ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 6 (309), 2014

Daniel Kosiorowski[*]

INCOME DISTRIBUTION MODELS AND INCOME INEQUALITY MEASURES FROM THE ROBUST STATISTICS PERSPECTIVE REVISITED[1]

1. INTRODUCTION

Considerations related to income distribution and income inequalities in populations of economic agents are at the core of modern economic theory. They appear also in public debates on taxation or pension politics, in theories of human capital creation or searching for regional development factors. Correctly estimating the parameters of income distribution and derivative measures of income inequality such as the Gini coefficient or Theil Index are important for several reasons – it is source of knowledge about income structure in society and also could be the basis for further economic issues such changing the taxation system or launching government aid programmes in order to redistribute some part of wealth. Underestimating the parameters of income distribution could lead to the conclusion that inequalities are too high and trigger corrective actions such as rising taxes in high income groups. If there is too much severity in changing the tax bracket, it may have influence on productivity and investment activities among well-paid citizens. Overestimating the parameters could have an opposite but also harmful effect on the health of the economy because an overly liberal taxation system would likely cause low-paid people to get insufficient public assistance. Moreover, income distribution effects economic growth, market demand and is an important factor in determining the amount of savings in a society (Kleiber, Kotz 2003).

In real economic data sets, it often happens that some observations are different to the majority. These outlying observations cause problems because they may strongly influence the results of an economic analysis. Robust

[*] Cracow University of Economics, Faculty of Management, Department of Statistics.

[1] The Author would like to thank NCS Poland for financial support DEC-011/03/B/HS4/01138 and CUE "*Badania statutowe 2015*" support.

statistics attempts to detect outliers by searching for a model that fits a majority of the data. All classical statistical methods (e.g.: discriminant analysis, factor analysis, regression analysis, estimation of time series models parameters) can be severely distorted by outliers. It should be stressed that statistical inferences (an important part of each economic analysis) are based only in part upon observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplest cases, there are explicit or implicit assumptions about randomness and independence, distributional models, possibly prior distributions for some unknown parameters, etc.

This paper deals selected aspects of robust estimation of the income distribution. Attention shall be focused on two well-known models for income distribution, namely the Pareto and log-normal distributions, as well as on popular income inequality measures, namely on the Lorentz curve and the Gini coefficient. The presented arguments, however, are applicable to a wide class of over 100 models used for income distributions modelling, which are by default estimated using the maximal likelihood methodology.

The rest of the paper is organised as follows. In Section 2, the selected income distribution models are presented. In Section 3, the selected robust estimators of income distribution are briefly presented. In Section 4, popular income distribution inequality measures are recalled. In Section 5, the results of simulation as well as empirical studies of statistical properties of the considered estimators are presented. The paper ends with conclusions and references.

2. SELECTED INCOME DISTRIBUTION MODELS

Modern concern about income distribution began with Pareto's research during his discussions with French and Italian socialists, who were insisting on institutional reforms to reduce inequality in income distribution. Pareto studied the income distribution of economic agents for tax purposes. The distribution was truncated to the left at the point x_m, the maximum non-taxable income, $x_m > 0$. He found a regularity of the observed income distribution obtained from tax records – a stable linear relation of the form $\log N(x)=A-\alpha \log x$, $x \geq x_m > 0$, $\alpha > 1$, where $N(x)$ is the number of economic units with income $X > x$ and X being the income variable with the range $[x_m, \infty)$. The Pareto type I model is the solution of that linear relationship. In the same context, in 1898, March proposed the gamma probability density function (PDF) and fitted it to the distribution of wages in France, Germany and the United States. Today, there are over 100 models used for income distribution modelling (see: Kleiber, Kotz 2002).

The Pareto distribution for modelling high-income groups and dealing with positive asymmetric distributions that have heavy weight tails with either finite or infinite variance still stands at the centre of income distribution considerations.

This is mainly due to its elegance, facility of interpretation and its relation to the popular income distribution inequality measures. Along with others, Pareto distribution skewed size distributions also appear in the context of economic data stream analysis, e.g.: for modelling data packages sizes on the Internet (see: Kosiorowski 2012).

For purposes of this paper, it is enough to consider a broad classification of income distribution according to the tail behaviour: Pareto type distributions (polynomially decreasing tails), log-normal distribution (intermediate case) and gamma-type distribution (exponentially decreasing tails).We shall focus our attention on two estimation difficulties which are good illustrations for the robust analysis of income distribution.

We shall start with the Pareto model *P* (x_m, α), which is suitable to model relatively high probability in the upper tail (right-skewed tail) where a lower α shape parameter determines the lower probability mass at x_m point. Thanks to this property, the model is useful and relatively effective to apply in actuarial applications, risk management and Economy of Welfare.

A simple Pareto distribution *P* (x_m, α) is given by its cumulative distribution function (*CDF*):

$$F(x) = 1 - \left(\frac{x_m}{x} \right)^{\alpha}, \tag{1}$$

for $x > x_m$, where α is the shape parameter that characterises the tail of the distribution and $x_m > 0$ is the scale parameter.

The Pareto distribution has a PDF of $\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$ for $x > x_m$ and the following formulas for the expected value:

$$E(X) = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha - 1} & \alpha > 1 \end{cases},$$

and variance:

$$D^2(X) = \begin{cases} \infty & \alpha \in (1,2] \\ \dfrac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \alpha > 2 \end{cases},$$

with the median $x_m \sqrt[\alpha]{2}$ and mode x_m.

If the sample observations follow the postulated model $P(x_m, \alpha)$, then it is well known that for large data sets, the maximum likelihood estimator (*MLE*) shall attain the minimum possible variance among a large class of competing estimators:

$$\hat{\alpha}_{ML} = \frac{n}{\sum_{i=1}^{n} \log(X_i / x_m)}. \tag{2}$$

It can be easily found that $2n\alpha / \hat{\alpha}_{ML}$ has a CDF of χ^2_{2n} (see: Brazauskas, Serfling 2000). Although $\hat{\alpha}_{ML}$ is biased, it is easy to find its unbiased version (*MLE*):

$$MLE = \frac{n-1}{\sum_{i=1}^{n} \log(X_i / x_m)}. \tag{3}$$

For large sample size n, *MLE* is approximately $N(\alpha, \frac{\alpha^2}{n})$ distributed. In case of the scale estimator, we have following maximal likelihood formula:

$$MLE(x_m) = \min_i \{X_i\}. \tag{4}$$

The Pareto distribution is widely used in economics due to its elegance and clear relations with the popular measure of income inequality known as the Gini coefficient $GINI = 1/(2\alpha-1)$ for $\alpha \geq 1$ or popular risk measures such as value at risk. It should be stressed, however, that even small relative errors in the estimation of α in $P(x_m, \alpha)$ may lead to a large relative error in the estimated quantiles or tail probabilities based on α. For the quantile q_ε corresponding to the upper tail probability ε, it follows that $q_\varepsilon = x_m \varepsilon^{-1/\alpha}$ For $\varepsilon = 0.001$, the underestimation of $\alpha = 1$ by only 5% leads to an overestimation of $q_{0.001}$ by 58%. Errors in the estimation of α may result in errors in the estimation of basic measures of social inequity and lead to incorrect social politics.

Next, an important distribution for modelling incomes is the log-normal distribution, which was developed for economic purposes by Gibrat (1931). The random variable *Y* has the log-normal distribution *L* (μ, σ) if $X = \log Y$ has the normal distribution *N* (μ, σ^2).

Three parameter form *L* (μ, σ, τ) is the distribution of $Y = \tau + e^X$, where τ represents a threshold value and *X* is a random variable with mean μ and standard deviation σ.

In many applications, a problem of efficient and robust estimation of the expected value of this distribution $E(Y) = e^{\mu+\sigma^2/2}$ appears (we assume the threshold τ is known). The problem leads to a nontrivial issue of the joint robust estimation of μ and σ in the context of the corresponding model *N* (μ, σ). For the sample $Y^n = \{Y_1, \ldots, Y_n\}$ from the model *L* (μ, σ), a transformation to the equivalent model *N* (μ, σ) yields the well-known *ML* estimators of the location μ and σ scale parameter:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \log Y_i \,, \tag{5}$$

$$\hat{\sigma}_{ML} = \left(\frac{1}{n}\sum_{i=1}^{n} (logY_i - \hat{\mu}_{ML})^2\right)^{1/2} , \tag{6}$$

and the estimator of the expected value:

$$E(Y) = e^{\hat{\mu}_{ML} + \hat{\sigma}^2_{ML}/2} \,. \tag{7}$$

Estimators 5, 6 and 7 have good statistical properties, i.e.: minimal asymptotic variances, but they fail to be robust, i.e. their breakdown point (*BP*) equals 0 and their influence function (*IF*) is unbounded.

As a last landmark distribution for incomes modelling, consider the generalised gamma distribution with *PDF*:

$$f(x) = \frac{a}{\beta^{ap}\Gamma(p)} x^{ap-1} e^{-(x/\beta)^a} , \tag{8}$$

where $x > 0$, $\beta = b^{1/a}$ scale parameters, *a*, *p* shape parameters.

Model 8 is usually estimated *via* maximal likelihood methodology, which leads to estimators which are not robust.

Each of the above distributions, particularly their parameters, have interesting economic interpretations expressed in terms of their elasticity of survival function, maximisation of entropy, probability of increasing an agent's income under some conditions, etc. Discerning between these three landmark distributions in cases where inliers or outliers within data are present using classical model selection methods may be a very difficult task. Empirical justifications of theoretical concepts explaining the form of income distribution may be shaky. Let us take, for instance, Mandelbrot (1960) who argued that incomes follow what he called a Pareto-Levy distribution – a maximally skewed stable distribution with a characteristic exponent α between 1 and 2.

3. ROBUST ESTIMATORS OF THE INCOME DISTRIBUTIONS

Kalecki (1945) found that income increments are proportional to excess in ability of the given members of the distribution over the lowest (or median) member. He considered the log-normal distribution for personal incomes in the United Kingdom for 1938-1939 and found that the log-normal distribution fits well only when certain parts of the data are omitted. He introduced, therefore, three parameter log-normal distributions. Kalecki can be treated as a pioneer of the robust approach to income distribution analysis.

Robust estimation of the bounded influence function of income distribution parameters was extensively studied by Victoria-Fezer (2000) based on the M-estimation approach (see: Marona et al. 2006). We shall focus our attention on a less known, but very interesting, approach related to Brazauskas and Serfling's studies.

We understand robustness of the estimator in terms of the influence function (*IF*) and in terms of the finite sample breakdown point (*BP*) – for further details see: Maronna et al. (2006).

Let us recall that for a given distribution F in $\Re$ and an $\varepsilon > 0$, the version of F contaminated by an ε amount of an arbitrary distribution G in $\mathbb{R}$ is denoted by $F(\varepsilon, G)=(1-\varepsilon)\,F+\varepsilon\, G$. The influence function (*IF*) of the estimator T at a given $x \in \Re$ for a given F is defined:

$$IF(x;T,F) = \lim_{\varepsilon \to 0^+} \left(T(F(\varepsilon,\delta_x)) - T(F)\right)\big/\varepsilon\,, \tag{9}$$

where δ_x is the point-mass probability measure at the point $x \in \Re$.

IF(x, T, F) describes the relative effect (influence) on T of an infinitesimal point-mass contamination at point x and measures the local robustness of T.

An estimator with the bounded *IF* (with respect to a given norm) is, therefore, robust (locally, as well as globally) and very desirable.

Let $X^n = \{X_1, \ldots, X_n\}$ be a sample of size n from X in $\Re$. The replacement breakdown point (*BP*) of an estimator T for the sample X^n is defined as:

$$BP(T, X^n) = \left\{ \frac{m}{n} : \left\| T(X_m^n) - T(X^n) \right\| > \delta \right\}, \tag{10}$$

where: X_m^n is a contaminated sample resulting from replacing m points of X^n with arbitrary values, $\| \ \|$ denotes a norm, δ is certain content-related threshold, i.e.: for the Gini coefficient we can take $\delta = 0.3$ if that value is faced with different social politics based on the Gini coefficient.

The *BP* point serves as a measure of global robustness, while the *IF* function captures the local robustness of estimators. In the context of the simple Pareto, log-normal or gamma distribution estimations, it is useful to discriminate between sample contamination with lower values (*LBP*) and sample contamination with upper values (*UBP*).

It is beyond the scope of this paper to introduce the reader into the formal details of robust statistics. An excellent introduction into the matter could be found for example in Huber and Ronchetti (2009) or Marona et al. (2006). For our purposes it is enough to intuitively understand the following simple example. Suppose we have five measurements of five monthly salaries (in PLN) in Poland from 2011: 3 225 PLN; 3 103 PLN; 2 944 PLN; 3 100 PLN; 1 123 PLN. Our aim is to estimate the true value of the "*centre salary*" in Poland in 2011. Calculating the mean, we obtain 2 699 PLN but when calculating the median we get 3 100 PLN. The median is the middle value and, in contrast to the mean, is not affected by outlying salary of 1 123 PLN. We can say that the median is more robust against the outlier than the mean. Similarly, calculating a typical measure of dispersion, the standard deviation (*SD*), we get 886.63, but calculating robust measure of dispersion – the median of absolute deviations from the median (*MAD*) – we get 185.23. We can say that the *MAD* shows the differences in salaries in a robust manner in contrast to the *SD*. The mean and *SD* have unbounded influence functions and their *BP* are equal to zero. The median and the *MAD* have bounded *IF* and maximal *BP* values.

3.1 Robust estimators of Pareto and log-normal distribution

Let us recall that for specified β_1 and β_2 satisfying $0 \leq \beta_1,\ \beta_2 < ½$, the trimmed mean is formed by discarding the population of lowest observations β_1 and the proportion of uppermost observations β_2 and averaging the remaining ones in some sense. In particular, for estimating α with a known x_m.

Brasauskas and Serfling (2000) proposed the trimmed mean estimator:

$$\hat{\alpha}_{TM} = \left(\sum_{i=1}^{n} c_{ni} \log\left(X_{(i)} / x_m \right) \right)^{-1}, \tag{11}$$

with $c_{ni} = 0$ for $1 \leq i \leq$, $c_{ni} = 0$ for $n - [n\beta_2] + 1 \leq i \leq n$ and $c_{ni} = 1/d\ (\beta_1,\ \beta_2,\ n)$ for $[n\beta_1] + 1 \leq i \leq n - [n\beta_1]$, where: [] denotes the “*greatest integer part*” and $d\ (b_1, b_2,\ n) = \sum_{j=[n\beta_1]+1}^{n-[n\beta_2]} \sum_{i=0}^{j-1} (n-i)^{-1}$.

The next robust estimator appeals to the idea of the generalised median (*GM*) statistic**.** The *GM* statistics are defined by taking median of the $\binom{n}{k}$ evaluations of a given kernel $h(x_1, \ldots, x_k)$ over all k-element subsets of the data. Brazauskas and Serfling (2002) proposed the following estimator for the parameter α in Pareto model in case of a known x_m:

$$\hat{\alpha}_{GM} = MED\left\{ h\left(X_{i1}, ..., X_{ik} \right) \right\}, \tag{12}$$

with a particular kernel $h(x_1,\ldots,x_k)$:

$$h(x_1, ..., x_k; x_m) = \frac{1}{C_k} \frac{k}{\sum_{j=1}^{k} \log\left(x_j / x_m \right)}, \tag{13}$$

where: C_k is a multiplicative, the median an unibasing factor, i.e.: chosen so that the distribution of $h(x_1, \ldots, x_k; x_m)$ has a median α – and the values of C_k for $k = 2$, $C_2 = 1.1916$, $k = 3$, $C_3 = 1.1219$.

For the log-normal distribution $L\ (\mu, \sigma)$, Serfling (2004) introduced *GM* estimators and obtained their properties. A kernel for the *GM* location estimator takes the form:

$$h_1(x_1,...,x_k) = \frac{1}{k}\sum_{i=1}^{k}\log x_i \ , \tag{14}$$

$$\hat{\mu}_{GM}(k) = \text{median}\left\{h_1\left(X_1,...,X_k\right)\right\}. \tag{15}$$

This estimator has a $BP(\hat{\mu}_{GM}(k)) = 1-(1/2)^{1/k}$ and smooth and bounded *IF*.

In the case of a scale estimator, Serfling (2004) proposes using the following kernel:

$$h_2(x_1,...,x_m) = \frac{1}{mM_{m-1}}\sum_{1\le i<j\le m}(x_i - x_j)^2 \ , \tag{16}$$

which leads to the following robust estimator of scale in the log-normal model:

$$\hat{\sigma}^2_{GM}(m) = \text{median}\left\{h_2\left(X_1,...,X_m\right)\right\}. \tag{17}$$

4. MEASURES OF INCOME INEQUALITY

Measuring income inequality within a population of economic agents is very closely related to estimating the probability of income distribution. Incorrect estimates of the distribution may lead to incorrect evaluations of inequalities and incorrect social politics. It should be stressed that we can evaluate the degree of income inequality assuming a certain model (e.g.: the Pareto model), estimate it and then use known relations between the parameters of this model and an inequality measure for evaluating of the degree of inequality in a population. From another point of view, it is possible to estimate a nonparametric degree of inequality – i.e.: without assuming the probability distribution generating the data. The first method is commonly said to be more elegant and easier for economic interpretations. The second method, however, is generally "*closer to the reality*" of the observed data.

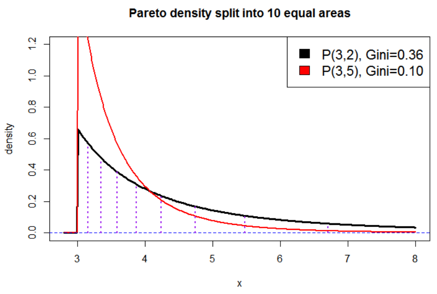

Figure 1. Pareto densities and corresponding Gini inequality coefficients

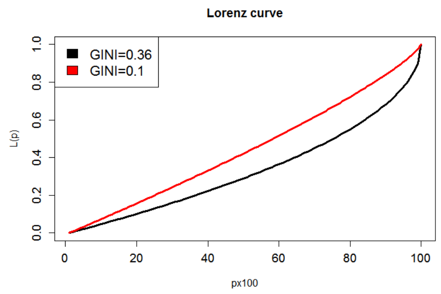

Figure 2. Lorenz curves for Pareto densities and corresponding Gini coefficients

Source: own elaborations.

Although there are at least twenty popular measures of income inequalities used, the benchmark measure is the Lorentz curve, a graphical representation of the *CDF* of the empirical probability of wealth. For the discrete probability function $f(y)$, let y_i, $i = 1, \ldots, n$ be points with non-zero probabilities indexed in increasing order $y_i < y_{i+1}$. The Lorentz curve is the continuous piecewise linear function connecting the points (F_i, L_i), $i = 1, \ldots, n$, where $F_0 = 0$, $L_0 = 0$, and $F_i = \sum_{j=1}^{i} f(x_i)$, $S_i = \sum_{j=1}^{i} f(x_j)x_j$, $L_i = S_i / S_n$. For the *PDF* function $f(x)$ with the *CDF* $F(x)$, the Lorentz curve $L(F(x))$ is given by:

$$L(F(x)) = \frac{\int_{-\infty}^{x} tf(t)dt}{\int_{-\infty}^{\infty} tf(t)dt} = \frac{\int_{-\infty}^{x} tf(t)dt}{\mu}, \tag{18}$$

with *CDF* F and expected value μ. The next popular measure of income inequality is the Gini coefficient, which is half the relative mean difference and usually defined based on the Lorentz Curve. For the random nonzero variable X with *CDF* F and expected value μ, the Gini coefficient is defined as:

$$G = 1 - \frac{1}{\mu}\int_{0}^{\infty}\left(1 - F(x)\right)^{2} dx = \int_{0}^{\infty} F(x)\left(1 - F(x)\right) dx. \tag{19}$$

The mean difference is defined as the expected value of the absolute difference of two random variables X and Y independently and identically distributed with the same unknown distribution $MD = E[|\,X - Y\,|]$. For the sample $X^n = \{x_1, \ldots, x_n\}$ it means:

$$MD = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| \tag{20}$$

and the relative mean difference is defined as:

$$RMD = \frac{MD}{\bar{x}} = 2 \cdot \mathrm{GINI}\,. \tag{21}$$

Other popular measures involve the Pietra coefficient, variance of logarithms, Zenga curve, Atchison generalised entropy measure.

Looking at models 18, 19, 20 and 21, it is easy to notice that robustness of the sample Lorentz curve is related to the robustness of the sample mean and robustness of the probability density estimator. The Gini coefficient may be calculated on several ways, which may give different results in case of the existence of outliers or inliers within the data. The popular method of "*robustifying*" an estimator involving, for example, trimming the data is applicable for model 20. We should notice, however, that the Gini coefficient takes a value from a bounded interval and its breakdown should be understood in the spirit of a certain decision process based on the Gini estimates. The theory for inequality measures may be obtained within the theory of empirical processes, where the Gini coefficient is treated as a function of the empirical Lorenz process or within the theory of sample quantiles so the theory for their robustness may be obtained at the same time.

Let us only briefly recall that the Lorenz curve may be generalised to a multivariate case within a data depth concept. The generalisation was proposed by Mosler (2013). The data depth concept was originally introduced as a way to generalise the concepts of median and quantiles to the multivariate framework. The depth function $D(\mathbf{x}, F)$ associates with any $\mathbf{x} \in \mathbb{R}^d$; the measure $D(\mathbf{x}, F) \in [0,1]$ with its centrality with regard to the probability measure $F \in \mathcal{P}$ over $\mathbb{R}^d$ or with regard to the empirical measure $F_n \in \mathcal{P}$ calculated from the sample $\mathbf{X}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is with regard to F or F_n. As an example of depth, let us recall the weighted L^p depth from the sample $\mathbf{X}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and is computed as follows:

$$L^p D(\mathbf{x}, \mathbf{X}^n) = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} w\left(\left\|\mathbf{x} - \mathbf{X}_i\right\|_p\right)}, \tag{22}$$

where w is suitable, non-decreasing and continuous on the weight function $[0, \infty)$, and $\| \|_p$ stands for the L^p norm (when $p = 2$ we have the usual Euclidean norm and so-called spatial depth).

The set of points for which depth takes a value not smaller than $\alpha \in [0, 1]$ is a multivariate analogue of the quantile and is called the α – central region:

$$D_\alpha(\mathbf{X}) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, \mathbf{X}) \geq \alpha\}. \tag{23}$$

The multivariate Lorentz curve is defined as the proportion of the mean confined to the central region $D_\alpha(\mathbf{X})$ to the overall mean. Let $f(x)$ denote the wealth of a point $\mathbf{x} = (x_1, \ldots, x_n)$, i.e.: the coordinates of points may represent amounts of d goods at an agent's disposal. We can define the multivatiate Lorenz Curve as:

$$L(\alpha) = \alpha \times \frac{E\left(f(\mathbf{x}) \mid \mathbf{x} \in D_\alpha(\mathbf{X})\right)}{E(f(\mathbf{x}))}. \tag{24}$$

Please note that the parameter $\alpha \in (0, 1)$ expresses the outlyingness of a point with regard to centre, i.e. a multivariate median induced by a depth function. It is, however, possible to use depth regions consisting of a probability mass not smaller than $\alpha \in (0, 1)$ and hence order them by probability.

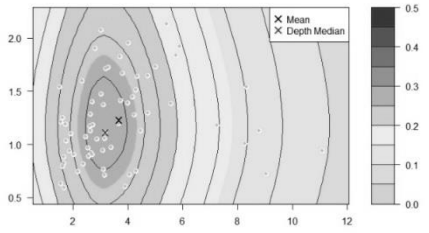

Figure 3. Contour plot for sample L^2 depth

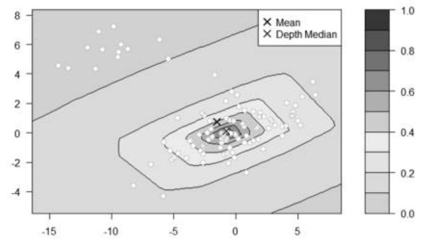

Figure 4. Contour plot for sample projection

Source: DepthProc R package.

Figure 3 presents a contour plot for the L^2 sample depth and Figure 4 presents a contour plot for projection sample depth. It is easy to notice that model 24 shows an allocation of wealth with respect to a departure from the central object (a multivariate median) – which for several socio-economic reasons may be more interesting than the relation of the object to group of very rich or very poor objects.

5. PROPERTIES OF THE ROBUST ESTIMATORS OF INCOME DISTRIBUTION

In order to critically study the performance of known robust estimators of income distributions and income inequalities, we conducted intensive simulation as well as empirical studies. Only a small part of the results are presented below.[2] In the context of the Pareto model estimation, we considered *MLE*, *TM* and *GM* estimators, which were compared with Victoria-Faser bounded *IF* proposals as well as with constrained local polynomial estimator proposed by Hyndman and Yao (2002). We performed a similar analysis for the log-normal distribution estimators, Dagum distribution estimators and the generalised gamma distribution.

In the case of the Pareto distribution, we performed intensive simulation studies involving simulated datasets with 500 observations from the following mixtures of distributions:

1. Mixture of P (1, 5)×10% and P (10, 5)×90%.
2. Mixture of lognormal distribution LN (2.14, 1)×10% and P (7, 2)×90%.
3. Mixture of normal distribution N (3 300, 500)×10% and P (2 500, 4)×90%
4. Mixture of uniform U [0, 0.1]×10% distribution and P (2 500, 4)×90% distribution.

Figures 5–8 present the estimated log densities for the mixtures, with x_m taken as minimum. It is easy to notice that the estimator of x_m has a crucial issue for the performance of the estimators. With the classical *MLE* estimator for x_m, all estimators of the parameter shape perform relatively poorly.

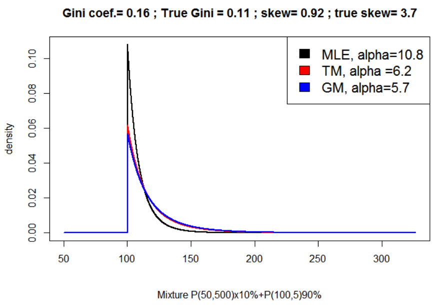

Figure 5. Estimated densities for the first mixture and x_m taken as the 12% quantile

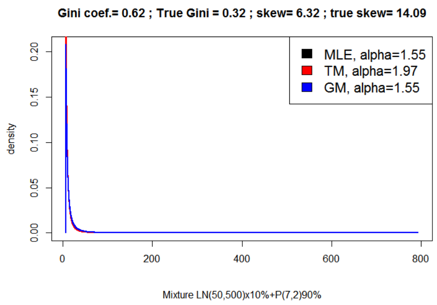

Figure 6. Estimated densities for the second mixture and x_m taken as the 12% quantile

Source: own elaborations.

[2] The rest of the results and R codes for calculating the robust estimators are available on request.

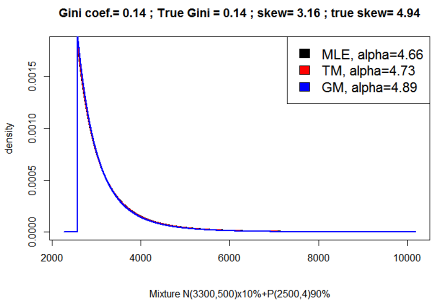

Figure 7. Estimated densities for the third mixture and x_m taken as the 12% quantile

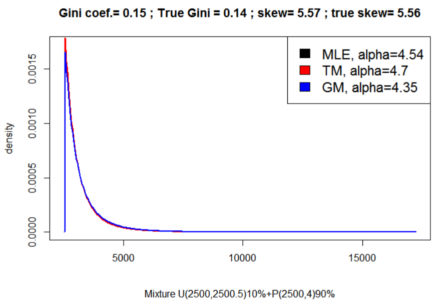

Figure 8. Estimated densities for the fourth mixture and x_m taken as the 12% quantile

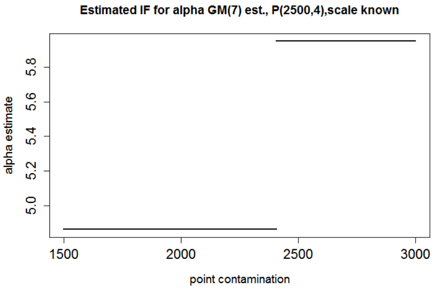

Figure 9. Estimated *IF* for the *MLE* estimator and stylised sample of 100 obs.

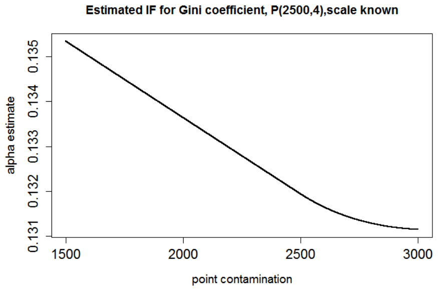

Figure 10. Estimated IF for the MLE estimator and stylised sample of 100 obs.

Source: own elaborations.

Figure 9 presents the stylised empirical influence function for the *GM* estimator in the case of subsamples consisting of 7 points, the Pareto *P* (2 500,4) model and scale estimator taken as quantile of order 0.12. In this case, the *GM* estimator can be treated as robust. Figure 10 presents the stylised empirical influence function for the Gini coefficient. It is easy to notice that this measure of inequality is not robust. The results of the simulation led to similar conclusions, which are also similar for other well-known income distribution models, estimators and popular inequality measures. The conclusions may be summarised as follows:

1. The GM estimators with scale (threshold in three parameter log-normal model) estimated as quantile of order $\beta \in (0, 0.3)$, where β is optimised using the Kolmogorov-Smirnov goodness of fit statistics outperforms the classical *MLE* and *TM* estimators. The estimators are computationally intensive, however. We recommend using the *GM* estimator for estimating scale.
2. Estimating the income distribution nonparametrically is worth considering – we recommend the constrained local polynomial estimator proposed by Hyndman and Yao (2002), which also provides estimates of the density derivatives, at least on the explanatory step of the research.

3. We recommend calculating the Gini coefficient "*nonparametrically*", i.e.: without using an assumption of the Pareto, log-normal, gamma distributed data. For popular scalar measures of inequality involving the Gini coefficient or Pierta coefficient, it is possible to apply the generalised median approach (see: Kosiorowski, Tracz 2014b).

For evaluating the considered robust estimators in the case of real data, we focused our attention on the data considered in Kosiorowski et al. (2014) – census data from MINNESOTA POPULATION CENTER[3]. We considered data on TOTAL INCOME from the following countries:

- **Panama**: 1960, 1970, 1980, 1990, 2000, 2010;
- **Mexico**: 1960, 1970, 1990, 1995, 2000, 2005, 2010;
- **Puerto Rico**: 1970, 1980, 1990, 2000, 2005;
- **Canada**: 1971, 1981, 1991, 2001;
- **Brazil**: 1960, 1970, 1980, 1991, 2000, 2010;
- **USA**: 1960, 1970, 1980, 1990, 2000, 2005, 2010.

Each time, we estimated the density using *GM*, *TM* and *M*-type estimators (parametrically) after selecting the models using the information criterion and value of the Kolmogorov goodness of fit statistic. Figures 11–16 present densities obtained using the constrained local polynomial method, which in our opinion is the best counterpart to both classical and robust estimators. The empirical data showed us a rich set of difficulties related to the robust model selection issue. These difficulties are automatically omitted in the case of the considered nonparametric method application. It is worth noticing that a kernel used within this method locally protects us against outliers. Using the *k*-nearest neighbours' type kernel protects us against inliers as well. In each case, the density was estimated using a local linear polynomial estimator in an equally spaced grid of 500 points.

Figures 15–16 presents the estimated results for the data divided by median incomes. The nonparametric estimator better underlies the heterogeneity of the incomes and should be considered at least as a preliminary research step.

[3] https://international.ipums.org/international

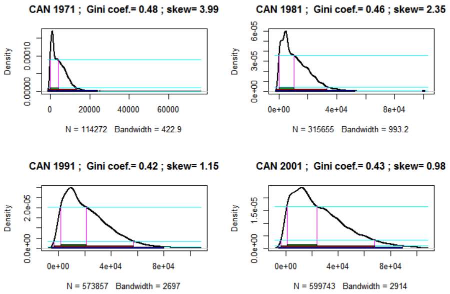

Figure 11. Estimated income densities in Canada 1971, 1981, 1991, 2001

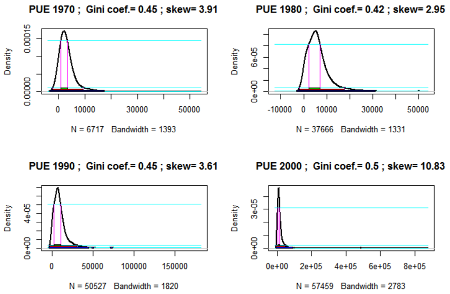

Figure 12. Estimated income densities in Puerto Rico 1970, 1980, 1990, 200.

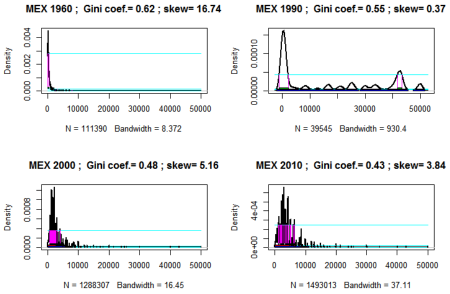

Figure 13. Estimated income densities in Mexico 1960, 1990, 2000, 2010

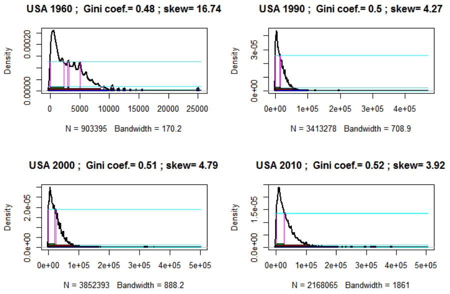

Figure 14. Estimated income densities in USA 1960, 1990, 2000, 2010

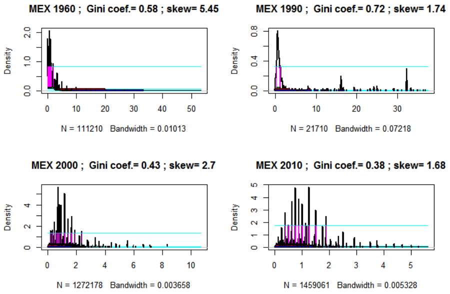

Figure 15. Estimated income/median (income) densities in Mexico 1960, 1990, 2000, 2010

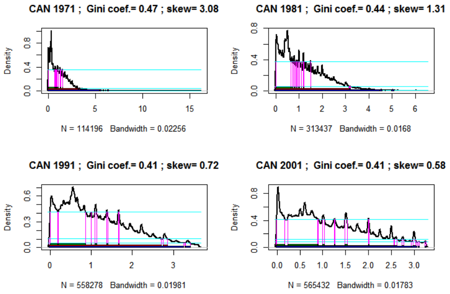

Figure 16. Estimated income/median (income) densities in Canada 1971, 1981, 1991, 2001

Source: own elaborations.

6. CONCLUSIONS

Considerations related to a nature of allocation of wealth within a population have a central position in the economic and public debate related to social justice and social solidarity. Arguments used within these debates strongly depend on the properties of the statistical procedures used for estimating income distributions and income distribution measures. Classical maximal likelihood estimators of the income distribution parameters are not robust to outliers or inliers in the data. There are good robust and/or nonparametric alternatives for them, however. We recommend using the generalised median approach proposed by Brazauskas and Serfling in the case of the existence of some knowledge on the considered phenomena and the constrained local polynomial estimator in case of a lack of knowledge on the subject of study.

REFERENCES

Brazauskas V., Serfling R. (2000), *Robust and Efficient Estimation of the Tail Index of a Single-Parameter Pareto Distribution*, “North American Actuarial Journal”, 4, pp. 12-27.

Brazauskas V., Serfling R. (2001), *Robust Estimation of Tail Parameters for Two-Parameter Pareto and Exponential Models via Generalized Quantile Statistics*, “Extremes”, 3, pp. 231-249

Brazauskas V., Serfling R. (2004), *Favorable Estimators for Fitting Pareto Models: A Study Using Goodness-of-Fit Measures with Actual Data*, ASTIN Bulletin, 2, pp. 365-381.

Dagum C. (2001), *A systemic approach to the generation of income distribution models*, (in:) Sattinger M. (ed.), *Income Distribution*, vol. I, E. Elgar, Northampton, pp. 32-53.

Hyndman J. R., Yao Q. (2002), *Nonparametric estimation and symmetry tests for conditional density functions*, “Journal of Nonparametric Statistics”, 14 (3), pp. 259 278.

Kalecki M. (1945), *On the Gibrat distribution*, “Econometrica”, 13, pp. 161-170.

Kleiber C., Kotz S. (2002), *A characterization of income distributions in terms of generalized Gini coefficients*, “Social Choice and Welfare”, 19, pp. 789-794.

Kleiber C., Kotz S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, New Jersey.

Kosiorowski D., Zawadzki Z. (2014). *DepthProc: An R package for robust exploration of multidimensional economic phenomena*. Submitted.

Kosiorowski D., Tracz, D. (2014a), *On robust estimation of Pareto models and its consequences for government aid programs evaluation*, (in:) Lula P., Rojek T. (eds.), *Knowledge-Economy-Society Contemporary Tools of Organizational Management*, pp. 253-267.

Kosiorowski D., Mielczarek D., Rydlewski J., Snarska M. (2014), *Applications of the functional data analysis for extracting meaningful information from families of yield curves and income distribution densities*, (in:) Lula P., Rojek T. (eds.), *Knowledge-Economy-Society Contemporary Tools of Organizational Management*, pp. 309-321.

Maronna R. A., Martin R. D., Yohai V. J. (2006), *Robust Statistics – Theory and Methods*, Wiley, Chichester.

Mosler K. (2013), *Depth statistics*, (in:) Becker C., Fried R. S. K. (eds.), *Robustness and Complex Data Structures*, Festschrift in Honour of Ursula Gather. Springer, pp. 17-34.

Pawlak W., Sztaudynger J. J. (2008), *Wzrost gospodarczy a optymalne zróżnicowanie dochodów w USA i Szwecji*, "Annales – Etyka w życiu gospodarczym", 1, pp. 259-271

Serfling R. (2002), *Efficient and Robust Fitting of Lognormal Distributions*.

Victoria-Feser M. P. (2000), *Robust Methods for the Analysis of Income Distribution, Inequality and Poverty*, "International Statistical Review", 68, pp. 277-293.

ABSTRACT

Considerations related to income distribution and income inequalities in populations of economic agents belong to the core of the modern economic theory. They appear also in a public debate concerning postulates as to taxation or pension politics, in theories of a human capital creation or searching for regional development factors.

Results of statistical inference conducted for giving arguments pro or against particular hypotheses, strongly depend on properties of statistical procedures used within this process. We mean here for example: a quality of probability density estimator in case of missing data, a quality of skewness measure in multivariate case departing from normality, or a quality of dimension reduction algorithm in case of existence of outliers.

In this paper from the robust statistics point of view, we analyse difficulties related to statistical inference on income distribution models and income inequalities measures. Theoretical considerations are illustrated using real data obtained from Eurostat and Minessota Population Center (IMPUS).

WYBRANE ZAGADNIENIA MODELOWANIA ROZKŁADU DOCHODU ORAZ POMIARU NIERÓWNOŚCI DOCHODOWYCH ROZPATRYWANE Z PUNKTU WIDZENIA STATYSTYKI ODPORNEJ

ABSTRAKT

Rozważania dotyczące rozkładów dochodów oraz nierówności dochodowych bez wątpienia należą o tzw. jądra ekonomii teoretycznej. Rozważania tego typu pojawiają się w debacie publicznej dotyczącej polityki podatkowej, polityki transferów społecznych, w teoriach tworzenia kapitału intelektualnego bądź w typowaniu czynników rozwoju regionalnego.

Warto zauważyć, że wyniki badań statystycznych prowadzonych, aby dostarczyć argumentów za bądź przeciw hipotezom stawianym w debatach ekonomistów zależą krytycznie od własności metod statystycznych wykorzystywanych w tych badaniach.

Mamy tutaj przykładowo na uwadze, jakość estymatora gęstości w przypadku brakujących danych, jakość wielowymiarowej miary skośności w przypadku odstępstwa od normalności populacji, bądź jakość algorytmu zmniejszającego wymiar zagadnienia statystycznego w przypadku występowania obserwacji odstających.

W sytuacji, gdy w badaniach tego typu uwzględniamy dodatkowo pewien wymiar przestrzenny bądź społecznoekonomiczny – przeprowadzenie dobrej jakości wnioskowania statystycznego wydaje się stanowić szczególnym wyzwanie.

W niniejszej pracy w krytyczny sposób analizujemy trudności związane z wnioskowaniem statystycznym dotyczącym wybranych modeli dochodu i wybranych miar nierówności dochodowych.

Z perspektywy statystyki odpornej badamy m.in. powszechnie wykorzystywane estymatory parametrów modeli Pareto, Pearsona, D'Addario oraz Daguma. Proponujemy odporne i nieparametryczne alternatywy dla popularnych miar nierówności dochodowych oraz pokazujemy jak zredukować liczbę predyktorów dla agregatów dochodowych w odporny sposób. Zwracamy szczególną uwagę na przestrzenny wymiar naszych badań.

Rozważania teoretyczne ilustrujemy m.in. wykorzystując dane empiryczne pochodzące z Eurostatu i Minnesota Population Center (IMPUS).