UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Ungrouping Income Distributions – The Italian Doxa Survey of 1948

Matteo Santi

Supervisor: Ola Olsson (Univ. of Gothenburg), Giovanni Vecchi (Univ. of Rome Tor Vergata)
Master's thesis in Economics, 30 hec
Spring 2020
Graduate School, School of Business, Economics and Law
University of Gothenburg, Sweden

Abstract

This paper investigates alternative statistical approaches to ungroup data in tabular form. After a theoretical discussion on the interpolation problem and on the features of ungrouping techniques, a non-parametric version of the algorithm of Shorrocks and Wan is introduced. The effectiveness of the different methods is assessed using recent microdata on Italian incomes available in the Bank of Italy's *Survey on Households Income and Wealth*. Taking advantage of this evaluation, the most suitable ungrouping methods are applied to the Doxa Survey of 1948, the first research on Italian households' incomes based on a probability sampling procedure. Lastly, the reconstructed samples of historical microdata are used to compute inequality and poverty measures.

Keywords: Data Ungrouping, Quantitative Economic History, Inequality, Poverty

1 Introduction

The estimation of inequality and poverty levels is a crucial part of the assessment of the economic well-being of a society. The most widely used metrics to measure the dispersion of income and the proportion of poor in the population, respectively the Gini (1912) index and the headcount ratio (the proportion of individuals that cannot satisfy their basic needs), share a common drawback: their estimation requires a significant amount of data, namely the full distributions of earnings or, alternatively, a representative sample of the population. However, in both cases, microdata (namely, information at unit level) on income are needed.

Unfortunately, the availability of information on the distribution of income is sometimes limited: instead of presenting data at the unit level, documents report them in *grouped* form, so that what is available is only the number of people in each interval of income and the mean income of each class. This situation is common when data on income come from developing or authoritarian countries, or when independent researchers are forced for some reason to summarize their findings by making use of tables and histograms (Shorrocks and Wan, 2008). In economic history, microdata on large groups of units drawn with a probabilistic sampling procedure are rare: when there are information on a large number of people, they are usually available only in tabular form. In these cases, reseachers have two possibilities: the first consists in limiting their analysis to measures that do not require data at unit level, such as the top-decile income or wealth shares, or the interquantile range. This approach is widely used when ancient/limited data are studied, as for instance in Piketty (2014), and does not require any further modification of the available data, nor assumptions on the shape of the function describing the distribution of income. On the other hand, the original distribution of income can be reconstructed from grouped data following several different approaches, in order to obtain a simulated sample of microdata that researchers can use to compute inequality and poverty measures, so to draw more precise conclusions on the distribution of income in remote times.

Following the second of these approaches, this paper analyses alternative techniques to ungroup data in tabular form: the parametric algorithm of Shorrocks and Wan and two non-parametric methods, the Hermite spline interpolation and the bootstrap kernel density estimation. Moreover, a non-parametric version of the algorithm of Shorrocks and Wan is introduced. These techniques are firsty presented from a theoretical point of view, and then evaluated on samples of microdata on Italian income earners; subsequently, the most appropriate techniques are then used to ungroup the data of the Doxa Survey of 1948, the first research on Italian households' incomes based on a probability sampling procedure (Brandolini, 1999; Vecchi, 2017). The obtained samples are finally used to compute inequality levels and absolute poverty rates. Therefore, besides being a statistical study on data ungrouping, the paper also contributes to the literature on the Italian economic history of inequality and poverty.

Since the data gathered in the Doxa Survey of 1948 are available only in tabular form, the estimates of inequality and poverty metrics for Italy based on the survey may vary according to the chosen ungrouping technique; the benchmark of this part of the analysis are the estimates by Vecchi (2017). Given

the relevance of this period of the Italian contemporary economic history (a few years after World War II, and immediately preceding the huge industrial development that the country experienced between the fifties and the seventies), it is of great importance to clarify the levels of inequality and poverty in this year, both to understand the distributional implications of a long period of severe conflict and to see in a new light the effects of the subsequent phase of rapid economic development (known in Italy as *miracolo economico*, i.e. "economic miracle"). The features of the dataset allow not only to reconstruct such measures for the country as a whole, but also to provide insights on which regions presented a more concentrated distribution of income and which ones had a higher percentage of poor households. Moreover, it is possible to carry out an analysis of the North-South divide in the post-war period, in terms of inequality levels and poverty rates.

The paper is structured as follows. The Literature Review gives an overview of the most used statistical techniques to ungroup data in tabular form and of their applications in international researches; in the Theoretical Framework, the problem is presented from a theoretical point of view and the techniques used later in the following analyses are explained. After the preliminary explanations presented in the Methodology and Data, in the Analysis of the SHIW Database recent samples of income earners in Italy are used to assess how well the original distribution of income is reconstructed. Once obtained a valid understanding of the most suitable methods, in the Analysis of the Doxa Survey, the study focuses on the grouped data referred to the survey of 1948, and interprets the results in terms of inequality and poverty across different regions of Italy after the Second World War. The Conclusion sums up the most relevant statistical and historical findings of the research.

2 Literature Review

The literature on the estimation of inequality measures in the presence of data in tabular form is vast and diverse. In one of the first contributions, Morgan (1962) assessed that the lower bound for the estimates of any inequality measure in presence of grouped data is represented by the dispersion *between* classes: this minimum estimate is correct only assuming that there is perfect equality *within* each of the groups. On the contrary, Gastwirth (1972) found an upper bound for dispersion, represented by the case in which the spread within each interval is maximum. The approach of Kakwani (1976) consists in fitting a polynomial function of third degree to represent the Lorenz curve and derive an estimated income density function, that can be used to compute inequality measures; finally, a correction is made for the extreme income ranges, that are treated by fitting a Pareto curve.
Following a similar approach, many functions have been proposed to approximate the distribution of income: Thurow (1970) used a Beta distribution, while Bartels and van Metelen (1975) proposed to use a Weibull. Singh and Maddala (1976) developed a new function, that bears their names, and so did Dagum (1977) who created a function known as Burr 3. McDonald (1984) introduced the Generalized Beta distributions, of which the previously mentioned functions are particular or limiting cases (Bandourian *et al.*, 2002). Another widely applied function, the Generalized Quadratic distribution, was proposed by Villaseñor and Arnold (1989).

In their study on the estimation of inequality, Cowell and Mehta (1982) compared the results obtained by using three different interpolation methods: (*1*) a piecewise Paretian interpolation, (*2*) a polynomial interpolation, and (*3*) a split-histogram interpolation. They found that these approaches provided similar estimates for the Gini coefficient, arguing in favour of the least computationally demanding one, the split-histogram interpolation. On the other hand, estimates of measures with higher inequality aversion (such as the Atkinson index with high values of ε) were found to be unreliable in the absence of available microdata. A similar technique, based on a quadratic interpolation of the Lorenz curve, is followed by Tillé and Langel (2012).

The approach of Shorrocks and Wan (2008) consists in drawing random samples from different distributions fitted on the grouped data in order to create simulated samples of unit data. Subsequently, these samples are corrected to obtain sample statistics that match the original figures. Since the algorithm requires an assumption on the shape of the starting distribution from which the samples are drawn, they test the performance of a group of functional forms in two ways. First, they compare a series of estimated inequality measures (Gini index, mean logarithmic deviation, Theil coefficient and squared coefficient of variation) reconstructed from grouped data with their known values, obtaining very encouraging results and small differences between the chosen distributions. Then, they directly compare all the reconstructed observations with their original counterparts, in order to identify the most problematic deciles to reconstruct. The tested distributions are the log-normal (Aitchinson and Brown, 1957), the Singh-Maddala (1976) and the Generalized Beta (McDonald, 1984), starting from deciles and quintiles shares. The analyses of this paper are similar to those of Shorrocks and Wan, but the study is extended in terms of grouping patterns (four, five, ten, fifteen and twenty groups) and of starting samples, that are here obtained non-parametrically as well.

All the methods described so far require a starting assumption on the distribution that has to be reconstructed. Higher degrees of complexity of the fitted distributions, obtained by increasing the number of parameters to create generalized versions of known density function, are a possibility that has been widely explored in the literature. The log-normal and the Beta distributions, both with two parameters have been compared to three-parameter distribution such as the Singh-Maddala. Moreover, generalizations of this latter have been developed in order to gain flexibility in the definition of the shape of the density function, therefore complicating the functional form and multiplicating the number of parameters to estimate. McDonald (1984) proposed the four-parameter family of Generalized Beta functions, of which the Beta, the Singh-Maddala and the Generalized Gamma are particular or limiting cases. The sophistication of this family of function went even further with the six-parameter compound confluent hypergeometric distribution, introduced by Gordy (1998).

However, the disaggregation of data in grouped form can be carried out also without specifying assumptions on the functional form of the distribution to be reconstructed. Non-parametric approaches to data ungrouping can be divided, as Rizzi *et al.* (2016) do, into three major groups. The first is kernel density estimators, namely those who make use of a kernel density obtained from a histogram and draw samples from it with different procedures, such as bootstrapping (Wang and Wertelecki, 2013). An alternative is

the use of spline interpolations of second or third degree (Gastwirth and Glauberman, 1976), also with the use of a Hyman (1983) filter to apply monotonicity constraints on the reconstructed distribution. Finally, in a medical framework, Rizzi *et al.* (2015) use a penalized composite link model that reconstructs a distribution of aggregated counts (such as deaths), treating them as realizations of a Poisson distribution. A radically different strategy is chosen by Cannari and D'Alessio (2018) in their study of inequality in Italy between 1968 and 1975: instead of starting from a theoretical function, they gather observations from an observed distribution of income that is supposed to resemble the one that has to be reconstructed (in this case, the distribution of income in the same country in more recent years). Then, the sample is reproportioned and the observations are reweighted according to some known characteristics of the population. First, the weights are adjusted according to the (known in grouped form) totals of the marginal distribution of one variable of interest, such as income. Then, these weights are adjusted according to the totals of another variable (for instance, a demographic one), but this second procedure may lead to inconsistencies with respect to the first one. Therefore, this cycle of iteration is repeated until all the constraints posed by the marginal distributions of the variables of interest are satisfied. This procedure, known as *raking* (Deville and Sarndal, 1992; Anderson and Fricker, 2015) delivers a sample of microdata that allows to consistently estimate the inequality measures of interest. Raking procedures have been criticized by Brick *et al.* (2003), who mainly base their judgements on the possibility that the structure imposed to the survey estimates does not correspond to the actual structure of the survey data.

Applications of parametric ungrouping techniques are very common in international researches. The World Bank estimates poverty and inequality measures with the POVCAL software, that fits the distribution of income using Generalized Quadratic and Beta distributions in order to obtain simulated microdata starting from grouped observations. Datt (1998) summarizes the formulae to compute the Gini index and the first and second derivatives of the Lorenz Curve. A critical review of this method based on Monte Carlo simulations can be found in Minoiu and Reddy (2009), who find that this technique loses much of its precision in the presence of multimodal distribution functions.
The study on poverty and inequality in Africa by Boukaka *et al.* (2018) is an example of a concrete application of the algorithm developed by Shorrocks and Wan: in this case, since the research uses data that are unlikely to be representative of the population, a post-stratification procedure follows the ungrouping of the data in tabular form.
An example of application of non-parametric methods to ungroup data is the estimation of the world income distribution carried out by Sala-i-Martin (2006), who estimates first a series of kernel densities (one for each country), and then collapses these functions into a unique one, with a procedure that he defines "kernel of kernels". The accuracy of this technique is strongly criticized by Miniou and Reddy (2008), whose study focuses on the precision of the kernel ungrouping method in the estimation of poverty measures.

As previously stated, this work extends the analysis of Shorrocks and Wan (2008) by studying the precision of their ungrouping algorithm starting from data in four, five, ten, fifteen and twenty groups; this assessment is carried out for the first time on samples of Italian income earners (the SHIW database), in terms of inequality levels. Moreover, the technique is evaluated also for its accuracy in an aspect that

had not studied before: the estimation of poverty levels. The non-parametric extension of this algorithm, finally, allows to obtain results that have the desired properties of the original technique, but that do not need any parametric assumption on the underlying distribution. The choice of the other tested techniques (Hermite spline interpolation and bootstrap kernel density estimation) is motivated both by a theoretical interest in the use of non-parametric methods and by issues on the availability of data. Alternative non-parametric techniques to ungroup data, such as raking, require a sample of microdata from a distribution that is supposed to be similar to the one that has to be reconstructed, as mentioned earlier: since the final objective of this work is studying the Doxa Survey of 1948, this method has not been analysed for the lack of representative samples of Italian income earners in unit form before 1977. On the other hand, simpler techniques as spline interpolations and kernel methods require less information and, especially if associated with some adjustments, deliver satisfying results in many cases. These methods are commonly used in heterogeneous fields, as for instance in medicine (Rizzi *et al.*, 2016), where aspects such as the estimation of measures like the Gini index are seldom deemed important: this work assesses their precision for inequality and poverty metrics. As Section 5 will show, the structure of the Doxa database, that reports data in 21 classes at national level and in 18 at sub-regional level, allows to obtain precise estimates of inequality and poverty measures, whose robustness is increased by the fact that the different ungrouping techniques provide results that differ in a very negligible way. This dataset has been analysed at household level by Brandolini (1999), who ungrouped it with a linear interpolation technique and a Paretian correction for the top interval in order to compute inequality measures at household level, and by Vecchi (2017), who used the parametric algorithm of Shorrocks and Wan. This paper ungroups the dataset using eight different techniques, and therefore attempts to clarify the conclusions that can be drawn from it in terms of inequality and poverty, at national as well as at regional level.

3 Theoretical Framework

This section begins by introducing the general problem of ungrouping of data and interpolation of a density function from units in tabular form. Subsequently, the approaches to the problem that will be used in the analysis of the SHIW database and of the Doxa Survey are discussed from a theoretical point of view.

3.1 General setting

The theoretical framework of the research is a general problem of interpolation of an unknown continuous distribution. Following the description of Cowell and Mehta (1982), the analysis begins from a set S of grouped observations of points in the interval $[0, \infty)$, and aims to approximate a continuous distribution defined as $\hat{f}(x)$.

The values are assumed to be non-negative ($y \in \mathbb{R}^+ \ \forall y \in S$), since the aim is to reconstruct a distribution of incomes; they represent the support of an unknown density function, but only their belonging to one of the ω classes of income is observed. These classes of income are defined as exclusive sets closed to the left:

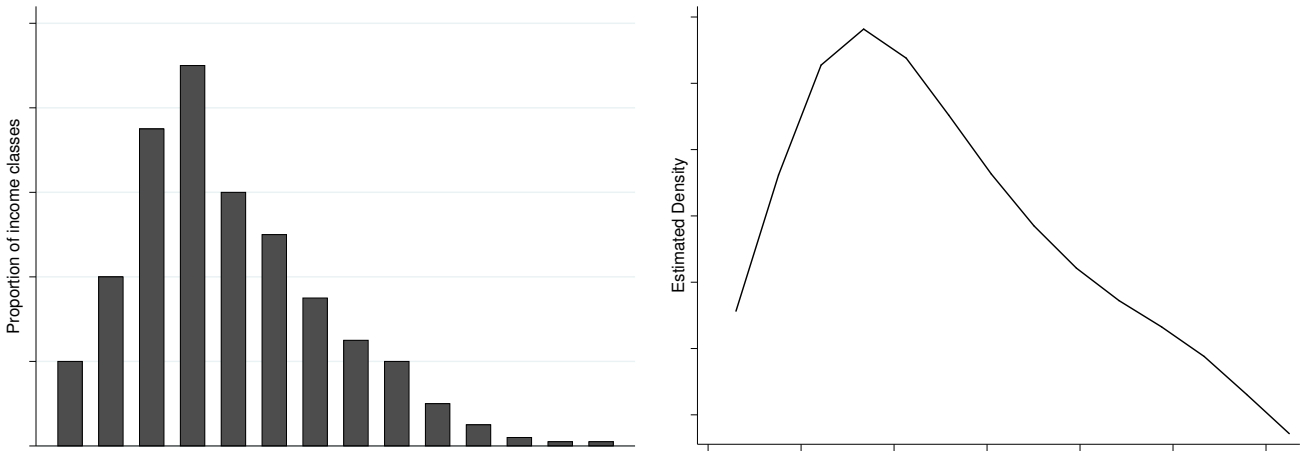
$$[a_1,\ a_2),\ [a_2,\ a_3),\ [a_3,\ a_4),\ ..\ ,\ [a_\omega,\ a_{\omega+1}) \quad \text{with} \quad 0 \leq a_1 < a_2 < .. < a_\omega < a_{\omega+1} \leq \infty$$

Together with the class boundaries a_θ, with $\vartheta = 1, .., \omega+1$, the additional available information are the two vectors representing, respectively, the number of units belonging to each class n_θ and the mean income of each of these groups, μ_θ. From a very general perspective, observations can be assigned in any possible manner to each level of income y, with the limit represented by the two following constraints. For $\vartheta = 1, .., \omega$:

(a) The number of units in interval ϑ has to equal n_θ

(b) Each interval mean has to be equal to μ_θ

Cowell and Mehta state a list of other desirable properties for the reconstructed density function: it should assume non-negative values in all the points of its support, be continuous within each interval and differentiable in each class boundary a_θ. Moreover, its limit should be 0 for $a_{\omega+1} \to \infty$; on the other hand, the condition of existence of a closed form integration of inequality measures may sound obsolete, given the development of computational algorithms since the publication of the paper[1].



(a) Histogram from a dataset in tabular form

(b) Reconstructed continuous density

Figure 1: An example of data ungrouping

The two "minimal" requirements for the hypothetical reconstructed distribution (a) and (b) permit to have a lower and an upper bound for any inequality measure I. Before stating these limits, it is appropriate to define more precisely the inequality measures we are dealing with. An inequality measure I is a S-convex function[2] from the space of incomes to the real axis (Cowell and Mehta, 1982):

$$I = f : \mathbb{R}_+^n \to \mathbb{R}$$

[1]Since 1982, namely when Cowell and Mehta published their article "The Estimation and Interpolation of Inequality Measures", a large number of computational methods to estimate inequality and poverty measures have been developed. For the aims of this paper, the estimation of the Gini index from samples of data has been carried out using the Stata command *igini*, available in the *DASP* package, developed by Araar Abdelkrim.

[2]As stated by Hudzik and Maligranda (1994, p.1), "a function f: $\mathbb{R}_+ \to \mathbb{R}$, where $\mathbb{R}_+ \equiv [0, \infty)$, is said to be *s-convex in the first sense* if $f(\alpha u + \beta v) \leq \alpha^s f(u) + \beta^s f(v)$ for all $u, v \in \mathbb{R}_+$ and all $\alpha, \beta \geq 0$ with $\alpha^s + \beta^s = 1$." In the particular case in which $s = 1$, S-convexity is equivalent to "ordinary" convexity.

The space of incomes $\mathbb{R}_+^n$ is defined only for non-negative values for the aforementioned reason that they represent incomes, and has a n-infinite dimension, since the measure I is required to allow for the evaluation of every possible combination of earnings across the n units.

Among the alternative inequality measures I, following the approach of Cowell and Mehta, we focus on the subset of metrics that are decomposable by non-overlapping population subgroups. These latter, among which there are the Gini Index and the class of Generalized Entropy Indices, are defined as *NODI* (Non-Overlapping Decomposable Inequality) measures.

$$I^G = \frac{1}{2\mu} \int_0^\infty \int_0^\infty |y - z| \; f(y) \; f(z) \; dy \; dz \qquad \text{(Gini Index)}$$

$$I^\beta = \frac{1}{\beta[\beta+1]} \left[\int_0^\infty \left[\frac{y}{\mu} \right]^{\beta+1} f(y) \; dy - 1 \right] \qquad \text{(Generalized Entropy)}$$

Many commonly used metrics can be derived from these formulae. For instance, the Theil Index is a particular case of the class of Generalized Entropy Indices with $\beta = 0$, the Atkinson Index is obtained by setting $\varepsilon = -\beta$, while the coefficient of variation corresponds to $\left[2I^2\right]^{\frac{1}{2}}$.

For any I belonging to the *NODI* set, the lower bound in the presence of grouped observations corresponds to the extreme case in which there is perfect equality within each class ϑ. In this case, the total inequality level among units is equal to the dispersion between the groups. This situation corresponds to the density function:

$$\begin{aligned} f_1(x) &= \frac{n_\theta}{n} \quad \text{if } y = \mu_\theta, & \vartheta = 1, .., \omega \\ &= 0 \quad \text{otherwise} \end{aligned}$$

The antithetical extreme case corresponds to a situation in which there is *full* dispersion within each of the ω classes. The related density function is:

$$\begin{aligned} f_2(x) &= \lambda_\theta \; \tfrac{n_\theta}{n} & \text{if } y = a_\theta, & & \vartheta = 1, .., \omega \\ &= [1 - \lambda_\theta] \; \tfrac{n_\theta}{n} & \text{if } y = a_{\theta+1}, & & \vartheta = 1, .., \omega \\ &= 0 & \text{otherwise} & & \\ \text{where} \quad & n \equiv \textstyle\sum_{\theta=1}^{\omega} n_\theta, \quad \lambda_\theta \equiv \dfrac{a_{\theta+1} - \mu_\theta}{a_{\theta+1} - a_\theta} & & & \end{aligned}$$

Under conditions (a) and (b), any estimated metrics $\hat{I}$ will be such that $I_1 \leq \hat{I} \leq \mathrm{I}_2$.

For what regards poverty measures, it is not possible to state similar inequalities. In this paper, the analysis will focus on the most commonly used metrics, the headcount ratio, defined as the proportion of population below the poverty line. The measure is defined in formal terms as:

$$HCR = \int_0^c f(y) \; dy$$

Where c is a poverty threshold that can be defined in absolute or relative terms. As made clear by the formula, this measure can either be under- or over-estimated both by assuming full equality within each class of income as in $f_1(x)$, and in the opposite case $f_2(x)$, in which the dispersion within each group is maximum. In both cases, the estimate of the ratio can be too high or too low, depending on the choice of the poverty threshold, the mean income of the class in which this threshold falls and on the grouping pattern, namely on the number and size of classes.

3.2 Statistical approaches to data ungrouping

Once having established some *minimal* requirements for the hypothetical, reconstructed density function $f(x)$, in this section some approaches to estimate and draw a sample from it in the presence of data in grouped form are discussed. In particular, the analysis focuses on the techniques used in Section 5 to study the SHIW dataset and the Doxa Survey.
This part of the paper is structured as follows. First, the algorithm developed by Shorrocks and Wan is explained in detail. Then, two non-parametric methods are discussed: the Hermite spline interpolation and the bootstrap kernel density estimation. Finally, a non-parametric version of Shorrocks and Wan's algorithm is introduced, developed by adapting the structure of the original method to a non-parametric setting.

3.2.1 Shorrocks and Wan's algorithm

The algorithm of Shorrocks and Wan (2008) consists of two stages. First, a parametric distribution is fitted to the grouped data, and a synthetic sample of microdata is drawn from it. Then, the observations $(x_1, x_2, .., x_n)$ are divided into ω exclusive sets and adjusted with a standardized two-step procedure.

Parametric distributions In the beginning, a distribution is fitted using the available grouped data. In the analysis, the precision of four different distributions has been tested: the log-normal, the Singh-Maddala, the Beta (used to model the Lorenz Curve) and the Generalized Beta of the Second Kind. Ideally, choosing more elaborated density functions could allow to model the distribution of incomes with greater precision. Table 1 summarizes the characteristics of the analysed distributions, that are further treated in the Appendix.

Table 1: Tested density functions

Distribution	Parameters
Log-Normal	μ, σ
Singh-Maddala	c, k, λ
Beta (Lorenz curve)	α, β
Generalized Beta 2	a, b, p, q

The flipside to the increased flexibility of distributions as the Generalized Beta of the Second Kind is the proliferation of parameters to be estimated, that are in this case four. As the analysis of the SHIW database will prove, the maximum likelihood method (Jenkins, 2009) to find those values is likely to be

imprecise in the case of wide grouping patterns (for instance, when the information on the distribution of income is available only in quartiles or quintiles).

The adjustment stage Once obtained a sample of microdata from the aforementioned parametric distributions, the units are divided up into ω groups and adjusted with a two-step procedure.
In the first step, each value x_i, $i = 1, 2, .., n$ is transformed into an "intermediate" value $\hat{x}_i$[3]. The objective of this part of the algorithm is letting the true mean of each interval μ^*_θ lie within the range of sample values of each subgroup. More formally,

$$\min_i \hat{x}_{\theta i} \leq \mu^*_\theta \leq \max_i \hat{x}_{\theta i}, \qquad \vartheta = 1, .., \omega$$

The second step is needed in order to equal the group means of the simulated sample to the true ones, by compressing the gaps between the sample values and the bounds of the group. Keeping the bounds unchanged, the intermediate values $\hat{x}_{\theta i}$ are transformed into the final values $x^*_{\theta i}$[4]. Finally, the algorithm delivers a sample of microdata from a continuous distribution that can be used to estimate inequality and poverty measures.

3.2.2 Hermite spline interpolation

A second interpolation approach, mentioned by Cowell and Mehta (1982), consists in fitting a polynomial spline using ω functions of degree K - one for each interval - to reconstruct the original distribution.

$$f(x) = \sum_{k=0}^{K} \gamma_{\theta k}\ x^k, \qquad x \in [a_\theta, a_{\theta+1})$$

Choosing a high K allows to obtain more precise figures for the inequality indices, but at the same time complicates the function (that can have $\omega[K-1]$ turning points, as underlined by Cowell and Mehta) and its estimation. On the other hand, easier functional forms (such as straight lines, obtained by setting K equal to 1) often cause the density function to assume negative values in some points of its support[5]. Tillé and Langel (2012) use a piecewise quadratic interpolation of the Lorenz curve, while Kakwani (1976) proposes a combination of different techniques: a third degree polynomial function to estimate the Lorenz curve in $\omega - 2$ of the subgroups, and two Pareto curves for the first and the last income classes.

[3]This first transformation is the following:

$$\begin{aligned}
\hat{x_i} &= \mu^*_\theta + \tfrac{\mu^*_{\theta+1}-\mu^*_\theta}{\mu_{\theta+1}-\mu_\theta}\ (x_i - \mu_\theta) && \text{if } x_i \in [\mu_\theta, \mu_{\theta+1}), && \vartheta = 1, ..,\ \omega - 1 \\
\hat{x_i} &= \tfrac{\mu^*_1}{\mu_1}\ x_i && \text{if } x_i < \mu_1 && \text{(first group)} \\
\hat{x_i} &= \tfrac{\mu^*_m}{\mu_m}\ x_i && \text{if } x_i \geq \mu_\theta && \text{(last group)}
\end{aligned}$$

[4]Defining as a_θ the bounds that separate the groups, this second transformation consists of the following passage:

$$\begin{aligned}
x^*_{\theta i} &= a_{\theta+1} - \tfrac{a_{\theta+1}-\mu^*_\theta}{a_{\theta+1}-\hat{\mu}_\theta}\ (a_{\theta+1} - \hat{x}_{\theta i}) && \text{if } \mu^*_\theta > \hat{\mu}_\theta \ \text{ and } \vartheta < \omega \\
x^*_{\theta i} &= a_\theta - \tfrac{\mu^*_\theta - a_\theta}{\hat{\mu}_\theta - a_\theta}\ (\hat{x}_{\theta i} - a_{\theta+1}) && \text{if } \mu^*_\theta < \hat{\mu}_\theta \ \text{ or } \vartheta = \omega
\end{aligned}$$

[5]This would contradict the very definition of "density function".

Following the approach of Rizzi *et al.* (2016), the analyses of Section 5 use a piecewise cubic Hermite interpolation. The algorithm starts from the two vectors defining the ω known points of a Lorenz curve, namely the vector of cumulative proportions of population p and that of cumulative proportions of incomes L. Then, it fits a piecewise spline made by ω polynomials of third degree, one for each class of income for which there is availability of data: in this way, a "smooth" Lorenz curve is generated.

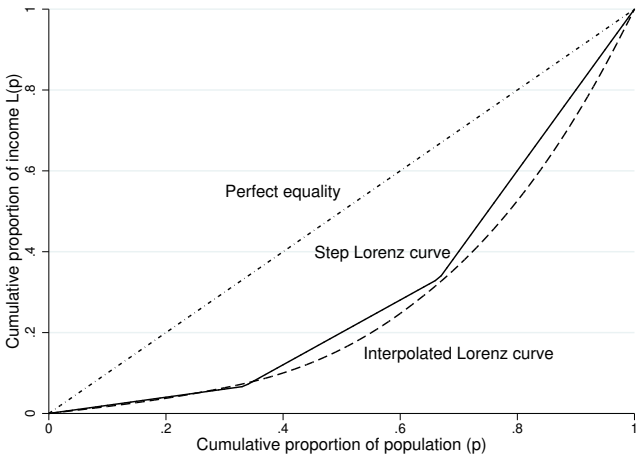


Figure 2: Spline interpolation of a Lorenz curve

The curve interpolated using this technique has a series of desirable properties: first, it intersects all the known points of the Lorenz curve and is a continuous function of class C^1 (namely, its first derivative is continuous). Moreover, as underlined by Cox (2012), this interpolation is shape-preserving: local minima and maxima do not change after the operation, and the same applies to increasing and decreasing sections of the function.
The final step of the procedure is the generation of a sample of data in unit form from the interpolated Lorenz curve.

3.2.3 Bootstrap kernel density estimation

"Naive" and kernel density estimators The kernel method is a widely used non-parametric technique to estimate an unknown density function $f(x)$ starting from a set of observed data. Its functioning can be explained, as in Silverman (1986), starting from the concept of "naive" estimator. From the very definition of density function it follows that:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

For each point x and any interval $(x - h,\ x + h)$, the density function is defined as the proportion of units falling within its boundaries. From this definition, the "naive" density function estimator can be expressed as:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h}\ w\left(\frac{x - X_i}{h}\right)$$

$$\text{with } w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

In this estimator, the function w increases the value of the density in point x if there are units in the interval defined by the parameter h[6].

The kernel estimator is a direct extension of this method that follows the same approach. The only difference is in the choice of the weight function w, that is in this case a symmetric probability function K - such as a normal density - that allows to estimate a "smooth" and continuous probability density function instead of a histogram. The kernel density estimator is therefore:

$$\hat{f}_k(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

The features of the estimated density function depend on two choices: (*1*) the shape of the kernel function K and (*2*) the choice of the bandwidth parameter h. For (*1*), a variety of choices have been developed in the literature: in this paper the choice is the Epanechnikov kernel[7], the optimal one in terms of efficiency (Peracchi, 2001). For what regards (*2*), two different criteria for the choice of h have been tested: Silverman's "rule of thumb" and the more formal method of Sheather and Jones (1991)[8]. The results and the conclusions in terms of inequality and poverty measures differ very little, and the results shown in the following part of the paper are those obtained with the Sheather and Jones bandwidth .

Bootstrapping the kernel density The procedure described so far provides an estimate of an univariate density function, composed by two vectors: a vector $\mathbf{x}$, representing the support of the kernel (the points in which it was evaluated) and a vector $\mathbf{f(x)}$, containing the estimated values of the density in each point of $\mathbf{x}$. However, in order to compute inequality and poverty measures, a sample of observations from the estimated density is required. Since the kernel density is not parametric, this sample cannot be randomly drawn with a standard procedure. This is the motivation for choosing a bootstrap approach.

Given a statistic of interest $\vartheta\mathbf{(x)}$, an estimate thereof can be obtained by drawing with replacement a series of samples from the original vector $\mathbf{x}$, and then by studying the bootstrap distribution of $\vartheta\mathbf{(x)}$, as described by Hansen (2020). In this paper, the statistics included in $\vartheta\mathbf{(x)}$ are the Gini index and the absolute poverty rate. The procedure is the following:

i) The vectors $\mathbf{x}$ and $\mathbf{f(x)}$ are estimated with the kernel method

[6]Silverman (1986, p.12) illustrates the approach by describing the naive estimator as "placing a box of width $2h$ and height $(2nh)^{-1}$ on each observation and then summing to obtain the estimate of the density". In this sense, the "naive" density estimator can be thought of as a histogram with bin size equal to $2h$.

[7]The Epanechnikov (1969) kernel is obtained from a problem of minimization of the mean integrated square error under the assumption that the bandwidth h is chosen optimally. However, as Silverman (1986) underlines, the efficiency loss of using kernels as the Gaussian or the cosine is negligible.

[8]Silverman's "rule of thumb" is based on the standard deviation and on the interquantile range of the available data, while the approach of Sheather and Jones follows a two-step procedure that minimizes the estimated mean integrated squared error.

ii) N samples are drawn from $\mathbf{x}$ with replacement, using $\mathbf{f(x)}$ as a vector of weights[9]

iii) $\vartheta\mathbf{(x)}$ is computed for each of the N samples

Finally, the obtained distribution of $\vartheta\mathbf{(x)}$ is studied to obtain an estimate of the Gini index and of the absolute poverty rate[10].

3.2.4 A non-parametric extension of the algorithm of Shorrocks and Wan

A hybrid approach whose accuracy is tested in this paper is a non-parametric version of the algorithm of Shorrocks and Wan. As aforementioned, this algorithm has proved to be a precise method to estimate inequality measures in the presence of data in grouped form and, no less important, the samples it produces have the desirable property of being equal to the known grouped distribution in the number of units that belong to each of the groups (n_θ) and in the mean income of each class (μ_θ).

However, a parametric assumption on the shape of the income distribution is required in the first stage of the procedure, in order to fit a density function and draw a sample from it. This non-parametric extension of the technique differs from the original in this aspect: first, a sample is generated using non-parametric methods (spline interpolation or kernel density estimation), and then the obtained sample is adjusted with the formulae of the second stage of the standard version of the algorithm, shown in 3.2.1. Finally, this adjusted sample is used to estimate inequality and poverty measures.

4 Methodology and Data

In this section an overview of the characteristics and the sources of the data used in the research is followed by a description of the methodology of analysis.

4.1 Features of the data

The SHIW database The precision of the ungrouping methods described in Section 3 is tested on the samples of Italian income earners contained in the SHIW ("Survey on Households Income and Wealth") database. This data is publicly available in the Bank of Italy website, which has been conducting these enquiries since 1977. The sample was originally composed of 3,000 households, but its size was extended to 8,000 in 1986; in terms of individuals, the most recent samples include information on 12,000-14,000 units for each wave. The data are collected every two years, and include information on income and wealth, as well as some other data on variables such as age, gender or working hours[11].

[9] In the analysis, two choices for the number of repetitions have been tested: 500 and 1000. The estimated values of inequality and poverty measures have been found to be approximatively the same in both cases. This procedure has been carried out using the R package *kernelboot*.

[10] An issue faced using this approach are the negative values that are generated from the kernel distribution. Since in this case the x variable represents incomes, a transformation of the support of the kernel density has been applied in order to obtain non-negative values in the simulated samples, with the following procedure. First, the support of the kernel (the vector $\mathbf{x}$) has been transformed in its natural logarithm. Then, the N samples have been drawn, and the obtained observations have been transformed back in their original scale, so to avoid negative values.

[11] The data on net disposable income has been collected since 1987, and therefore the analysis focuses on the period 1987-2016.

The Doxa Survey The Survey of 1948 is the result of the efforts of the pioneering work of Pierpaolo Luzzatto Fegiz and his collaborators of Doxa Institute. Luzzatto Fegiz, a professor of Statistics at the University of Trieste, founded the Institute in Milan, Italy, right after the end of World War II, in January 1946.

The research was started upon a request that the President of Italy of the time, Enrico De Nicola, had made to the Ministers for the Budget, Finance and Treasury in 1947. The idea was to provide the new-born Republic with a register of the conditions of its inhabitants, on the model of the British "White Book". In fact, reliable statistics on the income of citizens and on the levels of poverty in the country were extremely rare during the era of Mussolini's dictatorship (1922-1943). On the contrary, the regime tended to minimize the magnitude of social issues as poverty and illiteracy, even by physically "hiding" the poor from the view of bourgeois commentators: as an example of such an approach, the act of begging was considered a crime (Vecchi, 2017), punished with the reclusion in charity institutions.

In December 1947, a Decree assigned a contribute of 16 millions of Italian lire (that would correspond today to around 300,000 euros) to the Doxa institute, that started the relevations. In order to obtain representative samples of the population, a two-level clustering procedure was followed: first, the Italian households were divided into 13 groups, according to the region (or more correctly, to the group of regions) in which they lived, extrapolating the information on the population structure from the 1931 and 1936 Censuses, partially updated with more recent data. This choice was due to the absence of more adequate data and, as noted by Brandolini (1999) may have affected the accuracy of the sampling design, since more than ten years had passed, during which Italy had participated to the largest conflict of its history. Then, in each region 8 classes were defined, based on the economic and professional condition of the family. The following step of the procedure was drawing 104 samples, one from each defined class. Overall, the sample is composed of 10,755 households.

The results of the survey were published the following year by Luzzatto Fegiz (1949) and in an article in the *Giornale degli Economisti e Annali di Economia* in July-August, 1950.
The database provides information in grouped form on: (*1*) the absolute frequency distribution (n_θ) of the units of analysis into the 21 subsets and (*2*) the total income (y_θ) corresponding to each of these groups . Moreover, these data are also available for 13 geographical subgroups; although the data for some of the 20 actual Italian regions are blended, this feature can be useful to get some insights on the distribution of income both within and between the different areas of the country.

4.2 Methodology

The methodology of the analysis consists of two steps. The first one is intended to evaluate the reliability of the aforementioned ungrouping techniques, while in the latter the most suitable methods are applied to the grouped Doxa Survey of 1948.

Evaluation of the alternative techniques The observations of a sample of recent Italian microdata on income are sorted in ascending order, and divided into ω_1 groups[12]. For each group, the sample mean μ_θ and the absolute frequency n_θ is computed, so to create a sample in tabular form similar to that of the Doxa Survey. Starting from it, the methods described in 3.2 are applied to reconstruct samples of microdata. The reliability of such techniques is assessed comparing the estimated measures of inequality (Gini index) and poverty (the headcount ratio) to the true ones. Moreover, as in Shorrocks and Wan (2008), the absolute deviations of each reconstructed observation from its true counterpart are computed, in order to identify which are the most problematic quintiles of the distribution to reconstruct.

Ungrouping the Doxa survey database The techniques that are found to be the most trustworthy are applied to ungroup the Doxa Survey database. This source reports information on the income of Italian households, both at a national and at regional level. In order to take individuals, rather than households, as unit of analysis, the data are transformed using some additional information: exact demographic data for 1948 are not available, but the General Census of Population of 1951, carried out by the Italian National Institute of Statistics (ISTAT), is a reasonable source of data for the aims of this study.

The 1951 Census reports the average number of individuals in a family according to the occupation of the breadwinner, while the 1948 Survey provides the distribution of the households' incomes for each professional status of the family's head. Combining these two sources of information it is possible to obtain the average number of family components for each class of income, so to transform the grouped distribution of *households* into a grouped distribution of *individuals*. The last step of this procedure consists of an adjustment to take into account the differences in the demographic structure of the Italian regions[13].
Ungrouping this final distribution, a sample of Italian households in 1948 can be obtained and analysed to compute inequality and poverty measures among individuals.

[12] As in Shorrocks and Wan (2008), the analysis starts from quintiles and deciles. Moreover, it is extended to data in four, fifteen and twenty groups. This last situation is the closest to the Doxa Survey.

[13] According to the Italian Census of 1951, the average number of household components was 3.97. However, the degree of heterogeneity among regions was very high: while the average household in Piedmont was composed by 3.14 people, this figure rose to 4.56 in Umbria and to 4.70 in Veneto.

5 Results and analysis

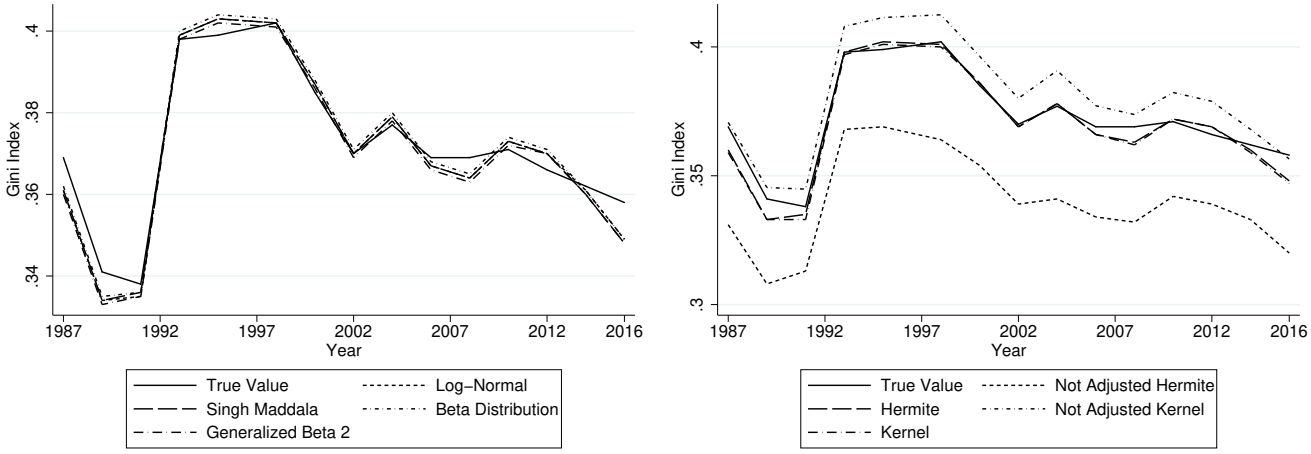
The structure of this section is the following: in 5.1, the samples of the SHIW database are analysed to assess the reliability of the alternative ungrouping techniques, in terms of precision in the estimation of inequality and poverty levels; in 5.2, the data of the Doxa Survey of 1948 is ungrouped and analysed from a historical perspective.

5.1 Analysis of the SHIW Database

The analysis first evaluates the precision of the standard parametric algorithm of Shorrocks and Wan, using as starting distributions a log-normal, a Singh-Maddala, a Beta and a Generalized Beta of the Second Kind. Then, the study moves to two non-parametric methods: the Hermite spline of third degree and the Kernel estimation. Finally, starting from the samples generated using these two latter techniques, there is an assessment of the accuracy of the non-parametric version of the algorithm of Shorrocks and Wan.

5.1.1 Inequality

The plots of Figure 3 show the time series of the Gini index reconstructed using the different techniques described so far, starting from units divided into twenty groups. In this situation, that is the most similar to the case of the Doxa Survey, all the parametric versions of the algorithm of Shorrocks and Wan provide extremely reliable estimates of this inequality measure, with a maximum deviation of half a percentage point. This is indicated by the fact that the solid line, representing the true value of the index, is almost perfectly superposable to all the dashed lines, that denote its reconstructed values.
As far as non-parametric methods are concerned, the not adjusted Hermite tends to underestimate the level of dispersion, while the Kernel overestimates it by around one percentage point. These methods become much more precise with the adjustment step, and in this case their performance is analogous to their parametric counterparts.

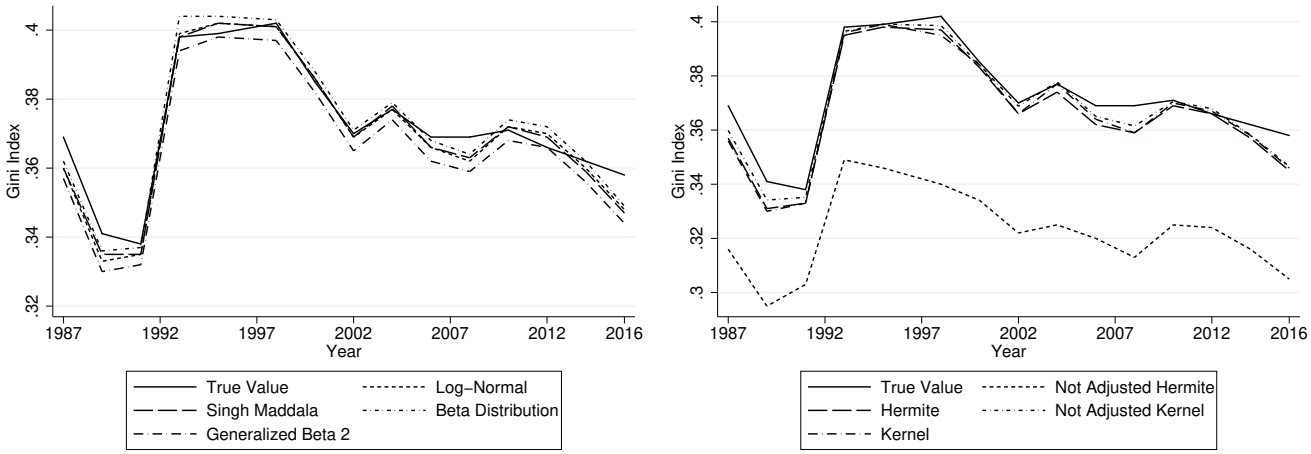


(a) Parametric methods

(b) Non-parametric methods

Figure 3: Gini index computed starting from ventiles

Similarly to the previous one, Figure 4 shows the Gini index computed starting from grouped data and compared to its true value: in this case, the index is reconstructed from deciles. In this case, the dashed lines are slightly more distant from each other. The most precise parametric functions are the Singh-Maddala and the Beta, but the advantage of using them in lieu of other distributions is very little, since the maximum deviation is in the order of one percentage point.
For what regards non-parametric methods, their precision is very similar to the parametric ones, except for the not adjusted Hermite, that significantly underestimates the value of the index.

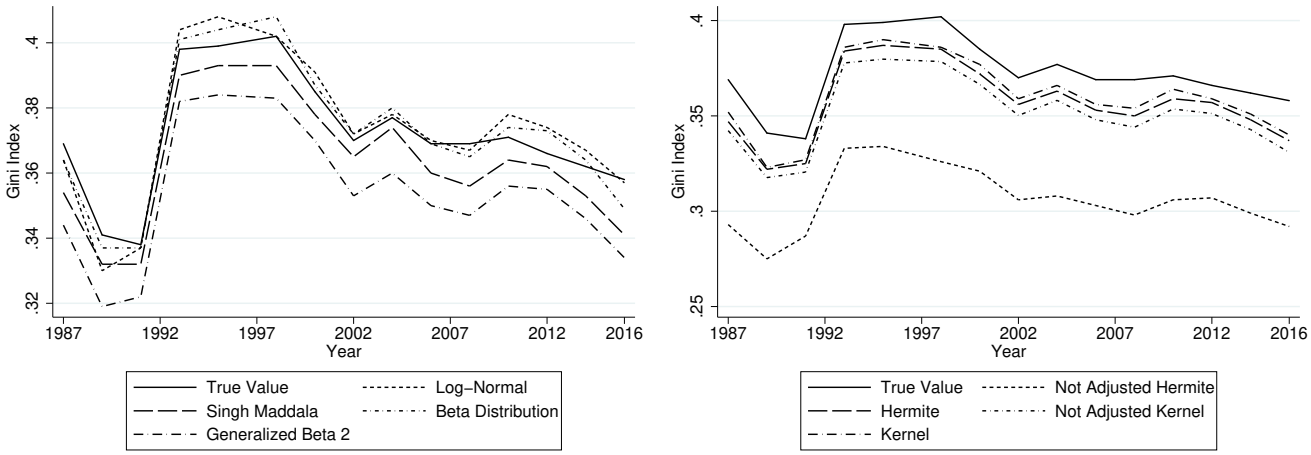


(a) Parametric methods

(b) Non-parametric methods

Figure 4: Gini index computed starting from deciles

Figure 5 shows the precision of the ungrouping techniques in estimating the Gini index starting from quintiles of data. As expected, the values obtained using alternative techniques differ more than in the previous cases. The Beta and the log-normal distributions prove to be very precise also in this case, while the most complex function, the Generalized Beta of the Second Kind, is the least accurate.
As far as non-parametric methods are concerned, while the not adjusted Hermite interpolation is very imprecise the other techniques are quite reliable, even if they tend to underestimate the levels of dispersion. Their average deviation, that ranges between one and two percentage points, is comparable to the one of the Singh-Maddala.

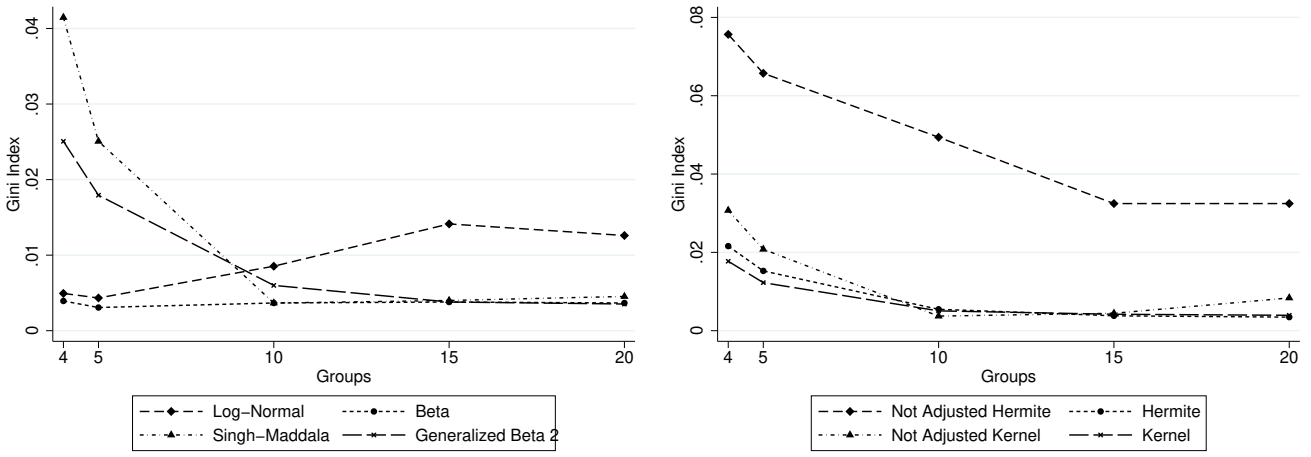


(a) Parametric methods

(b) Non-parametric methods

Figure 5: Gini index computed starting from quintiles

Figure 6 plots the mean absolute deviations of the estimated Gini Index from its true value against the number of groups from which the distribution is reconstructed. This latter, shown along the x-axis, has been reproduced starting from 4, 5, 10, 15 and 20 classes. Being the accuracy of the different techniques a function of the grouping pattern of the dataset, it is not possible to state that a technique is superior to another in every situation. On the contrary, their efficacy has to evaluated case-by-case, by looking at the number of classes and considering the objective of the analysis (in this case, reconstructing the value of an index of dispersion). As expected, the distributions with a larger number of parameters to estimate (the Singh-Maddala and the Generalized Beta of the Second Kind) become much more precise as the number of classes increases (and so does the available amount of information). On the other hand, the two simpler distributions are preferable when data are available in quartiles or quintiles, but comparably less precise when there are fifteen or twenty classes of income (this is particularly true for the log-normal distribution).



(a) Parametric methods

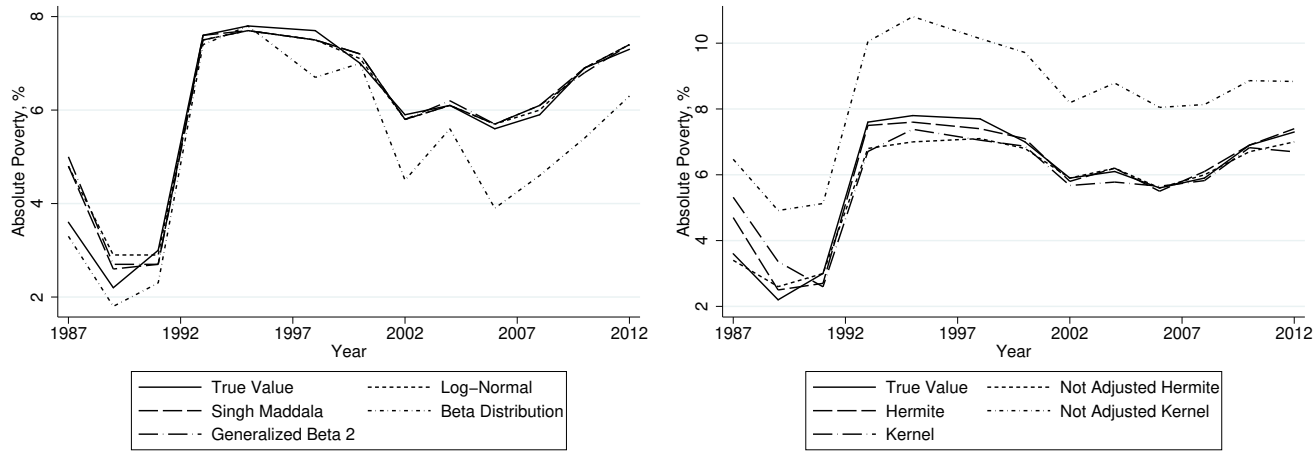
(b) Non-parametric methods

Figure 6: Absolute deviations, Gini Index

For what regards non-parametric methods, the not adjusted Hermite is surely the worst technique in all the analysed circumstances. The adjustment step improves the precision of the estimates, whose accuracy is comparable to that of the best parametric techniques in the case of ten classes or more. Conversely, they are not very reliable in case of data in quartiles or quintiles.

5.1.2 Poverty

Figure 7 reports the absolute poverty rates estimating starting from data in twenty classes, compared with the true time series of the variable[14].
Parametric methods are very precise for different choices of the starting density function: the only exception is represented by the Beta distribution, that tends to underestimate the proportion of poor in the population. With regard to non-parametric methods, the graph shows that the not adjusted kernel is certainly the worst technique. The reliability of the others is very similar to that of the most efficient parametric techniques. In particular, the adjusted Hermite approximation is the most precise among the evaluated non-parametric techniques.



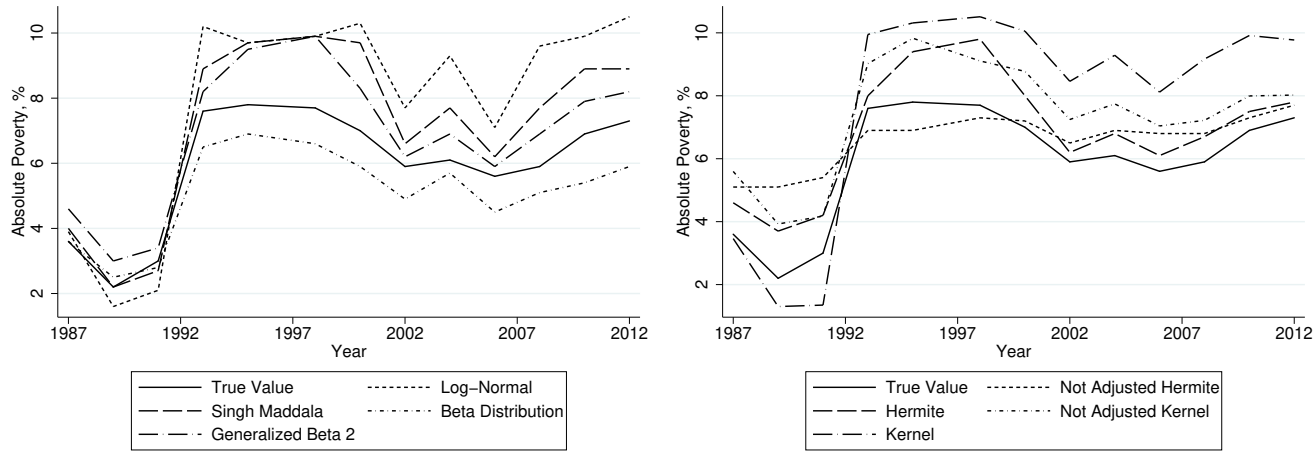
(a) Parametric methods

(b) Non-parametric methods

Figure 7: Poverty rates computed starting from ventiles

As shown by Figure 8, in the presence of data available in deciles, the estimation of poverty rates is more complicated. The log-normal distribution is the least precise, but also the Singh-Maddala and the Generalized Beta 2 tend to overestimate the number of poor in the population, with errors that are in the order of one or two percentage points. On the contrary, the Beta distribution underestimates the proportion of poor, and its average bias is the smallest under these circumstances, but remains significant. Similarly, non-parametric methods are not reliable in the estimation of the headcount ratio in the presence of ten classes of income: all the analysed methods overestimates this metrics. However, the right panel of Figure 8 clearly displays that the Hermite interpolation is preferable to the Kernel method.

[14]The poverty thresholds used in this section are those estimated by Vecchi (2017).

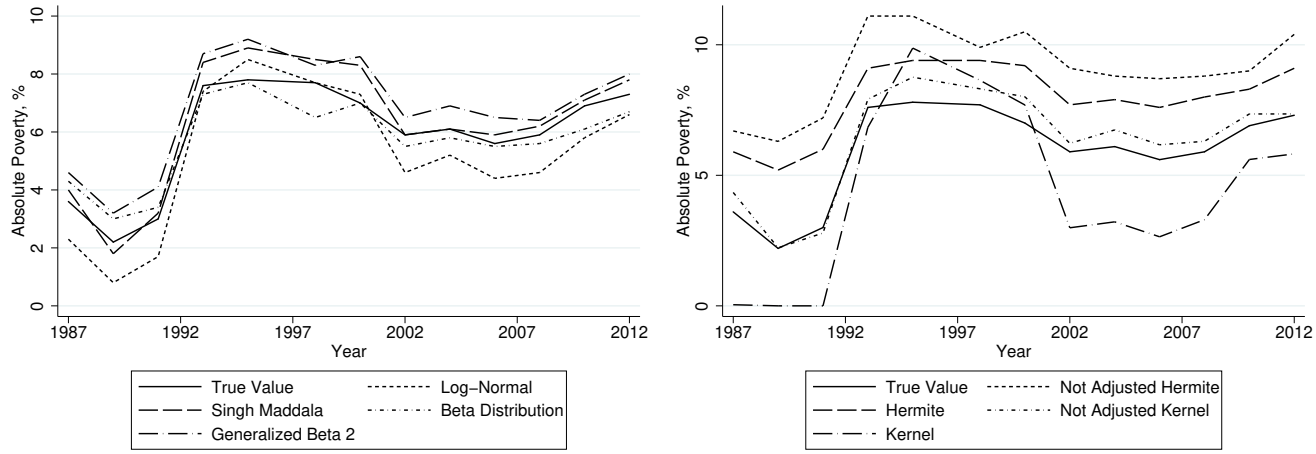


(a) Parametric methods

(b) Non-parametric methods

Figure 8: Poverty rates computed starting from deciles

Figure 9 reports the estimated absolute poverty indices computed from data in quintiles.
Many of the considerations made for data grouped in ten classes remain valid also in this case. Surely, obtaining precise estimates of the headcount ratio is difficult in the presence of such a limited information: parametric methods are not precise, and the most accurate choice for the starting density function is the Beta distribution.
Non-parametric methods are even less reliable: the estimates can either be significantly too high or too low, and the adjustment step of the algorithm of Shorrocks and Wan does not improve their precision at all.



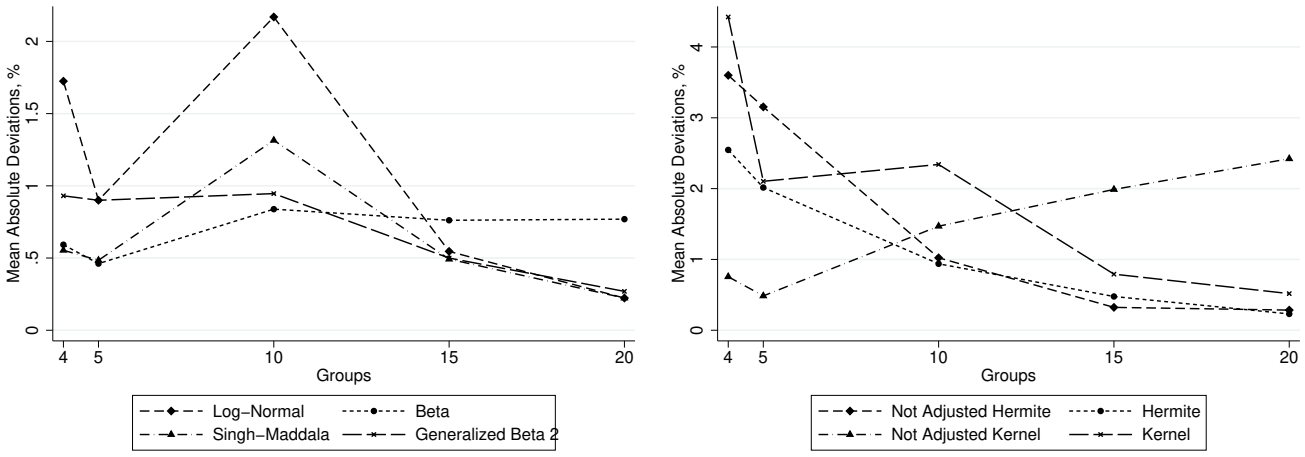
(a) Parametric methods

(b) Non-parametric methods

Figure 9: Poverty rates computed starting from quintiles

Similarly to Figure 6, Figure 10 shows how the average absolute deviation of the estimated headcount ratio from its true value varies depending on the number of classes from which it is computed. As for the levels of inequality, the analyses has been carried out for 4, 5, 10, 15 and 20 groups.

When the information on the distribution is conspicuous (fifteen or twenty groups), parametric methods are quite precise: in particular, the average deviation of the headcount ratio is less then half a percentage point for the log-normal, the Singh-Maddala and the Generalized Beta of the Second Kind. The Beta distribution provides instead the best estimates when the units are divided into four or five groups: in this case, the other distributions are extremely unreliable. As shown by the right panel of Figure 10, non-parametric methods produce reasonably precise estimates only when then groups are fifteen or twenty: in this case, the Hermite interpolation is superior to the kernel. Instead, when the groups are less than ten, mean deviations are very high, and the adjustment step does not reduce them.



(a) Parametric methods

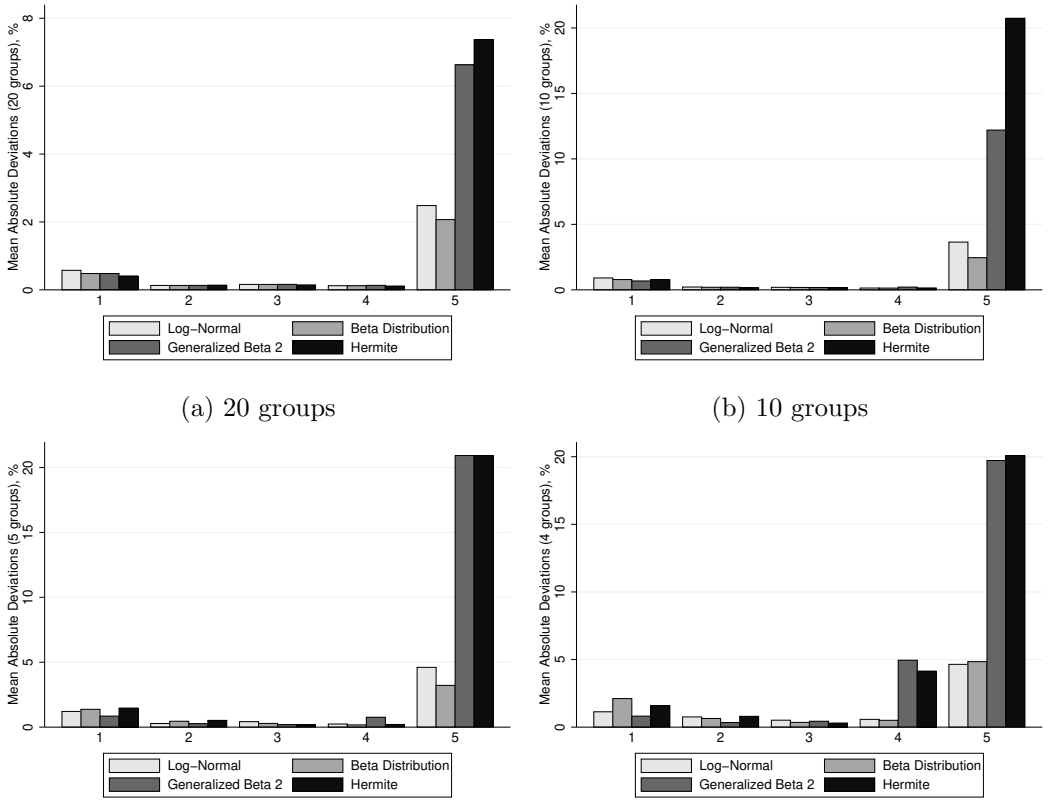
(b) Non-parametric methods

Figure 10: Percent deviations, poverty rates

5.1.3 Summary of the results

Figure 11 displays the average deviations (expressed in percentages) of the simulated single values from their true counterparts, for different grouping patterns and alternative methods.
Once reconstructed a sample, the units that compose it have been sorted in ascending order, and compared with the original observations they should resemble. Then, the obtained deviations have been summarized by taking the means of each quintile of income, so to understand which portion of the distribution is more difficult to reconstruct.

For all the employed methods and for every grouping pattern, the quintile that causes most of the issues is the last one. The right tail of the distribution generally contains a group of outliers (the super-rich) whose earnings are difficult to predict with any of the analysed methods, and the deviations of the samples generated using an adjusted Hermite interpolation are the most significant. On the other hand, the central quintiles are those that are reconstructed with the highest degree of accuracy in every circumstance.
As expected, the degree of imprecision increases as the number of classes shrinks: in particular, the Generalized Beta 2 distribution becomes much less precise than the others. This finding can be explained by the higher number (four) of parameters that its estimation requires.



(a) 20 groups

(b) 10 groups

(c) 5 groups

(d) 4 groups

Figure 11: Percent absolute deviations, by quintiles

It is worth noticing that, clearly, the precision of an ungrouping technique can be judged in this way, but for the aims of this study the most relevant measure of its accuracy is the capability to consistently estimate inequality and poverty measures. As the following Tables expose, a technique can generate samples that are imprecise in the last quintile without jeopardizing the accuracy of the estimates of the Gini index and of the headcount ratio.

Table 2 summarizes the mean absolute deviations of the Gini index for all the tested methods and for every grouping pattern. If the data are in quartiles or quintiles, the most reliable choice is the Beta distribution, and the log-normal is very precise as well. On the other hand, if the groups are ten or more, there is a quite large possibility of choice in terms of method, both parametric and non-parametric: apart from the not adjusted Hermite, all the other methods are very precise.

Table 2: Mean absolute deviations, Gini Index

Groups	4	5	10	15	20
Log-Normal	0.005	0.004	0.009	0.014	0.013
Singh-Maddala	0.041	0.025	**0.004**	**0.004**	0.005
Beta	**0.004**	**0.003**	**0.004**	**0.004**	0.004
Gen. Beta 2	0.025	0.018	0.006	**0.004**	0.004
Not Adj. Hermite	0.076	0.066	0.049	0.032	0.032
Hermite	0.022	0.015	0.005	**0.004**	**0.003**
Not Adj. Kernel	0.031	0.021	**0.004**	**0.004**	0.008
Kernel	0.018	0.012	0.005	**0.004**	0.004

Similarly, Table 3 deals with poverty rates. In this case, the most accurate techniques are the parametric ones: we see that also here the Beta distribution is often the best choice in case of a small number of classes. As for the Gini index, when the groups are many (fifteen or twenty), almost all the techniques provide reliable estimates. Interestingly, the adjustment does not increase the precision of the non-parametric methods when the groups are less than ten.

Table 3: Mean absolute deviations, Poverty rates (%)

Groups	4	5	10	15	20
Log-Normal	1.72	0.90	2.17	0.55	**0.22**
Singh-Maddala	**0.55**	0.48	1.32	0.49	**0.22**
Beta	0.59	**0.46**	**0.84**	0.76	0.77
Gen. Beta 2	0.93	0.90	0.95	0.50	0.27
Not Adj. Hermite	3.60	3.15	1.02	**0.32**	0.28
Hermite	2.55	2.02	0.94	0.48	0.23
Not Adj. Kernel	0.76	0.48	1.47	1.99	2.42
Kernel	4.42	2.10	2.34	0.79	0.52

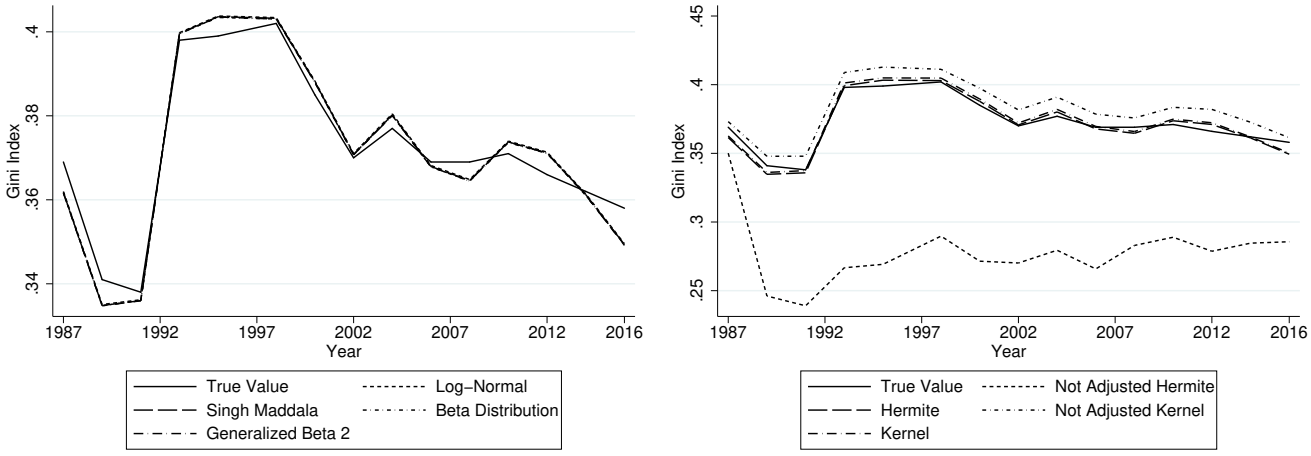
5.1.4 A further investigation

All the analyses carried out up to this point have evaluated the precision of the chosen ungrouping techniques starting from ω homogeneous classes, each one with the same relative frequency of units. However, the Doxa Survey of 1948 presents the data on households' income in 21 considerably heterogeneous groups: the first ones are very wide, being the 80% of the units in the first six classes, while the last ones are much smaller, so that the information is concentrated on the right tail of the distribution of income. Table 4 reports the relative frequency and the cumulative distribution function for each class of the Doxa Survey.

Table 4: Grouping Pattern, Doxa 1948

Class	$f(x_i)_{\%}$	$F(x_i)_{\%}$	Class	$f(x_i)_{\%}$	$F(x_i)_{\%}$
1	2.8	2.8	12	1.5	97.6
2	15.9	18.7	13	0.6	98.2
3	23.1	41.8	14	0.5	98.7
4	17.8	59.6	15	0.2	98.9
5	13.4	73.0	16	0.2	99.1
6	7.9	80.9	17	0.2	99.3
7	5.3	86.2	18	0.2	99.5
8	3.3	89.5	19	0.2	99.7
9	2.0	91.5	20	0.1	99.8
10	2.5	94.0	21	0.2	100
11	2.1	96.1			

In the presence of such an unbalanced grouping structure, the conclusions drawn in the previous sections on the precision of the different ungrouping techniques may end up being invalid. Therefore, the Gini index and the absolute poverty rates for the SHIW data have been reconstructed also starting from 21 groups "à la Doxa", specifically created to resemble the situation that will be faced in Section 5.2[15].



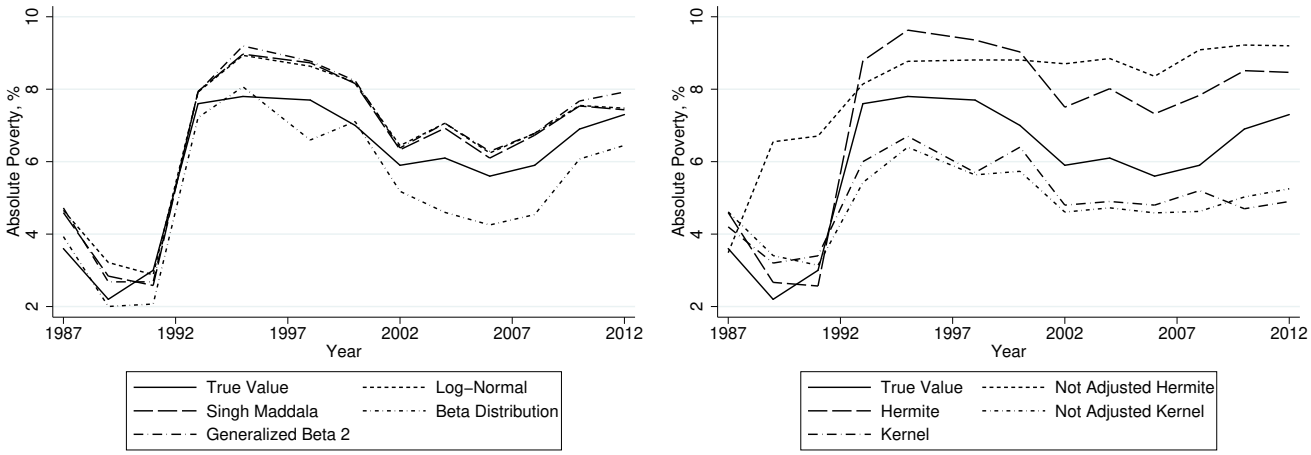
(a) Parametric methods

(b) Non-parametric methods

Figure 12: Gini index computed starting from 21 groups "à la Doxa"

Figure 12 reports the Gini index computed from 21 groups "à la Doxa". As in the ungrouping from ventiles shown in 5.1.1, the parametric version of the algorithm of Shorrocks and Wan provides very reliable estimates, with negligible differences among the different starting distributions. As far as non-parametric methods are concerned, the conclusions are similar to those of the study on ventiles: while the adjusted Hermite and Kernel are as precise as their parametric counterparts, the same is not true for the not adjusted versions of these techniques. In particular, the bootstrap kernel tends to overestimate the index by one point, while the Hermite spline underestimates it significantly.

[15]This procedure of ungrouping has been carried out also starting from 18 groups, as in the regional data of the Doxa Survey, finding equivalent results.



(a) Parametric methods

(b) Non-parametric methods

Figure 13: Poverty rates computed starting from 21 groups "à la Doxa"

Similarly, Figure 13 reports the reconstructed time series for absolute poverty rates. As the both panels show, this measure is much less precisely estimated starting from groups "à la Doxa" than it was from ventiles. In particular, as displayed by the left panel, some parametric methods allow to limit the bias (in particular the Singh-Maddala and the Log-Normal distributions); on the contrary, non-parametric methods are completely unreliable.

Table 5: Mean absolute deviations, Gini and HCR
(starting from groups "à la Doxa")

Method	Gini	HCR
Log-Normal	0.004	0.73
Singh-Maddala	0.004	**0.70**
Beta	**0.003**	0.76
Gen. Beta 2	0.004	0.82
Not Adj. Hermite	0.095	2.20
Hermite	**0.003**	1.43
Not Adj. Kernel	0.010	1.41
Kernel	0.004	1.38

Table 5 reports the mean absolute deviations of the Gini index and of the headcount ratio from their true values, when the SHIW data are ungrouped from 21 groups "à la Doxa". The conclusions for the upcoming analysis of the Doxa Survey are the following:

1. Inequality measures can be estimated with a very satisfying degree of precision both with the tested parametric techniques - all of them - and with the non-parametric versions of the algorithm of Shorrocks and Wan. The not adjusted Kernel slightly overestimates the measure, while the not adjusted Hermite is unreliable.

2. Poverty measures are more precisely estimated with parametric techniques: the Singh-Maddala distribution minimizes the bias from the original time series, while all the tested non-parametric techniques are not preferable. However, this metrics shows more significant deviations than the Gini index.

This last result - the reduced precision of poverty estimates in the presence of groups "à la Doxa" - is not surprising: the grouping pattern of the Survey provides only limited information on the left tail of the distribution, the part that is relevant for the estimation of poverty rates. For this reason, the results on poverty of the following section have to be treated with particular caution.

5.2 Analysis of the Doxa Survey of 1948

The analysis moves now to ungroup the Doxa Survey of 1948. Table 6 reports the information made available by Luzzatto Fegiz (1950) at national level.

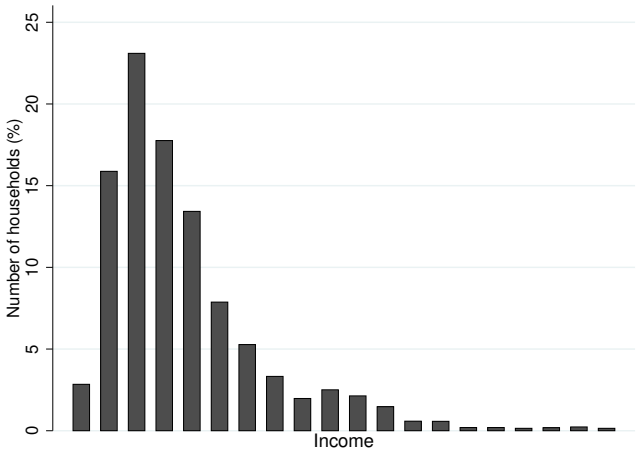
Table 6: The Italian Doxa Survey of 1948

Income Class (Thousands of Lire)	Households	Income (Millions of Lire)	Average income (Thousands of Lire)
0-130	305	30,460	99.87
130-260	1704	340,760	199.98
260-390	2479	818,268	330.08
390-520	1906	876,668	459.96
520-650	1441	850,308	590.08
650-780	845	608,688	720.34
780-910	566	481,185	850.15
910-1040	357	349,958	980.27
1040-1170	212	234,876	1107.91
1170-1300	269	334,056	1241.84
1300-1625	229	336,364	1468.84
1625-1950	158	283,178	1792.27
1950-2275	63	133,668	2121.71
2275-2600	62	150,548	2428.19
2600-2925	21	56,582	2694.38
2925-3250	21	65,508	3119.43
3250-3575	16	55,664	3479.00
3575-3900	20	73,678	3683.90
3900-5200	25	111,930	4477.20
5200-6500	16	92,430	5776.86
6500-	17	415,000	24411.77
Total	10,732	6,699,877	624.29

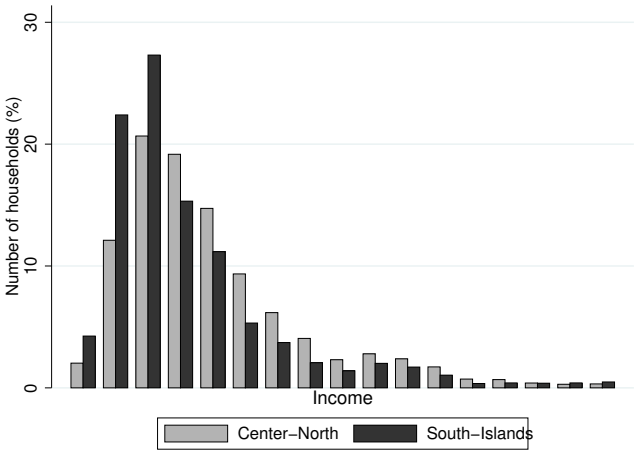
This is the starting point of the procedure: the 10,732 households (sampled with an approximate proportion of 1 to 1000) are divided into twenty-one groups in ascending order of income. For each of these classes, the table reports the absolute frequency of households (second column) and the total income earned by each of these groups of families (third column). The fourth column, representing the distribution of the households' average income, is obtained simply as the ratio of the third to the second column.

Figure 14a reports the relative frequencies of households for each of the income classes of the Survey.

As usual when dealing with income, the distribution is positively skewed: the large majority of families are in the first five groups, and the median income is lower than the mean.



(a) Italian households, 1948



(b) Italian households by region, 1948

The Survey provides information also at sub-regional level (for 13 regions or groups of regions), thus allowing for an analysis of the differences between the different parts of the country. Before moving to the main object of interest of the research - inequality and poverty levels - it is worth observing the distribution of households in the two macroregions in which Italy is usually divided: the Center-North and the South-Islands. Figure 14b is analogous to the previous one, but it plots separately the distribution of income in these two areas, showing important differences between them: southern households tend to concentrate more in the first three classes (the poorest), while the number of families who belong to the middle class is much more significant in the North. The percentage of rich families is approximately the same all over the country, thus suggesting the existence of an *élite* in every region.

5.2.1 Inequality

Moving to the treatment of measures of inequality, Table 7 reports the Gini Index computed on the samples obtained by ungrouping the Doxa Survey, both at national and regional level[16][17].
The overall level (0.404) conceals relevant differences among macro-regions: the South is found to be much more unequal than the rest of the country, being its Gini index seven points higher than that of the rest of the country[18].

[16]The values for the Gini Index, as well as the poverty rates presented in 5.2.2, are computed taking individuals as units of analysis.

[17]The reported values are those obtained using the Shorrocks and Wan's algorithm modelling the Lorenz curve with a Beta distribution, but the results obtained with the other parametric and non-parametric version of the algorithm differ from these in a very negligible way. The complete results are available in the Appendix.

[18]Here and in the remaining part of the paper, the 13 subregions of the Doxa Survey are organized as follows: the "Center-North" includes Piedmont and Liguria, Lombardy, Veneto and Venezia Tridentina, Emilia, Toscana, Marche and Umbria, and Lazio; the "South-Islands" is composed by Abruzzi and Molise, Campania, Puglia, Lucania and Calabria, Sicily and Sardinia. The survey does not provide information on Valle d'Aosta and the province of Trieste.

Table 7: Inequality, Italy 1948

Region	Gini Index
ITALY	0.404
Center-North	0.383
South-Islands	0.453
Piedmont and Liguria	0.328
Lombardy	0.370
Veneto and Venezia Tridentina	0.429
Emilia	0.458
Toscana	0.355
Marche and Umbria	0.326
Lazio	0.376
Abruzzi and Molise	0.360
Campania	0.445
Puglia	0.394
Lucania and Calabria	0.517
Sicily	0.498
Sardinia	0.448

In the most egalitarian regions (Piedmont/Liguria, that is also one of the richest, and Marche/Umbria), the Gini index is around 0.32-0.33, while in the most unequal ones (Lucania/Calabria and Sicily) it arrives to 0.50-0.51. Another indication of the gap between the different areas of the country is given by the proportion of income dispersion that is determined by differences *between* regions: if the average income were the same in every region, the Gini index would decrease by more than 5 points.

5.2.2 Poverty

Table 8 reports the absolute poverty rates computed from the data presented in the Doxa Survey[19][20]. As for inequality levels, limiting the analysis to the synthetic figure related to the national poverty rate (24.9%) would neglect the significant differences between the different areas of the country, that in this case are even more impressive.

For what regards poverty rates, the gap between the two macroregions is extremely wide: the poverty rate in the South is almost 15 percent points higher than that in the Center-North, and there are also significant differences within these groups of regions. While North-Western regions (Piedmont/Liguria and Lombardy) present relatively "low" levels of poverty (16-18%), the figures are much higher in the North-East and in the Center. The proportion of poor people gradually increases while moving south, up to the extremes of Lucania/Calabria, where this rate is the highest in the country, and of the islands: according to the Survey, in these regions more than 37% of the population lived in absolute poverty in 1948.

[19]The poverty threshold used is estimated by Vecchi (2017), according to whom a person living in Italy in 1948 is considered poor if its income is lower than 69,470 lire per year, corresponding to 1,338 euros of 2020 (for an extremely low monthly amount, 110 euros).

[20]The reported values are obtained using the Shorrocks and Wan's algorithm starting from a Singh-Maddala distribution, the one that proved to be the most precise in Section 5.1.4, but the results obtained with the other parametric techniques differ from these in a very negligible way. The complete results are available in the Appendix.

Table 8: Poverty, Italy 1948

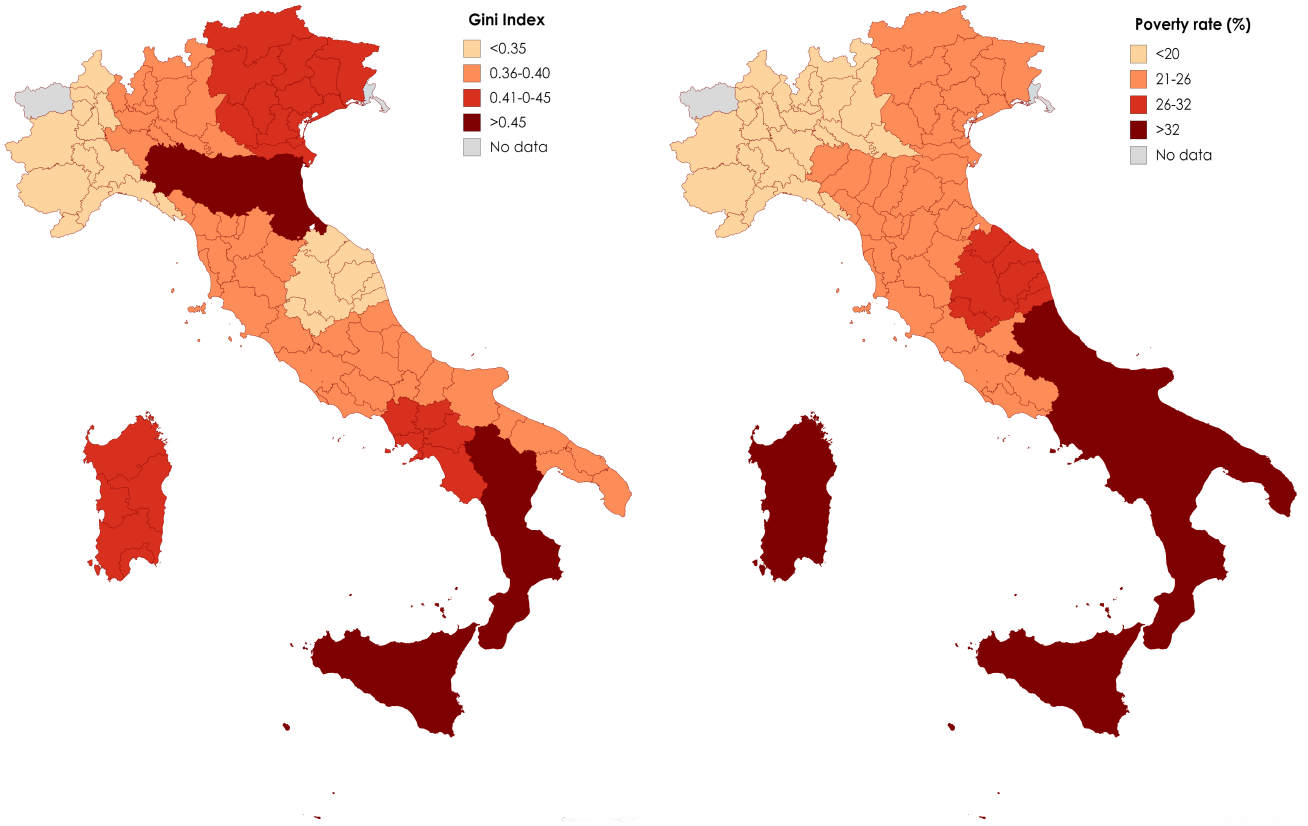
Region	Poverty rate (%)
ITALY	24.9
Center-North	20.8
South-Islands	35.3
Piedmont and Liguria	17.9
Lombardy	16.3
Veneto and Venezia Tridentina	22.7
Emilia	22.5
Toscana	22.6
Marche and Umbria	29.2
Lazio	23.9
Abruzzi and Molise	34.4
Campania	32.0
Puglia	34.5
Lucania and Calabria	38.8
Sicily	37.3
Sardinia	39.0

5.2.3 Historical reflections

The findings of the previous sections deliver an image of a remarkably heterogeneous country, as in the levels of inequality as, more relevantly, in the poverty rates. The maps of Figure 15 allow to see how large these differences were, and how in most cases the most unequal regions were the poorest, while the richest were usually those with a lower degree of income dispersion in the population. This division is roughly parallel to the North-South one, especially for what regards poverty.

To provide an idea of the difference between the different areas of the country, it can be useful to compare the results obtained at a regional level for 1948 with the national ones in other moments of the Italian economic history. According to Vecchi (2017), the Gini Index in Italy immediately after the full independence of the country in 1871[21] was 0.45, while the absolute poverty rate was around 39%. These conditions are very similar to those observed by looking at the data of the Doxa Survey that refer to southern regions as Lucania and Calabria, or to the islands of Sardinia and Sicily. On the contrary, in order to find national poverty rates comparable to those of the richest regions in 1948 (namely Piedmont/Liguria and Lombardy), we have to look at the figures of the late sixties of the twentieth century. For instance, the Italian poverty rate was 18% in 1967 and 15.8% in 1969; in 1948, this rate in Piedmont/Liguria and Lombardy was, respectively, 17.9% and 16.3%. Therefore it is possible to say that in a way, if the analysis focuses on inequality and poverty levels, in the post-war Italy some southern regions resembled the country in the late nineteenth century, while the levels of development of the North-West would have been reached by the country on average only after the *economic miracle*, at the end of the 1960s. Moreover, the poverty levels of the richest region in 1948 (Lombardy) were to be reached by the South only after 1975.

[21] Proclaimed its sovereignty in 1861 after two wars of Independence, Italy fought a third one against the Austro-Hungarian Empire in 1866 to gain control of Veneto, but the country was completely unified only in 1870; one year later, Rome became the capital of the new-born State.



(a) Gini Index

(b) Poverty Rates

Figure 15: Inequality and poverty in Italy, 1948

Back to the 1948 Survey, it is worth noticing that for what regards poverty rates the country can be divided into three groups of regions: a relatively developed North-West, a Center/North-East in the midst and a poorer South/Islands. As far as inequality levels are concerned, however, this classification is not so valid: richer regions are in general more egalitarian, but there are some relevant exceptions and differences. For instance, the levels of poverty in the North-East are very similar to those in the central regions (Toscana and Lazio), but these latter show much lower levels of inequality. This difference is particularly striking if we compare two neighboring regions as Toscana and Emilia: given the same poverty rate (22.5%), the Gini index of Emilia is ten points higher, and the highest of all regions but the extreme southern ones (0.458). This fact can be surprising, especially knowing that Emilia became in the post-war period one of the richest parts of the country, and the most egalitarian Italian region. The other significant exception from this association between high poverty and inequality levels is represented by Marche and Umbria. While the poverty rate of these central regions is much higher than that of (geographically and, to a certain extent, historically) close regions as Emilia, Toscana and Lazio, the inequality levels are the lowest in the whole country. This fact may be due to the small size of the upper-class in these rural regions, but caution is needed in this judgement: the quality of the sampling procedure of the Doxa Survey has been criticized by Brandolini (1999), and even the author of the research admitted some imprecisions due to the lack of recent Census information and to the attitudes of the interviewers (Luzzatto Fegiz, 1950). Moreover, it would have been interesting to produce distinct figures for the regions of Marche and

Umbria, but unfortunately at the time of the data collection they were treated as one sub-region, so that all the available information is merged together.

In a historical perspective, a Gini Index of 0.404 for 1948 is an argument for the thesis that the Italian *economic miracle* of the fifties and sixties did not have significant implications on the distribution of income in the country, at least at national level. According to Vecchi (2017) and Cannari and D'Alessio (2018), the levels of inequality in the late 1960s - when the Bank of Italy started producing grouped data on Italian incomes - were approximately the same as at the beginning at this period of massive economic growth (between 0.39 and 0.41). It can be therefore concluded that in the first twenty years of the Republic (1948-1968), the rapid development of the Italian economy was not accompanied by a reduction in inequality levels. These latter shrank significantly in the two subsequent decades (1968-1992), before starting to increase after the currency crisis of 1992. As far as poverty levels are concerned, they surely contracted in the first twenty years, reaching a level of 19.4% in 1971 (Vecchi, 2017); however, the most important improvements arrived only in the seventies, a period characterized by lower growth rates and extremely high inflation, but during which poverty rates sharply declined, reaching a value of only 3.8% in 1982. Also in this case, poverty started increasing once more from the nineties. Another rather sad conclusion is that the North-South gap did not become smaller in the last seventy years, both in inequality and poverty levels. It seems that the advent of democracy did not reduce the extent of this long-standing issue, that has existed since the unification of the country in 1861: the *questione meridionale* (i.e. "southern question") remains unsolved.

6 Conclusions

This paper has investigated a series of methods to ungroup data in tabular form, first by presenting them from a theoretical perspective and then by testing their precision in estimating inequality and poverty values. Latterly, the exposition moved to a historical analysis of the Doxa Survey: therefore, there are concluding remarks for both of these lines of the research.

For what regards the solution to the ungrouping problem, a conclusion of this research is that most of the tested methods are equally valid in the presence of a large number of groups (fifteen or more). When the available information is so abundant, inequality and poverty levels can be precisely computed using the standard Shorrocks and Wan's procedure with any parametric distribution. Non-parametric techniques as the bootstrap kernel density estimation are a valid choice for the estimation of the Gini index as well, but have proved to be less reliable for what regards poverty levels. The newly introduced non-parametric version of the algorithm of Shorrocks and Wan is as precise as the standard one for the estimation of the Gini index, but not for poverty rates, especially when there are no more than five groups. In this last case of "wide" grouping patterns, it is convenient to keep the procedure simple: the best choice seems to be the standard parametric version of the algorithm, starting from density functions whose estimation is not too demanding in terms of number of parameters, such as the log-normal, the Singh-Maddala or the Beta distributions. Finally, the analysis has proved the importance of testing the effectiveness of the techniques on structures of grouped data that differ not only in the number of classes,

but also in the characteristics of these latter: the further investigation made by ungrouping the SHIW database from heterogeneous groups has allowed to choose the most precise methods for the analysis of the grouped Doxa Survey.

Concerning the historical analysis, the most striking finding of the research is the extreme polarization of post-war Italy in terms of poverty levels. Some parts of the country showed poverty rates that were similar to the national ones in the second half of the nineteenth century, while others were way more developed: these disparities have not ceased to exist in the following seventy years. Inequality levels differed significantly among the different regions, with a generally more unequal south and a more egalitarian north (with the exceptions of Emilia and Veneto), and the national value of the Gini index (0.404) for 1948 suggests that the following twenty years of massive growth did not change significantly the distribution of income at national level. A possible line of future research would start from this result to investigate why this index did not change during the years of the *economic miracle*. Perhaps, the slow growth of incomes (Toniolo, 2013) due to the abundant availability of labour has been a specific determinant of the Italian development: this would have surely impeded the process of reduction of inequality and poverty. Clearly, the choices of economic policy had a role in the abatement of inequality and poverty in the 1970s: it would be of extreme interest to understand which were the new directions of economic policy, lacking during the years of the "miracle", that allowed to obtain these results. However, caution is needed in these judgements. The quality of the data of the Survey is questionable from different points of view, such as the accuracy of the sampling procedure, the choice made by pollsters on the people to interview, the reliability of the answers and the lack of data on non-monetary incomes. The Survey of 1948 was the first attempt to investigate with modern methods the living conditions of Italian households and, as noted by Brandolini (1999), it was considered by its very author primarily a useful experiment for future work. Finally, the processing of the data operated in the paper to consider individuals as unit of analysis is based on the Census data of 1951, that may not correspond exactly to the demographic structure of the country three years before.
Therefore, the inequality and poverty levels presented in this paper are more to be interpreted as an instrument for relative comparisons between regions rather than as perfect estimates in absolute terms. Unfortunately, the accuracy of the employed ungrouping techniques does not address the issues regarding the availability of information and its quality, but is nevertheless a fundamental requirement for the analysis of historical data, that are very commonly presented in tabular form.

Appendix

Distributions for Shorrocks and Wan's algorithm

The precision of the "parametric" version of the algorithm of Shorrocks and Wan has been tested using four different density functions as starting point: the log-normal[22], the Singh-Maddala, the Beta (for the Lorenz Curve) and the Generalized Beta of the Second Kind[23].

$$f(x;\mu,\sigma) = \frac{1}{x}\ \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left(-\frac{(lnx-\mu)^2}{2\sigma^2}\right) \qquad \text{(Log-Normal)}$$

$$f(x;c,k,\lambda) = \frac{ck}{\lambda}\left(\frac{x}{\lambda}\right)^{c-1}\left(1+\left(\frac{x}{\lambda}\right)^{c}\right)^{-k-1} \qquad \text{(Singh-Maddala)}$$

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\ x^{\alpha-1}\ (1-x)^{\beta-1} \qquad \text{(Beta distribution)}$$

$$\text{with } B(\alpha,\beta) = \int_0^1 t^{\alpha-1}\ (1-t)^{\beta-1}dt$$

$$f(x;a,b,p,q) = \frac{ax^{p-1}}{b^{ap}B(p,q)\left[1+\left(\frac{x}{b}\right)^{a}\right]^{p+q}} \qquad \text{(Generalized Beta 2)}$$

Inequality and poverty levels according to the different techniques

Table A1 and A2 report the Gini indices and the absolute poverty rates computed from the Doxa database using different alternative techniques. The methods are denoted as: LN (Shorrocks and Wan's algorithm from a log-normal distribution), SM (Shorrocks and Wan, Singh-Maddala), Beta (Shorrocks and Wan, modelling the Lorenz Curve with a Beta distribution), GB2 (Shorrocks and Wan, Generalized Beta of the Second Kind), NAH (Not adjusted Hermite interpolation), Her. (Hermite interpolation with Shorrocks and Wan's adjustment), NAK (Bootstrap kernel density estimation), Ker. (Kernel density estimation with Shorrocks and Wan's adjustment)[24].

[22]For the log-normal, Beta and Singh-Maddala distributions, the density fit has been carried out using the command *ungroup* of the Stata package *DASP*.

[23]The estimation of the parameters of the GB2 function has exploited the Stata package *gbgfit*, developed by Austin Nichols. Once obtained those parameters, the final synthetic sample has been obtained by drawing samples from two χ^2 distributions and with a series of transformations on the F distribution obtained by taking their ratio (Xi'an, 2015).

[24]The estimation of the parameters of the Hermite spline function has been carried out using the Stata package *pchipolate*, developed by Nicholas Cox, while the estimation of the kernel density starting from the dataset in tabular form has exploited the *density* function, available in the R package *stats*.

Table A1: Inequality, Italy 1948

Region	LN	SM	Beta	GB2	NAH	Her.	NAK	Ker.
ITALY	0.405	0.404	0.404	0.404	0.445	0.405	0.412	0.404
Center-North	0.383	0.383	0.383	0.383	0.442	0.384	0.392	0.383
South-Islands	0.455	0.454	0.453	0.454	0.447	0.455	0.459	0.454
Piedmont and Liguria	0.328	0.328	0.328	0.328	0.436	0.329	0.341	0.406
Lombardy	0.371	0.370	0.370	0.370	0.439	0.371	0.380	0.372
Veneto and Venezia Tridentina	0.430	0.429	0.429	0.429	0.445	0.430	0.436	0.439
Emilia	0.459	0.458	0.458	0.458	0.448	0.459	0.460	0.459
Toscana	0.355	0.355	0.355	0.355	0.446	0.356	0.365	0.355
Marche and Umbria	0.325	0.325	0.326	0.325	0.435	0.326	0.336	0.325
Lazio	0.374	0.374	0.376	0.375	0.447	0.375	0.384	0.375
Abruzzi and Molise	0.353	0.352	0.360	0.352	0.439	0.355	0.361	0.357
Campania	0.446	0.445	0.445	0.445	0.449	0.446	0.444	0.447
Puglia	0.394	0.393	0.394	0.428	0.445	0.395	0.400	0.398
Lucania and Calabria	0.518	0.517	0.517	0.516	0.443	0.518	0.513	0.517
Sicily	0.499	0.498	0.498	0.498	0.451	0.500	0.503	0.498
Sardinia	0.447	0.446	0.448	0.445	0.448	0.448	0.453	0.461

Table A2: Poverty, Italy 1948

Region	LN	SM	Beta	GB2	NAH	Her.	NAK	Ker.
ITALY	25.9	24.9	25.0	25.1	35.1	26.5	27.2	26.8
Center-North	21.4	20.8	20.8	21.9	33.9	22.2	22.3	21.4
South-Islands	36.3	35.3	35.3	34.9	37.1	36.3	36.6	34.6
Piedmont and Liguria	18.2	17.9	18.7	17.6	34.5	19.3	19.2	29.0
Lombardy	16.9	16.3	16.1	15.9	32.0	17.8	17.7	17.7
Veneto and Venezia Tridentina	23.4	22.7	23.0	23.0	32.8	24.0	24.1	23.5
Emilia	23.5	22.5	22.5	22.6	31.6	24.0	24.0	22.5
Toscana	22.9	22.6	23.3	23.4	36.1	23.9	24.2	22.9
Marche and Umbria	29.6	29.2	30.1	29.7	40.6	30.6	30.8	29.4
Lazio	24.4	23.9	22.6	24.0	35.7	25.4	25.5	24.6
Abruzzi and Molise	34.6	34.4	30.4	33.8	41.0	35.6	35.8	35.4
Campania	32.9	32.0	31.9	32.4	36.8	33.1	32.7	31.4
Puglia	35.0	34.5	35.2	35.5	39.5	35.5	35.6	34.9
Lucania and Calabria	40.2	38.8	39.1	38.5	34.7	39.4	38.3	39.0
Sicily	38.3	37.3	37.1	35.6	36.0	38.0	37.8	36.5
Sardinia	39.3	39.0	37.1	38.2	38.0	39.3	39.5	40.3

Family components by income class and region

Tables A3 and A4 report, respectively, the average number of family components, according to the income class of the household and to its region. These tables have been obtained with a processing of Doxa (1948) and National Census data (1951).

Table A3: Components, by income

Income Class (Thousands of Lire)	Components
0-130	3.69
130-260	4.04
260-390	4.13
390-520	4.12
520-650	4.15
650-780	4.05
780-910	4.18
910-1040	4.23
1040-1170	4.18
1170-1300	4.08
1300-1625	4.22
1625-1950	4.26
1950-2275	4.34
2275-2600	3.98
2600-3250	4.06
3250-3900	4.40
3900-6500	4.35
6500-	3.62

Table A4: Components, by region

Regions	Components
ITALY	3.97
Center-North	3.83
South-Islands	4.22
Piedmont and Liguria	3.16
Lombardy	3.64
Veneto and Venezia Tridentina	4.47
Emilia	4.01
Toscana	3.93
Marche and Umbria	4.55
Lazio	3.96
Abruzzi and Molise	4.32
Campania	4.43
Puglia	4.27
Lucania and Calabria	4.24
Sicily	3.91
Sardinia	4.39

Bibliography

Aitchison, J. and Brown, J.A., 1957. "The lognormal distribution with special reference to its uses in economics." Cambridge University Press.

Anderson, L. and Fricker Jr, R.D., 2015. "Raking: An important and often overlooked survey analysis tool." *Phalanx*, 48(*3*), pp.36-42.

Abdelkrim, A. and Duclos J.Y., 2007. "DASP: Distributive Analysis Stata Package". PEP, World Bank, UNDP and Université Laval.

Baldini, M. and Toso, S., 2009. "Diseguaglianza, povertà e politiche pubbliche" (pp. 1-262). Il Mulino.

Bandourian, R., McDonald, J. and Turley, R.S., 2002. "A comparison of parametric models of income distribution across countries and over time."

Bartels, C.P. and Van Metelen, H., 1975. "Alternative probability density functions of income: A comparison of the lognormal-, Gamma-and Weibull-distribution with Dutch data." Vrije Universiteit, Economische Faculteit.

Boukaka, S.A., Mancini, G. and Vecchi, G., 2018. "Poverty and Inequality in Francophone Africa, 1960s-2010s." (No. 16) *The Historical Household Budgets Project.*

Brandolini, A., 1999. "The distribution of personal income in post-war Italy: source description, data quality, and the time pattern of income inequality." *Giornale degli economisti e Annali di economia*, pp.183-239.

Brick, J.M., Montaquila, J. and Roth, S., 2003. "Identifying problems with raking estimators." In *Annual meeting of the American Statistical Association*, San Francisco, CA.

Cannari, L. and D'Alessio, G., 2018. "Wealth inequality in Italy: reconstruction of 1968-75 data and comparison with recent estimates." *Bank of Italy Occasional Paper*, (428).

Censimento generale della popolazione: 4 novembre 1951. (General Census of 1951). Istat. `https://ebiblio.istat.it/SebinaOpac/resource/9-censimento-generale-della-popolazione-4-novembre-1951/IST0069130?tabDoc=tabcontiene`

Cox, N.J., 2012. "PCHIPOLATE: Stata module for piecewise cubic Hermite interpolation," Statistical Software Components S457561, Boston College Department of Economics.

Cowell, F.A. and Mehta, F., 1982. "The estimation and interpolation of inequality measures." *The Review of Economic Studies*, 49(*2*), pp.273-290.

Dagum, C., 1977. "New model of personal income-distribution-specification and estimation." *Economie appliquée*, 30(*3*), pp.413-437.

Datt, G., 1998. "Computational tools for poverty measurement and analysis" (No. 583-2016-39573, pp. 1-29).

Deville, J.C. and Särndal, C.E., 1992 "Calibration estimators in survey sampling." *Journal of the Amer-*

ican statistical Association, 87(*418*), pp.376-382.

Epanechnikov, V.A., 1969. "Non-parametric estimation of a multivariate probability density." *Theory of Probability & Its Applications,* 14(*1*), pp.153-158.

Gastwirth, J.L., 1972. "The estimation of the Lorenz curve and Gini index." *The review of economics and statistics*, pp.306-316.

Gastwirth, J.L. and Glauberman, M., 1976. "The interpolation of the Lorenz curve and Gini index from grouped data." *Econometrica: Journal of the Econometric Society,* pp.479-483.

Gini, C., 1912. "Variabilità e mutabilità." Reprinted in *Memorie di metodologica statistica* (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.

Gordy, M.B., 1998. "A generalization of generalized beta distributions." Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board.

Hansen, B., 2020. "Econometrics". Publicly available at `https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf`

Hudzik, H. and Maligranda, L., 1994. "Some remarks on s-convex functions." *Aequationes Mathematicae*, 48(*1*), pp.100-111.

Hyman, J.M., 1983. "Accurate monotonicity preserving cubic interpolation." *SIAM Journal on Scientific and Statistical Computing,* 4(*4*), pp.645-654.

Jenkins, S.P., 2009. "Distributionally-sensitive inequality indices and the GB2 income distribution." *Review of Income and Wealth,* 55(*2*), pp.392-398.

Luzzatto-Fegiz, P., 1949. "I redditi delle famiglie italiane nel 1948: indagine rappresentativa sulla distribuzione del reddito nazionale e sulle entrade e spese delle famiglie italiane." Doxa.

Luzzatto-Fegiz, P., 1950. "La Distribuzione Del Reddito Nazionale." *Giornale Degli Economisti E Annali Di Economia, Nuova Serie*, 9, no. 7/8 (1950), pp. 341-54.

Kakwani, N., 1976. "On the estimation of income inequality measures from grouped observations." *The Review of Economic Studies*, 43(*3*), pp.483-492.

McDonald, J.B., 2008. "Some generalized functions for the size distribution of income". In *Modeling Income Distributions and Lorenz Curves* (pp. 37-55). Springer, New York, NY.

Microdata for the "Survey on Households Income and Wealth", Bank of Italy: `https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html`.

Minoiu, C. and Reddy, S., 2008. "Kernel density estimation based on grouped data: The case of poverty assessment" (No. 8-183). International Monetary Fund.

Minoiu, C. and Reddy, S.G., 2009. "Estimating poverty and inequality from grouped data: How well do parametric methods perform?" *Journal of Income Distribution*, 18(*2*).

Morgan, J., 1962. "The anatomy of income distribution." *The review of economics and statistics*, pp.270-283.

Nichols, A., 2010. "GBGFIT: Stata module to fit a Generalized Beta (Type 2) distribution to grouped data via ML." Statistical Software Components S457132, Boston College Department of Economics.

Peracchi, F., 2001. "Econometrics", John Wiley & Sons Ltd. Chichester, West Sussex.

Piketty, T, 2014. "Capital in the 21st Century."

PovCalNet Software, World Bank: `http://iresearch.worldbank.org/PovcalNet/home.aspx`

Rizzi, S., Gampe, J. and Eilers, P.H., 2015. "Efficient estimation of smooth distributions from coarsely grouped data." *American Journal of Epidemiology,* 182(*2*), pp.138-147.

Rizzi, S., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T.B., Vaupel, J.W. and Lindahl-Jacobsen, R., 2016. "Comparison of non-parametric methods for ungrouping coarsely aggregated data." *BMC medical research methodology,* 16(*1*), p.59.

Sala-i-Martin, X., 2006. "The world distribution of income: falling poverty and... convergence, period." *The Quarterly Journal of Economics,* 121(*2*), pp.351-397.

Shorrocks, A. and Wan, G., 2008. "Ungrouping income distributions: Synthesising samples for inequality and poverty analysis." (No. 2008/16) Research Paper, UNU-WIDER, United Nations University (UNU).

Signorell, A. *et mult.al.*, 2020. "DescTools: Tools for Descriptive Statistics." R package version 0.99.34, `https://cran.r-project.org/package=DescTools`.

Silverman, B.W., 1986. "Density estimation for statistics and data analysis" (Vol. 26). CRC press.

Singh, S.K. and Maddala, G.S., 1976. "A Function for the Size Distribution of Incomes." *Econometrica*, 44, pp.9632–970.

Thurow, L.C., 1970. "Analyzing the American income distribution." *The American Economic Review*, 60(*2*), pp.261-269.

Tillé, Y. and Langel, M., 2012. "Histogram-based interpolation of the Lorenz curve and Gini index for grouped data." *The American Statistician,* 66(*4*), pp.225-231.

Toniolo, G., 2013. "La crescita economica italiana 1861-2011". *L'Italia e l'economia mondiale. Dall'Unità a oggi*, pp.5-51.

Xi'an (`https://stats.stackexchange.com/users/7224/xian`), "How to draw a random sample from a Generalized Beta distribution of the second kind", URL (version: 2015-10-15): `https://stats.stackexchange.com/q/140539`

Vecchi, G., 2017. "Measuring wellbeing: a history of Italian living standards." Oxford University Press.

Villaseñor, J. and Arnold, B.C., 1989. "Elliptical Lorenz curves." *Journal of econometrics*, 40(*2*), pp.327-

338.

Wang, B. and Wertelecki, W., 2013. "Density estimation for data with rounding errors." *Computational Statistics & Data Analysis,* 65, pp.4-12.

Wolodzko, T., 2020. "Kernelboot: Smoothed Bootstrap and Random Generation from Kernel Densities." R package version 0.1.7, `https://cran.r-project.org/web/packages/kernelboot/index.html`