Walter, Paul; Weimer, Katja

Working Paper

Estimating poverty and inequality indicators using interval censored income data from the German microcensus

Discussion Paper, No. 2018/10

Provided in Cooperation with:
Free University Berlin, School of Business & Economics

Suggested Citation: Walter, Paul; Weimer, Katja (2018) : Estimating poverty and inequality indicators using interval censored income data from the German microcensus, Discussion Paper, No. 2018/10, Freie Universität Berlin, School of Business & Economics, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:188-refubium-22216-8

This Version is available at:
https://hdl.handle.net/10419/179926

Freie Universität Berlin

Estimating Poverty and Inequality Indicators using Interval Censored Income Data from the German Microcensus

Paul Walter
Katja Weimer

School of Business & Economics
Discussion Paper

Economics

2018/10

Estimating Poverty and Inequality Indicators using Interval Censored Income Data from the German Microcensus

Paul Walter* and Katja Weimer*

*Institute of Statistics and Econometrics, Freie Universität Berlin, Germany

Abstract

Rising poverty and inequality increases the risk of social instability in countries all around the world. For measuring poverty and inequality there exists a variety of statistical indicators. Estimating these indicators is trivial as long as the income variable is measured on a metric scale. However, estimation is not possible, using standard formulas, when the income variable is interval censored (or grouped), as in the German Microcensus. This is the case for numerous censuses due to confidentiality constraints or in order to decrease item non-response. To enable the estimation of statistical indicators in these scenarios, we propose an iterative kernel density algorithm that generates metric pseudo samples from the interval censored income variable. Based on these pseudo samples, poverty and inequality indicators are estimated. The standard errors of the indicators are estimated by a non-parametric bootstrap. Simulation results demonstrate that poverty and inequality indicators from interval censored data can be unbiasedly estimated by the proposed kernel density algorithm. Also the standard errors are correctly estimated by the non-parametric bootstrap. The kernel density algorithm is applied in this work to estimate regional poverty and inequality indicators from German Microcensus data. The results show the regional distribution of poverty and inequality in Germany.

Keywords: direct estimation, interval censored data, grouped data, poverty, inequality, kernel density estimation, German Microcensus

1 Introduction

In its Global Risks Report 2017, the World Economic Forum proclaims rising income and wealth disparity as the number one trend to determine global developments, governing the risks of, among others, profound social instability and unemployment (World Economic Forum, 2017). Also Germany has faced an increase in income inequality since its reunification in 1990 (Fuchs-Schündeln et al., 2010; Bönke et al., 2014). Yet, the question of how poverty and inequality is defined and can accurately be measured or diagnosed in a society remains debatable, see for example Lok-Dessallien (1999) and Hagenaars and Vos (1988). A common way to measure poverty and inequality is the estimation of statistical poverty and inequality indicators. However, computing them in practice is for several reasons not a trivial task. Since income information is not easily accessible governments or statistical offices need to conduct surveys or censuses to gain information about personal or household income. A difficulty remains the fact that income, in most societies, is considered a private topic. In the survey literature, questions about the aspects of income are referred to as a "sensitive question", therefore item non-response is high for these questions (Moore and Welniak, 2000; Hagenaars and Vos, 1988). To encounter this, many censuses, such as the German (Statistisches Bundesamt, 2017), the Australian (Australian Bureau of Statistics, 2011), the Columbian (Departamento Administrativo Nacional De Estadistica, 2005) and the Census from New Zealand (Statistics New Zealand, 2013), do not ask for the exact income of their citizens. They only ask for an income interval a person or household belongs to, thereby creating a sense of anonymity. The so obtained income data is not metric but rather interval censored (or grouped). This makes the use of standard formulas for the estimation of poverty and inequality indicators impossible because they rely on metric data.

To clarify terminology, depending on the author, the term grouped or censored income data can have different statistical meanings. Some authors such as Minoiu and Reddy (2008) and Milanovic (2003) use the term grouped data to refer to quantile means and Chotikapanich et al. (2007) consider population shares and class means. We use the term interval censored (or grouped) data, to refer to data that has the form of a frequency table, as in Chen (2017) or Hall and Wand (1996). This type of data is obtained by the aforementioned censuses.

A common parametric approach for density estimation from interval censored data is the use of the multinomial distribution, see for example Reed and Wu (2008) and Kleiber (2008). From the estimated parametric density any poverty and inequality indicator can be estimated. Chen (2017) proposes a generalised approach to multinomial maximum likelihood estimation for several types of grouped data, showing its consistency and supplying complementary simulation results.

With respect to inequality indicators, Kakwani and Podder (2008) argue against the parametric estimation of the income density from grouped data due to its lacking precision and present a method that can be utilised to estimate the Lorenz curve directly from the interval censored data in order to compute inequality indicators.

While many authors agree with Kakwani and Podder (2008) critique on the estimation of parametric distributions, they resolve these issues by using non-parametric estimators to model income instead. The popularity of these estimators comes from the fact that they do not require any prior assumptions about the theoretical distribution or its family. Although most authors do not directly address the topic of interval censored or grouped data, there is much literature about rounded data, which is easily obtained from interval censored data by substituting the intervals with their centres. Hall (1982), Scott and Sheather (1985), and Hall and Wand (1996) study the effects of rounded and interval censored data on standard, non-parametric kernel density estimation (KDE). In contrast to uncensored data, they derive that the mean integrated squared error of the KDE for rounded data depends on the smoothness of the used kernel function. Moreover, they find that censoring affects rather the bias than the variance of the estimate. Additionally, Hall and Wand (1996) present minimum grid sizes for KDE which are needed to achieve a given degree of accuracy. Grid size corresponds to the amount of points and therefore to the amount of intervals when the interval centres are used on which the density is estimated.

Wang and Wertelecki (2013) point out that standard KDE leads to increasingly spiky density estimates at rounded points with a growing sample size. KDE becomes smoother when larger bandwidths are used, thus an oversmoothed bandwidth selection has been proposed by Wand and Jones (1995) and implemented in the R package `KernSmooth` (Wand, 2015). Nevertheless, Wang and Wertelecki (2013) argue that this mostly leads to flatter estimates that underestimate the true density. Instead, they propose a bootstrap type kernel density estimator and show in a simulation study that the estimator provides better accuracy than standard KDE and over-smoothed KDE.

Groß et al. (2017) melt the principle of stochastic expectation maximization algorithms (Nielsen et al., 2000) with KDE to create a new density estimation alogrithm for rounded two-dimensional data. Its superiority compared to standard KDE is made apparent in a simulation study (Groß and Rendtel, 2016). Their algorithm can be seen as a generalisation of Wang and Wertelecki (2013) estimator.

Although a correctly estimated density leads to correctly estimated poverty and inequality indicators Lenau and Münnich (2016) focus their analysis on the impact of different estimation methods on the direct performance of the estimated statistical indicators. They evaluate three different estimation methods: Non-parametric splines, estimating the generalised beta distribution of the second kind (GB2) and linear interpolation. Linear interpolation is the method used by the German statistical offices to estimate indicators from interval censored data. This approach is similar to assuming a uniform distribution within each interval. They conclude that the performance of the different methods depends highly on the censoring schemes and none of the methods showed adequate results in terms of bias and variance for all analysed censoring schemes.

To overcome the disadvantage of the different estimation methods, we propose a non-parametric KDE-algorithm that is based on the algorithm of Groß et al. (2017). The KDE-algorithm enables the estimation of poverty and inequality indicators from interval censored data under different censoring schemes. In order to obtain representative results, the KDE-algorithm can incorporate survey weights. The standard errors of the statistical indicators are estimated by a non-parametric bootstrap.

The paper is structured as follows. In Section 2, the KDE-algorithm and the proposed non-parametric bootstrap are introduced. In Section 3, the properties of the KDE-algorithm and the bootstrap are evaluated using Monte Carlo simulation studies under different interval censoring schemes and different theoretical distributions. In Section 4, the algorithm is used to estimate regional poverty and inequality indicators from the German Microcensus. A final discussion of the major results, their implications and an outlook is given in Section 5.

2 Methodology

In order to estimate poverty and inequality indicators, we propose a novel KDE-algorithm to estimate metric pseudo samples from the observed interval censored true data. By using the pseudo samples, poverty and inequality indicators can be estimated applying standard formulas. In the next two subsections, the novel KDE-algorithm is introduced and a non-parametric bootstrap is proposed for variance estimation of the statistical indicators.

2.1 Kernel density estimation from interval censored data

Kernel density estimation is one of the most established non-parametric density estimation techniques in the literature and was first introduced by Rosenblatt (1956) and Parzen (1962). It is applied to estimate a continuous density from a random variable with density $f(x)$ directly from its independent and identically distributed observations without making any prior assumptions about its distributional family. Let $X = \{X_1, X_2, \ldots, X_n\}$ denote a sample of size n. For $i = 1, \ldots, n$ the KDE is defined as

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

where $k(\cdot)$ is a kernel function and the bandwidth is denoted by $h > 0$. For the shape and performance of the estimator, the choice of the bandwidth h is essential. The larger h, the smoother the estimated density, but also the more information about details, e.g. local extrema, may be lost (Zambom and Dias, 2012). Hence, bandwidth selection methods are widely discussed in literature with the two main categories being plug-in and cross-validation (Henderson and Parmeter, 2015; Jones et al., 1996; Loader, 1999). The basic idea of the first is to minimise the asymptotic mean integrated squared error whilst substituting the unknown density in the optimisation with a pilot estimate, whereas the second method is a more data-driven approach, for example utilising leave-one-out cross-validation.

In the presented *Naive* KDE, it is assumed that observations are taken directly from the continuous distribution that is to be estimated. Often, however, collecting continuous data is not possible due to various restrictions in practice, for example confidentiality concerns. In these situations we are left with interval cesnsored data, where only the interval information is observed. Thus, only the lower A_{K-1} and upper A_K interval bounds (A_{K-1}, A_K) of X is observed and its true value remains unknown. The continous scale is divided into n_K intervals. The variable K $(1 \leq K \leq n_K)$ indicates in which of the intervals an observation $K = \{K_1, K_2, \ldots, K_n\}$ falls into. Open ended intervals, thus $A_0 = -\infty$ or $A_{n_K} = \infty$ have to be replaced by a finite number (see Section 3.4). Applying KDE to the interval midpoints of the interval censored data falsely allocates too much probability mass to the midpoints and too little to the true unobserved X_i. This leads to spiky estimates, unless the bandwidth is chosen to be very large (Wang and Wertelecki, 2013). Increasing the bandwidth cannot be considered as a solution to this problem because this causes additional loss of information about the underlying true distribution. Wang and Wertelecki (2013) simulation study further found standard KDE to be very sensitive to sample size when interval censoring is ignored. Furthermore, Hall (1982) and Hall and Wand (1996) showed that, in contrast to uncensored data, the asymptotic performance of KDE for interval censored data depends on the smoothness of the kernel function in use.

These findings underline the necessity of using a more sophisticated non-parametric approach for density estimation from interval censored data. Wang and Wertelecki (2013) introduce a bootstrap type KDE based on a measurement error model and confirmed its superiority over the *Naive* estimator with

simulations. Groß et al. (2017) then generalised and extended the approach based on Stochastic Expectation Maximization (SEM) and iterative bootstrapping. Their newly proposed density estimator, abbreviated as GRSST, outperforms *Naive* KDE and a measurement error based estimator by Delaigle (2007), especially for stronger interval censoring. Since the GRSST estimator was formulated for two dimensional data with equally sized interval censoring, we reformulate the approach. The reformulated KDE-algorithm enables the density estimation for one dimensional data with unequally sized censoring. During the algorithm pseudo samples of the unobserved X_i are generated from whom the density and any statistical indicator can be estimated. Hence, for the estimation of poverty and inequality indicators the unobserved continuous distribution of the interval censored X is reconstructed. This is done by the use of the known interval information K. From Bayes theorem it follows that the conditional distribution of X given K is:

$$\pi(X|K) \propto \pi(K|X)\pi(X),$$

where $\pi(K|X)$ is defined by a product of Dirac distributions $\pi(K|X) = \prod_{i=1}^{n} \pi(K_i|X_i)$ with

$$\pi(K|X) = \begin{cases} 1 & \text{if } A_{K-1} \leq X_i \leq A_K, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$. By this formulation pseudo samples (imputations) of the unknown X_i are drawn that enable the estimation of any statistical indicator. Since $\pi(X) = \prod_{i=1}^{n} f(X_i)$ is initially unknown, an initializing estimate $\hat{f}_h(x)$ that is based on the interval midpoints, serves as a proxy. After that, the pseudo samples drawn from $\pi(X|K)$ are used to re-estimate $\pi(X)$. The following section focusses on the exact implementation of the proposed algorithm and discusses similarities to the popular EM-algorithm by Dempster et al. (1977) and the SEM-algorithm by Celeux and Dieboldt (1985) and Celeux et al. (1996).

Estimation and Computational Details

As in Groß et al. (2017) for fitting the model pseudosamples of X_i are drawn from the conditional distribution

$$\pi(X_i|K_i) \propto \mathbf{I}(A_{K-1} \leq X_i \leq A_K) f(X_i),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The conditional distribution of X_i given K_i is the product of a uniform distribution and density $f(x)$. As the density $f(x)$ is unknown it is replaced by $\hat{f}_h(x)$, an estimate that is obtained by the prior defined kernel density estimator. Hence, X_i is iteratively drawn from the known interval (A_{K-1}, A_K) with the current density estimate $\hat{f}_h(x)$ used as sampling weight. The steps of the iterative algorithm are described below.

Step 1: Use the midpoints of the intervals as pseudo $\tilde{X}_i$ for the unknown X_i. Obtain a pilot estimate of $\hat{f}_h(x)$, by applying KDE. Choose a sufficiently large bandwidth h, such that no rounding spikes occur.

Step 2: Evaluate $\hat{f}_h(x)$ on an equally spaced fine grid $G = \{g_1, \ldots, g_j\}$ with j grid points $g_1, \ldots, g_j$. The width of the grid is denoted by δ_g. It is given by:

$$\delta_g = \frac{|A_0 - A_{n_K}|}{j - 1}$$

The grid is defined as:

$$G = \{g_1 = A_0, g_2 = A_0 + \delta_g, g_3 = A_0 + 2\delta_g, \ldots, g_{j-1} = A_0 + (j-2)\delta_g, g_j = A_{n_K}\}$$

Step 3: Sample from $\pi(X|K)$ by drawing a pseudo sample $\tilde{X}_i$ randomly from $G \in (A_{K-1}, A_K)$ with sampling weights $\hat{f}_h(\tilde{X}_i)$ for $k = 1, \ldots, n_K$. The sample size within each interval is given by the number of observations within each interval.

Step 4: Estimate any statistical indicator of interest $\hat{I}$ using the pseudo $\tilde{X}_i$.

Step 5: Recompute $\hat{f}_h(x)$, using the pseudo samples $\tilde{X}_i$ obtained in iteration step 3.

Step 6: Repeat steps 2-5, with $B_{(KDE)}$ burn-in and $S_{(KDE)}$ additional iterations.

Step 7: Discard the $B_{(KDE)}$ burn-in iterations and estimate $\hat{I}$ by averaging the obtained $S_{(KDE)}$ estimates.

The KDE-algorithm estimates the distribution of an interval censored variable by only using the interval information. An algorithm that is widely used for models that depend on latent variables (in our case the unobserved interval censored X) is the EM-algorithm (Dempster et al., 1977). In the EM-algorithm the expectation of $X|K$ is obtained analytically. However, in the context of kernel density estimation this does not work because all values inside an interval would be concentrated at one point, the expectation. In a SEM-algorithm, the analytical E-step from the EM-algorithm is replaced by the drawing of pseudo samples (Celeux and Dieboldt, 1985; Celeux et al., 1996). In case of the KDE-algorithm, it is drawn from the distribution of $\pi(X|K)$. Hence, the proposed KDE-algorithm has similarities to a SEM-algorithm. In its common form, the EM- and SEM-algorithm are used for maximum likelihood (ML) estimation with unobserved data. McLachlan and Krishnan (2008) proposed a generalisation of the SEM-algorithm that can be used with surrogates for the ML estimation. In the KDE-algorithm the maximization of the asymptotic mean integrated squared error is used as such a surrogate. More information on the similarities between the KDE-algorithm, the EM-, SEM- algorithm and the GRSST estimator -on which the KDE-algorithm is based on- are given in (Groß et al., 2017).

2.2 Variance estimation

This section introduces a method for variance estimation of the statistical indicators that are estimated by the KDE-algorithm. A common way to estimate the variance, if X is observed on a continuous scale is linearisation. Taylor linearisation (Tepping, 1968; Woodruff, 1971; Wolter, 1985; Tille, 2001) is a well known and commonly applied method for the estimation of variance for non-linear indicators, such as ratios or correlations. However, the method cannot be applied for variance estimation of all non-linear indicators. For variance estimation of mathematically more complex indicators, e.g. the Gini, Deville (1999) introduced the generalised linearisation method. The generalised linearisation method is also used by Eurostat for the variance estimation of complex indicators (Osier, 2009). Nevertheless, linearisation cannot be applied when the variable of interest is observed as interval censored variable (Lenau and Münnich, 2016). To still produce variance estimates, resampling methods, such as bootstrapping can be applied (Münnich, 2008). Bootstrapping methods approximate the variance of an estimated indicator, in cases where the variance cannot be stated as closed form solution (Bruch et al., 2011). Therefore, the bootstrap introduced by Efron (1979), Shao and Tu (1995) is used for the variance estimation of the indicators estimated by the KDE-algorithm. Also any confidence interval can be estimated by using the quantiles from the bootstrap results (Pretson, 2008; Rao and Wu, 1988; Rao et al., 1992). The use of the bootstrapping permits to avoid theoretical calculations. However, the potential disadvantage is a long computational time. The non-parametric bootstrap is based on the assumption that the drawn sample is representative for the population. Therefore, the empirical distribution function $\hat{F}$ is a non-parametric estimate of the population distribution F. The desired poverty indicator of interest $\hat{I}$, is the empirical estimate of the true parameter. The bootstrap standard errors are calculated in the following way:

Step 1: Draw with replacement a bootstrap sample of the interval censored $X_i^{(b)}$ of size n from the sample dataset.

Step 2: Apply the KDE-algorithm to the bootstrap sample $X_i^{(b)}$ for the estimation of any indicator $\hat{I}^{(b)}$ of interest.

Iterate steps 1-2, $b = 1, \ldots, B$ times and estimate the standard error $s(\hat{I}) = \sqrt{\frac{\sum_{b=1}^{B}(\hat{I}^{(b)} - \overline{I}^{(b)})^2}{B}}$ with $\overline{I} = \frac{1}{B}\sum_{b=1}^{B}\hat{I}^{(b)}$

3 Simulation Results

This section presents extensive model-based simulation results in order to evaluate the performance of the KDE-algorithm in the context of estimating poverty and inequality from interval censored income data. The simulation study is set up with the following specifications. From a theoretical distribution $M = 500$ samples of simulated income data are drawn. The drawn samples are censored to specific intervals. The sample size for each sample is $n = 10000$. The KDE-algorithm is evaluated for large samples because interval censored income data is common for censuses which, in general, have very large sample sizes. For instance, in the application in Section 4, German Microcensus data is used which has a sample size of $n = 454852$. From the simulated interval censored income data different poverty and inequality indicators are estimated. The formulas are presented for metric data because the KDE-algorithm generates metric data from interval censored data that is used to estimate the statistical indicators. Consider $X = (X_1, \ldots, X_n)$ with $X_1 \leq \ldots \leq X_n$ and let $w = (w_1, \ldots, w_n)$ be the corresponding sampling weights. The weighted mean and the weighted quantiles (10%, 25%, 50%, 75%, 90%) are given by

$$\hat{I}_{\text{Mean}} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \tag{1}$$

$$\hat{I}_{\text{Q}_{(p)}} = \begin{cases} \frac{1}{2}(X_i + X_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = p \sum_{j=1}^{n} w_j; \\ X_{i+1} & \text{if } \sum_{j=1}^{i} w_j \leq p \sum_{j=1}^{n} w_j \leq \sum_{j=1}^{i+1} w_j, \end{cases} \tag{2}$$

where p denotes the quantile $p \in (0, 1)$. In the simulation study sampling weights are not included, because they are not needed to evaluate the performance of the KDE-algorithm. Therefore, $w_i = 1 \forall i$ in the simulation study. However, in the application in Section 4 weights are included for representative inference. The weighted poverty measures Headcount Ratio (HCR) and Poverty Gap (PGap) (Foster et al., 1984) are given by

$$\hat{I}_{\text{HCR}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \mathbf{I}(X_i \leq z), \tag{3}$$

$$\hat{I}_{\text{PGap}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(\frac{z - X_i}{z}\right) \mathbf{I}(X_i \leq z), \tag{4}$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The HCR and PGap include a threshold z that is known as the poverty line. For the simulation a relative poverty line, defined as 60% of the median of the simulated income variable is chosen. This corresponds to the EU definition (eurostat, 2014). The HCR is a measure for the percentage of observations (individuals or households) below the poverty line, whereas the PGap measures the average distance of those observations from the poverty line. Inequality is commonly measured by the Gini coefficient (Gini, 1912) and the quintile share ratio (QSR). The weighted indicators are estimated by

$$\hat{I}_{\text{Gini}} = \left[\frac{2\sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} w_i^2 X_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i X_i} - 1 \right], \tag{5}$$

$$\hat{I}_{\text{QSR}} = \frac{\sum_{i=1}^{n} \mathbf{I}(X_i \geq \hat{Q}_{0.8}) w_i X_i}{\sum_{i=1}^{n} \mathbf{I}(X_i \leq \hat{Q}_{0.2}) w_i X_i}. \tag{6}$$

The range of the Gini coefficient lies between 0 and 1. The higher its value, the higher the inequality. If the Gini is equal to 0 there is perfect equality in the data, whereas a Gini of 1 indicates perfect inequality. The QSR is the ratio of observations richer than 20% of the richest observations to the 20% of the poorest observations. Higher values of the QSR indicate higher inequality.

The indicators are estimated by the proposed KDE-algorithm (KDE). The number of burn-in iterations of the algorithm is set to $B_{(KDE)} = 80$, the number of additional iterations $S_{(KDE)} = 400$. Our experiences running several simulations show that 480 iterations are usually enough to ensure convergence. Nevertheless, we assure that the indicators in the presented simulations have converged by visually checking the convergence plots. The issue of convergence is discussed in more detail in Section 4. The number of grid points is set to $j = 4000$. In general, a higher number of grid points leads to more precise estimation results, because the number of grid points determines how many unique values the pseudo samples of the interval censored variable can consist of. However, the estimation time increases with increasing number of grid points. In the simulation, the number of grid points is chosen such that a further increase of the number of grid points does not lead to better estimation results. The presented poverty and inequality indicators are not only estimated by the KDE-algorithm (KDE). For comparison, the indicators are also estimated by linear interpolation. This method is used by the Federal Office of Statistics in Germany for the estimation of poverty and inequality indicators from the interval censored income variable of the German Microcensus (Information und Technik (NRW), 2009). This approach gives the same results as assuming a uniform distribution within the income classes (Uni). Furthermore, the statistical indicators are estimated by using the midpoints (Mid) of the intervals as proxy for the unobserved values within the income interval. Finally, the statistical indicators are also estimated with the true uncensored data (True). The results of the point estimates are evaluated by the relative bias (rB)

$$rB(\hat{I}) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{I}_m - I}{I}\right) \times 100,$$

and the empirical standard errors (se.emp)

$$se.emp(\hat{I}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{I}_m - \overline{I})^2},$$

with

$$\overline{I} = \frac{1}{M}\sum_{m=1}^{M}\hat{I}_m.$$

The proposed non-parametric bootstrap for the estimation of the standard errors is evaluated by comparing the estimated standard errors to the empirical standard errors. The bootstrap is run with $B = 100$. This number shows to be sufficient to obtain valid approximations of the standard errors.

The simulation study is divided into four subsections. In Section 3.1, the influence of different numbers of intervals on the performance of the KDE-algorithm is evaluated. In Section 3.2, different true distributions are evaluated and, in Section 3.3, the effect of equal vs. ascending interval width is studied. Section 3.4 summarizes the final results and discusses the issue of how to handle open ended intervals.

3.1 Different interval censoring scenarios

In this Section, the influence of the number of intervals on the performance of the KDE-algorithm is studied. As theoretical distribution the four parameter GB2 distributions that is often used to model income is specified such that the GB2 distribution well approximates the empirical German income distribution (Graf and Nedyalkova, 2014). The chosen parameter are given in Table 3. The drawn samples are interval censored using three different censoring scenarios. In scenario 1, the data is censored to 24 intervals as in the German Microcensus (Statistisches Bundesamt, 2017) that is used in the application in Section 4. The interval widths are chosen such that the interval censored theoretical distribution follows

the empirical distribution of the household income in the German Microcensus. This is visualized in Figure 1 in the upper two panels. The lower two panels show the GB2 distribution censored to 16 intervals (scenario 2) and 8 intervals (scenario 3). The performance of the algorithm with lower number of classes is studied because censuses from other countries censor the income variable to fewer than 24 intervals. For example, in the census from New Zealand the income variable is censored to 16 intervals (Statistics New Zealand, 2013), in the Australian census the data is censored to 12 intervals (Australian Bureau of Statistics, 2011) and in the Columbian census the income variable is censored to only 9 intervals (Departamento Administrativo Nacional De Estadistica, 2005).

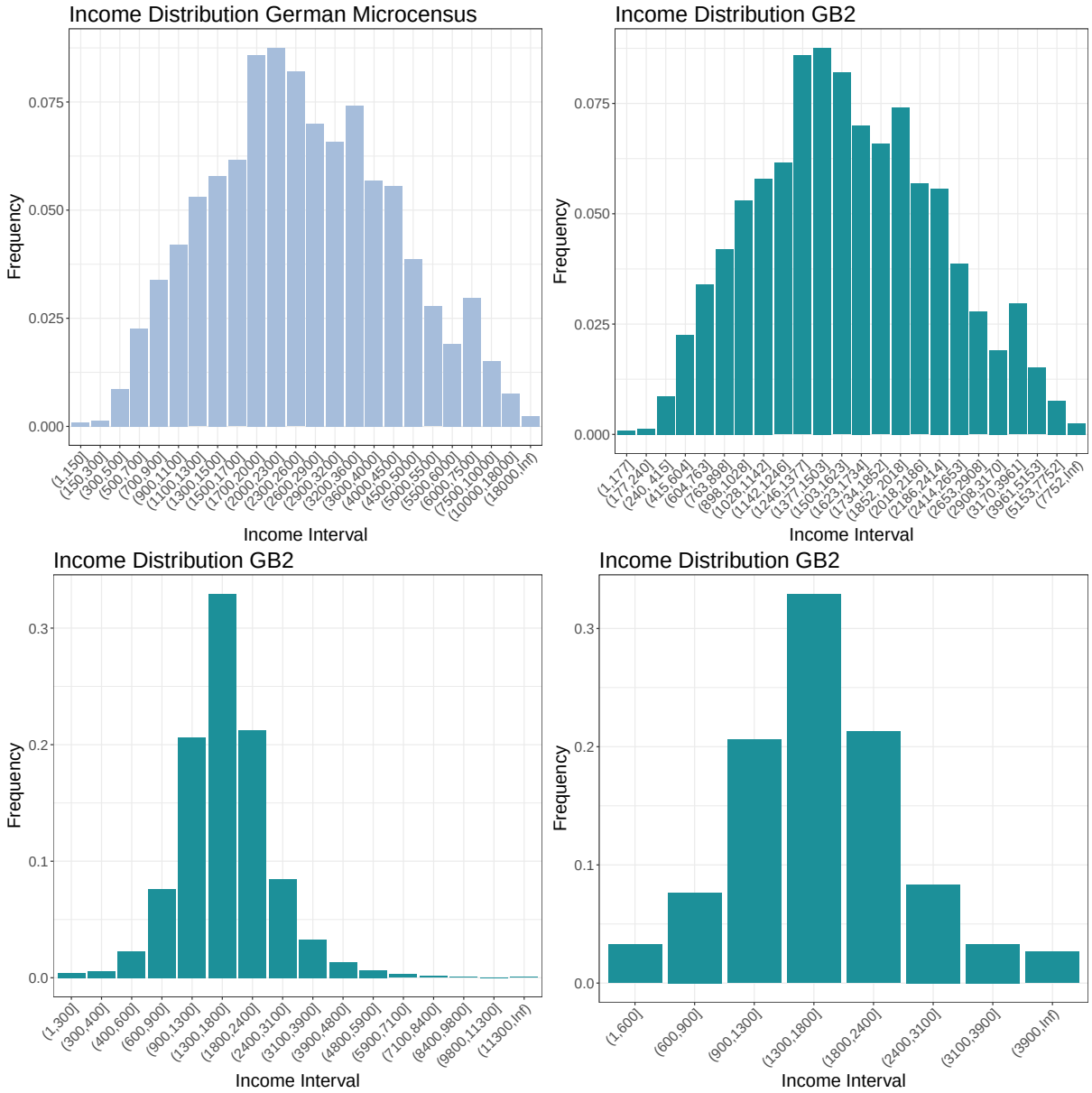


Figure 1: Interval censored income distribution of the German Microcensus (upper left) and theoretical GB2 distribution. The GB2 distribution is censored to 24 (upper right), 16 (lower left) and 8 intervals (lower right).

The results of the point estimates are given in Table 1. Using the true uncensored data for the estimation of the poverty and inequality indicators leads to unbiased results. This is not surprising, because the sample size ($n = 10000$) is very large. Using only the interval information, the KDE-algorithm outperforms the other approaches (Mid and Uni) in all three scenarios. The out-performance is especially stronger for indicators that rely on the whole shape of the distribution (Gini, Mean), for the more extreme quantiles (10% Quantile and 90 % Quantile) and for indicators that rely on more extreme quantiles (QSR). As the number of intervals decreases, the performance of the KDE-algorithm worsens.

Nevertheless, the bias is still under 1% for all indicators, except for the PGap and the Gini. The PGap shows a bias of 2.3% and the Gini a bias of -1.9% in the 8 interval scenario.

The estimated indicators using the other approaches (Mid and Uni) exhibit far larger biases as the number of intervals decreases. For example, in the 8 interval scenario the PGap has a bias of 22% and 20% and the Gini of 14% and 24% for the estimation approaches Uni and Mid, respectively. This shows the superiority of the KDE-algorithm.

The precision of the KDE-algorithm, measured by the empirical standard error (se.emp), is for all three scenarios close to the estimation results using the true data. This is the case, because the estimated indicators rely on the metric pseudo samples from the KDE-algorithm. However, the pseudo samples can -in rare circumstances- include very extreme values that lead to a higher variance when statistical indicators are estimated that rely on the whole distribution. This is for example the case for the mean in the 24 interval scenario. The KDE-algorithm almost looses no precision for a lower number of intervals. The methods Uni and Mid lead to less precise estimation results, especially with fewer intervals. For some of the estimated quantiles the empirical standard error of the Mid approach is 0. This is due to the fact, that the Mid approach estimates the indicators on the midpoints of the intervals. This leads to only 24, 16 or 8 unique values, respectively. With a sample size of ($n = 10000$) the estimated quantiles are likely to fall on the same midpoint for each of the 500 Monte-Carlo iterations. If the estimated quantile is constant over all Monte-Carlo iterations, the empirical standard error is 0.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method					**GB2: 24 intervals**					
rB	True	0.053	0.036	0.008	-0.003	0.017	0.023	-0.087	-0.005	-0.163	-0.005
	KDE	-0.102	-0.059	-0.033	-0.045	0.121	0.002	-0.141	0.720	0.181	-0.036
	Uni	-0.366	-0.086	0.065	0.080	0.171	1.104	1.087	3.751	2.628	3.374
	Mid	-4.654	0.003	-0.313	1.501	1.848	2.218	-11.962	35.517	1.529	6.161
se.emp	True	87.600	72.172	71.259	109.180	222.019	95.973	0.003	0.049	0.001	0.003
	KDE	84.944	68.284	69.756	112.048	227.883	121.231	0.003	0.067	0.001	0.004
	Uni	96.181	69.987	70.633	119.183	240.357	111.912	0.003	0.060	0.001	0.003
	Mid	83.717	0.000	0.000	738.583	1092.148	137.517	0.003	0.351	0.001	0.005
						GB2: 16 intervals					
rB	True	-0.007	0.012	0.022	0.021	0.014	-0.020	-0.030	-0.077	0.109	-0.102
	KDE	0.323	-0.021	0.260	0.190	-0.051	-0.018	0.478	0.699	0.034	-0.401
	Uni	-0.991	-1.832	0.823	3.492	3.543	1.154	4.522	5.113	7.699	3.691
	Mid	-14.210	-8.097	-1.200	3.499	3.098	1.536	-12.619	92.185	6.194	0.835
se.emp	True	90.029	72.505	78.428	113.178	232.863	101.242	0.003	0.048	0.001	0.003
	KDE	88.476	72.731	73.944	119.657	229.199	101.652	0.003	0.049	0.001	0.003
	Uni	120.142	84.036	81.005	131.425	248.381	110.794	0.003	0.055	0.001	0.003
	Mid	221.137	0.000	0.000	0.000	0.000	121.311	0.003	0.321	0.001	0.004
						GB2: 8 intervals					
rB	True	0.076	0.006	-0.016	0.021	0.017	-0.006	-0.103	-0.051	-0.131	-0.037
	KDE	0.106	-0.173	0.252	0.145	-0.141	-0.685	0.119	-1.151	2.329	-1.871
	Uni	-0.980	-1.850	0.820	3.519	3.587	4.190	4.323	17.586	21.758	13.522
	Mid	-13.972	-8.012	-1.155	3.582	3.092	10.187	-12.555	164.261	20.273	24.256
se.emp	True	92.276	75.720	71.976	111.044	240.443	100.286	0.003	0.050	0.001	0.003
	KDE	88.373	74.822	70.126	113.115	231.700	126.809	0.003	0.071	0.001	0.004
	Uni	120.998	86.888	73.876	128.360	253.586	132.150	0.003	0.075	0.001	0.004
	Mid	220.916	0.000	0.000	0.000	0.000	183.278	0.003	0.481	0.001	0.005

Table 1: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

In Table 2, the proposed bootstrap for the estimation of the standard errors is evaluated for the 3 different censoring scenarios. The standard errors estimated by the non-parametric bootstrap (se.est) offer a good approximation of the empirical standard errors (se.emp). This underlines the reliability of the proposed bootstrap method.

3.2 Different true distributions

While the previous Section evaluates the performance of the KDE-algorithm using different censoring schemes, this Section focuses on the evaluation of the performance using different theoretical distributions. A large number of theoretical distributions are suggested in the literature for modelling income

Measure	Estimation Method	$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
		GB2: 24 intervals									
se.emp	KDE	84.944	68.284	69.756	112.048	227.883	121.231	0.003	0.067	0.001	0.004
se.est		84.945	71.525	72.437	110.804	234.200	120.855	0.003	0.067	0.001	0.004
		GB2: 16 intervals									
se.emp	KDE	88.476	72.731	73.944	119.657	229.199	101.652	0.003	0.049	0.001	0.003
se.est		87.972	70.564	68.708	110.969	224.122	96.000	0.003	0.050	0.001	0.003
		GB2: 8 intervals									
se.emp	KDE	88.373	74.822	70.126	113.115	231.700	126.809	0.003	0.071	0.001	0.004
se.est		85.036	71.131	68.217	109.751	229.160	132.415	0.003	0.076	0.001	0.005

Table 2: Empirical and estimated standard error for the selected statistical indicators.

distributions (McDonald and Ransom, 1979; McDonald, 1984; McDonald and Xu, 1995; Bandourian et al., 2003; Kleiber and Kotz, 2003). According to McDonald (1984), McDonald and Xu (1995), Bordley et al. (1997), McDonald and Ransom (2008) the GB2 distribution is well suited for modelling income and it is superior to other parametric distributions (Kleiber and Kotz, 2003; Dastrup et al., 2007; Jenkins, 2009). Nevertheless, two special cases of the GB2 distribution are used for evaluations in order to illustrate the flexibility of the KDE-algorithm: the Dagum (Dagum, 1977) distribution and the Singh-Maddala (Singh and Maddala, 1976) distribution. The choice of parameters follows Bandourian et al. (2002) (see Table 3) in order to approximate empirical income distributions. The data is censored to 8 intervals and the interval width is chosen such that the relative frequency within each interval is similar to the 8 interval GB2 scenario from the previous section (Figure 2 and 1). The 8 interval scenario is chosen to evaluate the KDE-algorithm under extreme scenarios. By keeping the relative frequencies equal within each interval the effect of different distributions (GB2, Dagum and Singh-Maddala) on the estimation results is isolatedly evaluated.

Distribution	Parameter			
GB2	7.481	16351	0.4	0.468
Dagum	4.413	94075	0.337	
Singh-Maddala	1.771	500000	25.12	

Table 3: Distributions for the Model-based simulation

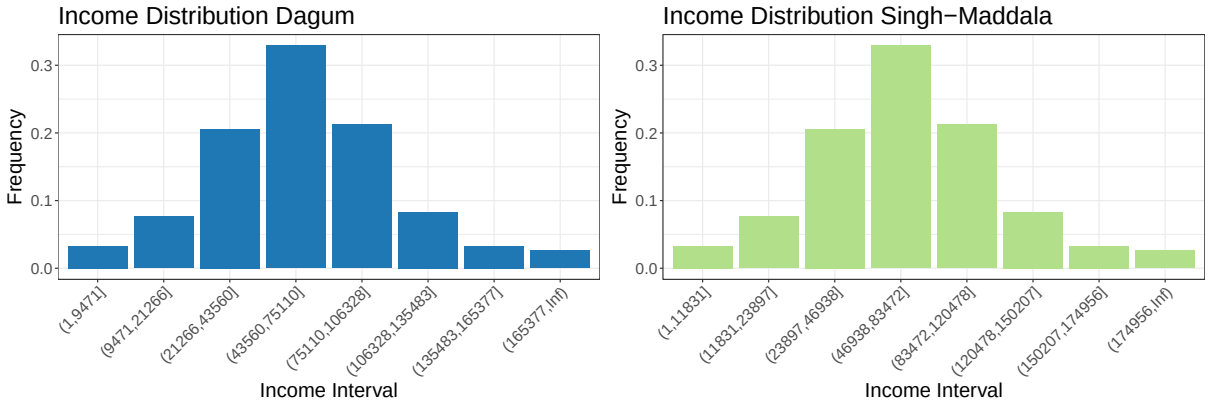


Figure 2: Dagum and Singh-Maddala distribution censored to 8 intervals.

The estimation results of the point estimates are given in Table 4 (Dagum and Singh-Maddala) and Table 1 (GB2). As expected, using the true data leads to unbiased estimation results. Also, the KDE-algorithm that only uses the interval information yields unbiased results for all indicators under the different scenarios. Hence, the performance of the KDE-algorithm is not impaired by the underlying theoretical distribution. The benchmark methods (Uni and Mid) give heavily biased estimation results, especially for indicators that depend on the whole distribution. For example, the QSR has a bias of

16.5% (Uni) and 209% (Mid) for the Dagum scenario and 18.5% (Uni) and 199% (Mid) for the Singh-Maddala scenario. These simulation results disqualify both estimation methods for the use in practical applications. Regarding the precision, the conclusions from the previous Section are transferable.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method	**Dagum: 8 intervals**									
rB	True	0.041	-0.014	0.020	0.003	0.005	0.015	0.032	0.072	0.036	0.028
	KDE	0.192	0.088	-0.146	0.225	0.038	-0.396	-0.126	-0.770	-0.084	-0.851
	Uni	-0.977	-1.719	0.675	3.150	2.883	5.454	2.579	16.532	4.163	9.840
	Mid	-23.304	-12.787	-2.552	3.227	2.420	12.042	29.230	209.641	-2.171	16.251
se.emp	True	399.449	437.440	455.249	584.052	988.153	442.182	0.004	0.128	0.002	0.003
	KDE	382.632	422.677	440.771	567.565	964.208	479.943	0.004	0.135	0.002	0.003
	Uni	459.406	461.163	456.904	645.016	1052.903	613.491	0.004	0.171	0.002	0.004
	Mid	0.000	0.000	0.000	0.000	0.000	826.842	0.005	1.024	0.002	0.005
		Singh-Maddala: 8 intervals									
rB	True	-0.070	0.001	0.035	0.014	-0.015	0.003	0.023	0.017	0.041	-0.006
	KDE	0.270	0.014	0.042	-0.039	-0.031	0.093	-0.039	0.714	0.085	0.213
	Uni	-1.031	-1.210	1.652	2.963	2.039	6.269	1.800	18.504	4.321	11.024
	Mid	-21.083	-11.797	-1.789	3.039	1.636	12.618	27.516	199.584	-1.651	17.009
se.emp	True	416.957	486.609	555.653	731.369	1049.186	443.818	0.004	0.099	0.002	0.002
	KDE	389.926	447.684	546.007	698.835	998.289	462.384	0.004	0.106	0.002	0.002
	Uni	467.696	502.097	547.127	784.601	1072.791	598.248	0.004	0.145	0.002	0.003
	Mid	784.213	0.000	0.000	0.000	0.000	784.707	0.005	0.930	0.002	0.004

Table 4: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

As given in Table 5, the estimated standard errors offer a good approximation of the empirical standard errors for the different scenarios.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Measure	Estimation Method	**Dagum: 8 intervals**									
se.emp	KDE	382.632	422.677	440.771	567.565	964.208	479.943	0.004	0.135	0.002	0.003
se.est		385.340	420.523	445.765	573.573	953.225	468.896	0.004	0.134	0.002	0.003
		Singh-Maddala: 8 intervals									
se.emp	KDE	389.926	447.684	546.007	698.835	998.289	462.384	0.004	0.106	0.002	0.002
se.est		386.539	430.594	523.137	691.090	983.671	460.726	0.004	0.110	0.002	0.002

Table 5: Empirical and estimated standard error for the selected statistical indicators.

3.3 Equal and ascending interval width

While the German (Statistisches Bundesamt, 2017), the Australian (Australian Bureau of Statistics, 2011), the Columbian (Departamento Administrativo Nacional De Estadistica, 2005) and the Census from New Zealand (Statistics New Zealand, 2013) use ascending class width, previous research shows that the performance of alternative estimation methods depends on the interval width (Lenau and Münnich, 2016). More precisely, performance depends on whether the data is censored to intervals of equal width or ascending width. Therefore, the GB2 distribution from Table 3 is now censored to 8 intervals with equal class width (except the last interval, which has an open ended upper interval bound). In all previous simulation scenarios ascending interval width is used. Figure 3 shows the censored GB2 distribution. The theoretical distribution is kept fixed in order to evaluate the influence of the censoring on the performance.

The results of the point estimates are given in Table 6. As before, using the true data leads to unbiased estimates. The estimates obtained by the KDE-algorithm are unbiased except for the QSR, PGap and Gini. These estimates exhibit a very small bias of -1.7%, 1.4% and -2.2%. However, the results are comparable to the estimation results from the GB2 scenario with 8 intervals with ascending interval width. Hence, the KDE-algorithm does not seem to be effected by the censoring scheme. The benchmark indicators Uni and Mid show, as before, large biases especially for indicators that rely on the whole shape of the distribution. With regard to precision, results and interpretation is the same as before.

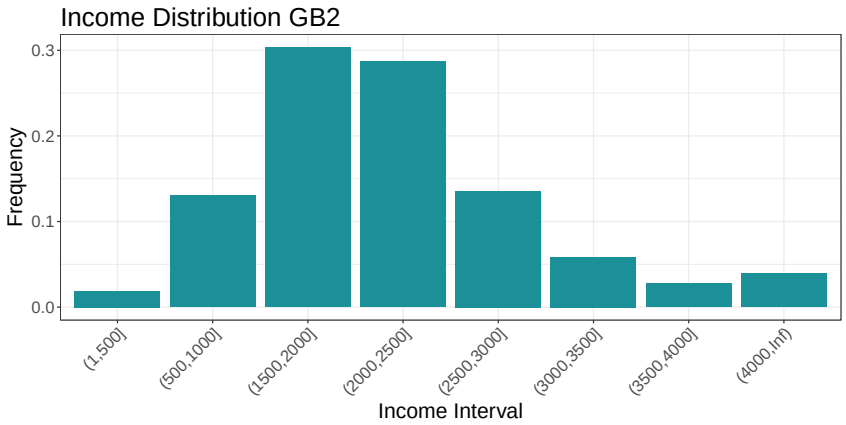


Figure 3: GB2 distribution censored to equally sized intervals (except the last -open ended- interval)

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method	**GB2: 8 intervals (equally sized)**									
rB	True	0.079	0.035	0.013	-0.026	-0.092	-0.024	-0.127	-0.144	-0.248	-0.110
	KDE	-0.005	-0.422	0.238	-0.066	0.050	-0.840	0.290	-1.706	1.370	-2.181
	Uni	-7.074	-2.388	0.909	0.560	1.704	4.648	7.351	21.365	30.052	16.251
	Mid	-14.151	4.640	11.598	10.730	3.174	12.498	19.720	73.226	28.594	30.467
se.emp	True	88.841	75.061	72.038	111.139	233.943	95.398	0.003	0.051	0.001	0.003
	KDE	86.255	70.621	76.142	109.506	223.898	128.012	0.003	0.076	0.001	0.005
	Uni	116.469	70.955	88.391	156.503	260.393	130.810	0.003	0.076	0.001	0.004
	Mid	0.000	0.000	0.000	544.426	0.000	180.793	0.004	0.281	0.001	0.005

Table 6: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

The proposed bootstrap also gives valid results with equally sized intervals (see Table 7).

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Measure	Estimation Method	**GB2: 8 intervals (equally sized)**									
se.emp	KDE	86.255	70.621	76.142	109.506	223.898	128.012	0.003	0.076	0.001	0.005
se.est		84.456	67.507	75.134	108.778	224.587	138.326	0.003	0.079	0.001	0.005

Table 7: Empirical and estimated standard error for the selected statistical indicators.

3.4 Conclusion and final remarks

The simulation results show that the KDE-algorithm outperforms other approaches (Uni and Mid) in terms of bias in all scenarios. The KDE method gives unbiased results under different censoring schemes and for different underlying theoretical distributions. The relative bias increases slightly whenever the number of intervals decreases. However, also in very extreme censoring scenarios (with only 8 intervals), the results are very precise. The relative bias is under 1% for almost all indicators. The KDE method shows comparable results in terms of precision to the direct estimation of the indicators from the true uncensored data. Additionally, it is superior to other approaches (Mid and Uni) that show worse precision for most indicators. Due to its easy usage, its ability to adapt to different underlying theoretical distributions and different censoring schemes and its precision practitioners should prefer the KDE-algorithm to other approaches.

The KDE-algorithm cannot handle open ended intervals. As mentioned before, lower bounds equal to $-\infty$ or upper bounds equal to ∞ have to be replaced by a finite number. The chosen value effects the performance of the KDE-algorithm. However, not all poverty and inequality indicators depend on the outer intervals. Indicators that depend on the outer intervals are indicators that depend, by their definition, on the whole distribution e.g., the mean or the Gini. These indicators are always influenced by the way how open ended intervals are handled, whereas other indicators, e.g. the median, are only affected if they fall into one of the open ended outer intervals. The replacement value used for open

ended upper and lower intervals has also an impact on the performance of the methods Uni and Mid. To make simulation results from the different estimation methods comparable to each other, we replace ∞ of the upper interval with a value of 3 times the value of the lower bound. For instance if the interval is $(4000, \infty)$ we replace the upper bound with $4000 * 3 = 12000$, resulting in the interval $(4000, 12000]$ which is used by the KDE-algorithm. In an application the practitioner should choose the interval bounds for open ended intervals with caution, with regard to content and to the censoring scheme. However, our experiences running several simulations indicate that a value of 3 times the value of the lower bound serves as good approximation when working with interval censored income data.

4 Estimating poverty and inequality indicators from the German Micro-Census

In this Section, the KDE-algorithm is applied to the problem of estimating poverty and inequality indicators from interval censored German Microcensus data. The relevance of poverty and inequality estimation becomes apparent when considering the rich amount of literature available on this topic. Germany's increasing inequality has sparked the interest of many scholars as well as governmental institutions. Known for stable wages in the 70s and 80s (Abraham and Houseman, 1995), Germany has faced growing income inequality since its reunification in 1990 (Fuchs-Schündeln et al., 2010; Bönke et al., 2014).

Most of these studies consider or focus on survey data such as the Socio-Economic Panel (SOEP) or the Income, Receipts, Expenditure survey (in German: Einkommens- und Verbrauchsstichprobe) (EVS). In contrast to the Microcensus, the participation is voluntary and participants are asked for their exact income (not interval censored), which enables the estimation of poverty and inequality indicators using standard formulas. However, since the German Microcensus is by far the biggest survey in Germany it would be favourable to use its data for the estimation of poverty and inequality. The proposed KDE-algorithm makes the valid and precise estimation of complex poverty and inequality indicator from interval censored data possible. This allows researchers and practitioners to use the German Microcensus for the further and more in depth investigation of the increasing income inequality in Germany. The following application presents estimation results from cross-sectional data for the year 2012. To investigate the spatial distribution of inequality, the different indicators are estimated for the 16 federal states.

4.1 Data and preparation

The German Microcensus is a representative household survey conducted by the Federal Statistical Office of Germany. About 1% of the German population is chosen randomly by a specified survey design and is asked about the living conditions. The Microcensus was first conducted in 1957 and provides data regarding the structure as well as the economic and social status of the population. Over the years the Microcensus has become one of the most important data sources regarding aspects such as partnership, family, labour market and eduction. For the estimation of poverty and inequality the variable household net income is used. For the analysis the Scientific-Use-File (SUF), a 70% sample of the Microcensus is used (Statistisches Bundesamt, 2017). After data cleaning, we are left with a sample size of $n_{Germany} = 454852$. Since interests also lie in the spatial distribution of poverty and inequality the statistical indicators are estimated for each federal state separately and for Germany. The sample size for each federal state and its location is given in Table 8 and Figure 7 in the Appendix. The sample sizes are very large for each federal state even for Bremen, the state with the smallest sample size $n_{Bremen} = 3356$. Thus, there are enough observations to directly (without covariates) estimate the statistical indicators with small standard errors. As previously mentioned, the variable household income is interval censored to 24 intervals. The distribution is visualized in Figure 1 in the upper left panel. To make the household income comparable between households of different sizes, the OECD household weights are used to estimate equivalised household income. Each households interval bound is divided by its corresponding OECD weight. For instance a household within interval $(1300, 1500]$ and with an OECD weight of 1.5 has equivalence interval bounds of $(867, 1000]$.

4.2 Estimation and results

In order to estimate the poverty and inequality indicators, the KDE-algorithm is applied to the equivalenced interval bounds. The open ended interval is handled as described in Section 3.4. Further, for representative results the extrapolation factors of the Microcensus are used for the estimation of the weighted statistical indicators (formulas are given in Equation 1-6). Therefore, the KDE-algorithm draws iteratively new metric pseudo samples plus the corresponding extrapolation weight from the equivalenced interval censored household income. As in the simulations, the number of burn-in iterations is $B_{(KDE)} = 80$, the number of additional iterations is $S_{(KDE)} = 400$ and the number of grid points $j = 4000$. The number of $B_{(KDE)}$ and $S_{(KDE)}$ is sufficiently large as it is seen in the convergence plot in Figure 4. Both indicators have converged after 480 iterations. While indicators that dependent on the whole distribution converge slower (e.g. Gini), indicators that do not depend on the whole distribution (e.g. HCR) converge faster. Also all other indicators are checked for convergence, but only two plots are shown exemplarily. The standard errors of the weighted indicators are estimated by the described non-parametric bootstrap as proposed by Alfons and Templ (2013). The number of bootstrap samples is set to $B = 100$ as in the simulation study.

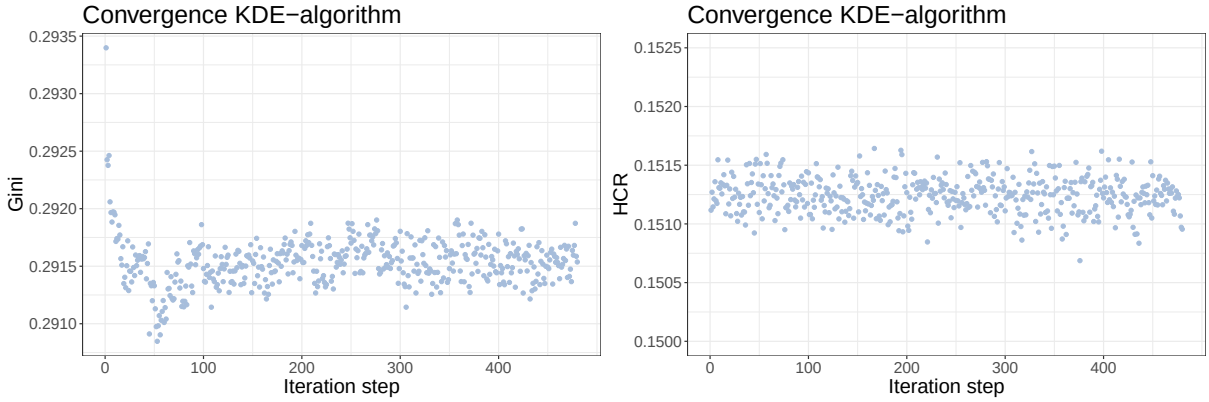
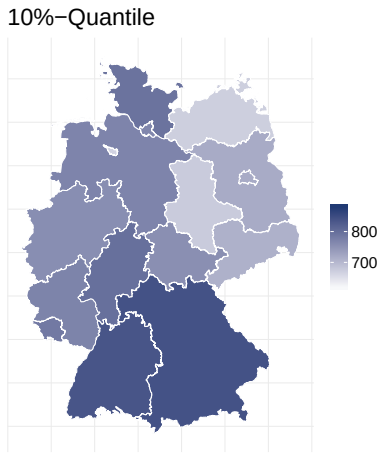


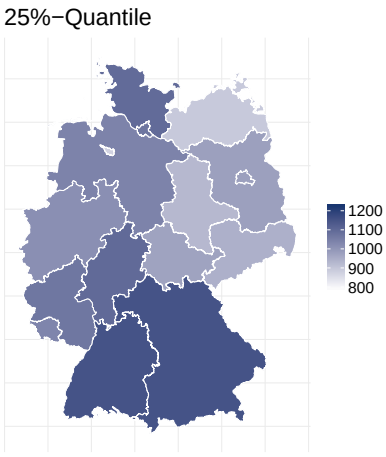
Figure 4: Convergence of the KDE-algorithm for the Gini and the HCR

The estimated indicators are presented in Figure 5 and 6 and the exact values and the estimated standard errors are given in the Appendix in Table 9. The estimated HCR $= 0.15$, the Gini $= 0.29$ and the QSR $= 4.31$. These results are comparable to the results from the EVS. The EVS reports the following values: HCR $= 0.16$, the Gini $= 0.27$ and the QSR $= 4.1$ (Statistisches Bundesamt, 2018). Because of the large sample size, valid estimates for smaller geographical areas can be estimated to evaluate the regional distribution of poverty and inequality in Germany. The quantiles and the mean indicate that the east (formerly German Democratic Republic DDR) is poorer than the west. This result is commonly known in Germany and not very surprising. Nevertheless, Brandenburg and Berlin have higher incomes than the rest of east Germany (Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia). Also Bremen, a federal state in the west, shows low income for the 10% and 25%-Quantile in comparison to the rest of west Germany, while for the higher quantiles Bremen shows similar results as the rest of Germany. The poorest states with a median of 1211.29 Euro and 1247.05 Euro are Mecklenburg-Vorpommern and Saxony-Anhalt and the richest ones with a median of 1580.43 Euro and 1580.35 Euro are Baden-Württemberg and Bavaria. For the estimation of the HCR and PGap, a regional poverty line (60% of the median) is used. The HCR indicates that in the east fewer people live under the regional poverty line than in the west. Also the people living under the poverty line live closer to it in the east, as shown by the PGap. When looking at the QSR and the Gini, the east-west trend is less striving. Nevertheless, the states in the east have lesser income inequality. The most unequal states with a Gini of 0.32 and 0.31 are Hamburg and Bremen and the most equal ones with a Gini of 0.25 and 0.25 are Saxony and Thuringia. The estimated standard errors of the indicators on state area are quite small. Therefore, estimating precise indicators for smaller geographical areas would probably also be possible, to get an even closer look at the geographical distribution of poverty and inequality.

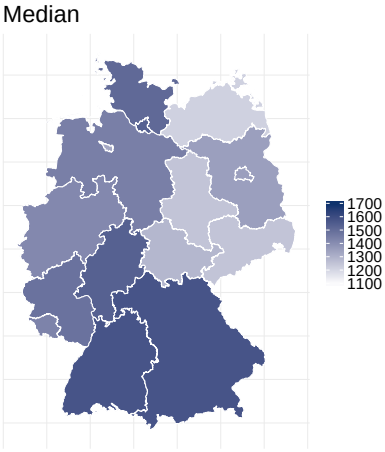
The application impressively demonstrates how the KDE-algorithm enables the estimation of poverty and inequality indicators from interval censored data. The precise estimations obtained by the KDE-



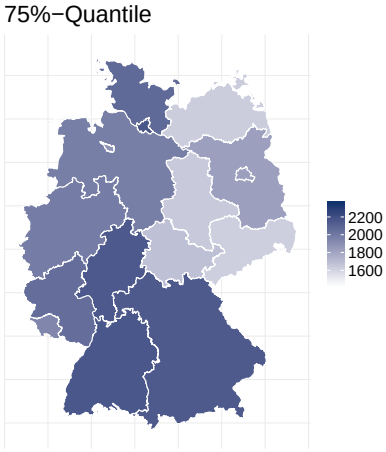
(a) Regional distribution of the 10%-Quantile



(b) Regional distribution of the 25%-Quantile



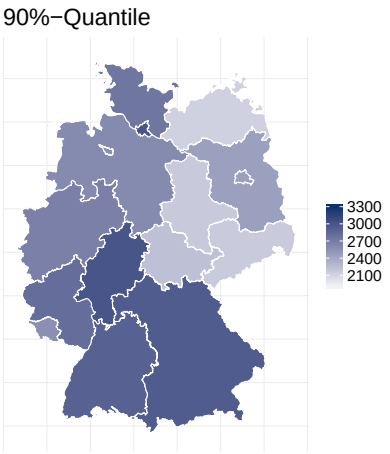
(c) Regional distribution of the Median



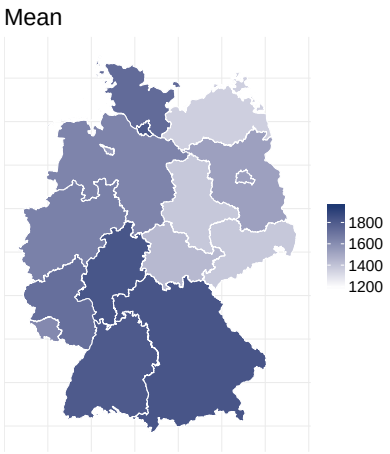
(d) Regional distribution of the 75%-Quantile

Figure 5: Regional distribution of different statistical indicators in Germany

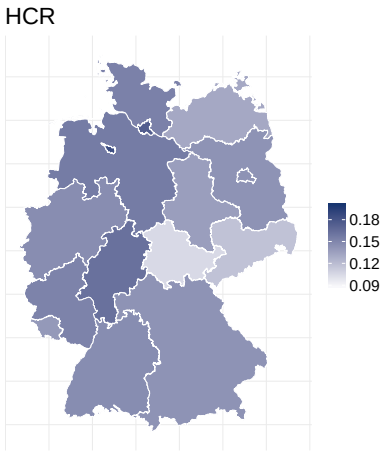
algorithm enable statisticians and statistical offices to report a variety of poverty and inequality indicators using the German Microcensus. The regional estimates will help to identify regions with lower income and higher inequality to target political activities more accurately for those in need.



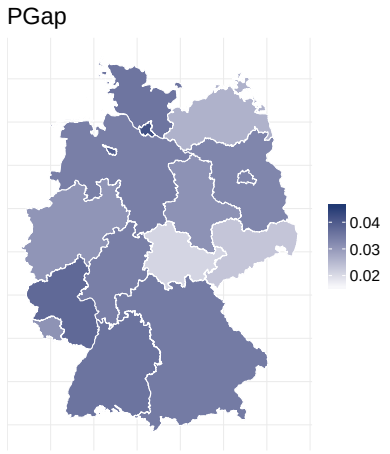
(a) Regional distribution of the 90%-Quantile



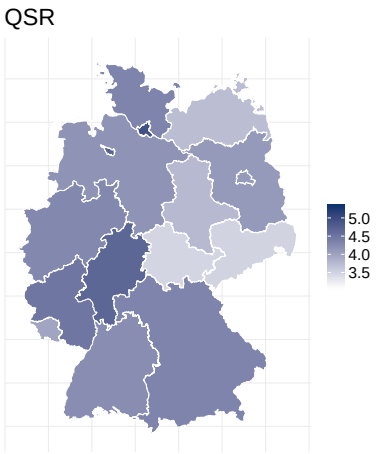
(b) Regional distribution of the Mean



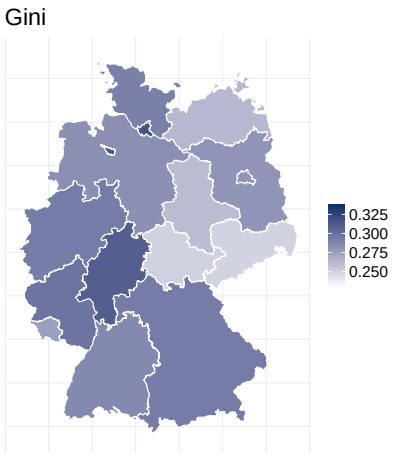
(c) Regional distribution of the HCR



(d) Regional distribution of the PGap



(e) Regional distribution of the QSR



(f) Regional distribution of the Gini

Figure 6: Regional distribution of different statistical indicators in Germany

5 Discussion and Outlook

In numerous censuses e.g., the German Microcensus or the Australian Census, the variable household or personal income is not observed on a continuous scale, but rather censored to specific intervals. This is due to confidentiality constraints or to reduce item non-response. Estimating poverty and inequality indicators from these kind of data requires more sophisticated statistical methods. As an estimation method we propose an iterative KDE-algorithm that enables the precise estimation of statistical indicators from interval censored data. The proposed KDE-algorithm has similarities to SEM-algorithms that are commonly used for the estimation of models that depend on latent unobserved variables (in our case the interval censored income). However, instead of maximizing the likelihood as it is common for SEM-algorithms, the asymptotic mean integrated squared error of the KDE is maximized. For the estimation of the standard errors of the statistical indicators a non-parametric bootstrap is proposed. The KDE-algorithm and the bootstrap work for different censoring scenarios and different underlying true distributions. The methodology is available in the R-Package `smicd` from the Comprehensive R Archive Network (Walter, 2018). Our simulation results demonstrate that the estimated poverty and inequality indicators outperform other estimation techniques (linear interpolation or the use of the midpoints of the intervals) in terms of bias. Also the standard errors of the estimates are close to the standard errors from the estimates that were obtained with the uncensored data, supporting the precision of the algorithm. Furthermore, the KDE-algorithm has the advantage of adapting to different interval censored theoretical distributions. Therefore, it is universally applicable for the estimation of poverty and inequality indicators from interval censored income data. We demonstrate the usefulness by estimating regional poverty and inequality indicators from the German Microcensus. To get representative results the algorithm is extended to take OECD equivalence weights and survey weights into account. The estimated regional indicators are plotted on maps that visualize the magnitude of poverty and inequality in Germany. With help of the KDE-algorithm statistical indicators can precisely be estimated from interval censored data in order to tackle the increasing problem of rising poverty and inequality in societies all over the world.

Further research should focus on convergence criteria that make the manual choice of the number of iteration obsolete.

Acknowledgements

We thank Timo Schmid and Marcus Groß for discussions and helpful comments on this paper.

6 Appendix A

State	Sample size	Number in Map
Germany	454852	
Schleswig-Holstein	15302	1
Hamburg	8630	2
Lower Saxony	45828	3
Bremen	3356	4
North Rhine-Westphalia	90778	5
Hesse	35730	6
Rhineland-Palatinate	21229	7
Baden-Württemberg	58685	8
Bavaria	75244	9
Saarland	5688	10
Berlin	19311	11
Brandenburg	15400	12
Mecklenburg-Vorpommern	8706	13
Saxony	24609	14
Saxony-Anhalt	13495	15
Thuringia	12861	16

Table 8: Sample size for Germany and each of the 16 federal states.

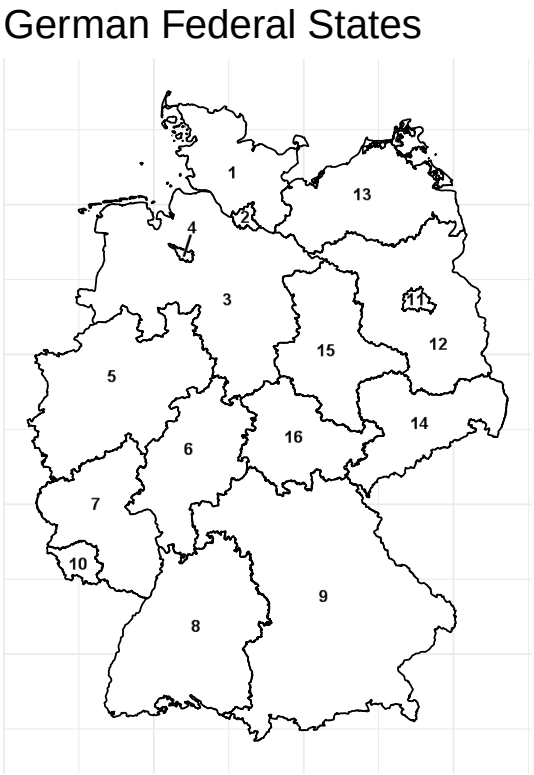


Figure 7: German Federal States, the names of the corresponding numbers are given in Table 8.

	Quant0.1	Quant0.25	Median	Quant0.75	Quant0.9	Mean	HCR	QSR	PGap	Gini
Germany	770.16	1040.23	1445.53	1998.96	2714.63	1675.88	0.15	4.31	0.03	0.29
	(0.00)	(3.28)	(4.93)	(2.59)	(3.69)	(1.85)	(0.00)	(0.01)	(0.00)	(0.00)
Schleswig-Holstein	794.21	1092.99	1512.96	2071.11	2743.79	1736.25	0.15	4.33	0.04	0.29
	(6.04)	(5.83)	(7.39)	(7.80)	(20.01)	(9.35)	(0.00)	(0.05)	(0.00)	(0.00)
Hamburg	765.68	1069.24	1540.20	2166.83	3002.79	1815.45	0.17	4.92	0.04	0.32
	(7.16)	(9.21)	(10.47)	(13.44)	(29.75)	(14.14)	(0.00)	(0.09)	(0.00)	(0.00)
Lower Saxony	770.08	1040.25	1445.04	1970.84	2603.04	1636.36	0.16	4.16	0.03	0.28
	(4.10)	(2.53)	(5.44)	(7.11)	(13.00)	(5.79)	(0.00)	(0.04)	(0.00)	(0.00)
Bremen	665.44	876.91	1328.23	1879.66	2564.10	1540.03	0.18	4.72	0.04	0.31
	(10.82)	(9.93)	(16.76)	(25.28)	(47.95)	(19.19)	(0.01)	(0.12)	(0.00)	(0.01)
North Rhine-Westphalia	756.85	1013.41	1418.50	1985.64	2674.27	1649.22	0.15	4.29	0.03	0.29
	(0.72)	(0.01)	(3.01)	(4.70)	(9.05)	(3.69)	(0.00)	(0.02)	(0.00)	(0.00)
Hesse	798.23	1094.75	1540.36	2149.61	2997.03	1825.06	0.16	4.66	0.03	0.31
	(4.56)	(4.88)	(6.03)	(7.49)	(14.23)	(7.54)	(0.00)	(0.04)	(0.00)	(0.00)
Rhineland-Palatinate	770.65	1067.23	1485.95	2052.43	2810.09	1720.30	0.15	4.49	0.04	0.30
	(5.39)	(5.36)	(6.36)	(8.97)	(17.17)	(8.40)	(0.00)	(0.06)	(0.00)	(0.00)
Baden-Württemberg	837.76	1148.33	1580.43	2160.84	2900.98	1806.40	0.15	4.24	0.04	0.29
	(2.23)	(4.10)	(4.08)	(6.62)	(10.50)	(5.52)	(0.00)	(0.03)	(0.00)	(0.00)
Bavaria	841.94	1148.28	1580.35	2147.52	2944.04	1826.62	0.14	4.33	0.03	0.29
	(5.37)	(4.62)	(4.99)	(5.19)	(8.81)	(4.75)	(0.00)	(0.03)	(0.00)	(0.00)
Saarland	784.70	1035.41	1434.20	1938.86	2559.91	1615.95	0.14	3.98	0.03	0.27
	(8.77)	(9.73)	(10.55)	(14.70)	(32.54)	(13.40)	(0.01)	(0.07)	(0.00)	(0.00)
Berlin	730.96	912.14	1328.50	1867.05	2552.45	1547.47	0.15	4.15	0.02	0.29
	(5.38)	(5.19)	(8.41)	(11.20)	(15.87)	(8.77)	(0.01)	(0.05)	(0.00)	(0.00)
Brandenburg	716.80	979.24	1351.02	1823.73	2446.72	1528.25	0.14	4.10	0.03	0.28
	(5.93)	(6.78)	(6.43)	(10.03)	(17.36)	(8.08)	(0.00)	(0.05)	(0.00)	(0.00)
Mecklenburg-Vorpommern	671.30	895.52	1211.29	1629.61	2120.56	1355.61	0.13	3.74	0.03	0.26
	(4.92)	(5.77)	(7.51)	(9.78)	(20.77)	(11.96)	(0.00)	(0.08)	(0.00)	(0.01)
Saxony	709.57	945.88	1247.40	1622.64	2155.08	1383.20	0.12	3.52	0.02	0.25
	(4.62)	(4.00)	(4.31)	(5.48)	(10.93)	(5.09)	(0.00)	(0.03)	(0.00)	(0.00)
Saxony-Anhalt	675.70	928.47	1247.05	1643.78	2161.33	1382.23	0.14	3.78	0.03	0.26
	(5.25)	(5.85)	(5.90)	(7.68)	(17.09)	(6.24)	(0.00)	(0.05)	(0.00)	(0.00)
Thuringia	755.17	973.23	1283.80	1683.44	2226.40	1435.52	0.11	3.50	0.02	0.25
	(5.32)	(4.23)	(5.50)	(7.11)	(16.00)	(7.45)	(0.00)	(0.04)	(0.00)	(0.00)

Table 9: Estimated statistical indicators for Germany and the 16 federal states. Standard errors are given in parentheses.

References

Abraham, K. and Houseman, S. (1995). Earnings inequality in Germany. In Freeman, R. B. and Katz, L. F., editors, *Differences and Changes in Wage Structures*, pages 371–404. Nber Comparative Labor Markets.

Alfons, A. and Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: the R package laeken. *Journal of Statistical Software*, 54(15):1–25.

Australian Bureau of Statistics (2011). Census household form. `https://unstats.un.org/unsd/demographic/sources/census/quest/AUS2011en.pdf`. Accessed: 2018-04-05.

Bandourian, R., McDonald, J., and Turley, R. (2003). Income distributions: an inter-temporal comparison over countries. *Estadistica*, 55(1):135 – 152.

Bandourian, R., McDonald, J., and Turley, R. S. (2002). A comparison of parametric models of income distribution across countries and over time. Technical report, Luxembourg Income Study.

Bönke, T., Corneo, G., and Lüthen, H. (2014). Lifetime earnings inequality in Germany. *Journal of Labor Economics*, 33(1):171–208.

Bordley, R. F., McDonald, J. B., and Mantrala, A. (1997). Something new, something old: parametric models for the size of distribution of income. *Journal of Income Distribution*, 6(1):91 – 103.

Bruch, C., Münnich, R., and Zins, S. (2011). Variance estimation for complex surveys. Technical report, European Commission.

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314.

Celeux, G. and Dieboldt, J. (1985). The SEM algorithm: a probalistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82.

Chen, Y.-T. (2017). A unified approach to estimating and testing income distributions with grouped data. *Journal of Business & Economic Statistics*, pages 1–18.

Chotikapanich, D., Griffiths, W. E., and Rao, D. S. P. (2007). Estimating and combining national income distributions using limited data. *Journal of Business & Economic Statistics*, 25(1):97–109.

Dagum, C. (1977). A new model of personal income distribution: specification and estimation. *Economie Appliquee*, 30:413–437.

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

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of berkson and classical errors. *Canadian Journal of Statistics*, 35(1):89–104.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

Departamento Administrativo Nacional De Estadistica (2005). Censo general 2005. `https://www.dane.gov.co/files/censos/libroCenso2005nacional.pdf?&`. Accessed: 2018-04-05.

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

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

eurostat (2014). Statistics explained: At-risk-of-poverty rate. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:At-risk-of-poverty_rate. Accessed: 2018-05-30.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52(3):761–766.

Fuchs-Schündeln, N., Krueger, D., and Sommer, M. (2010). Inequality trends for Germany in the last two decades: a tale of two countries. *Review of Economic Dynamics*, 13(1):103–132.

Gini, C. (1912). *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari. Tipogr. di P. Cuppini.

Graf, M. and Nedyalkova, D. (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*, 60(4):821–842.

Groß, M. and Rendtel, U. (2016). Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology*, 4(3):339–361.

Groß, M., Rendtel, U., Schmid, T., Schmon, S., and Tzavidis, N. (2017). Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. *Journal of the Royal Statistical Society: Series A*, 180(1):161–183.

Hagenaars, A. and Vos, K. D. (1988). The definition and measurement of poverty. *Journal of Human Resources*, 23(2):211–221.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM Journal on Applied Mathematics*, 42(2):390–399.

Hall, P. and Wand, M. P. (1996). On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis*, 56(2):165–184.

Henderson, D. J. and Parmeter, C. F. (2015). *Applied Nonparametric Econometrics*. Cambridge University Press, New York.

Information und Technik (NRW) (2009). Berechnung von Armutsgefährdungsquoten auf Basis des Mikrozensus. http://www.amtliche-sozialberichterstattung.de/pdf/Berechnung%20von%20Armutsgefaehrdungsquoten_090518.pdf. Accessed: 2018-04-09.

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

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407.

Kakwani, N. C. and Podder, N. (2008). Efficient estimation of the lorenz curve and associated inequality measures from grouped observations lorenz curve and associated inequality measures from grouped observations. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, pages 57–70. Springer.

Kleiber, C. (2008). A guide to the dagum distributions lorenz curve and associated inequality measures from grouped observations. In Chotikapanich, D., editor, *Modelig Income Distributions and Lorenz Curves*, pages 97–117. Springer.

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

Lenau, S. and Münnich, R. (2016). Estimating income poverty and inequality from income classes. In Münnich, R., editor, *InGRID Integrating Expertise in Inclusive Growth: Case Studies*, pages 60–90.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics*, 27(2):415–438.

Lok-Dessallien, R. (1999). Review of poverty concepts and indicators. Technical report, United Nations Development Programme.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663.

McDonald, J. B. and Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, pages 147–166. Springer.

McDonald, J. B. and Ransom, M. R. (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47(6):1513–1525.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1):133 – 152.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley & Sons, New York.

Milanovic, B. (2003). The ricardian vice: why sala-i-Martins calculations of world income inequality are wrong. Technical report, EconWPA.

Minoiu, C. and Reddy, S. (2008). Kernel density estimation based on grouped data: the case of poverty assessment. Technical report, International Monetary Fund.

Moore, J. C. and Welniak, E. J. (2000). Income measurement error in surveys: a review. *Journal of Official Statistics*, 16(4):331.

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

Nielsen, S. F. et al. (2000). The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3(3):167–195.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pretson, J. (2008). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, 35(2):227–234.

Rao, J., Wu, C., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18(2):209–217.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

Reed, W. J. and Wu, F. (2008). New four- and five-parameter models for income distributions. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, pages 211–224. Springer.

Rosenblatt, M. e. a. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Scott, D. W. and Sheather, S. J. (1985). Kernel density estimation with binned data. *Communications in Statistics - Theory and Methods*, 14(6):1353–1359.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics, New York.

Singh, S. and Maddala, G. (1976). A function for the size distribution of incomes. *Econometrica*, 44(5):963–970.

Statistics New Zealand (2013). New Zealand census of population and dwellings. https://unstats.un.org/unsd/demographic/sources/census/quest/NZL2013enIn.pdf. Accessed: 2018-05-13.

Statistisches Bundesamt (2017). Datenhandbuch zum Mikrozensus scientific use file 2012. http://www.forschungsdatenzentrum.de/bestand/mikrozensus/suf/2012/fdz_mz_suf_2012_schluesselverzeichnis.pdf. Accessed: 2017-07-22.

Statistisches Bundesamt (2018). Wirtschaftsrechnungen: Einkommens- und Verbrauchsstichsprobe Einkommensverteilung in Deutschland. https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/EinkommenVerbrauch/Einkommensverteilung2152606139004.pdf?__blob=publicationFile. Accessed: 2018-05-22.

Tepping, B. (1968). Variance estimation in complex surveys. *Proceedings of the American Statistical Association, Social Statistics Section*, pages 11–18.

Tille, Y. (2001). *Theorie des sondages: Echantillonnage et estimation en populations finies*. Dunod, Paris.

Walter, P. (2018). *smicd: Statistical Methods for Interval Censored Data*. R package version 1.0.0.

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing*. R package version 2.23-15.

Wand, M. and Jones, M. (1995). *Kernel smoothing*. Chapman & Hall, London.

Wang, B. and Wertelecki, M. (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65:4–12.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer, New York.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334):411–414.

World Economic Forum (2017). Global risks 2017. http://reports.weforum.org/global-risks-2017/part-1-global-risks-2017/. Accessed: 2017-09-28.

Zambom, A. Z. and Dias, R. (2012). A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2018 erschienen:

2018/1 BESTER, Helmut und Ouyang YAOFU
Optimal Procurement of a Credence Good under Limited Liability
Economics

2018/2 GROß, Markus, Ulrich RENDTEL, Timo SCHMID und Nikos TZAVIDIS
Switching between different area systems via simulated geo-coordinates: a case study for student residents in Berlin
Economics

2018/3 GROß, Markus, Ulrich RENDTEL, Timo SCHMID, Hartmut BÖMERMANN und Kerstin ERFURTH
Simulated geo-coordinates as a tool for map-based regional analysis
Economics

2018/4 KONYUKHOVSKIY, Pavel V. und Theocharis GRIGORIADIS
Proxy Wars
Economics

2018/5 FOX, Jonathan F. und Theocharis GRIGORIADIS
A Rural Health Supplement to the Hookworm Intervention in the American South
Economics

2018/6 VITOLAS,Alise und Theocharis GRIGORIADIS
Diversity & Emipre: Baltic Germans & Comparative Development
Economics

2018/7 GAENTZSCH, Anja
The distributional impact of social spending in Peru
Economics

2018/8 SCHREIBER, Sven
Are bootstrapped conintegration test findings unreliable?
Economics

2018/8 SCHREIBER, Sven
Are bootstrapped conintegration test findings unreliable?
Economics

2018/9 GRIGORIADIS, Theocharis
Aristotle vs. Plato: The Distributive Origins of the Cold War
Economics