm is shown. In Appendix C some theoretical results, supporting the procedure in Appendix A, are proven. Finally, Appendix D collects some tables regarding the stratification.

2 The stratification procedure

We consider a population of n individuals, $n \in \mathbb{N}$ and use y_i to denote the monetary disposable income of individual i, $i = 1, \ldots, n$. These observed incomes are ranked in ascending order, that is,

$$y_1 \leq y_2 \leq \ldots \leq y_n. \tag{1}$$

[1] The "modified OECD equivalent scale" assigns specific weights to each household member as follows: 1.0 to the first adult; 0.5 to the second and each subsequent person 14 or older; 0.3 to each child under 14.

In line with Pittau at al. (2010) we assume that the observed incomes belong to a population composed of a collection of groups, each with a homogeneous distribution. In what follows, we use the terms "income group" (or "group" for short) and "class" to denote, respectively, the set of incomes with homogeneous distribution and the set of individuals with incomes belonging to an income group.

The stratification procedure is based on two phases. In the first phase, an iterative scheme is developed to assign each income y_i, $i = 1, \ldots, n$, to the appropriate group, thereby partitioning the set of observed incomes into disjointed groups. In this first phase, the poverty line is not involved; it is, however, crucial in the second phase in that the poverty stratification is determined as the classes of individuals whose incomes are below the poverty line. Thus, the key ingredient of the stratification procedure is the iterative scheme to decompose the observed incomes into disjoint groups.

We briefly describe the iterative scheme, i.e., the first phase of the stratification. A detailed explanation of the scheme is presented in Appendix A.

Roughly speaking, the first phase implements an iterative top-down hierarchical clustering procedure that, at each iteration, splits a subset of the observed incomes into two disjoint sets. The splitting uses a bivariate mixture to model the distribution of the incomes in the considered subset and looks for an income where a "significant" change in the mixture distribution occurs. Specifically, in its first iteration, the hierarchical procedure starts from the set of all observed incomes $\mathcal{S}_n = \{y_1, y_2, \ldots, y_n\}$, and identifies a threshold value a^1 (i.e., the first change point) to split $\mathcal{S}_n$ into two disjoint groups: the left group $\mathcal{K}_1 = \{y \in \mathcal{S}_n \ \wedge \ y \in (0, a^1]\}$, made of incomes smaller than or equal to the threshold value, and a right group $\mathcal{R}_1 = \mathcal{S}_n \backslash \mathcal{K}_1$, made of incomes larger than the threshold value. In the second iteration, the procedure considers the subset of $\mathcal{S}_n$, $\mathcal{R}_1$, obtained in the first iteration, identifies a new threshold value $a^2 > a^1$, and splits $\mathcal{R}_1$ into two disjoint groups: the left group $\mathcal{K}_2 = \{y \in \mathcal{S}_n \ \wedge \ y \in (a^1, a^2]\}$ and the right group $\mathcal{R}_2 = \mathcal{S}_n \backslash \mathcal{K}_2$. In the k-th iteration the algorithm proceeds in a similar way by identifying the threshold a^k and splitting the set $\mathcal{R}_{k-1}$ into two groups: $\mathcal{K}_k = \{y \in \mathcal{S}_n \ \wedge \ y \in (a^{k-1}, a^k]\}$ and $\mathcal{R}_k = \{y \in \mathcal{S}_n \ \wedge \ y \in (a^k, +\infty)\} = \mathcal{S}_n \backslash \mathcal{K}_k$.

The thresholds a^1, a^2,..., are the boundaries separating the groups of incomes $\mathcal{K}_1$, $\mathcal{K}_2$, ... that constitute the stratification of the observed income $\mathcal{S}_n$. The procedure works under the assumption that the shifted income probability density function associated with the income sample considered in each iteration is drawn by a mixture of two log-normal distributions. At each iteration we compute the miss-identification error, i.e. the probability of wrongly classifying the incomes below the threshold as members of the left group. This error is associated with the group as a significance level. The procedure stops when a new threshold cannot be found.

Hence, the computation of the thresholds, i.e., the incomes where a "significant" change in the mixture distribution occurs, is crucial and should be detailed starting with some assumptions about the distribution of the observed incomes.

Let $g_0(x)$, $x \in \mathbb{R}_+$, be the probability density function describing the distribution of the observed incomes $\mathcal{S}_n$, and g_k, $x \in \mathbb{R}_+$. The function

$$g_k(x) = \frac{g_0\left(x + a^{k-1}\right)}{\displaystyle\int_{a^{k-1}}^{+\infty} g(y')dy'}, \quad x \in \mathbb{R}_+, \tag{2}$$

is the probability density function associated with the translated income sample $\widetilde{\mathcal{R}}_{k-1} = \{x = y - a^{k-1} \wedge y \in \mathcal{R}_{k-1}\}$. Note that Eq. (2) tells us that g_k is a conditional probability function and, for $k = 1, 2, \ldots$, we assume that g_k is given by a mixture of log-normal probability density functions:

$$g_k(x) = w_k f_{1,k}(x) + (1 - w_k) f_{2,k}(x), \quad x \in \mathbb{R}_+, \tag{3}$$

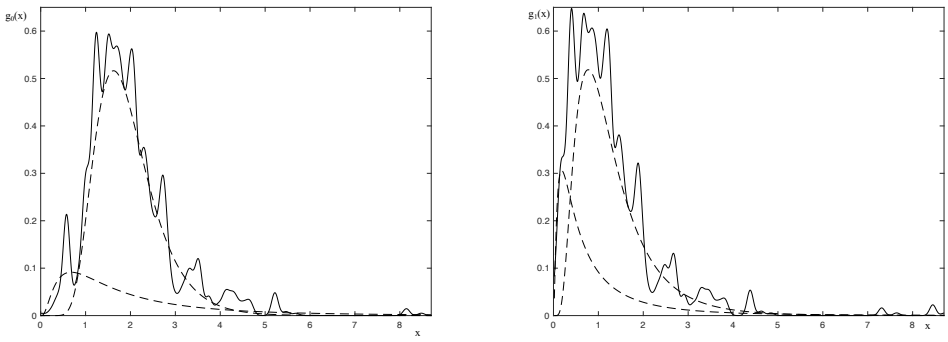

Figure 1: Components of the log-normal mixture, $w_k f_{1,k}$ and $(1-w_k)f_{2,k}$, associated with the observed incomes of Valle D'Aosta belonging to the set $\mathcal{R}_0 = \mathcal{S}_n$ in first iteration ($k=1$) (left panel) and to the set $\mathcal{R}_1$ in the second iteration ($k=2$) (right panel). Year 2015, income expressed in tens of thousands of euros. Kernel density of the income distribution (solid line) and components of the log-normal mixture (dashed line).

where $f_{1,k}(x)$ and $f_{2,k}(x)$, $x \in \mathbb{R}_+$, are the probability density functions associated with the two mixture components and $\underline{\Theta}_k = [w_k, \mu_{1,k}, \mu_{2,k}, \sigma_{1,k}, \sigma_{2,k}]^\top$ is the vector of unknown parameters. Note that $w_k \in [0,1]$ is the mixing weight representing the probability that the point $x = y - a^{k-1}$, $y \in \mathcal{K}_{k-1}$ belongs to the first component. For the sake of simplicity, we assume that $\mu_{1,k} < \mu_{2,k}$, that is, the first component of the mixture is the one with the smallest median, $e^{\mu_{1,k}}$, and the second component is the one with the largest median, $e^{\mu_{2,k}}$. For $j = 1, 2$ we assume that $f_{j,k}(x)$, $x \in \mathbb{R}_+$, is the log-normal density of parameters $\mu_{j,k}, \sigma_{j,k} \in \mathbb{R}$, that is,

$$f_{j,k}(x) = \frac{1}{x\sqrt{2\pi}\sigma_{j,k}} \exp\left\{-\frac{1}{2\sigma_{j,k}^2}(\ln(x) - \mu_{j,k})^2\right\}, \quad x \in \mathbb{R}_+,\ j = 1, 2. \tag{4}$$

The vector $\underline{\Theta}_k$ of the model parameters is unknown and is estimated using the incomes in the set $\mathcal{R}_{k-1}$ using the expectation maximization algorithm (see Appendixes A and B for further details). Fig. 1 shows the two components of the log-normal mixture in the first two iterations of the procedure in the case of Valle D'Aosta in 2015 (here x is the income expressed in tens of thousands of euros). We are now ready to explain the decision rule for membership, namely, the selection of the change points (see Appendix A for further details). At each step, k, once the parameter vector $\underline{\Theta}_k$ is estimated via the Expectation Maximization algorithm, we define the change point of the mixture at the k-th iteration as

$$a^k = \min\{y \in \mathcal{R}_{k-1} \ \wedge \ w_k f_{1,k}(y - a^{k-1}) = (1-w_k) f_{2,k}(y - a^{k-1})\}. \tag{5}$$

The change point a^k is the above-mentioned threshold value associated with iteration k; it represents the frontier separating the two groups $\mathcal{K}_k$ and $\mathcal{R}_k$ (for further details, see Appendix A). The term "change point" is inherited from statistical control theory (see Page, 1955) where it denotes the point at which a time series changes abruptly. Here, the term is used to denote the point separating a sample into two sub-samples with non-homogeneous distributions. In this sense, the purpose of detecting the change point is to segment the observations into statistically homogeneous contiguous regions (see Lung-Yut-Fong et al., 2015).

The definition of change point given in (5) and the assumption of log-normal distribution for $f_{1,k}$ and $f_{2,k}$ guarantee that

$$w_k f_{1,k}(y - a^{k-1}) \geq (1-w_k) f_{2,k}(y - a^{k-1}), \quad y \in \mathcal{R}_{k-1},\ y \leq a^k, \tag{6}$$

thereby implying that the incomes y in the set $\mathcal{R}_{k-1}$, and below a^k (i.e., $y \leq a^k$) belong to group $\mathcal{K}_k$. In fact, roughly speaking, Eq. (6) tells us that the first mixture component dominates the second one for any x such that $x = y - a^{k-1}$, $y \in \mathcal{R}_{k-1}$, $y \leq a^k$ (for the proof see Appendix C). It is worth noting that the choice of change point (5) is in line with the classification rule used in the discriminant analysis when the costs of miss-identification coincide (for further details see Klecka, 1980).

When the mixture under consideration is a mixture of log-normal distributions, the change point is given by an explicit and very simple formula, i.e., Eq. (16) (for further details see Appendix A). The stratification, $\mathcal{K}_1$, $\mathcal{K}_2, \ldots, \mathcal{K}_k, \ldots$ coming from this iterative approach is in line with the definition by Yitzhaki and Lerman (1991) according to which the stratification, unlike inequality, which measures similarities and differences within a class, captures the degree of overlap between members of different classes.

Moreover, we associate each left group $\mathcal{K}_k$ with a miss-identification error, adopting the definition by Pittau and Zelli (2014), which reads

$$F_{2,k}(a^k - a^{k-1}), \tag{7}$$

where $F_{2,k}(x) = \int_0^x f_{2,k}(s)ds$, $x \in \mathbb{R}_+$, is the cumulative distribution function associated with the second component of the mixture. The miss-identification error is the probability of wrongly classifying incomes below the change point a^k as members of the left group $\mathcal{K}_k$.

Finally, the iterative scheme stops when the change point does not exist. This condition corresponds to checking whether the discriminant of a second degree polynomial equation is negative, as illustrated in Appendix A.

The iterative procedure of the first phase presented here has three main advantages: i) it is parsimonious in terms of the number of parameters to be estimated at each iteration in that only a two-component mixture is considered at once; ii) it allows for explicit formulas for the change points; and iii) the iterative scheme does not require any exogenous parameter so the stratification — number of groups and thresholds — depends only on the income distribution.

The second phase of the procedure aims to stratify the poverty. To this end, we fix a poverty line a^P and consider the set $\mathcal{P} = \{y \in \mathcal{S}_n \ \wedge \ y \leq a^P\}$, that is, the set of observed incomes below the poverty line. We then define the poverty stratification as the groups $\mathcal{K}_1^P = \mathcal{K}_1$, $\mathcal{K}_2^P = \mathcal{K}_2$,...,$\mathcal{K}_{m-1}^P = \mathcal{K}_{m-1}$ with $a^k < a^P$, $k = 1, 2, \ldots, m-1$, where m is such that $a^m \geq a^P$, and $\mathcal{K}_m^P = \mathcal{P} \setminus \bigcup_{k=1}^{m-1} \mathcal{K}_k^P$. Note that the set $\{\mathcal{K}_1^P, \mathcal{K}_2^P, \ldots, \mathcal{K}_m^P\}$ constitutes a partition of the poor population $\mathcal{P}$. In the following, therefore, we refer to poor classes associated with the poverty line a^P as the classes made of individuals whose incomes are members of $\mathcal{K}_k^P$, $k = 1, 2, \ldots, m$.

In the second phase of the procedure to address $\mathcal{P}$, we compute the overall Gini index (see Gini, 1912) to measure inequality and we use the Liao index (see Liao, 2006) to measure the stratification of incomes belonging to the groups $\mathcal{K}_k^P$, $k = 1, 2, \ldots, m$. Specifically, the Gini index of $\mathcal{P}$ reads as

$$G = \frac{1}{2|\mathcal{P}|^2 \bar{y}} \sum_{y_i \in \mathcal{P}} \sum_{y_j \in \mathcal{P}} |y_i - y_j|, \tag{8}$$

where $|\mathcal{P}|$ is the cardinality of $\mathcal{P}$ and $\bar{y} = \frac{1}{|\mathcal{P}|} \sum_{y_i \in \mathcal{P}} y_i$ is the sample mean of poor population incomes.

The Liao index reads:

$$S = \frac{G_b}{G}, \tag{9}$$

where G_b is the between-class component of the Gini index given by

$$G_b = \frac{1}{2|\mathcal{P}|^2 \bar{y}} \sum_{k=2}^{m} \sum_{h=1}^{k-1} \sum_{y_i \in \mathcal{K}_k^P} \sum_{y_j \in \mathcal{K}_h^P} |y_i - y_j|. \tag{10}$$

The Liao index ranges from zero to one by virtue of the Dagum decomposition (see Dagum, 1997), which states that the overall Gini index is composed of three components: within-class, between-class, and an overlapping term. Here, the clusters $\mathcal{K}_k^P$, $k = 1, 2, \ldots, m$, do not overlap, so the overlapping term is zero and the Gini index is the sum of the within-class and between-class terms. The Liao index is equal to zero when there is no income variation between classes and the overall inequality is the sum of the inequalities within classes. By contrast, it is equal to one when the overall inequality reduces to the between-class inequality, so there is no variation within classes.

3 An empirical illustration

We briefly present the results of the iterative procedure, then we focus on poverty stratification, that is the stratification of classes with incomes below the poverty line, as described in Section 2. To this end, we fix the poverty line equal to the national poverty line, i.e., 60% of the median income of Italy.

We use cross-sectional EU-SILC data for Italy at regional level (NUTS2) in three different years, namely 2005, 2010, and 2015. The variable of interest is the equivalised disposable income, that is, the total household income after tax and other deductions, divided by the number of household members converted into equalised adults according the "modified OECD equivalence scale"[2]. According to what was suggested by Eurostat (2006) and Van Kerm (2007) about the treatment of EU-SILC data, we remove zero and negative incomes from the original sample.

A total of 158 cases were eliminated out of 22,032 in 2005, 142 out of 19,147 in 2010; and 191 out of 17,985 in 2015. For the sample size of data by region and year we refer to Table 3 in Appendix D. In this Appendix, we provide a comprehensive view of the income stratification in Tables 4–10. Specifically, Tables 4–6 report the income thresholds of the poverty stratification, Tables 7–9 report the income thresholds of poverty stratification as percentages of the median Italian income and Table 10 reports the income thresholds of the stratification and the miss-identification probabilities (in brackets).

As previously mentioned, the analysis is conducted at regional level.[3] The main advantage of regional analysis is the possibility of comparatively highlighting local income disparities (see Biggeri et al., 2018).

In this section, using the first phase of the iterative procedure detailed in Section 2, we determine the income stratification of the population in Italian regions in 2005, 2010, and 2015 and, later, we concentrate on the poverty stratification implementing the second phase of the procedure in Section 2.

The three panels in Fig. 2 show the income stratification of the regions up to the last threshold of each region along with the the national poverty line, set at 60% of the median income of Italy (dashed black line), the regional poverty lines, set at 60% of the median income of each region, (piece-wise solid red line), and the regional median income (piece-wise dotted blue line) in 2005 (top panel), 2010 (middle panel), and 2015 (bottom panel). We recall that the poverty stratification is composed of the classes below the national poverty line, that is, the classes below the dashed line in Fig. 2.

[2] A detailed definition of all the sources of income that constitute this variable were reported by Graf et al. 2011, in Table 2.1: Recommended definition of the variable total disposable household income (gross), on page 16.

[3] We use NUTS2 classification, see European Commission (2011). The Provincia Autonoma di Bolzano/Bozen and the Provincia Autonoma di Trento make up the region Trentino-Alto Adige. In this way, we use the Italian classification of the Regions.

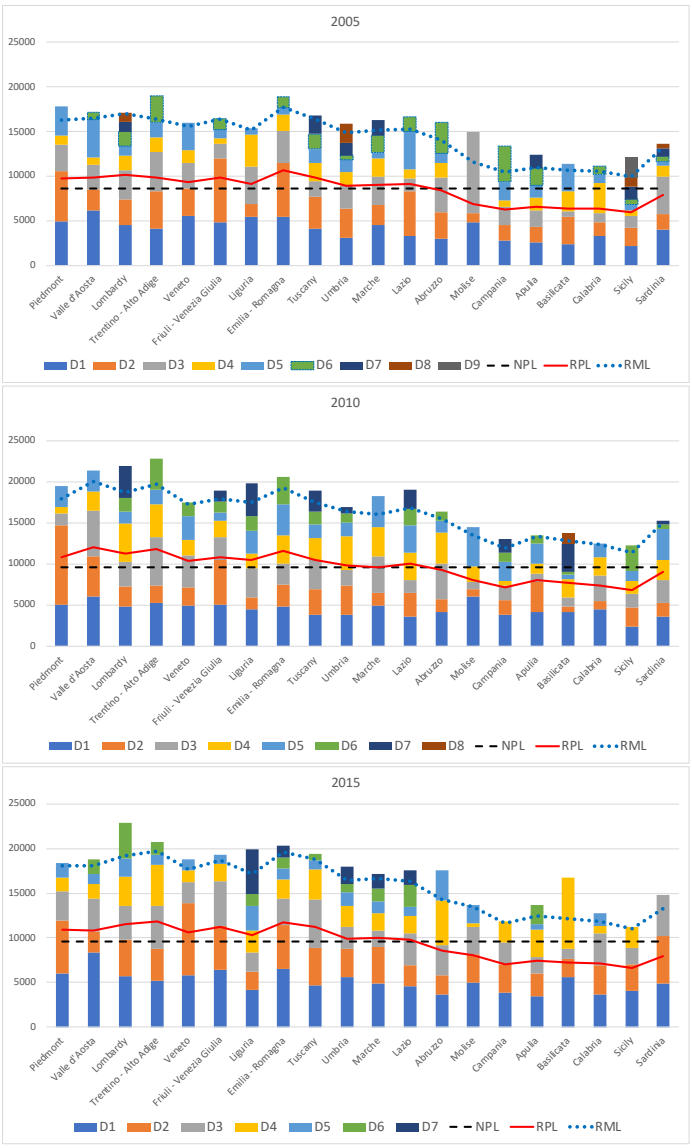

Figure 2: Stratification resulting from the iterative procedure (vertical bars). The last class represented is the one below the highest threshold of each region. Poverty stratification with respect to the national poverty line (NPL) – bars below horizontal dashed black line; poverty stratification with respect to the regional poverty line (RPL) – bars below the piece-wise solid red line. Here $Dk = a^k - a^{k-1}$, $k = 1, 2, \ldots, 8$, and the y-axis shows the individual equivalent disposable income. The dotted line denotes the regional median line (RML).

The three panels show that in the years considered , the regional poverty lines of the northern regions and Tuscany lie above the national poverty line while, as expected, those of the southern and insular regions lie below. Interestingly, these two lines very closely correspond for Umbria, Marche, and Lazio, indicating that these three regions reflect the aggregate data. We also observe that the income thresholds of the poorest classes lie far from the poverty line in many regions, which indicates that the poorest classes correspond to individuals living in extreme poverty and that there is evident heterogeneity among them. For instance, looking at the thresholds of the poorest class for the year 2005 in Tables 4–6 in Appendix D, we observe that in the northern regions the minimum value is held by Trentino-Alto Adige (EUR 4140.76) whereas the maximum is reached in Valle d'Aosta (EUR 6216.16). As expected, the thresholds of the poorest class for the central regions are lower: in 2005 Umbria has the minimum and Marche the maximum, with EUR 3070.54 and EUR 4580.09, respectively. Except for Molise, the

southern and insular regions have even lower values: Sicily has EUR 2168.33 and Sardinia has the highest value, namely, EUR 3982.08. The computation of national and sub-national poverty lines therefore highlights important geographical implications when evaluating poverty. Poverty stratification therefore illustrates the presence of very different poverty levels among regions and within regions, so it may represent a tool to implement class-tailored poverty-reducing policies.[4]

We now analyze the evolution of the poverty stratification, i.e., the classes below the dashed line in Fig. 2, focusing first on the poorest class. Looking at Figures 2 and 3, we note a reduction in the number of classes in the southern and insular regions from 2005 to 2015 except for Abruzzo and Molise, where this number remains constant (Fig. 2). We also note an upward shift in the thresholds of the poorest classes in many regions (Fig. 3). These two findings give rise to some questions.

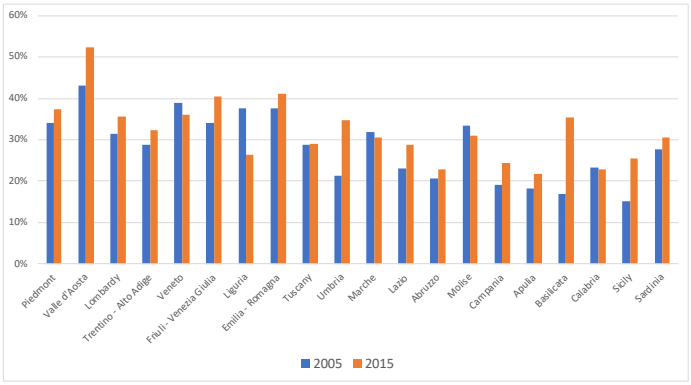

Figure 3: Individual equivalent disposable income threshold shares of the poorest classes (i.e., threshold to national median income) in 2005 and 2015.

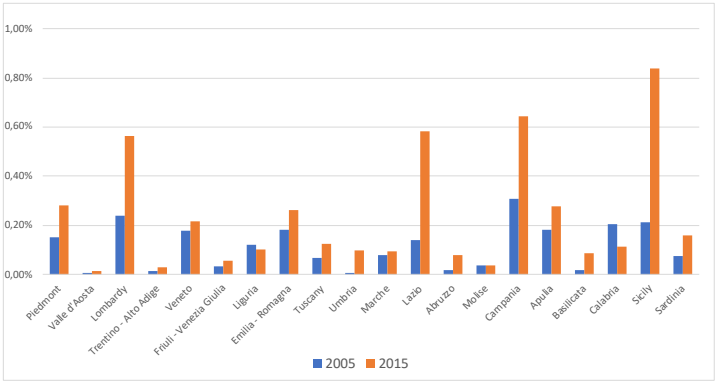

Figure 4: Shares of individuals in the poorest classes (i.e., poorest class size to the national population size) in 2005 and 2015.

Could the reduction in the number of classes in the income stratification across the southern and insular regions from 2005 to 2015 indicate an impoverishment of the poorest class? Could the upward shift in the poorest class thresholds from 2005 to 2015 allude to a positive signal?

We analyze the evolution of the poorest class from 2005 to 2015 to address these points. To do so, we simultaneously consider the evolution of the income threshold share, the share of individuals, and the Gini index of the poorest class from 2005 to 2015 as depicted, respectively, in Figure 3, Figure 4, and Table 1. The income threshold

[4]Of course, in order to compare regions in "real terms," we should also consider geographical variations in the cost of living in different areas by applying specific spatial price indices to the poverty rates.

share of the poorest class is calculated as the ratio of the income threshold of the poorest class to the national income median, while the size share is the ratio of the poorest class size to the national population size.

We expect that an increase in income threshold share and a decrease in the share of individuals in the poorest class is a positive signal, especially when accompanied by a stable or decreasing Gini index (i.e., not increasing inequality). In contrast, a decrease in income threshold share and an increase in the share of individuals in the poorest class is a negative signal, especially when the Gini index is stable or increasing (i.e., not decreasing inequality). None of the remaining combinations allow us to conclude that the poorest class has improved or not evolved.

Looking at Fig. 3, we first observe that Veneto, Liguria, Marche, Molise, and Calabria show a decrease in the income threshold share from 2005 to 2015 while all remaining regions show an increase in the poorest income threshold. In Veneto, Liguria, Marche, Molise, and Calabria the decrease in the threshold of the poorest class from 2005 to 2015 alludes to a worsened income condition for this class. However, only in Veneto and Marche do we observe signals confirming the worsened condition, that is, an increase in the share of individuals in the poorest class from 2005 to 2015 (see Fig. 4) and an increase in the Gini index of the poorest class.

With respect to Liguria, Molise, and Calabria, the observation of Fiure 3, Figure 4, and Table 1 do not allow us to conclude anything about the quality (improvement/worsening) of the poorest class; the same conflicting signals are also observed for all remaining regions.

Region	**Gini 2005**	**Gini 2010**	**Gini 2015**
Piedmont	0.243	0.297	0.250
Valle d'Aosta	0.102	0.250	0.124
Lombardy	0.248	0.321	0.258
Trentino - Alto Adige	0.388	0.322	0.426
Veneto	0.187	0.347	0.243
Friuli - Venezia Giulia	0.233	0.231	0.237
Liguria	0.257	0.306	0.307
Emilia - Romagna	0.222	0.289	0.256
Average	**0.235**	**0.311**	**0.261**

Region	**Gini 2005**	**Gini 2010**	**Gini 2015**
Tuscany	0.342	0.324	0.300
Umbria	0.207	0.341	0.316
Marche	0.346	0.325	0.349
Lazio	0.316	0.398	0.292
Average	**0.320**	**0.361**	**0.304**

Region	**Gini 2005**	**Gini 2010**	**Gini 2015**
Abruzzo	0.302	0.387	0.445
Molise	0.221	0.224	0.236
Campania	0.291	0.270	0.245
Apulia	0.347	0.251	0.385
Basilicata	0.337	0.264	0.117
Calabria	0.248	0.201	0.241
Sicily	0.343	0.298	0.227
Sardinia	0.236	0.339	0.243
Average	**0.307**	**0.279**	**0.277**

Table 1: Gini index of poorest population in northern (upper), central (middle), and southern and insular (bottom) Italian regions.

However, some insights into the evolution of the poorest classes may be gained from the average values in Table 1. We observe that from 2005 to 2015 in northern and central regions, the average value of the Gini index of the poorest class shows U-shaped dynamics. This is not observed in the southern and insular regions where the

average value of the Gini index of the poorest class decreases. These findings suggest that the poorest classes in the northern and central regions were affected by the crisis more than those in the southern and insular regions. Note that the average values in Tables 1 and 2 are computed as weighted arithmetic means of the regional indices using the regional sample weights. For this reason, the average behavior of the average indices is strongly affected by the behavior of the index in the most populated regions (Lombardy, Veneto and Emilia-Romagna for northern regions, Lazio for central regions and Campania and Sicily for southern and insular regions). Specifically, looking at the northern regions, we observe that the Gini index of the poorest class shows a peak in 2010 in all regions, except for Trentino-Alto Adige and Friuli-Venezia Giulia; moreover, in 2015 the inequality level in these regions returned to the values registered in 2005 except for Trentino-Alto Adige where the index doubled from 2005 to 2015. Regarding the southern and insular regions, the Gini index of the poorest class decreases appreciably in Campania, Basilicata, and Sicilia, thus indicating a reduction in inequality that, together with the increase in income thresholds (see Fig. 3), could be interpreted as a slight positive signal for the income condition of the poorest class. In Abruzzo, Molise, Apulia, and Sardinia, we register an increase in inequality in the poorest class. In the central regions we observe high values of the Gini index of the poorest class, especially in 2015, indicating a higher level of inequality in the poorest class in central Italy rather than in the North, South and islands, except for Abruzzo and Trentino-Alto Adige, where we observe a Gini index larger than 0.4. Indeed, with respect to social and economic factors, Abruzzo is not far from the central region, so the poorest class suffered from the crisis similarly to the poorest classes of the central regions.

	2005		**2010**		**2015**	
Region	**Gini index**	**Liao index**	**Gini index**	**Liao index**	**Gini index**	**Liao index**
Piedmont	0.158	0.620	0.188	0.713	0.201	0.746
Valle d'Aosta	0.142	0.761	0.202	0.772	0.148	0.399
Lombardy	0.160	0.863	0.187	0.861	0.215	0.735
Trentino - Alto Adige	0.135	0.689	0.179	0.860	0.217	0.876
Veneto	0.145	0.662	0.173	0.850	0.169	0.734
Friuli - Venezia Giulia	0.140	0.612	0.160	0.631	0.191	0.770
Liguria	0.196	0.878	0.174	0.718	0.225	0.907
Emilia - Romagna	0.170	0.718	0.160	0.856	0.204	0.750
Average	**0.159**	**0.748**	**0.178**	**0.815**	**0.202**	**0.754**

	2005		**2010**		**2015**	
Region	**Gini index**	**Liao index**	**Gini index**	**Liao index**	**Gini index**	**Liao index**
Tuscany	0.151	0.834	0.180	0.850	0.195	0.790
Umbria	0.133	0.767	0.146	0.863	0.258	0.860
Marche	0.224	0.893	0.192	0.828	0.226	0.765
Lazio	0.168	0.577	0.189	0.926	0.244	0.881
Average	**0.167**	**0.717**	**0.183**	**0.884**	**0.227**	**0.836**

	2005		**2010**		**2015**	
Region	**Gini index**	**Liao index**	**Gini index**	**Liao index**	**Gini index**	**Liao index**
Abruzzo	0.146	0.781	0.195	0.728	0.224	0.844
Molise	0.186	0.817	0.197	0.858	0.201	0.881
Campania	0.191	0.942	0.209	0.947	0.211	0.882
Apulia	0.190	0.947	0.179	0.866	0.226	0.938
Basilicata	0.167	0.868	0.189	0.921	0.172	0.930
Calabria	0.212	0.863	0.184	0.853	0.177	0.831
Sicily	0.198	0.970	0.191	0.958	0.233	0.923
Sardinia	0.172	0.820	0.181	0.912	0.200	0.667
Average	**0.189**	**0.918**	**0.192**	**0.906**	**0.215**	**0.880**

Table 2: Gini and Liao indices of the poor population in northern (upper), central (middle), southern and insular (bottom) Italian regions.

We now analyze the evolution of poverty stratification in Italy from 2005 to 2015, focusing on all the poor classes. This is done considering the evolution of the number of poor classes (see Fig. 2) along with the Gini and Liao indices (see Table 2), which are used to measure, respectively, inequality and stratification, or, using Liao's terminology (2006), the "individual inequality" and the "class inequality." In line with Stewart et al. (2005) and Jayaraj and Subramanian (2006), individual inequality, also called "vertical inequality," can be decomposed into between-class inequality (also called "horizontal inequality") and within-class inequality. Empirical research suggests that high individual inequality seems to be negatively correlated with economic growth (e.g., Persson and Tabellini, 1994; Perotti, 1993). On the other hand, high class inequality makes the achievement of social objectives more difficult and may be a source of social conflicts (see Steward et al., 2005). However, class inequality is crucial when implementing a poverty-reduction policy, since it is not realistic to improve the position of individuals without considering the position of the class.

With regard to the individual inequality of the poor population, we observe an increase on average in the Gini index from 2005 to 2015 in all regions except for Calabria, and the maximum value is, surprisingly, achieved in central Italy with 0.231. This increase in the average value of the Gini index alludes to a possible slowdown in economic growth (e.g., Persson and Tabellini, 1994; Perotti, 1993) probably generated by the crisis in 2008 and the sovereign debt crisis. In fact, the onset of the latter was in late 2009 (Greek government deficits) while culminating with the bailout of Greek, Spain, and Cyprus in 2012. Looking at Table 2, we observe that Umbria shows the lowest value of the Gini index in 2005 (0.133) and 2010 (0.146), while in 2015 the lowest value is observed in Valle d'Aosta (0.148). By contrast, the highest inequality is observed in Marche (0.224) in 2005, in Campania (0.209) in 2010, and in Lazio (0.244) in 2015.

As for class inequality, Table 2 shows that in all regions and in all years considered, the Liao index of the poor population, except for Valle d'Aosta in 2015, is larger than 0.5, and reaches the highest values in the southern and insular regions. This fact reveals that in most Italian regions, especially in the southern and insular regions, (individual) inequality in the poor population is mainly due to class inequality. Specifically for the northern regions, we observe that Friuli-Venezia Giulia shows the lowest value of the Liao index in 2005 (0.612) and 2010 (0.631), while in 2015 it is Valle d'Aosta that has the lowest value (0.399). For southern and insular regions, Sicily reaches 0.970 in 2005 and 0.958 in 2010 while Apulia is 0.938 in 2015. It is worth noting that in the northern and central regions, the average value of the Liao index of the poor population from 2005 to 2015 shows U-shaped dynamics already observed in Table 1 for the average value of the Gini index of the poorest class. Therefore, this Gini U-shaped dynamics combined with an unchanged number of poor classes from 2005 to 2015 in the northern and central regions except for Liguria (see Fig. 2), reveals a temporary increase (on average) in class inequality during the sovereign debt crisis.

In contrast to the northern and central regions, from 2005 to 2015, the southern and insular regions show on average a reduction of the Liao index. This finding could be due to the decrease in the number of poor classes observed in these regions, except for Abruzzo and Molise. However, countering this trend, Basilicata shows an increase in the Liao index, that is, a slight positive signal that class inequality is reduced.

A comparison of the Gini and Liao indices (see Table 2) shows that they do not necessarily move together. In fact, in 2005 Emilia-Romagna and Lazio show similar values of the Gini index (0.170, 0.168), while Emilia-Romagna shows a Liao index (0.718) appreciably higher than Lazio (0.577). This suggests that in Emilia-Romagna there is a more rigid, hierarchical order in income, thus exacerbating the social and political consequences of inequality (see Zhou and Wodtke, 2019; Wilkinson and Pickett, 2009). Moreover, Table 2 shows that the dynamics of both indices is very similar in Marche and Lazio, in Campania, Apulia, and Sicily, and in Lombardy and Trentino-Alto Adige. This combined use of these indices may help policy makers to better navigate the complexities of appropriate

policies to reduce inequality.

More interestingly, a comparison of the Gini index of the only poorest class (Table 1) with that of the poor classes together (Table 2) shows the evolution of inequality in the poor classes. As observed above, from 2005 to 2015 in all northern regions, except for Trentino-Alto Adige and Friuli-Venezia Giulia, the Gini index of the poorest class shows a peak in 2010. In contrast, the Gini index of the aggregated poor classes often shows a monotonic increase. This behavior of the index suggests an increase in inequality in the poor classes of Northern Italy, probably due to the restriction to bank loans after the Lehman Brothers bankruptcy, with a more marked effect on the poorest class in 2010 in Piedmont, Valle d'Aosta, Lombardy, Veneto, Liguria and Emilia-Romagna.

In conclusion, the analysis of behavior of the poor population over the three years considered reveals the ability of the stratification to respond to economic and financial distress.

4 Conclusions

This paper proposes an iterative approach for income stratification and analyzes the dynamics of poverty stratification in Italian regions in the years 2005, 2010, and 2015. The analysis shows that the Gini and Liao indices of poor classes do not necessarily move together. Furthermore, a comparison over time between the Gini index of the poor classes and the Gini index of the only poorest class allows for capturing the different evolution of the classes. This can lead to different perceptions of poverty within the poor population at the regional level, thereby exacerbating or ameliorating the social and political consequences of inequality. This fact highlights the importance of studying inequality not only from an individual perspective, but also from a class perspective. Moreover, this suggests future research to analyze the effects of the distribution of labour earnings and government transfers on income stratification. Finally, the dynamics of the poverty stratification over the three years analyzed shows that the number of poor classes changes differently over time, giving rise to different effects among regions. The dynamics of the Liao index and thresholds of the poorest classes at the regional level allow the poorest classes with a worsened poverty condition to be identified. That is, descriptive analysis of the poor classes over time enables the effect of financial and economic shock on poverty classes to be captured.

Appendix A: The stratification procedure in detail

In this appendix we detail phase one of the stratification procedure summarized in Section 2.

As previously mentioned, the first phase of the stratification is done by means of an iterative hierarchical clustering scheme based on a top-down or divisive approach applied starting from the set $\mathcal{S}_n = \{y_1, y_2, ..., y_n\}$, of the observed incomes ranked in ascending order.

In the final step, the procedure determines a partition of the set $\mathcal{S}_n$ into N groups,

$$\mathcal{K}_k = \{y \in \mathcal{S}_n \ \wedge \ y \in (a^{k-1}, a^k]\}, \ k = 1, 2, \ldots, N, \tag{11}$$

where $a^0 = 0$ and $a^{k-1} < a^k$, such that

$$\mathcal{K}_k \subset \mathcal{S}_n, \ k = 0, 1, \ldots, N, \ \bigcup_{k=1}^{N} \mathcal{K}_k = \mathcal{S}_n, \ \mathcal{K}_i \bigcap \mathcal{K}_j = \emptyset, \ i \neq j.$$

The number N of groups and the points a^1, a^2,...,a^N are not fixed in advanced but are determined by the stopping rule of the procedure itself.

Setting $\mathcal{R}_0 = \mathcal{S}_n$, for $k = 1, 2, \ldots, N$, the iterative scheme splits the set $\mathcal{R}_{k-1}$ into two disjoint sets $\mathcal{K}_k$ and $\mathcal{R}_k$ where $\mathcal{K}_k$ is defined in Eq. (11), while

$$\mathcal{R}_k = \left\{ y \in \mathcal{S}_n \ \wedge y \in (a^k, +\infty) \right\}. \tag{12}$$

The procedure stops at the N-th iteration if the stopping rule is satisfied and the last set $\mathcal{R}_N$ reads as

$$\mathcal{R}_N = \mathcal{S}_n \setminus \bigcup_{k=1}^{N} \mathcal{K}_k.$$

Hence, the first phase of the procedure provides the stratification $\mathcal{K}_1$, $\mathcal{K}_2$,...,$\mathcal{K}_N$, and $\mathcal{R}_N$ that includes all the observed incomes.

The second phase of the stratification requires a poverty line a^p, and the identification of two thresholds, a^{m-1}, a^m such that $a^{m-1} < a^P \leq a^m$. Thus, given the poverty line a^P and the m-th threshold, the poverty stratification is given by the groups: $\mathcal{K}_1^P = \mathcal{K}_1$, $\mathcal{K}_2^P = \mathcal{K}_2$,...,$\mathcal{K}_{m-1}^P = \mathcal{K}_{m-1}$ and $\mathcal{K}_m^P = \{\, y \in \mathcal{S}_n \ \wedge \ y \in (a^{m-1}, a^P] \,\}$.

We now focus on the splitting and stopping rules. We start by assuming that the overall income probability density function, $g_0(x)$, $x \in \mathbb{R}_+$, describing the distribution of the overall income sample $\mathcal{S}_n$ and all conditional probability density functions (see Eq. (2)),

$$g_k(x) = \frac{g_0\left(x + a^{k-1}\right)}{\displaystyle\int_{a^{k-1}}^{+\infty} g(y')dy'}, \quad x \in \mathbb{R}_+, \ k = 1, 2, \ldots, \tag{13}$$

are drawn from a mixture of two log-normal distributions. Specifically, we assume that the probability density functions $g_k(x)$ are given by

$$g_k(x) = w_k f_{1,k}(x) + (1 - w_k) f_{2,k}(x), \ x \in \mathbb{R}_+,$$

where $f_{j,k}$, $j = 1, 2$, $k = 1, 2, \ldots, N$ are, respectively, log-normal probability density functions of parameters $\mu_{j,k}$ and $\sigma_{j,k}$, $j = 1, 2$, while w_k and $1 - w_k$, $w_k \in [0, 1]$, are the mixing weights. Specifically, the weight w_k is the probability that x belongs to mixture component 1, while $1 - w_k$ is the probability of belonging to component 2. Furthermore, without loss of generality, we assume that for any k, $\mu_{1,k} < \mu_{2,k}$.

To keep the notation simple, we omit the dependence of the probability density function g_k from the vector of model parameters $\underline{\Theta}_k = [w_k, \mu_{1,k}, \mu_{2,k}, \sigma_{1,k}, \sigma_{2,k}]^\top$.

Starting from the set of the observed incomes, $\mathcal{R}_0 = \mathcal{S}_n$, at each step k, $k = 1, 2, \ldots, N$, we apply the Estimation Maximization (EM) method (see Dempster, 1977, and Appendix B for a brief description)[5] to estimate the parameter vector $\underline{\Theta}_k$ of the mixture used to approximate the conditional probability density function $g_k(x)$, $x \in \widetilde{\mathcal{R}}_{k-1}$, where the set $\widetilde{\mathcal{R}}_{k-1}$ is related to the observed incomes belonging to the set $\mathcal{R}_{k-1}$ via the following relation:

$$\widetilde{\mathcal{R}}_{k-1} = \left\{ x = y - a^{k-1}, \ \ y \in \mathcal{R}_{k-1} \right\}.$$

Once $\underline{\Theta}_k$ has been computed, we proceed by looking for the so called "change point',' a^k, defined as

$$a^k = a^{k-1} + \xi, \tag{14}$$

where $\xi \in \widetilde{\mathcal{R}}_{k-1}$ is the smallest solution to the equation:

$$w_k f_{1,k}(\xi) = (1 - w_k) f_{2,k}(\xi). \tag{15}$$

[5]Specifically, we use the Statistics and Machine Learning Toolbox of MatLab.

When the mixture is composed of log-normal distributions, the smallest solution ξ of Eq. (14), if **it** exists, is explicitly given by

$$a^k = a^{k-1} + \exp\left\{\frac{\sigma_{1,k}\sigma_{2,k}}{\sigma_{2,k}^2 - \sigma_{1,k}^2}\left[\left(\frac{\mu_{1,k}}{\sigma_{1,k}^2} - \frac{\mu_{2,k}}{\sigma_{2,k}^2}\right) - \sqrt{\Delta_k}\right]\right\}, \tag{16}$$

where

$$\Delta_k = \left(\frac{\mu_{1,k}}{\sigma_{1,k}^2} - \frac{\mu_{2,k}}{\sigma_{2,k}^2}\right)^2 + \left(2\ln\left(\frac{\sigma_{2,k}w_k}{\sigma_{1,k}(1-w_k)}\right) + \left(\frac{\mu_{1,k}}{\sigma_{1,k}}\right)^2 - \left(\frac{\mu_{2,k}}{\sigma_{2,k}}\right)^2\right)(\sigma_{2,k}^2 - \sigma_{1,k}^2). \tag{17}$$

The existence of the change point is guaranteed by $\Delta_k \geq 0$ and the proof of (16), (17) is presented in Appendix C.

When no solution to Eq. (15) exists (i.e., $\Delta_k < 0$) then the algorithm stops and $N = k$. Otherwise we split the set $\mathcal{R}_{k-1}$ into two sets (11) and (12). Each class $\mathcal{K}_k$, $k = 1, 2, \ldots, N$ is associated with the miss-identification error (see Pittau and Zelli, 2014) defined in (7). That is, the stopping rule reads as:

Stopping rule: The splitting procedure stops at the k-th iteration where no solution to Eq. (15) exists (i.e. $\Delta_k < 0$).

We conclude this appendix by providing further details on the splitting rule. For $k = 1, 2, \ldots$ and for any $x \in \widetilde{\mathcal{R}}_{k-1}$, we use z_x to denote the random variable

$$z_x = j, \quad \text{if } x \text{ belongs to component j}, \qquad j = 1, 2. \tag{18}$$

Note that

$$Pr(z_x = 1 \,|\, \underline{\Theta}_k) = w_k,\ Pr(z_x = 2 \,|\, \underline{\Theta}_k) = 1 - w_k, \quad x \in \widetilde{\mathcal{R}}_{k-1}, \tag{19}$$

are, respectively, the probability that $x \in \widetilde{\mathcal{R}}_{k-1}$ belongs to components 1 and 2 given that $\underline{\Theta}_k$ is the mixture parameter vector. Using Bayes' rule, we compute the conditional probability that x belongs to component j, given $x \in \widetilde{\mathcal{R}}_{k-1}$ and $\underline{\Theta}_k$:

$$\begin{aligned} Pr(z_x = j \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k) &= \lim_{h\to 0} Pr(z_x = j \mid [x-h, x+h] \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k) \\ &= \frac{Pr(z_x = j \,|\, \underline{\Theta}_k) f_{j,k}(x)}{Pr(z_x = 1 \,|\, \underline{\Theta}_k) f_{1,k}(x) + Pr(z_x = 2 \,|\, \underline{\Theta}_k) f_{2,k}(x)}, \quad j = 1, 2. \end{aligned} \tag{20}$$

By substituting (19) into (20), Eq. (20) can be rewritten as follows:

$$Pr(z_x = 1 \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k) = \frac{w_k f_{1,k}(x)}{w_k f_{1,k}(x) + (1-w_k) f_{2,k}(x)}, \tag{21}$$

$$Pr(z_x = 2 \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k) = \frac{(1-w_k) f_{2,k}(x)}{w_k f_{1,k}(x) + (1-w_k) f_{2,k}(x)}. \tag{22}$$

Finally, we associate the income $y = x + a^{k-1}$ with the class $\mathcal{K}_k$ when the conditional probabilities $Pr(z_x = j \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k)$, $j = 1, 2$, satisfy the inequality

$$Pr(z_x = 1 \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k) \geq Pr(z_x^k = 2 \mid x \in \widetilde{\mathcal{R}}_{k-1}, \underline{\Theta}_k),$$

which reads

$$w_k f_{1,k}(x) \geq (1-w_k) f_{2,k}(x). \tag{23}$$

Condition (23) is satisfied for all $x \in \widetilde{\mathcal{R}}_{k-1}$ such that $x + a^{k-1} \leq a^k$, where a^k is the change point at the k-th step in Eq. (14) (for further details see the corollary in Appendix C).

Hence, at each iteration, the splitting rule is based on the standard approach of the discriminant analysis in the case of two assigned populations with known probability densities and, roughly speaking, the splitting procedure stops when the weighted densities of the mixture overlap "too much." This occurs when the condition in the stopping rule mentioned above is verified.

In conclusion, the iterative approach is a tool for determining the stratification of the observed incomes and it is obtained without accounting for any poverty line. The incomes a^k, $k = 1, 2, \ldots, N$, constitute the border of the stratification and are determined without fixing any a priori percentage of median or mean of the income population. The poverty line is used only to define the poverty stratification as the stratification of the observed incomes below the poverty line.

Appendix B: EM in a nutshell

The Expectation-Maximization (EM) algorithm (see Dempster et al., 1977) is an iterative method in which the model depends on unobserved latent variables. It is usually used for hidden Markov models and mixture models. Here we focus on the EM algorithm for log-normal mixture models.

Let $x_1, x_2, \ldots, x_n$ be a sample of n independent observations from n independent and identically distributed (i.i.d.) variables $X_1, X_2, \ldots, X_n$ drawn from a mixture of two univariate log-normal distributions and let $Z_1, Z_2, \ldots, Z_n$ be latent variables that identify in which component the observations $x_1, x_2, \ldots, x_n$ originate. That is, we have

$$\ln(X_i) \,|\, Z_i = 1 \sim N(\mu_1, \sigma_1),$$
$$\ln(X_i) \,|\, Z_i = 2 \sim N(\mu_2, \sigma_2),$$

where:

$$P(Z_i = 1) = w_1 = w,$$
$$P(Z_i = 2) = w_2 = 1 - w.$$

The goal of the EM iterative scheme is:

Given the observations $\underline{x} = [x_1, x_2, \ldots, x_n]^\top$ *but no* $\underline{z} = [z_1, z_2, \ldots, z_n]^\top$ *observed, estimate the unknown parameters* $\underline{\Theta} = [w, \mu_1, \mu_2, \sigma_1, \sigma_2]^\top$*, bearing in mind the incomplete-data likelihood function:*

$$L(\underline{\Theta}; \underline{x}) = \prod_{i=1}^{n} \sum_{j=1}^{2} w_j f_j(x_i),$$

and the complete-data likelihood function,

$$L(\underline{\Theta}; \underline{x}, \underline{z}) = \prod_{i=1}^{n} \prod_{j=1}^{2} \left(w_j f_j(x_i)\right)^{\mathbb{I}(z_i = j)},$$

where $\mathbb{I}$ *is the indicator function and*

$$f_j(x) = \frac{1}{x\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2\sigma_j^2}(\ln(x) - \mu_j)^2\right\}, \quad x \in \mathbb{R}_+, \; j = 1, 2$$

is the log-normal probability density function of parameters μ_j, σ_j.

At each iteration h the EM algorithm iteratively applies the following two steps:

- E-step: Given a current estimate of the parameters $\underline{\Theta}_h = [w_h,\, \mu_{1,h},\, \mu_{2,h},\, \sigma_{1,h},\, \sigma_{2,h}]^\top$, the conditional probability of the Z_i (also called *membership probability*) is computed as follows:

$$P\left(Z_i = j \,|\, X_i = x_i, \underline{\Theta}^h\right) = \frac{w_{j,h} f_{j,h}(x_i)}{w_{1,h} f_{1,h}(x_i) + w_{2,h} f_{2,h}(x_i)},\ i = 1, 2, \ldots, n, \tag{24}$$

  where $f_{j,h}(x)$ is the log-normal probability density function of parameters $\mu_{j,h}$, $\sigma_{j,h}$, $w_{1,h} = w_h$, and $w_{2,h} = 1 - w_h$. Then using (24), the expected value of the log-likelihood function is computed as follows:

$$Q(\underline{\Theta} \,|\, \underline{\Theta}_h) = \quad E_{\underline{Z}\,|\underline{X},\underline{\Theta}_h}\left(\ln L(\underline{\Theta}; \underline{x}, \underline{Z})\right) = \sum_{i=1}^{n} \sum_{j=1}^{2} P\left(Z_i = j \,|\, X_i = x_i; \underline{\Theta}^h\right) \left[\ln w_j - \ln f_j(x_i)\right], \tag{25}$$

  where the expected value in (25) is computed with respect to the current conditional distribution of $\underline{Z} = [Z_1,\, Z_2,\, \ldots,\, Z_n]^\top$ given $\underline{X} = [X_1,\, X_2,\, \ldots,\, X_n]^\top$ and $\underline{\Theta}^h$.

- M-step: The new estimate of the parameter vector $\underline{\Theta}$ is determined by maximizing $Q(\underline{\Theta} \,|\, \underline{\Theta}_h)$ in (25):

$$\underline{\Theta}_{h+1} = \arg\max_{\underline{\Theta}} Q\left(\underline{\Theta} \,|\, \underline{\Theta}_h\right).$$

The EM algorithm stops at step h if the difference between the incomplete-data log-likelihood function at h-th and $(h-1)$-th steps is less than some small value $\epsilon > 0$:

$$\ln L(\underline{\Theta}_{h+1}; \underline{x}) - \ln L(\underline{\Theta}_h; \underline{x}) < \epsilon.$$

In our analysis, we use the Matlab function for the EM algorithm (see Ahmadzadeh, 2020).

Appendix C: Proofs

In this appendix we prove equations (16), (17) and (23) in Section 2.

Lemma *For* $j = 1, 2$ *let* $f_j(x)$, $x \in \mathbb{R}_+$, *be the log-normal probability density function of parameters* $\mu_j, \sigma_j \in \mathbb{R}$, *and let* $w \in (0, 1)$ *be a real number such that*

$$\Delta = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)^2 + \left(2 \ln\left(\frac{\sigma_2 w}{\sigma_1(1-w)}\right) + \left(\frac{\mu_1}{\sigma_1}\right)^2 - \left(\frac{\mu_2}{\sigma_2}\right)^2\right)(\sigma_2^2 - \sigma_1^2) \geq 0. \tag{26}$$

Then the solutions to the equation

$$w f_1(x) = (1-w) f_2(x), \quad x \in \mathbb{R}_+ \tag{27}$$

exist and are given by

$$a_\pm = \exp\left\{\frac{\sigma_1 \sigma_2}{\sigma_2^2 - \sigma_1^2}\left[\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) \pm \sqrt{\Delta}\,\right]\right\}. \tag{28}$$

Proof.

Eq. (27) can be rewritten as:

$$\frac{w}{a\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2\sigma_1^2}\left(\ln(a) - \mu_1\right)^2\right\} = \frac{1-w}{a\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2\sigma_2^2}\left(\ln(a) - \mu_2\right)^2\right\}. \tag{29}$$

Multiplying both sides of Eq. (29) by $a\sqrt{2\pi}$ and taking the logarithm, we obtain the following quadratic equation dependent on the variable $\ln(a)$:

$$\frac{1}{2\sigma_1^2\sigma_2^2}\left((\ln(a)-\mu_1)^2\sigma_2^2-(\ln(a)-\mu_2)^2\sigma_1^2\right)=\ln\left(\frac{\sigma_2 w}{\sigma_1(1-w)}\right),$$

whose roots are

$$\ln(a_\pm)=\frac{\sigma_1\sigma_2}{\sigma_2^2-\sigma_1^2}\left[\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)\pm\sqrt{\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)^2+\left(2\ln\left(\frac{\sigma_2 w}{\sigma_1(1-w)}\right)+\left(\frac{\mu_1}{\sigma_1}\right)^2-\left(\frac{\mu_2}{\sigma_2}\right)^2\right)(\sigma_2^2-\sigma_1^2)}\right]. \quad (30)$$

Finally, exponentiating (30) we obtain (28).

The following result follows easily from the lemma above.

Corollary *For $j=1,2$, let $f_j(x)$, $x\in\mathbb{R}_+$, be a log-normal probability density function of parameters $\mu_j,\sigma_j\in\mathbb{R}$, and let $w\in(0,1)$ be a real number such that (26) holds. Then the solutions of the inequality*

$$wf_1(x)\geq(1-w)f_2(x),\quad x\in\mathbb{R}_+ \quad (31)$$

exist and belong to the set $(0,a_-)\bigcup(a_+,+\infty)$, where $a_\pm$ are given by (28).

Appendix D: Some additional tables

Region	**2005**	**2010**	**2015**
Piedmont	4324012	4441186	4424390
Valle d'Aosta	122288	127195	128974
Lombardy	9413949	9787945	10009969
Trentino-Alto Adige	975379	1025573	1053374
Veneto	4695188	4885130	4926475
Friuli-Venezia Giulia	1203301	1230713	1225212
Liguria	1591298	1612152	1576738
Emilia-Romagna	4156718	4373262	4451351
Tuscany	3605143	3717578	3751371
Umbria	858498	899790	894983
Marche	1522956	1554629	1554833
Lazio	5256995	5669435	5887644
Abruzzo	1296301	1338438	1341122
Molise	320930	320896	312525
Campania	5777375	5818335	5861897
Apulia	4054992	4082747	4096612
Basilicata	589157	587980	576230
Calabria	2005670	2004930	1984742
Sicily	5002946	5041410	5114994
Sardinia	1644418	1671769	1669632
Italy	**58417514**	**60191093**	**60843068**

Region	**2005**	**2010**	**2015**
Piedmont	1508	1241	1319
Valle d'Aosta	405	344	316
Lombardy	2481	1895	1829
Trentino-Alto Adige	910	772	621
Veneto	1801	1445	1410
Friuli-Venezia Giulia	963	962	1100
Liguria	1006	929	996
Emilia-Romagna	1682	1424	1331
Tuscany	1560	1268	1196
Umbria	981	842	585
Marche	1095	913	1024
Lazio	1606	1376	1418
Abruzzo	541	434	434
Molise	421	390	260
Campania	1303	1254	1032
Apulia	986	954	855
Basilicata	486	439	329
Calabria	591	618	604
Sicily	1078	1112	889
Sardinia	628	535	437
Italy	**22032**	**19147**	**17985**

Region	**2005**	**2010**	**2015**
Piedmont	1504	1235	1310
Valle d'Aosta	404	343	315
Lombardy	2472	1887	1814
Trentino-Alto Adige	905	769	620
Veneto	1795	1441	1404
Friuli-Venezia Giulia	958	957	1092
Liguria	998	919	985
Emilia-Romagna	1675	1417	1322
Tuscany	1547	1261	1187
Umbria	975	836	583
Marche	1088	908	1020
Lazio	1585	1367	1399
Abruzzo	535	430	426
Molise	419	386	259
Campania	1284	1237	1016
Apulia	977	945	849
Basilicata	484	437	319
Calabria	583	608	595
Sicily	1063	1091	851
Sardinia	623	531	428
Italy	**21874**	**19005**	**17794**

Table 3: EU-SILC data: Italian Population (left panel), sample size of Italian regions (middle panel), sample size without negative and null incomes (right panel) in the years 2005, 2010, 2015.

Year	Piedmont	Valle d'Aosta	Lombardy	Trentino-Alto Adige	Veneto	Friuli-Venezia Giulia	Liguria	Emilia-Romagna
2005	4908.11	6216.16	4536.91	4140.76	5585.62	4897.16	5415.17	5414.01
	8634.28	8480.48	7453.13	8296.57	8585.55	8634.28	6870.18	8634.28
		8634.28	8634.28	8634.28	8634.28		8634.28	
2010	5009.95	6058.13	4792.20	5228.83	4920.13	5065.74	4451.33	4875.92
	9606.86	9606.86	7298.76	7384.21	7209.57	9606.86	5897.92	7534.09
			9606.86	9606.86	9606.86		9540.02	9606.86
							9606.86	
2015	5965.85	8332.25	5664.11	5169.06	5764.63	6434.26	4205.81	6559.22
	9569.71	9569.71	9569.71	8789.44	9569.71	9569.71	6183.27	9569.71
				9569.71			8362.06	
							9569.71	

Table 4: Individual equivalent disposable income thresholds of the poor population in northern Italian regions (in euro) in the years 2005, 2010, 2015.

Year	Tuscany	Umbria	Marche	Lazio
2005	4153.90	3070.54	4580.09	3314.29
	7701.67	6395.1	6786.07	8308.61
	8634.28	8634.28	8634.28	8634.28
2010	3869.58	3880.39	4931.61	3598.96
	6887.13	7335.09	6507.96	6515.77
	9606.86	9220.35	9606.86	7999.48
	.	9606.86		9606.86
2015	4641.09	5555.45	4879.49	4586.25
	8862.21	8817.98	9020.31	6942.88
	9569.71	9569.71	9569.71	9569.71

Table 5: Individual equivalent disposable income thresholds of the poor population in central Italian regions (in euro) in the years 2005, 2010, 2015.

Year	Abruzzo	Molise	Campania	Apulia	Basilicata	Calabria	Sicily	Sardinia
2005	2969.12	4797.92	2763.18	2640.11	2427.32	3358.35	2168.33	3982.08
	5983.66	5852.69	4571.80	4329.73	5458.44	4870.48	4252.35	5718.27
	8634.28	8634.28	6562.96	6205.91	6027.28	5831.45	5593.41	8634.28
			7290.40	7646.94	8365.32	8634.28	6192.81	
			8634.28	8634.28	8634.28		6838.85	
							7450.89	
							8634.28	
2010	4186.81	5998.71	3834.02	4169.84	4155.49	4540.57	2418.94	3591.45
	5763.45	6953.60	5604.60	7821.77	4877.06	5457.71	4725.04	5259.00
	9606.86	7873.07	7416.24	8876.74	5984.91	8656.00	6426.37	8054.05
	.	9606.86	7977.58	9606.86	8129.32	9606.86	7929.46	9606.86
			9606.86		8736.50		9201.47	
					9038.42		9606.86	
					9606.86			
2015	3650.78	4945.93	3904.33	3475.16	5634.82	3632.60	4064.92	4864.29
	5777.07	8130.93	6993.29	6020.47	7591.65	6962.70	7025.98	9569.71
	9167.71	9569.71	9525.18	7884.77	8811.24	9569.71	8845.98	
	9569.71		9569.71		9569.71		9569.71	

Table 6: Individual equivalent disposable income thresholds of the poor population in southern and insular Italian regions (in euro) in the years 2005, 2010, 2015.

Year	Piedmont	Valle d'Aosta	Lombardy	Trentino - Alto Adige	Veneto	Friuli - Venezia Giulia	Liguria	Emilia - Romagna
2005	34.11%	43.20%	31.53%	28.77%	38.81%	34.03%	37.63%	37.62%
	60.00%	58.93%	51.79%	57.65%	59.66%	60.00%	47.74%	60.00 %
		60.00%	60.00%	60.00%	60.00%		60.00%	
2010	31.29%	37.84%	29.93%	32.66%	30.73%	31.64%	27.80%	30.45%
	60.00%	60.00%	45.58%	46.12%	45.03%	60.00%	36.84%	47.05%
			60.00%	60.00%	60.00%		59.58%	60.00%
							60.00%	
2015	37.40%	52.24%	35.51%	32.41%	36.14%	40.34%	26.37%	41.12%
	60.00%	60.00%	60.00%	55.11%	60.00%	60.00%	38.77%	60.00%
				60.00%			52.43%	
							60.00%	

Table 7: Individual equivalent disposable income thresholds of thr poor population in northern Italian regions (as percentages of the median Italian income) in the years 2005, 2010, 2015.

Year	Tuscany	Umbria	Marche	Lazio
2005	28.87%	21.34%	31.83%	23.03%
	53.52%	44.44%	47.16%	57.74%
	60.00%	60.00%	60.00%	60.00%
2010	24.17%	24.24%	30.80%	22.48%
	43.01%	45.81%	40.65%	40.69%
	60.00%	57.59%	60.00%	49.96%
		60.00%		60.00%
2015	29.10%	34.83%	30.59%	28.75%
	55.56%	55.29%	56.55%	43.53%
	60.00%	60.00%	60.00%	60.00%

Table 8: Individual equivalent disposable income thresholds of the poor population in central Italian regions (as percentages of the median Italian income) in the years 2005, 2010, 2015.

Year	Abruzzo	Molise	Campania	Apulia	Basilicata	Calabria	Sicily	Sardinia
2005	20.63%	33.34%	19.20%	18.35%	16.87%	23.34%	15.07%	27.67%
	41.58%	40.67%	31.77%	30.09 %	37.93%	33.85%	29.55%	39.74%
	60.00%	60.00%	45.61%	43.13%	41.88%	40.52%	38.87%	60.00%
			50.66%	53.14%	58.13%	60.00%	43.03%	
			60.00%	60.00%	60.00%		47.52%	
							51.78%	
							60.00%	
2010	26.15%	37.47%	23.95%	26.04%	25.95%	28.36%	15.11%	22.43%
	36.00%	43.43%	35.00%	48.85%	30.46%	34.09%	29.51%	32.85%
	60.00%	49.17%	46.32%	55.44%	37.38%	54.06%	40.14%	50.30%
	.	60.00%	49.82%	60.00%	50.77%	60.00%	49.52%	60.00%
			60.00%		54.56%		57.47%	
					56.45%		60.00%	
					60.00%			
2015	22.89%	31.01%	24.48%	21.79%	35.33%	22.78%	25.49%	30.49%
	36.22%	50.98%	43.85%	37.75%	47.60%	43.65%	44.05%	60.00%
	57.48%	60.00%	59.72%	49.44%	55.24%	60.00%	55.46%	
	60.00%		60.00%	60.00%	60.00%			

Table 9: Individual equivalent disposable income thresholds of the poor population in southern and insular Italian regions (as percentages of the median Italian income) in the years 2005, 2010, 2015.

Region	Year	a^1	a^2	a^3	a^4	a^5	a^6
Piedmont	2005	4908.11(0.0031)	10527.79(0.0392)	13586.85(0.0578)	14536.95(0.0168)	17815.89(0.1376)	
Piedmont	2010	5009.95(0.0027)	14672.03(0.1229)	16211.3(0.0235)	16892.91(0.0097)	19466.77(0.0778)	
Piedmont	2015	5965.85(0.0051)	11923.59(0.042)	15204.99(0.0328)	16776.97(0.0238)	18426.36(0.0349)	
Valle d'Aosta	2005	6216.16(0.119)	8480.48(0.0051)	11259.4(0.0292)	12070.35(0.006)	16365.75(0.1904)	17121.42(0.0165)
Valle d'Aosta	2010	6058.13(0.0027)	10888.77(0.0157)	16477.45(0.0651)	18806.74(0.0311)	21428.34(0.1408)	
Valle d'Aosta	2015	8332.25(0.0127)	10745.29(0.0116)	14418.33(0.0563)	15990.66(0.0262)	17190.47(0.0246)	18829.94(0.043)
Lombardy	2005	4536.91(0.0024)	7453.13(0.0073)	10628.3(0.0241)	12326.83(0.0158)	13409.01(0.0114)	14981.08(0.0299)
Lombardy	2010	4792.2(0.002)	7298.76(0.0032)	10256.69(0.0138)	14906.52(0.0533)	16408.82(0.0158)	18077.17(0.0276)
Lombardy	2015	5664.11(0.0045)	9805.68(0.0156)	13595.7(0.0348)	16877.66(0.0505)	18891.63(0.032)	22881.33(0.1013)
Trentino-Alto Adige	2005	4140.76(8e-04)	8296.57(0.0171)	12719.32(0.0594)	14311.09(0.0312)	15991.57(0.0451)	19049.24(0.1537)
Trentino-Alto Adige	2010	5228.83(0.0013)	7384.21(0.0012)	13309.49(0.0452)	17318.34(0.0529)	19073.05(0.0291)	22778.22(0.1326)
Trentino-Alto Adige	2015	5169.06(0.0015)	8789.44(0.007)	13594.76(0.0416)	18158.08(0.0505)	19333.94(0.0142)	20717.5(0.0391)
Veneto	2005	5585.62(0.0057)	8585.55(0.0153)	11484.86(0.0374)	12937.54(0.0245)	15981.75(0.1072)	
Veneto	2010	4920.13(0.0015)	7209.57(0.002)	11066.27(0.0233)	12966.49(0.0176)	15846.14(0.0477)	17514.82(0.0259)
Veneto	2015	5764.63(0.0033)	13926.18(0.1079)	16181.74(0.0361)	17586.82(0.0261)	18741.91(0.0232)	
Friuli-Venezia Giulia	2005	4897.16(0.0030)	11963.99(0.0724)	13672.25(0.0224)	14259.34(0.0050)	15257.26(0.0175)	16447.71(0.0279)
Friuli-Venezia Giulia	2010	5065.74(0.0022)	10495.14(0.0293)	13259.76(0.0222)	15213.45(0.0218)	16213.16(0.0141)	17577.85(0.0321)
Friuli-Venezia Giulia	2015	6434.26(0.004)	11323.8(0.0256)	16309.08(0.0847)	18323.27(0.0384)	19339.81(0.0194)	
Liguria	2005	5415.17(0.0075)	6870.18(0.0033)	11117.1(0.0509)	14614.33(0.0808)	15418.24(0.0216)	
Liguria	2010	4451.33(0.0012)	5897.92(3e-04)	9540.02(0.0187)	11254.59(0.0109)	14072.85(0.0364)	15842.63(0.0268)
Liguria	2015	4205.81(0.0027)	6183.27(0.002)	8362.06(0.0068)	10834.75(0.0172)	13581.09(0.031)	14906(0.0121)
Emilia - Romagna	2005	5414.01(0.0034)	11492.77(0.0467)	15035.28(0.0489)	16943.9(0.0385)	17677.3(0.0143)	18912.96(0.0372)
Emilia-Romagna	2010	4875.92(0.0014)	7534.09(0.0033)	10008.3(0.0091)	13519.02(0.0352)	17294.06(0.0527)	20548.6(0.0879)
Emilia-Romagna	2015	6559.22(0.0059)	11465.2(0.0248)	14355.88(0.0228)	16582.08(0.0246)	17778.21(0.0139)	18974.5(0.0232)
Tuscany	2005	4153.9(9e-04)	7701.67(0.0098)	9487.97(0.0083)	11473.76(0.0198)	13052.76(0.0222)	14751.54(0.0313)
Tuscany	2010	3869.58(7e-04)	6887.13(0.0036)	10459.16(0.0208)	13187.93(0.0283)	14854.88(0.0199)	16374.84(0.0225)
Tuscany	2015	4641.09(0.0018)	8862.21(0.0139)	14336.78(0.0523)	17687.82(0.0556)	18659.71(0.0123)	19402.15(0.0106)
Umbria	2005	3070.54(7e-04)	6395.1(0.0088)	8908.68(0.0175)	10500.73(0.0138)	11878.23(0.0212)	12344.94(0.005)
Umbria	2010	3880.39(7e-04)	7335.09(0.008)	9220.35(0.0089)	13351.52(0.0634)	15036.2(0.025)	16163.97(0.0173)
Umbria	2015	5555.45(0.0063)	8817.98(0.0114)	11208.97(0.0163)	13577.91(0.0375)	15100.48(0.0202)	16028.75(0.0169)
Marche	2005	4580.09(0.0026)	6786.07(0.0042)	9929.1(0.0239)	12054.64(0.0273)	12659.68(0.0072)	14609.47(0.0531)
Marche	2010	4931.61(0.0025)	6507.96(7e-04)	10951.9(0.0475)	14476.33(0.0656)	18265.85(0.0606)	
Marche	2015	4879.49(0.0034)	9020.31(0.0182)	10807.37(0.0136)	12734.48(0.0229)	14071.38(0.0186)	15482.42(0.0255)
Lazio	2005	3314.29(0.0017)	8308.61(0.0385)	9782.58(0.0123)	10728.01(0.0107)	15025.14(0.0828)	16659.55(0.0353)
Lazio	2010	3598.96(0.0015)	6515.77(0.0059)	7999.48(0.0043)	11416.6(0.0382)	14681.59(0.0532)	16701.77(0.0338)
Lazio	2015	4586.25(0.0062)	6942.88(0.0059)	10479.14(0.0275)	12411.48(0.0184)	13421.1(0.0085)	15920.22(0.0528)
Abruzzo	2005	2969.12(8e-04)	5983.66(0.0089)	9829.73(0.0463)	11489.69(0.0221)	12578.93(0.0168)	16038.42(0.0721)
Abruzzo	2010	4186.81(0.0019)	5763.45(0.0012)	10098.34(0.0391)	13819.75(0.048)	15225.21(0.0188)	16391.04(0.0236)
Abruzzo	2015	3650.78(0.0018)	5777.07(0.0029)	9167.71(0.0334)	14235.19(0.106)	17570.72(0.0717)	
Molise	2005	4797.92(0.0108)	5852.69(0.0023)	15018.9(0.2701)			
Molise	2010	5998.71(0.0155)	6953.6(0.0017)	7873.07(0.0039)	9703.85(0.0213)	14467.26(0.2418)	
Molise	2015	4945.93(0.0096)	8130.93(0.0188)	11171.95(0.0362)	11663.99(0.0032)	13682.65(0.0599)	
Campania	2005	2763.18(0.0038)	4571.8(0.0072)	6562.96(0.0267)	7290.4(0.0085)	9350.15(0.0861)	13334.86(0.1637)
Campania	2010	3834.02(0.0081)	5604.6(0.0082)	7416.24(0.022)	7977.58(0.0055)	10316.43(0.0482)	11357.11(0.0202)
Campania	2015	3904.33(0.0096)	6993.29(0.0296)	9525.18(0.0545)	11808.07(0.051)		
Apulia	2005	2640.11(0.0025)	4329.73(0.0054)	6205.91(0.0148)	7646.94(0.0167)	9007.87(0.0232)	10886.69(0.041)
Apulia	2010	4169.84(0.0062)	7821.77(0.026)	8876.74(0.0073)	10008.18(0.0148)	12533.03(0.0504)	13542.25(0.0207)
Apulia	2015	3475.16(0.0045)	6020.47(0.0102)	7884.77(0.019)	10899.74(0.0761)	11537.55(0.0103)	13709.22(0.0694)
Basilicata	2005	2427.32(0.0014)	5458.44(0.0206)	6027.28(0.0021)	8365.32(0.0582)	11396.69(0.0868)	
Basilicata	2010	4155.49(0.0051)	4877.06(2e-04)	5984.91(0.0027)	8129.32(0.026)	8736.5(0.0052)	9038.42(0.0017)
Basilicata	2015	5634.82(0.0221)	7591.65(0.0098)	8811.24(0.0114)	16768.52(0.1082)		
Calabria	2005	3358.35(0.0101)	4870.48(0.0067)	5831.45(0.0071)	9239.28(0.0895)	10172.96(0.0153)	11118.93(0.018)
Calabria	2010	4540.57(0.0171)	5457.71(0.0025)	8656(0.0464)	10876.67(0.0465)	12455.14(0.0344)	
Calabria	2015	3632.6(0.0085)	6962.7(0.0319)	10478.81(0.0817)	11281.77(0.0111)	12801.1(0.0381)	
Sicily	2005	2168.33(0.0028)	4252.35(0.0127)	5593.41(0.015)	6192.81(0.0065)	6838.85(0.0109)	7450.89(0.0117)
Sicily	2010	2418.94(0.0027)	4725.04(0.0115)	6426.37(0.0187)	7929.46(0.022)	9201.47(0.023)	12251.3(0.0746)
Sicily	2015	4064.92(0.0193)	7025.98(0.032)	8845.98(0.0343)	11177.8(0.0538)		
Sardinia	2005	3982.08(0.0075)	5718.27(0.0051)	9961.25(0.0746)	11226.35(0.02)	11603.02(0.0048)	12240.14(0.0141)
Sardinia	2010	3591.45(8e-04)	5259(6e-04)	8054.05(0.0128)	10480.24(0.0286)	14306.07(0.0749)	14854.94(0.0054)
Sardinia	2015	4864.29(0.0092)	10167.06(0.0852)	14763.4(0.1241)			

Table 10: Individual equivalent disposable income stratification of the Italian regions and miss-identification errors (in brackets) in the years 2005, 2010, 2015.

References

[1] Ahmadzadeh, R. (2020). Expectation maximization algorithm (https://www.mathworks.com/matlabcentral/fileexchange/65772-expectation-maximization-algorithm), MATLAB Central File Exchange. Retrieved March 29, 2020.

[2] Anikin, V. A., Lezhnina, Y. P., Mareeva, S. V., Slobodenyuk, E. D., Tikhonova, N. N. (2016). Income stratification: key approaches and their application to Russia. Working paper. National Research University - Higher School of Economics, Moscow.

[3] Bellettini, G., Ceroni, C. B. (2007). Income distribution, borrowing constraints and redistributive policies. *European Economic Review* 51(3), 625–645.

[4] Biggeri, L., Giusti, C., Marchetti, S., Pratesi, M. (2018). Poverty indicators at local level: definitions, comparisons in real terms and small area estimation methods. *Statistics and Applications* 16(1), 351–364.

[5] Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio. *Empirical Economics* 22(4), 515–531.

[6] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39(1), 1–38.

[7] Dynan, K. E., Skinner, J., Zeldes, S. P. (2004). Do the rich save more?. *Journal of Political Economy* 112(2), 397–444.

[8] Eisenhauer, J. G. (2011). The rich, the poor, and the middle class: Thresholds and intensity indices. *Research in Economics* 65(4), 294–304.

[9] European Commission (EC). (2011). Regions in the European Union. Nomenclature of territorial units for statistics NUTS 2010 EU-27.

[10] Eurostat (2006). *Some proposals on the treatment of negative incomes*. EU-SILC Documents TFMC-15/06, European Commission, Eurostat.

[11] Feenberg, D. R., Poterba, J. M. (2000). The income and tax shares of very high-income households, 1960-1995. *American Economic Review* 90(2), 264–270.

[12] Foster, J. E. (1998). Absolute versus relative poverty. *The American Economic Review* 88(2), 335–341.

[13] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Springer Series in Statistics. Springer-Verlag, New York.

[14] Gini, C. (1912). *Variabilità e mutuabilità. Contributo allo studio delle distribuzioni e delle relazioni statistiche*. C. Cuppini, Bologna.

[15] Graf, M., Wenger, A., Nedyalkova, D. (2011). Deliverable 5.1: Quality of EU-SILC data. Report from the EU's FP7 Programme project "AMELI-Advanced Methodology for European Laeken Indicators", University of Trier.

[16] Jayaraj, D., Subramanian, S. (2006). Horizontal and vertical inequality: some interconnections and indicators. *Social Indicators Research* 75(1), 123–139.

[17] Jędrzejczak, A. (2014). Income inequality and income stratification in Poland. *Statistics in Transition new series* 15(2), 269–282.

[18] Klecka, W. R. (1980). *Discriminant Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-019, Beverly Hills, CA: Sage Publications.

[19] Liao, T. F. (2006). Measuring and analyzing class inequality with the Gini Index informed by model-based clustering. *Sociological Methodology* 36(1), 201–224.

[20] López-Calva, L., Ortiz-Juarez, E. (2014). A vulnerability approach to the definition of the middle class. *Journal of Economic Inequality* 12(1), 23–47.

[21] Lung-Yut-Fong, A., Lévy-Leduc, C., Cappé, O. (2015). Homogeneity and change point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique* 156(4), 133–162.

[22] Mann, M. (1984). *The International Encyclopedia of Sociology*. New York: Macmillan.

[23] McLachlan, G. J. , Peel, D. (2000). *Finite mixture models*, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. .

[24] McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification* 33(3), 331–373.

[25] Medeiros, M. (2006). *Poverty, inequality, and redistribution: a methodology to define the rich*. United Nations Development Programme, International Poverty Center Working Paper 18, Brazil.

[26] Niño-Zarazúa, M., Roope, L., Tarp, F. (2017). Global inequality: Relatively lower, absolutely higher. *Review of Income and Wealth* 63(4), 661–684.

[27] OECD (2019). *Under pressure: The squeezed middle class*. OECD Publishing, Paris.

[28] Page, E. S. (1955). Control charts with warning lines. *Biometrika* 42(1-2), 243–257.

[29] Peichl, A., Schaefer, T. , Scheicher, C. (2010). Measuring richness and poverty: a micro data application to Europe and Germany. *Review of Income and Wealth* 56(3), 597–619.

[30] Perotti, R. (1993). Political equilibrium, income distribution, and growth. *Review of Economic Studies* 60(4), 755-776.

[31] Persson, T., Tabellini, G. (1994). Is inequality harmful for growth? *American Economic Review* 84(3), 600-621.

[32] Pittau, M. G., Zelli, R., Johnson, P. A. (2010). Mixture models, convergence clubs, and polarization. *Review of Income and Wealth* 56(1), 102–122.

[33] Pittau, M. G., Zelli, R. (2014). Poverty status probability: a new approach to measuring poverty and the progress of the poor. *The Journal of Economic Inequality* 12(4), 469–488.

[34] Pressman, S. (2007). The decline of the middle class: an international perspective. *Journal of Economic Issues* 41(1), 181–200.

[35] Profeta, P. (2007). Political support and tax reforms with an application to Italy. *Public Choice* 131(1-2), 141–155.

[36] Schotte, S., Zizzamia, R., Leibbrandt, M. (2018). A poverty dynamics approach to social stratification: The South African case. *World Development* 110, 88–103.

[37] Stahl, D., Sallis, H. (2012). Model-based cluster analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 4(4), 341–358.

[38] Stewart, F., Brown, G., Mancini, L. (2005). Why Horizontal Inequalities Matter: Some Implications for Measurement. CRISE WORKING PAPER No. 19, University Oxford, 1–30.

[39] Van Kerm, P. (2007). *Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC.* IRISS Working Paper Series 01, CEPS/INSTEAD.

[40] Yitzhaki, S., Lerman, R. I. (1991). Income stratification and income inequality. *Review of Income and Wealth* 37(3), 313–329.

[41] Zhou, X., Wodtke, G.T. (2019). Income stratification among occupational classes in the United States. *Social Forces* 97 (3), 945–972.