Received: 3 April 2022 | Revised: 7 March 2023 | Accepted: 18 April 2023

DOI: 10.1002/sam.11623

WILEY

RESEARCH ARTICLE

A new parametric approach to gender gap with application to EUSILC data in Poland and Italy

Francesca Greselin[1] | **Alina Jędrzejczak**[2] | **Kamila Trzcińska**[2]

[1]Department of Statistics and Quantitative Methods, University of Milano Bicocca, Milano, Italy

[2]Department of Statistical Methods, University of Łódź, Łódź, Poland

Correspondence
Francesca Greselin, Department of Statistics and Quantitative Methods, University of Milano Bicocca, Milano, Italy.
Email: francesca.greselin@unimib.it

Abstract

Real income distribution comparisons are of interest to policy makers across European countries. Nowadays, a crucial component of income inequality remains the discrepancy between men and women, often called the gender gap. Since the gender gap is related to the whole distribution of incomes in a population, popular single metrics are not adequate, and previous studies applied the relative distribution method, a non-parametric approach to the comparison of distributions. Here, we propose a parametric approach for estimating the relative distribution. Then we extend it to assess the impact of selected covariates—related to the personal characteristics of the samples—on the existing gender gap in both countries. In more detail, models for income were fitted to empirical data from Poland and Italy, from the European Survey of Income and Living Conditions (wave 2018). Afterwards, their parameters were employed to obtain the estimates of relative distribution characteristics. The methods applied in the study turned out to be relevant to describe the gender gap over the entire income range. Finally, the results of the empirical analysis are discussed to reveal similarities and substantial differences between the countries.

KEYWORDS

Dagum, gender gap, income inequality, Italy, parametric inference, Poland, relative distribution method

JEL CLASSIFICATION

C1; C46; D63

1 | INTRODUCTION

Income disparity is still growing in OECD countries, and has reached its highest level in the past half-century. Several studies have been conducted on this issue; among them, it is worth recalling *the Divided We Stand. Why Inequality Keeps Rising* [1] and *In It Together: Why Less Inequality Benefits All* [2]. The trend of rising inequality has become a priority for policymakers, and calls for the analysis of various aspects of income inequality, including its measurement and decomposition by regional areas, by income sources and—recently—also across genders.

Gender equality, often understood only in terms of income, should be viewed as multidimensional. Gender equality also means equal economic independence for women and men; it refers to equality in decision-making and, in the broader setting, it requires equal dignity, integrity, and the ending of gender-based violence. Gender equality is one of the fundamental values of the European Union (EU). The European Commission's work on gender equality policy is based on the "Strategic Engagement for gender equality 2016–2019", which focuses on five priority areas, including increasing female labour-market participation, reducing the gender pay, earnings and pension gaps, combating gender-based violence, improving gender balance in decision-making and promoting gender equality within the Member States and across the world. Although, in EU countries, it is generally illegal for employers to pay different amounts of men and women to do the same job, there are many other reasons why, on average, substantial income differences between men and women can be observed. The observed differences capture differences along many possible dimensions, including education level, working hours, experience, occupation, and many others. Therefore, the gender gap is based only on income discrepancies, and does not account for underlying differences in the above-mentioned covariates. It measures gender inequality but not necessarily discrimination. The adjustment of the gender gap for the covariates can be helpful in recognizing the reasons for existing inequalities; however, further in-depth and country-specific analysis is often necessary to detect all cases and causes of possible discrimination. Many countries still have ineffective equal-pay legislation that regulates women's overall paid working hours. The distribution of the workforce across working hour bands is generally more even for women than for men, due largely to the higher incidence of part-time work among women and, thus, a greater proportion of women working fewer hours. Blau and Kahn developed in [3] an interesting in-depth analysis, showing that differences in pay are caused by many concurring factors. On one hand, occupational segregation is perhaps the main reason: men are prevalent in higher-paid industries, while women are mostly in lower-paid industries. Differences in remuneration across industry sectors all influence the gender pay gap [3]. Finally, some barriers to entry into the labour market are related to the education level and single parenting rate [4]. There is vertical segregation, too: few women work in senior, and hence better-paying positions.

For the sake of social and economic policies, it seems interesting to compare income inequality across EU countries. In this paper, the focus will be on income distributions across Poland and Italy. Their different economic backgrounds offer interesting perspectives. Poland is still suffering the effects of the transition from a centrally-planned to a market-based economy, while Italy is a former well-established market economy. Moreover, according to the Tárki European Social Report [5], there is a lower level of acceptance of inequality in the post-socialist bloc than in other European countries. Results of EU-SILC show that the popular Gini inequality indexes for net household incomes in the two countries were rather similar and equalled 0.34 and 0.35, respectively. Nevertheless, the comparative studies conducted by Jędrzejczak [6] and Zenga and Jędrzejczak [7] revealed substantial differences in "inequality patterns" much higher for the Italian macro-regions, as compared with the Polish ones. In particular, a relatively strong negative correlation has been observed between GDP per capita and income inequality measured by the Gini index in the Italian regions, while in Poland this correlation turned out to be slightly positive. As a result, in Italy, the highest inequality levels occur in the poorest regions, while in Poland the opposite situation has been observed.

A debated research issue regards the methodology of measuring the gender gap. In the Eurostat database, one can find an indicator called "unadjusted gender pay gap", defined as a relative difference between average gross hourly earnings, coming from the four-yearly Structure of Earnings Survey.[1] The gender pay gap in the EU in 2019 was 14.1% and had only changed minimally over the last decade—it means that women earn 14.1% less per hour than men on average. Another summary measure used by Eurostat, called "the gender overall earnings gap", stood at 36.7%. It measures the combined impact of the average hourly earnings, the monthly average of the number of hours paid (before any adjustment for part-time work) and the employment rate. Similar indicators can easily be obtained based on SILC (Survey of Income and Living Conditions) data, by comparing mean or median incomes, available for gender groups. Such an approach seems unsatisfactory, as the phenomenon of the gender gap is related to the entire distribution of incomes in a population, so it is difficult to capture the full range of experiences by means of the aforementioned single metrics. To uncover the factors contributing to the gender discrepancy, one should adopt a variety of tools, consider concomitant variables and move beyond the typical focus on average or median earnings differences, towards a full comparison of the entire distribution of women's earnings relative to men's.

An analytic study of the gender gap for Poland and Italy has been developed in [8], by comparing data provided by Eurostat for Poland and Italy in 2015 through the

[1] https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey

relative distribution approach introduced by Handcock and Morris [9]. Such method is a non-parametric complete summary of the information required for scale-invariant comparisons between distributions. The main objective of the present paper is to take a step further in the analysis, starting from the measurement of the differences observed in income distributions for men and women in both countries with parametric income distribution models, and extending the method to assess the effects of some important covariates. First, the parameters of the selected theoretical distributions were estimated from the data and the best-fit model was selected. Afterwards, empirical and theoretical distributions were compared through a relative approach. The next step of the analysis was the search for socio-economic factors which could explain the observed differences, by means of a newly introduced parametric decomposition for covariate adjustment.

The rest of the paper is organized as follows. Section 1 introduces the most used models for income data, while Section 2 presents the measures we adopted to evaluate the quality of our first inferential results. After recalling briefly the notion of relative distribution in Section 3, we introduce a parametric version of the approach in Section 4. In addition, the decomposition of covariates is summarized in Section 5, and its parametric version is presented in Section 6. A brief description of the EU-SILC data opens Section 7, that is then devoted to compare and discuss all the obtained results from the gender gap analysis, in the non-parametric and parametric approaches. Conclusions and final remarks end the paper in Section 8.

2 | MODELS FOR INCOME DATA

We recall here three economic size distributions widely employed in the literature for fitting income data, namely the Dagum, the Singh-Maddala and the Lognormal model. We provide their definitions, and basic information for making inference on survey data. For the interested reader, we suggest the books of Kleiber and Kotz [10] and Arnold [11] as invaluable resources on income models and various Pareto-type distributions, including statistical inference procedures.

2.1 | Dagum distribution

This model takes its name from Camilo Dagum, who introduced it in the 1970s, when working on a quest for a statistical distribution closely fitting empirical income and wealth distributions.

To mimic characteristic properties observed in such datasets, Dagum searched for a model permitting an interior mode (as the Lognormal) and able to handle heavy tails (like the Pareto), at the same time. Furthermore, he moved from characteristic properties of empirical income and wealth distributions: he stated a generating system where the income elasticity $\eta(F, y)$ of the cumulative distribution function (cdf) of income y is a decreasing and bounded function of F, and therefore of y. Let us recall here, briefly, that *elasticity* in economics is defined as the ratio between the percentage changes of two variables. It is a measure of the sensitivity of the first variable to changes in the second one. After decades of applications to real data (see Reference [10] and references wherein), we can deem the Dagum model as an excellent candidate for our purpose.

We say that F belongs to the Dagum family if its probability density function (pdf) is given by

$$f_D(y; a, b, p) = \frac{a p y^{ap-1}}{b^{ap}\left[1 + \left(\frac{y}{b}\right)^a\right]^{p+1}}, \qquad y > 0$$

for some $a, b, p > 0$, where a and p are shape parameters, while b is a scale parameter. The shape parameters are related to inequality, Lorenz and first stochastic dominance. For example, let $F_1 = f_D(a_1, b_1, p_1)$ and $F_2 = f_D(a_2, b_2, p_2)$ be two Dagum distributions, then the necessary and sufficient conditions for Lorenz dominance (i.e., non intersecting Lorenz curves) is $a_1 p_1 \leq a_2 p_2$ and $a_1 \leq a_2$.

This model allows for various degrees of positive skewness and leptokurtosis; moreover, it owns a built-in flexibility to be unimodal, to approximate income distributions; or zeromodal, to describe wealth distributions. For more details on this distribution, in the framework of economic size distributions, see Kleiber and Kotz (Reference [10], chap. 6.3) and references wherein.

The cumulative distribution function (cdf) for the Dagum is given by

$$F_D(y; a, b, p) = \left[1 + \left(\frac{y}{b}\right)^{-a}\right]^{-p}, \qquad y > 0. \tag{1}$$

We can invert the cdf F to obtain the quantile function, yielding

$$F_D^{-1}(u; a, b, p) = b\left[u^{-1/p} - 1\right]^{-1/a}, \qquad u \in (0, 1). \tag{2}$$

The Dagum model can be seen as a special case of the generalized beta distribution of the second kind (GB2; it is also a member of the the Burr family being equivalent the Burr type III distribution [12]).

Given an i.i.d. sample $\{y_1, y_2, \ldots, y_n\}$ drawn from the parent distribution Y, the likelihood function takes the form

$$L_D(a,b,p|y) = \left(\frac{ap}{b}\right)^n \prod_{i=1}^{n}\left(\frac{y_i}{b}\right)^{ap-1}\left(1+\left(\frac{y_i}{b}\right)^a\right)^{-p-1}.$$

Finally, the solution of the following system of equations provides the ML estimation

$$\begin{cases} \frac{n}{a} + p\sum_{i=1}^{n}\ln\left(\frac{y_i}{b}\right) - (p+1)\sum_{i=1}^{n}\frac{\ln\left(\frac{y_i}{b}\right)}{1+\left(\frac{b}{y_i}\right)^a} = 0 \\ np - (p+1)\sum_{i=1}^{n}\frac{1}{1+\left(\frac{b}{y_i}\right)^a} = 0 \\ \frac{n}{p} + a\sum_{i=1}^{n}\ln\left(\frac{y_i}{b}\right) - \sum_{i=1}^{n}\ln\left[1+\left(\frac{y_i}{b}\right)^a\right] = 0. \end{cases}$$

Unfortunately, no explicit solution of this system is known (see, among others, Reference [13]). This issue perhaps explains its relative unpopularity, despite the solid rationale on which the Dagum model is based. We developed our own code in Mathematica, to numerically solve the ML optimization.

2.2 | The Singh-Maddala distribution

Another probability distribution frequently applied to modeling income and wage distributions was proposed by Singh and Maddala [14]. The distribution has been derived based on the concept of hazard rate or failure rate, widely used for deriving probability distributions in reliability theory or for the analysis of lifetime distributions. Like the Dagum model, it is a special case of the four-parameter GB2 model, introduced by McDonald [15], and a member of the Burr family (Burr type XII model), both are therefore a special case of the Feller-Pareto family. Its pdf is given by

$$f_{SM}(y) = a\,b^{-a}\,q\,y^{a-1}\left[1+(y/b)^a\right]^{-q-1} \quad y>0,$$

where $a,b,q>0$. The cumulative distribution function takes the form

$$F_{SM}(y) = 1-\left[1+(y/b)^a\right]^{-q} \quad y>0.$$

The likelihood function for the Singh-Maddala distribution reads as follows:

$$L_{SM}(a,b,p|y) = \left(a\,b^{-a}\,q\right)^n \prod_{i=1}^{n} y_i^{a-1}\left[1+(y_i/b)^a\right]^{-q-1}.$$

To obtain the normal equations for the unknown parameters, we take partial derivatives of (1.2) with respect to a, b and q and equate them to zero:

$$\begin{cases} \frac{n}{a} + \sum_{i=1}^{n}\ln(y_i/b) - (q+1)\sum_{i=1}^{n}\ln(y_i/b)\left[1+(b/y_i)^a\right]^{-1} = 0 \\ n - (q+1)\sum_{i=1}^{n}\left[1+(b/y_i)^a\right]^{-1} = 0 \\ \frac{n}{q} - \sum_{i=1}^{n}\ln\left[1+(y_i/b)^a\right] = 0. \end{cases}$$

The solutions of the above equations are the maximum likelihood estimators of the Singh-Maddala model parameters a, b and q.

Very recently, Dutang et al. [16] provided functions in R for obtaining the MLE estimators for the whole family of Feller-Pareto distributions. Their asymptotic properties have been derived in Reference [17].

2.3 | The Lognormal distribution

A two-parameter model that has been frequently applied for fitting income distributions in many countries, mainly owing to its simplicity and the straightforward interpretation of its parameters, is the lognormal distribution. The Lognormal fits better than the Pareto distribution the lower income levels, but its fit towards the upper tail is far from satisfactory. Nevertheless, the Lognormal distribution can be applied to approximate selected empirical income distribution, especially in post-socialist countries [18].

A Lognormal random variable Y has the following density function

$$f_L(y) = \frac{1}{y\ \sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y-\mu)^2}{2\sigma^2}\right)$$

where $\mu\in\mathbb{R}$ and $\sigma>0$ are respectively the expected value and the standard deviation of $log(Y)$.

It is worth noting that also the Lognormal model can be obtained from the GB2 model as a limiting case, assuming special parameter values [19]. The maximum likelihood estimators of the Lognormal distribution parameters μ and σ, based on a random sample $\{y_1, y_2, \ldots, y_n\}$, are given by the following explicit formulas:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln y_i$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\ln y_i - \widehat{\mu}\right)^2$$

The estimators are most efficient and unbiased (in the case of $\widehat{\sigma^2}$ the latter property is true only asymptotically) and their respective large sample variances are: $V^2(\widehat{\mu}) = \sigma^2/n$ and $V^2\left(\widehat{\sigma^2}\right) = 2\sigma^4/(n-1) \sim 2\sigma^4/n$ (see Reference [20]).

3 | GOODNESS OF FIT MEASURES

Let us employ the own words of George Box [21] for the *incipit* of this section: "All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true. All models are wrong, but some models are useful. So the question you need to ask is not: Is the model true? (it never is), but: Is the model good enough for this particular application?"

So, the primary question is the *goodness of fit*. Goodness-of-fit measures are widely employed to assess how well the estimated models fit a set of observations, to provide an initial—partial—answer to George Box's question. In the following, we will not base our judgment on statistical tests, like Pearson's chi-square, Anderson-Darling, or Kolmogorov–Smirnov tests, because of the size (around tens of thousands of data) of the survey data on which our inferential results are based. In general, with huge sample sizes, we expect that all consistent tests would reject the H_0 hypothesis stating the equality between the empirical and the theoretical distribution (even when applied to data randomly generated from the model). Therefore, we rely on descriptive measures to summarize the discrepancy between the observed frequencies and those expected under the model.

Let the empirical data be arranged into a grouped frequency distribution with s-class intervals, and let n_j and $\widehat{n}_j$ be, respectively, the observed and the estimated (under the model) frequencies of the random variable Y over the j-th interval. Any measure of deviation, based on a synthesis of the absolute differences between $\widehat{n}_j$ and n_j can be a candidate for assessing the goodness of fit. We will evaluate the Mortara index A_1, the Pearson index A_2, the modified quadratic index A_2', and the coefficient of similarity W_p, here recalled:

$$A_1 = \frac{1}{n}\sum_{j=1}^{s}|n_j - \widehat{n}_j|, \tag{3}$$

$$A_2 = \sqrt{\frac{1}{n}\sum_{j=1}^{s}\frac{\left(n_j - \widehat{n}_j\right)^2}{\widehat{n}_j}}, \tag{4}$$

$$A_2' = \sqrt{\frac{1}{n}\sum_{j=1}^{s}\frac{\left(n_j - \widehat{n}_j\right)^2}{n_j}}, \tag{5}$$

$$W_p = \sum_{j=1}^{s}\min\left(n_j - \widehat{n}_j\right). \tag{6}$$

4 | THE RELATIVE DISTRIBUTION METHOD

The aim of this methodology, introduced in 2006 by Hancock and Morris in Reference [9], is to provide a solid methodology for the comparison of two distributions.

Let Y_0 be a random variable (r.v) representing a measurement for a population, with cdf $F_0(y)$, and pdf $f_0(y)$ (if it exists). We will call the population that generated Y_0 the *reference* population.

Suppose there is a second population, called the *comparison* population, on which the same measurement originates the r.v. Y, with cdf $F(y)$ and pdf $f(y)$ (if it exists).

The *relative* distribution of Y to Y_0 is defined as the distribution of the r.v. $R = F_0(Y)$:

- R is obtained from Y by transforming it by the cdf of Y_0, that is F_0,
- R measures the *relative rank of Y compared to Y_0*,
- R has cdf G such that $G(r) = F\left(F_0^{-1}(r)\right)$ for $0 \le r \le 1$.

We will call r, a realization of R, the *relative data*. The relative data can be interpreted as the *percentile rank that the original comparison value would have in the reference population*. The pdf of R, that is, the *relative density*, can be obtained as the derivative of $G(r)$

$$g(r) = \frac{f\left(F_0^{-1}(r)\right)}{f_0\left(F_0^{-1}(r)\right)} \qquad \text{for } 0 \le r \le 1. \tag{7}$$

The relative density can be interpreted as a density ratio. This can be seen more easily by expressing $g(r)$ explicitly in terms of the original measurement scale, y. Let the r-th quantile of R be denoted by the value y_r on the original measurement scale, so the y_r corresponding to r is $F_0^{-1}(r)$. The relative pdf is then:

$$g(r) = \frac{f(y_r)}{f_0(y_r)} \qquad \text{with } y_r = F_0^{-1}(r) \ge 0. \tag{8}$$

we want to remark, here, that the relative density is a proper pdf in the sense that it integrates to 1 over the unit interval, due to the rescaling imposed by the quantile function in the numerator and denominator of (7). Because pdfs are one of the basic building blocks of statistical theory, the fact that the relative density is a proper pdf provides the relative distribution with a firm basis for estimation, inference, and interpretation, and a general framework for methodological development [9].

The smoothness of F and F_0 ensure that $g(r)$ is continuous on (0,1). If the two distributions are identical, then the relative density is the uniform probability distribution on (0,1) and the cdf of the relative distribution is the 45° line from (0,0) to (1,1).

The relative distribution is an intuitively appealing approach to the comparison problem because the relative data, pdf and cdf have clear, simple interpretations.

The relative pdf $g(r)$ can be interpreted as a density ratio: *the ratio of the fraction of respondents in the*

comparison group to the fraction in the reference group, up to a given level of the outcome attribute Y, that is, up to $y = F^{-1}(r)$.

The relative cdf, $G(r)$, can be interpreted as the *proportion of the comparison group, whose attribute lies below the r-th quantile of the reference group*. Note that the implicit unit of comparison is the *value of the attribute on the original measurement scale*, where $y_0 = F_0^{-1}(G(r))$ represents the cut-point.

If the two r.v. Y and Y_0 are identical, then the cfd of the relative distribution is a 45° line and the pdf of the relative distribution is that of the uniform on $[0, 1]$.

5 | A PARAMETRIC APPROACH TO THE RELATIVE DISTRIBUTION

Since Pareto proposed his first income distribution model in 1896, based on empirical evidence from tax statistics, many economists and mathematicians have tried to describe empirical distributions by mathematical formulas with a few parameters.

What are the advantages of using a parametric model instead of relying directly on survey data? First, applying a theoretical model simplifies the analysis because a few parameters can subsume different distribution characteristics. In particular, the functional relationships between various inequality measures and the model parameters can be used to assess sensitivity of these measures to variations of location and shape of an underlying distribution. Second, a theoretical model well-fitted to wage or income data can be used to accurately predict wage and income distributions in different divisions. Moreover, the approximation of the empirical wage and income distributions through theoretical curves can smooth the irregularities coming from the data collecting method, what can be especially important for sparse data in high-income groups (see chap. 1 in Reference [10]). In our study, the last reason seems to be the most important, as the datasets based on sample surveys are subject to various sampling and nonsampling errors. Notice that standard imputation and calibration techniques cannot fully eliminate such kinds of errors.

A necessary condition that has to be assumed for a theoretical density to be applicable as an appropriate income distribution model is its empirical (socio-economic) or/and stochastic foundations [22, 23]. In the considerations devoted to income distribution models, it is generally accepted that the Pareto model provides the optimal fit for high-income groups, hence the convergence to the weak Pareto law [24] has become the standard requirement. There have been many attempts to explain the behavior of the Pareto model in stochastic terms, with the main one due to [25], based on Markov chains.

The Lognormal model makes use of Gibrat's law of proportionate effect to explain how income is distributed and how the distribution changes over time in a population. Despite this stochastic foundation of the model, limited flexibility due to having only two parameters and lack of convergence to the Pareto law make the model inadequate at the tails.

A stochastic mechanism leading to the Dagum distribution can be described by the solution of Kolmogorov forward equations which gives the equilibrium distribution of a diffusion process [26]. Moreover, the model has socio-economic solid fundaments from the contributions of Sylos Labini on social stratification [27]. The Dagum distribution is also the solution of a differential equation formulated on the basis of empirical evidence on income elasticity of the cumulative distribution function, based on income data from many countries and in different divisions [22].

Finally, the Singh-Maddala distribution—in contrast to the Lognormal and the Dagum models—is based on formal analogy, rather than on stochastic or economic foundations. This analogy is supported by the similarity to the lifetime distributions used in reliability theory. Despite the lack of direct foundations, which would have been interpreted in terms of socio-economic processes, it is a generalization of the Pareto distribution, and hence behaves as Pareto among the highest incomes.

Given a distribution that fits well some empirical data, a parametric version of the relative distribution method can be introduced. We will explicitly derive here the relative density and relative distribution for the Dagum model. Analogous results for the Singh-Maddala and the Lognormal distributions are available in the Appendix A.

Recalling the definition of the *relative* distribution of Y to Y_0, that is $G(r) = F\left(F_0^{-1}(r)\right)$ for $0 \le r \le 1$, and using the Dagum cdf and quantile function, given respectively in (1) and (2), we set

$$G_{D(a,b,p;a_0,b_0,p_0)}(r) = \left[1 + \left(\frac{b}{b_0}\right)^p \left(r^{-1/p_0} - 1\right)^{p/a_0}\right]^{-a}. \quad (9)$$

Analogously, the relative pdf can be derived as follows

$$g_{D(a,b,p;a_0,b_0,p_0)}(r) = \frac{a\,p\,b_0^{a_0 p_0}}{a_0\,p_0\,b^{ap}}\, R_0^{a\,p-a_0 p_0}\, \frac{\left[1 + (R_0/b_0)^{a_0}\right]^{p_0+1}}{\left[1 + (R_0/b)^a\right]^{p+1}}, \quad (10)$$

where $R_0 = b_0\left(r^{-1/p_0} - 1\right)^{-1/a_0}$.

6 | ADJUSTING FOR COVARIATES

Often some covariates vary systematically by the compared populations, and the impact of these covariates is of interest. For instance, suppose the education composition is different in the reference (men) and comparison (woman) population. We want to quantify the impact of this difference on income distribution. Further, there could be a different relationship between the covariate (say, education) and the response variable (say, income). Our purpose is to separate out the two effects.

We will follow the approach introduced by Handcock and Morris in [9], where the overall relative distribution is decomposed into:

- *A first term*: the composition effect, which measures the shift in the covariates from one population to the other;
- *A second term*: the residual effect, obtained by adjusting the reference (men) population to have the same marginal covariate composition as the comparison (women) population.

By holding the population composition constant across the gender groups, differences in the covariate-response relationships can be correctly identified. Let (Y_0, Z_0) and (Y, Z) denote the random vectors describing the reference and comparison populations, where Y_0 and Y are the response variables, while Z_0 and Z are the categorical covariates, with support $1, 2, \dots, K$.

Let π_k and π_{0k} be the probability mass functions of Z and Z_0, respectively, for $k = 1, \dots, K$. They represent *the population composition with respect to the covariate*.

The marginal density of Y can be written as

$$f(y) := \sum_{k=1}^{K} \pi_k f_{Y|Z}(y|k). \tag{11}$$

An analogous definition holds for the reference distribution Y_0

$$f_0(y) := \sum_{k=1}^{K} \pi_{0k} f_{Y_0|Z_0}(y|k). \tag{12}$$

The differences between $f(y)$ and $f_0(y)$ are also a result of the differences in the conditional densities $f_{Y_0|Z_0}(y|k)$ and $f_{Y|Z}(y|k)$, for $k = 1, \dots, K$. The latter conditional densities represent differences in the *covariate-response relationship* between the two populations.

Using these ideas, we can construct *a counter-factual distribution for the compositional difference*. We define the distribution of Y_0 *composition-adjusted* to Y to be Y_{0C} with density:

$$f_{0C}(y) := \sum_{k=1}^{K} \pi_k f_{Y_0|Z_0}(y|k). \tag{13}$$

Y_{0C} is the distribution of income from the comparison population, if they had the distribution of covariate of the reference distribution. Using the composition-adjusted response distribution, we can decompose the overall relative distribution into a component that represents the effect of changes in the marginal distribution of the covariate *the composition effect*, and a component that represents *the residual effect*. In terms of density ratios, we have:

$$\frac{f(y_r)}{f_0(y_r)} = \frac{f_{0C}(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0C}(y_r)}. \tag{14}$$

We would like to conclude with a short discussion about other econometric methods for assessing gender gap. Within a classical regression approach on income Y, the model can incorporate a qualitative predictor for gender via a dummy variable. Juhn, et al. [28] develop a regression method that separates changes in covariates from changes in the regression coefficients, to obtain estimated returns to the covariates, changes in the mean residual earnings gap between the groups and changes in the standard deviation of the men's residual earning variation. They apply their method to investigate the race-gap in wages, and Blau and Kahn [29] apply it to the gender-gap. However, the principal limitation of regression is related to the fact that it works only with average differences. On the other side, quantile regression [30] can track distributional changes. We opted for the relative distribution framework because it provides a fully distributional approach to location and/or shape differences among two distributions, and covariate decomposition. We briefly develop, in the sequel, a parametric version of covariate adjustment. Afterward, in Section 7 we apply such decompositions to real data, and show how they offer a deeper insight to gender gap.

7 | A PARAMETRIC APPROACH TO COVARIATE ADJUSTMENT

We want now to introduce a parametric version of the adjustment of covariates. Our aim here is to define the estimated marginal densities $\widehat{f}(y)$, $\widehat{f}_0(y)$ for the comparison and the reference population, and the estimated composition-adjusted density $\widehat{f}_{0C}(y)$: they will be based, again, on the Dagum model.

To this purpose, we will consider the subsample, drawn from the reference population, having value $Z = k$ for the covariate, and estimate a Dagum model on it, yielding

$$\widehat{f}_{Y|Z}(y|k) = f_D\left(y; \widehat{a}_k, \widehat{b}_k, \widehat{p}_k\right), \qquad \text{for } k = 1, \dots, K. \tag{15}$$

Similarly, we take the subsample from the comparison population, with covariate $Z = k$, and fit a Dagum model

to it, giving raise to

$$\widehat{f}_{Y_0|Z_0}(y|k) = f_D\left(y; \widehat{a}_{0k}, \widehat{b}_{0k}, \widehat{p}_{0k}\right), \qquad \text{for } k = 1, \ldots, K. \tag{16}$$

We may substitute π_k and π_{0k} in (11) and (12) by their natural estimators $\widehat{\pi}_k$ and $\widehat{\pi}_{0k}$, that are the empirical relative frequencies of observing the value k for covariate Z and Z_0, in the samples drawn from the reference and from the comparison populations, respectively. We obtain

$$\widehat{f}(y) := \sum_{k=1}^{K} \widehat{\pi}_k f_D\left(y; \widehat{a}_k, \widehat{b}_k, \widehat{p}_k\right), \qquad \text{and}$$
$$\widehat{f}_0(y) := \sum_{k=1}^{K} \widehat{\pi}_{0k} f_D\left(\widehat{a}_{0k}, \widehat{b}_{0k}, \widehat{p}_{0k}\right) \tag{17}$$

Finally, we can decompose the overall estimated density ratio $\widehat{f}(y_r)/\widehat{f}_0(y_r)$ into a product of two density ratios

- the first given by $\widehat{f}_{0C}(y_r)/\widehat{f}_0(y_r)$, which represents the effect on Y_0 due to the different marginal density of the covariate Z, and
- the second, $\widehat{f}(y_r)/\widehat{f}_{0C}(y_r)$, that expresses the residual effect.

8 | APPLICATIONS TO EUSILC DATA FOR POLAND AND ITALY

The *European Union Statistics on Income and Living Conditions* (EU-SILC) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions [31]. This instrument is anchored in the European Statistical System (ESS).

The *employee gross income* consists of.

- cash gross total income PY010G,
- cash benefits from self employment PY050G,
- gross pension individual plans PY080G,
- gross unemployment benefits PY090G,
- old age benefits PY100G,
- survival benefits PY110G,
- disability benefits PY130G,
- education related allowances PY140G.

Remarkable differences between Poland and Italy, primarily related to the discrepancy between men and women across regions, have been found in the literature [8]. To uncover the factors contributing to the gender discrepancy, it is useful to move beyond the typical focus on average or median earnings differences, towards a view on how the entire distribution of women's earnings (which generated the comparison Y_0) relative to men's (originating the reference Y) compares. The next natural step is hence to search for the socioeconomic factors that could explain the differences observed in the income distribution for men and women, employing the method based on the covariate decomposition introduced in Section 5.

In our analysis, we will employ EU-SILC data of wave 2018. We proceed as follows:

- *First step*: we fit the three models introduced in Section 1 to males and female empirical data, in both countries, and select the best fitting solution;
- *Second step*: we evaluate the relative distributions, based on the empirical data and on the estimated models;
- *Third step*: we decompose the relative distribution with respect to the following covariates:
  - Covariate 1: *Education level* (PE040),
  - Covariate 2: *Managerial position* (PL150),
  - Covariate 3: *Working time* (PL060 + PL100).

The estimated models, superimposed to the histograms of EU-SILC data, for income distribution of women (left panel) and men (right panel) for Italy and Poland, are given in Figures 1 and 2, respectively.

Table 1 presents the goodness-of-fit measures, and shows that in almost all cases, the Dagum distribution outperforms its competitors and provides one more piece of evidence that it can be considered as a good model for survey income data. The Singh-Maddala model yielded a good fit in some cases, while the Lognormal distribution turned out to be rather poor for our purposes. Notice that, while A_1, A_2 and A_2' indicate a measure of distance between the model and the binned data (hence best values are the lowest ones), W_p is a measure of similarity, therefore the better the fit, the higher the value.

In Figure 3 we observe, for each country, the differences between the Dagum density curves estimated for men and women.

We see that the estimated model for males income first-order stochastically dominates the one for females income, both in Poland and in Italy. This property can be easily checked by comparing the parameters of the Dagum models in Table 1, using results derived by Klonner [32].

Here, we prefer to fully exploit the relative distribution approach, as follows. First-order dominance $G \geq_1 F$ can also be written in terms of quantiles, requiring $G^{-1}(p) \leq F^{-1}(p)$ for all $p \in [0, 1]$, with strict inequality for at least one p [33]. This means that ranking the individuals of each population in terms of their income level,

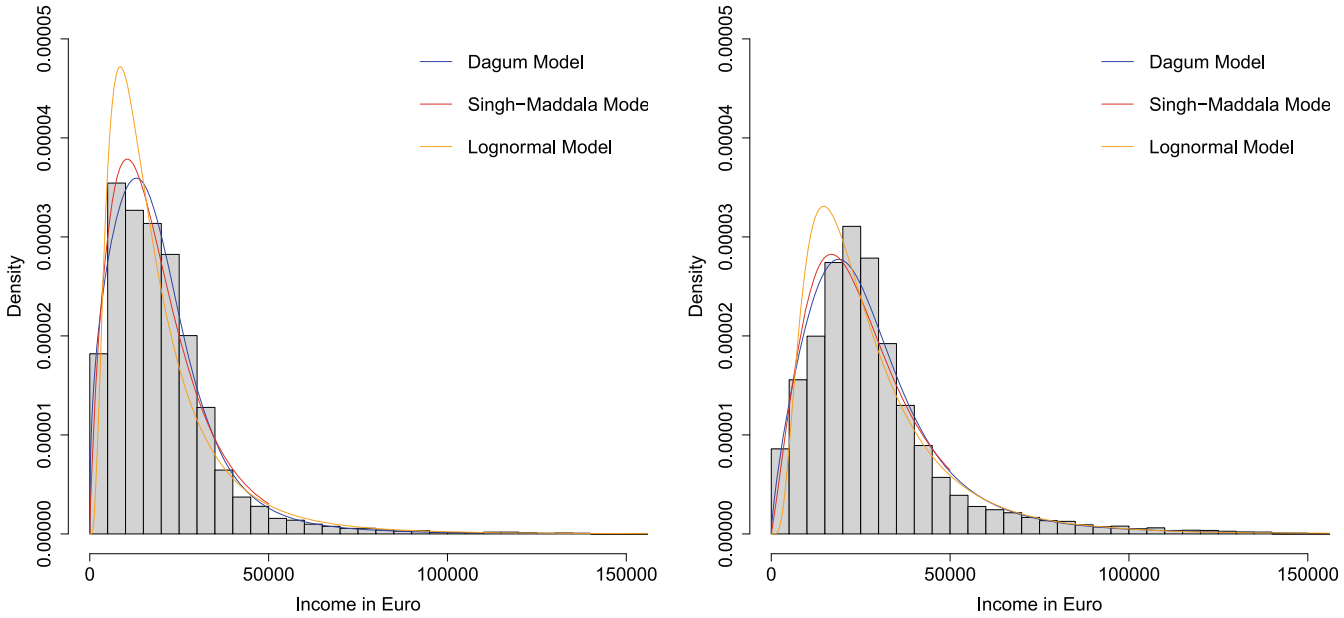

FIGURE 1 Estimated models, superimposed to the histogram of EU-SILC data, for income of Italian women (left panel) and men (right panel).

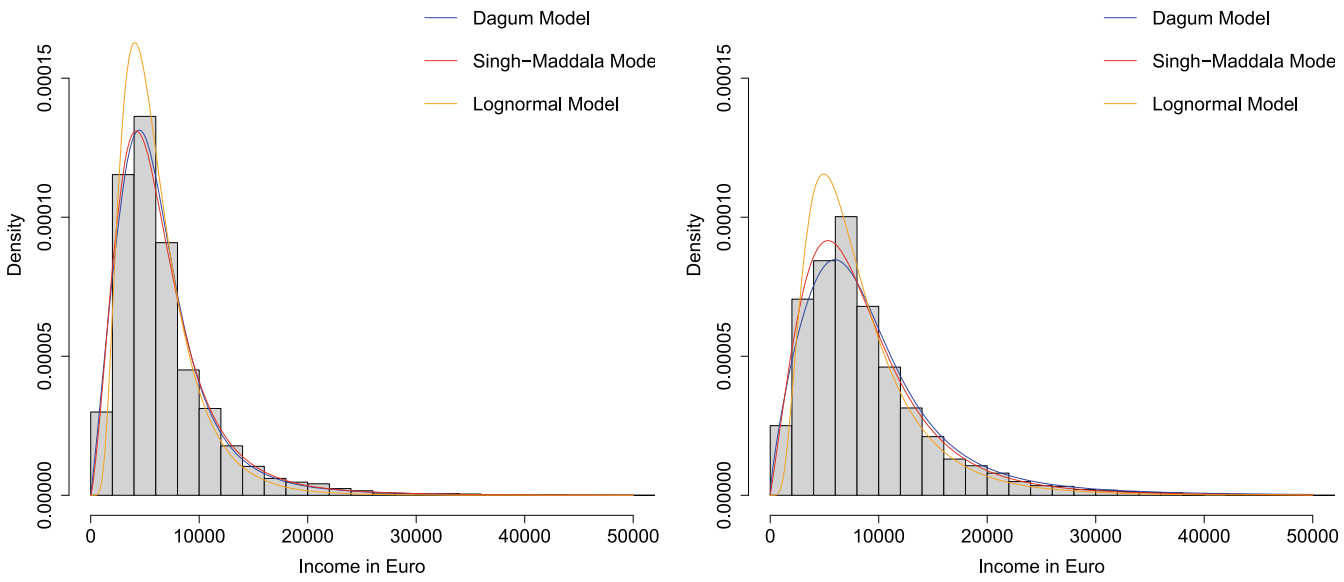

FIGURE 2 Estimated models, superimposed to the histogram of EU-SILC data, for income of Polish women (left panel) and men (right panel).

distribution G first-order stochastically dominates distribution F if the level of income in each position p in G is at least as high as the corresponding level of income in the same position in F. The plots of the relative parametric cdf in Figure 5 represent the estimated males distribution first-order dominating the females distribution of income. Moreover, first-order dominance implies all higher order dominances, among which the Lorenz dominance.

However, a part from this consideration, the distributions of gender groups differ in all aspects—scale, location, and shape. From a direct comparison of the density curves, it seems hard to grasp and characterize the gender gap.

Therefore, we employed the relative distribution approach to assessing the gender gap, first by evaluating it on the empirical data and then from the fitted Dagum Models. Figure 4 shows the pdf of the relative density for Italy and Poland, based on EUSILC 2018 data. The densities of relative distributions can be interpreted in terms of density ratios of the compared populations, that is, the ratios of the fraction of respondents in the comparison

TABLE 1 Estimated parameter values and goodness of fit measure A_1, A_2, A_2' and W_p (defined in (2.3), (2.4), (2.5), and (2.6)) obtained for EU-SILC data (2018) for Italy and Poland.

	Estimated parameter values			Indexes of goodness of fit			
Dagum	**a**	**b**	**p**	A_1	A_2	A_2'	W_p
Males Poland	3.2261	9928.55	0.5530	0.0803	0.1018	0.1257	0.9598
Females Poland	3.5579	8000.65	0.4941	**0.0205**	0.0538	0.0526	**0.9897**
Males Italy	3.1610	32478.90	0.5525	**0.1145**	**0.1747**	**0.1555**	**0.9428**
Females Italy	3.4373	26071.80	0.4091	**0.0769**	**0.0987**	**0.0931**	**0.9615**
Singh-Maddala	**a**	**b**	**q**	A_1	A_2	A_2'	W_p
Males Poland	2.0606	11541.10	2.0472	**0.0636**	**0.0748**	**0.0717**	**0.9682**
Females Poland	2.3524	7081.63	1.5319	0.0281	**0.0532**	**0.0506**	0.9860
Males Italy	2.0949	33096.48	1.6736	0.1483	0.2601	0.1912	0.9259
Females Italy	1.7463	34694.68	2.8934	0.1010	0.1147	0.1404	0.9495
Lognormal	μ	σ^2		A_1	A_2	A_2'	W_p
Males Poland	8.8489	0.5894		0.0802	0.2053	0.1245	0.9599
Females Poland	8.5855	0.5254		0.0865	0.5861	0.1539	0.9567
Males Italy	10.0297	0.6617		0.2289	0.7743	0.2713	0.8855
Females Italy	9.6114	0.7517		0.1843	0.2142	0.2014	0.9078

Note: Best values of each index are bolded, for each sample.

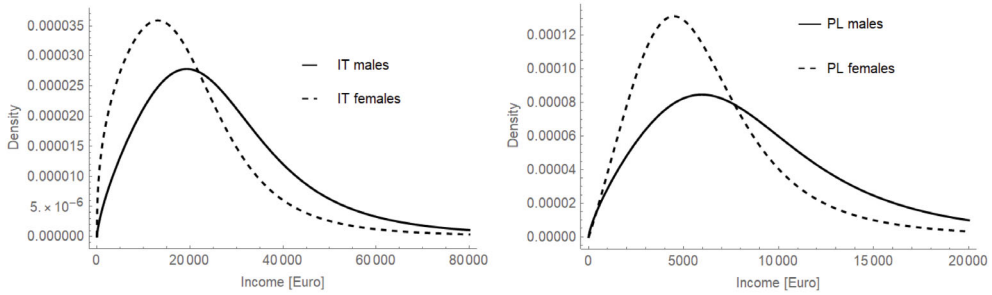

FIGURE 3 Dagum estimated models for Italian (left panel) and Polish (right panel) EU-SILC data.

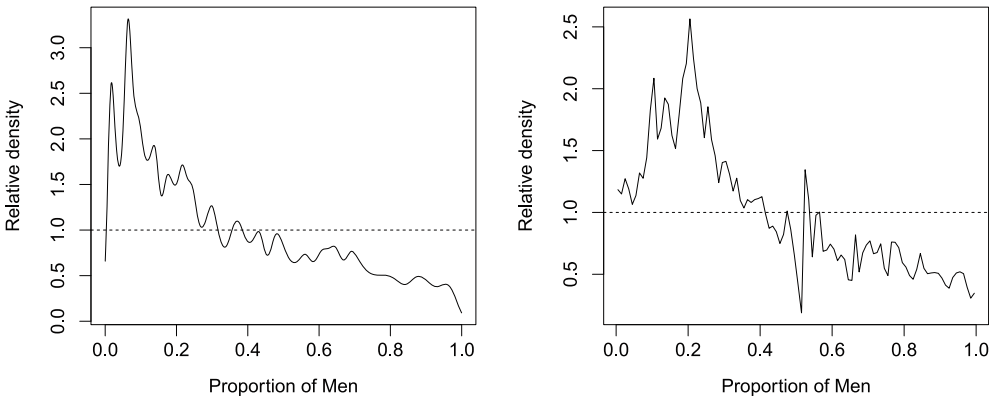

FIGURE 4 Relative density obtained from the (comparison) female income to the (reference) male income, for Italy (left panel) and Poland (right panel), based on empirical data.

group to the fraction in the reference group, at a given level of income. For example, for Poland, the relative density at the 2nd decile of men's income is about equal to 2, meaning that women are about twice as likely as men to fall at this income level. The respective ratio for Italy is about 1.5. In general, the relative density for Italy is more polarized than for Poland, having a rather hyperbolic shape. Therefore, the highest discrepancy can be observed in the bottom

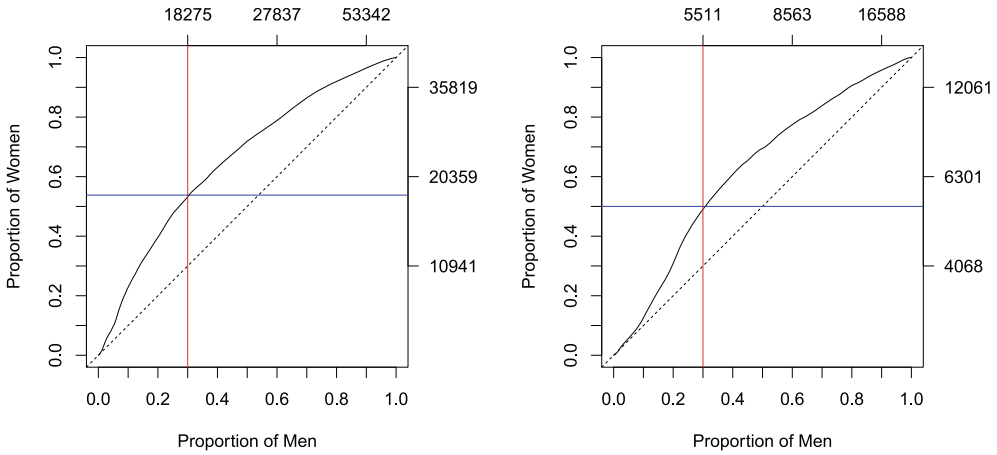

FIGURE 5 Relative distribution obtained from the (comparison) female income to the (reference) male income, for Italy (left panel) and Poland (right panel), based on empirical data. The values of the third, sixth and ninth decile of income (in Euro) are indicated on the right and upper axes.

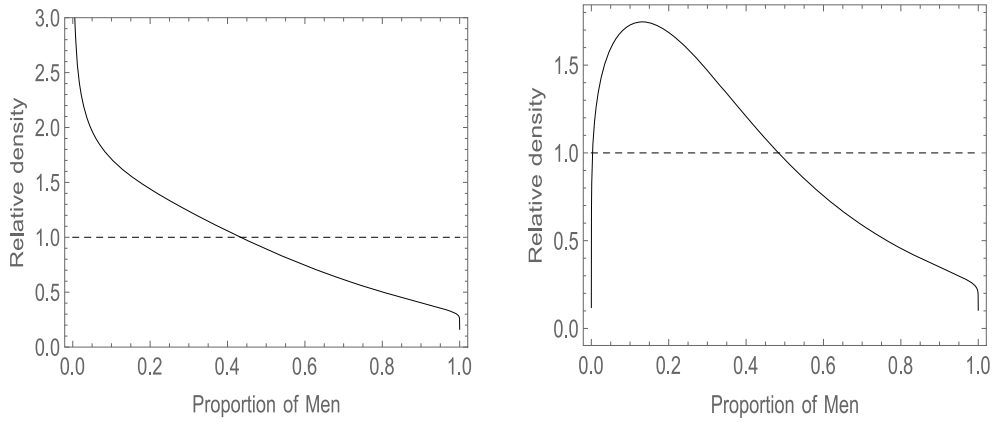

FIGURE 6 Relative density for Italian data (left panel) and Polish data (right panel), based on the estimated Dagum models (parametric approach).

and top income deciles. In contrast, in Poland the highest density ratio was observed closer to the middle of income range, namely at the 3rd decile group.

The curves of the relative income distribution, shown in Figure 5, provide a rich and detailed information. Each point on the curve has a precise interpretation. For instance, in the right panel, it can be seen that at the third decile of the Polish men's earnings distribution, that is $p = 0.3$, it holds $G(0.3) = 0.50$. This means that approximately 50 percent of women earn less than the third decile men's income. The situation is even worse in Italy (left panel), where as much as 54 percent of women earn less than the third decile men. Note that one of the peculiarities of the relative graphs is that the distance between Euro values on the right-hand scale is measured in units of persons rather than in euros. It is worth pointing out that the gender gap in Poland increased in 2018 comparing to 2015, despite the fact that income inequality in this country decreased over the same period [8].

In the next step, we smoothed out the irregularities coming from random selection of empirical data, using the parametric approach based on the Dagum model. The parametric versions of the relative density and relative distribution function are presented in the Figures 6 and 7, respectively. When comparing the empirical cdfs and pdfs (Figures 4 and 5) with the corresponding parametric ones (Figures 6 and 7), it can be noticed that the empirical curves are not only irregular but also visibly underestimated with respect to the right tails, especially for Poland. For this reason the parametric gender gaps are higher for upper deciles and relatively smaller for low-income groups. The theoretical curves present the model-based versions of the empirical ones so they can be viewed as maximum likelihood estimates of the population counterparts. Such an approach can be particularly helpful in the case of the relative densities having highly irregular estimates, even more where the samples were substantially smaller, that is, for Polish data.

Figure 8 shows the decomposition of the relative income distribution of women in relation to men, assuming the working position (managerial or not, variable PL150: Managerial position) as the explanatory variable. The first panel from the left shows the (uncorrected) relative density of income differences between men and women, the middle panel represents the effects of differences in the distributions of the explanatory variable, and the right panel represents the counterfactual distribution – that is, the expected relative density for men's and

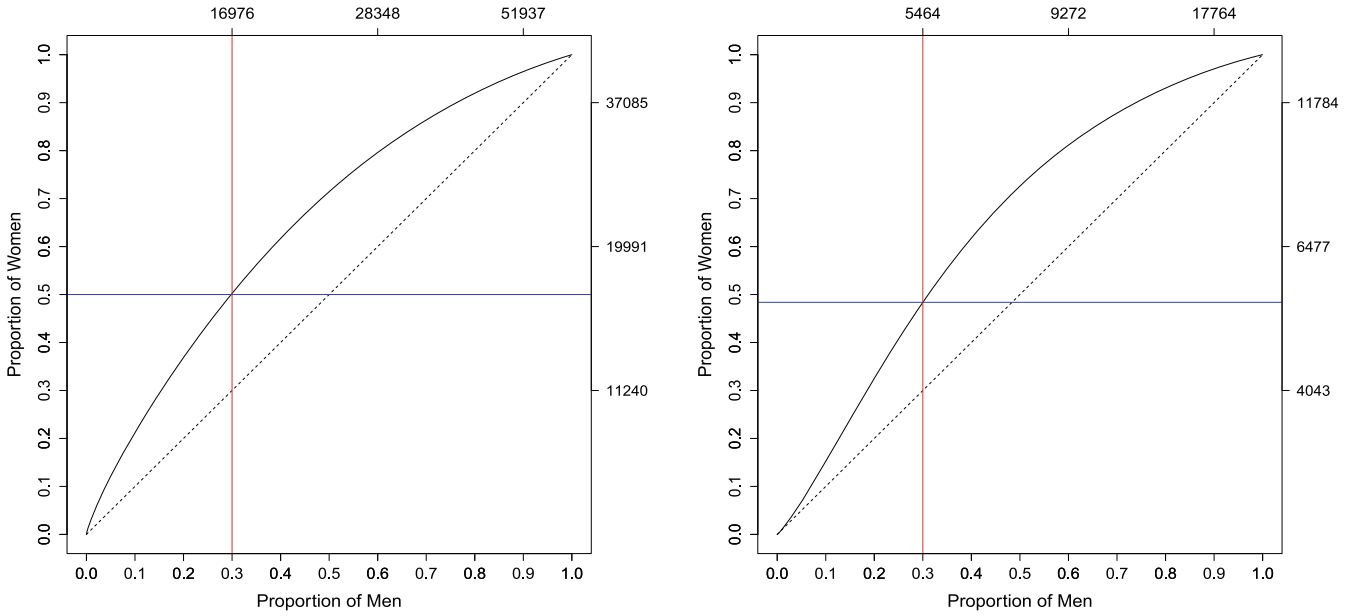

FIGURE 7 Relative distribution for income in Italy (left panel) and Poland (right panel), based on the estimated Dagum models (parametric approach).

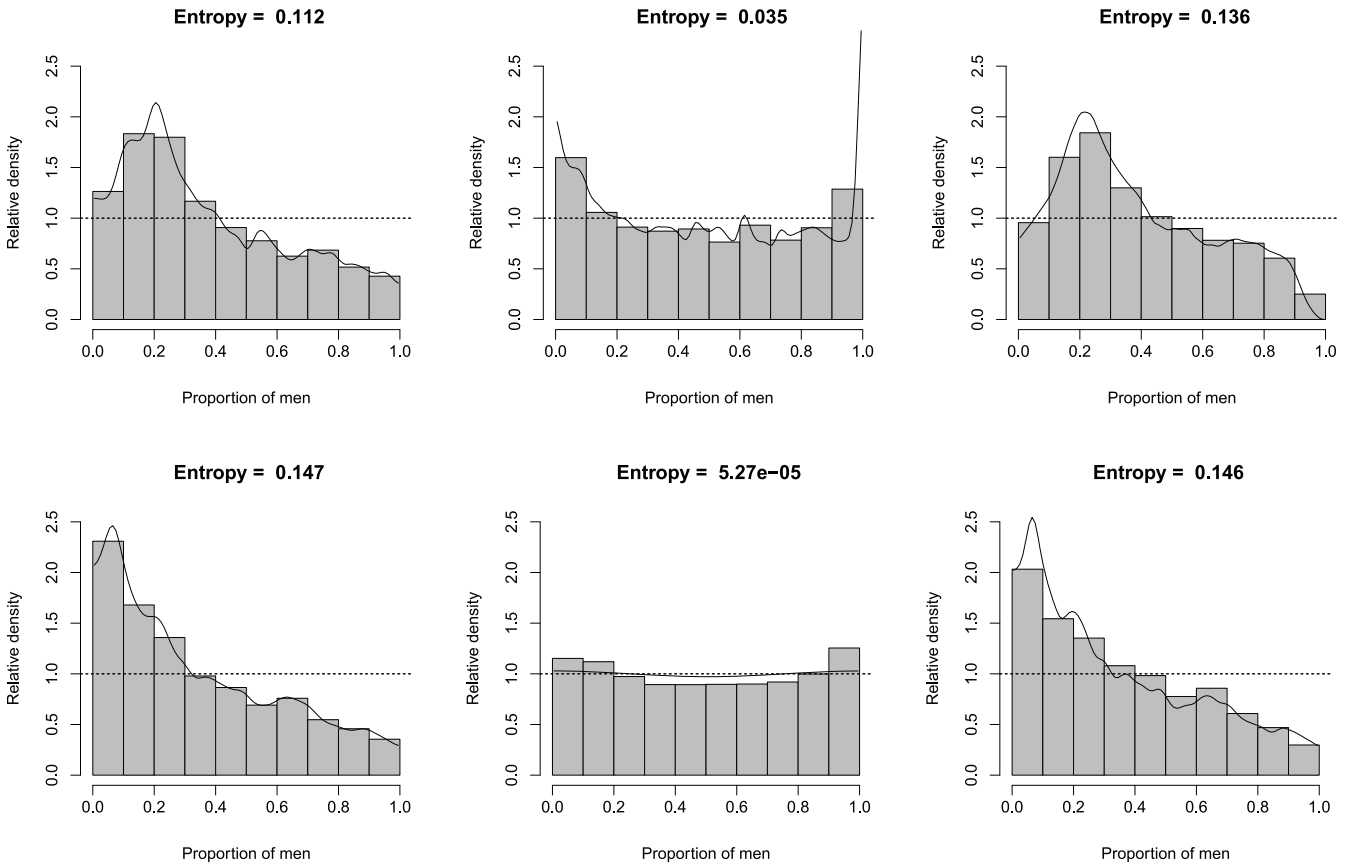

FIGURE 8 The three plots for Polish data (upper panel) and Italian data (lower panel) to assess the effect of managerial position on gender equality.

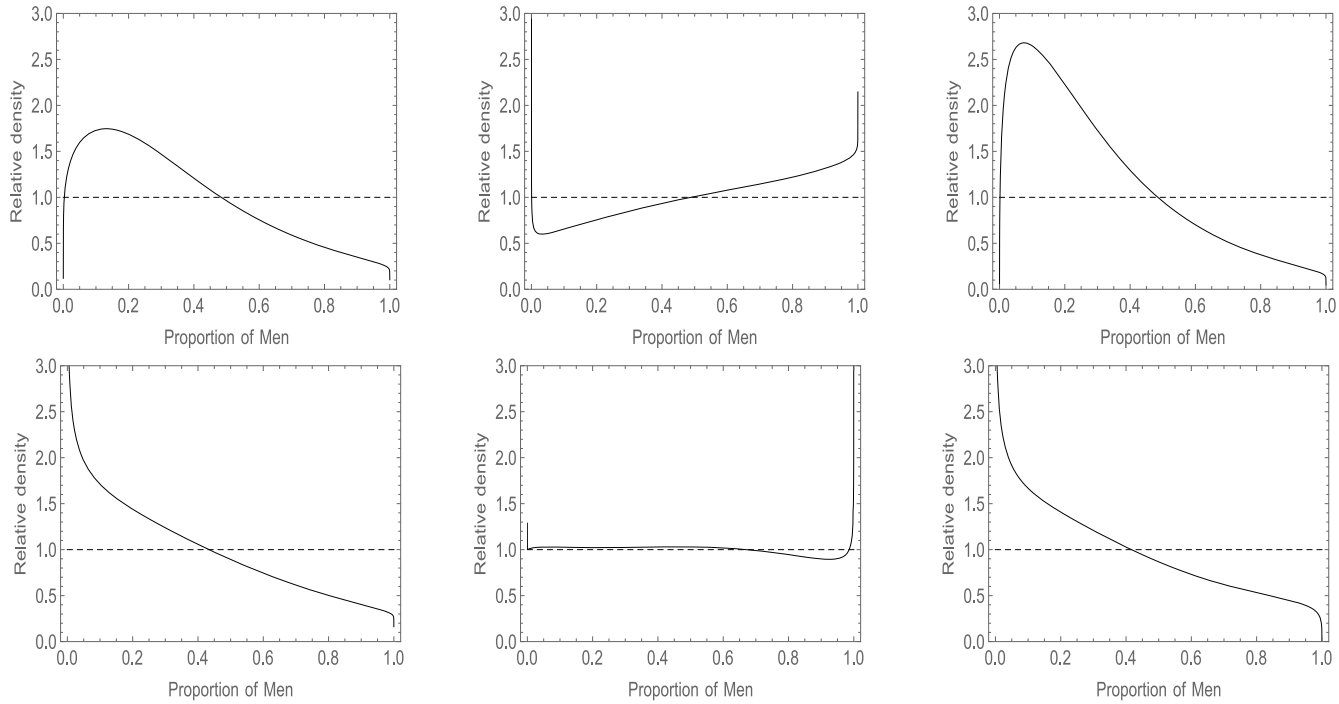

FIGURE 9 The three plots for Polish data (upper panel) and Italian data (lower panel) to assess the effect of managerial position on gender equality, obtained through the parametric approach.

women's income distributions when assuming the same profiles of positions held in both groups (upper panels refer to Poland, lower panels to Italy).

The comparison of the three relative densities provides a valuable tool for assessing the relative magnitude and nature of the impact of covariate distributions, and of the different covariate-to-response relationships in gender groups. The distribution in the middle panel for Poland is mildly U-shaped and, in the central part, it is close to the uniform. Therefore, the difference in the structure of management positions observed between the two cohorts in central deciles has little effect on the observed income gap. More significant differences occur in the extreme deciles, which suggests some income polarization of these parts of populations, in relation to the position held. Women from the last decile occupy higher positions; however, the latter does not translate into corresponding earnings. As a result, the income gap in these groups, adjusted by the type of position held in the counterfactual distribution, widens (right panel). On the other hand, results on Italian data are somehow different and indicate almost no impact of the managerial position on the gender gap in that country. The patterns observed in Figure 8 are confirmed by the parametric approach, whose results are shown in Figure 9.

The relative distribution, capturing all the information that is necessary and sufficient for strong scale-invariant comparison, provides a general framework for defining a variety of summary measures. Among the measures of distributional divergence based on the relative distribution, we consider the Kullback–Leibler divergence between the reference and the comparison distributions, due to its useful decomposition properties. It is given by $D(F;F_0) = \int_{-\infty}^{\infty} log\left(\frac{f(x)}{f_0(x)}\right) dF(x) = \int_0^1 \log(g(r))\, g(r)\, dr$. On the right hand side we find the (negative) entropy of the relative density, the quantity that we calculated and reported in Figures 8, 9, and 10. We can interpret $D(F;F_0)$ as the expected information for discriminating the relative density $g(r)$ from a uniform distribution, based on a single observation from R (see Handcock and Morris [9], chap. 5.3).

The effect of adjusting the relative income distribution of women to men in terms of education levels (variable PE140: Education level) is plotted in Figure 10. The left-hand panel shows the (unadjusted) relative density of income differences between women and men, the middle panel represents the effects of differences in the distribution of education levels, and the right-hand panel shows the expected relative density for male and female income distributions assuming the same educational profiles in both groups. In Poland (Figure 10, top panels), the relative distribution of education levels (middle panel) is almost uniform, so this covariate does not explain the income gap between women and men, except for the top decile group, where we can observe a much greater share of better-educated women. However, this situation does not translate into the amount of their income—therefore, the

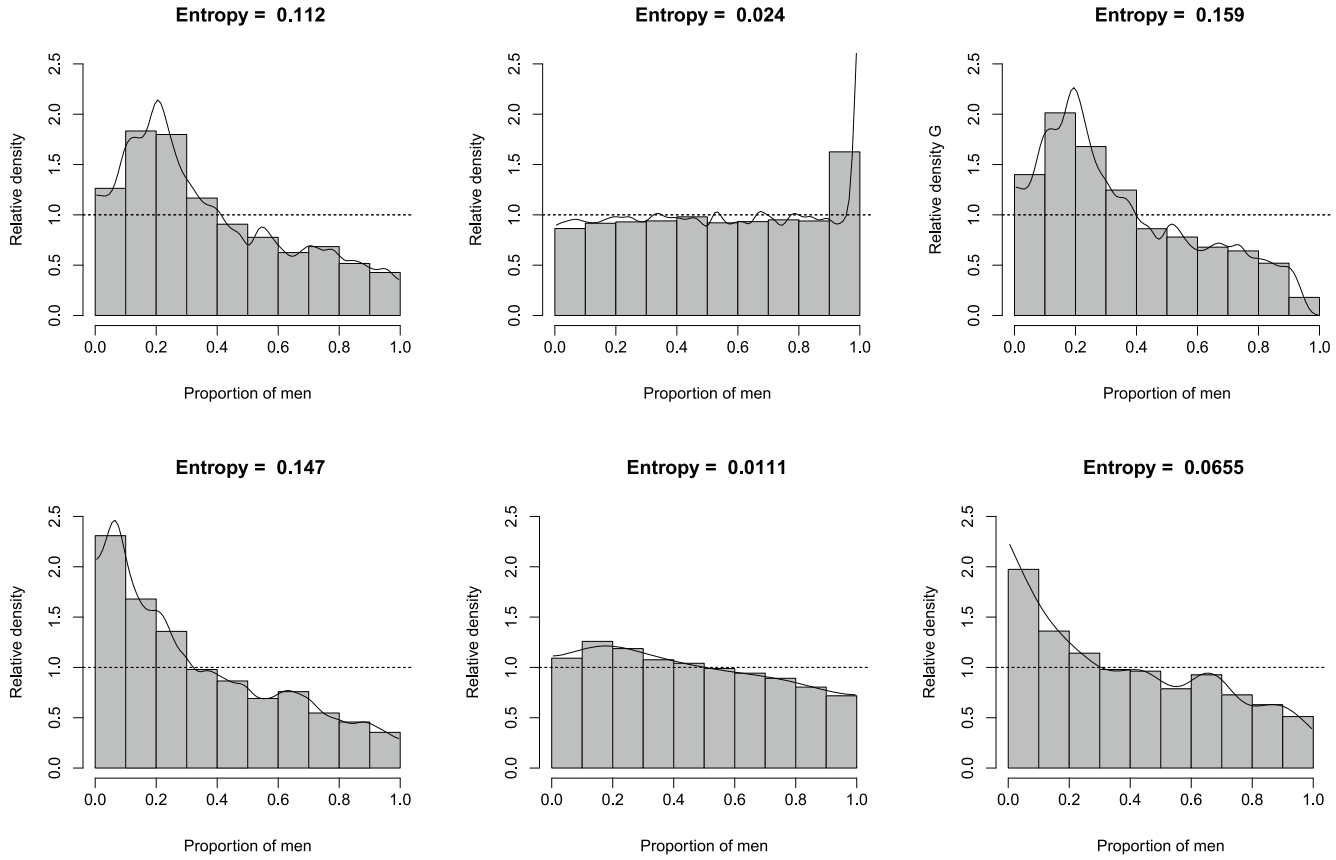

FIGURE 10 The three plots for Polish data (upper panel) and Italian data (lower panel) to assess the effect of education on gender equality.

last bar of the histogram in the counterfactual distribution is significantly lower (right panel). The situation is different in Italy (Figure 10, bottom panels)—the covariate Education level seems to have an impact on the gender gap in this country. Apart from the lowest income groups, the observed differences in education levels are in favor of men and grow with increasing incomes, with the most significant difference in the top decile (middle panel). Consequently, the counter-factual relative distribution, showing the hypothetical situation of having the same covariate structures in both gender groups (right panel) is visibly less dispersed than the actual distribution (left panel).

In the central top panel of Figure 11, we can observe the density of the random variable *R*, which is created by comparing the adjusted distribution of men's income in Poland (using the variable PL060 + PL100: Working time) in relation to the unadjusted distribution, which allows us to assess the impact of differences in the structure of hours worked on the observed income gap. The middle panel shows the portion of the income gap that can be attributed to the effects of changes in the distribution of weekly hours worked.

In general, it can be stated that a significant part of the observed income gap between men and women results from differences in their work schedules. After correcting the relative distribution with the use of this variable, we obtain a distribution (right panel) that is closer to the uniform than the original one (left panel). We can see that after considering working time, the relative situation of women improves the most in the last deciles, as shown in the right panel, and perhaps even more clearly, in the central plot. If women in these groups worked the same as men (i.e., more), with the current structure of their earnings, the income gap would decrease. An interesting situation is observed in the last decile group. After considering the covariate adjustment, the income gap increases so that the share of women's income in the highest decile group becomes minimal compared to men's pay. This may be because women in this income group work many more hours (or due to other concomitant variables not considered in our analysis). Therefore, the differences in working time do not explain the income differences between the top 10 per cent of the wealthiest people. In the first two deciles of the distribution, the income situation of women in terms of earnings to working time is also unfavorable—in the counter factual distribution we observe a decrease in the share of women in such decile groups.

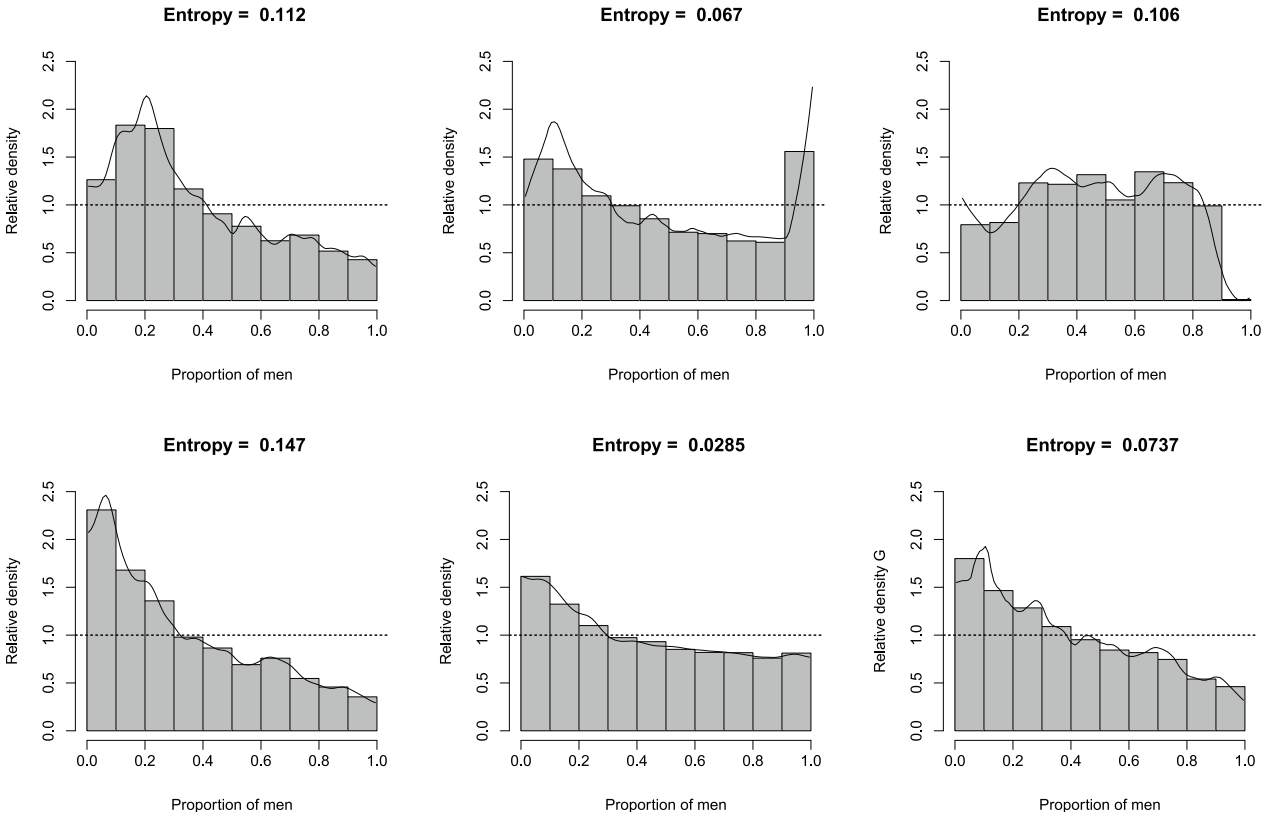

FIGURE 11 The three plots for Polish data (upper panel) and Italian data (lower panel) to assess the effect of working time on gender equality.

In Italy (Figure 11, bottom panels), the impact of differences in working time schedules on the gender gap is visible but much smaller than in Poland. The observed discrepancy in time schedules between gender groups may only partially explain the actual gender gap, so the counterfactual distribution is just a little less dispersed that the original one.

9 | CONCLUSIONS AND FURTHER WORK

We studied the gender gaps in Poland and in Italy, using the relative distribution method, a non-parametric approach to the comparison of distributions. We also assessed the impact of selected covariates, describing the personal or household characteristics of the samples, on the existing gender gaps in both countries.

We contributed to the literature by introducing a parametric version to estimating the relative distribution and to its decomposition with respect to covariates. The methods applied in the study turned out to be relevant to describe the gender gap for the entire income range and smoothed out the irregularities due to sample data. They also evaluated the impact of the main drivers on the income discrepancies between men and women. The parametric approach based on the Dagum model made it possible to better describe the existing gender gaps in both countries, especially at the tails. The natural covariates considered in the study, including education level, working time and the position held (managerial or not) partially account for the gender gaps in both countries. The observed gender gaps should also be attributed to a different relationship between the income and the covariates across the gender groups, or are due to the other factors not included in the study. Differences and similarities in the compared countries have been highlighted and discussed. Naturally, the construction of the counterfactual distribution for a single covariate, here considered, can be extended to the multivariate case, and may be the topic of future research. Further work could be devoted to implement a combination of different theoretical distributions for the comparison, which may be useful for some empirical data broken down by occupation or social group. There is no reason, in principle, for requiring that the reference and the comparison distribution come from the same parametric family. Another interesting issue, beyond the scope of the present paper, is the construction of parametric confidence intervals for the relative density and the relative distribution function.

ACKNOWLEDGMENTS

We thank two anonymous referees for their careful reading of the original version of the research. Their pertinent comments helped us to improve the paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from EUSILC2018. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of EUSILC2018.

ORCID

Francesca Greselin https://orcid.org/0000-0003-2929-1748

REFERENCES

1. OECD, *Divided we stand*, OECD Publishing Paris, Paris, 2011 https://www.oecd.org/els/soc/49170768.
2. OECD, *In it together: Why less inequality benefits all*, OECD Publishing Paris, Paris, 2015 https://www.oecd.org/els/soc/OECD2015-In-It-Together-Chapter1-Overview-Inequality.png.
3. F. D. Blau and L. M. Kahn, *Gender differences in pay*, J. Econ. Perspect. 14 (2000), no. 4, 75–99.
4. D. Leythienne and P. Ronkowski, *A decomposition of the unadjusted gender pay gap using structure of earnings survey data*, Vol 10, Luxembourg, Publications Office of the European Union, Luxembourg, 2018, 796328.
5. TÁRKI. 2009 European Social Report from. http://old.tarki.hu/en/research/european_social_report/european_social_report_2009_full.png.
6. A. Jędrzejczak, *Regional income inequalities in Poland and Italy*, Comp. Econ. Res. 18 (2015), no. 4, 27–45.
7. M. Zenga and A. Jędrzejczak, *Decomposition of the Zenga inequality index* I(Y) *into the contributions of macro-regions and income components - an application to data from Poland and Italy*, Argumenta Oeconomica 1 (2020), no. 44, 101–125.
8. F. Greselin and A. Jędrzejczak, *Analyzing the gender gap in Poland and Italy, and by regions*, Int. Adv. Econ. Res. 26 (2020), 433–447.
9. M. S. Handcock and M. Morris, *Relative distribution methods in the social sciences*, Springer Science & Business Media, New York, NY, 2006.
10. C. Kleiber and S. Kotz, *Statistical size distributions in economics and actuarial sciences*, Vol 470, John Wiley & Sons, Hoboken, New Jersey, 2003.
11. B. C. Arnold, *Pareto distributions*, 2nd ed., Chapman & Hall, New York, 2015.
12. P. R. Tadikamalla, *A look at the Burr and Related distributions*, Int. Stat. Rev. 48 (1980), 337–344.
13. C. Kleiber, "*A guide to the Dagum distributions*," *Modeling income distributions and Lorenz curves*, Springer, Berlin, 2008, pp. 97–117.
14. S. K. Singh and G. S. Maddala, *A function for size distribution of income*, Econometrica 44 (1976), 963–970.
15. J. B. McDonald, *Some generalized functions for the size distribution of income*, Econometrica 52 (1984), 647–663.
16. C. Dutang, V. Goulet, and N. Langevin, *Feller-Pareto and related distributions: Numerical implementation and actuarial applications*, J. Stat. Softw. 103 (2022), no. 1, 1–22.
17. V. Brazauskas, *Fisher information matrix for the feller-Pareto distribution*, Stat. Probab. Lett. 59 (2002), no. 2, 159–167.
18. C. Domański and A. Jedrzejczak, *Income inequality analysis in the period of economic transformation in Poland*, Int. Adv. Econ. Res. 8 (2002), no. 3, 212–220.
19. J. B. McDonald and Y. J. Xu, *A generalization of the Beta distribution with applications*, J. Econ. 66 (1995), no. 1–2, 133–152.
20. J. Aitchison and J. A. C. Brown, *The lognormal distribution with special reference to its uses in econometrics*, Cambridge University Press, Cambridge, 1957.
21. G. E. P. Box, A. Luceño, and M. del Carmen Paniagua-Quinones, *Statistical control: By monitoring and feedback adjustment*, John Wiley & Sons, Hoboken, New Jersey, 2011.
22. C. Dagum, *A new model of personal income distribution: Specification and estimation*, Econ. Appl. 30 (1977), no. 3, 413–437.
23. C. Metcalf, *An econometric model of the income distribution*, Markham Publishing Company, Chicago, 1972.
24. B. Mandelbrot, *The Pareto-Lévy law and the distribution of income*, Int. Econ. Rev. 1 (1960), no. 2, 79–106.
25. D. G. Champernowne, *A model of income distribution*, Econ. J. 53 (1953), 318–351.
26. L. Fattorini and A. Lemmi, *The stochastic interpretation of the Dagum personal income distribution: A tale*, Underst. Stat. 66 (2006), no. 3, 325–329.
27. P. Sylos Labini, "*Le stratificazioni sociali*," *Profili dell'Italia repubblicana*, O. Cecchi and E. Ghidetti (eds.), Editori Riuniti, Rome, 1985.
28. C. Juhn, K. M. Murphy, and B. Pierce, "*Accounting for the slowdown in black?White wage convergence*," *Workers and their wages*, M. Kosters (ed.), American Enterprise Institute Press, Washington, DC, 1991, pp. 107–143.
29. F. D. Blau and L. M. Kahn, *Rising wage inequality and the U.S. gender gap*, Am. Econ. Rev. 84 (1994), 23–28.
30. M. Buchinsky, *Quantile regression, Box-cox transformation model, and the U.S. wage structure, 1963–1987*, J. Econ. 65 (1995), 109–154.
31. Eurostat. European union statistics on income and living conditions (eu-silc). 2018 https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions, data retrieved from Eurostat (last accessed, Feb. 2020).
32. S. Klonner, *The first-order stochastic dominance ordering of the Singh-Maddala distribution*, Econ. Lett. 69 (2000), 123–128.
33. C. García-Gómez, A. Perez, and M. Prieto-Alaiz, *A review of stochastic dominance methods for poverty analysis*, J. Econ. Surv. 33 (2019), no. 5, 1437–1462.

How to cite this article: F. Greselin, A. Jędrzejczak, and K. Trzcińska, *A new parametric approach to gender gap with application to EUSILC data in Poland and Italy*, Stat. Anal. Data Min.: ASA Data Sci. J. (2023), 1–17. https://doi.org/10.1002/sam.11623

APPENDIX A

In this section we report the relative pdf/cdf for the Lognormal and the Singh-Maddala distributions.

Lognormal

Recalling that $F^{-1}(r) = \exp\big(\mu + \sigma\Phi^{-1}(r)\big)$, we have that the relative cdf is given by

$$G_{L(\mu,\sigma;\mu_0,\sigma_0)}(r) = F\left(F_0^{-1}(r)\right) = \Phi\left\{\frac{\mu_0 + \sigma_0\Phi^{-1}(r) - \mu}{\sigma}\right\}$$

and the relative pdf can be derived as follows

$$\begin{aligned} g_{L(\mu,\sigma;\mu_0,\sigma_0)}(r) &= \frac{f\left(F_0^{-1}(r)\right)}{f_0\left(F_0^{-1}(r)\right)} \\ &= \frac{\sigma_0}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[\mu_0 + \sigma_0\Phi^{-1}(r) - \mu\right]^2 \right. \\ &\quad \left. + \frac{1}{2\sigma_0^2}\left[\mu_0 + \sigma_0\Phi^{-1}(r) - \mu\right]^2\right\}. \end{aligned}$$

Singh-Maddala

We have that the relative cdf is given by

$$\begin{aligned} &G_{SM(a,b,q;a_0,b_0,q_0)}(r) \\ &\quad = 1 - \left\{1 + \left[\frac{b_0\left[(1-r)^{-1/q_0} - 1\right]^{1/a_0}}{b}\right]^a\right\}^{-q} \end{aligned}$$

and the relative pdf is given by

$$\begin{aligned} &g_{SM(a,b,q;a_0,b_0,q_0)}(r) \\ &\quad = \frac{a\,b^{-a}\,q}{a_0\,b_0^{-a_0}\,q_0}\left[F_0^{-1}(r)\right]^{a-a_0}\left[1 + \left(F_0^{-1}(r)/b\right)^a\right]^{-(q+1)} \\ &\qquad \times \left[1 + \left(F_0^{-1}(r)/b_0\right)^{a_0}\right]^{q_0+1}, \end{aligned}$$

where $F_0^{-1}(r) = b_0\left[(1-r)^{-1/q_0} - 1\right]^{1/a_0}$.