Diplomarbeit

zum Thema

Robust methods for the estimation of selected Laeken indicators

ausgeführt am
Institut für Statistik und Wahrscheinlichkeitstheorie
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser
und
Univ.-Ass. Dipl.-Ing. Matthias Templ
als
verantwortlich mitwirkendem Universitätsassistenten

durch

Josef Holzer
Matrikelnr.: 0326888
Schachen 19
8250 Vorau

Vorau, am 7. Mai 2009 Josef Holzer

Acknowledgments

First of all I would like to thank my supervisor Professor Dr. Peter Filzmoser, who inspired my interests in robust and computational statistics in several courses. As a result, an interesting topic for this diploma thesis was selected. Whenever I had questions, he took his time to give helpful suggestions and support. I also thank Dipl.-Ing. Matthias Templ for the great help to find suitable literature for this topic and for the support in the implementation of the methods. He often spent his sparse time to give new ideas for improvement and for managing the AMELI project, where this diploma thesis forms a small part.
Furthermore, I would like to thank Andreas Alfons and Stefan Kraft for their support and help with the statistical environment **R**. In addition to that, the good atmosphere in the office, made it easier to write this thesis.
In particular, I would like to thank Romana Schachner for her support and useful remarks. Special thanks go to my family for their permanent support. Without their help, the current lines would not have been written.

Josef Holzer

Abstract

This diploma thesis is a part of the *Advanced Methodology for European Laeken Indicators* (AMELI) project, in which the *Department of Statistics and Probability Theory of the Vienna University of Technology* is involved. The goal of this project is to develop new robust methods and improve the existing methods for the estimation and the monitoring of poverty, inequality and social cohesion. Furthermore, the analysis of the results should help policy makers in their decisions.
The main task of this thesis is to compare several methods to make selected Laeken indicators more robust. The robustness methods can be divided into three parts, non-parametric-, semi-parametric- and parametric approaches. For the non-parametric approach trimming and winsorising is examined. The other two parts need parametric assumptions. First, the parametric approach is to fit income distribution models. For example the log-normal distribution can be used. For the semi-parametric approach only the upper tail of the empirical distribution is used. Large income values have also a large influence on the estimation of the Laeken indicators, so it makes sense to fit the upper tail with a distribution. In this thesis the Pareto distribution is used for this purpose. A list of methods for the estimation of the parameters of the Pareto distribution is being described as well as methods for the estimation of a specific threshold, where above this threshold the Pareto distribution is applied. These different methods are compared to find an argumentation, which methods are useful for their application to the *European Survey on Income and Living Conditions* (EU-SILC) data set, which provides a basis for the AMELI project. The theory to the different statistical methods is also included in this diploma thesis.
The second main task of this thesis was the implementation of programs for the methods described above in the statistical environment **R** (see R Development Core Team 2009). The documentation of the developed software and the source code can be found in the appendix.

Contents

Chapter 1

Introduction

The *Lisbon Strategy* is a development plan for the European Union (EU). Its goal is to turn the EU into the most competitive and dynamic economy in the world by 2010. The strategy was developed in 2000 for a period of ten years in Lisbon by the European Council. For this reason, the European Commission formed some social indicators for the measurement of poverty, inequality and social cohesion. This set of indicators are called Laeken indicators. More details can be found in Kok (2004).
The project *Advanced Methodology for European Laeken Indicators* (AMELI) started in April 2008. The AMELI project is a joint project of statistical offices and some universities in the countries Germany, Switzerland, Finland, Austria, Slovenia and Estonia. The project is split into more than 10 parts, called work packages. The main goal of the project is to increase the quality of the methodology for the estimation of the Laeken indicators. Furthermore, the development of new robust methods and also new imputation methods should increase the precision of the indicators. Another part of the AMELI project is visualisation, which is also an important issue. The analysis of the results of the developed methods and graphics should help the policy makers in their decisions. The project will finish in 2011. This diploma thesis should be a small part of work package 4, which is called *robustness*.
The AMELI project is funded by the European Commission within the 7th Framework Programme for Research of the European Union and the project number is 217322.
For the estimation of the Laeken indicators the *European Survey on Income and Living Conditions* (EU-SILC) data set is used. This data set includes personal and household data and other topics, it is collected every year in all 27 member states of the European Union (EU) and in Norway, Iceland, Turkey and Switzerland as well. More information about EU-SILC can be found in EUR (2007). There is a direct impact between the quality of the data and the quality of the Laeken indicators. The EU-SILC data set is very important for the European social statistics and is taken as a basis for the estimation of the Laeken indicators (see Atkinson et al. 2002). In this diploma thesis the Austrian EU-SILC 2006 data are used.
It is well-known that social indicators can be very sensitive in presence of some extreme incomes. This is very problematic for most of the indicators. Therefore, robust

methods have to be developed.
The diploma thesis is structured as follows. In Section 2, a short description of the underlying mathematical and statistical theory, especially the log-normal- and Pareto distribution and the M-estimator, which is used later, is given. Section 3 introduces the basic definitions of three Laeken indicators, the at-risk-of-poverty-rate, the Gini coefficient and the quintile share ratio. The main part of this diploma thesis is included in Section 4, where many methods to obtain more robust estimations of the indicators are described. Here, non-parametric methods are listed, such as trimming and winsorising, as well as parametric modeling. First, a parametric fitting of income distribution models is done. Then, the indicators are computed from the estimated income model. The other procedure is parametric tail modeling. Here, often the Pareto distribution is used to fit the upper tail of the empirical distribution. In order to fix the threshold which divides the data into 'normal' data points and the upper tail including the extremes, a suitable threshold has to be estimated and also the parameters of the Pareto distribution. Hence, for parametric modeling assumptions must be made before analysis. For example, which income distribution is used or which distribution is used for the upper tail. Also the results of these methods are described in Section 4. The last Section 5 concludes.
The methods which are described in Section 4 are especially designed for the Gini coefficient and the quintile share ratio. Since, the at-risk-at-poverty-rate is more robust than the others, because outlying observations have less influence, this indicator is not considered in this thesis.
The methods will be implemented in the statistical environment **R** (see R Development Core Team 2009) and will be released in near future. Some helpful information on **R** can be found in Venables and Ripley (2002), Chambers (2008) and Chen et al. (2008).

Chapter 2

Mathematical background

2.1 Income distributions

2.1.1 Log-normal distribution

The log-normal distribution is often applied to income data. Details can be found in Limbert et al. (2001). The log-normal distribution is well described in the literature, such as in Crow and Shimizu (1988) or Kleiber and Kotz (2003).
Usually the notation $X \sim N(\mu, \sigma^2)$ implies that a positive random variable X originates from a normal distribution with parameters μ and σ^2 and $X \sim LN(\mu, \sigma^2)$ implies that X is log-normal distributed respectively.
The following equivalence holds,

$$X \sim LN(\mu, \sigma^2) \Leftrightarrow lnX \sim N(\mu, \sigma^2) \quad ,$$

whereas $-\infty < \mu < \infty$ and $\sigma > 0$. The probability density function is given by

$$f(x, \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\} & , \text{if} \quad x > 0 \\ 0 & , \text{if} \quad x \leq 0 \end{cases} .$$

In Figure 2.1 the probability density function with different choices of σ is shown. The density of the log-normal distribution is skewed to the right. For a given μ the skewness increases as σ increases.
The cumulative distribution function of the log-normal distribution is given by

$$F(x) = \begin{cases} \Phi\left(\frac{\ln x - \mu}{\sigma}\right) & , \text{if} \quad x > 0 \\ 0 & , \text{if} \quad x \leq 0 \end{cases} ,$$

whereas Φ is the cumulative distribution function of the standard normal distribution. The log-normal quantile function can be expressed as

$$F^{-1}(u) = e^{\mu + \sigma\Phi^{-1}(u)} \quad , \; 0 < u < 1 \; .$$

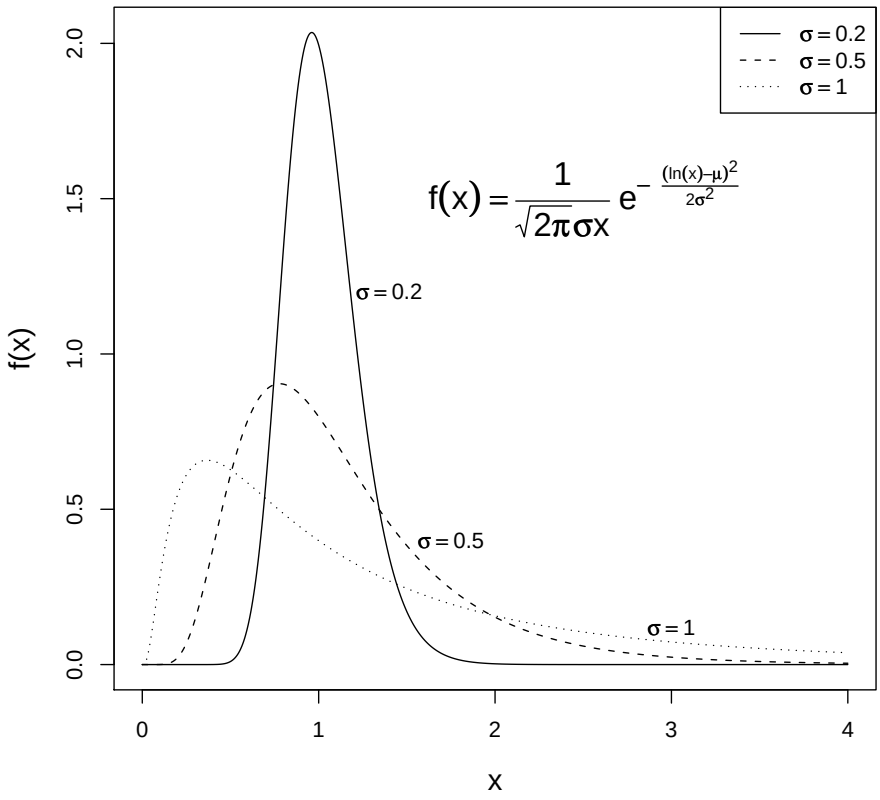

Figure 2.1: Probability densities of log-normal distributions with certain parameters for σ and $\mu = 0$

The moments exist for any real number r and can be expressed as

$$E(X^k) = e^{r\mu + \frac{1}{2}r^2\sigma^2} \quad .$$

A log-normal distribution is not uniquely determined by its moments. Other distributions can be defined with exactly the same moments. Further details and examples can be found in Crow and Shimizu (1988).
If X is a log-normally distributed variable, its expectation is given by

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \quad , \tag{2.1}$$

and the variance can be written as

$$Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \quad . \tag{2.2}$$

The formula for the median is

$$med(X) = e^{\mu} \quad ,$$

and the mode can be written as

$$mode(X) = e^{\mu - \sigma^2} \quad .$$

For a maximum likelihood (ML) estimation of the parameters μ and σ^2, firstly the likelihood function must be formulated:

$$L(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \quad .$$

In order to be able to find the maximum of the log-likelihood function, the following Equations (2.3) and (2.4) must be solved:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\ln x_i - \mu) = 0 \tag{2.3}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (\ln x_i - \mu)^2 = 0 \quad . \tag{2.4}$$

The solutions of Equation (2.3) and (2.4) yield the estimation of the parameters μ and σ^2,

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (\ln x_i - \hat{\mu}_{ML})^2 \quad .$$

2.1.2 Pareto distribution

The Pareto distribution is often used to parametrise the upper tail of income data. A detailed description can be found in Kleiber and Kotz (2003). The cumulative distribution function is given by

$$F(x) = 1 - \left(\frac{x}{x_0} \right)^{-\theta} \quad , \; x \geq x_0 \quad ,$$

whereas θ, $\theta > 0$, is the shape parameter and $x_0 > 0$ is the scale parameter. The notation of the Pareto distribution is given by

$$X \sim Par(x_0, \theta) \quad .$$

The probability density function can be expressed as

$$f(x, x_0, \theta) = \frac{\theta x_0^\theta}{x^{\theta+1}} \quad , \; x \geq x_0 \quad .$$

Figure 2.2 visualises the Pareto density function with different choices of the shape parameter θ. The Pareto distribution is a highly right skewed distribution. It is a heavy tailed distribution meaning that a random variable following a Pareto distribution can include extreme values. The effect of changing the shape parameter θ is clearly visible due to the maximum of the distribution. For a given x_0 the maximum increases as soon as θ increases.

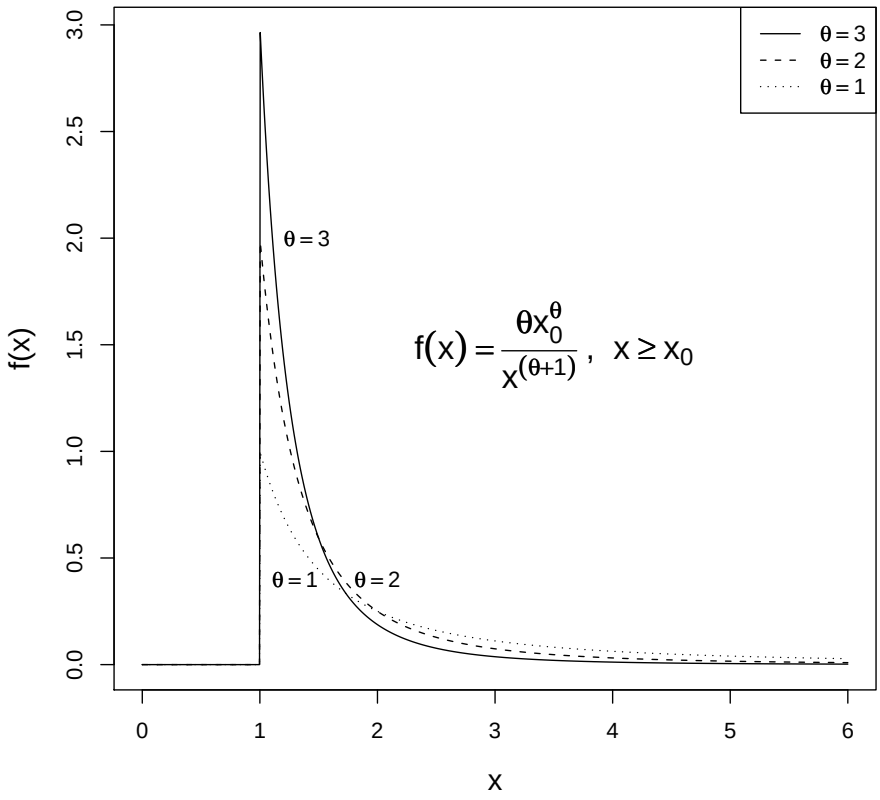

Figure 2.2: Probability densities of Pareto distributions with certain parameters for θ and $x_0 = 1$

The Pareto quantile function can be written as

$$F^{-1}(u) = x_0(1-u)^{-\frac{1}{\theta}} \quad , \; 0 < u < 1 \quad .$$

The k^{th} moment only exists if $k < \theta$. In that case the formula for the moments is

$$E(X^k) = \frac{\theta x_0^k}{\theta - k} \quad .$$

If X is a Pareto distributed variable, its expectation is given by

$$E(X) = \frac{\theta x_0}{\theta - 1} \quad , \ \theta > 1 \quad ,$$

and the variance can be written as

$$Var(X) = \frac{\theta x_0^2}{(\theta - 1)^2(\theta - 2)} \quad , \ \theta > 2 \quad .$$

The formula for the median is

$$med(X) = x_0 \sqrt[\theta]{2} \quad ,$$

and the mode is given by

$$mode(X) = x_0 \quad .$$

For a maximum likelihood estimation of the parameters x_0 and θ, firstly the likelihood function must be formulated:

$$L(x_1, \ldots, x_n; x_0, \theta) = \prod_{i=1}^{n} f(x_i; x_0, \theta) = \prod_{i=1}^{n} \frac{\theta x_0^\theta}{x_i^{\theta+1}} \quad .$$

The parameter x_0 can be estimated by

$$\hat{x}_{0_{ML}} = \min_i x_i \quad .$$

To determine the maximum of the log-likelihood function, the following equation (2.5) must be solved:

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} + n \ln x_0 - \sum_{i=1}^{n} \ln x_i = 0 \quad . \tag{2.5}$$

The solution of Equation (2.5) gives the estimation of the parameter θ

$$\hat{\theta}_{ML} = n \left[\sum_{i=1}^{n} \ln \left(\frac{x_i}{\hat{x}_{0_{ML}}} \right) \right]^{-1} . \tag{2.6}$$

The estimator given in Equation (2.6) is known as the Hill estimator and its detailed description can be found in Dupuis and Victoria-Feser (2006).

2.2 M-estimator of location

First of all, the M-estimates are a generalisation of the maximum likelihood (ML) estimates (see Maronna et al. 2006).
First, the log-likelihood function has to be formulated, given as

$$l(x_1, \ldots, x_n; T) = \sum_{i=1}^{n} \ln f(x_i - T) \quad ,$$

for a sample $x_1, \ldots, x_n$ and the location parameter T.
To determine the maximum, the log-likelihood function must be differentiated,

$$\sum_{i=1}^{n} \frac{f'}{f}(x_i - T) = 0 \quad .$$

The M-estimator can be expressed as

$$\arg\min_{T} \sum_{i=1}^{n} \rho(x_i - T) \quad . \tag{2.7}$$

If ρ is differentiable, the expression given in (2.7) can be differentiated with respect to T,

$$\sum_{i=1}^{n} \Psi(x_i - T) = 0 \quad , \tag{2.8}$$

with $\Psi = \rho'$. If Ψ is continuous, (2.8) can be solved exactly. Each maximum likelihood estimate is an M-estimate, if $\rho := \ln f$ and $\Psi := f'/f$, but not vice versa.
For example, if $\rho(x) = x^2/2$, then $\Psi(x) = x$. To obtain the M-estimate the following equation has to be solved

$$\sum_{i=1}^{n} (x_i - T) = 0 \quad ,$$

where the solution is given by $\hat{T} = \overline{x}$ (the arithmetic mean).
If $\Psi(0) = 0$ and $\Psi'(0)$ exists, the location M-estimate can be expressed as a weighted mean. The weight function is given by

$$w(x) = \begin{cases} \frac{\Psi(x)}{x} & , \text{if} \quad x \neq 0 \\ \Psi'(0) & , \text{if} \quad x = 0 \quad . \end{cases} \tag{2.9}$$

Then Equation (2.8) can be written as

$$\sum_{i=1}^{n} w(x_i - T)(x_i - T) = 0 \quad ,$$

yielding

$$\hat{T} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \text{with} \quad w_i = w(x_i - \hat{T}) \quad ,$$

where $w(.)$ is the weight function given in Equation (2.9).
The aim is that outlying observations get smaller weights. If Ψ is a suitable odd, bounded and monotonous increasing function, the breakdown point equals 0.5 in the univariate case. Some popular robust M-estimates are

- Huber-estimate
- Hampel estimate
- Andrews' wave
- Tukey's biweight

A lot of literature is available for robust M-estimates, for example, Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Staudte and Sheather (1990), Wilcox (1997) and Maronna et al. (2006).
The weight function $w(x)$ of the Huber M-estimator is given by

$$w(x) = \begin{cases} 1 & , \text{if} \quad |x| \leq k \\ \frac{k\, sgn(x)}{x} & , \text{if} \quad |x| > k \end{cases} \quad ,$$

where $sgn(x)$ can be written as

$$sgn(x) = \begin{cases} 1 & , \text{if} \quad x > 0 \\ 0 & , \text{if} \quad x = 0 \\ -1 & , \text{if} \quad x < 0 \end{cases} \quad .$$

Figure 2.3 shows the weight function of Huber's M-estimate with parameter $k = 1$. Observations within the interval $[-k, k]$ will not be weighted downwards and

$$\lim_{x \to \pm\infty} w(x) = 0 \quad .$$

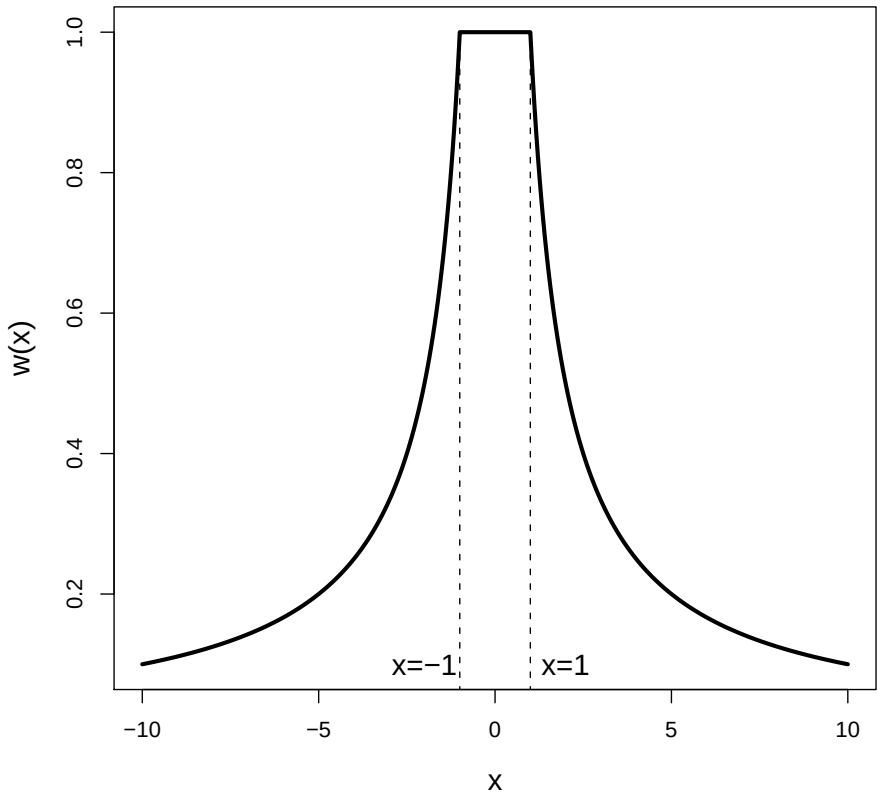

Figure 2.3: Weight function of the Huber M-estimator (k=1)

Chapter 3

Estimation of the Laeken indicators

A lot of research is devoted to social indicators. Detailed informations can be found in Atkinson et al. (2002), Cowell and Flachaire (2002), Jenkins and Van Kerm (2003) and the Council of Europe (2005). The Laeken indicators form some of these social indicators.
A description of a set of Laeken indicators can be found in the working papers EUROSTAT (2004a) and EUROSTAT (2004b). The *European Survey on Income and Living Conditions* (EU-SILC) is a data set with personal and household data and other topics, it is collected every year in all member states of the European Union (EU) and Switzerland.
The EU-SILC data set is very important for the European social statistics and is taken as a basis for the estimation of the Laeken indicators.

3.1 Equivalised disposable income (EQ_INC)

3.1.1 Definition

For each person of a household, the equivalised disposable income (EQ_INC) is defined as the total household disposable income divided by the equivalised household size (EQ_SS) .

3.1.2 Calculation of total disposable income

In order to be able to estimate the equivalised personal income for each person, the total diposable income has to be calculated first. The total disposable income is the sum of the personal income components for all household members plus income components of the household.
First, in Table 3.1 the personal income components are listed. The personal income

Table 3.1: Income components on personal level used for the calculation of EQ_INC

personal income components	
income variable	name
$PY010G$	employee cash or near cash income
$PY020G$	non-cash employee income
$PY030G$	employers's social insurance contribution
$PY050G$	cash benefits or losses from self-employment
$PY070G$	value of goods produced by own-consumption
$PY090G$	unemployment benefits
$PY100G$	old-age benefits
$PY110G$	survivor benefits
$PY120G$	sickness benefits
$PY140G$	education-related allowances

can be written as

$$\begin{aligned} pInc &= PY010G + PY020G + PY030G + PY050G + PY070G + PY090G + \\ &+ PY100G + PY110G + PY120G + PY130G + PY140G - PY030G \quad . \end{aligned}$$

The household income components are listed in Table 3.2. The household income is given by

$$\begin{aligned} hInc &= HY030G + HY040G + HY050G + HY060G + HY070G + HY080G + \\ &+ HY090G + HY110G - HY100G - HY120G - HY130G - HY140G \quad . \end{aligned}$$

Then the total disposable income ($HY020$) can be calculated using $pInc$ and $hInc$. Note that the inflation factor ($HY025$) equals 1 for the Austrian SILC survey,

$$HY020 * HY025 = (pInc + hInc) * HY025 \quad .$$

3.1.3 Calculation of equivalised household size (EQ_SS)

For the calculation of the equivalised household size, the variables shown in Table 3.3 are needed, which can be expressed as

$$EQ\_SS = 1 + 0.5 * (HM_{14+} - 1) + 0.3 * HM_{13-} \quad .$$

Table 3.2: Household income components used for the calculation of EQ_INC

household income components	
income variable	name
$HY030G$	imputed rent
$HY040G$	income from rental of a property or land
$HY050G$	family/children related allowances
$HY060G$	social exclusion not elsewhere classified
$HY070G$	housing allowances
$HY080G$	regular inter-household cash transfer received
$HY090G$	interest, dividends, profit from capital investments in unincorporated business
$HY100G$	interest repayments on mortgage
$HY110G$	income received by people aged under 16
$HY120G$	regular taxes on wealth
$HY130G$	regular inter-household cash transfer paid
$HY140G$	tax on income and social contributions

Table 3.3: Equivalised household size variables

variable	name
HM_{14+}	number of household members aged 14 and over
HM_{13-}	number of household members aged 13 or less

3.1.4 Calculation of equivalised disposable income (EQ_INC)

For each person, the equivalised disposable income (EQ_INC) is defined as the total household disposable income divided by the equivalised household size,

$$EQ\_INC = \frac{HY020 * HY025}{EQ\_SS} \quad .$$

Note that the equivalised disposable income is equal for each person who is living in the same household.

3.2 At-risk-of-poverty-rate

The 'at-risk-of-poverty rate' is given by the percentage of persons over the total population with equivalised disposable income below the 'at-risk-of-poverty threshold'. The 'at-risk-of-poverty rate' is defined (see Graf 2007) as

$$ARPR = P(x < 0.6 \cdot med(x)) = P(x' < 0.6) = F(0.6) \quad ,$$

where $x' = x/med(x)$ is the scaled equivalised income and F is the distribution function of the scaled income.
Since few information about the population is known, the 'at-risk-of-poverty rate' has to be estimated using the sample including the sample weights (u_i). The estimated 'at-risk-of-poverty threshold' is set at 60% of the national median equivalised disposable income and can be expressed as

$$ARPT = 0.6 * med(EQ\_INC) \quad .$$

Note that for the estimation of $med(EQ\_INC)$, the equivalised disposable income has to be sorted in increasing order and can be written as

$$med(EQ\_INC) = \begin{cases} \frac{1}{2}(EQ\_INC_j + EQ\_INC_{j+1}) & \text{, if} \quad \sum_{i=1}^{j} u_i = \frac{1}{2}U \\ EQ\_INC_{j+1} & \text{, if} \quad \sum_{i=1}^{j} u_i < \frac{1}{2}U < \sum_{i=1}^{j+1} u_i \end{cases} \quad ,$$

where $U = \sum_{i=1}^{n} u_i$ is the estimated population size and n is the number of observations.
Firstly, we need a variable y which can be expressed for all n persons in the sample such as

$$y_i = \begin{cases} 1 & \text{, if} \quad EQ\_INC_i < ARPT \quad i = 1, \ldots, n \\ 0 & \text{, otherwise} \quad . \end{cases}$$

The estimated 'at-risk-of-poverty-rate' is then given by

$$ARPR = \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i} * 100 \quad .$$

3.3 Gini coefficient

The Gini coefficient is defined as the relationship of cumulative shares of the population arranged according to the level of equivalised disposable income to the cumulative share of the equivalised total disposable income received by them.

3.3.1 Formula

The Gini coefficient is defined as (see Cowell and Victoria-Feser 2003)

$$G = 1 - 2 \int_0^1 \frac{C(F;q)}{C(F;1)} \, \mathrm{d}q \quad ,$$

where $C(F;q)$ is the cumulative income functional that can be expressed as

$$C(F;q) = \int_{\underline{x}}^{Q(F;q)} x \, \mathrm{d}F(x) \quad ,$$

where $\underline{x}$ is the lower limit of the support and the quantile functional $Q(F;q)$ is given by

$$Q(F;q) = \inf\{x | F(x) \geq q\} \quad .$$

For the estimation of the Gini coefficient the equivalised disposable income (EQ_INC) has to be sorted in increasing order.
The estimated Gini coefficient can be expressed as (see Hulliger and Münnich 2006)

$$\begin{aligned} \hat{G} &= \frac{1}{\hat{\tau}} \cdot \sum_{i \in S} u_i \cdot \left(2 \cdot \frac{1}{\hat{N}} \sum_{j \in S} u_j \cdot \mathbb{1}\{EQ\_INC_j \leq EQ\_INC_i\} - 1 \right) \cdot EQ\_INC_i \\ &= \frac{1}{\hat{\tau}} \cdot \sum_{i \in S} u_i \cdot \left(2 \cdot F_S(EQ\_INC_i) - 1 \right) \cdot EQ\_INC_i \quad , \end{aligned}$$

where $\hat{\tau}$ denotes the Horwitz-Thompson (HT) estimate for EQ_INC which is given by

$$\hat{\tau} = \frac{1}{\hat{N}} \sum_{i \in S} u_i \cdot EQ\_INC_i \quad ,$$

with $\hat{N} = \sum_{i \in S} u_i$ the number of individuals and S the index set referring to the sample.
The indicator function $\mathbb{1}\{EQ\_INC_j \leq EQ\_INC_i\}$ can be written as

$$\mathbb{1}\{EQ\_INC_j \leq EQ\_INC_i\} = \begin{cases} 1 & \text{, if} \quad EQ\_INC_j \leq EQ\_INC_i \\ 0 & \text{, otherwise} \quad . \end{cases}$$

A part in the formula of the Gini coefficient is the empirical cumulative distribution function $F_S(EQ\_INC_i)$ of incomes, which is given by

$$F_S(EQ\_INC_i) = \frac{\sum\limits_{i \in S} u_i \cdot \mathbb{1}\{EQ\_INC_j \leq EQ\_INC_i\}}{\sum\limits_{i \in S} u_i} \quad .$$

3.3.2 Lorenz curve

The Gini coefficient is the percentage of the area between the line of perfect equality and the observed Lorenz curve and the area between the line of perfect equality and line of perfect inequality (see Figure 3.1). The shaded area is related to the Gini coefficient. The higher the coefficient, the higher the disparity of the income of the population. Note that for the estimation of the Lorenz curve, the income has to be sorted in increasing order. Every point on the Lorenz curve represents a statement like 'the bottom 20% of all households have 10% of the total income'. A perfectly equal income distribution results when every person has the same income.

3.4 Quintile share ratio

The income quintile share ratio is given by the sum (or mean) of the upper 20% of the equivalised disposable income divided by the sum (or mean) of the lower 20% of the equivalised disposable income (see Hulliger and Münnich 2006).

3.4.1 Formula

The quintile share ratio (QSR) is defined (see Graf 2007) as

$$Q = \frac{E(x|x > Q(0.8))P(x > Q(0.8)}{E(x|x < Q(0.2))P(x < Q(0.2)} \quad ,$$

where $Q(0.2)$ and $Q(0.8)$ are the 0.2 and 0.8 quantiles of variable x.
The quintile share ratio can be estimated from the sample as follows

$$\hat{Q} = \frac{\sum\limits_{i \in S} u_i \cdot EQ\_INC_i \cdot \mathbb{1}\{EQ\_INC_i > Q_{EQ\_INC}(0.8)\}}{\sum\limits_{i \in S} u_i \cdot EQ\_INC_i \cdot \mathbb{1}\{EQ\_INC_i \leq Q_{EQ\_INC}(0.2)\}} \quad ,$$

where $\mathbb{1}$ is the indicator function and $Q_{EQ\_INC}(0.2)$ and $Q_{EQ\_INC}(0.8)$ are the 0.2 and 0.8 quantiles of the equivalised disposable income.

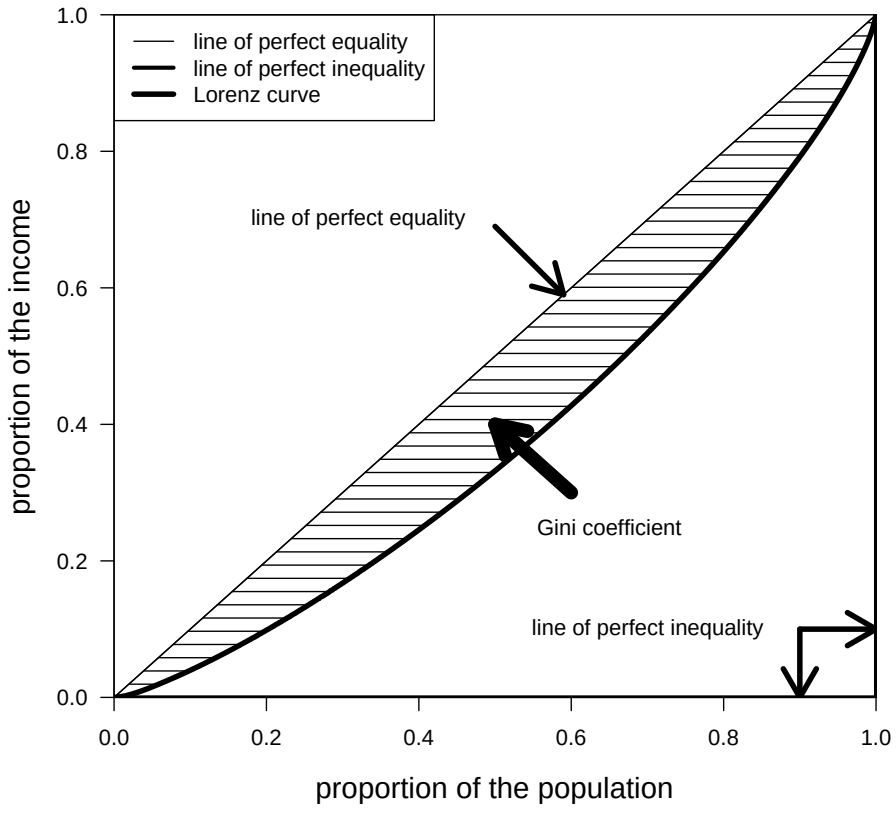

Figure 3.1: Lorenz curve

Chapter 4

Robustness

4.1 Introduction

In this Section we will introduce methods that lead to more robust versions of the Laeken indicators (Section 3). One definition of robust statistics is given in Hampel et al. (1986). *In a broad informal sense, robust statistics is a body of knowledge, partly formalised into 'theories of robustness', relating to deviations from idealised assumptions in statistics.*

4.1.1 Outlier

A definition of an outlier is given in Barnett and Lewis (1981). *We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.*
In Chambers (1986) there are two types of outliers. First the *representative outliers*, these are correctly measured observations that are outlying relative to the rest of the sample and for which there is no reason to believe that similar values do not exist in the non-sampled part of the survey population. These observations are handled in the survey editing process by the use of outlier robust estimation methods. The second group are the *non-representative outliers*, which are observations that are either incorrect, for example, if there are deficiencies in the survey process (coding). These errors have nothing to do with the values in the non-sampled part of the survey population. This type of outliers are detected and corrected during the survey editing process. For the robustness methods in this diploma thesis, an outlier is a data point, which has the potential to severely influence the Laeken indicators.

4.1.2 Outlier detection

In a first approach we bootstrap the original data set (EU-SILC 2006) and take a look at the Gini coefficient and the quintile share ratio.
Table 4.1 shows the results for the 99% bootstrap confidence intervals of the Gini coefficient and the 99% bootstrap confidence intervals of the quintile share ratio. For the simulation 1000 bootstrap samples are used and for each sample the Gini coefficient and the quintile share ratio (bootstrap replicates) are estimated. To determine the 99% confidence interval, the 0.005 and 0.995 quantiles has to be estimated. Values outside these boundaries are potential outliers. The aim of the next methods is that the indicators are as robust as possible.

Table 4.1: 99% bootstrap confidence interval of the Gini coefficient and the quintile share ratio

	Gini coefficient			quintile share ratio		
region	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
Austria	24.78	25.33	25.89	3.55	3.65	3.75
Burgenland	22.80	26.03	29.50	3.07	3.58	4.11
Lower Austria	24.46	25.81	27.12	3.55	3.83	4.11
Vienna	27.66	29.15	30.65	4.21	4.61	4.98
Carinthia	21.68	23.04	24.48	3.04	3.26	3.50
Styria	21.58	22.71	23.80	3.01	3.20	3.38
Upper Austria	23.08	24.33	25.62	3.19	3.41	3.62
Salzburg	21.39	23.56	25.77	2.92	3.33	3.70
Tyrol	21.91	23.62	25.45	2.98	3.24	3.52
Vorarlberg	22.31	24.30	26.28	3.08	3.42	3.72

Secondly, outliers are removed from the data set. Potential outliers are values outside the interval

$$\overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot s \quad ,$$

where $\overline{x}$ is an estimate of the arithmetic mean of the equivalised disposable income, s is the empirical standard deviation, α is the level of significance (e.g. $\alpha = 0.01$) and $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$ - quantile of the standard normal distribution $N(0,1)$ ($z_{0.995} = 2.58$). Results are shown in Table 4.2 (first line).
To obtain more robust estimates of the mean and standard deviation, the median and the median absolute deviation(MAD) could be used. The MAD is defined as

$$MAD(x) = 1.4826 \cdot med|x_i - med(x)|, \quad i = 1, \ldots, n \quad ,$$

where $med(x)$ is the median. Results are presented in Table 4.2 (second line).
Finally, the center can be estimated by the M-estimate (T) (see Section 2.2) and the weighted M-estimate (T_u). For the weighted M-estimate the sample weights u_i are

used. Table 4.2 shows the results of the Gini coefficient and the quintile share ratio of income data from Austria (line 3 and 4). The results obtained for each region in Austria are similar.

Table 4.2: Laeken indicators with trimmed data

estimation	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
$\overline{x} \pm 2.58s$	22.28	22.66	23.05	3.18	3.26	3.34
$med(x) \pm 2.58 \cdot MAD(x)$	20.82	21.18	21.54	2.98	3.05	3.12
$T + 2.58 \cdot MAD(x)$	20.81	21.18	21.54	2.97	3.04	3.10
$T_u \pm 2.58 \cdot MAD(x)$	18.34	18.63	18.93	2.57	2.62	2.66

Another possibility is to take the logarithm of the equivalised disposable income, use different estimates for the outlier detection rule, look which data are in the 99% interval, transform the data back, and afterwards the Laeken indicators are estimated. The results are given in Table 4.3.

Table 4.3: Laeken indicators for log-transformed and trimmed data

estimation	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
$\overline{x} \pm 2.58 \cdot s$	22.39	22.79	23.19	3.13	3.20	3.26
$med(x) \pm 2.58 \cdot MAD(x)$	21.48	21.83	22.19	2.98	3.04	3.10
$T + 2.58 \cdot MAD(x)$	21.50	21.85	22.22	2.99	3.05	3.10
$T_u \pm 2.58 \cdot MAD(x)$	18.58	18.84	19.10	2.56	2.60	2.63
boxplot outlier boundaries	21.12	21.48	21.85	3.02	3.09	3.16

A possible large bias is obtained, when applying classical estimates to trimmed data in order to estimate these Laeken indicators. It seems that the Laeken indicators with trimmed data have a considerable bias downwards, and thus other methods for robust estimation should be preferred.
In addition to that, a fifth possibility to detect outliers is to visualise the data with the help of boxplots. The numerical results are shown in Table 4.3 (last line) and the boxplot is shown in Figure 4.1.
The results in Table 4.3 are not much better than in Table 4.2. The indicators also have a considerable bias downwards and thus other methods are needed.
In this thesis for the variance estimation of the Laeken indicators bootstrap (see Efron and Tibshirani 1993) is used. Other methods for the variance estimation of indicators are described in Deville (1999), Canty and Davison (1999), Chambers (2005a), Osier (2006), Hulliger and Münnich (2006), Wolter (2007) and Münnich (2008).

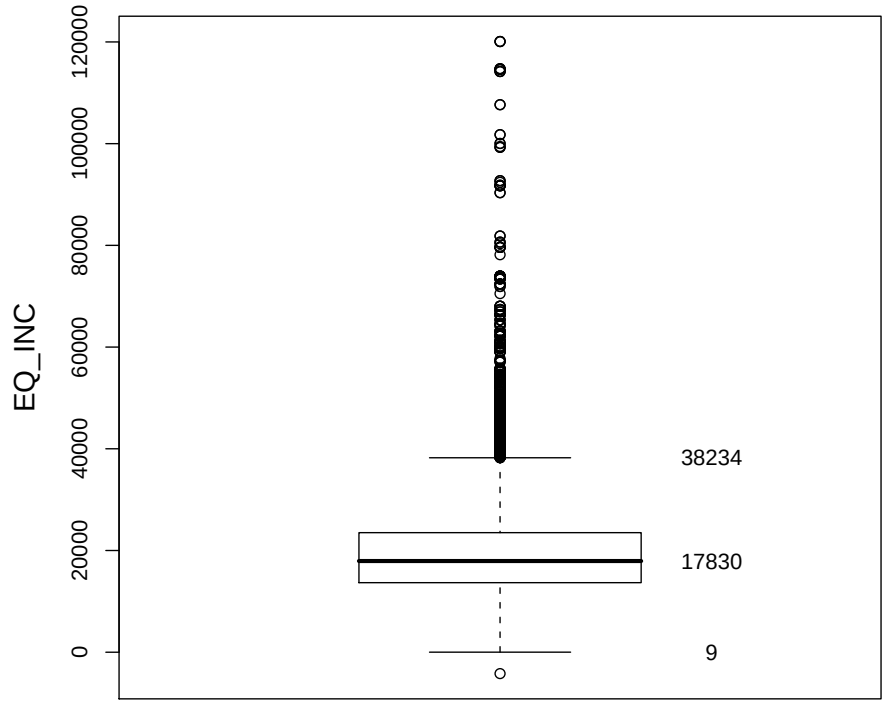

Figure 4.1: Boxplot of equivalised disposable income

4.2 Methods for the estimation of the threshold

As an alternative to rejecting outliers the equivalised disposable income can be modeled by a Pareto distribution (Section 2.1.2). Extreme values, however, will not be modeled by this distribution. Thus the task is to find the point x_0 where the 'extremes' start to occur, i.e. where the Pareto distribution is applied.
In this Section, several methods to detect this point will be presented.

4.2.1 Pareto quantile plot

The first method to evaluate x_0 is a graphical method, called the Pareto quantile plot, which is described in Dupuis and Victoria-Feser (2006) and Beirlant et al. (1999). First of all, the equivalised disposable income has to be sorted in increasing order. Then the values $log(EQ\_INC_i)$ are plotted against $-log(\frac{n+1-i}{n+1})$, for $i = 1, \ldots, n$. The plot typically shows a change-point, where values on the x-axis at the right of this

point flatten out. All values from the point after this change-point, where a nearly straight line begins, can be modeled by the Pareto distribution, but this point cannot be determined exactly. An example for the graphical detection of the threshold x_0 for the Austrian EU-SILC 2006 data set is given in Figure 4.2. The dashed lines indicates the straight line and the estimated point, where the Pareto distribution is considered as valid.

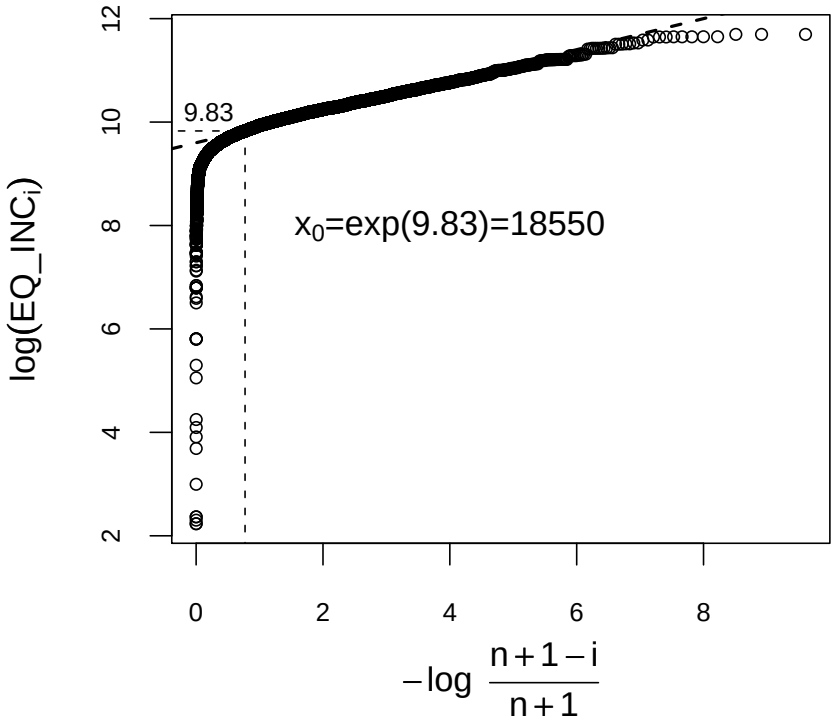

Figure 4.2: Pareto quantile plot

4.2.2 Mean excess plot

The second method to detect the threshold x_0 of the Pareto distribution is again a graphical method, called mean excess plot. The *excess function* is given by

$$e(x_0) = E(x - x_0 | x > x_0), \quad x_0 \geq 0 \quad .$$

First of all, the equivalised disposable income has to be sorted in increasing order. For the estimation of the excess function an operator z^+ is required, where z^+ is defined

as

$$z^+ = \max(z, 0) \quad .$$

The number of ordered observations, which are greater than the threshold x_0 are named k. There is a clear relationship between k and x_0, and k can be expressed as

$$k = \sum_{i=1}^{n} \mathbb{1}\{EQ\_INC_i > x_0\} \quad ,$$

where $\mathbb{1}$ is the indicator function.
The empirical function e_n, which can be expressed by

$$e_n(x_0) = \frac{1}{k} \sum_{i=1}^{n} (EQ\_INC_i - x_0)^+ \quad ,$$

is an estimate of the excess function.
For the mean excess plot, the values of $e_n(x_0)$ are drawn against EQ_INC_i. The Pareto distribution is valid for all points on the right-hand side of the mean excess plot, after the increasing values form a nearly straight line. A disadvantage of both, the Pareto quantile plot and the mean excess plot, is that the threshold cannot be determined exactly. So this methods should be used only for a first view. The mean excess plot is described in Borkovec and Klüppelberg (2000) and Vandewalle et al. (2007). An example of this plot with the Austrian EU-SILC 2006 data set is shown in Figure 4.3. The dashed lines indicates the straight line and the estimated point, where the Pareto distribution is valid.

4.2.3 Method of Van Kerm

In Van Kerm (2007) a cut-off point is given, which can be used as threshold x_0. The cut-off point can be determined as

$$x_0 = \max(\min(2.5\mu, Q(0.98)), Q(0.97)) \quad ,$$

where μ is the mean income and $Q(0.97)$ and $Q(0.98)$ are the quantiles of the equivalised income.
The cut-off point x_0 for the Austrian EU-SILC 2006 data set is 46515.31 €, for example.

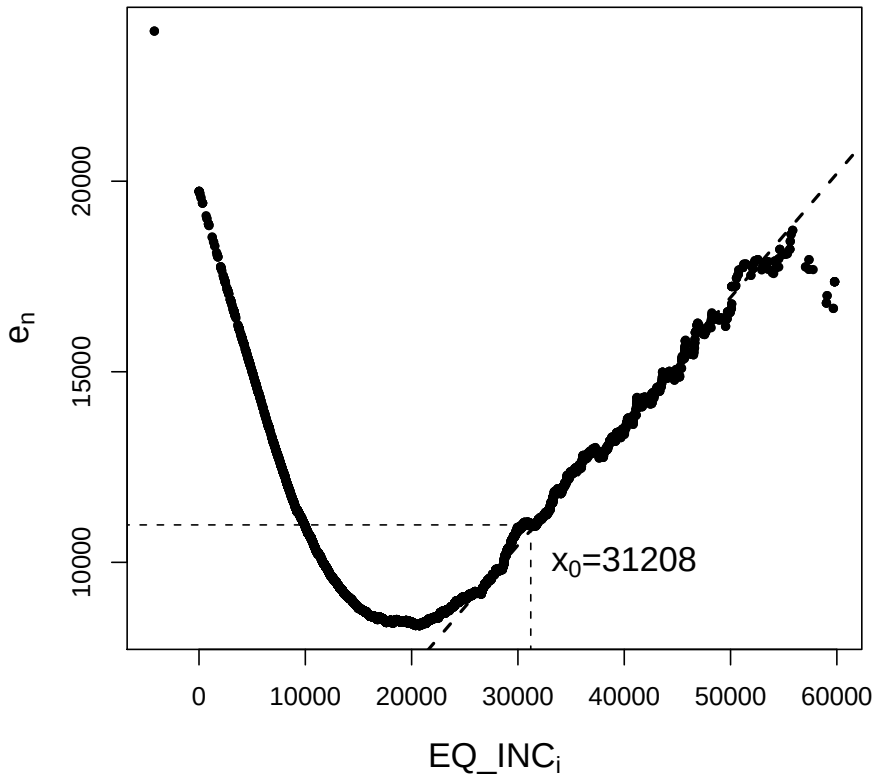

Figure 4.3: Mean excess plot

4.2.4 Prediction error criterion

The next method to detect outliers is the prediction error criterion (C - criterion). It can be found in Dupuis and Victoria-Feser (2006).
In the case of independent observations the prediction error criterion is given by

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\frac{Y_i - E[Y_i]}{\sigma_i}\right)^2\right] \quad , \tag{4.1}$$

where $Y = (Y_1, \ldots, Y_n)^T$ are the observations, $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$ the predicted values of Y and $\sigma_i^2 = Var(Y_i)$ are the variances of the n observations.
Criterion (4.1) is sensitive to outliers, and so weights $\hat{w}_i \in [0, 1]$ will be included. The

rescaled mean squared weighted prediction error can be written as

$$\begin{aligned}\Gamma_R &= \frac{1}{n}\sum_{i=1}^{n} E\left[\hat{w}_i^2\left(\frac{Y_i - E[Y_i]}{\sigma_i}\right)^2\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^{n}\hat{w}_i^2\left(\frac{Y_i-\hat{Y}_i}{\sigma_i}\right)^2\right] + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}Cov\left[\hat{w}_iY_i, \hat{w}_i\hat{Y}_i\right] - \\ &\quad \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}Var\left[\hat{w}_iY_i\right] \quad .\end{aligned}$$

An unbiased estimator of Γ_R is chosen, which is given by

$$C_R = \frac{1}{n}\sum_{i=1}^{n}\hat{w}_i^2\left(\frac{Y_i-\hat{Y}_i}{\sigma_i}\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}Cov\left[\hat{w}_iY_i, \hat{w}_i\hat{Y}_i\right] - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}Var\left[\hat{w}_iY_i\right] \quad , \tag{4.2}$$

with suitable estimators for σ_i^2, $Cov\left[\hat{w}_iY_i, \hat{w}_i\hat{Y}_i\right]$ and $Var\left[\hat{w}_iY_i\right]$.
To find suitable estimators of (4.2) the following results will be used. For a fixed threshold x_0 the ordered k quantiles $X^*_{[i]} \geq x_0$ have the density

$$\begin{aligned}f_{i;k}(x) &= (k-(i-1))\binom{k}{i-1}\left\{1-\left(\frac{x}{x_0}\right)^{-\theta}\right\}^{i-1}\left\{\left(\frac{x}{x_0}\right)^{-\theta}\right\}^{k-(i-1)}\theta x^{-1} \\ &= \frac{k+1-i}{x}\theta B\left[x=i-1; n=k; p=1-\left(\frac{x}{x_0}\right)^{-\theta}\right] \quad ,\end{aligned}$$

for $x \geq x_0$, where $B[x;n;p]$ is the value of the probability function of the Binomial distribution and the variance of Y_i is

$$\sigma_i^2 = \sum_{j=1}^{i}\frac{1}{\theta^2(k-i+j)^2} = \frac{1}{\theta^2}\left[\frac{1}{k^2}+\frac{1}{(k-1)^2}+\ldots+\frac{1}{(k+1-i)^2}\right] \quad . \tag{4.3}$$

The notation of the ordered largest k observations $X_i \geq x_0$ is $X^*_{[i]},\ i=1,\ldots,k$. Let

$$Y_i = \log\left(\frac{X^*_{[i]}}{x_0}\right) \quad , \tag{4.4}$$

and for the predicted values

$$\hat{Y}_i = -\frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right) \quad , \tag{4.5}$$

where $\hat{\theta}$ is the Hill estimator $\hat{\theta}_H = \left[\frac{1}{k}\sum_{i=1}^{k} \log X^*_{[i]} - \log x_0\right]^{-1}$. In this non-robust version all weights $\hat{w}_i$ equal 1. An estimator of Γ_R is given by

$$\begin{aligned}\hat{C}(x_0) &= \frac{\hat{\theta}^2}{k}\sum_{i=1}^{k}\left[\frac{1}{k^2}+\ldots+\frac{1}{(k+1-i)^2}\right]^{-1}\left[\log\left(\frac{X^*_{[i]}}{x_0}\right)+\frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right]^2 \\ &\quad +\frac{2}{k^2}\sum_{i=1}^{k}\left[\frac{1}{k^2}+\ldots+\frac{1}{(k+1-i)^2}\right]^{-1}\log\left(\frac{k+1-i}{k+1}\right)^2-1 \quad .\end{aligned}$$

A detailed derivation of the prediction error criterion and the proofs of the equations can be found in Dupuis and Victoria-Feser (2006). The ordered observations X_i, which minimise $\hat{C}(x_0)$ are considered for the estimation of the threshold x_0. All observations greater than x_0 can be modeled with the Pareto distribution.
In Table 4.4 some results of the prediction error criterion are shown, where k_{opt} is the estimated value of the highest observations, which are used for the Pareto distribution and $\hat{\theta}_H$ is the estimated Hill estimator of these observations. Figure 4.4 shows the mean squared prediction error for different numbers k, where the Pareto distribution is valid. The optimal k, k_{opt}, is given by the dashed line at 671. A more robust version

Table 4.4: Prediction error criterion

region	n	k_{opt}	x_0	$\hat{C}(x_0)$	$\hat{\theta}_H$
Austria	14883	671	37653.79	0.1212	3.88
Burgenland	549	51	28819.42	0.4134	3.18
Lower Austria	2820	510	26082.00	0.2180	3.67
Vienna	2334	271	33151.15	0.1011	3.69
Carinthia	1081	156	26346.71	0.2682	4.45
Styria	2305	24	50718.50	0.2630	6.59
Upper Austria	2817	26	59791.52	0.2305	4.42
Salzburg	924	35	34622.07	0.5509	2.56
Tyrol	1320	38	40040.76	0.0783	3.07
Vorarlberg	733	284	21130.80	0.3929	3.55

of the prediction error criterion can be obtained by a weight function and a robust estimation of the shape parameter θ of the Pareto distribution is used. Then Equation (4.2) can be used to estimate the robust prediction error. The robust criterion (4.2) can be written more generally. An estimator of $Var[\hat{w}_i Y_i]$ is

$$Var[\hat{w}_i Y] = E[\hat{w}_i^2 Y_i^2] - E[\hat{w}_i Y_i]^2 \quad ,$$

where

$$E[\hat{w}_i^j Y_i^j] = \int_{x_0}^{\infty} \hat{w}(x)^j \log\left(\frac{x}{x_0}\right)^j f_{i;k}(x)\mathrm{d}x \quad .$$

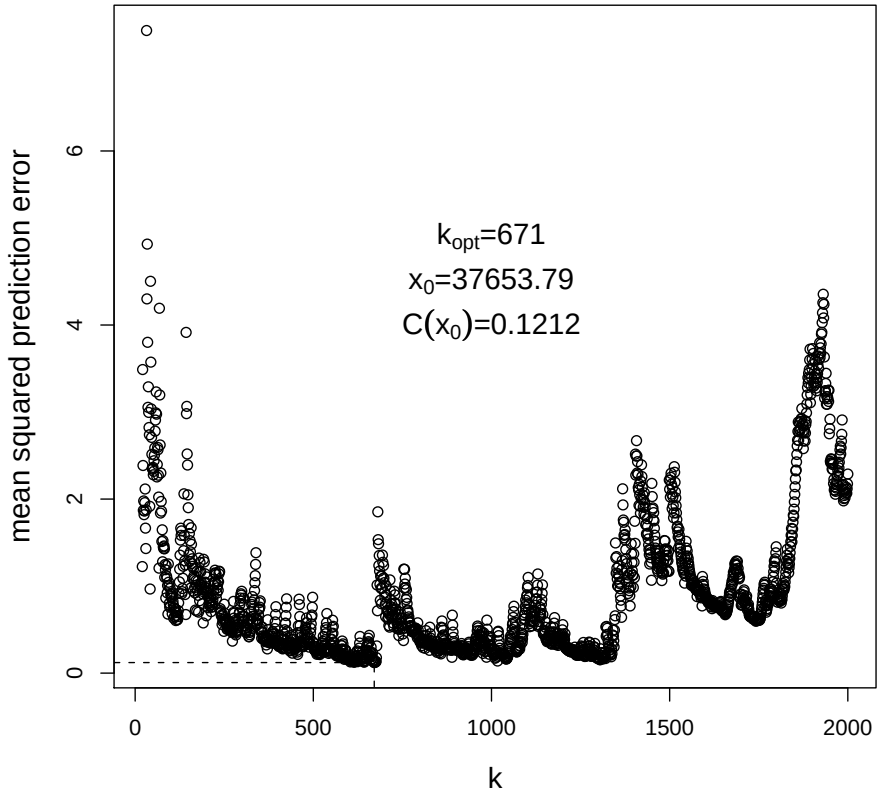

Figure 4.4: Mean squared prediction error criterion

An estimator of the $Cov[\hat{w}_i Y_i, \hat{w}_i \hat{Y}_i]$ is given by

$$Cov[\hat{w}_i Y_i, \hat{w}_i \hat{Y}_i] = E[\hat{w}_i^2 Y_i \hat{Y}_i] - E[\hat{w}_i Y_i] E[\hat{w}_i \hat{Y}_i] \quad ,$$

where

$$E[\hat{w}_i \hat{Y}_i] = -\int_{x_0}^{\infty} \frac{1}{\hat{\theta}} \log\left(\frac{k+1-i}{k+1}\right) \hat{w}(x) f_{i;k}(x) \mathrm{d}x \quad ,$$

and

$$E[\hat{w}_i^2 Y_i \hat{Y}_i] = -\int_{x_0}^{\infty} \frac{1}{\hat{\theta}} \log\left(\frac{k+1-i}{k+1}\right) \hat{w}(x) \log\left(\frac{x}{x_0}\right) f_{i;k}(x) \mathrm{d}x \quad .$$

In the same way as before, the robust prediction error $\hat{C}_R(x_0)$ is estimated. The difference to the non-robust method is, that robustness weights (w_i) and robust estimates

of the shape parameter of the Pareto distribution (θ) are used. The threshold x_0 which minimise $C_R(x_0)$ will be chosen.
Dupuis and Victoria-Feser (2006) noted, that $k = 20$ observations for the estimation of θ are sufficient. Both, $\hat{w}(x)$ and $f_{i;k}$ depend on θ, which is replaced by $\hat{\theta}$, where $\hat{\theta}$ can be estimated by robust methods. In Table 4.5 some results of the robust prediction error criterion are shown, where k_{opt} is the estimated value of of the highest observations, which are used for the Pareto distribution and $\hat{\theta}_{WM}$ is the estimated shape parameter of the Pareto distribution with the weighted maximum likelihood method of these observations. The weighted maximum likelihood estimator is described in Section 4.3.7. Figure 4.5 shows the mean squared prediction error for different numbers k. The optimal k, k_{opt}, is given by the dashed line at 82.

Table 4.5: Robust prediction error criterion

region	n	k_{opt}	x_0	$\hat{C}_R(x_0)$	$\hat{\theta}_{WM}$
Austria	14883	82	66191.68	0.1023	4.43
Burgenland	549	28	35463.32	1.7454	4.29
Lower Austria	2820	28	59018.25	0.6848	4.10
Vienna	2334	51	52108.52	1.9458	3.80
Carinthia	1081	27	40495.49	1.4875	10.20
Styria	2305	32	49955.00	1.3857	7.78
Upper Austria	2817	23	60656.44	1.1941	4.16
Salzburg	924	20	50503.74	0.4976	5.08
Tyrol	1320	36	41731.93	0.7837	3.43
Vorarlberg	733	21	41001.80	1.1424	5.49

The results of the robust prediction error criterion (C_R - criterion) is not so different for each region, so this method leads to better results than the non-robust prediction error criterion.
More details about the robust estimation of the shape parameter θ of the Pareto distribution can be found in Section 4.3.

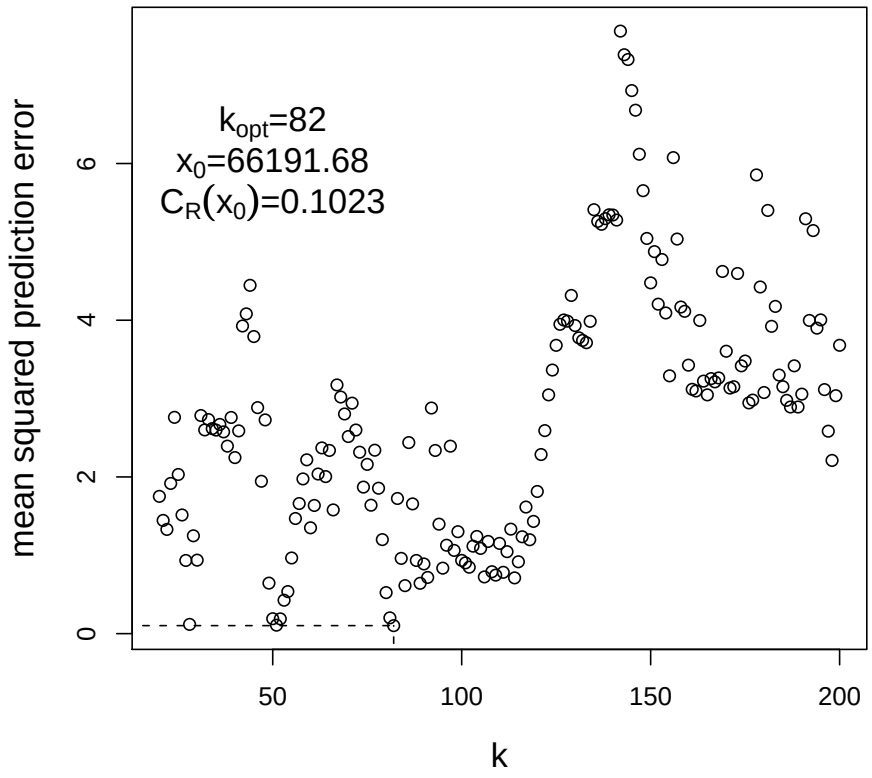

Figure 4.5: Robust mean squared prediction error criterion

4.2.5 Bootstrap method

This method can be found in Danielsson et al. (2000). It uses a bootstrap method to choose the sample fraction for the tail index estimation. First, the asymptotic mean squared error (AMSE) is used for this method, which is given for $\hat{\gamma}_n$ by

$$AMSE(n,k) = AsyE(\hat{\gamma}_n(k) - \gamma)^2 \quad ,$$

where n is the number of observations, k is the number of observations which are used for modeling the upper tail, $\hat{\gamma}_n$ is the estimated parameter, defined in Equation (4.6) and γ is the true parameter. The evaluation of the AMSE is done by using a bootstrap procedure. The aim is to minimise the estimated AMSE for finding the optimal k-value.
First n_1, where $n_1 < n$, resamples of size n_1 with replacement are drawn from the sample. Each resample must be ordered in increasing order ($X^*_{n_1,1} \leq \ldots \leq X^*_{n_1,n_1}$).

Then the inverse Hill estimator is defined as

$$\hat{\gamma}_{n_1}(k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1} \log X^*_{n_1,n_1-i+1} - \log X^*_{n_1,n_1-k_1} \quad . \tag{4.6}$$

The problem is to find a consistent estimator of the real parameter γ. Therefore a control variate will be constructed, which can be expressed as

$$\hat{M}_n(k) = \frac{1}{k}\sum_{i=1}^{k} (\log X_{n,n-i+1} - \log X_{n,n-k})^2 \quad .$$

$M_n(k)/(2\gamma_n(k))$ is a consistent estimator of γ.
Next, the estimate for the AMSE is given by

$$\hat{Q}(n_1,k_1) = E\left(\left(\hat{M}_{n_1}(k_1) - 2(\hat{\gamma}_{n_1}(k_1))^2\right)^2\right) \quad , \tag{4.7}$$

where $\hat{M}_{n_1}(k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1}(\log X^*_{n_1,n_1-i+1} - \log X^*_{n_1,n_1-k_1})^2$.
$\hat{Q}(n_1,k_1)$ is calculated for each possible value k_1. The estimated optimal value $\hat{k}_{1,0}(n_1)$ is the value which minimises the AMSE. The next step is to repeat this procedure for a smaller resample size n_2, where n_2 is given by

$$n_2 = \frac{n_1^2}{n} \quad .$$

The minimising AMSE for $Q(n_2,k_2)$ yields the optimal value of k_2, where the notation is $\hat{k}_{2,0}$. The optimal sample fraction for the tail index estimation can be expressed as

$$\hat{k}_0(n) = \frac{(\hat{k}_{1,0}(n_1))^2}{\hat{k}_{2,0}(n_2)}\left(\frac{(\log \hat{k}_{1,0}(n_1))^2}{(2\log n_1 - \log \hat{k}_{1,0}(n_1))^2}\right)^{\frac{\log n_1 - \log \hat{k}_{1,0}(n_1)}{\log n_1}} \quad .$$

The final step is to estimate γ by $\hat{\gamma}_n(\hat{k}_0)$ which is given by

$$\hat{\gamma}_n(\hat{k}_0) = \frac{1}{k}\sum_{i=1}^{k} \log X^*_{n,n-i+1} - \log X^*_{n,n-k} \quad .$$

If this method is used two tuning parameters have to be chosen, the number of bootstrap resamples and n_1. For the number of bootstrap samples a stopping criterion can be chosen, where the resampling is stopped if the change of the AMSE falls under a defined level. A good choice for n_1 is given by

$$n_1 = n^{1-\epsilon} \quad ,$$

for $0 < \epsilon < 0.5$. An estimator for the AMSE is the ratio

$$R(n_1) = \frac{(\hat{Q}(n_1, \hat{k}_{1,0}))^2}{\hat{Q}(n_2, \hat{k}_{2,0})} \quad ,$$

where $\hat{Q}(n_1, \hat{k}_{1,0})$ and $\hat{Q}(n_2, \hat{k}_{2,0})$ can be estimated by Equation (4.7). The sample n_1, where $R(n_1)$ is minimal, is chosen for the sample size n_1. The proofs and further details can be found in Danielsson et al. (2000). The shape parameter $\hat{\theta}$ of the Pareto distribution can be evaluated by

$$\hat{\theta} = \frac{1}{\hat{\gamma}_n(\hat{k}_0)} \quad .$$

This bootstrap method is developed for the Hill estimator, so it is non-robust. In Table 4.6 the results for Austria, and the federal states of Austria are shown, where k_{opt} is the estimated value of the highest observations, which are used for the Pareto distribution and $\hat{\theta}_H$ is the estimated Hill estimator of these observations. It can be seen, that the thresholds are very low and in addition to that, the computational costs are high. So this method gives poor results compared to the prediction error criterion.

Table 4.6: Bootstrap method for the estimation of the threshold

region	n	k_{opt}	x_0	$\hat{\theta}_H$
Austria	14883	658	37916.91	3.91
Burgenland	549	144	22500.8	3.52
Lower Austria	2820	159	35973.33	3.85
Vienna	2334	240	34394.87	3.75
Carinthia	1081	186	25237.3	4.39
Styria	2305	182	29916.15	4.26
Upper Austria	2817	667	24600.59	4.15
Salzburg	924	335	20274.33	3.65
Tyrol	1320	249	23542.01	3.48
Vorarlberg	733	226	22700.73	3.69

4.2.6 Weighted asymptotic mean squared error

This robust method is described in Beirlant et al. (2002), Beirlant et al. (1996b) and Beirlant et al. (1996a). It is a nonparametric estimate of the AMSE of the Hill estimator $\hat{\theta}_H$. The optimal value k is obtained, when the weighted mean squared error (WMSE) is minimal, which is given by

$$MSE_{opt}(k) = \frac{1}{k}\sum_{i=1}^{k} w_{i,k}^{opt}\left(\log\frac{X_{[i]}^*}{x_0} + \frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right)^2 \quad ,$$

where $w_{i,k}^{opt}$ is a sequence of weights. For the optimal weight $w_{i,k}^{opt}$ a linear combination of two sequences of weights $w_{i,k}^{(1)}$ and $w_{i,k}^{(2)}$ is used, which can be expressed as

$$w_{i,k}^{opt} = \delta_{1,k} w_{i,k}^{(1)} + \delta_{2,k} w_{i,k}^{(2)} \quad ,$$

where $\delta_{1,k}$ and $\delta_{2,k}$ are scaling constants that depend on k. The chosen weight functions are given by

$$\begin{aligned} w_{i,k}^{(1)} &= 1 \\ w_{i,k}^{(2)} &= \frac{k+1-i}{k+1} \quad . \end{aligned}$$

Finally, the MSE can be written as

$$\begin{aligned} MSE_{opt}(k) &= \frac{1}{k}\sum_{i=1}^{k}\left(\delta_{1,k} + \delta_{2,k}\left(\frac{k+1-i}{k+1}\right)\right)\left[\log\frac{X_{[i]}^*}{x_0} + \frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right]^2 \\ &= \frac{\delta_{2,k}}{k(k+1)}\sum_{i=1}^{k}(k+1-i)\left(\log\frac{X_{[i]}^*}{x_0} + \frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right)^2 \\ &\quad + \frac{\delta_{1,k}}{k}\sum_{i=1}^{k}\left(\log\frac{X_{[i]}^*}{x_0} + \frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right)^2 \quad . \end{aligned}$$

The values of the constants $\delta_{1,k}$ and $\delta_{2,k}$ can be obtained numerically for simple choices of the weights $w_{i,k}^{(1)}$ and $w_{i,k}^{(2)}$, which are given by

$$\begin{aligned} \delta_{1,k} &= \frac{b_k^{(2)} - a_k^{(2)}}{a_k^{(1)}b_k^{(2)} - b_k^{(1)}a_k^{(2)}} \\ \delta_{2,k} &= \frac{a_k^{(1)} - b_k^{(1)}}{a_k^{(1)}b_k^{(2)} - b_k^{(1)}a_k^{(2)}} \quad , \end{aligned}$$

where $a_k^{(j)}$ and $b_k^{(j)}$ $(j = 1, 2)$ are corresponding scaling constants. They can be expressed as

$$\begin{aligned} a_k^{(j)} &= \frac{1}{k}\sum_{i=1}^{k} w_{i,k}^{(j)} \left(\left(\frac{i}{k+1} \right)^{-1} - 1 \right) \\ b_k^{(j)} &= (1-\rho)^2 \frac{1}{k}\sum_{i=1}^{k} w_{i,k}^{(j)} \left(\frac{\left(\frac{i}{k+1}\right)^{-\rho} - 1}{\rho} \right)^2 \quad , \end{aligned}$$

where ρ is a nuisance parameter, which is set to -1. The parameter ρ is an index of regular variation, which is negative. For the computation of this estimator an algorithm can be found in Caers et al. (1998).
Another possible weight function is given in Beirlant et al. (1996a) and can be expressed by

$$\begin{aligned} w_{i,k}^{(1)} &= \frac{1}{k+1} \\ w_{i,k}^{(2)} &= -\log \frac{i}{k+1} \quad . \end{aligned}$$

For the Austrian EU-SILC 2006 this weight function leads to better results. Table 4.7 highlights the results of this method with robust estimation of the shape parameter θ.

Table 4.7: Estimation of the threshold with the WMSE

region	n	k_{opt}	x_0	$\hat{\theta}_{WM}$
Austria	14883	129	59672.37	4.49
Burgenland	549	54	28813.41	3.38
Lower Austria	2820	93	41713.52	4.00
Vienna	2334	87	45405.36	3.98
Carinthia	1081	28	40495.49	10.22
Styria	2305	20	54529.71	9.69
Upper Austria	2817	89	39849.45	3.57
Salzburg	924	87	29489.25	3.72
Tyrol	1320	86	31950.53	3.27
Vorarlberg	733	64	33733.32	4.99

Figure 4.6 shows the weighted mean squared error (MSE_{opt}) for different choices of k. The optimal k, k_{opt}, which is the value, where the mean squared error is minimal, is given by the dashed line at 129. A disadvantage of this method is, that there are many parameters that have to be estimated and also a tuning constant must be chosen, which are important to obtain reasonable results.

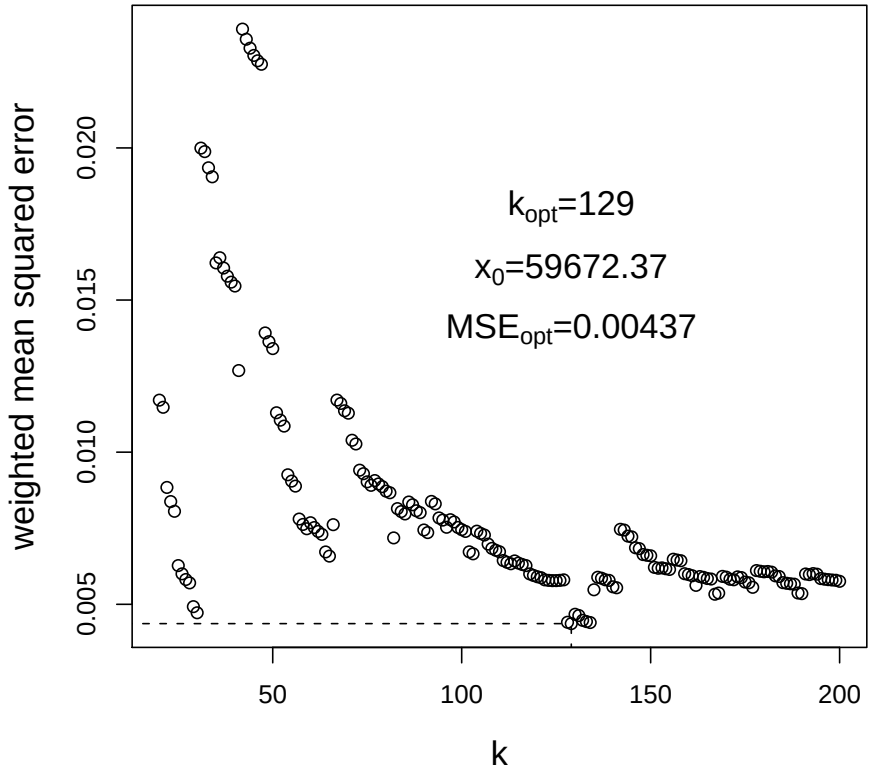

Figure 4.6: Weighted mean squared error for estimation of the threshold

Other non-robust methods for the estimation of the threshold x_0 are published in Drees and Kaufmann (1998), Beirlant et al. (1999), Guillou and Hall (2001), Vandewalle et al. (2004) and Beirlant et al. (2005).

4.3 Methods for the estimation of the parameters

In Section 2.1.2 the maximum likelihood estimator of the scale parameter θ of the Pareto distribution was mentioned. The aim of this Section is to find more robust methods to estimate the scale parameter θ.

4.3.1 QQ estimator

The description of this method for the estimation of the location parameter θ of the Pareto distribution can be found in Kratz and Resnick (1995). The procedure is similar to the Pareto quantile plot in Section 4.2.1, where the points above the threshold x_0 should lay on a straight line. First, the n observations have to be sorted in increasing order. The QQ-estimator is defined as

$$\hat{\theta}_{QQ} = \left[\frac{\sum_{i=1}^{k} -\log\left(\frac{i}{k+1}\right) \left[k \log(x_{n+1-i}) - \sum_{i=1}^{k} \log(x_{n-k+i}) \right]}{k \sum_{i=1}^{k} \left(\log\left(\frac{i}{k+1}\right)\right)^2 - \left(\sum_{i=1}^{k} \log\left(\frac{i}{k+1}\right) \right)^2} \right]^{-1} \quad ,$$

with k the amount of observations above x_0.
The QQ-estimator is not robust and the results can be dramatically influenced by outliers.

4.3.2 Pickands estimator

Another non-robust estimator of the shape parameter θ of the Pareto distribution is the Pickands estimator, which can be found in Pickands (1975). This estimator is defined as

$$\hat{\theta}_P = \left[\frac{1}{\log 2} \log \left(\frac{x_{n-\lfloor k/4 \rfloor} - x_{n-\lfloor k/2 \rfloor}}{x_{n-\lfloor k/2 \rfloor} - x_{n-k}} \right) \right]^{-1} \quad ,$$

where x_i are the sorted observations in increasing order and $\lfloor z \rfloor$ denotes the largest integer less than or equal to z.

4.3.3 Moment estimator

This method can be found in Dekkers et al. (1989). Similarly as before, the observations have to be ordered in increasing order. For the estimation, two variables has to be constructed. First, the inverse Hill estimator is expressed by

$$M_n^{(1)} = \frac{1}{k} \sum_{i=1}^{k} \log x_{n-k+1} - \log x_{n-k} \quad ,$$

and the second one can be expressed as

$$M_n^{(2)} = \frac{1}{k}\sum_{i=1}^{k}(\log x_{n-k+1} - \log x_{n-k})^2 \quad .$$

The moment estimator is then defined as

$$\hat{\theta}_M = \left[M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1}\right]^{-1} \quad .$$

Again this method is also non-robust against outlying observations.

4.3.4 Least squares estimator

This method can be found in Brazauskas and Serfling (2000a) and Brazauskas and Serfling (2000b) and is also a non-robust method. To reduce the influence of outlying observations the logarithm of the observations ($Z_i = \log X_i$) is used. First of all, the least squares regression estimators are defined as

$$p_i = \begin{cases} \frac{i}{k} & \text{, if} \quad 1 \leq i \leq k-1 \\ \frac{k}{k+1} & \text{, if} \quad i = k \quad . \end{cases}$$

The next step is to define

$$\begin{aligned} c_i &= -\log(1-p_i) \\ \overline{c}_k &= \frac{1}{k}\sum_{i=1}^{k} c_i \quad . \end{aligned}$$

Then the shape parameter θ of the Pareto distribution is given by

$$\hat{\theta}_{LS} = \left(\frac{\frac{1}{k}\sum_{i=}^{k} c_i Z_i - \overline{c}_k \overline{Z}_k}{\frac{1}{k}\sum_{i=}^{k} c_i^2 - (\overline{c}_k)^2} \right)^{-1} \quad ,$$

where $\overline{Z}_k = \frac{1}{k}\sum_{i=1}^{k} Z_i$.
With the least squares (LS) method the threshold x_0 can be estimated as well. It can be expressed as

$$\hat{x}_{0_{LS}} = e^{\overline{Z}_k - \hat{\theta}_{LS}^{-1}\overline{c}_k} \quad .$$

4.3.5 Generalised median

This more robust method can be found in Brazauskas and Serfling (2000a) and Brazauskas and Serfling (2000b). Again, k observations, which form amount of observations above x_0 are used. For a choice of the integer k_1 (kernel size) ($k > k_1 \geq 2$) a kernel is defined as

$$h_0(x_1, \ldots, x_{k_1}) = \left[\frac{1}{k_1} \sum_{i=1}^{k_1} \log x_i - \log x_0 \right]^{-1} \quad .$$

To make h_0 *median unbiased* for the estimation of the shape parameter θ, the fact that

$$\frac{2k_1\theta}{h_0(x_1, \ldots, x_{k_1})}$$

has cumulative distribution function (cdf) $\chi^2_{2k_1}$ is used, where $\chi^2_{2k_1}$ denotes the chi-square distribution with $2k_1$ degrees of freedom. With this assumption it follows that the kernel

$$h(x_1, \ldots, x_{k_1}) = \frac{M_{2k_1}}{2k_1} h_0(x_1, \ldots, x_{k_1})$$

is median unbiased, where M_{2k_1} is a multiplicative correction factor. The procedure is now, to generate all $\binom{k}{k_1}$ subsets of kernel size k_1. This yields the generalised median estimator, which is given by

$$\hat{\theta}_{GM} = med(h(x_{i_1}, \ldots, x_{i_{k_1}})) \quad ,$$

where $i = 1, \ldots, \binom{k}{k_1}$. This method has a high efficiency and is highly robust, if k_1 is between 2 and 10. For kernel sizes from 2 to 10 the multiplicative correction factors (M_{2k_1}) are given in Table 4.8.

Table 4.8: Multiplicative correction factor of the generalised median estimator method

k_1	2	3	4	5	6
M_{2k_1}	3.3567	5.3481	7.3441	9.3418	11.3403
k_1	6	7	8	9	10
M_{2k_1}	11.3403	13.3393	15.3385	17.3379	19.3374

The problem of this method is, that for computing the generalised median estimator the the computational cost is $O(k^{k_1})$, which makes it impossible to calculate it for large k.

4.3.6 Trimmed median

This method is also described in Brazauskas and Serfling (2000a) and Brazauskas and Serfling (2000b). First of all, the observations have to be sorted in increasing order. The method needs two trimming proportions, β_1 and β_2, where $0 \leq \beta_1 < 1$ and $0 \leq \beta_2 < 1 - \beta_1$. β_1 is the proportion for the lower part and β_2 is used for the upper part of the ordered sample. Then the trimmed mean estimator is defined as

$$\hat{\theta}_{TM} = \left[\sum_{i=1}^{k} c_{k,i} (\log x_{n-k+i} - \log x_{n-k}) \right]^{-1} \quad ,$$

where the mean unbiasing factors $c_{k,i}$ are given by

$$c_{k,i} = \begin{cases} 0 & , \text{if} \quad 1 \leq i \leq \lfloor k\beta_1 \rfloor, \\ \frac{1}{d(\beta_1, \beta_2, k)} & , \text{if} \quad \lfloor k\beta_1 \rfloor + 1 \leq i \leq k - \lfloor k\beta_2 \rfloor, \\ 0 & , \text{if} \quad k - \lfloor k\beta_2 \rfloor + 1 \leq i \leq k \quad . \end{cases}$$

$\lfloor z \rfloor$ denotes the largest integer less than or equal to z and $d(\beta_1, \beta_2, k)$ is defined as

$$d(\beta_1, \beta_2, k) = \sum_{j=\lfloor k\beta_1 \rfloor + 1}^{k - \lfloor k\beta_2 \rfloor} \sum_{i=1}^{j} \frac{1}{k - i + 1} \quad ,$$

such that $\hat{\theta}_{TM}$ is mean unbiased.

4.3.7 Weighted maximum likelihood estimator

This method is a robust one, which estimates the parameter θ of the Pareto distribution for a given value of the threshold x_0 and can be found in Cowell and Victoria-Feser (2003). The weighted maximum likelihood (WMLE) approach of Dupuis and Morgenthaler (2002) is implemented. For the estimation the $k = \sum_{i=1}^{n} \mathbb{1}\{X_{[i]} \geq x_0\}$ largest observations will be used. The method is an M-estimator, which is defined as

$$\sum_{i=1}^{k} \Psi(X^*_{[i]}; \theta) = 0 \quad ,$$

where $\hat{\theta}_{WM}$ is the solution. The function Ψ is expressed as

$$\Psi(x; \theta) = w(x; \theta) \frac{\partial}{\partial \theta} \log f(x; \theta) = w(x; \theta) \left(\frac{1}{\theta} - \log \frac{x}{x_0} \right) \quad ,$$

if the Pareto distribution (Section 2.1.2) is used and $w(x; \theta)$ is a weight function with values in $[0, 1]$. Dupuis and Morgenthaler (2002) also defined the bias-corrected

WMLE, which depends on the model and the choice of the weight function. The bias-corrected WMLE ($\tilde{\theta}$) is given by

$$\tilde{\theta}_{WM} = \hat{\theta}_{WM} - B(\hat{\theta}_{WM}) \quad ,$$

where $B(\hat{\theta}_{WM})$ is given by

$$B(\hat{\theta}_{WM}) = -\frac{\int (w(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta))|_{\hat{\theta}} \mathrm{d}F_{\hat{\theta}}(x)}{\int (\frac{\partial}{\partial\theta}w(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta) + w(x;\theta)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta))|_{\hat{\theta}} \mathrm{d}F_{\hat{\theta}}(x)} \quad .$$

Dupuis and Victoria-Feser (2006) describe two possible weight functions. The first one is a probability based weighting where $w(x;\theta)$ is given by

$$w_F(x;\theta) = \begin{cases} \frac{F_\theta(x)}{p_1} & \text{, if} \quad F_\theta(x) < p_1, \\ 1 & \text{, if} \quad p_1 < F_\theta(x) < 1 - p_2, \\ \frac{1-F_\theta(x)}{p_2} & \text{, if} \quad F_\theta(x) > 1 - p_2 \quad , \end{cases}$$

where $F_\theta(x)$ is the empirical distribution function and p_1 and p_2 are constants, which regulate the robustness properties.
If the Pareto distribution is valid, the bias correction term $B(\theta)$ for the weight function $w_F(x;\theta)$ can be computed easily and is given by

$$B(\theta) = \left(\frac{\theta}{2}\right) \frac{2(1-p_1)^2\log(1-p_1) + p_1(1-p_1) + p_1(1-p_2) + 2p_1p_2\log p_2}{[(1-p_1)\log(1-p_1)]^2 - p_1(1-p_1) - p_1(1-p_2) + p_1p_2(\log p_2)^2} \quad .$$

A good choice of p_1 and p_2 is 0.005, this value indicates an efficiency of about 95%. If the values of p_1 and p_2 are larger, the estimator of θ is more robust, but less efficient.
An other possibility for a weight function is to use the standardised residual r_i, which is given by

$$r_i = \frac{Y_i - \hat{Y}_i}{\sigma_i} \quad ,$$

with Y_i as in (4.4), $\hat{Y}_i$ as in (4.5), σ_i as in (4.3) and $\hat{\theta}$ is the WMLE. The weight function $w_R(x;\theta)$ is defined as a Huber type approach which is given by

$$w_R(X^*_{[i]};\theta) = \begin{cases} 1, & \text{if} \quad |r_i| < c, \\ \frac{c}{|r_i|} & \text{if} \quad |r_i| > c \quad , \end{cases}$$

where c is a constant which is used for regulating the robustness. In Figure 4.7 the weight function can be seen for a tuning constant c=2.5. If the residuals are large, the weights go to 0.
If the Pareto distribution is valid, θ is the WMLE, and for the weight function $w_R(X_{[i]}*;\theta)$ above the approximate bias correction term $B(\theta)$ is equal to

$$-\frac{\sum_{i=1}^{k}(w(X^*_{[i]};\theta)\frac{\partial}{\partial\theta}\log f(X^*_{[i]};\theta))|_{\hat{\theta}}(F_{\hat{\theta}}(X^*_{[i]}) - F_{\hat{\theta}}(X^*_{[i-1]}))}{\sum_{i=1}^{k}(\frac{\partial}{\partial\theta}w(X^*_{[i]};\theta)\frac{\partial}{\partial\theta}\log f(X^*_{[i]};\theta) + w(X^*_{[i]};\theta)\frac{\partial^2}{\partial\theta^2}\log f(X^*_{[i]};\theta))|_{\hat{\theta}}(F_{\hat{\theta}}(X^*_{[i]}) - F_{\hat{\theta}}(X^*_{[i-1]}))},$$

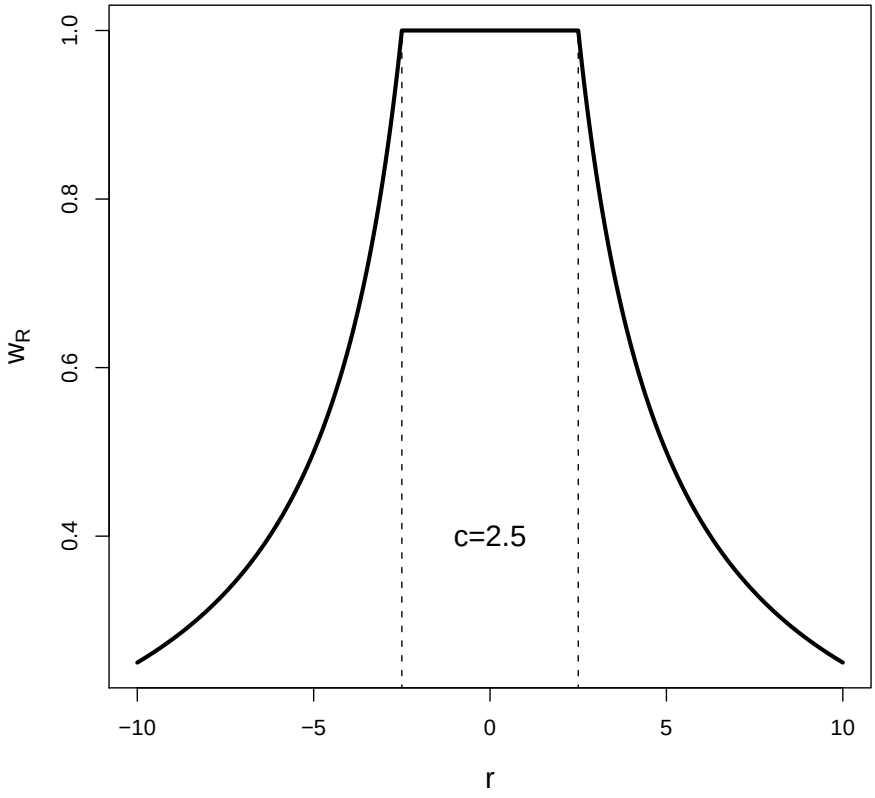

Figure 4.7: Weight function of standardised residuals (c=2.5)

where $X^*_{[0]}$ is set to the threshold x_0.
The tuning constant c depends on the data set. A good choice of c is a value in the interval $[1, 2.5]$. If a large value of c is used, the efficiency is smaller.

4.3.8 Integrated squared error (ISE) estimator

This method is described in Vandewalle et al. (2007) and is based on the relative excesses Y_{x_0} over a threshold x_0 which are given by

$$Y_{x_0} = \left(\frac{X}{x_0} > y \middle| X > x_0 \right) \quad ,$$

where $X = \{X_1, \ldots, X_n\}$ are independent and identically distributed (i.i.d) random variables. The distribution function can be expressed as

$$F_{Y_{x_0}}(y) = P\left(\frac{X}{x_0} > y \middle| X > x_0\right) \quad .$$

Here we use the equivalised disposable income (EQ_INC) for the variable X. First of all, the equivalised disposable income has to be sorted in increasing order. This method is developed for Pareto-type distributions. The density function of the Pareto distribution for the relative excesses is given by

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

This method estimates the parameter θ which brings the estimated density $f(y|\hat{\theta})$ close to the true unknown density $f(y)$. The integrated squared distance criterion is used as the minimum distance criterion. It can be expressed as

$$\begin{aligned} \hat{\theta} &= \arg\min_\theta \left[\int (f(y|\theta) - f(y))^2 \mathrm{d}y\right] \\ &= \arg\min_\theta \left[\int f^2(y|\theta)\mathrm{d}y - 2\int f(y|\theta)f(y)\mathrm{d}y + \int f^2(y)\mathrm{d}y\right] \quad . \end{aligned} \tag{4.8}$$

The middle term is the expected value of the density $f(Y|\theta)$ and the last term does not depend on θ. So Equation (4.8) can be written as

$$\hat{\theta} = \arg\min_\theta \left[\int f^2(y|\theta)\mathrm{d}y - 2E[f(Y|\theta)]\right] \quad .$$

If the empirical mean is used as an unbiased estimator of $E[f(Y|\theta)]$, the integrated squared error estimator is derived, which is given by

$$\hat{\theta}_{k,ISE} = \arg\min_\theta \left[\int f^2(y|\theta)\mathrm{d}y - \frac{2}{k}\sum_{j=1}^{k} f(Y|\theta)\right] \quad .$$

4.3.9 Partial density component (PDC) estimator

If the model is an incomplete mixture model $\omega f(.|\theta)$, the partial density component (PDC) estimator could be used. This method can be found in Vandewalle et al. (2007). The PDC estimator is given by

$$\hat{\theta}_k^\omega = \arg\min_{\theta,\omega} \left[\omega^2 \int f^2(y|\theta)\mathrm{d}y - \frac{2\omega}{k}\sum_{j=1}^{k} f(Y|\theta)\right] \quad ,$$

where two parameters, ω and θ, have to be estimated. $1-\omega$ is the estimated part of outliers, ω can be estimated as follows

$$\hat{\omega}_k = \frac{\frac{1}{k}\sum_{j=1}^{k} f(Y_j|\hat{\theta}_k)}{\int f^2(y|\hat{\theta}_{k,P})\mathrm{d}y} \quad .$$

The partial density components estimator $\hat{\theta}_k^{\omega}$ is given by

$$\hat{\theta}_{k,PDC}^{\omega} = \arg\min_{\theta,\omega} \left[\omega^2 \int_1^{\infty} f_\theta^2(y)\mathrm{d}y - \frac{2\omega}{k}\sum_{j=1}^{k} f_\theta(Y_{jk}) \right] \quad ,$$

where $Y_{jk} = X_{n-j+1}/x_0$ for $j = 1,\ldots,k$ are the relative excesses. The integral $\int_1^{\infty} f_\theta^2(y)\mathrm{d}y$ can be calculated for the Pareto distribution in closed form as

$$\int_1^{\infty} f_\theta^2(y)\mathrm{d}y = \frac{\theta^2}{2\theta+1} \quad .$$

More details can be found in Scott (2004).

4.3.10 Optimal bias robust estimator (OBRE)

This robust estimator of the shape parameter of the Pareto distribution θ can be found in Victoria-Feser and Ronchetti (1994), Dupuis (1997) and Dupuis and Field (1998).
The OBRE $\hat{\theta}_{OB}$ is an M-estimator, which is the solution of the following expression

$$\sum_{i=1}^{k} \Psi(x;\theta) = 0 \quad ,$$

see Section 4.3.7. The score function Ψ is given by

$$\Psi(x;\theta) = (s(x;\theta) - a(\theta))W_c(x;\theta) \quad ,$$

where $s(x;\theta)$ for the Pareto distribution can be expressed as

$$s(x;\theta) = \frac{1}{\theta} - \log\frac{x}{x_0} \quad ,$$

which is the score function of Section 4.3.7. The weights W_c are given by

$$W_c(x;\theta) = \min\left\{1; \frac{c}{\parallel A(\theta)[s(x-a(\theta)] \parallel}\right\} \quad , \tag{4.9}$$

where $\| \cdot \|$ denotes the Euclidean norm, and the matrix $A(\theta)$ and vector $a(\theta)$ are defined implicitly by

$$\begin{aligned} E[\Psi(x;\theta)\Psi'(x;\theta)] &= [A(\theta)'A(\theta)]^{-1} \\ E[\Psi(x;\theta)] &= 0 \quad . \end{aligned}$$

The weights in Equation (4.9) are downweight extreme values and the constant c is a robustness parameter. The lower c the more robust is the OBRE but also less efficient. A value of $c = 2$ is a good choice.
For the estimation of the shape parameter θ of the Pareto distribution with different methods, the value of k is chosen from the prediction error criterion in Section 4.2.4 for the Austrian EU-SILC sample 2006 as $k = 671$. The results are shown in Table 4.9.

Table 4.9: Estimation of θ with different methods

method	$\hat{\theta}$
Hill estimator	3.88
QQ estimator	3.83
Pickands estimator	4.41
Moment estimator	4.52
LS estimator	4.07
Trimmed median estimator	3.89
WMLE with residuals	3.88
WMLE with distribution function	3.83
ISE estimator	3.88
PDC estimator	3.89

4.4 Simulation study

The simulation study compares the different methods for the estimation of the threshold and is designed as in Dupuis and Victoria-Feser (2003).
First, the Burr distribution has to be considered and samples of sizes 500 and 1500 are drawn from this distribution. The distribution function is given by

$$F(x) = 1 - (1 + x^{-\rho})^{1/\rho} \quad ,$$

for some parameter $\rho < 0$. The value of ρ is chosen as -1. For this chosen value, the shape parameter θ of the Pareto distribution is 1, because the Burr distribution is a generalisation of the Pareto distribution. The Burr distribution is also called Singh-Maddala distribution.

The results of the estimation of θ can be seen in Table 4.10, where the best results are reached with the robust prediction error criterion, because the results are closest to 1.

Table 4.10: Estimation of θ and the threshold x_0 with a Burr distribution

method	n	k	$\hat{\theta}$	x_0
van Kerm	500	15	0.87	28.70
C - criterion	500	46	0.83	7.01
C_R - criterion	500	21	0.91	18.95
bootstrap method	500	109	0.87	21.07
MSE_{opt}	500	78	0.85	5.66
van Kerm	1500	45	0.82	32.52
C - criterion	1500	35	0.74	38.71
C_R - criterion	1500	84	0.94	18.94
bootstrap method	1500	607	0.79	1.51
MSE_{opt}	1500	154	0.82	7.87

In Figure 4.8 the results of the simulation, which should indicate the shape parameter θ, are presented via boxplots. The dashed line indicates the true shape parameter $\theta = 1$. For the threshold the result of the C_R - criterion of Table 4.10 with sample size 500 (k=21) was chosen. All methods give good results. However, the problem of the PDC method (see Section 4.3.9) is, that PDC is implemented for the Pareto distribution, but in this simulation study the Burr distribution is used. To reach better results, the method must be adapted also for the Burr distribution.
Another simulation study is carried out by using a mixture distribution with Pareto- and triangular distributions. The generated data include $\lfloor \alpha n \rfloor$ randomly drawn values from a Pareto distribution with parameters $x_0 = 2.5$ and $\theta = 1$ and $n - \lfloor \alpha n \rfloor$ values from a triangular distribution, where $\lfloor x \rfloor$ is the interger part of x and $\alpha = 0.1$. The density of the triangular distribution is given by

$$f(x) = 2(x - 1.5) \quad \text{for} \quad 1.5 \leq x \leq 2.5 \quad .$$

The results of the simulation study are shown in Table 4.11. Also for this mixed distribution, the robust prediction error criterion (C_R - criterion) leads to the best results.
In Figure 4.9 the 50 values of the Pareto distribution in the upper tail were multiplied with 10. So the methods should yield results around 50, which is indicated with the dashed line as the threshold. Both robust methods, the robust prediction error criterion (C_R - criterion) and the weighted asymptotic mean squared error method (MSE_{opt}), results in good estimates of this threshold.
In both simulations the estimated θ should be close to 1. The conclusion of the simulation study is that, in general, reasonable results are obtained, but the results

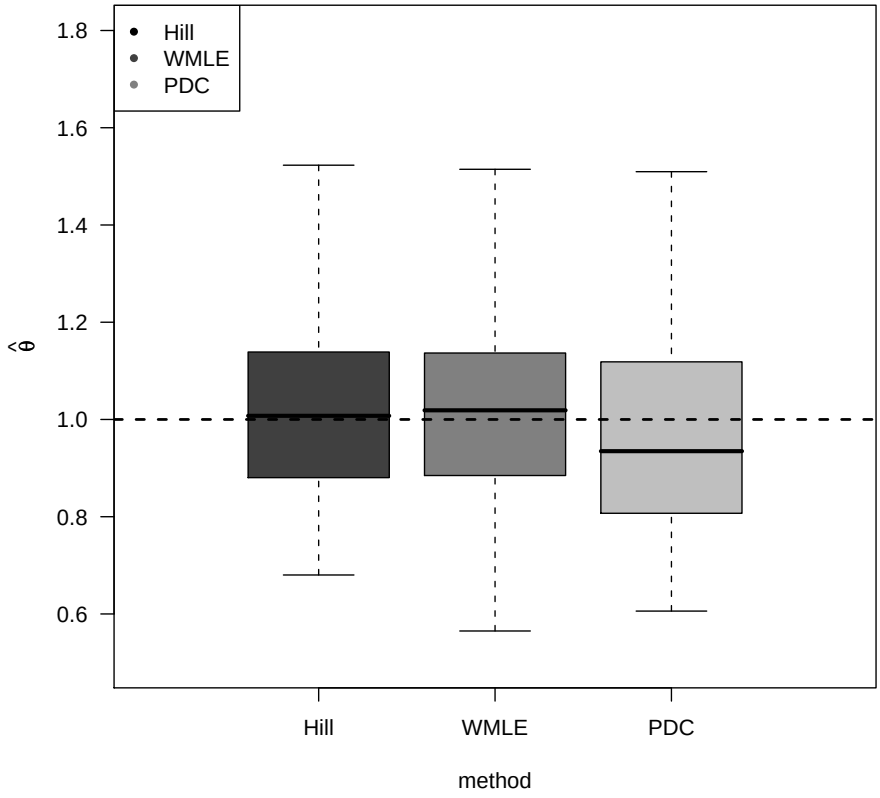

Figure 4.8: Boxplots of the shape parameter θ of the Burr distribution

Table 4.11: Estimation of θ and the threshold x_0 with a mixed distribution

method	n	k	$\hat{\theta}$	x_0
van Kerm	500	15	0.81	8.39
C - criterion	500	24	0.86	4.34
C_R - criterion	500	30	0.92	3.68
bootstrap method	500	489	2.62	1.66
MSE_{opt}	500	29	0.90	3.72

depend extremely on the parameters of the methods. If bad parameter values are chosen, also the results are poor.

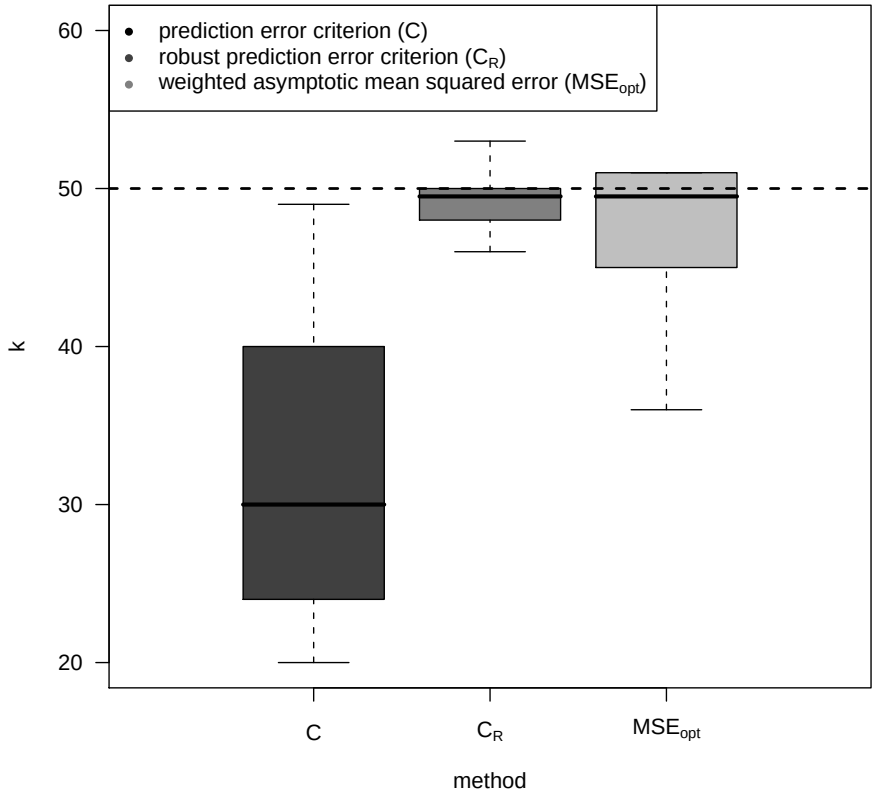

Figure 4.9: Boxplots of the location parameter x_0 for the mixed distribution

4.5 Non-parametric approach

4.5.1 Trimming

A lot of information about this topic can be found in Cowell and Victoria-Feser (2006). The idea is, that a proportion of originally observed data are removed to make the Laeken indicators more robust. The fixed percentage of data can be removed at either or both ends of the income distribution. In Van Kerm (2007) trimming percentages of 0.25%, 0.5%. 0.75% and 1%, both one-sided and two-sided, are used.
In Table 4.12 the results of some trimming percentages are shown. The problem of trimming is, that the trimmed indicator depends extremely on the sample. The percentages of trimming for good results are different in each case.
A procedure to estimate a confidence interval of the Gini coefficient is described in Cowell and Victoria-Feser (2003). The notation of the proportions, which are removed from the bottom and the top of the income distribution, is $\underline{\alpha}$ and $1 - \overline{\alpha}$. The total

Table 4.12: Trimming

$\underline{\alpha}$	$\overline{\alpha}$	$\hat{G}$	$\hat{Q}$
0	0	25.33	3.65
0.5%	0.5%	25.32	3.65
1%	1%	25.32	3.65
0%	0.5%	24.15	3.48
0%	1%	23.60	3.40
0.5%	1%	25.32	3.65
1%	0.5%	25.32	3.65

trimmed proportion is defined as

$$\alpha = \underline{\alpha} + (1 - \overline{\alpha}) \quad .$$

Furthermore the trimmed mean of an ordered, untrimmed sample $x_{[i]}$ with n observations is given by

$$\hat{\mu}_\alpha = \frac{1}{n(1-\alpha)} \sum_{i=\kappa(n,\underline{\alpha})}^{\kappa(n,\overline{\alpha})} x_{[i]} \quad ,$$

where $\kappa(n, \underline{\alpha})$ and $\kappa(n, \overline{\alpha})$ can be expressed as

$$\kappa(n, q) = \lfloor (n-1)q + 1 \rfloor \quad ,$$

where $\lfloor z \rfloor$ denotes the largest integer less than or equal to z. Additionally, n_α, the trimmed sample size, is defined as

$$n_\alpha = n(1-\alpha) \quad .$$

Then a $100(1-\beta)\%$ confidence interval for the Gini coefficient of an ordered trimmed sample $x_{[1]}, \ldots, x_{[n_\alpha]}$ is given by

$$\left(1 - \frac{2}{\hat{\mu}_\alpha n_\alpha^2} \sum_{i=1}^{n_\alpha} \sum_{j=1}^{i} x_{[j]} \pm z_{\beta/2} \sqrt{\frac{4\hat{\vartheta}_\alpha}{n_\alpha (1-\alpha) \hat{\mu}_\alpha^4}} \right) \quad ,$$

where $z_{\beta/2}$ is the quantile of the standard normal distribution $N(0,1)$ and the asymptotic variance ϑ_α can be expressed as

$$\vartheta_\alpha = \frac{1}{(1-\alpha)^2 \hat{\mu}_\alpha^4} [\mu_\alpha^2 \overline{\omega}_{qq'} + c_{\alpha,q} c_{\alpha,q'} \overline{\omega}_{\overline{\alpha}\overline{\alpha}} - \mu_\alpha c_{\alpha,q} \overline{\omega}_{q'\overline{\alpha}} - \mu_\alpha c_{\alpha,q'} \overline{\omega}_{q\overline{\alpha}}] \quad .$$

Therefore the empirical income cumulation $c_{\alpha,q}$ can be estimated by

$$\hat{c}_{\alpha,q} = \frac{1}{n_\alpha} \sum_{i=\kappa(n,\underline{\alpha})}^{\kappa(n,q)} x_{[i]}$$

and $\overline{\omega}_{q_i q_j}$ can be estimated by

$$\begin{aligned}\hat{\overline{\omega}}_{q_i q_j} &= \Big[q_i x_i - \underline{\alpha} x_{[1]} - (1-\alpha)\sum_{k=1}^{i} x_{[k]}\Big] \cdot \\ &\Big[(1-q_j)x_{[j]} - (1-\underline{\alpha})x_{[1]} - (1-\alpha)\frac{1}{n_\alpha}\sum_{k=1}^{j} x_{[k]}\Big] - \\ &\Big[x_{[i]}(1-\alpha)\frac{1}{n_\alpha}\sum_{k=1}^{i} x_{[k]} - (1-\alpha)\frac{1}{n_\alpha}\sum_{k=1}^{i} x_{[k]}^2\Big] + \\ &x_{[1]}\Big[q_i x_{[i]} - \underline{\alpha} x_{[i]} - (1-\alpha)\frac{1}{n_\alpha}\sum_{k=1}^{i} x_{[k]}\Big]\end{aligned}$$

for a set of proportions $\Theta = \{q_i = \underline{\alpha} + \frac{i(1-\alpha)}{n_\alpha} \quad ; i = 1, n_\alpha\}$.

4.5.2 Winsorising

In this case, the data below a lower threshold $\underline{y}$ and above an upper threshold $\overline{y}$ are set to a constant, for example to $\underline{y}$ and $\overline{y}$. In Van Kerm (2007) winsorising percentages of 0.25%, 0.5%. 0.75% and 1%, both one-sided and two-sided, are used. More about this topic can be found in Hulliger (1999).
In Table 4.13 the results of some winsorising percentages are shown. The problem of winsorising is, that the winsorised indicator depends extremely on the thresholds. The conclusion is that also winsorising is very problematic for the estimation of Laeken indicators.

Table 4.13: Winsorising

$\underline{\alpha}$	$\overline{\alpha}$	$\hat{G}$	$\hat{Q}$
0	0	25.33	3.65
0.5%	0.5%	26.42	3.87
1%	1%	26.85	3.98
0%	0.5%	24.94	3.60
0%	1%	24.60	3.55
0.5%	1%	26.63	3.91
1%	0.5%	26.64	3.94

4.5.3 Minimum estimated risk (MER) estimator

This method consists of a robust estimation of the mean of an asymmetric distribution and can be found in Hulliger and Schoch (2008). An M-estimator $T(F_S, k)$ (see Section 2.2) with asymmetric Huber Ψ-function, $\Psi(x, k) = \min(x, k)$, with a tuning constant $k > 0$ is used to estimate the mean of the highest quintile for the estimation of the quintile share ratio given in Section 3.4. The M-estimator is defined as the solution of

$$\sum_{i \in S} w_i \mathbb{1}\{Q_{F_S}(0.8) < EQ\_INC_i\} \Psi(EQ\_INC_i - T, k) = 0 \quad .$$

Note, that a scale estimator is used to standardise $EQ\_INC_i - T$. To estimate the tuning constant k the estimated mean squared error is used, where the bias is estimated by the difference of the M-estimator and the mean m_5 of the highest quintile and the estimated sample variance by $\hat{V}(T(F_S, k))$. The estimated mean squared error is given by

$$\hat{r}(k) = \hat{V}(T(F_S, k)) + (T(F_S, k) - m_5)^2 \quad .$$

The M-estimator $T(F_S, k_0)$ with the minimising tuning constant k_0 (MER estimator) is chosen to estimate the mean of the highest quintile.
For the EU-SILC 2006 data set from Austria, the minimum estimated risk is obtained as $k_0 = 12.6$. The estimated risk is shown in Figure 4.10.
Figure 4.11 shows the sensitivity curve of the quintile share ratio with one contaminated value of the equivalised income. The influence of outliers is unbounded without M-estimation.
This method is good for samples which includes only few outlying observations.

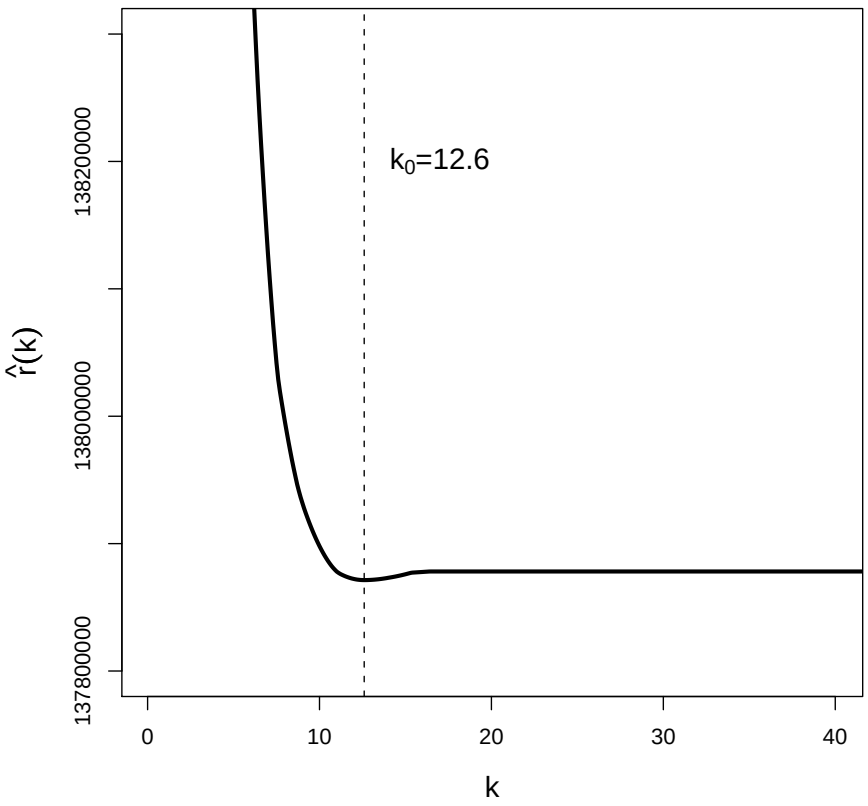

Figure 4.10: Estimated risk with several tuning constants k

4.6 Parametric approach

Information about a parametric approach with income distributions can be found in Cowell (1998), Bandourian et al. (2002) Cowell and Flachaire (2004), Jenkins and Van Kerm (2004) and Davidson and Flachaire (2006).
Various distributions could be used for modeling income distributions. In Figure 4.12 a diagram with the income distributions and their parameters used is visualised (see also Dastrup et al. 2007). All these distributions can be expressed with the generalised Beta distribution (GB) . The probability density function (pdf) of the generalised Beta distribution is given by

$$f(y;a,b,c,p,q) = \frac{|a|y^{ap-1}(1-(1-c)(y/b)^a)^{q-1}}{b^{ap}B(p,q)(1+c(y/b)^a)^{p+q}} \quad \text{for} \quad 0 < y^a < \frac{b^a}{1-c} \quad ,$$

where $0 \leq c \leq 1$, $b, p, q \geq 0$, and $B(p,q)$ is the Beta function. Outside the boundaries 0 and $\frac{b^a}{1-c}$ the probability density function $f(y;a,b,c,p,q) = 0$.

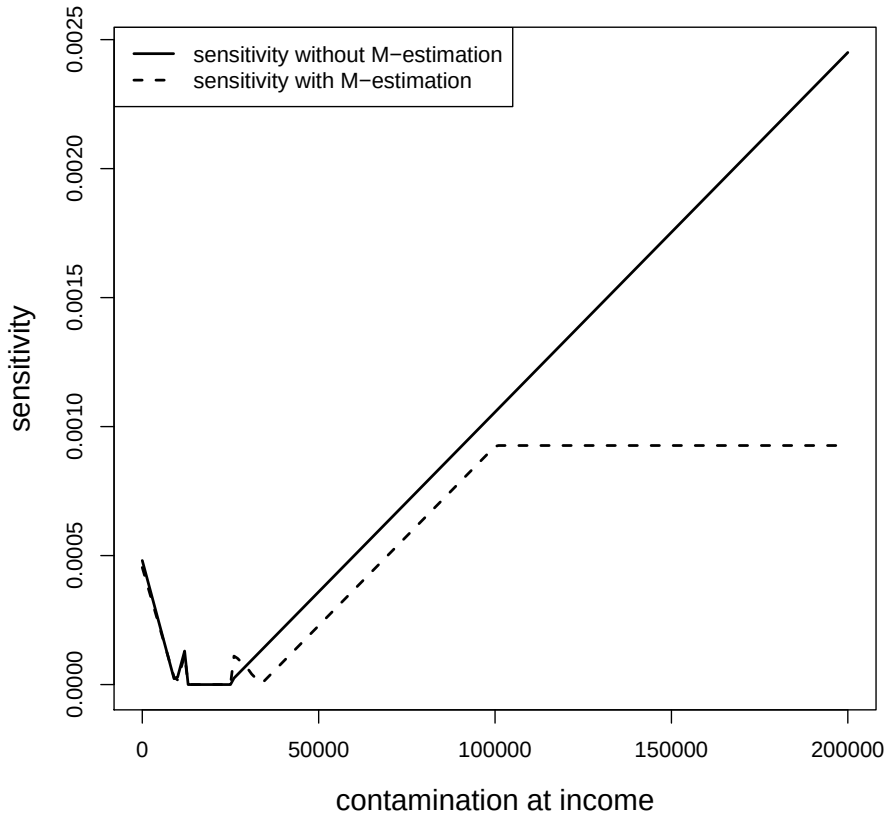

Figure 4.11: Sensitivity curve for the quintile share ratio

4.6.1 Modeling the income distribution with the log-normal distribution

In this diploma thesis, the log-normal- and the Pareto distribution are used for modeling the income distribution. First of all, the log-normal distribution is used. μ and σ^2 are needed for fitting the log-normal distribution, which result from Equations (2.1) and (2.2),

$$\begin{aligned} \sigma^2 &= \ln\left(1 + \frac{Var(X)}{(E(X))^2}\right) \\ \mu &= \ln(E(X)) - \frac{1}{2}\ln\left(1 + \frac{Var(X)}{(E(X))^2}\right) \\ &= \ln(E(X)) - \frac{\sigma^2}{2} \quad . \end{aligned}$$

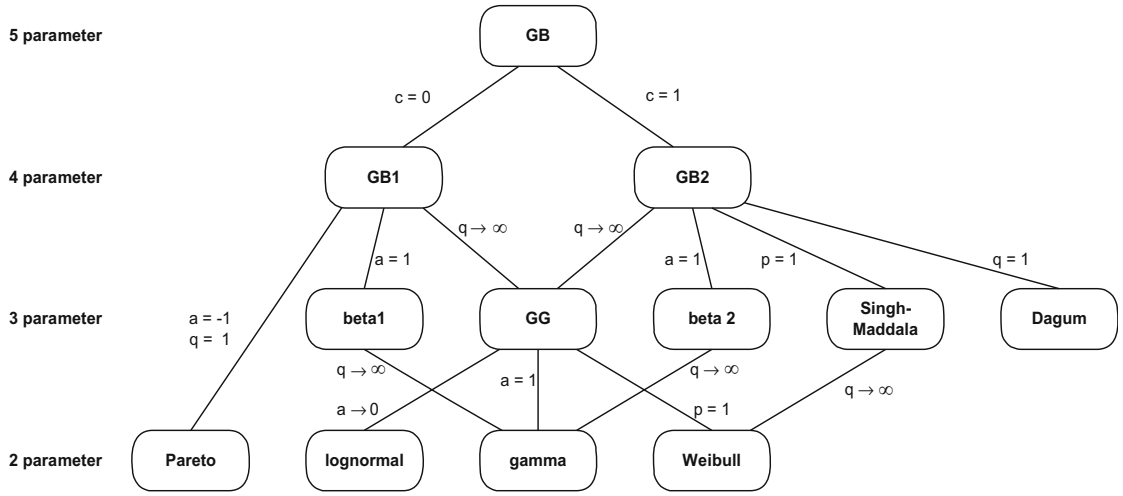

Figure 4.12: Income distribution tree

If a log-normal distribution is used for modeling the empirical income distribution, the Gini coefficient is given by

$$G = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1 \quad ,$$

where Φ is the cumulative distribution function of the standard normal distribution. Then the Gini coefficient can be estimated asymptotically (see Kleiber and Kotz 2003),

$$\hat{G} = 2\Phi\left(\frac{\hat{\sigma}}{\sqrt{2}}\right) - 1 \approx N\left[G, \frac{\sigma^2 e^{\sigma^2/2}}{2\pi n}\right] \quad .$$

The results according to this parametric approach are shown in Table 4.14. Note, that there is a large difference between the estimation in Table 4.1 and the results of the parametric approach in Table 4.14. This indicates that this method is highly non-robust. Furthermore, the Austrian EU-SILC 2006 sample is used for the estimation of the parameters of the log-normal distribution.
Also in the density plots of Figure 4.13, which show the estimated densities of income in each region in Austria, it can be seen that the fitted density differs a lot.
If the sample includes outlying observations, the results gets worse. A possibility to get better results is to robustly estimate the parameters of the log-normal distribution. More information about efficient and robust fitting of log-normal distributions can be found in Serfling (2002).

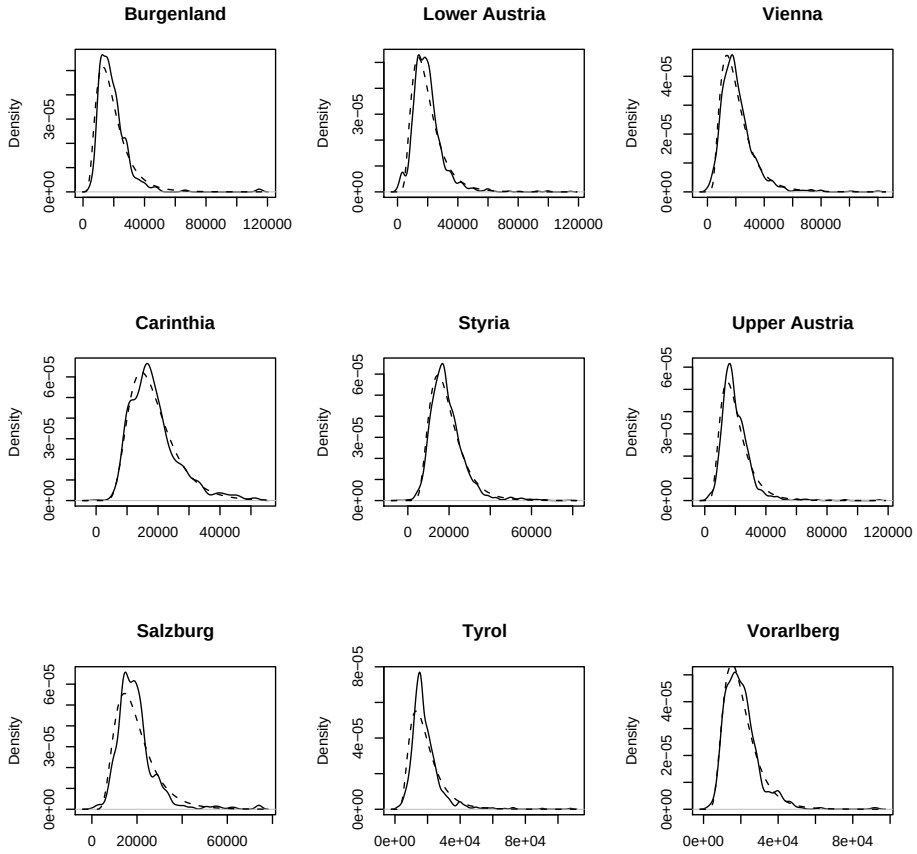

Figure 4.13: Density plots of sample density and log-normal density with estimated parameters from the sample for all regions in Austria

4.6.2 Modeling the income distribution with a mixture of distributions

The next approach of parametric modeling is a mixed income distribution. As before, a log-normal distribution for the main part of the data and a Pareto distribution for the upper tail of the distribution is used. First of all, the threshold x_0 has to be estimated with methods from Section 4.2. The results of the robust prediction error criterion in Table 4.5 are used as threshold x_0. The next step is to estimate the parameters of the log-normal distribution below this threshold and the robust parameters of the Pareto distribution above x_0.
The results are shown in Table 4.15. In Figure 4.14 it can be seen, that the fit in the lower tail is not really good, where the dashed line is the theoretical density with the log-normal distribution. For the upper tail the Pareto distribution is used.

Table 4.14: 99% bootstrap confidence interval of the Gini coefficient and the quintile share ratio with generated log-normal data

	Gini coefficient			quintile share ratio		
region	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
Austria	25.76	26.18	26.63	3.69	3.78	3.86
Burgenland	27.34	29.71	32.21	3.95	4.50	5.08
Lower Austria	26.02	26.94	27.85	3.77	3.97	4.17
Vienna	28.02	29.13	30.31	4.14	4.39	4.66
Carinthia	20.76	21.94	23.17	2.85	3.04	3.22
Styria	22.12	23.05	24.06	3.04	3.19	3.35
Upper Austria	24.09	25.02	25.95	3.42	3.60	3.77
Salzburg	24.04	25.61	27.32	3.32	3.61	3.90
Tyrol	25.47	26.81	28.25	3.64	3.91	4.18
Vorarlberg	22.46	24.08	25.80	3.12	3.42	3.72

Table 4.15: 99% bootstrap confidence interval of the Gini coefficient and the quintile share ratio with generated log-normal and Pareto data

	Gini coefficient			quintile share ratio		
region	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
Austria	23.53	24.01	24.49	3.23	3.30	3.38
Burgenland	21.53	23.61	25.83	2.86	3.20	3.55
Lower Austria	23.71	24.66	25.64	3.30	3.46	3.62
Vienna	24.94	25.97	27.06	3.49	3.68	3.88
Carinthia	20.21	21.56	22.94	2.73	2.94	3.13
Styria	20.97	22.06	23.17	2.84	2.99	3.14
Upper Austria	22.60	23.82	25.05	3.11	3.29	3.46
Salzburg	21.02	23.33	25.75	2.81	3.14	3.47
Tyrol	21.27	22.94	24.65	2.86	3.11	3.35
Vorarlberg	20.70	23.00	25.35	2.83	3.13	3.45

4.6.3 Remarks

To get a better fit, also for the lower tail an inverse Pareto distribution can be used. A possible choice is described in Van Kerm (2007). The cumulative distribution function is given by

$$F^L(x, \theta^l, x^l) = \left(\frac{2x^l - x}{x^l}\right)^{-\theta^l} \quad ,$$

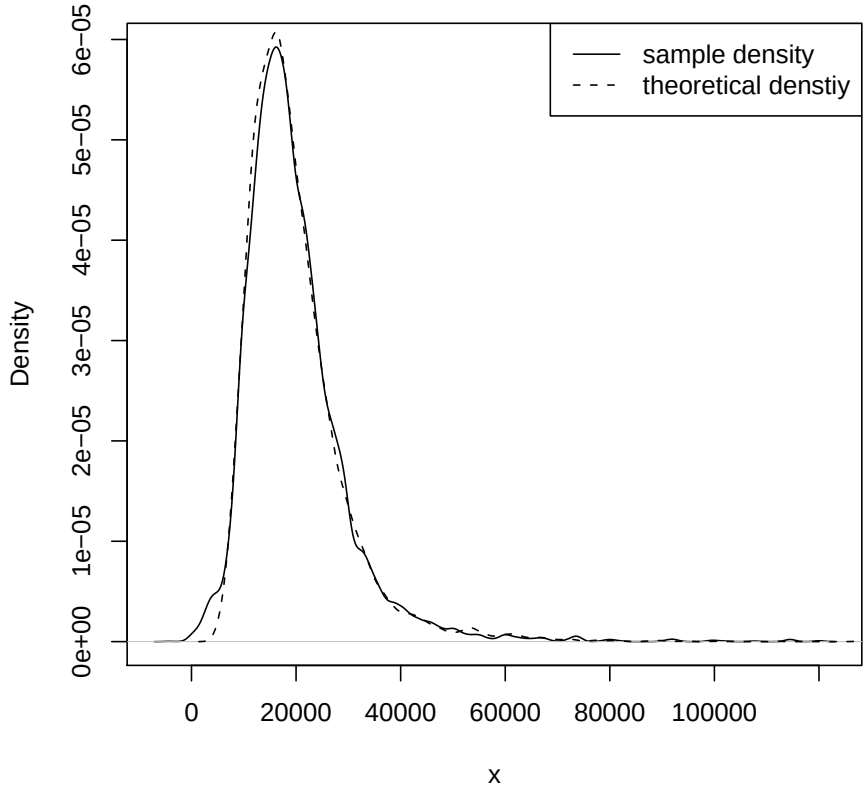

Figure 4.14: Density plot of sample density and estimated log-normal and Pareto density for all Austrian data

where x^l is the threshold and θ^l is the parameter of the inverse Pareto distribution. Another problem of the estimation of an income distribution is that the equivalised income (EQ_INC) also includes zeros and negative values. This problem is not really considered in the literature. Often trimming and winsorising are used to get rid of the problem with negative income. If a semi-parametric approach is applied, the data below the threshold are not changed and the data above this threshold are modeled with the Pareto distribution. The advantage is that negative values and 0 must not be changed and also the fit of the lower tail does not change. In general, the lower tail is not a big problem, because these values have not a big influence on the estimation of the Laeken indicators.

4.7 Semi-parametric approach

Here, only data above a certain threshold (see Section 4.2) are modeled with a suitable distribution and values below this threshold are left unchanged. More information and details can be found in Cowell and Victoria-Feser (2007). The results of this method are shown in Table 4.16 and the conclusion is that this method provides the best fit compared to the estimation according to the original data. Moreover, this conclusion holds if Tables 4.1, 4.14, 4.15 and 4.16 are compared.

Table 4.16: 99% bootstrap confidence interval of the Gini coefficient and the quintile share ratio with generated Pareto data for the upper tail

	Gini coefficient			quintile share ratio		
region	lower	$\hat{G}$	upper	lower	$\hat{Q}$	upper
Austria	24.49	25.04	25.59	3.51	3.61	3.71
Burgenland	22.46	24.67	27.06	3.04	3.40	3.76
Lower Austria	24.16	25.28	26.43	3.50	3.75	4.02
Vienna	27.26	28.49	29.77	4.13	4.49	4.82
Carinthia	21.77	23.24	24.78	3.05	3.28	3.53
Styria	21.53	22.55	23.63	2.98	3.18	3.36
Upper Austria	23.12	24.57	26.05	3.19	3.44	3.67
Salzburg	21.16	23.22	25.32	2.90	3.28	3.65
Tyrol	21.48	23.11	24.81	2.91	3.18	3.45
Vorarlberg	22.46	24.82	27.24	3.11	3.49	3.86

4.7.1 Robustness issues

The aim is to contaminate values of the equivalised income with high values in order to investigate the robustness of the semi-parametric approach. First, a few observations will be contaminated with high values (in this case 1000000 €). In Figure 4.15 it is shown, that the robust methods (PDC, Section 4.3.9 and WMLE, Section 4.3.7)(third and fourth boxplot), are much better than the ML estimation (Hill) (second boxplot). The first boxplot for any number of outliers represents the estimated Gini coefficient with the contaminated data. The dashed line indicates the Gini coefficient of the non-contaminated sample.
Results obtained with the parametric approach are similar, but the semi-parametric approach has the advantage that only few values have to be estimated. Also for the quintile share ratio, the robust methods (PDC and WMLE) are much better than the ML estimation (Hill), which can be seen in Figure 4.16. The dashed line indicates the quintile share ratio of the non-contaminated sample. If many outlying observations

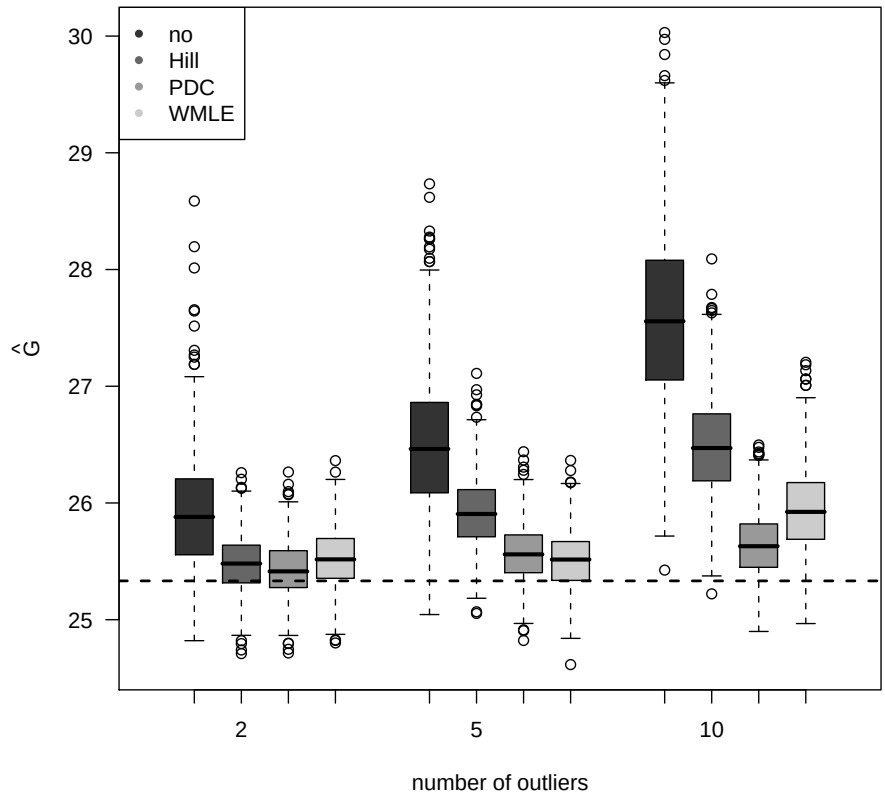

Figure 4.15: Boxplots of the Gini coefficient with contaminated values

are included in the sample, the PDC method is more robust than the WMLE. Also the MER method (see Section 4.5.3) of Hulliger and Schoch (2008) performs well.
For the last analysis, one observation (EQ_INC) is contaminated with different large values starting from 100000 € up to 10000000 € . The results of the two robust methods, visualised in Figure 4.17 show that WMLE and PDC are constant for any contamination. The dashed line indicates the Gini coefficient of the uncontaminated sample.
The same results are obtained for the quintile share ratio. Figure 4.18 shows that the two robust methods, WMLE and PDC, perform best. The dashed line indicates the quintile share ratio of the uncontaminated sample. The tuning constant c for the WMLE method was set to 2.5. Other tuning constants (1, 2, 3 and 5) give worse results. The advantage of the PDC method is, that no tuning constant must be chosen.

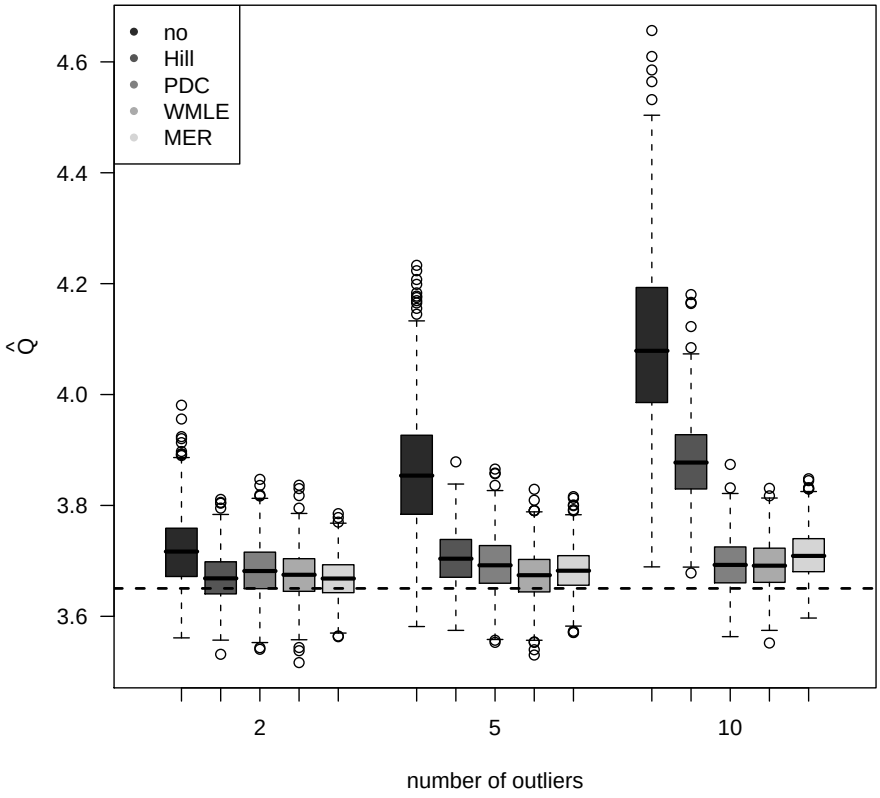

Figure 4.16: Boxplots of the quintile share ratio with contaminated values

4.7.2 Remarks

As a next step, other distributions can be used for modeling the upper tail. However, when applying other distributions more parameters have to be estimated and more complicated methods for the robust estimation of the Laeken indicators are needed, which also leads to higher computational costs. Another problem is the estimation of small domains. If only a few observations for these domains are available, the methods described are not useful. For example, for the estimation of the robust shape parameter of the Pareto distribution, 20 observations are sufficient. This implies, that much more than 20 observations below the upper tail are needed, i.e. the sample must be sufficiently large. If the Laeken indicators for this small domains are estimated using robust procedures, other methods are required.

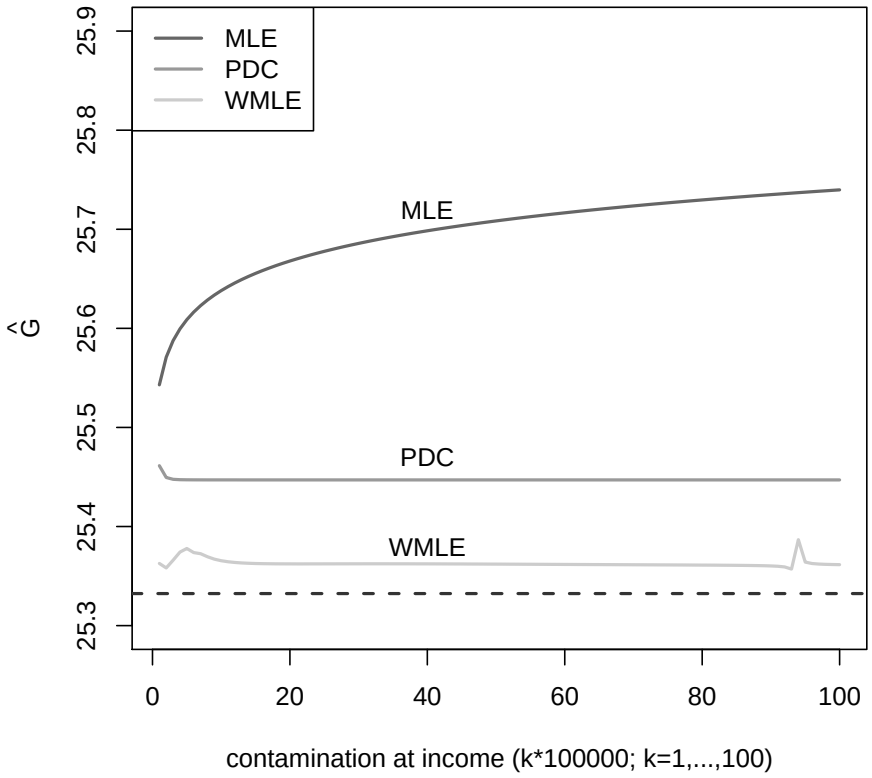

Figure 4.17: Gini coefficient of a contaminated sample

4.8 Outlook to small area estimation (SAE)

The term *small area* is commonly used for a small geographical area, such as a county or a municipality. However, in official statistics, a *small domain* is a (small) subpopulation, for example a age-gender-citizenship subgroup within a large geographical area. These two terms are used interchangeably. More about the explanation of a small are can be found in Rao (2003) and Ghosh and Rao (1994).
For policy makers it is often important to have relevant information for such small areas. This problem leads to the topic of small area estimation (SAE), because in most cases only a few observations for such small domains are available and therefore the estimated statistic is not applicable without improving the estimation using additional information.
Many small area methods to estimate totals, means, ratios and other parameters exist, some of these methods can be found in Särndal et al. (1992), Pfeffermann

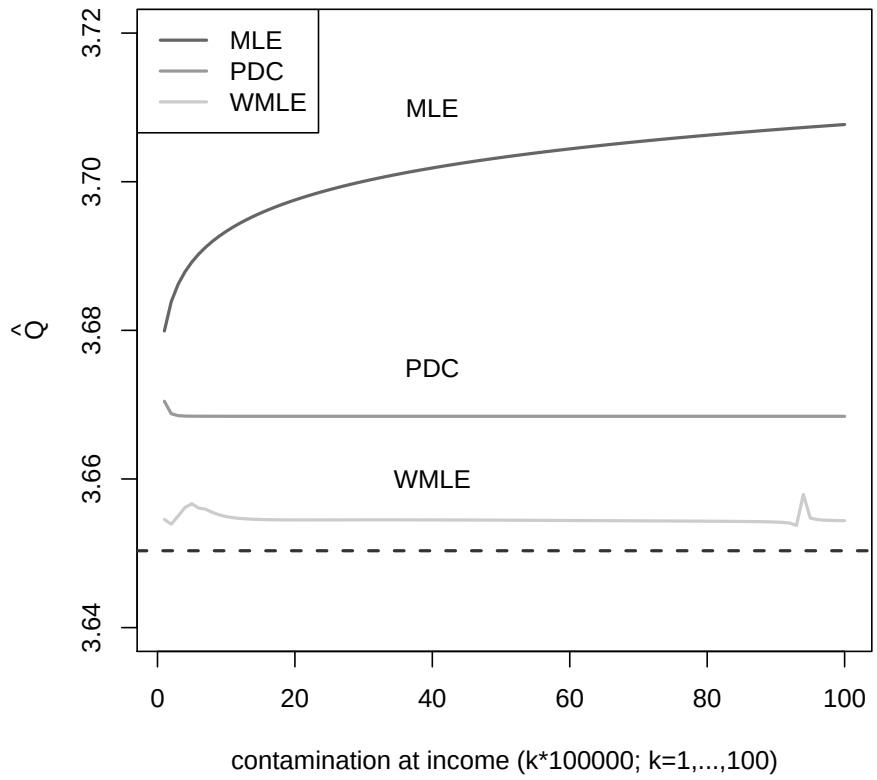

Figure 4.18: Quintile share ratio of a contaminated sample

(2002), Lehtonen et al. (2003), Lehtonen and Pahkinen (2004) and Lehtonen and Veijanen (2008). Popular small area estimators are the generalised regression (GREG) estimator or the empirical best linear unbiased prediction (EBLUP) estimator. Many other methods can be found in Rao (2003), Chambers (2005b), Chambers and Tzavidis (2005), Tzavidis and Chambers (2007) and Tzavidis et al. (2008).
A future task is to implement robust small area estimators in **R**.

Chapter 5

Summary and Conclusions

The motivation of this diploma thesis was to investigate in robust methods for the estimation of the Laeken indicators. From the beginning, the problem was the appropriate modeling of the upper tail of the empirical income distribution. The influence of only a few extreme values on the Gini coefficient and the quintile share ratio is large when using standard estimators. Therefore, robust methods, where these extreme values have less influence on the estimation of certain Laeken indicators, have to be used. In the literature a lot of different possibilities are already described.
The first one is the non-parametric approach (see Section 4.5), especially trimming. The problem that occurs within that approach is to fix the amount of trimming. However, with trimming a lot of information is lost and often the results get worse, since large (non-representative) outliers have zero influence on the estimation. Therefore, other methods should be considered.
The next possibility is the parametric approach (see Section 4.6), where a theoretical distribution of the income is used. The goal is to fit the income distribution model as good as possible. For the estimation of the parameters of the distribution also robust methods can be used. The problem now is to get a good fit over the whole income distribution, which turns out to be not realistic with the underlying data.
Finally, the semi-parametric approach (see Section 4.7) is described. The advantage is, that only the upper tail of the empirical income distribution has to be estimated. Moreover, two further problems occur. First, a threshold from which the chosen distribution for the upper tail is applied has to be found. Different methods for the estimation of this threshold are compared (see Section 4.2) within a simulation study (see Section 4.4); the prediction error criterion, the robust prediction error criterion, a bootstrap method and the weighted asymptotic mean squared error. The best results are obtained with the robust prediction error criterion. The second problem is to estimate the parameters for modeling the upper tail. In this diploma thesis the Pareto distribution was used for modeling. Thus the shape parameter has to be estimated. Different methods were compared, the classical maximum likelihood estimator (Hill estimator), and two robust estimators, the weighted maximum likelihood estimator (WMLE) and the partial density component (PDC) estimator (see Section 4.3). The

Hill estimator performs reasonable with only few extreme values (outlying observations), but in general, the robust methods perform better. The difference between the two robust estimators is that a tuning constant must be estimated for the WMLE estimator. If this tuning constant is worse, also the results gets worse. The big advantage of the PDC estimator is that it does not need a tuning constant. Different results of the methods for the estimation of the shape parameter of the Pareto distribution are obtained in the simulation study. As a conclusion, both robust methods can be used in practice.
As a final remark, the semi-parametric methods are only suitable if enough observations for the estimation of the shape parameter are available. 20 observations above the threshold are sufficient for both methods. This problem leads to the topic small area estimation (see Section 4.8). However, this diploma thesis only gives a short outlook to this topic.
All methods are implemented in the statistical environment **R**. A freely available open-source R package including all the methods and their documentation will be released in the near future.

Appendix A

R-Code

The code related to this thesis was written in **R**. Some helpful information about **R** can be found in Venables and Ripley (2002), Chambers (2008) and Chen et al. (2008).

A.1 Prediction error criterion

A.1.1 Hill estimator

```
> hill <- function(data, k) {
+     n <- length(data[, "eqInc"])
+     eqInc <- sort(data[, "eqInc"])
+     x0 <- eqInc[n - k]
+     theta <- k/(sum(log(eqInc[(n + 1 - k):n])) - k * log(x0))
+     return(theta)
+ }
```

A.1.2 Mean squared prediction error

```
> C_x0 <- function(data, k) {
+     hillv <- hill(data, k = k)
+     sumh1 <- sumh2 <- c(1:k)
+     n <- length(data[, "eqInc"])
+     eqInc <- sort(data[, "eqInc"])
+     hf1 <- function(eqInc = eqInc, k = k, i, hillv = hillv) {
+         hk <- c(1:k)
+         x0 <- eqInc[n + 1 - k]
+         h1 <- 1/(sum(1/hk[(k - i + 1):k]^2)) * (log(eqInc[n -
+             k + i]/x0) + 1/hillv * log((k + 1 - i)/(k + 1)))^2
+         return(h1)
+     }
+     hf2 <- function(eqInc = eqInc, k = k, i, hillv = hillv) {
```

```
+         hk <- c(1:k)
+         h2 <- 1/(sum(1/hk[(k - i + 1):k]^2)) * (log((k + 1 -
+             i)/(k + 1)))^2
+         return(h2)
+     }
+     sumh1 <- sapply(sumh1, function(x) hf1(eqInc, k, x, hillv))
+     sumh2 <- sapply(sumh2, function(x) hf2(eqInc, k, x, hillv))
+     val <- hillv^2/k * sum(sumh1[1:k]) + 2/k^2 * sum(sumh2[1:k]) - 1
+ }
```

A.1.3 Estimation of the threshold x_0

```
> predErr <- function(data, max = (dim(data)[1] - 5),start=20) {
+     n <- length(data[, "eqInc"])
+     c0 <- c(start:max)
+     c0 <- sapply(c0, function(x) C_x0(data, x))
+     k <- order(c0)[1] + 20 - 1
+     c1 <- c0[order(c0)[1]]
+     eqInc <- sort(data[, "eqInc"])
+     list(kopt = k, cmin = c1, val = eqInc[n - k], predErr = c0,
+         k = c(start:max))
+ }
```

A.2 Robust prediction error criterion

A.2.1 Robust estimation of the shape parameter with standardised residuals

```
> MTheta <- function(data,k,const = 1.5,tol = .Machine$double.eps^0.5,
+     interval = c(0.001, 30), bias = TRUE) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Y <- hY <- hhsig <- hsig <- hk <- c(1:k)
+     Y <- log(x/x0)
+     hY <- -log((k + 1 - hk)/(k + 1))
+     hhsig <- (1/(k + 1 - hk)^2)
+     hsig <- sapply(hk, function(x) sqrt(sum(hhsig[1:x])))
+     zeroTheta <- function(x) {
+         ri <- ni <- wi <- c(1:k)
+         ri <- (Y - hY/x)/(hsig/x)
```

```
+         weightTheta <- function(ni, ri, const = const) {
+             if (abs(ri[ni]) < const)
+                 wi[ni] = 1
+             else wi[ni] = const/abs(ri[ni])
+         }
+         wi <- sapply(ni, function(y) weightTheta(y, ri, const))
+         sum(sapply(ni, function(z) wi[z] * (1/x - Y[z])))
+     }
+     val <- uniroot(f = zeroTheta, tol = tol, interval = interval,
+         maxiter = 1000)
+     if (bias == TRUE) {
+         ri <- ni <- wi <- c(1:k)
+         ri <- (Y - hY/val$root)/(hsig/val$root)
+         weightTheta <- function(ni, ri, const = const) {
+             if (abs(ri[ni]) < const)
+                 wi[ni] = 1
+             else wi[ni] = const/abs(ri[ni])
+         }
+         wi <- sapply(ni, function(y) weightTheta(y, ri, const))
+         F <- c(1:(k + 1))
+         x0 <- eqInc[n - k]
+         x <- eqInc[(n - k):n]
+         F <- sapply(F, function(s) 1 - (x[s]/x0)^(-val$root))
+         FmF <- wlgf <- lgwf <- num <- den <- c(1:k)
+         FmF <- sapply(ni, function(t) F[t + 1] - F[t])
+         wlgf <- sapply(ni, function(u) wi[u] * (1/val$root -
+             Y[u]))
+         lgwf <- sapply(ni, function(v) wi[v] * (-1/val$root^2))
+         ablwi <- function(ni, wi, Y = Y, hY = hY, hsig = hsig,
+             theta = val$root, const = const) {
+             hval <- c(1:k)
+             if (wi[ni] == 1)
+                 hval[ni] = 0
+             else hval[ni] = (-const) * hsig[ni] * Y[ni]/(theta^2 *
+                 (Y[ni] - hY[ni]/theta)^2) * (1/theta - Y[ni])
+         }
+         wiabl <- c(1:k)
+         wiabl <- sapply(ni, function(q) ablwi(q, ri, Y = Y, hY = hY,
+             hsig = hsig, theta = val$root, const = const))
+         num <- wlgf * FmF
+         den <- (wiabl + lgwf) * FmF
+         biasval <- -sum(num)/sum(den)
```

```
+         theta <- val$root - biasval
+     }
+     else theta <- val$root
+     return(theta)
+ }
```

A.2.2 Robust estimation of the shape parameter with distribution function

```
> MThetaF <- function(data, k, p1 = 0.005, p2 = 0.005,
+     tol =.Machine$double.eps^0.5,interval = c(0.001, 30),bias=TRUE){
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Y <- c(1:k)
+     Y <- log(x/x0)
+     zeroTheta <- function(t) {
+         Fi <- ni <- wi <- c(1:k)
+         Fi <- 1 - (x/x0)^(-t)
+         weightTheta <- function(ni, Fi, p1 = p1, p2 = p2) {
+             if (Fi[ni] < p1)
+                 wi[ni] = Fi[ni]/p1
+             else {
+                 if (Fi[ni] >= p1 && Fi[ni] <= (1 - p2))
+                   wi[ni] = 1
+                 else wi[ni] = (1 - Fi[ni])/p2
+             }
+         }
+         wi <- sapply(ni, function(y) weightTheta(y, Fi, p1 = p1,
+             p2 = p2))
+         sum(sapply(ni, function(z) wi[z] * (1/t - Y[z])))
+     }
+     val <- uniroot(f = zeroTheta, tol = tol, interval = interval,
+         maxiter = 1000)
+     if (bias == TRUE) {
+         biasval <- (val$root/2) * (2 * (1 - p1)^2 * log(1 - p1) +
+             p1 * (1 - p1) + p1 * (1 - p2) + 2 * p1 * p2 * log(p2))/
+             (((1 -p1) * log(1 - p1))^2 - p1 * (1 - p1) - p1 *
+             (1 - p2) + p1 * p2 * (log(p2))^2)
+         theta <- val$root - biasval
+     }
```

```
+     else theta <- val$root
+     return(theta)
+ }
```

A.2.3 Robust mean squared prediction error

```
> Cr_x0 <- function(data, k, theta, method = "MRes", const = 2.5,
+     p1 = 0.005, p2 = 0.005) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Yi <- hatYi <- hsig <- sigi <- ri <- ni <- c(1:k)
+     Yi <- log(x/x0)
+     hatYi <- -1/theta * log((k + 1 - ni)/(k + 1))
+     hsig <- (1/(k + 1 - ni)^2)
+     sigi <- sapply(ni, function(x) 1/theta^2 * (sum(hsig[1:x])))
+     ri <- (Yi - hatYi)/sqrt(sigi)
+     wi <- c(1:k)
+     Fi <- c(1:k)
+     Fi <- 1 - (x/x0)^(-theta)
+     if (method == "MRes") {
+         weightTheta <- function(ni, ri, const = const) {
+             if (abs(ri[ni]) < const)
+                 wi[ni] = 1
+             else wi[ni] = const/abs(ri[ni])
+         }
+         wi <- sapply(ni, function(y) weightTheta(y, ri, const))
+     }
+     else {
+         if (method == "MF") {
+             weightTheta <- function(ni, Fi, p1 = p1, p2 = p2) {
+                 if (Fi[ni] < p1)
+                   wi[ni] = Fi[ni]/p1
+                 else {
+                   if (Fi[ni] >= p1 && Fi[ni] <= (1 - p2))
+                     wi[ni] = 1
+                   else wi[ni] = (1 - Fi[ni])/p2
+                 }
+             }
+             wi <- sapply(ni, function(y) weightTheta(y, Fi, p1,
+                 p2))
+         }
```

```
+         else wi <- 1
+     }
+     fik <- matrix(1, nrow = k, ncol = k)
+     f_x <- function(ni, theta = theta, x0 = x0, x = x) fik[,ni]
+         = (k - ( ni- 1))* choose(k,(ni-1)) * ((1 - (x/x0)^
+         (-theta))^(ni - 1))* (((x/x0)^(-theta))^(k - (ni - 1))) *
+         theta * x^(-1)
+     fik <- sapply(ni, function(z) f_x(z, theta = theta, x0 = x0,
+         x = x))
+     hi <- c(1:k)
+     hi <- log((k + 1 - ni)/(k + 1))
+     xcor <- c(x0, x)
+     diff <- c(1:k)
+     diff <- sapply(ni, function(s) xcor[s + 1] - xcor[s])
+     covwiYi <- c(1:k)
+     covwiYi <- sapply(ni, function(u) -1/theta * sum(diff[1:u] *
+         hi[1:u] * wi[1:u] * Yi[1:u] * fik[(1:u), u]) + 1/theta *
+         sum(wi[1:u] * diff[1:u] * Yi[1:u] * fik[(1:u), u]) *
+         sum(diff[1:u] * hi[1:u] * wi[1:u] * fik[(1:u), u]))
+     varwiYi <- c(1:k)
+     varwiYi <- sapply(ni, function(v) sum(diff[1:v] * wi[1:v]^2 *
+         Yi[1:v]^2 * fik[(1:v), v]) - (sum(diff[1:v] * wi[1:v] *
+         Yi[1:v] * fik[(1:v), v])))^2
+     wiri <- c(1:k)
+     wiri <- wi^2 * ri^2
+     Cr <- 1/k * sum(wiri) + 2/k * sum(1/sigi * covwiYi) - 1/k *
+         sum(1/sigi * varwiYi)
+     return(Cr)
+ }
```

A.2.4 Estimation of the threshold x_0

```
> robpredErr <- function(data,max = 950,method = "MRes",const = 1.5,
+     p1 = 0.005, p2 = 0.005, interval = c(0.001, 30), start=20) {
+     n <- length(data[, "eqInc"])
+     cr0 <- hv <- theta <- c(start:max)
+     if (method == "MRes") {
+         theta <- sapply(hv, function(z) MTheta(data, k = z,
+         const = const, interval = c(0.001, 100), bias = TRUE))
+     }
+     if (method == "MF") {
+         theta <- sapply(hv, function(z) MThetaF(data, k = z,
+         p1 = p1, p2 = p2, interval = c(0.001, 10), bias = TRUE))
```

```
+     }
+     cr0 <- sapply(hv, function(x) Cr_x0(data, hv[x - start + 1],
+         theta[x - start + 1], method = method, const = const,
+         p1 = p1, p2 = p2))
+     k <- order(cr0)[1] + start - 1
+     c1 <- cr0[order(cr0)[1]]
+     eqInc <- sort(data[, "eqInc"])
+     list(kopt = k, cmin = c1, val = eqInc[n - k],
+        theta = theta[order(cr0)[1]],predErr = cr0,k = c(start:max))
+ }
```

A.3 Bootstrap method

```
> kboot <- function(data, n1 = NULL) {
+     eqInc <- data[, "eqInc"]
+     n <- length(eqInc)
+     n1min <- trunc(sqrt(10 * n)) + 1
+     if (is.null(n1))
+         n1 <- n1min
+     else n1 <- min(max(n1min, n1), (n - 1))
+     n2 <- round(n1^2/n)
+     if (n1 < sqrt(n) || n2 < 10)
+         cat("n1 zu gering!!!\n")
+     else {
+         data <- matrix(1, ncol = n1, nrow = n1)
+         a <- c(1:(n1 - 5))
+         b <- c(1:n1)
+         data <- sapply(b, function(f) sort(sample(eqInc, n1,
+             replace = TRUE)))
+         gn <- matrix(1, ncol = n1, nrow = (n1 - 5))
+         M <- matrix(1, ncol = n1, nrow = (n1 - 5))
+         Q <- matrix(1, ncol = n1, nrow = (n1 - 5))
+         gn <- sapply(b, function(g) sapply(a, function(y)
+             Gamman(data[, g], y)))
+         M <- sapply(b, function(h) sapply(a, function(z) Mn(data[,
+             h], z)))
+         Q <- t(sapply(a, function(p) sapply(b, function(s) (M[p,
+             s] - 2 * gn[p, s]^2)^2)))
+         Qn <- c(1:(n1 - 5))
+         Qn <- sapply(a, function(t) 1/n1 * sum(Q[t, ]))
+         k1 <- order(Qn)[1]
+         val1 <- Qn[k1]
```

```
+         data1 <- matrix(1, ncol = n2, nrow = n2)
+         a <- c(1:(n2 - 2))
+         b <- c(1:n2)
+         data1 <- sapply(b, function(k) sort(sample(eqInc, n2,
+             replace = TRUE)))
+         gn1 <- matrix(1, ncol = n2, nrow = (n2 - 2))
+         M1 <- matrix(1, ncol = n2, nrow = (n2 - 2))
+         Q1 <- matrix(1, ncol = n2, nrow = (n2 - 2))
+         gn1 <- sapply(b, function(l) sapply(a, function(u)
+             Gamman(data1[, l], u)))
+         M1 <- sapply(b, function(m) sapply(a, function(v)
+             Mn(data1[,m], v)))
+         Q1 <- t(sapply(a, function(q) sapply(b, function(w)
+             (M1[q,w] - 2 * gn1[q, w]^2)^2)))
+         Qn1 <- c(1:(n2 - 2))
+         Qn1 <- sapply(a, function(x) 1/n2 * sum(Q1[x, ]))
+         k2 <- order(Qn1)[1]
+         val2 <- Qn1[k2]
+         k <- round(k1^2/k2*((log(k1))^2/((2 * log(n1) -
+             log(k1))^2))^((log(n1) - log(k1))/log(n1)))
+         if (k > (n-5))
+             k <- n-5
+         x0 <- sort(eqInc)[n - k]
+         theta <- 1/Gamman(eqInc, k)
+         list(k = k, theta = theta, x0 = x0, k1 = k1, k2 = k2,
+             n1 = n1, n2 = n2)
+     }
+ }
```

A.4 Weighted mean squared error

A.4.1 Mean squared error

```
> MSEopt <- function(data, k, rho = -1, const = 2.5, p1 = 0.005,
+     p2 = 0.005, interval = c(0.01, 100), method = "MRes") {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     i <- hv <- c(1:k)
+     w1 <- i/(k + 1)
+     w2 <- -log(i/(k + 1))
+     hv <- i/(k + 1)
+     ak1 <- 1/k * sum(hv^(-1) - 1)
```

```
+     ak2 <- 1/k * sum(w2 * (hv^(-1) - 1))
+     bk1 <- (1 - rho)^2 * 1/k * sum(((hv^(-rho) - 1)/rho)^2)
+     bk2 <- (1 - rho)^2 * 1/k * sum(w2 * ((hv^(-rho) - 1)/rho)^2)
+     delta1 <- (bk2 - ak2)/(ak1 * bk2 - bk1 * ak2)
+     delta2 <- (ak1 - bk1)/(ak1 * bk2 - bk1 * ak2)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Y <- hY <- hhsig <- hsig <- hk <- c(1:k)
+     Y <- log(x/x0)
+     hY <- -log((k + 1 - i)/(k + 1))
+     if (method == "MRes")
+         theta <- MTheta(data, k = k, const = const,
+             interval = interval, bias = TRUE)
+     if (method == "MF")
+         theta <- MThetaF(data, k = k, p1 = p1, p2 = p2,
+             interval = interval, bias = TRUE)
+     if (method == "ISE")
+         theta <- ThetaISE(data, k = k, interval = interval)$minimum
+     if (method == "PDC")
+         theta <- ThetaPDC(data, k = k, interval = interval)$minimum
+     hatY <- 1/theta * hY
+     MSE <- delta2/k * sum(w2 * (Y - hatY)^2) + delta1/k * sum((Y -
+         hatY)^2)
+     list(MSE = MSE, theta = theta, x0 = x0, k = k)
+ }
```

A.4.2 Estimation of the threshold x_0

```
> minMSE <- function(data, max = 1000, const = 2.5, p1 = 0.005,
+     p2 = 0.005, interval = c(0.01, 30), method = "MRes", start=20) {
+     n <- dim(data)[1]
+     MSE <- k <- c(start:max)
+     MSE <- sapply(k, function(x) MSEopt(data, k = x, const = const,
+         p1 = p1, p2 = p2, interval = interval, method = method)$MSE)
+     kopt <- order(MSE)[1] + start - 1
+     MSEmin <- MSE[order(MSE)[1]]
+     eqInc <- sort(data[, "eqInc"])
+     theta <- MTheta(data,k = kopt,const = const,interval = interval,
+         bias = TRUE)
+     list(kopt = kopt, MSEmin = MSEmin, x0 = eqInc[n - kopt],
+         theta = theta, MSE = MSE, k = k)
+ }
```

A.5 Estimation of the shape parameter θ of the Pareto distribution

A.5.1 QQ estimator

```
> ThetaQQ <- function(data, k) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     i <- c(1:k)
+     den <- sum(-log(i/(k + 1)) * (k * log(eqInc[n + 1 - i]) -
+         sum(log(eqInc[n - k + i]))))
+     nom <- k * sum((log(1/(k + 1)))^2) - (sum(log(1/(k + 1))))^2)
+     QQ <- den/nom
+     return(QQ)
+ }
```

A.5.2 Pickands estimator

```
> ThetaPickands <- function(data, k) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     Pickands <- ((1/log(2)) * log((eqInc[n - trunc(k/4)] - eqInc[n -
+         trunc(k/2)])/(eqInc[n - trunc(k/2)] - eqInc[n - k])))^(-1)
+     return(Pickands)
+ }
```

A.5.3 Moment estimator

```
> ThetaMoment <- function(data, k) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     i <- c(1:k)
+     M1 <- 1/k * sum(log(eqInc[n - i + 1])) - log(eqInc[n - k])
+     hM <- (log(eqInc[n - i + 1]) - log(eqInc[n - k]))^2
+     M2 <- 1/k * sum(hM)
+     Moment <- (M1 + 1 - 1/2 * (1 - M1^2/M2)^(-1))^(-1)
+     return(Moment)
+ }
```

A.5.4 LS estimator

```
> ThetaLS <- function(data, k) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     z <- log(x)
+     ci <- pi <- i <- c(1:k)
+     pi <- i/k
+     pi[k] <- k/(k + 1)
+     ci <- -log(1 - pi)
+     cim <- mean(ci)
+     zm <- mean(z)
+     theta <- ((1/k * sum(ci * z) - cim * zm)/(1/k * sum(ci^2) -
+         cim^2))^(-1)
+     x0est <- exp(zm - theta^(-1) * cim)
+     list(theta = theta, x0 = x0est)
+ }
```

A.5.5 Trimmed median estimator

```
> ThetaTM <- function(data, k, beta1 = 0.05, beta2 = 0.05) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     ci <- diff <- c(1:k)
+     kl <- trunc(k * beta1)
+     kh <- k - trunc(k * beta2)
+     ci[1:kl] <- 0
+     ci[(kh + 1):k] <- 0
+     hind <- herg <- c((kl + 1):kh)
+     hv <- i <- c(1:kh)
+     hv <- 1/(k - i + 1)
+     herg <- sapply(hind, function(x) sum(hv[1:x]))
+     ci[(kl + 1):kh] <- 1/sum(herg)
+     diff <- log(x) - log(x0)
+     theta <- (sum(ci * diff))^(-1)
+     return(theta)
+ }
```

A.5.6 Integrated squared error (ISE) estimator

```
> ThetaISE <- function(data, k, tol = .Machine$double.eps^0.5,
+     interval = c(0.001, 10)) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Y <- c(1:k)
+     Y <- x/x0
+     ISE <- function(t) {
+         sumf <- c(1:k)
+         sumf <- t * Y^(-1 - t)
+         t^2/(2 * t + 1) - 2/k * sum(sumf)
+     }
+     val <- optimize(f = ISE, tol = tol, interval = interval)
+     return(val)
+ }
```

A.5.7 Partial density component (PDC) estimator

```
> ThetaPDC <- function(data, k, tol = .Machine$double.eps^0.5,
+     interval = c(0.001, 10)) {
+     eqInc <- sort(data[, "eqInc"])
+     n <- length(eqInc)
+     x0 <- eqInc[n - k]
+     x <- eqInc[(n - k + 1):n]
+     Y <- c(1:k)
+     Y <- x/x0
+     ISE <- function(t) {
+         sumf <- c(1:k)
+         sumf <- t * Y^(-1 - t)
+         what <- (1/k * sum(sumf))/(t^2/(2 * t + 1))
+         what^2 * (t^2/(2 * t + 1)) - 2 * what/k * sum(sumf)
+     }
+     val <- optimize(f = ISE, tol = tol, interval = interval)
+     return(val)
+ }
```

A.6 Minimum estimated risk (MER) estimator

A.6.1 Estimation of the minimum estimated risk (MER)

```
> kmer <- function(data, range = c(0.1, 50)) {
+     corvar <- function(data, mean) {
+         n <- dim(data)[1]
+         h <- c(1:n)
+         h <- sapply(h, function(t) (data$eqInc[t] - mean)^2)
+         nvar <- 1/(n - 1) * sum(h)
+     }
+     y <- seq(range[1], range[2], 0.1)
+     n <- length(y)
+     T <- sT <- r <- c(1:n)
+     m5 <- sum(data$weight * data$eqInc)/sum(data$weight)
+     T <- sapply(T,function(u) Mmer(data$eqInc, k=y[u],
+           weights=data$weight)$mu)
+     sT <- sapply(sT, function(v) corvar(data, T[v]))
+     r <- sapply(r, function(w) sT[w] + (T[w] - m5)^2)
+     minimum <- order(r)[1]
+     kmin <- y[minimum]
+     list(kmin=kmin,r=data.frame(k=y,risk=r,M=T,Mkmin=T[minimum]))
+ }
```

A.6.2 M-estimator

```
> Mmer <- function(x, k = 10, weights = NULL, tol = 1e-06,
+     mu = if (is.null(weights)) median(x) else
+     wgt.himedian(x, weights), s = if (is.null(weights))
+     mad(x, center = mu) else wgt.himedian(abs(x - mu),
+     weights),warn0scale=getOption("verbose")){
+     if (any(i <- is.na(x))) {
+         x <- x[!i]
+         if (!is.null(weights))
+             weights <- weights[!i]
+     }
+     n <- length(x)
+     sum.w <- if (!is.null(weights)) {
+         stopifnot(is.numeric(weights), weights >= 0,
+             length(weights) == n)
+         sum(weights)
+     }
+     else n
```

```
+     it <- 0:0
+     if (sum.w == 0)
+         return(list(mu = NA, s = NA, it = it))
+     if (s <= 0) {
+         if (s < 0)
+             stop("negative scale `s'")
+         if (warn0scale && n > 1)
+             warning("scale `s' is zero -- returning initial `mu'")
+     }
+     else {
+         sc <- x - mu
+         st <- mad(sc)
+         wsum <- if (is.null(weights))
+             sum
+         else function(u) sum(u * weights)
+         repeat {
+             it <- it + 1:1
+             yh <- pmin((x - mu)/st, k)
+             y <- yh * st + mu
+             mu1 <- wsum(y)/sum.w
+             if (abs(sum(weights * yh)) < tol)
+                 break
+             mu <- mu1
+         }
+     }
+     list(mu = mu, s = s, it = it)
+ }
```

A.7 Parametric modeling

A.7.1 Estimation of the sample weights

```
> weightdata <- function(data, region = TRUE, N) {
+     if (region == FALSE) {
+         part <- data
+         n <- 1
+         weight <- len <- ind <- c(1:n)
+         len <- dim(data)[1]
+         weight <- sum(N)/len
+     }
+     else {
+         part <- split(data, list(data[, "region"]))
```

```
+         n <- length(levels(data[, "region"]))
+         weight <- len <- ind <- c(1:n)
+         len <- sapply(ind, function(x) dim(part[[x]])[1])
+         weight <- sapply(ind, function(y) N[y]/len[y])
+     }
+     return(weight)
+ }
```

A.7.2 Estimation of a log-normal distributed sample

```
> fiteqIncLN <- function(data, region = TRUE, N) {
+     eqInc <- data[, "eqInc"]
+     part <- split(data, list(data[, "region"]))
+     if (region == TRUE) {
+         m <- c(by(data$eqInc, data$region, mean))
+         v <- c(by(data$eqInc, data$region, var))
+         sdlog <- sqrt(log(v/m^2 + 1))
+         meanlog <- log(m) - sdlog^2/2
+         wval <- weightdata(data, N = N)
+         for (i in 1:length(levels(data$region))) {
+             ma <- data.frame(eqInc = rlnorm(dim(part[[i]])[1],
+                 meanlog[i], sdlog[i]), region =
+                 levels(data$region)[i], weight = wval[i])
+             if (i == 1)
+                 eqIncSample <- ma
+             else eqIncSample <- rbind(eqIncSample, ma)
+         }
+         eqIncSample <- data.frame(eqIncSample,
+             pID = c(1:dim(data)[1]))
+     }
+     else {
+         sdlog <- sqrt(log(var(data$eqInc)/mean(data$eqInc)^2 +
+             1))
+         meanlog <- log(mean(data$eqInc)) - sdlog^2/2
+         wval <- weightdata(data, region = FALSE, N = N)
+         eqIncSample <- data.frame(eqInc = rlnorm(dim(data)[1],
+             meanlog, sdlog), weight = wval, pID = c(1:dim(data)[1]),
+             region = "Austria")
+     }
+ }
```

A.7.3 Estimation of a log-normal distributed and Pareto distributed sample

```
> fitParLN <- function(data, N, k, method = "Hill", part = TRUE) {
+     if (part == TRUE) {
+         part <- split(data, list(data[, "region"]))
+         wval <- weightdata(data, N = N)
+         for (i in 1:length(levels(data$region))) {
+             eqInc <- sort(part[[i]]$eqInc)
+             n <- length(eqInc)
+             xm <- eqInc[(n - k[i])]
+             xi <- eqInc[(n - k[i] + 1):n]
+             if (method == "Hill")
+                 theta <- hill(part[[i]], k[i])
+             if (method == "WML")
+                 theta <- MTheta(part[[i]], k[i])
+             if (method == "PDC")
+                 theta <- ThetaPDC(part[[i]], k[i])$minimum
+             eqIncP <- data.frame(eqInc = VGAM::rpareto(k[i],
+                 loc = xm,shape = theta),region =
+                 levels(data$region)[i], weight = wval[i])
+             m <- mean(eqInc[1:(n - k[i])])
+             v <- var(eqInc[1:(n - k[i])])
+             sdlog <- sqrt(log(v/m^2 + 1))
+             meanlog <- log(m) - sdlog^2/2
+             eqIncLN <- data.frame(eqInc = rlnorm((n - k[i]),
+                 meanlog, sdlog), region = levels(data$region)[i],
+                 weight = wval[i])
+             eqIncPopB <- rbind(eqIncLN, eqIncP)
+             if (i == 1)
+                 eqIncSample <- eqIncPopB
+             else eqIncSample <- rbind(eqIncSample, eqIncPopB)
+         }
+         eqIncSample <- data.frame(eqIncSample,
+             pID = c(1:dim(data)[1]))
+     }
+     else {
+         eqInc <- sort(data$eqInc)
+         n <- length(eqInc)
+         wval <- weightdata(data, region = FALSE, N = N)
+         xm <- eqInc[(n - k)]
+         xi <- eqInc[(n - k + 1):n]
+         if (method == "Hill")
```

```
+             theta <- hill(data, k)
+         if (method == "WML")
+             theta <- MTheta(data, k)
+         if (method == "PDC")
+             theta <- ThetaPDC(data, k)$minimum
+         eqIncP <- data.frame(eqInc = VGAM::rpareto(k, loc = xm,
+             shape = theta), region = "Austria", weight = wval)
+         m <- mean(eqInc[1:(n - k)])
+         v <- var(eqInc[1:(n - k)])
+         sdlog <- sqrt(log(v/m^2 + 1))
+         meanlog <- log(m) - sdlog^2/2
+         eqIncLN <- data.frame(eqInc = rlnorm((n - k), meanlog,
+             sdlog), region = "Austria", weight = wval)
+         eqIncSample <- rbind(eqIncLN, eqIncP)
+         eqIncSample <- data.frame(eqIncSample,
+             pID = c(1:dim(data)[1]))
+     }
+ }
```

A.8 Semi - parametric modeling

A.8.1 Estimation of the upper tail with a Pareto distribution

```
> fiteqIncPar <- function(data, N, k, method = "Hill", part = TRUE) {
+     if (part == TRUE) {
+         part <- split(data, list(data[, "region"]))
+         for (i in 1:length(levels(data$region))) {
+             eqInc <- part[[i]]$eqInc
+             n <- length(eqInc)
+             orddata <- order(eqInc)
+             ind <- orddata[(length(orddata)-k[i]+1):length(orddata)]
+             xm <- sort(eqInc)[(n - k[i])]
+             xi <- sort(eqInc)[(n - k[i] + 1):n]
+             if (method == "Hill")
+                 theta <- hill(part[[i]], k[i])
+             if (method == "WML")
+                 theta <- MTheta(part[[i]], k[i])
+             if (method == "PDC")
+                 theta <- ThetaPDC(part[[i]], k[i])$minimum
+             pardata <- VGAM::rpareto(k[i], xm, theta)
+             y <- eqInc
+             y[ind] <- pardata
```

```
+             eqIncR <- data.frame(eqInc = y,
+                 region = part[[i]]$region,
+                 weight = part[[i]]$weight, pID = part[[i]]$pID)
+             if (i == 1)
+                 eqIncSample <- eqIncR
+             else eqIncSample <- rbind(eqIncSample, eqIncR)
+         }
+     }
+     else {
+         eqInc <- data$eqInc
+         n <- length(eqInc)
+         xm <- sort(eqInc)[(n - k)]
+         xi <- sort(eqInc)[(n - k + 1):n]
+         orddata <- order(eqInc)
+         ind <- orddata[(length(orddata) - k + 1):length(orddata)]
+         if (method == "Hill")
+             theta <- hill(data, k)
+         if (method == "WML")
+             theta <- MTheta(data, k)
+         if (method == "PDC")
+             theta <- ThetaPDC(data, k)$minimum
+         pardata <- VGAM::rpareto(k, xm, theta)
+         y <- eqInc
+         y[ind] <- pardata
+         eqIncSample <- data.frame(eqInc = y, region = data$region,
+             weight = data$weight, pID = data$pID)
+     }
+ }
```

List of Tables

List of Figures

Bibliography

T. Atkinson, B. Cantillon, E. Marlier, and B. Nolan. *Social Indicators: The EU and Social Inclusion*. Oxford University Press, New York, 2002.

R. Bandourian, J.B. McDonald, and R.S. Turley. A comparison of parametric models of income distributions across countries and over time. Technical Report 305, Luxembourg Income Study Working Paper, 2002.

V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley & Sons, New York, 1981.

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

J. Beirlant, P. Vynckier, and J.L. Teugels. Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667, 1996b.

J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200, 1999.

J. Beirlant, G. Dierckx, A. Guilliou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180, 2002.

J. Beirlant, G. Dierckx, and A. Guillou. Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6):949–970, 2005.

M. Borkovec and C. Klüppelberg. Extremwerttheorie für Finanzzeitreihen - ein unverzichtbares Werkzeug im Risikomanagement. In *Handbuch Risikomanagement*, pages 219–242, Bad Soden, 2000. Uhlenbruch.

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

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

J. Caers, J. Beirlant, and P. Vynckier. Bootstrap confidence intervals for tail indices. *Computational Statistics & Data Analysis*, 26(3):259–277, January 1998.

A.J. Canty and A.C. Davison. Resampling-based variance estimation for labour force surveys. *The Statistician*, 48(3):379–391, 1999.

J.M. Chambers. *Software for Data Analysis: Programming with R*. Springer, New York, 2008.

R.L. Chambers. What if... ? robust prediction intervals for unbalanced samples. S3RI methodology working papers, Southampton Statistical Sciences Research Institute, 2005a.

R.L. Chambers. Calibrated weighting for small area estimation. S3RI methodology working papers, Southampton Statistical Sciences Research Institute, 2005b.

R.L. Chambers. Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069, 1986.

R.L. Chambers and N. Tzavidis. M-quantile models for small area estimation. S3RI methodology working papers, Southampton Statistical Sciences Research Institute, 2005.

C. Chen, W. Härdle, and A. Unwin. *Handbook of Data Visualization*. Springer, Berlin, 2008.

Council of Europe. *Concerted development of Social Cohesion Indicators: Methodological guide*. Council of Europe Publishing, Strasbourg, 2005.

F. A. Cowell. Measurement of Inequality. STICERD - Distributional Analysis Research Programme Papers 36, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE, 1998.

F.A. Cowell and E. Flachaire. Sensitivity of inequality measures to extreme values. STICERD - distributional analysis research programme papers, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE, 2002.

F.A. Cowell and E. Flachaire. Income distribution and inequality measurement : the problem of extreme values. Cahiers de la Maison des Sciences Economiques v04101, Universit Panthon-Sorbonne (Paris 1), July 2004.

F.A. Cowell and M.-P. Victoria-Feser. Distributional dominance with trimmed data. *Journal of Business & Economic Statistics*, 24(3):291–300, July 2006.

F.A. Cowell and M.-P. Victoria-Feser. Robust stochastic dominance: A semi-parametric approach. *Journal of Economic Inequality*, 5(1):21–37, April 2007.

F.A. Cowell and M.-P. Victoria-Feser. Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality*, 1(3): 191–219, December 2003.

E. L. Crow and K. Shimizu. *Lognormal Distributions: Theory and Applications*. Dekker, New York, 1988.

J. Danielsson, de Haan, L. Peng, and C.G. de Vries. Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. Econometric Institute Report 197, Erasmus University Rotterdam, Econometric Institute, 2000.

S.R. Dastrup, R. Hartshorn, and J.B. McDonald. The impact of taxes and transfer payments on the distribution of income: A parametric comparison. *Journal of Economic Inequality*, 5(3):353–369, December 2007.

R. Davidson and E. Flachaire. Asymptotic and bootstrap inference for inequality and poverty measures. Departmental Working Papers 2005-06, McGill University, Department of Economics, September 2006.

A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17(4):1833–1855, 1989.

J.C. Deville. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25(2):193–203, 1999.

H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172, 1998.

D.J. Dupuis. Extreme value theory based on the r largest annual events: a robust approach. *Journal of Hydrology*, 200(1-4):295–306, 1997.

D.J. Dupuis and C.A. Field. Robust estimation of extremes. *Canadian Journal of Statistics*, 26(2):199–215, 1998.

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

D.J. Dupuis and M.-P. Victoria-Feser. A prediction error criterion for choosing the lower quantile in Pareto index estimation. Cahiers de Recherche HEC 10, University of Geneva, 2003.

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

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

EUROSTAT. Common Cross-sectional EU indicators based on EU-SILC; the gender pay gap. Technical Report EU-SILC 131/04, Working Group on Statistics on Income and Living Conditions (EU-SILC), 2004a.

EUROSTAT. Description of target variables: Cross-sectional and Longitudinal. Technical Report EU-SILC 065/04, Working Group on Statistics on Income and Living Conditions (EU-SILC), 2004b.

Comparative EU statistics on Income and Living Conditions: Issues and Challenges, Methodologies and working papers, Luxembourg, 2007. EUROSTAT.

M. Ghosh and J.N.K Rao. Small Area Estimation: An Appraisal. *Statistical Science*, 9(1):55–76, 1994.

M. Graf. Use of Distributional Assumptions for the Comparison of four Laeken Indicators on EU-SILC Data. Swiss Federal Statistical Office, Statistical Methods Unit, November 2007.

A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of The Royal Statistical Society Series B*, 63(2):293–305, 2001.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W. Stahel. *Robust Statistics. The Approach Based on Influence Functions*. Wiley & Sons, New York, 1986.

P.J. Huber. *Robust Statistics*. Wiley & Sons, New York, 1981.

B. Hulliger. Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, pages 54–63. American Statistical Association, 1999.

B. Hulliger and R. Münnich. Variance estimation for complex surveys in the presence of outliers. In *Proceedings of the Section on Survey Research Methods*, pages 3153–3161. American Statistical Association, 2006.

B. Hulliger and T. Schoch. Robustification of the Quintile Share Ratio. Technical report, University of Applied Sciences Northwestern Switzerland, 2008.

S.P. Jenkins and P. Van Kerm. Trends in income inequality, pro-poor income growth and income mobility. Discussion Papers of DIW Berlin 377, DIW Berlin, German Institute for Economic Research, 2003.

S.P. Jenkins and P. Van Kerm. Accounting for income distribution trends: A density function decomposition approach. IZA Discussion Papers 1141, Institute for the Study of Labor (IZA), May 2004.

Ch. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley & Sons, New Jersey, 2003.

W. Kok. The lisbon strategy for growth and employment. High level group, Office for Official Publications of the European Communities, 2004.

M. F. Kratz and S. I. Resnick. The QQ-Estimator and Heavy Tails. ORIE Technical Reports, Cornell University Operations Research and Industrial Engineering, March 1995.

R. Lehtonen and R. Pahkinen. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley & Sons, Chichester, 2 edition, 2004.

R. Lehtonen and A. Veijanen. Design-based methods of estimation for domains and small areas. In D. Pfeffermann and J.N.K Rao, editors, *Sample Surveys: Inference and Analysis*, volume 29B of *Handbook of Statistics*. Elsevier, 2008.

R. Lehtonen, C.-E. Särndal, and A. Veijanen. The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29(1):33–44, 2003.

E. Limbert, W.A. Stahel, and M. Abbt. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51(5):341–352, May 2001.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theary and Methods*. Wiley & Sons, Chichester, 2006.

R. Münnich. Varianzschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, 37(3&4):319–334, 2008.

G. Osier. Variance estimation: The linearization approach applied by eurostat to the 2004 silc operation. Helsinki, 2006. Eurostat and Statistics Finland Methodological Workshop on EU-SILC.

D. Pfeffermann. Small are estimation - new developments and directions. *International Statistical Review*, 70(1):125–143, 2002.

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131, 1975.

R Development Core Team. An introduction to R. Technical Report V 2.9.0, R Development Core Team, 2009.

J.N.K Rao. *Small Area Estimation*. Wiley & Sons, New Jersey, 2003.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley & Sons, Chichester, 1987.

C.-E. Särndal, B. Swensson, and J.H. Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

D.W. Scott. Partial mixture estimator and outlier detection in data and regression. In *Theory and Applications of Recent Robust Methods*, pages 297–306, Basel, 2004. Birkhauser.

R. Serfling. Efficient and robust fitting of lognormal distributions. *North American Actuarial Journal*, 4(8):95–109, 2002.

R.G. Staudte and S.J. Sheather. *Robust estimation and testing*. Wiley & Sons, New York, 1990.

N. Tzavidis and R.L. Chambers. Robust prediction of small area means and distributions. CCSR working paper series, Centre for Census and Survey Research, 2007.

N. Tzavidis, N. Salvati, M. Pratesi, and R. Chambers. M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17(3):393–411, July 2008.

P. Van Kerm. Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. IRISS Working Paper 1, CEPS/INSTEAD, 2007.

B. Vandewalle, J. Beirlant, and M. Hubert. A Robust Estimator of the Tail Index based on an Exponential Regression Model. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology, pages 367–376. Birkhauser, Basel, 2004.

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

W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002.

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

R.R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego, 1997.

K.M. Wolter. *Introduction to Variance Estimation*. Springer, New York, 2007.