Robust estimation of economic indicators from survey samples based on Pareto tail modeling

Andreas Alfons
KU Leuven, Leuven, Belgium
Vienna University of Technology, Vienna, Austria

Matthias Templ
Vienna University of Technology, Vienna, Austria
Statistics Austria, Vienna, Austria

Peter Filzmoser
Vienna University of Technology, Vienna, Austria

Summary. Motivated by a practical application, this paper investigates robust estimation of economic indicators from survey samples based on a semiparametric Pareto tail model. Economic performance is typically measured by a set of indicators, which are often estimated from survey data – the motivating example being the European indicators on social exclusion and poverty computed from the well known EU-SILC survey. Since economic data typically contain variables with heavily tailed distributions and additional extreme outliers, the idea is to use robust Pareto tail modeling to detect the extreme outliers and reduce their influence on the indicators. In the survey context, however, sample weights need to be considered when modeling the tail with a Pareto distribution such that the true distribution on the population level is accurately reflected. Therefore, the main methodological contribution is to adapt commonly used robust estimators for the parameters of the Pareto distribution to take sample weights into account. The resulting approach for robust estimation of indicators is then evaluated by means of a simulation study and applied in the context of estimating the Gini coefficient from EU-SILC data.

Keywords: EU-SILC; Gini coefficient; Pareto distribution; Survey statistics; Tail modeling

1. Introduction

Economic indicators monitor the economic performance of administrative units such as countries or regions for analysis and prediction purposes. In many cases, economic indicators are estimated from survey data due to a lack of availability of suitable population data. The motivating example for this paper is the *European Union Statistics on Income and Living Conditions* (EU-SILC), which is an annual panel survey conducted in European Union member states and other European countries. This survey is used as data basis for a set of indicators to measure risk-of-poverty and social exclusion in Europe. However, many of these economic indicators are highly sensitive to outlying observations, in particular the *Gini coefficient* (Gini, 1912) and other indicators of inequality.

E-mail: andreas.alfons@kuleuven.be
E-mail: templ@tuwien.ac.at
E-mail: p.filzmoser@tuwien.ac.at

In economic data, the distributions of variables such as income or sales turnover usually have heavy tails. In addition, even more extreme outliers deviating from the rest of the tail are a common problem. Heavy tails are frequently modeled by a Pareto distribution (e.g. Kleiber and Kotz, 2003), while robust estimation allows to identify extreme outliers. So far many estimators for the parameters of a Pareto distribution have been proposed. However, those procedures are usually designed for samples from infinite populations, survey samples are typically not considered in the literature on the Pareto model. Finite population survey sampling is in general based on complex sampling designs with unequal inclusion probabilities for the observations in the population, which leads to unequal weights for the observations in the sample (see, e.g., Tillé, 2006). The initial weights are also often further modified by techniques such as calibration (e.g., Deville et al., 1993) so that the sample weights of the observations in certain subsets sum up to known population totals. The idea behind Pareto tail modeling for survey data is that the upper tail of the population data follows a Pareto distribution. Hence sample weights need to be considered for fitting the distribution in order to avoid bias in the estimation of the parameters.

The proposed method for estimating indicators is based on modeling the heavy tails of economic survey data with a Pareto distribution in order to identify the extreme outliers. Therefore we adapt promising robust estimators to take sample weights into account, which is the main methodological contribution of this paper. Then we propose two strategies to reduce the influence of the extreme outliers on the indicators: downweighting the outlying observations, or replacing their values. This general approach has the advantage that it can be applied to many indicators. Nevertheless, in order to keep the paper concise, it is focused on applying the developed methodology to the Gini coefficient estimated from EU-SILC data. An extensive simulation study including results for other indicators can be found in a technical report (Hulliger et al., 2011). Furthermore, it should be noted that more general Pareto-type distributions or other complex distributions could be considered to model the data, but the Pareto distribution is chosen due to its simple form (see, e.g., Kleiber and Kotz, 2003, for an overview of statistical distributions in economics).

In the following, we present an overview of the literature on the Pareto model. In Pareto tail modeling, typically the shape of the Pareto distribution is estimated for points over a large threshold. Possibly the most widely known estimator was suggested by Hill (1975) and follows a maximum likelihood approach. Other classical estimators were introduced by Pickands (1975), Dekkers and de Haan (1989), and Kratz and Resnick (1996). Brazauskas and Serfling (2000a,b) examined various estimators with respect to their robustness properties. More advanced robust estimators were proposed by Victoria-Feser and Ronchetti (1994, 1997) following an optimal bias-robust approach, or by Dupuis and Morgenthaler (2002) and Dupuis and Victoria-Feser (2006) following a weighted maximum likelihood approach. Vandewalle et al. (2007), on the other hand, developed a promising robust estimator based on an integrated squared error criterion.

For the choice of the threshold, various proposals have been made in the literature as well. Beirlant et al. (1996a,b) and Danielsson et al. (2001) introduced procedures to determine the optimal choice of the number of observations in the tail for the Hill estimator based on minimizing the asymptotic mean squared error (AMSE). Nevertheless, those two procedures are not robust as they are designed for the non-robust Hill estimator. A robust prediction error criterion for simultaneously choosing the number of observations in the tail and estimating the shape parameter was introduced by Dupuis and Victoria-Feser (2006).

Concerning robustness in survey statistics, Chambers (1986) introduced the notion of *representative* and *nonrepresentative* outliers. Keep in mind that each observation in a survey sample represents a number of observations in the population as given by its sample weight. Representative outliers are observations whose values are correctly recorded and are not unique in the population. Therefore they contain relevant information and need to be considered in the estimation of quantities of interest. Nonrepresentative outliers are observations that either contain incorrect values or can in some sense be considered unique in the population. Consequently, they may corrupt the estimation of quantities of interest and need to be excluded or downweighted. In economic survey data, representative outliers are the observations forming the heavy tails, whereas nonrepresentative outliers are even more extreme observations that deviate from the observations in the tails. It is important to note that nonrepresentative outliers may very well belong to the true distribution on the population level, but including them in the estimation of quantities of interest from the sample may have too high an influence on the estimates. Cowell and Flachaire (2007) use the term *high-leverage* observations for such data points and stress their frequent occurrence in economic data.

The rest of the paper is organized as follows. Brief descriptions of the Gini coefficient and the Pareto distribution are given in Sections 2 and 3, respectively. Section 4 presents the Pareto quantile plot for the case of survey data. Afterwards, selected estimators for the shape parameter of the Pareto distribution are adapted for sample weights in Section 5. How to use the semiparametric Pareto model for robust estimation of economic indicators is described in Section 6. In Section 7, the estimators are evaluated by means of simulation. The application to EU-SILC data is then presented in Section 8. Finally, Section 9 concludes.

2. Gini coefficient

The motivating application of this paper is given by the European indicators on social exclusion and poverty, which have been defined by the European Union for monitoring and evaluating economic policies of its members and other European countries. A large subset of these indicators is estimated from the well known panel survey *European Union Statistics on Income and Living Conditions* (EU-SILC). We focus on one of the indicators that is particularly influenced by extreme outliers: the *Gini coefficient*.

Originally proposed by Gini (1912), the Gini coefficient is a well known measure of inequality of a distribution and is widely applied in many fields of research. In the context of EU-SILC, it is used to measure inequality of income. Eurostat (2004, 2009) defines the Gini coefficient as the relationship of cumulative shares of the population arranged according to the level of income, to the cumulative share of the income received by them. All members of a household are thereby assigned the same *equivalized disposable income* (see Eurostat, 2004, 2009, for details on its computation).

For a definition of the Gini coefficient in mathematical terms, let $\boldsymbol{x} := (x_1, \ldots, x_n)'$ be the income with $x_1 \leq \ldots \leq x_n$ and let $\boldsymbol{w} := (w_1, \ldots, w_n)'$ be the corresponding sample weights, where n denotes the number of observations. Then the Gini coefficient is estimated by

$$\widehat{Gini} := 100 \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} {w_i}^2 x_i}{\left(\sum_{i=1}^{n} w_i \right) \sum_{i=1}^{n} (w_i x_i)} - 1 \right]. \tag{1}$$

The Gini coefficient is closely related to the Lorenz curve (Lorenz, 1905), which plots the cumulative proportion of the total income against the corresponding proportion of the

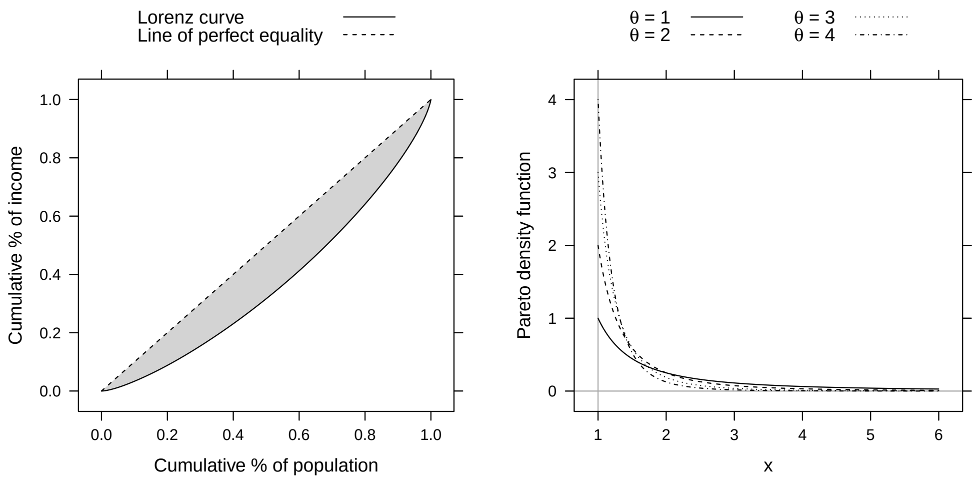

Fig. 1. Left: Example for the Lorenz curve. Right: Probability density function of the Pareto distribution with parameters $x_0 = 1$ and $\theta = 1, 2, 3, 4$.

population. As for the Gini coefficient, the data are first sorted by income in non-decreasing order. An example for the Lorenz curve is shown in Figure 1 (left). The line at the angle of 45° thereby corresponds to perfect equality of incomes. The Gini coefficient can then be written as

$$Gini = 100 \cdot 2A, \tag{2}$$

where A denotes the area between the Lorenz curve and the line of perfect equality. In the example in Figure 1 (left), the area A is shaded in grey.

3. Pareto distribution

In this paper, the *Pareto distribution* is used to model the upper tail of economic survey data in order to identify extreme outliers that may highly influence indicators. The Pareto distribution is well studied in the statistics and economics literature. It is defined in terms of its cumulative distribution function

$$F_\theta(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}, \qquad x \geq x_0, \tag{3}$$

where $x_0 > 0$ is the scale parameter and $\theta > 0$ is the shape parameter (Kleiber and Kotz, 2003). The corresponding density function is given by

$$f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \qquad x \geq x_0. \tag{4}$$

Figure 1 (right) displays the density function of the Pareto distribution with scale parameter $x_0 = 1$ and different values of the shape parameter θ. The effect of changing the shape parameter θ is thereby clearly visible: the lower θ, the lower the probability mass at

x_0 and the longer the tail. In extreme value theory, the *tail index* is a measure of the tail heavyness of a distribution. For the Pareto distribution, the tail index is in fact given by $\gamma = 1/\theta$.

In the semiparametric Pareto tail model, the cumulative distribution function on the whole range of x is modeled as

$$F(x) = \begin{cases} G(x), & \text{if } x \leq x_0, \\ G(x_0) + (1 - G(x_0))F_\theta(x), & \text{if } x > x_0, \end{cases} \tag{5}$$

where G is an unknown distribution function (Dupuis and Victoria-Feser, 2006).

Let n be the number of observations and let $\boldsymbol{x} = (x_1, \dots, x_n)'$ denote the observed values with $x_1 \leq \dots \leq x_n$. If k is the number of observations to be used for tail modeling, the threshold x_0 is estimated by

$$\hat{x}_0 := x_{n-k}. \tag{6}$$

If, on the other hand, an estimate $\hat{x}_0$ for the scale parameter of the Pareto distribution is available, k is given by the number of observations larger than $\hat{x}_0$. In this way, the estimation of x_0 and k directly corresponds with each other.

4. Pareto quantile plot

In applied data analysis, visual exploration is an important first step to gain insight into the data at hand. For our purpose, the *Pareto quantile plot* allows to check whether the Pareto model for the upper tail of the data is suitable. Moreover, it is a graphical method for inspecting the parameters of a Pareto distribution. For the case without sample weights, it is described in detail in Beirlant et al. (1996a).

If the Pareto model holds, there exists a linear relationship between the logarithms of the observed values and the quantiles of the standard exponential distribution, since the logarithm of a Pareto distributed random variable follows an exponential distribution. Hence the logarithms of the observed values, $\log(x_i)$, $i = 1, \dots, n$, are plotted against the theoretical quantiles.

In the case without sample weights, the theoretical quantiles of the standard exponential distribution are given by

$$-\log\left(1 - \frac{i}{n+1}\right), \qquad i = 1, \dots, n, \tag{7}$$

i.e., by dividing the range into $n + 1$ equally sized subsets and using the resulting n inner gridpoints as probabilities for the quantiles. For survey data, the range of the exponential distribution needs to be divided according to the weights of the n observations. The Pareto quantile plot is thus generalized by using the theoretical quantiles

$$-\log\left(1 - \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j} \frac{n}{n+1}\right), \qquad i = 1, \dots, n, \tag{8}$$

where the correction factor $n/(n + 1)$ ensures that the quantiles reduce to (7) if all sample weights are equal.

If the tail of the data follows a Pareto distribution, those observations form almost a straight line. The leftmost point of a fitted line can thus be used as an estimate of the

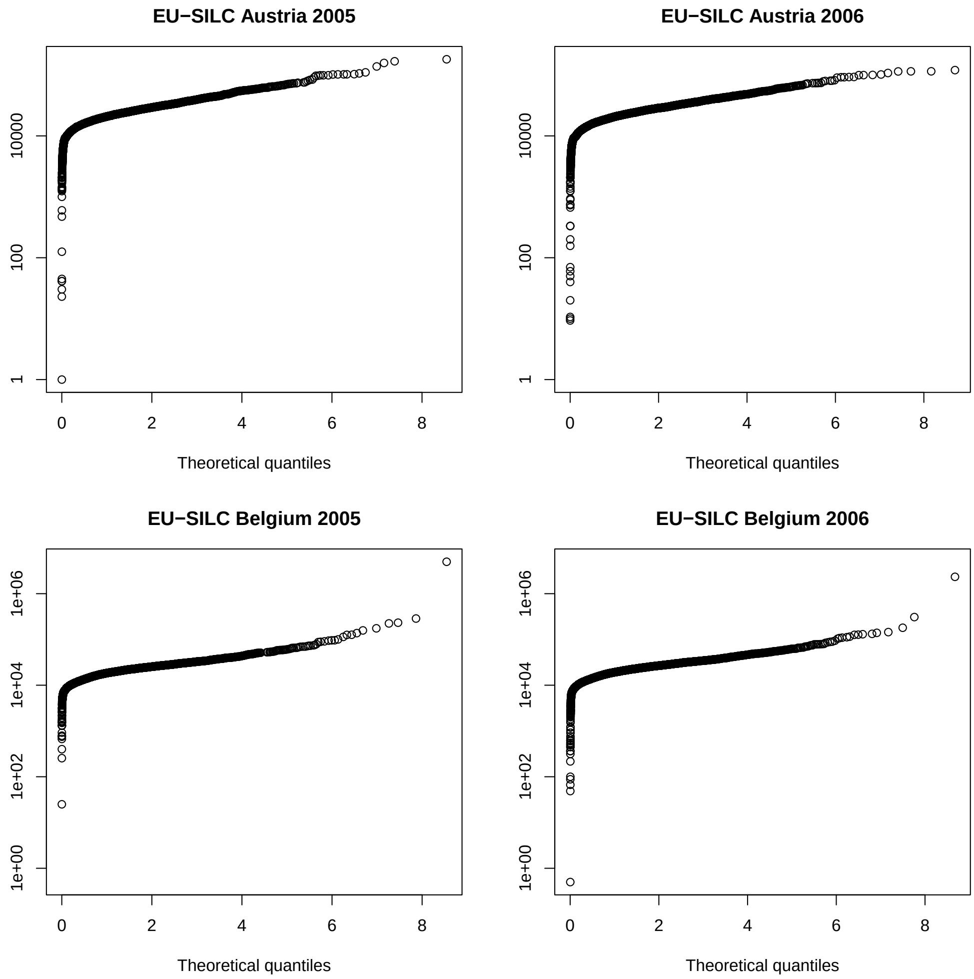

Fig. 2. Pareto quantile plots of Austrian (top) and Belgian (bottom) EU-SILC income survey data from 2005 (left) and 2006 (right).

threshold x_0, the scale parameter. All values starting from the point after the threshold may be modeled by a Pareto distribution, but of course this point cannot be determined exactly by graphical means. Furthermore, the slope of the fitted line is in turn an estimate of $1/\theta$, the reciprocal of the shape parameter. Another advantage of the Pareto quantile plot is that nonrepresentative outliers, i.e., extreme observations in the upper tail that deviate from the Pareto model, are clearly visible.

Figure 2 shows Pareto quantile plots for Austrian and Belgian EU-SILC income survey data from 2005 and 2006. These data sets were provided by Eurostat and are used in the application for the estimation of the Gini coefficient in Section 8. Note that the Austrian data are clean despite some minor irregularities in the upper tail of the 2005 data, whereas the Belgian data sets each contain one clear outlier from the Pareto tail model.

5. Robust estimation of the shape parameter

In order to detect nonrepresentative outliers that deviate from the Pareto model for the upper tail, the shape parameter of the Pareto distribution needs to be estimated in a robust manner. This section therefore describes promising estimators for the shape parameter of a Pareto distribution. Since the original proposals do not take sample weights into account, the estimators are adjusted for the case of survey samples.

5.1. Integrated squared error (ISE) estimator

Terrell (1990) first proposed estimation based on an integrated squared error minimum distance criterion as a more robust alternative to the maximum likelihood framework. Intuitively speaking, this estimation method reduces the influence of outliers by trying to find largest proportion of the data that matches the assumed parametric model (Vandewalle et al., 2007). A detailed discussion on this behavior can be found in Scott (2001).

For the *integrated squared error* (ISE) estimator in the case of the semiparametric Pareto tail model (Vandewalle et al., 2007), the Pareto distribution is modeled in terms of the relative excesses

$$y_i := \frac{x_{n-k+i}}{x_{n-k}}, \qquad i = 1, \dots, k. \tag{9}$$

Then the density function of the Pareto distribution for the relative excesses is approximated by

$$f_\theta(y) = \theta y^{-(1+\theta)}. \tag{10}$$

With this density, the integrated squared error criterion to find an estimate of the parameter θ is given by

$$\begin{aligned}\hat{\theta} &= \arg\min_\theta \left[\int \left(f_\theta(y) - f(y)\right)^2 dy \right] & (11)\\ &= \arg\min_\theta \left[\int f_\theta^2(y)dy - 2\int f_\theta(y)f(y)dy + \int f^2(y)dy \right], & (12)\end{aligned}$$

where $f(y)$ denotes the unknown true density. Since the last term is constant with respect to θ, it can be omitted. Furthermore, the middle term denotes the expected value of the model density. Hence Equation (12) can be rewritten as

$$\hat{\theta} = \arg\min_\theta \left[\int f_\theta^2(y)dy - 2\mathbb{E}(f_\theta(Y)) \right]. \tag{13}$$

If there are no sample weights in the data, the ISE estimator is obtained by using the mean as an unbiased estimator of $\mathbb{E}(f_\theta(Y))$:

$$\hat{\theta}_{\mathrm{ISE}} = \arg\min_\theta \left[\int f_\theta^2(y)dy - \frac{2}{k}\sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{14}$$

For survey samples, the mean in Equation (14) is simply replaced by a weighted mean. This leads to the *weighted integrated squared error* (wISE) estimator

$$\hat{\theta}_{\mathrm{wISE}} = \arg\min_\theta \left[\int f_\theta^2(y)dy - \frac{2}{\sum_{i=1}^{k} w_{n-k+i}}\sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right]. \tag{15}$$

5.2. Partial density component (PDC) estimator

In an application of the integrated squared error criterion to outlier detection and regression, Scott (2004) noticed that this criterion only requires the true density f to be a real density, but not f_θ. Vandewalle et al. (2007) use this result to define the *partial density component* (PDC) estimator for the Pareto model. This estimator minimizes the integrated squared error criterion based on an incomplete density mixture model uf_θ. If the data do not contain sample weights, the PDC estimator is thus given by

$$\hat{\theta}_{\mathrm{PDC}} = \arg\min_\theta \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{k} \sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{16}$$

In order to obtain an estimate for the parameter u, the expression between brackets in Equation (16) is differentiated with respect to θ and evaluated at $\hat{\theta}_{\mathrm{PDC}}$. Equating to zero and solving the resulting equation then leads to the estimate

$$\hat{u} = \frac{1}{k} \sum_{i=1}^{k} f_{\hat{\theta}}(y_i) \bigg/ \int f_{\hat{\theta}}^2(y)dy. \tag{17}$$

A detailed discussion on the interpretation of $\hat{u}$ can be found in Vandewalle et al. (2007).

Taking survey sample weights into account, the *weighted partial density component* (wPDC) estimator is obtained by replacing the mean as an estimator of $\mathbb{E}(f_\theta(Y))$ by the weighted mean. Thus Equations (16) and (17) are generalized to

$$\hat{\theta}_{\mathrm{wPDC}} = \arg\min_\theta \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right], \tag{18}$$

$$\hat{u} = \frac{1}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_{\hat{\theta}}(y_i) \bigg/ \int f_{\hat{\theta}}^2(y)dy. \tag{19}$$

6. Robust estimation of indicators based on Pareto tail modeling

With all the pieces of the puzzle now in place, this section introduces two general approaches for reducing the influence of outliers on economic indicators. The basic idea is to first detect nonrepresentative outliers based on the semiparametric Pareto tail model. Then we propose two strategies to reduce their influence on the indicators: downweighting the outlying observations and recalibrating the remaining observations, or replacing the outlying values with values drawn from the fitted distribution.

In mathematical terms, we first define the outlier indicator O_i, $i = 1, \ldots, n$, based on the Pareto distribution $F_{\hat{\theta}}$ fitted to the upper tail of the data as

$$O_i := \begin{cases} 1, & \text{if } x_i > F_{\hat{\theta}}^{-1}(1-\alpha), \\ 0, & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n, \tag{20}$$

where $F_{\hat{\theta}}^{-1}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the fitted distribution. In principle, any estimator $\hat{\theta}$ could be used, but we propose to use a weighted estimator to avoid bias (cf. Section 7.1). Based on comprehensive experience from simulations, $\alpha = 0.005$ or $\alpha = 0.01$ seem to be suitable choices for the tuning parameter; cf. the extensive collection of simulation results in a technical report (Hulliger et al., 2011). If α is chosen too low, some nonrepresentative outliers may not be detected, whereas too high a value may lead to too many observations being declared as outliers. For the Gini coefficient, for instance, too low a value of α may thus lead to overestimation and too high a value to underestimation of the true population value. It should be noted that $\alpha = 0.005$ is used throughout this paper. In an application to a specific data set, the Pareto quantile plot can be used as a diagnostic tool to check whether a certain value for α is suitable. For this purpose, observations with $O_i = 1$ can be highlighted in the plot with a different plot symbol or color. Because outliers are clearly visible in the Pareto quantile plot (cf. Figure 2), it is possible to visually check whether the choice for α yields reasonable outlier detection performance.

Once nonrepresentative outliers are detected, they can be treated with one of the following two strategies.

Calibration for nonrepresentative outliers (CN): Since nonrepresentative outliers are considered to be unique to the population data in some sense, the sample weights of the corresponding observations are set to 1 and the weights of the remaining observations are adjusted accordingly by calibration. Hence we first define weights $w_i^* := 1$ for all observations with $O_i = 1$. In addition, let $I_j = (I_{1j}, \ldots, I_{nj})'$, $j = 1, \ldots, p$, be a set of indicator variables defining subgroups of the data such that $I_{ij} = 1$ if observation i belongs to subgroup j and $I_{ij} = 0$ otherwise. With corresponding population totals $N_j = \sum_{i=1}^{n} I_{ij} w_i$, calibration of the remaining observations with $O_i = 0$ then seeks weights w_i^* that are close to the original w_i while satisfying

$$\sum_{i:O_i=0} I_{ij} w_i^* = N_j - \sum_{i:O_i=1} I_{ij}, \qquad j = 1, \ldots, p. \tag{21}$$

If each observation i belongs to exactly one subgroup defined by the I_j, denoted by j_i, the calibrated sample weights can be written explicitly as

$$w_i^* = \frac{N_j - \sum_{l:O_l=1} I_{lj_i}}{\sum_{l:O_l=0} I_{lj_i} w_l} w_i, \qquad i : O_i = 0. \tag{22}$$

Nevertheless, in practice the I_j are often derived from more than one auxiliary variable (e.g., region and gender) such that each observation belongs to more than one subgroup. In that case, (21) yields a more complex optimization problem. Details on calibration can be found in, e.g., Deville et al. (1993). If the original sample weights w_i have already been obtained by calibration, it is a natural choice to use the same indicator variables for obtaining the weights w_i^*. Otherwise for stratified sampling designs, using the indicator variables giving the strata seems reasonable. In any case, an indicator is then computed using the standard formula with the original values x_i and the modified weights w_i^*, $i = 1, \ldots, n$. For the Gini coefficient, the formula from (1) is simply applied with w_i^* instead of w_i.

Replacement of nonrepresentative outliers (RN): The nonrepresentative outliers are replaced by values drawn from the fitted Pareto distribution, thereby preserving the order of the original values. Let $k^* := \sum_{i=1}^{n} O_i$ denote the number of detected nonrepresentative outliers, and let $i_1, \ldots, i_{k^*}$ denote their indices such that $x_{i_1} \leq \ldots \leq x_{i_{k^*}}$. Furthermore, let $z_1, \ldots, z_{k^*} \sim F_{\hat{\theta}}$ be random values drawn from the fitted distribution such that

$z_1 \leq \ldots \leq z_{k^*}$. The modified values x_i^* are then given by

$$x_i^* := \begin{cases} z_j & \text{if } i = i_j \text{ for any } j \in \{1, \ldots, k^*\}, \\ x_i & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n. \tag{23}$$

An indicator is then computed using the standard formula with the modified values x_i^* and the original weights w_i, $i = 1, \ldots, n$. For the Gini coefficient, the formula from (1) is simply applied with x_i^* instead of x_i.

Note that the application in this paper is focused on finite population estimation and inference. In such situations, the CN approach may be conceptually preferred since it modifies the sample weights rather than drawing values from the model distribution. On the other hand, if the theoretical income distribution is of interest as well, for instance for model-based superpopulation inference, the RN strategy may be the more natural choice.

In addition, it would also be possible to derive semiparametric estimators based on the fitted Pareto distribution. However, this would require new estimators to be derived for different indicators. To give an example, Cowell and Flachaire (2007) use moments to derive semiparametric estimators for a generalized entropy class of inequality indicators. The advantage of the proposed approaches based on outlier detection is that they can directly be applied to many indicators.

7. Simulation studies

The simulations presented in this section are performed in R (R Development Core Team, 2011) using the simulation framework from package **simFrame** (Alfons et al., 2010; Alfons, 2012). All considered methods are all available in package **laeken** (Alfons et al., 2012).

7.1. Estimation of the shape parameter

The first simulation experiment compares the weighted and unweighted estimators for the shape parameter of the Pareto distribution presented in Section 5. Its aim is to demonstrate the importance of considering the sample weights in the finite population sampling context.

First, 100 population data sets of size $N = 10\,000$ are generated. Values in the variable of interest $x = (x_1, \ldots, x_N)'$ are drawn from a Pareto distribution with scale parameter $x_0 = 1$ and shape parameter $\theta = 4$. The scale parameter x_0 is thereby assumed to be known throughout the simulation study. In addition, an auxiliary variable $p = (p_1, \ldots, p_N)'$ giving probability weights for sampling is created for each population. It takes $s = 100$ equally spaced values between 1 and 10 and is constructed for each observation $i = 1, \ldots, N$, as

$$p_i := \begin{cases} 1, & x_i > F_\theta^{-1}(\frac{s-1}{s}) \\ 10 - \frac{9}{s-1} j, & F_\theta^{-1}\left(\frac{j}{s}\right) < x_i \leq F_\theta^{-1}\left(\frac{j+1}{s}\right) \text{ for any } 1 \leq j < s-1, \\ 10, & x_i \leq F_\theta^{-1}(\frac{1}{s}) \end{cases}$$

where F_θ is the cumulative distribution function of the Pareto distribution from (3). Second, 100 samples of size $n = 200$ observations are drawn from each of the populations, resulting in a total number of 10 000 simulation runs. The samples are taken using Midzuno's method for unequal probability sampling (Midzuno, 1952) with inclusion probabilities determined by the probability weights p. Hence observations with lower values in the variable of interest have higher inclusion probabilities, which in turn results in lower sample weights. This is motivated by EU-SILC, where the equivalized income is the main variable of interest.

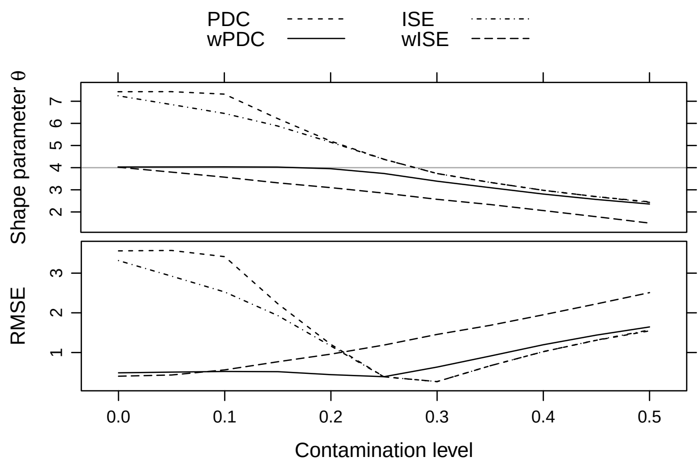

Fig. 3. Average simulation results (top) and RMSE (bottom) for the estimation of the shape parameter θ with contamination level ε varying between 0 and 50%.

In EU-SILC, higher inclusion probabilities are often assigned to larger households, which typically have lower equivalized income than smaller ones. Moreover, the above definition of the probability weights yields realistically large variation among the sample weights. Then a proportion ε of randomly selected observations in the samples are replaced by outliers. The contamination level ε is varied from 0 to 0.5 in steps of 0.05, and the values of the selected observations are drawn from a normal distribution $N(\mu, \sigma)$ with mean $\mu = 10$ (the 99.99% quantile of the Pareto distribution of the true values) and standard deviation $\sigma = 1$.

Figure 3 displays the average simulation results (top) and the root mean squared error (RMSE; bottom) for varying contamination level ε, where the true shape parameter $\theta = 4$ is indicated by the grey horizontal line. Clearly, the unweighted methods overestimate the shape parameter if there is no contamination. With increasing contamination level, the large outliers have a decreasing effect on the shape parameter, resulting in underestimation for higher contamination levels. The weighted estimators are very close to the true shape parameter in the case of no contamination. As contamination increases, the wISE estimator gradually moves away from the true value, but the robust wPDC remains accurate until about 20% contamination. Furthermore, the results for the bias are strongly reflected in the RMSE, although wISE exhibits a slightly lower RMSE than wPDC for very low contamination levels. To summarize, the sample weights need to be taken into account in the finite population sampling context, and the wPDC estimator is clearly favorable.

7.2. Estimation of the Gini coefficient

In this simulation study, robust estimation of the Gini coefficient based on the Pareto tail model is investigated in a close-to-reality setting. The basis for the simulations are synthetic population data generated from Austrian EU-SILC data from 2006, the latter of which were provided by Statistics Austria. The population data are thereby simulated with the methodology described in Alfons et al. (2011) and implemented in the R package **simPopulation** (Alfons and Kraft, 2012). In total, the population consists of 8 182 222 individuals from 3 505 145 households. For each individual, information on demographics and income is available. It is important to note that the synthetic population data do not contain outliers in the income data, as these are generated in the samples for full control over the amount of contamination (cf. Alfons, 2011).

Concerning the sampling design, 1 000 samples of 6 000 households are drawn via stratified cluster sampling. While the strata are given by the nine federal states of Austria, the primary sampling units (PSUs) correspond to the households. Within each stratum, households are drawn with unequal probabilities using Midzuno's method (Midzuno, 1952). The numbers of households selected from the strata are thereby proportional to the numbers of households in the population of the strata, and the probabilities of selection for households within each stratum are determined by the household sizes.

Since the Gini coefficient in the case of EU-SILC is estimated from an equivalized household income (see Section 2), contamination is inserted on the household level. First, households are selected by simple random sampling, then the equivalized household income of the selected households is drawn from a normal distribution $N(\mu, \sigma)$ with $\mu = 1\,000\,000$ and $\sigma = 20\,000$. All individuals within a contaminated household are assigned the same value. Moreover, EU-SILC data typically contain only a small amount of outliers. For a realistic scenario, the contamination level is set to $\varepsilon = 0.0025$, which in this case corresponds to 15 contaminated households. Additionally, the contamination level $\varepsilon = 0.01$ is investigated. This corresponds to 60 contaminated households out of a total 6 000 households, which in official statistics would be considered rather poor data quality.

In this simulation setting, the methods from Section 6 are evaluated for varying number of households k used for Pareto tail modeling. Only the weighted estimators of the shape parameter from Section 5 are thereby considered. Note that households are used as observations for fitting the Pareto distribution and detecting outliers. In the CN approach, the sample weight of each individual in an outlying household is set to 1. Calibration on the remaining individuals is then performed on the strata of the sampling procedure. In the RN approach, all individuals within the same outlying household receive the same replacement value. It is important to note that this simulation study is focused on exploring the behavior of the proposed methods for different choices of the threshold for tail modeling. The estimation of the threshold is not addressed in this paper. Furthermore, standard estimation of the Gini coefficient (see Section 2) is investigated for comparison.

Figure 4 (left) presents the simulation results for the case without contamination. The proposed methods are evaluated by the average over the simulation runs (top) and the root mean squared error (RMSE; bottom) for varying number of households k used for tail modeling. In addition, reference lines are drawn for the standard estimation (dash-dotted grey lines), as well as for the true population value of the Gini coefficient (solid grey line in the top plots). For almost the entire investigated range of k, the methods based on the Pareto tail model are very close to the standard estimation and the true population value (note the scale of the y-axes in the plots, differences occur only in the second digit after the comma). Only at $k \approx 350$, the estimates start to drift away from the true value and their RMSEs increase.

The results for the scenario with contamination level $\varepsilon = 0.0025$ are displayed in Figure 4 (center). For low values of k, the behavior of the methods based on the Pareto tail model is similar to the standard estimation until they reach enough observations to reduce the influence of the outliers. Afterwards, there is a smooth transition towards the true population value. The estimators thereby reach this changepoint very quickly at $k \approx 20$. In particular for the wPDC estimator, there is a large drop in bias and RMSE until $k \approx 30$,

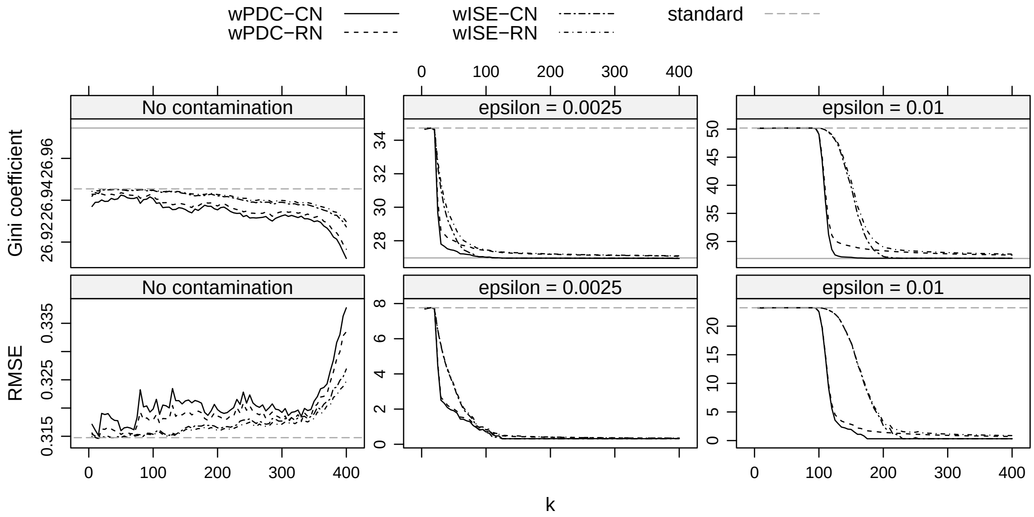

Fig. 4. Simulation results for the estimation of the Gini coefficient without contamination (left), contamination level $\varepsilon = 0.0025$ (center) and contamination level $\varepsilon = 0.01$ (right): average results (top) and RMSE (bottom). Reference lines are drawn for the standard estimation (dash-dotted grey lines), as well as for the true population value of the Gini coefficient (solid grey lines in the top plots).

after which point the curves continue with a slower rate of reduction. With the wISE estimator, the transition is smoother, hence not as steep in the beginning. Both methods stabilize at $k \approx 130$ and then remain very close to the true population value. Note that the CN approach performs slightly better with respect to bias than the RN approach, but differences with respect to RMSE are too small to state a clear preference.

In Figure 4 (right), the simulation results for the configuration with contamination level $\varepsilon = 0.01$ are shown. Due to the increased number of outlying households, a larger number of households in the tail is necessary before the wPDC and wISE estimators are able to reduce the influence of the outliers ($k \approx 100$ for wPDC, $k \approx 110$ for wISE). The wPDC estimator in this case also stabilizes much earlier than the wISE estimator ($k \approx 170$ for wPDC, $k \approx 240$ for wISE). Thus the methods based on the wPDC estimator clearly perform best in this scenario. However, the CN approach leads to very accurate results, whereas there is some remaining bias and a slightly larger RMSE for the RN method.

To summarize the simulation results for the Gini coefficient, robust estimation based on the wPDC and wISE estimators produces excellent results for a wide range of the number of households k used for tail modeling. In particular under heavier contamination in the upper tail, the wPDC estimator is preferable. Furthermore, the CN approach is favorable over RN since it leads to more accurate results and does not require drawing random values from the fitted distribution.

8. Application to real data

In this section, the proposed methods for the estimation of the Gini coefficient are applied to real Austrian and Belgian EU-SILC survey data from 2005 and 2006, which have already

Table 1. Gini coefficient estimated from Austrian and Belgian EU-SILC survey data from 2005 and 2006. Standard deviations estimated via stratified bootstrap with calibration are given in parenthesis.

		Austria		**Belgium**	
Method	Type	2005	2006	2005	2006
standard		26.13 (0.26)	25.33 (0.22)	28.53 (1.59)	27.82 (1.02)
wPDC	CN	26.13 (0.26)	25.33 (0.22)	26.91 (0.39)	26.78 (0.26)
wPDC	RN	26.13 (0.26)	25.33 (0.22)	26.92 (0.38)	26.79 (0.26)
wISE	CN	26.13 (0.26)	25.33 (0.22)	26.91 (0.40)	26.78 (0.26)
wISE	RN	26.13 (0.26)	25.33 (0.22)	26.91 (0.39)	26.79 (0.26)

been used for the Pareto quantile plots in Figure 2. Furthermore, the threshold for Pareto tail modeling is determined by Van Kerm's formula (Van Kerm, 2007). It is given by

$$\hat{x}_0 := \min(\max(2.5\bar{x}, Q(0.98)), Q(0.97)), \tag{24}$$

where $\bar{x}$ is the weighted mean and $Q(.)$ denotes weighted quantiles. Note that this formula was developed specifically for the equivalized disposable income in EU-SILC data and has more of a rule-of-thumb nature. A drawback of the formula is that it is designed to handle only very few outliers, but that is not a problem in this application. Robust estimation of the Gini coefficient based on Pareto tail modeling is then done in the same manner as in the simulation study from the previous section, except that the CN approach uses calibration according to regional information on a more aggregated level, as information on the strata used for sampling is not available in the data used here.

Table 1 shows the results for standard and robust estimation of the Gini coefficient for the Austrian and Belgian EU-SILC data from 2005 and 2006. Standard deviations estimated via stratified bootstrap with calibration are thereby given in parenthesis. Details on the computation of this bootstrap estimator are out of scope for this paper and can be found in Templ and Alfons (2011). For the methods based on Pareto tail modeling, all steps are performed for each bootstrap sample to account for the additional uncertainty. Nevertheless, stratification cannot be replicated exactly in the bootstrap samples as the data only contain more aggregated information than the strata used for sampling in practice. Hence estimates of the standard deviation may not be the most accurate, but they illustrate the importance of robust estimation of the Gini coefficient.

Regarding the Austrian data, the estimates based on the Pareto model are identical to the standard estimation of the Gini coefficient in Table 1. For neither the wPDC or the wISE estimator, the outlier indicator from Equation (20) detects any outliers in the 2005 or 2006 data. This is in accordance with the Pareto quantile plots in Figure 2 (top), which do not reveal any clear outliers either, only some irregularities in the upper tail for 2005. Since the 2005 and 2006 data appear very similar otherwise, those irregularities may be responsible for the differences in the corresponding Gini coefficient estimates.

The situation is different for the Belgian data. For both the wPDC and wISE estimator, the outlier indicator from Equation (20) detects the largest observation as the only outlier in 2005, and the two largest observations in 2006. Hence the respective estimates based on the Pareto model in Table 1 are all very similar and considerably lower than the corresponding standard estimates. However, the estimated standard deviations are much larger for the standard estimation of the Gini coefficient than for the robust methods. Considering that the sample size is 5 137 households for 2005 and 5 860 households for 2006, this suggests a disproportionally high influence of only 1 or 2 households with extremely large equivalized income. The robust estimates may therefore be considered more reliable. In addition, the outlier detection can again be checked by looking at the Pareto quantile plots in Figure 2 (bottom). The plot of the 2005 data also reveals the largest observation as the only outlier, otherwise the upper tail follows the Pareto model quite nicely. For 2006, the largest observation is also a clear outlier. While the second largest observation slightly deviates from the Pareto model, it is debatable whether it constitutes an outlier. Except for the outliers, the 2005 and 2006 data are very similar, which may explain the small differences between the two years in the robust estimates.

Note that even though the Pareto quantile plot could be used to graphically find outliers in the first place, such a procedure would be highly subjective. Detecting outliers based on a robustly fitted Pareto distribution is objective and should thus be preferred. Nevertheless, the Pareto quantile plot is an important diagnostic tool to determine whether the Pareto model is appropriate and to visually check the detected outliers.

9. Conclusions

Motivated by the estimation of the Gini coefficient from EU-SILC data, this paper introduces robust methods for the estimation of economic indicators from survey samples. More specifically, two approaches based on identifying extreme outliers via Pareto tail modeling are considered. For this purpose, commonly used estimators for the shape parameter of the Pareto distribution are adapted to allow for sample weights. The importance of taking sample weights into account is demonstrated by a small simulation experiment. Moreover, a close-to-reality simulation study and an application to real EU-SILC data demonstrate the excellent performance of the robust procedure in the case of the Gini coefficient. In particular the weighted partial density component (wPDC) estimator for the shape parameter of the Pareto distribution is favorable, as it still leads to excellent results in the simulations with an unrealistically high amount of contamination.

Even though the Gini coefficient is used as an example in this paper, the developed approaches can be applied to other economic indicators as well. For instance, an extensive collection of results for several European indicators on social exclusion and poverty from a wide range of simulation settings can be found in a technical report (Hulliger et al., 2011). Neither is the developed methodology restricted to EU-SILC data, as heavily tailed distributions frequently occur in economic surveys. The proposed adaptation of the Pareto quantile plot can thereby be used to check whether the Pareto tail model is appropriate for the data at hand.

Last but not least, all methods presented in this paper are available in the statistical environment R through the contributed package **laeken**.

Acknowledgments

This work was partly funded by the European Union (represented by the European Commission) within the 7$^{\text{th}}$ framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI, Grant Agreement No. 217322). For more information on the project, visit `http://ameli.surveystatistics.net`. Furthermore, we thank the Joint

Editor, the Associate Editor and the referee for their constructive remarks that led to an improvement of the paper.

References

Alfons, A. (2011). *Simulation and Robust Statistics: Application to Laeken Indicators and Quality of Life Research.* Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften. ISBN 978-3-8381-2706-4.

Alfons, A. (2012). ***simFrame**: Simulation framework.* R package version 0.5.0.

Alfons, A., J. Holzer, and M. Templ (2012). ***laeken**: Laeken indicators for measuring social cohesion.* R package version 0.3.3.

Alfons, A. and S. Kraft (2012). ***simPopulation**: Simulation of synthetic populations for surveys based on sample data.* R package version 0.4.0.

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications 20*(3), 383–407.

Alfons, A., M. Templ, and P. Filzmoser (2010). An object-oriented framework for statistical simulation: The R package **simFrame**. *Journal of Statistical Software 37*(3), 1–36.

Beirlant, J., P. Vynckier, and J. Teugels (1996a). Excess functions and estimation of the extreme-value index. *Bernoulli 2*(4), 293–318.

Beirlant, J., P. Vynckier, and J. Teugels (1996b). Tail index estimation, Pareto quantile plots, and regression diagnostics. *J. Am. Statist. Ass. 31*(436), 1659–1667.

Brazauskas, V. and R. Serfling (2000a). Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal 4*, 12–27.

Brazauskas, V. and R. Serfling (2000b). Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes 3*, 231–249.

Chambers, R. (1986). Outlier robust finite population estimation. *J. Am. Statist. Ass. 81*(396), 1063–1069.

Cowell, F. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*(2), 1044–1072.

Danielsson, J., L. de Haan, L. Peng, and C. de Vries (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis 76*(2), 226–248.

Dekkers, A.L.M., E.-J. and L. de Haan (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist. 17*(4), 1833–1855.

Deville, J.-C., C.-E. Särndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *J. Am. Statist. Ass. 88*(423), 1013–1020.

Dupuis, D. and S. Morgenthaler (2002). Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics 30*(1), 17–36.

Dupuis, D. and M.-P. Victoria-Feser (2006). A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics 34*(4), 639–658.

Eurostat (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg.

Eurostat (2009). Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics, Eurostat, Luxembourg.

Gini, C. (1912). Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. *Studi Economico-Giuridici della R. Università di Cagliari 3*, 3–159.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist. 3*(5), 1163–1174.

Hulliger, B., A. Alfons, C. Bruch, P. Filzmoser, M. Graf, J.-P. Kolb, R. Lehtonen, D. Lussmann, A. Meraner, R. Münnich, D. Nedyalkova, T. Schoch, M. Templ, M. Valaste, A. Veijanen, and S. Zins (2011). Report on the simulation results. Deliverable D7.1, AMELI Project.

Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken: John Wiley & Sons. ISBN 0-471-15064-9.

Kratz, M. and S. Resnick (1996). The QQ-estimator and heavy tails. *Stochastic Models 12*(4), 699–724.

Lorenz, M. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association 9*(70), 209–219.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics 3*(2), 99–107.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist. 3*(1), 119–131.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Scott, D. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics 43*(3), 274–285.

Scott, D. (2004). Partial mixture estimation and outlier detection in data and regression. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst (Eds.), *Theory and Applications of Recent Robust Methods*, pp. 297–306. Basel: Birkhäuser. ISBN 3-7643-7060-2.

Templ, M. and A. Alfons (2011). Variance estimation of social inclusion indicators using the R package **laeken**. Research Report CS-2011-3, Department of Statistics and Probability Theory, Vienna University of Technology.

Terrell, G. (1990). Linear density estimates. In *Proc. of the Statistical Computing Section*, pp. 297–302. American Statistical Association.

Tillé, Y. (2006). *Sampling Algorithms.* New York: Springer. ISBN 0-387-30814-8.

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

Vandewalle, B., J. Beirlant, A. Christmann, and M. Hubert (2007). A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics & Data Analysis 51*(12), 6252–6268.

Victoria-Feser, M.-P. and E. Ronchetti (1994). Robust methods for personal-income distribution models. *The Canadian Journal of Statistics 22*(2), 247–258.

Victoria-Feser, M.-P. and E. Ronchetti (1997). Robust estimation for grouped data. *J. Am. Statist. Ass. 92*(437), 333–340.